Globally Convergent Derivative-Free Methods in Nonconvex Optimization with and without Noise

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

July 1, 2024

Abstract. This paper addresses the study of nonconvex derivative-free optimization problems, where only information of either smooth objective functions or their noisy approximations is available. General derivative-free methods are proposed for minimizing differentiable (not necessarily convex) functions with globally Lipschitz continuous gradients, where the accuracy of approximate gradients is interacting with stepsizes and exact gradient values. Analysis in the noiseless case guarantees convergence of the gradient sequence to the origin as well as global convergence with constructive convergence rates of the sequence of iterates under the Kurdyka-Łojasiewicz property. In the noisy case, without any noise level information, the designed algorithms reach near-stationary points with providing estimates on the required number of iterations and function evaluations. Addressing functions with locally Lipschitzian gradients, two algorithms are introduced to handle the noiseless and noisy cases, respectively. The noiseless version is based on the standard backtracking linesearch and achieves fundamental convergence properties similarly to the global Lipschitzian case. The noisy version is based on a novel bidirectional linesearch and is shown to reach near-stationary points after a finite number of iterations when the Polyak-Łojasiewicz inequality is imposed. Numerical experiments are conducted on a diverse set of test problems to demonstrate more robustness of the newly proposed algorithms in comparison with other finite-difference-based schemes and some highly efficient, production-ready codes from the SciPy library.

Key words: derivative-free optimization, nonconvex smooth objective functions, finite differences, blackbox optimization, noisy optimization, zeroth-order optimization, globally convergent algorithms

Mathematics Subject Classification (2020) 90C25, 90C26, 90C30, 90C56

1 Introduction

This paper is devoted to the development of novel *derivative-free methods* of solving unconstrained optimization problems given in the form

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

where $f : \mathbb{R}^n \to \mathbb{R}$ is a continuously differentiable ($\mathcal{C}^1$-smooth) function, not necessarily convex. In the context of derivative-free optimization, we assume that only information of either $f(x)$ (noiseless case) or its *noisy approximation* $\phi(x) = f(x) + \xi(x)$ (noisy case) is available, where $\xi(x) : \mathbb{R}^n \to \mathbb{R}$ is the noise function bounded by a positive constant ξ_f. These problems have received much attention with a variety of methods being developed over the years [5, 16]. The major developments in this vein are provided by the *Nelder-Mead simplex method* [42, 45], *direct search methods* [29, 32], *conjugate direction method* [50], *trust-region methods* [17, 50], and *finite-difference-based methods* [6, 8, 46, 56, 58]. Applications of derivative-free optimization methods [2, 5, 16] have also gained a lot of interest since many efficient methods, including Nelder-Mead, Powell (a short name of Powell's conjugate direction method) [50], COBYLA [51], and L-BFGS-B, are implemented as production-ready codes in SciPy [59], a well-known Python library. More recently, numerous empirical results conducted by Shi et al. in [58] show that derivative-free optimization methods based on *finite differences* are accurate, efficient, and in some cases superior to other state-of-the-art derivative-free optimization methods developed in the literature. Meanwhile, extensive numerical comparisons in Berahas et al. [7], together with further analysis by Scheinberg [54], also tell us that the accuracy of gradients obtained from standard finite differences is significantly higher than from *randomized schemes* [19, 26, 27, 46]. These empirical results suggest that

*Group of Analysis and Applied Mathematics, Department of Mathematics, Ho Chi Minh City University of Education, Ho Chi Minh City, Vietnam. E-mails: pdkhanh182@gmail.com, khanhpd@hcmue.edu.vn

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

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

the methods using standard finite differences have much to be recommended, and that the research in this direction should be strongly encouraged.

When no error is present within the function evaluations, implementing finite difference approximations for gradient descent methods is rather simple because it is possible to use a fixed, sufficiently small finite difference interval (referred to as *GD (fixed)* for the sake of brevity). However, dealing with noisy problems is more challenging since finding the optimal finite difference interval requires not only the *noise level* information but also the higher-order derivatives of the function that are often unavailable. Hence this topic attracts many studies, which develop finite-difference-based methods under different types of noise. Kelley et al. [15, 24, 36] proposed the *implicit filtering* algorithm based on a finite difference approach to deal with *noisy smooth box-constrained* optimization problems with the noise being decayed near local minimizers. Berahas et al. [6, 8] developed finite-difference-based linesearch methods for the minimization of *smooth functions* with *bounded noise*. The schemes to adapt the *finite difference intervals* were also studied by Gill et al. [23], Moré and Wild [44], and recently by Shi et al. [56, 57].

Motivations. Although methods of this type are often used in practice to solve derivative-free smooth problems with and without noise, there are still some significant concerns related to their theoretical and practical developments that should be addressed.

- *Analysis in the noiseless case*: In the noiseless case, due to the usage of a fixed finite difference interval, GD (fixed) methods do not obtain sufficient convergence properties compared to standard gradient descent methods. These properties include the *stationarity of accumulation points* and the *convergence* of the *sequence of iterates* to *nonisolated stationary points* under the *Kurdyka-Łojasiewicz* (KL) condition [3, 37, 38, 39, 41, 43], which is a rather mild regularity condition satisfied for the vast majority of objective functions in practice. Note that simply letting the finite difference interval approach zero is not a favorable approach, since without careful adaptations the approximate gradient obtained may not be even a descent direction.
- *Dealing with small noise without any noise level information*: The practical implementations of finite-difference-based algorithms also face issues in this case since choosing sufficiently small finite difference intervals makes GD (fixed) methods perform *poorly* due to the roundoff error, while using an adaptive scheme as in [23, 44, 56] becomes *impossible* since the noise level is unknown.
- *Assumption on the gradient global Lipschitz continuity, i.e., the $\mathcal{C}_L^{1,1}$ property of objective functions*: This assumption seems to be omnipresent in derivative-free linesearch methods; see, e.g., [15, Theorem 2.1], [6, Assumption A1], and [8, Assumption 1.1]. For general derivative-free trust-region methods, Conn et al. [17] proved global convergence results for $\mathcal{C}_L^{1,1}$ minimization problems under the additional assumption on the Lipschitz continuity of the Hessian of the objective function. Such properties were also employed in the proximal point method adapted to derivative-free smooth optimization problems by Hare and Lucet [30, Assumption 1]. In [55], the class of smooth functions with Hölderian gradients, being larger than the class of $\mathcal{C}_L^{1,1}$ functions, was investigated. However, we are not familiar with any efficient finite-difference-based method considering specifically the class of smooth functions with *locally Lipschitzian gradients*, i.e., the class of $\mathcal{C}^{1,1}$ functions, which is much broader than the class of $\mathcal{C}_L^{1,1}$ ones. This is in contrast to the exact versions of gradient descent methods that obtain various convergence properties including the stationarity of accumulation points for the version with backtracking stepsizes in the class of $\mathcal{C}^1$-smooth functions [9, Proposition 1.2.1] and the global convergence for the version with sufficiently small stepsizes in the class of definable $\mathcal{C}^{1,1}$ functions [34]. This raises the need for the design and analysis of finite-difference-based methods concerning the class of $\mathcal{C}^{1,1}$ functions, which is the best we can hope in this context since the error bounds for finite differences are not available outside of this class.

Contributions. Having in mind the above discussions, we first address $\mathcal{C}_L^{1,1}$ optimization problems and propose the *derivative-free method with constant stepsize* (DFC). The method offers *generalizations* and *improvements* in both theoretical and practical aspects compared to GD (fixed) algorithms. The *generalizations* imply that other gradient approximation methods can be employed in DFC besides finite differences, provided that the approximation methods adhere to general conditions outlined in Definition 3.1. The *improvements* of DFC in comparison with GD (fixed) algorithms are as follows.

- In the *theoretical aspects*, DFC comes with a detailed convergence analysis, which is presented in Section 4 and contains the following.
  - In the *noiseless* case, our analysis establishes the *convergence to the origin* for the gradient sequence, *global convergence* of iterates under the KL property, and *constructive convergence rates* depending on the KL exponents. Note that none of these properties can be achieved by GD (fixed) algorithms.
  - In the *noisy case*, the *finite convergence* of the sequence of iterates to a *near-stationary point* is established, along with estimates on the number of iterations and function evaluations

needed to reach the near-stationary point. The construction of DFC and all of its convergence properties in the noisy case do *not require* the knowledge of *noise levels*, although it does require some mild conditions for initialization.

- In the *practical aspects*, DFC achieves at least similar, or even better, numerical performance in comparison with GD (fixed) methods in the *noiseless* case being more efficient in the presence of small noise with an *unknown noise level.* These numerical results are presented in Section 6.

Note that the main feature in the algorithmic constructions of DFC, which allows us to achieve the above goals, is the *adaptivity* of the finite difference interval. Contrarily to using a fixed small interval as in GD (fixed) methods, we start with a *much larger* finite difference interval and decrease it along the sequence of iterates if a descent condition is not satisfied. The finite difference interval in DFC also interacts with the approximate Lipschitz constant (or equivalently, the stepsize), which creates more robustness for the algorithm. This interaction distinguishes DFC from the methods in [23, 44, 56, 57], where the finite difference interval is constructed independently from the stepsize. By adopting this approach, we are able to theoretically derive the *fundamental convergence properties* of DFC for both noiseless and noisy functions. Practically this approach helps DFC *avoiding roundoff errors* as much as possible to ensure the quality of the gradient approximation, which leads us to the numerical performance highlighted above.

Next we address the class of $\mathcal{C}^{1,1}$ functions. Due to the complex structure of functions in this class, we introduce two different algorithms as follows.

- DFB: *Derivative-free method with backtracking linesearch* to deal with noiseless problems and problems with small noise. This algorithm is inspired by DFC, with the primary difference that a backtracking linesearch step is performed in *each iteration*, similarly to the standard approach of gradient descent methods when dealing with $\mathcal{C}^{1,1}$ functions. The analysis is conducted in the noiseless case and establishes the *stationarity of the accumulation point*, *global convergence* under the KL property, and *constructive convergence rates* depending on the KL exponents.
- DFBD: *Derivative-free method with bidirectional linesearch* to deal with problems with large noise and known noise level. To be more specific, in each iteration DFBD uses a bidirectional linesearch to approximate the local Lipschitz constant of the gradient in a region around the current iterate. The approximate Lipschitz constant is then used for determining both the stepsize and the finite difference interval. By employing this approach, DFBD exhibits more favorable numerical behavior compared to other finite difference schemes [24, 46, 56] as demonstrated in Figure 2. This example also shows that the standard backtracking linesearch does not work well for functions of $\mathcal{C}^{1,1}$ class with *large noise*, which is the main motivation for us to implement the bidirectional linesearch in DFBD. The global analysis of DFBD demonstrates that the algorithm always makes progress whenever the gradient at the current iterate is not near the origin. It is also established that the sequence of iterates finds a near-stationary point after a *finite number of iterations*, with a constructive estimate given when the *Polyak-Łojasiewicz inequality* (cf. [49] and [43]) is satisfied.

To demonstrate the practical aspects of our study, *extensive numerical experiments* on synthetic problems with and without noise are conducted in Section 6. The results when the noise is small show that DFC and DFB methods do improve the performance of GD (fixed) methods as well as the performance of the implicit filtering (IMFIL) algorithm [24] and random gradient-free (RG) algorithm [46]. When the noise is large, our DFBD demonstrates its numerical reliability in comparison with SciPy [59] production-ready codes, including Powell, COBYLA and L-BFGS-B algorithms.

Related Works. The adaptivity of the finite difference interval to ensure the quality of the approximate gradient is a main feature employed in many finite-difference-based methods. Cartis and Scheinberg [13] analyzed a general linesearch algorithm for smooth functions without noise under the major condition that the gradient estimates are sufficiently accurate with a certain probability. This analysis is then extended in Berahas et al. [6] to the case where the function values are noisy. The practical schemes to choose the finite difference intervals adaptively based on testing ratios were also studied by Gill et al. [23], Shi et al. [56, 57], and heuristically by Moré and Wild [44]. The adaptivity in the selection of the finite difference interval is also related to the dynamic accuracy of gradient approximations, which is considered for adaptive regularization algorithms without noise in [12, 28] and with noise in [14, 11]. Among the aforementioned publications, [13] and [6] are the most related to our DFC development for $\mathcal{C}^{1,1}_L$ functions. These results are discussed in more detail in Remark 4.4 and Remark 4.10.

Organization. The rest of the paper is organized as follows. Section 2 presents some basic definitions and preliminaries used throughout the entire paper. Section 3 examines two types of gradient approximations that include finite differences. The main parts of our work, concerning the design and convergence properties of general derivative-free methods under the global and local Lipschitz continuity of the gradient, are given in Section 4 and Section 5, respectively. Numerical experiments, which compare the efficiency of the proposed methods with other derivative-free methods for both noisy and noiseless functions, are conducted in Section 6. Concluding remarks on the main contributions of this paper together with some open questions and perspectives of our future research are presented in Section 7.

2 Preliminaries

First we recall some basic notions and notation frequently used in the paper. All our considerations are given in the space $\mathbb{R}^n$ with the Euclidean norm $\|\cdot\|$. For any $i = 1, \ldots, n$, let e_i denote the i^{th} basic vector in $\mathbb{R}^n$. As always, $\mathbb{N} := \{1, 2, \ldots\}$ signifies the collection of natural numbers. For any $x \in \mathbb{R}^n$ and $\varepsilon > 0$, let $\mathbb{B}(x, \varepsilon)$ and $\overline{\mathbb{B}}(x, \varepsilon)$ stand for the open and closed balls centered at x with radius ε, respectively. When $x = 0$, these balls are denoted simply by $\varepsilon\mathbb{B}$ and $\varepsilon\overline{\mathbb{B}}$.

Recall that a mapping $G \colon \mathbb{R}^n \to \mathbb{R}^m$ is *Lipschitz continuous on a subset* D of $\mathbb{R}^n$ if there exists a constant $L > 0$ such that we have

$$\|G(x) - G(y)\| \leq L \|x - y\| \text{ for all } x, y \in D.$$

If $D = \mathbb{R}^n$, the mapping G is said to be *globally Lipschitz continuous*. The *local Lipschitz continuity* of G on $\mathbb{R}^n$ is understood as the Lipschitz continuity of this mapping on every compact subset of $\mathbb{R}^n$. The latter is equivalent to saying that for any $x \in \mathbb{R}^n$ there is a neighborhood U of x such that G is Lipschitz continuous on U. In what follows, we denote by $\mathcal{C}^{1,1}$ the class of $\mathcal{C}^1$-smooth mappings that have a *locally Lipschitz continuous gradient* on $\mathbb{R}^n$ and by $\mathcal{C}^{1,1}_L$ the class of $\mathcal{C}^1$-smooth mappings that have a *globally Lipschitz continuous gradient with the constant* $L > 0$ (i.e., L-Lipschitz continuous) on the entire space.

Our convergence analysis of the numerical algorithms developed in the subsequent sections largely exploits the following important results and notions. The first result taken from [33, Lemma A.11] presents a simple albeit very useful property of real-valued functions with Lipschitz continuous gradients.

Lemma 2.1. *Let $f : \mathbb{R}^n \to \mathbb{R}$, let $x, y \in \mathbb{R}^n$, and let $L > 0$. If f is differentiable on the line segment $[x, y]$ with its derivative being L-Lipschitz continuous on this segment, then*

$$|f(y) - f(x) - \langle \nabla f(x), y - x\rangle| \leq \frac{L}{2} \|y - x\|^2. \tag{2.1}$$

The second lemma established in [37, Section 3] is crucial in the convergence analysis of the general linesearch methods developed in this paper.

Lemma 2.2. *Let $\{x^k\}$ and $\{d^k\}$ be sequences in $\mathbb{R}^n$ satisfying the condition*

$$\sum_{k=1}^{\infty} \|x^{k+1} - x^k\| \cdot \|d^k\| < \infty. \tag{2.2}$$

If $\bar{x}$ is an accumulation point of $\{x^k\}$ and if the origin is an accumulation point of $\{d^k\}$, then there exists an infinite set $J \subset \mathbb{N}$ such that

$$x^k \xrightarrow{J} \bar{x} \text{ and } d^k \xrightarrow{J} 0. \tag{2.3}$$

Next we recall the classical results from [20, Section 8.3.1] that describe important properties of accumulation points generated by a sequence satisfying the limit condition introduced by Ostrowski [48].

Lemma 2.3. *Let $\{x^k\} \subset \mathbb{R}^n$ be a sequence satisfying the Ostrowski condition*

$$\lim_{k \to \infty} \|x^{k+1} - x^k\| = 0. \tag{2.4}$$

Then the following assertions are fulfilled:

(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 $\bar{x}$, then this sequence converges to $\bar{x}$.*

The version of the fundamental *Kurdyka-Łojasiewicz* (*KL*) *property* formulated below is taken from Absil et al. [1, Theorem 3.4].

Definition 2.4. 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

$$\|\nabla f(x)\| \geq \psi\big(f(x) - f(\bar{x})\big) \text{ for all } x \in U \text{ with } f(\bar{x}) < f(x) < f(\bar{x}) + \eta. \tag{2.5}$$

Remark 2.5. If f satisfies the KL property at $\bar{x}$ with a neighborhood U, it is clear that the same property holds for any $x \in U$ where $f(x) = f(\bar{x})$. It has been realized that the KL property is satisfied in broad settings. In particular, it holds at every *nonstationary point* of f; see [3, Lemma 2.1 and Remark 3.2(b)]. Furthermore, it is proved in the seminal paper by Łojasiewicz [43] that any analytic function $f : \mathbb{R}^n \to \mathbb{R}$ satisfies the KL property at every point $\bar{x}$ with $\psi(t) = Mt^q$ for some $q \in [0, 1)$. As demonstrated in [37, Section 2], the KL property formulated in Attouch et al. [3] is stronger than the one in Definition 2.4. Typical smooth functions that satisfy the KL property from [3], and hence the one from Definition 2.4, are smooth *semialgebraic* functions and also those from the more general class of functions known as *definable in o-minimal structures*; see [3, 4, 41]. The latter property is fulfilled, e.g., in important models arising in deep neural networks, low-rank matrix recovery, principal component analysis, and matrix completion as discussed in [10, Section 6.2].

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

Proposition 2.6. *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 $k \to \infty$.

When the sequence under consideration is generated by a linesearch method and satisfies some conditions stronger than (H1) and (H2) in Proposition 2.6, its convergence rates are established in [37, Proposition 2.4] under the KL property with $\psi(t) = Mt^q$ as given below.

