Non-convergence Analysis of Probabilistic Direct Search

Cunxin Huang* Zaikun Zhang†

June 4, 2026

Abstract

We present a non-convergence theory for probabilistic direct search, a randomized derivative-free optimization method, where non-convergence means the failure to produce iterates that achieve stationarity asymptotically. The motivation is to understand whether the submartingale-like assumption in the existing convergence theory is essential or merely an artifact of the analysis techniques. For convex objectives, we prove that the probability of non-convergence is positive, provided that the polling directions satisfy a probabilistic ascent condition that is essentially the opposite of the submartingale-like convergence condition. Furthermore, we establish a lower bound for the non-convergence probability. For the typical implementation of this method, where each iteration draws a fixed number of random polling directions independently and uniformly from the unit sphere, our theory implies that the method is not globally convergent if the number of directions is below the threshold specified in the convergence theory, and the submartingale-like assumption is confirmed to be essential for convergence. Our theory is obtained by examining two random series that control the distance from any iterate to the starting point and estimating the probability for these series to stay below certain bounds.

Keywords: Derivative-free optimization, Direct search, Submartingale-like assumption, Non-convergence analysis, Randomized methods

1 Introduction

When does your algorithm fail to converge? This question is arguably as fundamental as asking when it converges. A theoretical investigation of this issue can potentially deepen our understanding of the algorithm, inform its practical implementation, and, in particular, provide a framework for assessing whether the assumptions in existing convergence theory are genuine necessities or merely technical artifacts. However, this question has received far less attention than its convergence counterpart.

*Department of Applied Mathematics, The Hong Kong Polytechnic University, Hong Kong, China. Email: cun-xin.huang@connect.polyu.hk.

†School of Mathematics, Sun Yat-sen University, Guangzhou, China. Email: zhangzaikun@mail.sysu.edu.cn.

We address this issue for a randomized derivative-free optimization (DFO) algorithm known as probabilistic direct search (PDS) [24, 26]. Our answer is provided in the form of a *non-convergence analysis* of PDS. We will focus on the unconstrained optimization problem

$$\min_{x\in\mathbb{R}^n} f(x), \tag{1.1}$$

where $f : \mathbb{R}^n \to \mathbb{R}$ is the objective function.

Direct search [23, 29] is a class of DFO methods that define iterates based on comparisons of function values sampled following a certain scheme without building models for the objective. We focus on directional direct search methods requiring sufficient decrease [18, Section 7.7]. The deterministic variant of these methods evaluates at least $n + 1$ function values per iteration in the worst case, which is prohibitively expensive even for modestly big n.

To alleviate the computational burden per iteration, Gratton et al. [24] propose PDS, which searches along a set of directions randomly generated at each iteration. In the typical implementation of PDS, each iteration draws m i.i.d. random polling directions uniformly from the unit sphere with no dependence on existing polling directions or iterates. PDS is shown to converge globally if the random directions form a sequence of p_0-probabilistic κ-descent sets [24, Definition 3.1] with $\kappa > 0$ and

$$p_0 \;=\; \frac{\log\theta}{\log(\gamma^{-1}\theta)}, \tag{1.2}$$

where $\theta \in (0, 1)$ and $\gamma \geq 1$ are parameters that the algorithm uses to update the step size. In particular, the above-mentioned typical implementation of PDS converges globally if $\gamma > 1$ and

$$m \;>\; \log_2\left(1 - \frac{\log\theta}{\log\gamma}\right), \tag{1.3}$$

where the right-hand side is independent of problem (1.1), especially its dimension n.

A natural question then arises: what if (1.3) is not satisfied? More fundamentally, is the p_0-probabilistic κ-descent assumption essential for the convergence of PDS or is it only a technical artifact in the existing convergence analysis? Establishing a theory (rather than an example) to answer these questions is the main goal of our non-convergence analysis.

However, the motivation for our investigation is not limited to PDS. The probabilistic descent assumption is a representative of the *submartingale-like assumptions* first proposed in [6] and now widely used in the convergence theory of randomized optimization algorithms, including probabilistic trust region [6, 24, 55], line search [9, 11], cubic regularization [11], and subspace methods [10, 47]. We hope that our work will not only shed light on the necessities of such assumptions, but also provide inspiration and tools for the non-convergence analysis of other randomized methods, thus contributing to a deeper understanding of these methods.