Proposition 2.7. *Let $f : \mathbb{R}^n \to \mathbb{R}$ be a $\mathcal{C}^1$-smooth function, and let the sequences $\{x^k\} \subset \mathbb{R}^n, \{\tau_k\} \subset [0, \infty), \{d^k\} \subset \mathbb{R}^n$ satisfy the iterative condition $x^{k+1} = x^k + \tau_k d^k$ for all $k \in \mathbb{N}$. Assume that for sufficiently large $k \in \mathbb{N}$, we have $x^{k+1} \neq x^k$ together with the estimates*

$$f(x^k) - f(x^{k+1}) \geq \beta\tau_k \left\|d^k\right\|^2 \quad \text{and} \quad \left\|\nabla f(x^k)\right\| \leq \alpha \left\|d^k\right\|, \tag{2.6}$$

where α, β are some positive constants. Suppose in addition that the sequence $\{\tau_k\}$ is bounded away from 0 (i.e., there exists some $\bar{\tau} > 0$ such that $\tau_k \geq \bar{\tau}$ for sufficiently large $k \in \mathbb{N}$), 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 $q \in [1/2, 1)$. Then the following convergence rates are guaranteed:

(i) *If $q = 1/2$, then the sequence $\{x^k\}$ converges linearly to $\bar{x}$.*

(ii) *If $q \in (1/2, 1)$, then we have the estimate*

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

Remark 2.8. Observe that the two conditions in (2.6) together with the boundedness away from 0 of $\{\tau_k\}$ yield assumptions (H1), (H2) in Proposition 2.6. Indeed, (H1) is verified by the following inequalities:

$$\begin{aligned} f(x^k) - f(x^{k+1}) &\geq \beta\tau_k \left\|d^k\right\|^2 = \beta \left\|\tau_k d^k\right\| \cdot \left\|d^k\right\| \\ &\geq \frac{\beta}{\alpha} \left\|x^{k+1} - x^k\right\| \cdot \left\|\nabla f(x^k)\right\|. \end{aligned}$$

In addition, since $\{\tau_k\}$ is bounded away from 0, there exists $\bar{\tau} > 0$ such that $\tau_k \geq \bar{\tau}$ for sufficiently large $k \in \mathbb{N}$. Then for such k, the condition $f(x^{k+1}) = f(x^k)$ implies that $d^k = 0$ by the first inequality in (2.6), and hence $x^{k+1} = x^k$ by the iterative procedure $x^{k+1} = x^k + \tau_k d^k$, which therefore verifies (H2).

3 Global and Local Approximations of Gradients

This section is devoted to analyzing several methods for approximating gradients of a smooth function $f : \mathbb{R}^n \to \mathbb{R}$ by using only information about the function values that frequently appears in derivative-free optimization. Methods of this type include, in particular, finite differences [47, Section 9], the Gupal estimation [31], and gradient estimation via linear interpolation [7]. We construct two types of approximations that cover all these methods.

Definition 3.1. Let $f : \mathbb{R}^n \to \mathbb{R}$ be a $\mathcal{C}^1$-smooth function. A mapping $\mathcal{G} : \mathbb{R}^n \times (0, \infty) \to \mathbb{R}^n$ is:

(i) A *global approximation* of ∇f if there is a constant $C > 0$ such that

$$\|\mathcal{G}(x, \delta) - \nabla f(x)\| \le C\delta \ \text{ for any } \ (x, \delta) \in \mathbb{R}^n \times (0, \infty). \tag{3.1}$$

(ii) A *local approximation* of ∇f if for any bounded set $\Omega \subset \mathbb{R}^n$ and any $\Delta > 0$, there is $C > 0$ with

$$\|\mathcal{G}(x, \delta) - \nabla f(x)\| \le C\delta \ \text{ for any } \ (x, \delta) \in \Omega \times (0, \Delta]. \tag{3.2}$$

Remark 3.2. We have the following observations related to Definition 3.1:

(i) If $\mathcal{G}$ is a global approximation of ∇f, then it is also a local approximation of ∇f.

(ii) Assume that $\mathcal{G}$ is a local approximation of ∇f and that $x \in \mathbb{R}^n$. Then we deduce from (3.2) with $\Omega = \{x\}$ and any $\Delta > 0$ the condition

$$\limsup_{\delta \downarrow 0} \frac{\|\mathcal{G}(x, \delta) - \nabla f(x)\|}{\delta} < \infty. \tag{3.3}$$

Next we formulate the two standard types of finite differences taken from [47, Section 9], which serve as typical examples of the approximations in Definition 3.1.

- *Forward finite difference*:

$$\mathcal{G}(x, \delta) := \frac{1}{\delta} \sum_{i=1}^{n} \left(f(x + \delta e_i) - f(x) \right) e_i \text{ for any } (x, \delta) \in \mathbb{R}^n \times (0, \infty). \tag{3.4}$$

- *Central finite difference*:

$$\mathcal{G}(x, \delta) := \frac{1}{2\delta} \sum_{i=1}^{n} \left(f(x + \delta e_i) - f(x - \delta e_i) \right) e_i \text{ for any } (x, \delta) \in \mathbb{R}^n \times (0, \infty). \tag{3.5}$$

Remark 3.3. Let us now recall some results on the error bounds for the two types of finite differences that are mentioned above.

(i) The global error bound for the forward finite difference (see, e.g., [7, Theorem 2.1] and [47, Section 8]) shows that it is a global approximation of ∇f when $f \in \mathcal{C}_L^{1,1}$. The local error bound for the forward finite difference is also given in [47, Exercise 9.13].

(ii) On the other hand, the global error bound for the central finite difference (see, e.g., [7, Theorem 2.2] and [47, Lemma 9.1]) requires that f is twice continuously differentiable with a Lipschitz continuous Hessian, which is a rather restrictive assumption.

For completeness, we present a short proof showing that both types of finite differences are global approximations of ∇f when $f \in \mathcal{C}_L^{1,1}$ and are local approximations of ∇f when $f \in \mathcal{C}^{1,1}$.

Proposition 3.4. *Let $f : \mathbb{R}^n \to \mathbb{R}$ be a $\mathcal{C}^1$-smooth function. Then the following hold:*

(i) *Given $x \in \mathbb{R}^n$ and $\delta > 0$, if the gradient ∇f is Lipschitz continuous on $\overline{\mathbb{B}}(x, \delta)$ with the constant $L > 0$, then both forward finite difference (3.4) and central finite difference (3.5) satisfy the estimate*

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

(ii) *If the gradient ∇f is globally Lipschitz continuous with the constant $L > 0$, then both forward finite difference (3.4) and central finite difference (3.5) satisfy the estimate*

$$\|\mathcal{G}(x, \delta) - \nabla f(x)\| \le \frac{L\sqrt{n}\delta}{2} \ \textit{ for any } \ (x, \delta) \in \mathbb{R}^n \times (0, \infty). \tag{3.7}$$

(iii) *If the gradient ∇f is locally Lipschitz continuous, then for any bounded set $\Omega \subset \mathbb{R}^n$ and for any $\Delta > 0$, there exists a positive number L such that both forward finite difference (3.4) and central finite difference (3.5) satisfy the estimate*

$$\|\mathcal{G}(x, \delta) - \nabla f(x)\| \le \frac{L\sqrt{n}\delta}{2} \ \textit{ for any } \ (x, \delta) \in \Omega \times (0, \Delta]. \tag{3.8}$$

Proof. We begin with verifying (i) for each type of the aforementioned finite differences and then employ (i) to justify (ii) and (iii) for both types.

(i) Take any $x \in \mathbb{R}^n, \delta > 0$ and assume that ∇f is Lipschitz continuous on $\overline{\mathbb{B}}(x, \delta)$ with the constant $L > 0$. Consider first the case where $\mathcal{G}$ is given by the forward finite difference (3.4). Then for any $i = 1, \ldots, n$, we get by employing Lemma 2.1 that

$$|f(x + \delta e_i) - f(x) - \langle \nabla f(x), x + \delta e_i - x\rangle| \le \frac{L}{2} \|x + \delta e_i - x\|^2 = \frac{L\delta^2}{2}, \tag{3.9}$$

which is clearly equivalent to

$$\left| \frac{1}{\delta} \left(f\left(x + \delta e_i\right) - f\left(x\right)\right) - \frac{\partial f}{\partial x_i}(x) \right| \le \frac{L\delta}{2}.$$

Since the latter inequality holds for all $i = 1, \ldots, n$, we deduce that

$$\|\mathcal{G}(x, \delta) - \nabla f(x)\| = \sqrt{\sum_{i=1}^{n} \left(\frac{1}{\delta} \left(f(x + \delta e_i) - f\left(x\right)\right) - \frac{\partial f}{\partial x_i}(x) \right)^2} \le \frac{L\sqrt{n}\delta}{2},$$

which therefore verifies estimate (3.6).

Assume now that $\mathcal{G}$ is given by the central finite difference (3.5). Employing Lemma 2.1 gives us for any $i = 1, \ldots, n$ the two estimates

$$|f(x + \delta e_i) - f(x) - \langle \nabla f(x), (x + \delta e_i) - x\rangle| \le \frac{L\delta^2}{2},$$
$$|f(x) - f(x - \delta e_i) - \langle \nabla f(x), x - (x - \delta e_i)\rangle| \le \frac{L\delta^2}{2}.$$

Summing up the above estimates and using the triangle inequality, we deduce that

$$|f(x + \delta e_i) - f(x - \delta e_i) - 2\langle \nabla f(x), \delta e_i\rangle| \le L\delta^2,$$

which implies in turn the conditions

$$\left| \frac{1}{2\delta} \left(f\left(x + \delta e_i\right) - f\left(x - \delta e_i\right)\right) - \frac{\partial f}{\partial x_i}(x) \right| \le \frac{L\delta}{2}$$

for all $i = 1, \ldots, n$. Therefore, we get

$$\|\mathcal{G}(x, \delta) - \nabla f(x)\| = \sqrt{\sum_{i=1}^{n} \left(\frac{1}{2\delta} \left(f(x + \delta e_i) - f\left(x - \delta e_i\right)\right) - \frac{\partial f}{\partial x_i}(x) \right)^2} \le \frac{L\sqrt{n}\delta}{2},$$

which brings us to (3.6) and thus justifies (i).

Assertion (ii) follows directly from (i). To verify (iii), pick some $\Delta > 0$ and a bounded set $\Omega \subset \mathbb{R}^n$, and then find $r > 0$ such that $\Omega \subset r\overline{\mathbb{B}}$. Defining $\Theta := (r + \Delta)\overline{\mathbb{B}}$, it is clear that Θ is compact. Since ∇f is locally Lipschitzian, it is Lipschitz continuous on Θ with some constant $L > 0$. Taking any $(x, \delta) \in \Omega \times (0, \Delta]$, we get that $\overline{\mathbb{B}}(x, \delta) \subset \Theta$, and thus ∇f is Lipschitz continuous on $\overline{\mathbb{B}}(x, \delta)$ with the same constant L. Employing finally (i) justifies assertion (iii). □

The following example shows that when the local Lipschitz continuity of ∇f is replaced by merely the continuity of ∇f, the finite differences may not be a local approximation of ∇f.

Example 1. Define the univariate real-valued function f by

$$f(x) := \begin{cases} \frac{2}{3}\sqrt{x^3} & \text{if } x \ge 0, \\ -\frac{2}{3}\sqrt{-x^3} & \text{if } x < 0. \end{cases}$$

The derivative of f is calculated by

$$\nabla f(x) = \begin{cases} \sqrt{x} & \text{if } x \ge 0, \\ \sqrt{-x} & \text{if } x < 0 \end{cases}$$

being clearly continuous on $\mathbb{R}$ while not Lipschitz continuous around 0. If we suppose that $\mathcal{G}(x,\delta)$ is the forward finite difference approximation of $\nabla f(x)$ from (3.4), we get that

$$\mathcal{G}(0,\delta)=\frac{f(\delta)-f(0)}{\delta}=\frac{\frac{2}{3}\sqrt{\delta^3}}{\delta}=\frac{2\sqrt{\delta}}{3}\quad\text{for all}\quad \delta>0,$$

which implies that $\mathcal{G}(0,\delta)/\delta\to\infty$ as $\delta\downarrow 0$. It follows from (3.3) that $\mathcal{G}(x,\delta)$ is not a local approximation of the derivative ∇f. Supposing now that $\mathcal{G}(x,\delta)$ is the central finite difference approximation of $\nabla f(x)$, we deduce from (3.5) the expression

$$\mathcal{G}(0,\delta)=\frac{f(\delta)-f(-\delta)}{2\delta}=\frac{4\sqrt{\delta}}{3}\quad\text{for all}\quad \delta>0,$$

which also tells us that $\mathcal{G}(x,\delta)$ is not a local approximation of ∇f.

4 General Derivative-Free Methods for $\mathcal{C}^{1,1}_L$ Functions

This section addresses the optimization problem (1.1) when $f\in\mathcal{C}^{1,1}_L$ for some $L>0$. By employing gradient approximation methods that satisfy the global error bound (3.1), we propose the general *derivative-free method with constant stepsize* (DFC) to solve this problem for both noiseless and noisy cases, providing its convergence analysis. The DFC algorithm is described as follows.

4.1 Algorithm Construction

Algorithm 1 (DFC)**.**

Step 0. Choose a global approximation $\mathcal{G}$ of ∇f under condition (3.1). Select an initial point $x^1\in\mathbb{R}^n$, an initial sampling radius $\delta_1>0$, a constant $C_1>0$, a reduction factor $\theta\in(0,1)$, and scaling factors $\mu>2,\eta>1,\kappa>0$. Set $k:=1$.

Step 1 (approximate gradient)**.** Find g^k and the smallest nonnegative integer i_k such that

$$g^k=\mathcal{G}(x^k,\theta^{i_k}\delta_k)\ \text{and}\ \left\|g^k\right\|>\mu C_k\theta^{i_k}\delta_k.$$

Then set $\delta_{k+1}:=\theta^{i_k}\delta_k$.

Step 2 (update)**.** If $f\Big(x^k-\dfrac{\kappa}{C_k}g^k\Big)\le f(x^k)-\dfrac{\kappa(\mu-2)}{2C_k\mu}\left\|g^k\right\|^2$, then $x^{k+1}:=x^k-\dfrac{\kappa}{C_k}g^k$ and $C_{k+1}:=C_k$. Otherwise, $x^{k+1}:=x^k$ and $C_{k+1}:=\eta C_k$.

Remark 4.1. Let us present some observations concerning Algorithm 1. The first observation clarifies the existence of g^k and i_k in Step 1. Observation (ii) explains the iteration updates in Step 2 while observation (iii) interprets the term "constant stepsize" in the name of our method.

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

$$g^k:=\mathcal{G}(x^k,\delta_k). \tag{4.1}$$

While $\left\|g^k\right\|\le\mu C_k\theta^{i_k}\delta_k$, increase i_k by 1 and recalculate g^k under (4.1). When $\nabla f(x^k)\neq 0$, the existence of g^k and i_k in Step 1 is guaranteed. Indeed, otherwise we get a sequence $\{g^k_i\}$ with

$$g^k_i=\mathcal{G}(x^k,\theta^i\delta_k)\ \text{ and }\ \left\|g^k_i\right\|\le\mu\theta^i\delta_k\ \text{ for all }\ i\in\mathbb{N}. \tag{4.2}$$

Since $\theta\in(0,1)$, the latter means that $g^k_i\to 0$ as $i\to\infty$. Remembering that $\mathcal{G}$ is a global approximation of ∇f, we get for $C>0$ given in (3.1) that

$$\left\|g^k_i-\nabla f(x^k)\right\|\le C\theta^i\delta_k\ \text{ whenever }\ i\in\mathbb{N}.$$

Letting $i\to\infty$ with taking into account that $g^k_i\to 0$, the latter inequality implies that $\nabla f(x^k)=0$, which is a contradiction.

(ii) The condition $f(x^k-C_k^{-1}\kappa g^k)\le f(x^k)-\dfrac{\kappa(\mu-2)}{2C_k\mu}\left\|g^k\right\|^2$ determines whether C_k is a good approximation for C in the sense that the objective function f is sufficiently decreasing when the iterate moves from x^k to $x^{k+1}:=x^k-C_k^{-1}\kappa g^k$. If this condition fails, we increase C_k by setting $C_{k+1}:=rC_k$ to get a better approximation for C and stagnate the iterative sequence by setting $x^{k+1}:=x^k$.

(iii) It will be shown in Proposition 4.2 that there exists a positive number $\bar{C}$ such that $C_k=\bar{C}$ for sufficiently large $k\in\mathbb{N}$, which also implies that $x^{k+1}=x^k-\kappa\bar{C}^{-1}g^k$ for such k. This explains the term "constant stepsize" in the name of our algorithm.

4.2 Analysis for Noiseless Functions

In this subsection, we derive convergence properties of DFC in Algorithm 2 for noiseless functions, i.e., when $f(x)$ is available for all $x \in \mathbb{R}^n$. Our analysis begins with a crucial result showing that the tail of the sequence $\{C_k\}$ generated by Algorithm 1 is constant.

Proposition 4.2. *Let $\{C_k\}$ be the sequence generated by Algorithm 1. Assume that $\nabla f(x^k) \neq 0$ for all $k \in \mathbb{N}$. Then there exists a number $N \in \mathbb{N}$ such that $C_{k+1} = C_k$ whenever $k \geq N$.*

Proof. Since $\mathcal{G}$ is a global approximation of ∇f under condition (3.1), there exists $C > 0$ such that

$$\|\mathcal{G}(x,\delta) - \nabla f(x)\| \leq C\delta \text{ for all } (x,\delta) \in \mathbb{R}^n \times (0,\infty). \tag{4.3}$$

By the imposed assumption, we find $L > 0$ such that ∇f is Lipschitz continuous with the constant L on $\mathbb{R}^n$. Arguing by contradiction, suppose that the number N asserted in the proposition does not exist. By Step 2 of Algorithm 1, this implies that $C_{k+1} = \eta C_k$ for infinitely many $k \in \mathbb{N}$, and hence $C_k \to \infty$ as $k \to \infty$. Therefore, there exists a number $K \in \mathbb{N}$ such that $C_{K+1} = \eta C_K$ and $C_K > \max\{C, L\kappa\}$. Using Step 2 of Algorithm 1 together with the update $C_{K+1} = \eta C_K$, we deduce that

$$f\Big(x^K - \frac{\kappa}{C_K} g^K\Big) > f(x^K) - \frac{\kappa(\mu-2)}{2C_K\mu}\left\|g^K\right\|^2. \tag{4.4}$$

Combining $g^K = \mathcal{G}(x^K, \delta_{K+1})$ and $\left\|g^K\right\| \geq \mu C_K \delta_{K+1}$ from Step 1 of Algorithm 1 with (4.3) and $C_K > C$ as above, we get the relationships

$$\begin{aligned}\left\|g^K - \nabla f(x^K)\right\| &= \left\|\mathcal{G}(x^K, \delta_{K+1}) - \nabla f(x^K)\right\| \\ &\leq C\delta_{K+1} \leq C_K\delta_{K+1} \leq \mu^{-1}\left\|g^K\right\|.\end{aligned}$$

By the Cauchy-Schwarz inequality, the latter tells us that

$$\begin{aligned}\left\langle \nabla f(x^K), g^K \right\rangle &= \left\langle \nabla f(x^K) - g^K, g^K \right\rangle + \left\|g^K\right\|^2 \\ &\geq -\left\|\nabla f(x^K) - g^K\right\| \cdot \left\|g^K\right\| + \left\|g^K\right\|^2 \\ &\geq (1-\mu^{-1})\left\|g^K\right\|^2.\end{aligned}$$

Combining this with Lemma 2.1 and taking into account the global Lipschitz continuity of ∇f with the constant L as well as the condition $C_K > L\kappa$ as above, we get that

$$\begin{aligned} f\Big(x^K - \frac{\kappa}{C_K}g^K\Big) - f(x^K) &\leq -\frac{\kappa}{C_K}\left\langle \nabla f(x^K), g^K\right\rangle + \frac{L}{2}\left\|\frac{\kappa}{C_K}g^K\right\|^2 \\ &\leq -\frac{\kappa}{C_K}\Big(1 - \frac{1}{\mu}\Big)\left\|g^K\right\|^2 + \frac{\kappa}{2C_K}\left\|g^K\right\|^2 \\ &= -\frac{\kappa}{C_K}\left\|g^K\right\|^2\Big(\frac{1}{2} - \frac{1}{\mu}\Big) = -\frac{\kappa(\mu-2)}{2C_K\mu}\left\|g^K\right\|^2,\end{aligned}$$

which clearly contradicts (4.4) and thus completes the proof of the proposition. $\square$

Now we are ready to establish the convergence properties of Algorithm 1 in the noiseless case.

Theorem 4.3. *Let $\{x^k\}$ be the sequence generated by Algorithm 1 and assume that $\nabla f(x^k) \neq 0$ for all $k \in \mathbb{N}$. Then either $f(x^k) \to -\infty$ as $k \to \infty$, or we have the assertions:*

(i) *The gradient sequence $\{\nabla f(x^k)\}$ converges to 0 as $k \to \infty$.*

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

(iii) *If f satisfies the KL property at some accumulation point $\bar{x}$ of $\{x^k\}$ with $\psi(t) = Mt^q$ for $M > 0$ and $q \in [1/2, 1)$, then the following convergence rates are guaranteed for $\{x^k\}$:*

- *If $q = 1/2$, then $\{x^k\}$, $\{\nabla f(x^k)\}$, and $\{f(x^k)\}$ converge linearly to $\bar{x}$, 0, and $f(\bar{x})$, respectively.*
- *The setting of $q \in (1/2, 1)$ ensures the estimates*

$$\left\|x^k - \bar{x}\right\| = \mathcal{O}\Big(k^{-\frac{1-q}{2q-1}}\Big),\ \left\|\nabla f(x^k)\right\| = \mathcal{O}\Big(k^{-\frac{1-q}{2q-1}}\Big),\ \ \textit{and}\ \ f(x^k) - f(\bar{x}) = \mathcal{O}\Big(k^{-\frac{2-2q}{2q-1}}\Big).$$

Proof. Since $\mathcal{G}$ is a global approximation of ∇f under condition (3.1), there exists $C > 0$ such that

$$\|\mathcal{G}(x,\delta) - \nabla f(x)\| \leq C\delta \text{ for all } (x,\delta) \in \mathbb{R}^n \times (0,\infty). \tag{4.5}$$

By $f \in \mathcal{C}_L^{1,1}$, we find $L > 0$ such that the gradient mapping ∇f is Lipschitz continuous with the constant L on $\mathbb{R}^n$. Taking the number $N \in \mathbb{N}$ from Proposition 4.2 ensures that $C_k = C_N$ for all $k \geq N$. This implies by Step 2 of Algorithm 1 that

$$f(x^{k+1}) = f\Big(x^k - \frac{\kappa}{C_N} g^k\Big) \leq f(x^k) - \frac{\kappa(\mu-2)}{2C_N\mu} \left\|g^k\right\|^2 \quad \text{for all} \quad k \geq N, \tag{4.6}$$

which tells us that $\{f(x^k)\}_{k\geq N}$ is decreasing. If $f(x^k) \to -\infty$, there is nothing to prove, so we assume that $f(x^k) \not\to -\infty$, which implies that $\{f(x^k)\}$ is convergent. As a consequence, we get $f(x^k) - f(x^{k+1}) \to 0$ as $k \to \infty$. Then (4.6) tells us that $g^k \to 0$. From Step 1 of Algorithm 1 it follows that

$$\left\|g^k\right\| > \mu C_k \delta_{k+1} = \mu C_N \delta_{k+1} \quad \text{for all} \quad k \geq N \tag{4.7}$$

ensuring that $\delta_{k+1} \downarrow 0$ as $k \to \infty$. It further follows from $g^k = \mathcal{G}(x^k, \delta_{k+1})$ and (4.5) that

$$\left\|g^k - \nabla f(x^k)\right\| = \left\|\mathcal{G}(x^k, \delta_{k+1}) - \nabla f(x^k)\right\| \leq C\delta_{k+1} \text{ for all } k \in \mathbb{N}, \tag{4.8}$$

which yields $\nabla f(x^k) \to 0$ as $k \to \infty$ and thus justifies (i).

To verify (ii), take any accumulation point $\bar{x}$ of $\{x^k\}$ and assume that f satisfies the KL property at $\bar{x}$. By (4.7) and (4.8), we obtain that

$$\begin{aligned} \left\|\nabla f(x^k)\right\| &\leq \left\|g^k\right\| + \left\|\nabla f(x^k) - g^k\right\| \leq \left\|g^k\right\| + C\delta_{k+1} \\ &\leq \left\|g^k\right\| + \frac{C\left\|g^k\right\|}{\mu C_N} = \alpha\left\|g^k\right\| \quad \text{for all} \quad k \geq N, \end{aligned}$$

where $\alpha := \frac{\mu C_N + C}{\mu C_N}$. This together with (4.6) brings us to condition (2.6). By Remark 2.8(i), assumptions (H1) and (H2) in Proposition 2.6 hold. Therefore, $x^k \to \bar{x}$ as $k \to \infty$, which justifies (ii).

To proceed with the proof of assertion (iii) under the KL property at $\bar{x}$ with $\psi(t) = Mt^q$, we use the iterations $x^{k+1} = x^k - C^{-1}\kappa g^k$ as in Step 2 of Algorithm 1 together with $\|g^k\| > 0$ from Step 1 of Algorithm 1. This gives us $x^{k+1} \neq x^k$ for $k \geq N$. Combining the latter with (4.6) and (4.9), we see that all the assumptions in Proposition 2.7 are satisfied. This verifies the convergence rates of $\{x^k\}$ to $\bar{x}$ stated in (iii). Since $\bar{x}$ is an accumulation point of $\{x^k\}$, it follows from (i) that $\bar{x}$ is a stationary point of f, i.e., $\nabla f(\bar{x}) = 0$. Hence the usage of Lemma 2.1 and the decreasing property of $\{f(x^k)\}_{k\geq N}$ yields

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

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

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

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

This therefore completes the proof of the theorem. □

Remark 4.4. Here we present a comparison between our analysis for DFC with the analysis in [13]. While both [13] and our paper address the noiseless case and [13] considers a more general approach, our analysis provides additional developments that are not studied in [13]. Specifically:

- **(i)** Our DFC method (Algorithm 1) explicitly specifies how to construct the gradient approximation. In contrast, [13, Algorithm 3.1] assumes a more general construction and requires the gradient approximation to be sufficiently accurate (as per [13, Assumption 3.1]). In order to make the gradient approximation in DFC satisfying [13, Assumption 3.1], the constant C_k should be larger than C, which is not required in and not ensured by our analysis. Furthermore, the finite difference interval and the stepsize are interacting each other in our DFC method, while they are considered separately in [13, Algorithm 3.1].
- **(ii)** Additionally, [13, Algorithm 3.1] employs a different rule for choosing stepsize, allowing it to increase after each iteration, while our DFC does not allow this. The numerical experiments in Subsection 6.1.1 demonstrate that this small change significantly affects the numerical performance of the methods with a more favourable result for DFC.
- **(iii)** Apart from the differences in algorithmic constructions, our analysis takes a distinct direction by demonstrating the convergence of the gradient sequence to 0 and the convergence of the sequence of iterates to a stationary point. On the other hand, [13, Theorem 3.1] reveals the number of iterations required to reach a near-stationary point, and it also establishes that $\liminf_{k\to\infty} \left\|\nabla f(x^k)\right\| = 0$. Although we do not conduct the complexity analysis for DFC in this section, our asymptotic results look to be transparently stronger.

4.3 Analysis for Noisy Functions

In this part, we provide the convergence analysis with error bounds for DFC in Algorithm 1 addressing problem (1.1) when only a *noisy approximation* $\phi(x) = f(x) + \xi(x)$ of f is available, where $\xi : \mathbb{R}^n \to \mathbb{R}$ is a *noise function* bounded by some constant $\xi_f > 0$, i.e.,

$$|\xi(x)| \le \xi_f \text{ for all } x \in \mathbb{R}^n. \tag{4.9}$$

Due to the design of DFC, we *do not* assume that ξ_f is known. For brevity, consider only the *forward finite difference approximation*, while other gradient approximation methods can be employed via modifications of the inexact conditions in Definition 3.1 for noisy functions. We first construct the gradient approximation for f via the forward finite difference with the noisy function ϕ defined by

$$\widetilde{\mathcal{G}}(x,\delta) := \frac{1}{\delta}\sum_{i=1}^{n}\big(\phi(x+\delta e_i) - \phi(x)\big)e_i \text{ for any } (x,\delta) \in \mathbb{R}^n \times (0,\infty), \tag{4.10}$$

where the positive number δ in (4.10) is named the *finite difference interval.* Recall the following noisy version of Proposition 3.4, which is well known and can be found in, e.g., [7, Theorem 2.1].

Proposition 4.5. *Let $f : \mathbb{R}^n \to \mathbb{R}$ be a $\mathcal{C}_L^{1,1}$ function. Then the noisy forward finite difference (4.10) satisfies the error bound*

$$\left\|\widetilde{\mathcal{G}}(x,\delta) - \nabla f(x)\right\| \le \frac{L\sqrt{n}\delta}{2} + \frac{2\sqrt{n}\xi_f}{\delta} \text{ for all } \delta > 0. \tag{4.11}$$

For a better exposition, consider DFC with specific parameters $\mu = 4$ and $\kappa = \frac{L\sqrt{n}}{2}$, although other general selections of $\mu > 2$ and $\kappa > 0$ still work with the same analysis. We also define $L_k := \frac{C_k}{\kappa}$ for each $k \in \mathbb{N}$ as approximate Lipschitz constants. In order to deal with noise, a relaxation is required in the descent condition in Step 2 of DFC, which leads us to the following algorithm.

Algorithm 2 (DFC for noisy functions).

Step 0. Select some $x^1 \in \mathbb{R}^n, \delta_1 > 0,\ L_1 > 0,\ \theta \in (0,1)$, and $\eta > 1$.

Step 1 (approximate gradient). Find g^k and the smallest nonnegative integer i_k such that

$$g^k = \widetilde{\mathcal{G}}(x^k, \theta^{i_k}\delta_k) \text{ and } \left\|g^k\right\| > 2L_k\sqrt{n}\theta^{i_k}\delta_k. \tag{4.12}$$

Then set $\delta_{k+1} := \theta^{i_k}\delta_k$.

Step 2 (update). If $\phi\Big(x^k - \dfrac{1}{L_k}g^k\Big) \le \phi(x^k) - \dfrac{1}{24L_k}\left\|g^k\right\|^2$, then $x^{k+1} := x^k - \dfrac{1}{L_k}g^k$ and $L_{k+1} := L_k$. Otherwise, $x^{k+1} := x^k$ and $L_{k+1} := \eta L_k$.

Due to the presence of noise, there is no guarantee that Step 1 of Algorithm 2 will terminate after a finite number of trials for i_k. Therefore, we say that Step 1 is *successful* if such an i_k is found, and is *unsuccessful* otherwise. Let us begin our analysis with a result that ensures the success of Step 1 of Algorithm 2 and provides a bound for the approximate Lipschitz constant L_k, whenever the gradient of the current iterate is not near 0.

Proposition 4.6. *At the k^{th} iteration of Algorithm 2, if the conditions*

$$\left\|\nabla f(x^k)\right\| \ge (4\theta^{-1}\eta + \theta^{-1} + 1)\sqrt{Ln\xi_f}, \quad \delta_k \ge \sqrt{\frac{4\xi_f}{L}}, \quad L_k < \eta L$$

are satisfied, then Step 1 is successful with $\delta_{k+1} \ge \sqrt{\frac{4\xi_f}{L}}$.

Proof. Let $i := \lfloor \log_\theta\Big(\frac{1}{\delta_k}\sqrt{\frac{4\xi_f}{L}}\Big)\rfloor$, where $\lfloor\cdot\rfloor$ stands for the *floor/greatest integer function.* Since we have $\delta_k \ge \sqrt{\frac{4\xi_f}{L}}$ and $\theta \in (0,1)$, the number i is a nonnegative integer satisfying the inclusion

$$i \in \left(\log_\theta\left(\frac{1}{\delta_k}\sqrt{\frac{4\xi_f}{L}}\right) - 1, \log_\theta\left(\frac{1}{\delta_k}\sqrt{\frac{4\xi_f}{L}}\right)\right]$$

while implying in turn that $\theta^i\delta_k \in \left[\sqrt{\frac{4\xi_f}{L}}, \theta^{-1}\sqrt{\frac{4\xi_f}{L}}\right)$. We now show that inequality (4.12) in Step 1 of Algorithm 2 is satisfied for $i_k = i$, which yields the success of the step with $\delta_{k+1} \geq \theta^i\delta_k \geq \sqrt{\frac{4\xi_f}{L}}$. Indeed, with $g_i^k := \widetilde{\mathcal{G}}(x^k, \theta^i\delta_k)$, the error bound (4.11) and $\theta^i\delta_k \in \left[\sqrt{\frac{4\xi_f}{L}}, \theta^{-1}\sqrt{\frac{4\xi_f}{L}}\right)$ tell us that

$$\begin{aligned}\left\|g_i^k - \nabla f(x^k)\right\| &\leq \frac{L\sqrt{n}}{2}\theta^i\delta_k + \frac{2\sqrt{n}\xi_f}{\theta^i\delta_k} \\ &\leq \theta^{-1}\sqrt{Ln\xi_f} + \sqrt{Ln\xi_f} = (1+\theta^{-1})\sqrt{Ln\xi_f}.\end{aligned}$$

Since $\left\|\nabla f(x^k)\right\| \geq (4\theta^{-1}\eta + \theta^{-1} + 1)\sqrt{Ln\xi_f}$ and $L_k < \eta L$, we get that

$$\left\|g_i^k\right\| \geq \left\|\nabla f(x^k)\right\| - \left\|g_i^k - \nabla f(x^k)\right\| \geq 4\theta^{-1}\eta\sqrt{Ln\xi_f} = 2\eta L\sqrt{n}\theta^{-1}\sqrt{\frac{4\xi_f}{L}} > 2L_k\sqrt{n}\theta^i\delta_k.$$

Therefore, Step 1 of Algorithm 2 is successful with $\delta_{k+1} \geq \theta^i\delta_k \geq \sqrt{\frac{4\xi_f}{L}}$. □

Proposition 4.7. *At the k^{th} iteration of Algorithm 2, suppose that $\left\|\nabla f(x^k)\right\| \geq 16\sqrt{Ln\xi_f}$ and that Step 1 is successful with $\delta_{k+1} \geq \sqrt{\frac{4\xi_f}{L}}$. The following assertions hold:*

- **(i)** $\left\|g^k - \nabla f(x^k)\right\| \leq \frac{4L+L_k}{15L_k}\left\|g^k\right\|$.
- **(ii)** *If in addition $L_k \geq L$, then $\left\langle\nabla f(x^k), g^k\right\rangle \geq \frac{2}{3}\left\|g^k\right\|^2$ and $\left\|\nabla f(x^k)\right\| \leq \frac{4}{3}\left\|g^k\right\|$.*
- **(iii)** *If in addition $L_k < \eta L$, then $L_{k+1} < \eta L$.*

Proof. (i) Since Step 1 of Algorithm 2 is successful, we deduce that $\left\|g^k\right\| \geq 2L_k\sqrt{n}\delta_{k+1}$. Combining this with $g^k = \widetilde{\mathcal{G}}(x^k, \delta_{k+1})$, (4.11), and the estimates $\delta_{k+1} \geq \sqrt{\frac{4\xi_f}{L}}$, $\left\|\nabla f(x^k)\right\| \geq 16\sqrt{Ln\xi_f}$ brings us to

$$\begin{aligned}\left\|g^k - \nabla f(x^k)\right\| &\leq \frac{L\sqrt{n}}{2}\delta_{k+1} + \frac{2\sqrt{n}\xi_f}{\delta_{k+1}} \\ &\leq \frac{L\sqrt{n}}{2}\frac{\left\|g^k\right\|}{2L_k\sqrt{n}} + \sqrt{Ln\xi_f} \\ &\leq \frac{L}{4L_k}\left\|g^k\right\| + \frac{1}{16}\left\|\nabla f(x^k)\right\| \\ &\leq \frac{L}{4L_k}\left\|g^k\right\| + \frac{1}{16}\left\|g^k\right\| + \frac{1}{16}\left\|g^k - \nabla f(x^k)\right\|.\end{aligned}$$

The latter inequality yields $\left\|g^k - \nabla f(x^k)\right\| \leq \frac{4L+L_k}{15L_k}\left\|g^k\right\|$, which verifies (i).

(ii) Using (i) and $L_k \geq L$ gives us $\left\|g^k - \nabla f(x^k)\right\| \leq \frac{1}{3}\left\|g^k\right\|$. Combining this estimates with the Cauchy-Schwarz inequality, we arrive at

$$\begin{aligned}\left\langle\nabla f(x^k), g^k\right\rangle &\geq \left\langle\nabla f(x^k) - g^k, g^k\right\rangle + \left\|g^k\right\|^2 \\ &\geq -\left\|\nabla f(x^k) - g^k\right\|\left\|g^k\right\| + \left\|g^k\right\|^2 \geq \frac{2}{3}\left\|g^k\right\|^2.\end{aligned}$$

In addition, it follows from $\left\|g^k - \nabla f(x^k)\right\| \leq \frac{1}{3}\left\|g^k\right\|$ that

$$\left\|g^k\right\| \geq \left\|\nabla f(x^k)\right\| - \left\|g^k - \nabla f(x^k)\right\| \geq \left\|\nabla f(x^k)\right\| - \frac{1}{3}\left\|g^k\right\|,$$

which justifies the claimed estimates $\left\|\nabla f(x^k)\right\| \leq \frac{4}{3}\left\|g^k\right\|$.

(iii) This assertion obviously holds if $L_{k+1} = L_k$, so we consider the case where $L_{k+1} = \eta L_k$. Suppose on the contrary that $L_{k+1} \geq \eta L$, which yields $L_k \geq L$. Then it follows from assertion (ii) that $\left\langle\nabla f(x^k), g^k\right\rangle \geq \frac{2}{3}\left\|g^k\right\|^2$. Furthermore, Lemma 2.1 tells us that

$$\begin{aligned}f\left(x^k - \tfrac{1}{L_k}g^k\right) - f(x^k) &\leq -\frac{1}{L_k}\left\langle\nabla f(x^k), g^k\right\rangle + \frac{L}{2}\left\|\frac{1}{L_k}g^k\right\|^2 \\ &\leq -\frac{1}{L_k}\frac{2}{3}\left\|g^k\right\|^2 + \frac{1}{2L_k}\left\|g^k\right\|^2 = -\frac{1}{6L_k}\left\|g^k\right\|^2.\end{aligned} \tag{4.13}$$

Since Step 1 is successful with $\delta_{k+1} \geq \sqrt{\frac{4\xi_f}{L}}$ and $L_k \geq L$, we deduce that

$$\left\|g^k\right\| > 2L_k\delta_{k+1} \geq 2L_k\sqrt{\frac{4\xi_f}{L}} \geq 4\sqrt{L_k\xi_f},$$

which means that $\xi_f \leq \frac{1}{16L_k}\left\|g^k\right\|^2$. Combining the latter with (4.9) and (4.13) gives us

$$\begin{aligned}\phi\Big(x^k - \frac{1}{L_k}g^k\Big) - \phi(x^k) &\leq f\Big(x^k - \frac{1}{L_k}g^k\Big) - f(x^k) + 2\xi_f \\ &\leq -\frac{1}{6L_k}\left\|g^k\right\|^2 + \frac{1}{8L_k}\left\|g^k\right\|^2 = -\frac{1}{24L_k}\left\|g^k\right\|^2.\end{aligned}$$

By Step 2 of Algorithm 2, it follows that $L_{k+1} = L_k$, a contradiction, which justifies $L_{k+1} < \eta L$. □

In the propositions above, we can choose $\theta \in (0,1)$ and $\eta > 1$ to get $4\theta^{-1}\eta + \theta^{-1} + 1 < 16$ by taking, e.g., $\theta = \frac{\sqrt{2}}{2}$ and $\eta = 2$. To simplify the presentation, we make such a *selection of parameters* in the results below. Under this choice, the following property of Algorithm 2 can be deduced immediately from Proposition 4.6 and Proposition 4.7(iii).