The main discoveries of our non-convergence analysis are as follows. If the polling direction sets of PDS form a sequence of p-probabilistic ascent sets (Definition 3.1) with $p > p_*$, where

$$p_* \;=\; 1 - p_0 \;=\; \frac{\log\gamma}{\log(\theta^{-1}\gamma)}, \tag{1.4}$$

then PDS does not converge globally on convex objectives (Theorem 3.2). For the aforementioned typical implementation of PDS, the non-convergence condition reduces to

$$m \ < \ \log_2 \left(1 - \frac{\log \theta}{\log \gamma}\right),$$

a requirement opposite to the convergence condition (1.3) except for the boundary situation with $m = \log_2(1 - \log\theta/\log\gamma)$, which cannot be covered by the existing convergence analysis [24] or our theory. As a highlight, we not only prove that PDS fails to converge globally under the above-mentioned conditions, but also establish a lower bound on the non-convergence probability (Theorem 3.3), which appears to be sharp in numerical experiments.

We must stress that the word "non-convergence" in our paper does not mean divergence. Here, non-convergence signifies that the iterates do not achieve any kind of stationarity asymptotically, but they may converge to a non-stationary point (see Remark 3.2). Indeed, non-convergence investigations are not uncommon in optimization research. For instance, BFGS [21], ADMM [13], and Adam [57] are all shown to be non-convergent under certain conditions. Particularly, Audet [1] exhibits non-convergence examples of Generalized Pattern Search (GPS), a classic direct search method, thereby confirming that the conditions imposed in its convergence theory are all indispensable. Other papers on non-convergence include [7, 27, 34, 39, 45, 52, 56]. Most of these works, however, focus on constructing *non-convergence examples*, whereas our paper aims to establish a systematic *non-convergence theory*. Examples are valuable, but isolated examples, especially pathological ones, may fail to reflect typical behavior. A theory, by contrast, can often provide a broader picture and reveal deeper mechanisms.

The remaining sections are organized as follows. In Section 2, we provide a concise review of DFO and introduce preliminary concepts about PDS. Section 3 contains the major results of this paper. It establishes the non-convergence theory and shows that the typical implementation of PDS is not globally convergent if the number of polling directions is less than $\log_2(1-\log\theta/\log\gamma)$, with the non-convergence probability quantified. We extend our theory to the nonsmooth case in Section 4 and conclude with some perspectives in Section 5. The appendices contain some lemmas, proofs, and discussions, all of which can be skipped without affecting the understanding of the main theory.

Notations. For an event E, we use $\mathbb{1}(E)$ to denote the random variable such that

$$\mathbb{1}(E) \ = \ \begin{cases} 1, & \text{if } E \text{ happens,} \\ 0, & \text{otherwise.} \end{cases}$$

The abbreviation "a.s." stands for "almost surely" and "i.o." for "infinitely often". The Euclidean norm is denoted by $\|\cdot\|$, and $\mathcal{B}(x, r)$ represents the open Euclidean ball centered at $x \in \mathbb{R}^n$ with radius $r > 0$. For the objective function f of problem (1.1), we denote

$$\begin{aligned} \inf f \ &= \ \inf_{x\in\mathbb{R}^n} f(x), \\ \mathcal{S}(f) \ &= \ \{x \in \mathbb{R}^n : f(x) = \inf f\}. \end{aligned}$$

Note that $\inf f$ may be $-\infty$ and $\mathcal{S}(f)$ may be empty. As in [49, page 113], we define the gap distance between two sets $A, B \subseteq \mathbb{R}^n$ as

$$\operatorname{gap}(A, B) = \inf\{\|a - b\| : a \in A, b \in B\},$$

which is supposed to be ∞ if $A = \emptyset$ or $B = \emptyset$; if A is a singleton $\{a\}$, then we write $\operatorname{gap}(a, B)$ instead of $\operatorname{gap}(\{a\}, B)$. As a convention, we define the summation and product over an empty index set as 0 and 1, respectively. In particular, $\sum_{k=i}^{j} a_k = 0$ and $\prod_{k=i}^{j} a_k = 1$ for any real sequence $\{a_k\}$ whenever $i > j$. We write $\lim_k$ instead of $\lim_{k\to\infty}$ in inline equations for brevity, which applies to lower and upper limits as well.