Proposition 4.8. *Let $\{x^k\}$ be generated by Algorithm 2 with $L_1 < \eta L$ and $\delta_1 \geq \sqrt{\frac{4\xi_f}{L}}$. If for some $K \in \mathbb{N}$ we have $\left\|\nabla f(x^k)\right\| \geq 16\sqrt{Ln\eta\xi_f}$ whenever $k = 1,\ldots,K$, then Step 1 is successful with $\delta_{k+1} \geq \sqrt{\frac{4\xi_f}{L}}$ and $L_{k+1} < \eta L$ for all $k = 1,\ldots,K$.*

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

Theorem 4.9. *Let $\{x^k\}$ be generated by Algorithm 2 with $\delta_1 \geq \sqrt{\frac{4\xi_f}{L}}$ and $L_1 < \eta L$. Then the number N of iterations that Algorithm 2 takes until $\left\|\nabla f(x^N)\right\| < 16\sqrt{Ln\xi_f}$ is bounded by*

$$N \leq N_{\text{opt}} := 1 + \left\lfloor \frac{f(x^1) - f^* + 2\xi_f}{M\xi_f} \right\rfloor + \left\lfloor \log_\eta\left(\frac{\eta L}{L_1}\right)\right\rfloor, \quad \textit{where} \quad M := \frac{15nL_1^2}{\eta(L+4L_1)^2}.$$

The total number N_{fval} of function evaluations needed to achieve this goal is bounded by

$$N_{\text{fval}} \leq (n+2)N_{\text{opt}} + n\left\lfloor \log_\theta\left(\frac{2\sqrt{\xi_f}}{\delta_1\sqrt{L}}\right)\right\rfloor.$$

Proof. If Step 1 is unsuccessful for the first time at $K \leq N_{\text{opt}}$, then it follows from Proposition 4.8 that $\left\|\nabla f(x^N)\right\| < 16\sqrt{Ln\xi_f}$ for some $N \leq K$, which verifies the claimed bound. Now we suppose that Step 1 is successful for all $k = 1,\ldots,N_{\text{opt}}$ and assume on the contrary that

$$\left\|\nabla f(x^k)\right\| \geq 16\sqrt{Ln\xi_f} \quad \text{for all} \quad k = 1,\ldots,N_{\text{opt}}.$$

Proposition 4.8 tells us that $\delta_{k+1} \geq \sqrt{\frac{4\xi_f}{L}}$ and $L_{k+1} < \eta L$ for all $k = 1,\ldots,N_{\text{opt}}$, and thus it follows from Proposition 4.7(i) and the construction of $\{L_k\}$ that

$$\left\|g^k - \nabla f(x^k)\right\| \leq \frac{4L + L_k}{15L_k}\left\|g^k\right\| \leq \frac{4L + L_1}{15L_1}\left\|g^k\right\|.$$

This gives us in turn the estimates

$$16\sqrt{Ln\xi_f} \leq \left\|\nabla f(x^k)\right\| \leq \left\|g^k\right\| + \left\|g^k - \nabla f(x^k)\right\| \leq \frac{4L + 16L_1}{15L_1}\left\|g^k\right\| \text{ for all } k = 1,\ldots,N_{\text{opt}}. \tag{4.14}$$

Define $I := \{k \in \mathbb{N} \mid 1 \leq k \leq N_{\text{opt}},\ L_{k+1} = L_k\}$ and deduce from the construction of $\{L_k\}$ with $L_{k+1} < \eta L$ as $k = 1,\ldots,N_{\text{opt}}$ that there are at most $\left\lfloor \log_\eta\left(\frac{\eta L}{L_1}\right)\right\rfloor$ iterations for which $L_{k+1} = \eta L_k$. This yields

$$|I| \geq N_{\text{opt}} - \left\lfloor \log_\eta\left(\frac{\eta L}{L_1}\right)\right\rfloor = 1 + \left\lfloor \frac{f(x^1) - f^* + 2\xi_f}{M\xi_f}\right\rfloor. \tag{4.15}$$

Take any $k = 1,\ldots,N_{\text{opt}}$. If $k \notin I$, we get that $\phi(x^{k+1}) = \phi(x^k)$. For any $k \in I$, it follows from Step 2 of Algorithm 2, $L_k < \eta L$ and (4.14) that

$$\phi(x^{k+1}) - \phi(x^k) \leq -\frac{1}{24L_k}\left\|g^k\right\|^2 \leq -\frac{1}{24\eta L}\left\|g^k\right\|^2 \leq -M\xi_f,$$

where M is defined in the statement of the theorem. Since $\phi(x^k) = \phi(x^{k+1})$ when $k \notin I$, we have

$$f^* - \xi_f \le \phi(x^{N_{\rm opt}+1}) = \phi(x^1) + \sum_{k=1}^{N_{\rm opt}} (\phi(x^{k+1}) - \phi(x^k)) \le f(x^1) + \xi_f - |I| M \xi_f,$$

which yields $|I| \le \frac{f(x^1) - f^* + 2\xi_f}{M\xi_f}$ and thus contradicts (4.15).

(iii) By (ii), at most $N_{\rm opt}$ iterations are needed to reach the near stationary point. For each iteration of Algorithm 2, we need at least one approximate gradient evaluation g^k in Step 1. Since $\{\delta_k\}$ is nonincreasing with $\delta_1 \ge \delta_k \ge \sqrt{\frac{4\xi_f}{L}}$, the number of additional approximate gradient evaluations g^k required to adjust the finite difference intervals δ_k throughout all the iterations is at most $\left\lfloor \log_\theta \left(\frac{2\sqrt{\xi_f}}{\delta_1 \sqrt{L}} \right) \right\rfloor$. Employing the forward finite difference, we can reuse $\phi(x^k)$ for additional gradient evaluations. This tells us that the total number of function evaluations for determining the approximate gradient g^k is at most $(n+1)N_{\rm opt} + n \left\lfloor \log_\theta \left(\frac{2\sqrt{\xi_f}}{\delta_1 \sqrt{L}} \right) \right\rfloor$. We also need one additional function evaluation to check the descent condition in Step 2 of Algorithm 2 at each iteration, which results in at most $(n+2)N_{\rm opt} + n \left\lfloor \log_\theta \left(\frac{2\sqrt{\xi_f}}{\delta_1 \sqrt{L}} \right) \right\rfloor$ total function evaluations. □

Remark 4.10. Let us briefly discuss relationships between our analysis for DFC and the analysis in [6]. First observe that the noise level is unknown for our DFC algorithm, while it is required to be known for the analysis in [6] as mentioned after [6, Assumption 1.3]. The algorithmic construction of [6, Algorithm 2.1] shares many similarities with [13, Algorithm 3.1], and so it has some major differences with our DFC as mentioned above in Remark 4.4(i,ii).

5 General Derivative-Free Methods for $\mathcal{C}^{1,1}$ Functions

In this section, we consider problem (1.1), where f is of class $C^{1,1}$ and develop new derivative-free optimization methods in both cases of noiseless and noisy objective functions.

5.1 Backtracking Linesearch for Noiseless Functions

Here we propose and justify the novel *derivative-free method with backtracking stepsize* (DFB) to solve the optimization problem (1.1) in the noiseless setting. The main result of this subsection establishes the *global convergence* with convergence rates of the following algorithm, which employs gradient approximations satisfying the local error bound estimate (3.2).

Algorithm 3 (DFB).

Step 0 (initialization). Choose a local approximation $\mathcal{G}$ of ∇f under condition (3.2). Select an initial point $x^1 \in {\rm I\!R}^n$ and initial radius $\delta_1 > 0$, a constant $C_1 > 0$, factors $\theta \in (0,1)$, $\mu > 2$, $\eta > 1$, linesearch constants $\beta \in (0, 1/2), \gamma \in (0,1)$, $\bar\tau > 0$, and an initial bound $t_1^{\min} \in (0, \bar\tau)$. Choose a sequence of manually controlled errors $\{\nu_k\} \subset [0,\infty)$ such that $\nu_k \downarrow 0$ as $k \to \infty$. Set $k := 1$.

Step 1 (approximate gradient). Select g^k and the smallest nonnegative integer i_k so that

$$g^k = \mathcal{G}\big(x^k, \min\big\{\theta^{i_k}\delta_k, \nu_k\big\}\big) \quad \text{and} \quad \big\|g^k\big\| > \mu C_k \theta^{i_k} \delta_k. \tag{5.1}$$

Then set $\delta_{k+1} := \theta^{i_k}\theta_k$.

Step 2 (linesearch). Set $t_k := \bar\tau$. While

$$f(x^k - t_k g^k) > f(x^k) - \beta t_k \big\|g^k\big\|^2 \text{ and } t_k \ge t_k^{\min}, \tag{5.2}$$

set $t_k := \gamma t_k$.

Step 3 (stepsize and parameters update). If $t_k \ge t_k^{\min}$, then set $\tau_k := t_k$, $C_{k+1} := C_k$, and $t_{k+1}^{\min} := t_k^{\min}$. Otherwise, set $\tau_k := 0$, $C_{k+1} := \eta C_k$, and $t_{k+1}^{\min} := \gamma t_k^{\min}$.

Step 4 (iteration update). Set $x^{k+1} := x^k - \tau_k g^k$. Increase k by 1 and go back to Step 1.

Remark 5.1. **(i)** Fix any $k \in \mathbb{N}$. The procedure of finding g^k and i_k that satisfies Step 1 of Algorithm 3 can be described as follows. Set $i_k := 0$ and calculate g^k as

$$g^k = \mathcal{G}\big(x^k, \min\{\theta^{i_k}\delta_k, \nu_k\}\big). \tag{5.3}$$

While $\|g^k\| \le \mu C_k \theta^{i_k}\delta_k$, increase i_k by 1 and recalculate g^k by formula (5.3). We now show that when $\nabla f(x^k) \neq 0$, this procedure stops after a *finite number of steps* giving us g^k and i_k as desired. Indeed, assuming on the contrary that the procedure does not stop, we get a sequence of $\{g_i^k\}$ with

$$g_i^k = \mathcal{G}\big(x^k, \min\{\theta^i\delta_k, \nu_k\}\big) \ \text{ and } \ \|g_i^k\| \le \mu C_k\theta^i\delta_k \ \text{ for all } \ i \in \mathbb{N}. \tag{5.4}$$

Since $\mathcal{G}$ is a local approximation of ∇f, for any fixed $\Delta > 0$ condition (3.2) with $\Omega = \{x^k\}$ ensures the existence of a positive number C such that

$$\|\mathcal{G}(x^k, \delta) - \nabla f(x^k)\| \le C\delta \ \text{ whenever } \ 0 < \delta \le \Delta. \tag{5.5}$$

By $\theta \in (0, 1)$, there is $N \in \mathbb{N}$ with $\theta^i\delta_k \le \Delta$ for all $i \ge N$. Combining this with (5.4) and (5.5) yields

$$\|g_i^k - \nabla f(x^k)\| \le C\theta^i\delta_k \ \text{ and } \ \|g_i^k\| \le \mu C_k\theta^i\delta_k \ \text{ for all } \ i \ge N.$$

Letting $i \to \infty$, we arrive at $\nabla f(x^k) = 0$, which is a contradiction.

(ii) It follows directly from the construction of δ_k in Step 1 of Algorithm 3 that

$$g^k = \mathcal{G}\big(x^k, \min\{\delta_{k+1}, \nu_k\}\big) \ \text{ and } \ \|g^k\| > \mu C_k\delta_{k+1}. \tag{5.6}$$

To proceed further with the convergence analysis of Algorithm 3, we obtain two results of their independent interest. The first one reveals some *uniformity* of general linesearch procedures with respect to the selections of reference points, stepsizes, and directions.

Lemma 5.2. *Let $f : \mathbb{R}^n \to \mathbb{R}$ be a function with a locally Lipschitz continuous gradient, and let $\beta \in (0, 1/2)$. Then for any nonempty bounded set $\Omega \subset \mathbb{R}^n$, there exists $\bar{t} > 0$ such that*

$$f(x - tg) \le f(x) - \beta t\|g\|^2 \quad \textit{whenever} \ \ x \in \Omega, \ 2\|g - \nabla f(x)\| \le \|g\|, \ \ \textit{and} \ \ t \in (0, \bar{t}].$$

Proof. The boundedness of Ω gives us $r > 0$ such that $\Omega \subset r\mathbb{B}$. Using the continuity of ∇f and the compactness of $r\mathbb{B}$, define $r' := \max\{\|\nabla f(x)\| \mid x \in r\mathbb{B}\}$. Since $f \in \mathcal{C}^{1,1}$, there exists $L > 0$ such that ∇f is Lipschitz continuous with the constant L on $\Theta := (r + 2r')\mathbb{B}$. By $\beta < 1/2$, we find $\bar{t} > 0$ with $\bar{t} < \min\{1, L^{-1}(1 - 2\beta)\}$. Now take some $x \in \Omega \subset \Theta$ and $g \in \mathbb{R}^n$ such that $2\|g - \nabla f(x)\| \le \|g\|$ and $t \in (0, \bar{t}]$. The choice of g gives us by the Cauchy-Schwarz inequality that

$$\begin{aligned} \langle \nabla f(x), g\rangle &= \langle \nabla f(x) - g, g\rangle + \|g\|^2 \ge -\|\nabla f(x) - g\|\,\|g\| + \|g\|^2 \\ &\ge -\tfrac{1}{2}\|g\|^2 + \|g\|^2 = \tfrac{1}{2}\|g\|^2, \end{aligned} \tag{5.7}$$

and by using the triangle inequality that

$$\|\nabla f(x)\| \ge \|g\| - \|g - \nabla f(x)\| \ge \tfrac{1}{2}\|g\|.$$

Combining the latter with the choice of t, $\bar{t}$, $x \in \Omega \subset r\mathbb{B}$ and the construction of r' yields

$$t\|g\| \le \bar{t}\|g\| \le 2\bar{t}\|\nabla f(x)\| \le 2\bar{t}r' < 2r',$$

which ensures that $x - tg \in \Theta$. The convexity of Θ tells us that the entire line segment $[x, x - tg]$ lies on Θ. Remembering that ∇f is Lipschitz continuous with the constant L on Θ, we employ Lemma 2.1 by taking into account that $t \le \bar{t} < L^{-1}(1 - 2\beta)$ and that (5.7). This gives us

$$\begin{aligned} f(x - tg) - f(x) &\le \langle x - tg - x, \nabla f(x)\rangle + \frac{L}{2}\|x - tg - x\|^2 \\ &= -t\langle g, \nabla f(x)\rangle + \frac{Lt^2}{2}\|g\|^2 \le -\frac{t}{2}\|g\|^2 + \frac{Lt^2}{2}\|g\|^2 \\ &= -\beta t\|g\|^2 + t\|g\|^2\,\frac{2\beta - 1 + Lt}{2} \le -\beta t\|g\|^2 \end{aligned}$$

and thus completes the proof of the lemma. □

Employing the obtained lemma, we derive the next result showing that unless the stationary point is found, Algorithm 3 always makes a progress after a finite number of iterations.

Proposition 5.3. *Let $\{x^k\}$ and $\{\tau_k\}$ be the sequences generated by Algorithm 3, and let $K \in \mathbb{N}$ be such that $\nabla f(x^K) \neq 0$. Then we can choose a number $N \geq K$ so that $\tau_N > 0$.*

Proof. Assume on the contrary that $\tau_k = 0$ for all $k \geq K$. Steps 3 and 4 of Algorithm 3 give us

$$t_{k+1}^{\min} = \gamma t_k^{\min} \text{ and } x^k = x^K \text{ for all } k \geq K. \tag{5.8}$$

Therefore, $\nabla f(x^k) = \nabla f(x^K) \neq 0$ for all $k \geq K$, which implies that g^k and i_k in Step 1 of Algorithm 3 exist for all $k \geq K$. Since $\mathcal{G}$ is a local approximation of ∇f, for any fixed $\Delta > 0$ condition (3.2) with $\Omega = \{x^K\}$ ensures the existence of $C > 0$ with

$$\left\|\mathcal{G}(x^K, \delta) - \nabla f(x^K)\right\| \leq C\delta \text{ whenever } 0 < \delta \leq \Delta. \tag{5.9}$$

It follows from Lemma 5.2 with $\Omega = \{x^K\}$ that there exists some $\bar{t} > 0$ such that

$$f(x^K - tg) \leq f(x^K) - \beta t \|g\|^2 \text{ whenever } 2\left\|g - \nabla f(x^K)\right\| \leq \|g\| \text{ and } t \in (0, \bar{t}]. \tag{5.10}$$

Using $\nu_k \downarrow 0$, $t_k^{\min} \downarrow 0$, $\nabla f(x^K) \neq 0$, and (5.8) gives us $N \geq K$ for which $\nu_N < \min\left\{\Delta, \frac{1}{3C}\left\|\nabla f(x^K)\right\|\right\}$ and $t_N^{\min} < \gamma\bar{t}$. Then we get from (5.9) with taking into account $x^N = x^K$ that

$$\left\|\mathcal{G}(x^N, \min\{\delta_{N+1}, \nu_N\}) - \nabla f(x^N)\right\| \leq C \min\{\delta_{N+1}, \nu_N\} \leq C\nu_N \leq \frac{1}{3}\left\|\nabla f(x^N)\right\|.$$

Combining this with $g^N = \mathcal{G}(x^N, \min\{\delta_{N+1}, \nu_N\})$ from (5.6) provides the estimate

$$\left\|g^N - \nabla f(x^N)\right\| \leq \frac{1}{3}\left\|\nabla f(x^N)\right\|,$$

which implies by the triangle inequality that

$$\left\|g^N\right\| \geq \left\|\nabla f(x^N)\right\| - \left\|g^N - \nabla f(x^N)\right\| \geq 2\left\|g^N - \nabla f(x^N)\right\|.$$

Employing the latter together with (5.10) and $x^N = x^K$ yields

$$f(x^N - tg^N) \leq f(x^N) - \beta t \left\|g^N\right\|^2 \text{ for all } t \in (0, \bar{t}]. \tag{5.11}$$

It follows from (5.11) and the choice of parameters that

$$\max\left\{t \mid f(x^N - tg^N) \leq f(x^N) - \beta t\left\|g^N\right\|^2, \ t = \bar{\tau}, \bar{\tau}\gamma, \bar{\tau}\gamma^2, \ldots\right\} > \gamma\bar{t} > t_N^{\min},$$

which implies in turn by Step 2 of Algorithm 3 that

$$t_N = \max\left\{t \mid f(x^N - tg^N) \leq f(x^N) - \beta t\left\|g^N\right\|^2, \ t = \bar{\tau}, \bar{\tau}\gamma, \bar{\tau}\gamma^2, \ldots\right\} > t_N^{\min}.$$

By Step 3 of Algorithm 3, we conclude that $\tau_N = t_N > 0$, a contradiction completing the proof. $\square$

Now we are ready the establish the convergence properties of Algorithm 3.

Theorem 5.4. *Let $\{x^k\}$ be the sequence generated by Algorithm 3 and assume that $\nabla f(x^k) \neq 0$ for all $k \in \mathbb{N}$. Then either $f(x^k) \to -\infty$ as $k \to \infty$, or the following assertions hold:*

(i) *Every accumulation point of $\{x^k\}$ is a stationary point of f.*

(ii) *If the sequence $\{x^k\}$ is bounded, then the set of accumulation points of $\{x^k\}$ is nonempty, compact, and connected in $\mathbb{R}^n$.*

(iii) *If $\{x^k\}$ has an isolated accumulation point $\bar{x}$, then this sequence converges to $\bar{x}$.*

Proof. First it follows from Steps 2 and 3 of Algorithm 3 that

$$\beta\tau_k\left\|g^k\right\|^2 \leq f(x^k) - f(x^{k+1}) \text{ for all } k \in \mathbb{N}, \tag{5.12}$$

which tells us that $\{f(x^k)\}$ is nonincreasing. If $f(x^k) \to -\infty$, there is nothing to prove; so we assume that $f(x^k) \not\to -\infty$, which implies by the nonincreasing property of $\{f(x^k)\}$ that $\inf f(x^k) > -\infty$. Summing up the inequalities in (5.12) over $k = 1, 2, \ldots$ with taking into account that $x^{k+1} = x^k - \tau_k g^k$ from the update in Step 3 of Algorithm 3 gives us

$$\sum_{k=1}^{\infty}\tau_k\left\|g^k\right\|^2 < \infty \text{ and } \sum_{k=1}^{\infty}\left\|g^k\right\| \cdot \left\|x^{k+1} - x^k\right\| < \infty. \tag{5.13}$$

We divide the proof of (i) into two parts by showing first that the origin is an accumulation point of $\{g^k\}$ and then employing Lemma 2.2 to establish stationarity of all the accumulation points of $\{x^k\}$.

Claim 1. *The origin $0 \in \mathbb{R}^n$ is an accumulation point of the sequence $\{g^k\}$.*

Arguing by contradiction, suppose that there are numbers $\varepsilon > 0$ and $K \in \mathbb{N}$ such that

$$\|g^k\| \geq \varepsilon \text{ for all } k \geq K. \tag{5.14}$$

Combining this with (5.13) gives us $\tau_k \downarrow 0$ and $\sum_{k=1}^{\infty} \|x^{k+1} - x^k\| < \infty$. The latter implies that $\{x^k\}$ converges to some $\bar{x} \in \mathbb{R}^n$. By taking a larger K, we can assume that $\tau_k < \bar{\tau}$ for all $k \geq K$. Let $\mathcal{N}$ be the set of all $k \in \mathbb{N}$ such that $\tau_k > 0$. It follows from Proposition 5.3 that $\mathcal{N}$ is infinite. Hence we can take any $k \geq K$ with $k \in \mathcal{N}$ and get that $\tau_k \in (0, \bar{\tau})$. Step 3 of Algorithm 3 ensures that $\tau_k = t_k \in [t_k^{\min}, \bar{\tau})$. Fixing such an index k, we get from the exit condition in Step 2 of Algorithm 3 that

$$-\gamma^{-1}\beta\tau_k \|g^k\|^2 < f(x^k - \gamma^{-1}\tau_k g^k) - f(x^k). \tag{5.15}$$

The classical mean value theorem gives us $\widetilde{x}^k \in [x^k, x^k - \gamma^{-1}\tau_k g^k]$ such that

$$f(x^k - \gamma^{-1}\tau_k g^k) - f(x^k) = -\gamma^{-1}\tau_k \langle g^k, \nabla f(\widetilde{x}^k)\rangle. \tag{5.16}$$

Combining this with (5.15) yields

$$-\gamma^{-1}\beta\tau_k \|g^k\|^2 < -\gamma^{-1}\tau_k \langle g^k, \nabla f(\widetilde{x}^k)\rangle,$$

which implies by dividing both sides of the inequality by $-\gamma^{-1}\tau_k < 0$ that

$$\langle g^k, \nabla f(\widetilde{x}^k)\rangle < \beta \|g^k\|^2 \quad \text{for all} \quad k \geq K, \ k \in \mathcal{N}. \tag{5.17}$$

Take some neighborhood Ω of $\bar{x}$ and $\Delta > 0$. Since $\mathcal{G}$ is a local approximation of ∇f under condition (3.2), there exists $C > 0$ such that

$$\|\mathcal{G}(x, \delta) - \nabla f(x)\| \leq C\delta \quad \text{whenever} \quad 0 < \delta \leq \Delta \quad \text{and} \quad x \in \Omega. \tag{5.18}$$

Since $\nu_k \downarrow 0$ and $x^k \to \bar{x}$, by taking a larger K we can assume that $\nu_k < \Delta$ and $x^k \in \Omega$ for all $k \geq K$. Using this together with (5.18) and $g^k = \mathcal{G}(x^k, \min\{\delta_{k+1}, \nu_k\})$ in (5.6) tells us that

$$\|g^k - \nabla f(x^k)\| = \|\mathcal{G}(x^k, \min\{\delta_{k+1}, \nu_k\}) - \nabla f(x^k)\| \leq C \min\{\delta_{k+1}, \nu_k\} \leq C\nu_k.$$

Combining the latter with $x^k \to \bar{x}$, $\nu_k \downarrow 0$ as $k \to \infty$, and the continuity of ∇f gives us

$$g^k \to \nabla f(\bar{x}) \text{ as } k \to \infty, \tag{5.19}$$

which yields $\|\nabla f(\bar{x})\| > 0$ by (5.14). It follows from (5.19), $\tau_k \downarrow 0$, $x^k \to \bar{x}$, and $\widetilde{x}^k \in [x^k, x^k - \gamma^{-1}\tau_k g^k]$ for all $k \geq K$ with $k \in \mathcal{N}$ that $\widetilde{x}^k \xrightarrow{\mathcal{N}} \bar{x}$. Letting $k \xrightarrow{\mathcal{N}} \infty$ in (5.17) and taking into account the convergence above and (5.19) bring us to the estimate

$$\|\nabla f(\bar{x})\|^2 \leq \beta \|\nabla f(\bar{x})\|^2.$$

This contradicts $\beta < \frac{1}{2}$ and $\|\nabla f(\bar{x})\| > 0$. Thus the origin is an accumulation point of $\{g^k\}$ as claimed.

Claim 2. *Every accumulation point of $\{x^k\}$ is a stationary point of f.*

Pick any accumulation point $\bar{x}$ of $\{x^k\}$. Using Claim 1, the second inequality in (5.13), and Lemma 2.2 tells us that there exists an infinite set $J \subset \mathbb{N}$ such that

$$x^k \xrightarrow{J} \bar{x} \text{ and } g^k \xrightarrow{J} 0.$$

Take a neighborhood Ω of $\bar{x}$ and $\Delta > 0$. Since $\mathcal{G}$ is a local approximation of ∇f under condition (3.2), there exists $C > 0$ for which

$$\|\mathcal{G}(x, \delta) - \nabla f(x)\| \leq C\delta \quad \text{whenever} \quad 0 < \delta \leq \Delta \quad \text{and} \quad x \in \Omega. \tag{5.20}$$

Since $\nu_k \downarrow 0$ and $x^k \xrightarrow{J} \bar{x}$, we can select $K \in \mathbb{N}$ so that $\nu_k \leq \Delta$ and $x^k \in \Omega$ for all $k \geq K$, $k \in J$. This ensures together with (5.20) that

$$\|g^k - \nabla f(x^k)\| = \|\mathcal{G}(x^k, \min\{\delta_{k+1}, \nu_k\}) - \nabla f(x^k)\| \leq C\nu_k \quad \text{for all} \quad k \geq K, \ k \in J.$$

Employing $g^k \xrightarrow{J} 0$ and $\nu_k \downarrow 0$ as above, we deduce that $\nabla f(x^k) \xrightarrow{J} 0$, and hence $\nabla f(\bar{x}) = 0$. Therefore, $\bar{x}$ is a stationary point of f, which justifies (i).

Now we verify (ii) and (iii) simultaneously. It follows from (5.13) and $\tau_k \le 1$ for all $k \in \mathbb{N}$ by the choice of τ_k in Step 3 of Algorithm 3 that

$$\sum_{k=1}^{\infty} \left\|x^{k+1} - x^k\right\|^2 = \sum_{k=1}^{\infty} \tau_k^2 \left\|g^k\right\|^2 \le \bar{\tau} \sum_{k=1}^{\infty} \tau_k \left\|g^k\right\|^2 < \infty,$$

which implies that $\left\|x^{k+1} - x^k\right\| \to 0$. Then both assertions (ii) and (iii) follow from Lemma 2.3. □

The next result establishes the global convergence with convergence rates of the iterates $\left\{x^k\right\}$ in Algorithm 3 under the KL property and the boundedness of $\left\{x^k\right\}$. We have already discussed the KL property in Remark 2.5. The boundedness of $\left\{x^k\right\}$ is also a standard assumption that appears in many works on gradient descent methods; see, e.g., [4, Theorem 4.1], [34, Theorem 1], and [40, Assumption 7].

Theorem 5.5. *Let $\left\{x^k\right\}$ be the sequence of iterates generated by Algorithm 3. Assuming that $\nabla f(x^k) \neq 0$ for all $k \in \mathbb{N}$ and that $\left\{x^k\right\}$ is bounded yields the assertions:*

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

(ii) *If in addition to* (i)*, the KL property at $\bar{x}$ is satisfied with $\psi(t) = Mt^q$ for some $M > 0, q \in [1/2, 1)$, then the following convergence rates are guaranteed:*

- *If $q = 1/2$, then the sequence $\left\{x^k\right\}$ converges linearly to $\bar{x}$.*
- *If $q \in (1/2, 1)$, then we have the estimate*

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

Proof. Let $\Omega := \left\{x^k\right\}$, and let $\Delta > 0$. Since $\mathcal{G}$ is a local approximation of ∇f satisfying condition (3.2), there exists a positive number C such that

$$\|\mathcal{G}(x, \delta) - \nabla f(x)\| \le C\delta \quad \text{whenever} \quad x \in \Omega \quad \text{and} \quad 0 < \delta \le \Delta. \tag{5.21}$$

Select $K \in \mathbb{N}$ so that $\nu_k < \Delta$ for all $k \ge K$, which implies by (5.21) and the choice of g^k in Step 1 of Algorithm 3 the relationships

$$\begin{aligned}\left\|g^k - \nabla f(x^k)\right\| &= \left\|\mathcal{G}(x^k, \min\{\delta_{k+1}, \nu_k\}) - \nabla f(x^k)\right\| \\ &\le C \min\{\delta_{k+1}, \nu_k\} \le C\delta_{k+1} \quad \text{for all} \quad k \ge K.\end{aligned} \tag{5.22}$$

We split the proof of the result into two parts by showing first that the sequences $\{C_k\}$ and $\left\{t_k^{\min}\right\}$ are constant after a finite number of iterations and verifying then the convergence of $\left\{x^k\right\}$ in (i) with the rates in (ii) by using Propositions 2.6 and 2.7.

Claim 1. *There exists $k_0 \in \mathbb{N}$ such that $C_k = C_{k_0}$ and $t_k^{\min} = t_{k_0}^{\min}$ for all $k \ge k_0$.*

Arguing by contradiction, suppose that such a number k_0 does not exist. By the construction of $\{C_k\}$ and $\left\{t_k^{\min}\right\}$ in Step 3 of Algorithm 3, we deduce that $C_k \uparrow \infty$ and $t_k^{\min} \downarrow 0$ as $k \to \infty$. Since Ω is bounded, Lemma 5.2 allows us to find $\bar{t} \in (0, 1)$ for which

$$f(x - tg) \le f(x) - \beta t \|g\|^2 \quad \text{whenever} \quad x \in \Omega, \ \|g - \nabla f(x)\| \le \frac{1}{2}\|g\|, \quad \text{and} \quad t \in (0, \bar{t}]. \tag{5.23}$$

Using the aforementioned properties of $\{C_k\}$ and $\{t_k^{\min}\}$, we get $N \ge K$ such that $C_k > C$ and $t_k^{\min} < \gamma\bar{t}$ for all $k \ge N$. Fix such a number k and then combine the condition $\left\|g^k\right\| > \mu C_k \delta_{k+1}$ from (5.1) with $C_k > C$, $\mu > 2$, and (5.22). This gives us the inequalities

$$\left\|g^k\right\| > \mu C_k \delta_{k+1} \ge \mu C \delta_{k+1} \ge 2\left\|g^k - \nabla f(x^k)\right\|,$$

which imply together with $x^k \in \Omega$ and (5.23) the estimate

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

and thus tell us that $t_k > \gamma\bar{t} > t_k^{\min}$. Employing Step 3 of Algorithm 3 yields $t_{k+1}^{\min} = t_k^{\min}$. Since the latter holds whenever $k \ge N$, we conclude that the equality $t_k^{\min} = t_N^{\min}$ is satisfied for all $k \ge N$. This contradicts the condition $t_k^{\min} \downarrow 0$ as $k \to \infty$ and hence justifies the claimed assertion.

Claim 2. *All the assertions in* (i) *and* (ii) *are fulfilled.*

From Step 2 and Step 3 of Algorithm 3, we deduce that

$$f(x^k) - f(x^{k+1}) \geq \beta\tau_k \left\|g^k\right\|^2 \quad \text{for all} \quad k \in \mathbb{N}. \tag{5.24}$$

Defining $N := \max\{K, k_0\}$ with k_0 taken from Claim 1 gives us the equalities

$$C_k = C_N \quad \text{and} \quad t_k^{\min} = t_N^{\min} \quad \text{whenever} \quad k \geq N. \tag{5.25}$$

Combining $C_k = C_N$ with (5.22) and $\|g^k\| > \mu C_k \delta_{k+1}$ from (5.1) ensures that

$$\begin{aligned}\|\nabla f(x^k)\| &\leq \|g^k\| + C\delta_{k+1} \\ &\leq \|g^k\| + \frac{C}{\mu C_N}\|g^k\| = \alpha\|g^k\| \quad \text{for all} \quad k \geq N,\end{aligned} \tag{5.26}$$

where $\alpha := 1 + \frac{C}{\mu C_N}$. In addition, we have $t_{k+1}^{\min} = t_k^{\min} = t_N^{\min}$ in (5.25), which implies together with Step 3 of Algorithm 3 the relationships

$$\tau_k = t_k \geq t_k^{\min} = t_N^{\min} \quad \text{as} \quad k \geq N \tag{5.27}$$

confirming the boundedness of $\{\tau_k\}$ from below. If the KL property of f holds at the accumulation point $\bar{x}$ of $\{x^k\}$, it follows from Remark 2.8(i), (5.24), (5.26), and (5.27) that assumptions (H1) and (H2) in Proposition 2.6 hold. Thus $x^k \to \bar{x}$ as $k \to \infty$, which verifies (i).

Assume finally that the KL property at $\bar{x}$ is satisfied with $\psi(t) = Mt^q$, $M > 0$, and $q \in [1/2, 1)$. The iterative procedure $x^{k+1} = x^k - \tau_k g^k$ in Step 4 of Algorithm 3 together with (5.27) and $g^k > 0$ from Step 1 therein tells us that $x^{k+1} \neq x^k$ for $k \geq N$. Combining this with (5.24), (5.26), and (5.27) verifies all the assumptions of Proposition 2.7 and therefore completes the proof of the theorem. □

5.2 Bidirectional Linesearch for Noisy Functions

In this subsection, we continue the study of problem (1.1) with the objective function $f : \mathbb{R}^n \to \mathbb{R}$ of class $\mathcal{C}^{1,1}$. Similarly to Subsection 4.3, assume that only a *noisy approximation* $\phi(x) = f(x) + \xi(x)$ of f is available, where $\xi : \mathbb{R}^n \to \mathbb{R}$ is a noise function bounded by some known constant $\xi_f > 0$. However, differently from Subsection 4.3 dealing with $\mathcal{C}_L^{1,1}$ functions, we suppose here that the noise level ξ_f is known. Considering only the forward finite difference given by

$$\widetilde{\mathcal{G}}(x, \delta) = \frac{1}{\delta}\sum_{i=1}^{n}\big(\phi(x + \delta e_i) - \phi(x)\big)e_i \quad \text{for any} \quad (x, \delta) \in \mathbb{R}^n \times (0, \infty), \tag{5.28}$$

we state the following noisy version of Proposition 3.4 that can be verified similarly.

Proposition 5.6. *Let $f : \mathbb{R}^n \to \mathbb{R}$ be a $\mathcal{C}^1$-smooth function such that ∇f is Lipschitz continuous on $\overline{\mathbb{B}}(x, \Delta)$ with constant $\ell > 0$. Then the noisy forward finite difference (5.28) satisfies the error bound*

$$\left\|\widetilde{\mathcal{G}}(x, \delta) - \nabla f(x)\right\| \leq \frac{\ell\sqrt{n}\delta}{2} + \frac{2\sqrt{n}\xi_f}{\delta} \quad \textit{for all} \quad \delta \in (0, \Delta]. \tag{5.29}$$

Now we are ready to design the algorithm, which main feature is to be a *bidirectional linesearch*, for determining both the *stepsize* and the *finite difference interval.*

Algorithm 4 (DFBD for noisy functions)**.**

Step 0 (initialization)**.** Select an initial point $x^1 \in \mathbb{R}^n$, $\eta > 1$, and $L_1 > 0$. Set $k := 1$.

Step 1 (bidirectional linesearch)**.** Find $i_k \in \mathbb{Z}$ with the smallest absolute value such that for $g^k := \widetilde{\mathcal{G}}\left(x^k, \sqrt{\frac{4\xi_f}{\eta^{i_k}L_k}}\right)$ and $\tau_k = \frac{1}{\eta^{i_k}L_k}$, it holds that

$$\phi\left(x^k - \tau_k g^k\right) \leq \phi(x^k) - \frac{\tau_k}{9}\left\|g^k\right\|^2. \tag{5.30}$$

Step 2 (stepsize and parameters update)**.** Set $x^{k+1} := x^k - \tau_k g^k$ and $L_{k+1} := \eta^{i_k}L_k$.

Remark 5.7 (Discussion on bidirectional linesearch)**.**

- In Step 1 of Algorithm 4, we employ bidirectional linesearch to find an approximation $L_{k+1} := \eta^{i_k}L_k$ for the Lipschitz constant ℓ_k of ∇f locally around the current iterate x^k. Then we use L_{k+1} to determine both the stepsize $\tau_k = \frac{1}{L_{k+1}}$ and the finite difference interval $\sqrt{\frac{4\xi_f}{L_{k+1}}}$, which is the minimizer of the right-hand side of (5.29) with respect to δ when $\ell = L_{k+1}$.