2 Preliminaries

Within the existing literature, DFO methods are broadly classified into two primary categories: model-based methods and direct search methods [4, 18]. Model-based methods construct local models of the problem based on function values and exploit such models under a trust region [17] or line search [8] framework. A wealth of classical literature on model-based methods can be found in [8, 17, 41–44], to name but a few. Direct search methods do not explicitly build models for the problem, but instead, they define iterates based on comparisons of function values sampled following a certain scheme [23, 29]. There exist multiple types of direct search methods, examples including the Nelder–Mead simplex method [36], GPS [2, 33, 53], the MADS family [3, 5, 31], and BFO [37, 38]. More extensive surveys on DFO can be found in the monographs [4, 18], the review papers [19, 30, 46], and the references therein.

Probabilistic techniques have been introduced to both categories of methods in the past decade [6, 10, 11, 24–26, 47], PDS being a result in the direct search category. In what follows, we first present a basic framework of direct search adopted from [18, Section 7.7] and [54], and then introduce the probabilistic variant of this framework proposed in [24] for PDS, around which our investigation will revolve.

2.1 Direct search based on sufficient decrease

Algorithm 2.1 presents a direct search method for solving problem (1.1). Inequality (2.1) is the sufficient decrease condition, where the forcing function $\rho : (0, \infty) \to (0, \infty)$ is nondecreasing and $\rho(\alpha) = o(\alpha)$ when $\alpha \to 0^+$, a typical choice being $\rho(\alpha) = c\alpha^2/2$ with a positive constant c.

Step 2 of Algorithm 2.1 is known as *polling* [18, Chapter 7], and the directions in $\mathcal{D}_k$ are called the *polling directions*. In practice, a search step may be taken at the beginning of each iteration (see [29, Algorithm 3.2]). As in [24], we omit such an option and focus on polling.

Algorithm 2.1 represents only one of the many types of direct search. Instead of the sufficient decrease condition (2.1), there are frameworks that adopt the simple decrease condition

$$f(x_k) - f(x_k + \alpha_k d_k) \;>\; 0 \tag{2.2}$$

Algorithm 2.1 Deterministic direct search based on sufficient decrease

Select $x_0 \in \mathbb{R}^n$, $\alpha_0 > 0$, $\theta \in (0,1)$, $\gamma \in [1,\infty)$, and a forcing function ρ.

For $k = 0, 1, 2, \ldots$, do the following.

1. Generate a set of directions $\mathcal{D}_k \subseteq \mathbb{R}^n$ deterministically.
2. If there exists a direction $d_k \in \mathcal{D}_k$ such that

$$f(x_k) - f(x_k + \alpha_k d_k) \ > \ \rho(\alpha_k), \tag{2.1}$$

then set $x_{k+1} = x_k + \alpha_k d_k$ and $\alpha_{k+1} = \gamma\alpha_k$; otherwise, set $x_{k+1} = x_k$ and $\alpha_{k+1} = \theta\alpha_k$.

and guarantee global convergence by integrality and rationality requirements on the polling directions and the step sizes [1, 2, 53] (see also [18, Section 7.6]).

To implement Algorithm 2.1, a polling strategy is needed to select the direction d_k if there are multiple candidates satisfying (2.1). Two common strategies exist: complete polling chooses the direction that decreases the function value the most, picking the first in case of a tie; opportunistic polling takes the first direction fulfilling (2.1). We also need to set an order for evaluating $\{f(x_k + \alpha_k d) : d \in \mathcal{D}_k\}$. A strategy suggested in [20, Section 4] decides the order by an oracle that can help us rank the prospective decreases of f along the polling directions, the oracle in [20] being a direction of potential descent. For generality, Algorithm 2.1 deliberately keeps the strategies of polling and ordering unspecified.

The analysis of Algorithm 2.1 depends on the concept of the cosine measure defined below.