- The idea of using bidirectional linesearch to adjust the stepsize is not new as it has been already employed in the recent works [22, 53] for gradient descent methods with the exact gradient, and in [7, 8] for derivative-free optimization methods. To the best of our knowledge, using this procedure to determine both *stepsize* and *finite-difference interval* appears for the first time in this paper.
- The bidirectional linesearch procedure plays an important role not only in our convergence analysis but also in practical modeling. *Theoretically*, condition (5.30) is necessary for the value $\eta^{i_k} L_k$ in Step 1 being a good approximation of the Lipschitz constant ℓ_k of ∇f locally around the reference iterate x^k, which is confirmed by Proposition 5.9. *Numerically*, by automatically approximating the local Lipschitz constant of the gradient, our DFBD has a better performance in comparison with other finite-difference-based algorithms for noisy $\mathcal{C}^{1,1}$ functions with complex structures. To illustrate this claim, we consider the function $f(x, y) := (e^{2x+3y-1} + e^{3x-y} + e^{x-y-6} - 3)^2$ with the graph in $\mathbb{R}^3$ depicted below.

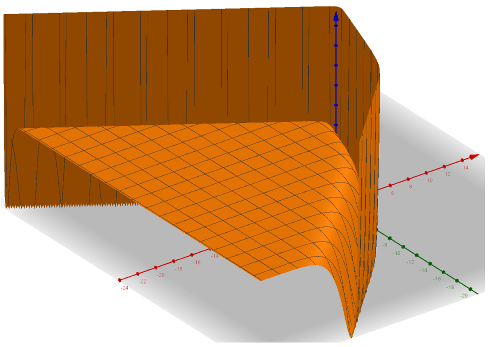


Figure 1: Graph of $f(x, y) = (e^{2x+3y-1} + e^{3x-y} + e^{x-y-6} - 3)^2$

This function is inspired by a univariate function in [56, Section 4.1.2]. It provides a challenging example for finite-difference-based methods in both cases of approximate gradients and finding minimizers. This is because f has very small first- and second-order derivatives at points belonging to most of the second and third parts of the plane. As in the context of noisy DFO, we manually inject into f uniformly distributed stochastic noises with different levels. The plots below show the trajectories of iterates generated by our DFBD (Algorithm 4) and the following algorithms:

- IMFIL: The implicit filtering algorithm [24, Algorithm 2.2] with the forward finite difference.
- RG: The random gradient-free method [46, Section 5].
- L-BGFS (Ada): The noise-tolerant quasi-Newton algorithm [57, Algorithm 2.1], where the gradient is approximated by the forward finite difference with the adaptive finite difference interval estimation from [56, Algorithm 2.1].
- GD-BD (Ada): Gradient descent with bidirectional linesearch, where the gradient is approximated by the forward finite difference with the adaptive finite difference interval estimation from [56, Algorithm 2.1].
- DF-backtracking: A modified version of our basic DFBD, where the bidirectional linesearch is replaced by the standard backtracking linesearch, i.e., the condition $i \in \mathbb{Z}$ is replaced by $i \in \mathbb{N}$ in Step 1 of Algorithm 4.

In the algorithms above, only our DFBD method (Algorithm 4) uses the bidirectional linesearch to determine both *stepsize* and *finite difference interval*. The selections of stepsize and finite difference interval for each method are listed in Table 1, where GS in the selections of RG means grid search. Details for the settings of the algorithms and additional numerical results on this experiment can be found in Appendix A, where different noise levels are addressed.

Method	IMFIL	RG	GD-BD (Ada)	L-BFGS (Ada)	DF-backtracking	DFBD
Stepsize	Backtracking	GS	Bidirectional	Armijo + Wolfe	Backtracking	Bidirectional
FD interval	Decreasing	GS	Adaptive	Adaptive	Backtracking	Bidirectional

Table 1: Stepsize and finite difference interval selections

It can be observed from Figure 2 addressing the noise level 0.01 that only the last points (red stars) generated by our DFBD method successfully identify the minimum region (depicted in dark blue) regardless of the choice of initial points (blue circles). L-BFGS (Ada) locates the minimum region in the only case when the initial point is $(-4,-4)$. Other algorithms including RG, IMFIL, GD-BD (Ada), and DF-backtracking perform even worse since they remain stuck at the initial points in all the scenarios. Additional graphs in Appendix A also show that these results are stable with respect to different levels of noise ranging from 1 or 10^{-3}. The failure of GD-BD (Ada) and DF-backtracking emphasizes the crucial role of using bidirectional linesearch to determine both the stepsize and the finite difference interval in the construction of DFBD.

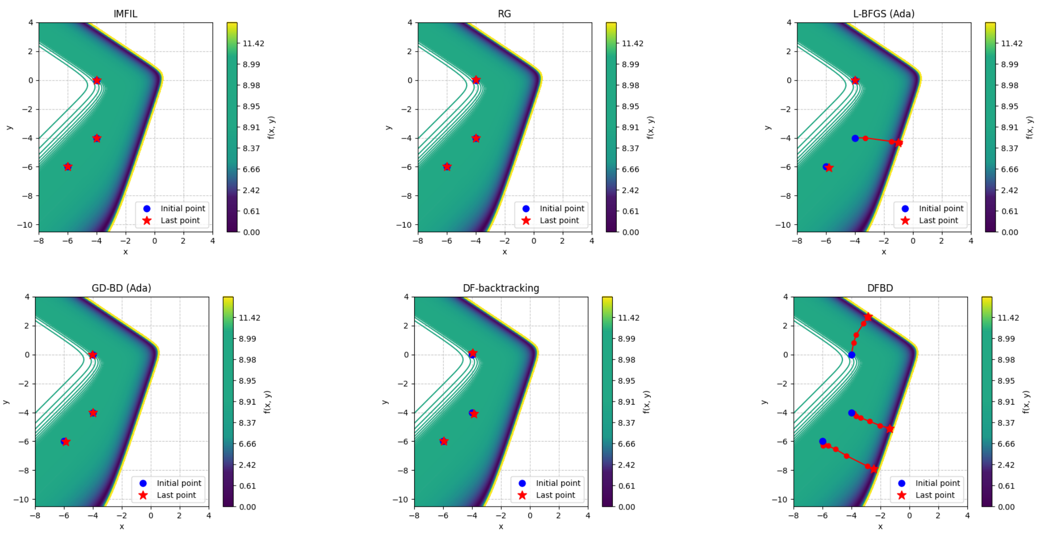


Figure 2: Finite-difference-based methods on minimizing a $\mathcal{C}^{1,1}$ function with complex structure

The rest of this subsection is devoted to deriving the fundamental convergence properties of Algorithm 4 for noisy smooth functions. We begin with a simple albeit useful lemma about the *optimal local Lipschitz constant* of the gradient of a $\mathcal{C}^{1,1}$ function.

Lemma 5.8. *Let $\Omega \subset \mathbb{R}^n$ be a nonempty bounded set. Then for any $\xi > 0$, there exists some $\ell > 0$ such that ℓ is the Lipschitz constant of ∇f on $\bigcup_{x\in\Omega} \mathbb{B}(x, \delta_x)$, where*

$$\delta_x := \max\left\{\frac{3}{2\ell}\left\|\nabla f(x)\right\|, \sqrt{\frac{4\xi}{\ell}}\right\}.$$

Proof. Define the number $M := \sup\left\{\left\|\nabla f(x)\right\| \;\middle|\; x \in \Omega\right\} \in \mathbb{R}$ and deduce from the assumed $\mathcal{C}^{1,1}$ property of f that f is Lipschitz continuous on the set $\bigcup_{x\in\Omega} \mathbb{B}(x, \max\left\{\frac{3}{2}M, 2\sqrt{\xi}\right\})$ with some Lipschitz constant $L > 0$. Denoting $\ell := \max\{1, L\}$, we get

$$\delta_x = \max\left\{\frac{3}{2\ell}\left\|\nabla f(x)\right\|, \sqrt{\frac{4\xi}{\ell}}\right\} \le \max\left\{\frac{3}{2}M, 2\sqrt{\xi}\right\} \quad \text{for all} \quad x \in \Omega.$$

This tells us that ∇f is Lipschitz continuous with the constant ℓ on $\bigcup_{x\in\Omega} \mathbb{B}(x, \delta_x)$ as claimed. □

The next result plays a crucial technical role in deriving the convergence properties in what follows.

Proposition 5.9. *Let $\ell > 0$ and $x \in \mathbb{R}^n$ be such that ∇f is Lipschitz continuous with some constant $\ell > 0$ on $\mathbb{B}\big(x, \max\left\{\frac{3}{2\ell}\left\|\nabla f(x)\right\|, \sqrt{\frac{4\xi_f}{\ell}}\right\}\big)$, and let $\widetilde{\ell} > 0$, $i \in \mathbb{Z}$, $\eta > 1$ be selected so that $\ell \in (\eta^{i-1}\widetilde{\ell}, \eta^i\widetilde{\ell}]$. Define $g \in \mathbb{R}^n$ and $\tau > 0$ by*

$$g := \widetilde{\mathcal{G}}\left(x, \sqrt{\frac{4\xi_f}{\eta^i\widetilde{\ell}}}\right) \quad \text{and} \quad \tau := \frac{1}{\eta^i\widetilde{\ell}},$$

where $\widetilde{\mathcal{G}}$ is taken from (5.28). If $\left\|\nabla f(x)\right\| \ge 8\sqrt{\ell\eta n\xi_f}$, then we have the estimates

(i) $f(x - \tau g) \le f(x) - \frac{3\tau}{32}\left\|\nabla f(x)\right\|^2$,

(ii) $\phi(x - \tau g) \le \phi(x) - \frac{\tau}{9}\left\|g\right\|^2$.

Proof. Since ℓ is the Lipschitz constant of ∇f on $\mathbb{B}\left(x, \sqrt{\frac{4\xi_f}{\ell}}\right)$, we deduce from Proposition 5.28 that

$$\left\|\widetilde{\mathcal{G}}(x,\delta) - \nabla f(x)\right\| \le \frac{\ell\sqrt{n}\delta}{2} + \frac{2\sqrt{n}\xi_f}{\delta} \text{ for all } \delta \in \left(0, \sqrt{\frac{4\xi_f}{\ell}}\right]. \tag{5.31}$$

Combining this with $g = \widetilde{\mathcal{G}}\left(x, \sqrt{\frac{4\xi_f}{\eta^i L}}\right)$ and $\ell \le \eta^i L$ tells us that

$$\begin{aligned}\|g - \nabla f(x)\| &\le \frac{\ell\sqrt{n}}{2}\sqrt{\frac{4\xi_f}{\eta^i L}} + 2\sqrt{n}\xi_f\sqrt{\frac{\eta^i\widetilde{\ell}}{4\xi_f}} \\ &\le \frac{\eta^i\widetilde{\ell}\sqrt{n}}{2}\sqrt{\frac{4\xi_f}{\eta^i L}} + 2\sqrt{n}\xi_f\sqrt{\frac{\eta^i L}{4\xi_f}} = 2\sqrt{\eta^i\widetilde{\ell}n\xi_f}.\end{aligned}$$

Using the triangle inequality and $\eta^{i-1}\widetilde{\ell} < \ell$ yields

$$\begin{aligned}\|g\| &\ge \|\nabla f(x)\| - \|g - \nabla f(x)\| \\ &\ge 8\sqrt{\eta\ell n\xi_f} - 2\sqrt{\eta^i\widetilde{\ell}n\xi_f} \\ &> 6\sqrt{\eta^i\widetilde{\ell}n\xi_f} \ge 3\|g - \nabla f(x)\|,\end{aligned} \tag{5.32}$$

which being combined with the Cauchy-Schwarz inequality ensures that

$$\begin{aligned}\langle\nabla f(x), g\rangle &= \langle\nabla f(x) - g, g\rangle + \|g\|^2 \\ &\ge -\|\nabla f(x) - g\|\,\|g\| + \|g\|^2 \ge \frac{2}{3}\|g\|^2.\end{aligned}$$

Thus we arrive at $\|g\| \le \frac{3}{2}\|\nabla f(x)\|$ implying together with $\tau = \frac{1}{\eta^i L} \le \frac{1}{\ell}$ that

$$x - \tau g \in \mathbb{B}\Big(x, \frac{3}{2\eta^i L}\|\nabla f(x)\|\Big) \subset \mathbb{B}\Big(x, \frac{3}{2\ell}\|\nabla f(x)\|\Big).$$

By the Lipschitz continuity of ∇f with constant ℓ on the ball above and Lemma 2.1, we get

$$\begin{aligned}f(x - \tau g) &\le f(x) + \langle x - \tau g - x, \nabla f(x)\rangle + \tfrac{\ell}{2}\|x - \tau g - x\|^2 \\ &= f(x) - \tau\langle g, \nabla f(x)\rangle + \tfrac{\ell\tau^2}{2}\|g\|^2 \\ &\le f(x) - \tfrac{2\tau}{3}\|g\|^2 + \tfrac{\tau}{2}\|g\|^2 = f(x) - \tfrac{\tau}{6}\|g\|^2.\end{aligned} \tag{5.33}$$

It also follows from (5.32) that

$$\|g\| \ge \|\nabla f(x)\| - \|g - \nabla f(x)\| \ge \|\nabla f(x)\| - \frac{1}{3}\|g\|,$$

which yields $\|\nabla f(x)\| \le \frac{4}{3}\|g\|$ and, being combined with (5.33), verifies (i).

(ii) Using (5.33) and the construction of the noisy approximation ϕ gives us the estimate

$$\phi(x - \tau g) \le \phi(x) - \frac{\tau}{6}\|g\|^2 + 2\xi_f, \tag{5.34}$$

which implies together with $\eta^{i-1}\widetilde{\ell} < \ell$, $n \ge 1$, and $\|\nabla f(x)\| \ge 8\sqrt{\ell\eta n\xi_f}$ that

$$\frac{\tau}{18}\|g\|^2 \ge \frac{1}{18\eta^i\widetilde{\ell}}\frac{9}{16}\|\nabla f(x)\|^2 \ge \frac{2}{64\eta^i\widetilde{\ell}}64\eta\ell n\xi_f \ge 2\xi_f.$$

Combining the latter with (5.34) leads us the conclusion in (ii) and thus completes the proof. $\square$

Similarly to Subsection 4.3, we say that Step 1 of Algorithm 4 is *successful* if the integer number i_k is found, and *unsuccessful* otherwise. It follows directly from Proposition 5.9 that Step 1 of Algorithm 4 is successful whenever $\left\|\nabla f(x^k)\right\|$ is not near 0 as stated below.

Corollary 5.10. *At the k^{th} iteration of Algorithm 4, let ℓ_k be such that ∇f is Lipschitz continuous on $\mathbb{B}\big(x^k, \max\big\{\frac{3}{2\ell_k}\|\nabla f(x^k)\|, \sqrt{\frac{4\xi_f}{\ell_k}}\big\}\big)$ with some constant $\ell_k > 0$. If the condition*

$$\left\|\nabla f(x^k)\right\| \ge 8\sqrt{\ell_k\eta n\xi_f} \tag{5.35}$$

is satisfied, then Step 1 of Algorithm 4 is successful.

Employing the obtained corollary, we arrive at the next proposition, which is useful in the proof of the main convergence results below.

Proposition 5.11. *At some k^{th} iteration of Algorithm 4, let $L > 0$ be such that ∇f is Lipschitz continuous with constant L on $\mathbb{B}\big(x^k, \max\big\{\frac{3}{2L}\left\|\nabla f(x^k)\right\|, \sqrt{\frac{4\xi_f}{L}}\big\}\big)$ and assume that*

$$\left\|\nabla f(x^k)\right\| \geq 8\sqrt{L\eta n \xi_f}.$$

The following assertions hold:

(i) *If $L_k < \eta L$ then $L_{k+1} < \eta L$.*

(ii) *If $L_k \geq L$ then $L_{k+1} \geq L$.*

(iii) *If $L_k \in [L, \eta L)$ then $L_{k+1} = L_k$.*

Proof. (i) By the construction of $\{L_k\}$, we find $m \in \mathbb{Z}$ such that $L_{k+1} = \eta^m L_k$. If $m \leq 0$, then $L_{k+1} \leq L_k < \eta L$, and so we assume that $m > 0$. Then the exit condition in Step 1 of Algorithm 4 yields

$$\phi\Big(x^k - \frac{1}{\eta^i L_k} g_i^k\Big) > \phi(x^k) - \frac{1}{\eta^i L_k}\left\|g_i^k\right\|^2 \text{ for all } i \in \{0, \ldots, m-1\}, \tag{5.36}$$

where $g_i^k := \widetilde{\mathcal{G}}\big(x^k, \sqrt{\frac{4\varepsilon_f}{\eta^i L_k}}\big)$. Observe that condition (5.35) holds for $\ell_k = L$, which is a Lipschitz constant of ∇f on $\mathbb{B}\Big(x^k, \max\{\frac{3}{2L}\left\|\nabla f(x^k)\right\|, \sqrt{\frac{4\xi_f}{L}}\}\Big)$ by the assumptions made. Combining this with Corollary 5.10 and estimate (5.36), we deduce that $\eta^i L_k \notin [L, \eta L)$ for all $i \in \{0, \ldots, m-1\}$. This fact together with $L_k < \eta L$ tells us that $L_{k+1} = \eta^m L_k < \eta L$.

(ii) By the construction of $\{L_k\}$, we find some $m \in \mathbb{Z}$ such that $L_{k+1} = \eta^m L_k$. If $m \geq 0$, then $L_{k+1} \geq L_k \geq L$, and so we assume that $m < 0$. Then the exit condition in Step 1 of Algorithm 4 yields

$$\phi\Big(x^k - \frac{1}{\eta^i L_k} g_i^k\Big) > \phi(x^k) - \frac{1}{\eta^i L_k}\left\|g_i^k\right\|^2 \text{ for all } i \in \{0, -1, \ldots, m+1\}, \tag{5.37}$$

where $g_i^k := \widetilde{\mathcal{G}}\big(x^k, \sqrt{\frac{4\varepsilon_f}{\eta^i L_k}}\big)$. Observe that condition (5.35) holds for $\ell_k = L$, which is also a Lipschitz constant of ∇f on $\mathbb{B}\big(x^k, \max\big\{\frac{3}{2L}\left\|\nabla f(x^k)\right\|, \sqrt{\frac{4\xi_f}{L}}\big\}\big)$ by the assumptions made. Combining this with Corollary 5.10 and (5.37), we get that $\eta^i L_k \notin [L, \eta L)$ for all $i \in \{0, -1, \ldots, m+1\}$. This fact together with $L_k \geq L$ verifies that $L_{k+1} = \eta^m L_k \geq L$. $\square$

Now we in a position to derive convergence properties of DFBD from Algorithm 4. Consider first the case where at some K^{th} iteration, Step 1 of Algorithm 4 is not successful, i.e., we cannot find $i_K \in \mathbb{Z}$ that ensures the descent condition (5.30). Then Corollary 5.10 tells us that $\left\|\nabla f(x^K)\right\| < 8\sqrt{\ell_K \eta n \xi_f}$, where ℓ_K is a Lipschitz constant of ∇f around x^K. In this case, Algorithm 4 finds a point near a stationary one after a finite number of iteration. In practice, to avoid the process of finding i_k in Step 1 of Algorithm 4 from running infinitely to cause a computational error, the users can add a lower bound sufficiently small and an upper bound sufficiently large for i_k in the loop.

The main theorem of this section concerns the case where Step 1 of Algorithm 4 is successful for all $k \in \mathbb{N}$. In this scenario, we can find a point near a stationary one, along the sequence of iterates generated by the algorithm, if just one of the Lipschitz approximations is appropriate.

Theorem 5.12. *Assume that Step 1 of Algorithm 4 is successful for all $k \in \mathbb{N}$ and that there exists $L > 0$ such that ∇f is Lipschitz continuous on $\bigcup_{k=1}^{\infty} \mathbb{B}\big(x^k, \max\big\{\frac{3}{2L}\left\|\nabla f(x^k)\right\|, \sqrt{\frac{4\xi_f}{L}}\big\}\big)$. If $\inf_{k\in\mathbb{N}} f(x^k) > -\infty$ and for some $K \in \mathbb{N}$ we have $L_K \in [L, \eta L)$, then the following assertions hold:*

(i) *There exists $N \in \mathbb{N}$ for which*

$$\left\|\nabla f(x^N)\right\| < 8\sqrt{L\eta n \xi_f}. \tag{5.38}$$

(ii) *Assume in addition that f has a global minimizer with the minimum value f^*, that $f(x^K) > f^*$, and that f satisfies the Polyak-Łojasiewicz inequality with some constant $\mu > 0$, i.e.,*

$$\mu(f(x) - f^*) \leq \frac{1}{2}\left\|\nabla f(x)\right\|^2 \quad \textit{for all } \; x \in \mathbb{R}^n. \tag{5.39}$$

Then the number N from (5.38) admits the upper estimate

$$N \leq \max\Big\{1 + K, 1 + K + \log_{1-\frac{3\mu}{16\eta L}}\Big(\frac{32\eta n \xi_f}{f(x^K) - f^*}\Big)\Big\}. \tag{5.40}$$

Proof. (i) Assume on the contrary that $\|\nabla f(x^k)\| \geq 8\sqrt{L\eta n\xi_f}$ as $k \in \mathbb{N}$. It follows from Proposition 5.11 and $L_K \in [L, \eta L)$ that $L_{k+1} = L_k$ whenever $k \geq K$. Using Proposition 5.9(i) with

$$\ell := L, \quad x := x^k, \quad \widetilde{\ell} := L_k, \quad \text{and} \quad i := 0,$$

we get the relationship below between two subsequent iterations

$$f(x^{k+1}) \leq f(x^k) - \frac{3}{32L_K}\|\nabla f(x^k)\| \quad \text{whenever} \quad k \geq K,$$

which tells us that $\{f(x^k)\}$ is a strictly decreasing sequence. By $\inf_{k\in\mathbb{N}} f(x^k) > -\infty$, this sequence is convergent, and hence $\nabla f(x^k) \to 0$ as $k \to \infty$. We arrive at a contradiction with $\|\nabla f(x^k)\| \geq 8\sqrt{L\eta n\xi_f}$ for all $k \in \mathbb{N}$, and thus justify (5.38) in (i).

To verify now assertion (ii), let N be the first iteration for which (5.38) holds, i.e.,

$$\|\nabla f(x^k)\| \geq 8\sqrt{L\eta n\xi_f} \text{ for } k \in \{1, \ldots, N-1\}.$$

If $N \leq K + 1$, estimate (5.40) is obviously satisfied, and thus we suppose that $N > K + 1$. It follows from Proposition 5.11 that $L_k = L_K \in [L, \eta L)$ for all $k \in \{K, \ldots, N-1\}$. Fixing such a number k and employing Proposition 5.9(i) for

$$\ell := L, \quad x := x^k, \quad \widetilde{\ell} := L_k, \quad \text{and} \quad i := 0$$

clearly bring us to the estimates

$$f(x^{k+1}) \leq f(x^k) - \frac{3}{32L_K}\|\nabla f(x^k)\|^2 \leq f(x^k) - \frac{3}{32\eta L}\|\nabla f(x^k)\|^2.$$

Combining this with the Polyak-Łojasiewicz inequality from (5.39), we obtain the condition

$$f(x^{k+1}) \leq f(x^k) - \frac{3\mu}{16\eta L}(f(x^k) - f^*),$$

which can be equivalently rewritten as

$$f(x^{k+1}) - f^* \leq \left(1 - \frac{3\mu}{16\eta L}\right)(f(x^k) - f^*).$$

Using the latter condition for $k = K, K+1, \ldots, N-2$ gives us

$$f(x^{N-1}) - f^* \leq \left(1 - \frac{3\mu}{16\eta L}\right)^{N-1-K}(f(x^K) - f^*). \tag{5.41}$$

Since ∇f is Lipschitz continuous on $\overline{\mathbb{B}}\left(x^{N-1}, \frac{1}{L}\nabla f(x^{N-1})\right)$ with constant L, Lemma 2.1 yields

$$\begin{aligned}
f^* &\leq f\left(x^{N-1} - \frac{1}{L}\nabla f(x^{N-1})\right) \\
&\leq f(x^{N-1}) + \left\langle x^{N-1} - \frac{1}{L}\nabla f(x^{N-1}) - x^{N-1}, \nabla f(x^{N-1})\right\rangle + \frac{L}{2}\left\|x^{N-1} - \frac{1}{L}\nabla f(x^{N-1}) - x^{N-1}\right\|^2 \\
&= f(x^{N-1}) - \frac{1}{L}\|\nabla f(x^{N-1})\|^2 + \frac{1}{2L}\|\nabla f(x^{N-1})\|^2,
\end{aligned}$$

which ensures in turn the fulfillment of

$$f(x^{N-1}) - f^* \geq \frac{1}{2L}\|\nabla f(x^{N-1})\|^2 \geq \frac{1}{2L}64L\eta n\xi_f = 32\eta n\xi_f.$$

Combining the obtained estimates with (5.41) tells us that

$$32\eta n\xi_f \leq \left(1 - \frac{3\mu}{16\eta L}\right)^{N-1-K}(f(x^K) - f^*),$$

and thus verifies the claimed conclusion (5.40). □

Remark 5.13. The existence of the constant L in the assumptions of Theorem 5.12 is guaranteed under the fulfillment of either one of the following conditions:

- The objective function f is of class $\mathcal{C}^{1,1}_L$.
- The level set $\{x \mid f(x) \leq f(x^1) + 2\xi_f\}$ is bounded. Indeed, it follows from (5.30) that the sets

  $$\{x^k\} \subset \{x \in \mathbb{R}^n \mid \phi(x) \leq \phi(x^1)\} \subset \{x \in \mathbb{R}^n \mid f(x) \leq f(x^1) + 2\xi_f\}$$

  are bounded as well. Combining this with Proposition 5.8 for $\Omega := \{x^k\}$ verifies the existence of L.

6 Numerical Experiments

In this section, we present numerical experiments demonstrating the efficiency of our methods in solving derivative-free optimization problems with and without the presence of noise. This section is split into two subsections addressing different noise levels: small noise, which also includes the noiseless case, and large noise. For each type of the noise level, we compare the performance of our newly developed methods with various well-known algorithms to ensure the diversity of the numerical experiments. In total, 786 test problems and 10 algorithms are considered in what follows.

6.1 Finite-Difference-Based Algorithms for Functions with Small Noise

Here we compare the performance of our DFC (Algorithm 1) and DFB (Algorithm 3) methods with other finite-difference-based algorithms to minimize smooth (convex and nonconvex) functions either without noise, or with small noise. The results in this subsection suggest that, in addition to the theoretical guarantees, our methods are more robust than the standard implementations of gradient descent methods with a *constant/backtracking stepsize* and with finite difference gradient for a *fixed finite difference interval*. The presented results also confirm the practicality of DFC and DFB methods in comparison with other well-known algorithms as in [24, 46].

6.1.1 Experiments with $\mathcal{C}_L^{1,1}$ Functions

The first part of the subsection compares the performance of our DFC method using forward finite differences with some other well-known derivative-free methods for minimizing $\mathcal{C}_L^{1,1}$ functions. Since our DFC method is of the gradient descent type, we choose the set of testing algorithms as follows:

(i) GDC (fixed), i.e., the standard gradient descent with a *constant stepsize* and gradients obtained from forward finite differences with a *fixed finite difference interval.*

(ii) GD-ada, a variant of DFC with the stepsize being update by the rule in [6, Algorithm 2.2].

(iii) IMFIL, i.e., the implicit filtering algorithm with forward finite differences [24].

(iv) RG, i.e., a random gradient-free algorithm for smooth optimization proposed in [46].

The testing objective functions f are chosen as follows.

1. *Least-square (LS) regression:* $f(x) := \|Ax - b\|^2$, where A is an $n \times n$ matrix and $b \in \mathbb{R}^n$.
2. *A smooth nonconvex (NC) objective:* $f(x) := \sum_{i=1}^n \log(1 + (Ax - b)_i^2)$, where A is an $n \times n$ matrix and b is a vector in $\mathbb{R}^n$. This problem is considered in [52, Section 5.5] and [39, Section 4] with a nonsmooth term added to the objective function.

Random datasets are generated with different sizes for the testing purpose. To be more specific, an $n \times n$ matrix A and a vector $b \in \mathbb{R}^n$ are generated randomly with i.i.d. (independent and identically distributed) standard Gaussian entries. The dimension n is chosen from the set $\{10i, i = 1, \ldots, 20\}$. We inject a uniformly distributed random noise with level $\xi_f \geq 0$, i.e., $\xi(x) \sim U(-\xi_f, \xi_f)$ to the function f and assume only the access to $\phi(x) := f(x) + \xi(x)$ for all the objective functions. The noise level is chosen from the set $\xi_f \in \{10^i,\ i = -9,\ \ldots, -4\}$. The initial points are chosen as the zero vector for all the tests and algorithms. We also assume that the noise level is unknown in these numerical experiments. For that reason, the settings for DFC and GDC (fixed) are chosen as follows:

- DFC, GD-ada: The initial finite difference interval is $\delta_1 = 10^{-2}$. Other parameters are chosen as: $\mu = 2.5,\ r = 2,\ \kappa = \sqrt{n}/2,\ \theta = 0.5$.
- GDC (fixed): The finite difference interval is chosen as $\delta = 10^{-8}$ for the noiseless case and $\delta = 2\sqrt{\xi_f}$ for the noisy case, which is of the same order as the optimal finite difference interval. Note that the latter selection is for testing purposes only since GDC (fixed) does not perform well with $\delta = 10^{-8}$ in the presence of noise. Of course, when the noise level is unknown, choosing a good finite difference interval for GDC (fixed) is not an easy task. To ensure a fair comparison, the stepsize of GDC (fixed) is chosen by a grid search on the set $\left\{\frac{1}{n}, \frac{0.2}{n}, \frac{0.1}{n}\right\}$, where n is the dimension of the problem.

The setting of IMFIL is similar to the one given in Appendix A. The setting of RG is also similar to that in Appendix A, except that the approximate Lipschitz constant is chosen by a grid search on $\{n, 5n, 10n\}$, where n is the dimension of the problem, to ensure a fair comparison. All the methods are executed until they reach the maximum number of function evaluations of $200n$.

In order to illustrate the performance of the algorithms, we use the performance profiles [18] with the measure $f_p^s - f_p^*$, where f_p^s is the function value obtained by method s for problem p, and where f_p^* is the optimal value of the problem p. To be more specific, we assume that the set of problem tests is P. For each method s, we plot the graph of the function

$$\rho_s(\tau) := \frac{1}{|P|}\left|\left\{p \in P \;\middle|\; \frac{f_p^s - f_p^*}{f_p^{\text{best}} - f_p^*} \leq \tau\right\}\right| \quad \text{for} \quad \tau \geq 1,$$

where $|P|$ is the size of P, and where f_p^{best} is the smallest function value obtained by all the methods in problem p. For example, $\rho_s(1)$ represents the percentage of problems where the method s *performs the best*. Due to the structure of the problems, f_p^* is always chosen to be 0. The results for different levels are presented in Figure 3. It can be seen that DFC performs the best in most tests. The robustness of DFC is also good for most selections of performance ratios and is increasing when the noise level is increasing.

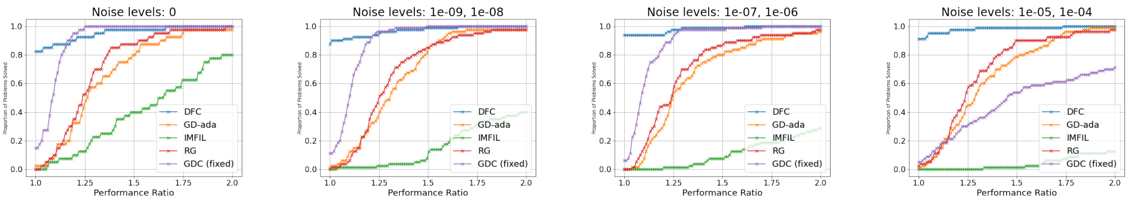


Figure 3: Performance profiles of finite-difference-based methods in minimizing $\mathcal{C}_L^{1,1}$ functions

6.1.2 Experiments with $\mathcal{C}^{1,1}$ Functions

In this subsection, we illustrate the performance of DFB method, i.e., Algorithm 3 with forward finite differences on a subset of CUTEst problems [21, 25] with the details given in Table 2. We also inject uniformly distributed stochastic noise as before, with the noise level ξ_f is either 0, or is chosen from the set $\{10^i, i = -9, \dots, -4\}$ while being unknown to the tested algorithms. In addition to DFB, the methods considered in this numerical experiment are IMFIL, RG with the same setting as in Subsection 6.1.1, and GDB (fixed), i.e., the standard gradient descent method with *backtracking stepsize*, where the approximate gradient is obtained from the forward finite difference with a *fixed finite difference interval*.

Problem	n	**Problem**	n	**Problem**	n	**Problem**	n
ALLINITU	4	DIXMAANB	90	HIMMELBG	2	SPARSINE	100
ARWHEAD	100	DQRTIC	10	HIMMELBH	2	TOINTGSS	50
BARD	3	ENGVAL1	50	HUMPS	2	TOINTGSS	100
BDQRTIC	100	ENGVAL1	100	LOGHAIRY	2	TQUARTIC	100
BOX3	3	FLETBV3M	10	NCB20B	100	TRIDIA	100
BOXPOWER	100	FLETBV3M	100	NONDIA	100	VARDIM	10
BRKMCC	2	FLETCBV2	10	NONDQUAR	100	VAREIGVL	50
BROWNAL	100	FLETCBV3	10	PENALTY3	50	VAREIGVL	100
COSINE	10	FLETCBV3	100	POWELLSG	4	WOODS	100
CRAGGLVY	4	FLETCHCR	100	ROSENBRTU	2	ZANGWIL2	2
CURLY30	100	GULF	3	SENSORS	3		
DIXMAANB	15	HIMMELBCLS	2	SISSER	2		

Table 2: A set of unconstrained problems from CUTEst

The settings for DFB and GDB (fixed) are chosen as follows:

- DFB: The initial finite difference interval $\delta_1 = 10^{-2}$. Other parameters are chosen as: $\theta = 0.5,\ \mu = 2.1,\ \eta = 2, \beta = 0.1,\ \gamma = 0.5,\ C_1 = \frac{\sqrt{n}}{2},\ t_1^{\min} = 10^{-6},\ \bar{\tau} = 1,\ \nu_k = 1/k$.
- GDB (fixed): The finite difference interval is chosen as $\delta = 10^{-8}$ for the noiseless case and $\delta = 2\sqrt{\xi_f}$ for the noisy case, similarly to the selection in GDC (fixed) in previous numerical experiments. The linesearch reduction factor is 0.5, the linesearch constant is 0.1, and the lower bound of the linesearch stepsize is 10^{-10}.

All the methods are executed until they reach the maximum number of function evaluations of 200^*n. Similarly to the previous experiments, the results here are illustrated by the performance profiles with the same measure as in Subsection 6.1.1. Since the exact optimal value is unknown, we approximate it by running DFB and Powell algorithm from SciPy library [59] with the maximum number of function evaluations of 400^*n on the noiseless function. The performance profiles with different levels are presented in Figure 4 showing that DFB achieves the best performance for most of the performance ratios.

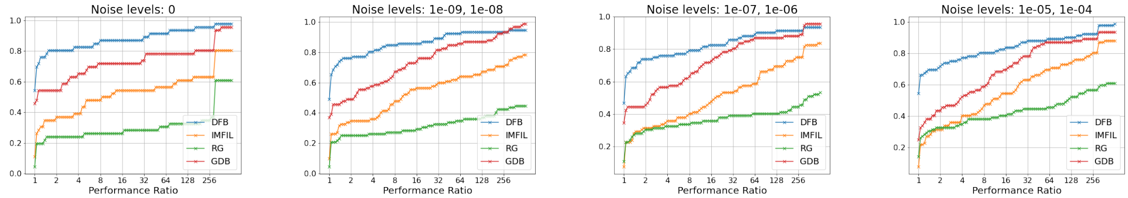


Figure 4: Performance profiles of finite-difference-based methods on minimizing $\mathcal{C}^{1,1}$ functions

6.2 SciPy Production-Ready Algorithms for Functions with Large Noise

This subsection contains some illustrations of the performance of DFBD (Algorithm 4) on the same subset of CUTEst problems [25] with the details given in Table 2. To demonstrate the efficiency of DFBD in handling large noise, we inject the uniformly distributed stochastic noise into the tested problems as in the previous experiments with the high levels of noise $\xi_f \in \{1, 10^{-1}, 10^{-2}, 10^{-3}\}$. In this experiment, the performance of DFBD is compared with *efficient production-ready codes* from the well-known SciPy library [59] of Python; namely, L-BFGS-B, Powell, and COBYLA algorithms. To the best of our knowledge, these methods are among the most popular, efficient, and state-of-the-art derivative-free methods for smooth functions. Although the Nelder-Mead method is also presented in the SciPy library, we do not consider it here due to its poor performance on smooth functions, since it does not take smooth structures into account in the algorithmic design.

All the algorithms are executed until they reach the maximum number of function evaluations of 200^*n. The setting of DFBD is similar to the one given in Appendix A, while the settings of L-BFGS-B, Powell, and COBYLA algorithms are chosen to be standard without any modifications.

The illustration of the results is similar to the one mentioned in Subsection 6.1.2 and is presented in Figure 5. While we found that the Powell and COBYLA algorithms usually work well for the smallest noise $\xi_f = 10^{-3}$, our DFBD method exhibits better results when the noise is larger, i.e., $\xi_f \geq 10^{-2}$. For this reason, we illustrate in Figure 6 below the results for few representative problems in 100-dimensional spaces with the noise levels 1 and 10^{-3}. Since the L-BFGS-B method does not achieve a comparable performance with other methods due to the large noise, we do not plot the results obtained by L-BFGS-B.