Definition 2.1 (Cosine measure)**.** Let $\mathcal{D}$ be a finite and nonempty set of nonzero vectors in $\mathbb{R}^n$. The cosine measure of $\mathcal{D}$ with respect to a nonzero vector v, denoted by $\mathrm{cm}(\mathcal{D}, v)$, is defined as

$$\mathrm{cm}(\mathcal{D}, v) \ = \ \max_{d \in \mathcal{D}} \frac{d^{\mathsf{T}} v}{\|d\|\|v\|}.$$

The cosine measure of $\mathcal{D}$, denoted by $\mathrm{cm}(\mathcal{D})$, is defined as $\mathrm{cm}(\mathcal{D}) = \min_{v \in \mathbb{R}^n \setminus \{0\}} \mathrm{cm}(\mathcal{D}, v)$.

Remark 2.1. *Definition 2.1 does not specify the value of* $\mathrm{cm}(\,\cdot\,, 0)$*. As a convention, we suppose that it is defined to be a constant in* $[-1, 1]$*. For example,* $\mathrm{cm}(\,\cdot\,, 0) = 1$ *in [24]. We choose not to particularize this constant, because its value will not affect our non-convergence analysis. See Remark B.1 for more details.*

If f is smooth and there exists a constant $\kappa > 0$ such that $\mathrm{cm}(\mathcal{D}_k) \geq \kappa$ for each $k \geq 0$, then Algorithm 2.1 converges under some technical assumptions. See [29, Theorem 3.11].

2.2 Probabilistic direct search

Algorithm 2.2 presents the PDS method proposed in [24]. It is the same as Algorithm 2.1 except that the polling directions in Step 1 are random vectors over a probability space $(\Omega, \mathcal{F}, \mathbb{P})$. Consequently, the iterates and the step sizes are also random, although the starting point and the initial step size are still chosen deterministically.

Algorithm 2.2 Probabilistic direct search (PDS)

Identical to Algorithm 2.1 except that the polling directions in Step 1 are generated randomly.

For a clear discussion of Algorithm 2.2, it is necessary to use different notations for random elements and their realizations. Similar to [24], we adopt the notations summarized in Table 1.

Table 1: Notations for random elements and their realizations at iteration k

	Polling direction set	Iterate	Step size	Gradient at the iterate
Random element	$\mathfrak{D}_k$	X_k	A_k	G_k
Realization	$\mathcal{D}_k$	x_k	α_k	g_k

We make the following blanket assumption on the sequence of polling direction sets $\{\mathfrak{D}_k\}$ to simplify our presentation, although our analysis remains valid after slight modifications if the lengths of the polling directions are only uniformly bounded.

Blanket Assumption. *For each $k \geq 0$, the set $\mathfrak{D}_k$ is nonempty and consists of finitely many unit random vectors.*

The study of Algorithm 2.2 heavily relies on the concept of σ-algebras and conditional probability with respect to them [22, Section 4.1]. For each $k \geq 0$, we define

$$\mathcal{F}_k \ = \ \sigma(\mathfrak{D}_0, X_1, \ldots, \mathfrak{D}_k, X_{k+1}), \tag{2.3}$$

which is the σ-algebra generated by $\mathfrak{D}_0, X_1, \ldots, \mathfrak{D}_k, X_{k+1}$. Note that $\mathcal{F}_k$ does not involve X_0, which is deterministic. In addition, we define

$$\mathcal{F}_{-1} \ = \ \{\emptyset, \Omega\}.$$

Roughly speaking, for $k \geq 0$, $\mathcal{F}_k$ captures the information about the polling directions and iterates up to the end of iteration k, when X_{k+1} has been generated but $\mathfrak{D}_{k+1}$ has not. Clearly, $\mathfrak{D}_k$ is $\mathcal{F}_k$-measurable and X_k is $\mathcal{F}_{k-1}$-measurable. In addition, G_k is $\mathcal{F}_{k-1}$-measurable if f is continuously differentiable, and A_k is $\mathcal{F}_{k-1}$-measurable by mathematical induction based on the recurrence

$$A_{k+1} \ = \ \gamma^{\mathbb{1}(X_{k+1} \neq X_k)} \theta^{\mathbb{1}(X_{k+1} = X_k)} A_k, \tag{2.4}$$

which holds because $\{A_{k+1} = \gamma A_k\} = \{X_{k+1} \neq X_k\}$ and $\{A_{k+1} = \theta A_k\} = \{X_{k+1} = X_k\}$. In the language of probability theory, $\{\mathfrak{D}_k\}$ is adapted to $\{\mathcal{F}_k\}$, while $\{X_k\}$, $\{G_k\}$, and $\{A_k\}$ are predictable with respect to $\{\mathcal{F}_k\}$ (see [22, Section 4.1]).