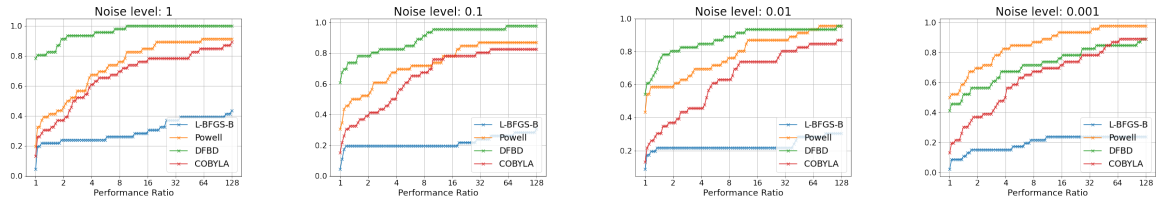
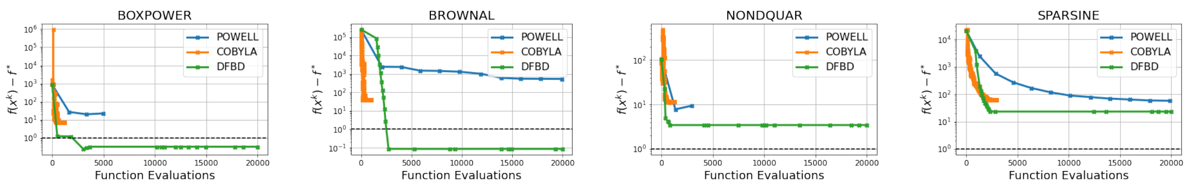
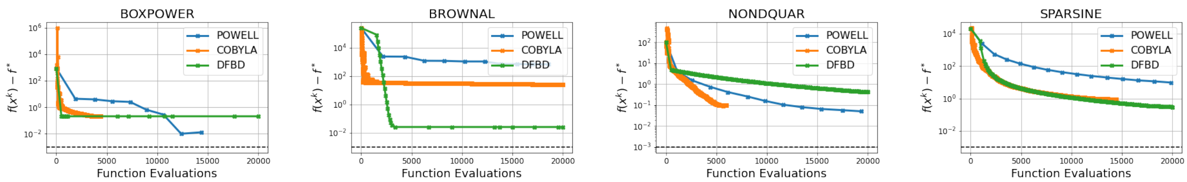


Figure 5: Performance profiles of derivative-free methods on $\mathcal{C}^{1,1}$ functions with large noise



(a) $\xi_f = 1$



(b) $\xi_f = 10^{-3}$

Figure 6: Comparison of forward-difference DFBD with Powell and COBYLA algorithms from SciPy library. The exact function values against the function evaluations are presented. The dashed black line shows the noise level ξ_f of the function.

7 Concluding Remarks

This paper addresses derivative-free optimization problems with smooth and not necessarily convex objectives. A general derivative-free optimization method with a constant stepsize (DFC) is proposed to deal with $\mathcal{C}_L^{1,1}$ problems. This novel method is shown to achieve the fundamental convergence properties of standard gradient descent in the noiseless case and reach a near-stationary point in the noisy case without demanding any noise level information. Constructive estimates of the number of required iterations and function evaluations are established in the paper .

To deal with $\mathcal{C}^{1,1}$ problems, a general derivative-free optimization method with backtracking stepsize (DFB) is proposed. The analysis of DFB in the noiseless case recovers convergence properties of the standard gradient descent method with a backtracking stepsize. To handle $\mathcal{C}^{1,1}$ problems with large noise, a derivative-free optimization method with bidirectional linesearch (DFBD) is proposed. It is revealed that DFBD offers greater robustness than other finite-difference-based schemes to solve $\mathcal{C}^{1,1}$ problems with complex structure. The conducted analysis shows that under certain conditions, DFBD reaches a near-stationary point after a finite number of iterations.

Numerical results demonstrate that DFC and DFB achieve higher efficiency and robustness in comparison with other well-known finite-difference-based schemes in solving noiseless problems and problems with small noise. Moreover, DFBD provide favorable results compared to some production-ready codes from SciPy library when the noise is large.

Our future research includes convergence analysis of the newly developed algorithms coupled with quasi-Newton methods for noisy smooth functions or accelerations. We also intend to establish efficient conditions to ensure local and global convergence to local minimizers of iterative sequences generated by derivative-free methods for problems of nonsmooth unconstrained/constrained optimization.

Declarations

Conflict of interest The authors declare that they have no conflict of interest.

References

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

[2] A. Addis, A. Cassioli, M. Locatelli and F. Schoen, A global optimization method for the design of space trajectories, Comput. Optim. Appl. 48 (2011), 635–652.

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

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

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

[6] A. S. Berahas, R. H. Byrd and J. Nocedal, Derivative-free optimization of noisy functions via quasi-Newton methods, SIAM J. Optim. 29 (2019), 965–993.

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

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

[9] D. P. Bertsekas, Nonlinear Programming, 3rd edition, Athena Scientific, Belmont, MA, 2016.

[10] J. Bolte and E. Pauwels, Conservative set-valued fields, automatic differentiation, stochastic gradient methods and deep learning, Math. Program. 188 (2021), 19–51.

[11] S. Bellavia, G. Gurioli, B. Morini and P. L. Toint, Adaptive regularization algorithms with inexact evaluations for nonconvex optimization, SIAM J. Optim. 29 (2019), 2881–2915.

[12] C. Cartis, N. I. M. Gould and P. Toint, On the oracle complexity of first-order and derivative-free algorithms for smooth nonconvex minimization, SIAM J. Optim. (2012), 66–86.

[13] C. Cartis and K. Scheinberg, Global convergence rate analysis of unconstrained optimization methods based on probabilistic models, Math. Program. 169 (2018), 337–375.

[14] C. Cartis, N. I. M. Gould and P. L. Toint, Evaluation Complexity of Algorithms for Nonconvex Optimization: Theory, Computation and Perspectives, SIAM, Philadelphia, PA, 2022

[15] T. D. Choi and C. T. Kelley, Superlinear convergence and implicit filtering, SIAM J. Optim. 10 (2000), 1149–1162.

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

[17] A. R. Conn, K. Scheinberg and L. N. Vicente, Global convergence of general derivative-free trust-region algorithms to first- and second-order critical points, SIAM J. Optim. 20 (2009), 387–415.

[18] E. Dolan and J. Moré, Benchmarking optimization software with performance profiles, Math. Program. 91 (2022), 201–213.

[19] J. C. Duchi, M. I. Jordan, M. J. Wainwright and A. Wibisono, Optimal rates for zero-order convex optimization: The power of two function evaluations, IEEE Trans. Inf. Theory 61 (2015), 2788–2806.

[20] F. Facchinei and J.-S. Pang, Finite-Dimensional Variational Inequalities and Complementarity Problems, Vol. II, Springer, New York, 2003.

[21] J. Fowkes, L. Roberts and A. Burmen, PyCUTEst: An open source Python package of optimization test problems, J. Open Source Softw. 7 (2022), 4377.

[22] S. Fridovich-Keil and B. Recht, Choosing the stepsize: Intuitive linesearch algorithms with efficient convergence, The 11th Workshop on Optimization for Machine Learning (2019).

[23] P. E. Gill, W. Murray, M. A. Saunders and M. H. Wright, Computing forward-difference intervals for numerical optimization, SIAM J. Sci. Comput. 4 (1983), 310–321.

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

[25] N. I. Gould, D. Orban and P. L. Toint, CUTEst: A constrained and unconstrained testing environment with safe threads for mathematical optimization, Comput. Optim. Appl. 60 (2015), 545–557.

[26] A. Gasnikov, D. Dvinskikh, P. Dvurechensky, E. Gorbunov, A. Beznosikov and A. Lobanov, Randomized gradient-free methods in convex optimization, arXiv preprint arXiv:2211.13566, 2022.

[27] E. Gorbunov, P. Dvurechensky and A. Gasnikov, An accelerated method for derivative-free smooth stochastic convex optimization, SIAM J. Optim. 32 (2022), 1210–1238.

[28] G. N. Grapiglia, Worst-case evaluation complexity of a derivative-free quadratic regularization method, Optim. Lett. 18 (2024), 195–213.

[29] G. A. Gray and T. G. Kolda, Algorithm 856: Appspack 4.0: Asynchronous parallel pattern search for derivative-free optimization, ACM Trans. Math. Softw. 32 (2006), 485–507.

[30] W. Hare and Y. Lucet, Derivative-free optimization via proximal point methods, J. Optim. Theory Appl. 160 (2014), 204–220.

[31] W. Hare and J. Nutini, A derivative-free approximate gradient sampling algorithm for finite minimax problems, Comput. Optim. Appl. 56 (2013), 1–38.

[32] R. Hooke and T. A. Jeeves, Direct search solution of numerical and statistical problems, J. ACM. 8 (1961), 212–229.

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

[34] C. Josz, Global convergence of the gradient method for functions definable in o-minimal structures, Math. Program. 202 (2023), 355–383.

[35] C. Josz, L. Lai and X. Li, Convergence of the momentum method for semi-algebraic functions with locally Lipschitz gradients, SIAM J. Optim. 33 (2023), 2988–3011.

[36] C. T. Kelley, Iterative Methods for Optimization, SIAM, Philadelphia, PA, 1999.

[37] P. D. Khanh, B. S. Mordukhovich and D. B. Tran, Inexact reduced gradient methods in smooth nonconvex optimization, J. Optim. Theory Appl. (2023), https://doi.org/10.1007/s10957-023-02319-9.

[38] P. D. Khanh, B. S. Mordukhovich and D. B. Tran, A new inexact gradient descent method with applications to nonsmooth convex optimization, Optim. Methods Softw. (2024), https://doi.org/10.1080/10556788.2024.2322700.

[39] P. D. Khanh, B. S. Mordukhovich, V. T. Phat and D. B. Tran, Inexact proximal methods for weakly convex functions, arXiv:2307.15596 (2023).

[40] J. D. Lee, I. Panageas, G. Piliouras, M. Simchowitz, M. I. Jordan and B. Recht, First-order methods almost always avoid strict saddle points, Math. Program. 176 (2019), 311–337.

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

[42] J. C. Lagarias, J. A. Reeds, M. H. Wright and P. E. Wright, Convergence properties of the Nelder–Mead simplex method in low dimensions, SIAM J. Optim. 9 (1998), 112–147.

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

[44] J. J. Moré and S. M. Wild, Estimating derivatives of noisy simulations, ACM Trans. Math. Softw. 38 (2012), 1–21.

[45] J. A. Nelder and R. Mead, A simplex method for function minimization, Comput. J. 7 (1965), 308–313.

[46] Y. Nesterov and V. Spokoiny, Random gradient-free minimization of convex functions, Found. Comput. Math. 17 (2017), 527–566.

[47] J. Nocedal and S. J. Wright, Numerical Optimization, 2nd edition. New York, 2006.

[48] A. Ostrowski, Solution of Equations and Systems of Equations, 2nd edition, Academic Press, New York, 1966.

[49] B. T Polyak, Gradient methods for the minimization of functionals, USSR Comput. Math. Math. Phys. 3 (1963), 864–878.

[50] M. J. D. Powell, An efficient method for finding the minimum of a function of several variables without calculating derivatives, Comput. J. 7 (1964), 155–162.

[51] M. J. D. Powell, A direct search optimization method that models the objective and constraint functions by linear interpolation, in Advances in Optimization and Numerical Analysis, pp. 51–67, Springer, New York, 1994.

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

[53] T. T. Truong and H.-T. Nguyen, Backtracking gradient descent method and some applications in large scale optimisation. Part 2: Algorithms and experiments, Appl. Math. Optim. 84 (2021), 2557–2586.

[54] K. Scheinberg, Finite difference gradient approximation: To randomize or not?, INFORMS J. Comput. 34 (2022), 2384–2388.

[55] I. Shibaev, P. Dvurechensky and A. Gasnikov, Zeroth-order methods for noisy Hölder-gradient functions, Optim. Lett. 16 (2022), 2123–2143.

[56] H. M. Shi, Y. Xie, M. Q. Xuan and J. Nocedal, Adaptive finite-difference interval estimation for noisy derivative-free optimization, SIAM J. Sci. Comput. 44 (2022), 2302–2321.

[57] H. M. Shi, Y. Xie, R. Byrd and J. Nocedal, A noise-tolerant quasi-Newton algorithm for unconstrained optimization, SIAM J. Optim. 32 (2022), 29–55.

[58] H. M. Shi, M. Q. Xuan, F. Oztoprak and J. Nocedal, On the numerical performance of finite-difference-based methods for derivative-free optimization, Optim. Methods Softw. 38 (2023), 289–311.

[59] P. Virtanen, R. Gommers and T. E. Oliphant et al. SciPy 1.0: Fundamental algorithms for scientific computing in Python. Nat. Methods 17 (2020), 261–272.

A Numerical Results on Bivariate Functions

In this appendix, we present additional results for the experiment conducted in Remark 5.7. The setup for the experiment is the following:

- DFBD: The parameters are chosen as $\eta = 2$, $L_1 = 1$.
- IMFIL: The setting of IMFIL in this experiment follows the original development at [24, Page 279], with $\bar{\alpha} = 10^{-10}; \beta = 0.1; \gamma = 0.5$ and $h_k = 2^{1-k}$.
- RG: The parameters of RG in this experiment also obey the equations (55) and (58) in the original paper [46], i.e., $h = \frac{1}{4(n+4)L}$ and $\mu = \frac{5}{3(n+4)}\sqrt{\frac{\varepsilon}{2L}}$. Since the function in question does not have a globally Lipschitz continuous gradient, we tune the Lipschitz constant L by grid search on the set $\{0.1, 1, 10\}$ and choose the best one corresponding to the smallest function value at the last iterate.
- DF-backtracking: The parameters are chosen as $\eta = 2$ and $L_1 = 1$ similarly to DFBD.
- GD-BD (Ada): The code for the adaptive finite difference interval estimation is given in [56, Algorithm 2.1]. The parameters for bidirectional linesearch are similar to DFBD.
- L-BFGS (Ada): The L-BFGS code [1] is provided is taken from [57], while the code for the adaptive finite difference interval estimation is provided by [56, Algorithm 2.1].

All the algorithms are executed for 200 function evaluations with the three different initial points $(-4, 0), (-4, -4), (-6, 0)$. We also choose the noise levels $\{1, 0.1, 0.01, 0.001\}$. Since the result with a noise level of 0.01 is already presented in Remark 5.7, we do not represent it here. In addition, while conducting the experiments, due to the randomness of the noisy objective function, there are some cases where the iterative sequence generated by the RG method explodes to extremely large numbers (around 10^{26}) and does not find the minimum region properly. For this reason, we exclusively plot points generated by the methods within a ball centered at the origin with the radius 20. It can be seen that our DFBD is stable with respect to different levels of noise, and fails only in one over nine cases when the noise is 0.1 and the initial point is $(-4, 0)$.

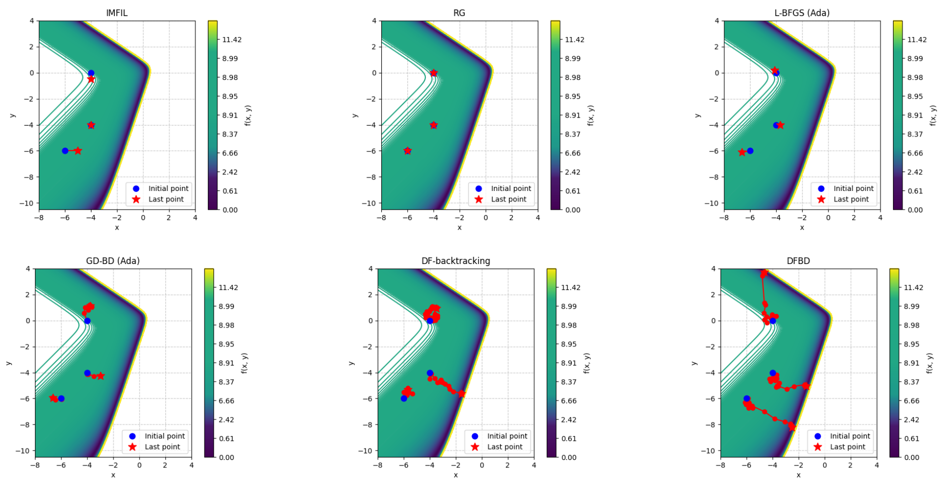


Figure 7: Finite-difference-based methods on minimizing a bivariate $\mathcal{C}^{1,1}$ function (noise level 1)

[1] https://github.com/hjmshi/noise-tolerant-bfgs

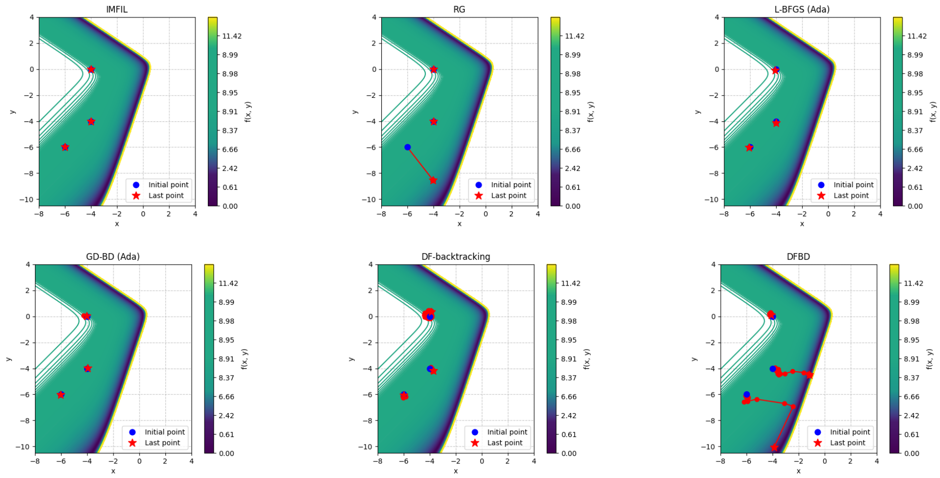


Figure 8: Finite-difference-based methods on minimizing a bivariate $\mathcal{C}^{1,1}$ function (noise level 10^{-1})

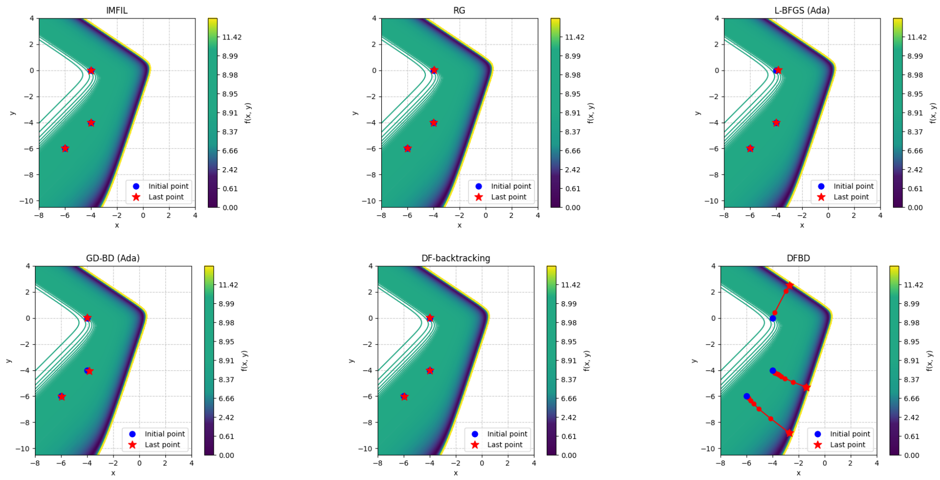


Figure 9: Finite-difference-based methods on minimizing a bivariate $\mathcal{C}^{1,1}$ function (noise level 10^{-3})