2.3 Global convergence of PDS

The global convergence theory [24] of Algorithm 2.2 is summarized below for later reference.

Definition 2.2 ([24, Definition 3.1])**.** Let $p \in [0,1]$ and $\kappa \in [-1,1]$. Consider Algorithm 2.2 with f being continuously differentiable on $\mathbb{R}^n$. The sequence $\{\mathfrak{D}_k\}$ is said to be a sequence of p-probabilistic κ-descent sets if

$$\mathbb{P}\left(\operatorname{cm}\left(\mathfrak{D}_k, -G_k\right) \geq \kappa \mid \mathcal{F}_{k-1}\right) \geq p \quad \text{for each } k \geq 0. \tag{2.5}$$

Theorem 2.1 ([24, Theorem 3.4])**.** *Consider Algorithm 2.2 with f being continuously differentiable and bounded below on $\mathbb{R}^n$, and ∇f being Lipschitz continuous on $\mathbb{R}^n$. If $\{\mathfrak{D}_k\}$ is a sequence of p_0-probabilistic κ-descent sets with p_0 being defined in (1.2) and κ being a positive constant, then $\mathbb{P}(\liminf_k \|G_k\| = 0) = 1$.*

Remark 2.2. *The probability in (2.5) is a conditional probability with respect to a σ-algebra. It is a random variable and is only defined up to almost sure equivalence (e.g., [51, Remark 1 on page 213] and [28, Theorem 8.12]). Consequently, following the convention in probability theory (e.g., [22, page 179] and [28, page 195]), the inequality in Definition 2.2 should be understood in the almost sure sense. Henceforth, all the equalities and inequalities should be interpreted in this way if they involve conditional probabilities or expectations with respect to a σ-algebra (e.g., condition (3.2) in Definition 3.1), and we will not repeat this point every time.*

Remark 2.3. *The σ-algebra $\mathcal{F}_k$ defined in (2.3) reduces to $\mathcal{F}_k^{\mathfrak{D}} = \sigma(\mathfrak{D}_0, \ldots, \mathfrak{D}_k)$ if we assume that X_ℓ is measurable with respect to $\mathcal{F}_{\ell-1}^{\mathfrak{D}}$ for each $\ell \geq 1$. As will be clarified in Lemma D.1, this assumption is fulfilled by the implementations of Algorithm 2.2 considered in [24], but it may fail if we allow the unspecified polling strategy in the algorithm to involve randomness beyond the polling directions (Example D.1). Hence, we choose not to impose such an assumption. In this sense, Theorem 2.1 is a slight generalization of [24, Theorem 3.4], although the proof is essentially the same.*

Corollary 2.1 is the specialization of Theorem 2.1 for the typical implementation of PDS mentioned in Section 1.

Corollary 2.1 ([24, Corollary B.4])**.** *Consider Algorithm 2.2 with f satisfying the assumptions in Theorem 2.1. Let each $\mathfrak{D}_k$ be a set of m i.i.d. random vectors uniformly distributed on the unit sphere in $\mathbb{R}^n$ with no dependence on existing polling directions or iterates. If $\gamma > 1$ and $m > \log_2(1 - \log\theta/\log\gamma)$, then $\mathbb{P}(\liminf_k \|G_k\| = 0) = 1$.*

The global convergence rate of PDS can also be established under a probabilistic descent assumption [24, Section 4], but it is not the focus of this paper.

3 Probabilistic ascent and non-convergence analysis of PDS

How will Algorithm 2.2 behave if the polling direction sets $\{\mathfrak{D}_k\}$ fail to satisfy the probabilistic descent condition in Theorem 2.1? We now address this question by introducing the concept of probabilistic ascent and developing the non-convergence theory of Algorithm 2.2 based on it.

Our discussion begins with two motivating examples to build intuition. We then define probabilistic ascent and prove a key lemma (Lemma 3.2) that establishes the framework for the subsequent analysis. This framework yields two non-convergence results for Algorithm 2.2: a basic one using Markov's inequality (Theorem 3.1) and a refined characterization via a conditional Chernoff bound (Theorems 3.2 and 3.3). The latter constitutes the main technical contribution of the section and enables us to show that the probabilistic descent assumption in Theorem 2.1 is necessary for the global convergence of the algorithm under some conditions (Theorem 3.4). We then revisit our motivating examples and close by proposing a condition weaker than probabilistic ascent to broaden our theory.

3.1 Motivating examples

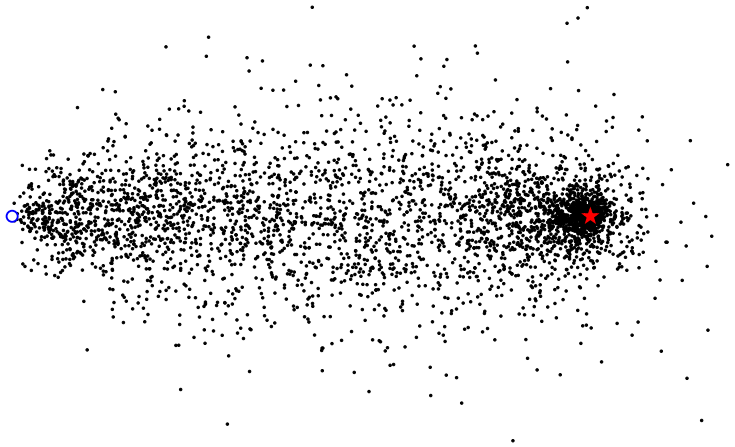
3.1.1 Failure of global convergence: a numerical illustration

Before diving into the analysis, we conduct a simple test to illustrate the behavior of Algorithm 2.2 when the probabilistic descent condition in the convergence theory is not satisfied. We will focus on the typical implementation of the algorithm specified in Corollary 2.1.

As a simple illustration, let the objective function be $f(x) = x^\mathsf{T}x$ with $x \in \mathbb{R}^2$. We set the forcing function $\rho(\alpha) = 10^{-3}\alpha^2$, the starting point $x_0 = (-10, 0)^\mathsf{T}$, the initial step size $\alpha_0 = 1$, the shrinking factor $\theta = 1/4$, and the expanding factor $\gamma = 3/2$. At each iteration, the polling direction set consists of $m = 2$ random vectors that are i.i.d. and uniformly distributed on the unit circle with no dependence on existing polling directions or iterates. Note that $\log_2(1 - \log\theta/\log\gamma) \approx 2.14 > m$, violating the condition in Corollary 2.1 for convergence. The polling strategy is complete polling. The algorithm is terminated when the step size drops below the machine epsilon ($\approx 2 \times 10^{-16}$) or the number of iterations reaches 10^3.

We run the algorithm for 10^4 times independently. The results are shown in Figure 1, where the circle represents the starting point x_0, the pentagram represents the global minimizer $(0, 0)^\mathsf{T}$, and each dot represents the best iterate (i.e., the one with the lowest function value) obtained in a run of the algorithm.

As we can observe, many of these dots are far away from the global minimizer. Even though we cannot draw any rigorous conclusion about the asymptotic behavior of Algorithm 2.2 based on this finite-time observation, it motivates us to conjecture that the algorithm is not globally convergent under this setting. We will confirm this conjecture in the subsequent analysis, and furthermore, estimate the probability that the iterates of the algorithm stay away from the



○ Starting point ★ Global minimizer · Best iterate of a run of the algorithm

Figure 1: A test illustrating failure of convergence of Algorithm 2.2

minimizer (see Corollary 3.1). The estimation will be verified numerically by a refined version of the above experiment in Subsection 3.6.1.

3.1.2 A tiny example reflecting the big picture

As a preview, Example 3.1 presents a particularly simple instance that illustrates our non-convergence theory, highlighting the sharp dichotomy between the convergence and non-convergence conditions. We will provide detailed explanations about this example in Subsection 3.6.2, although the convergence part follows from [24] (see also Corollary 2.1).

Example 3.1. *Let $n = 1$, $f(x) = x^2$, and x_0 be a nonzero number. Consider Algorithm 2.2 with $\mathfrak{D}_k = \{\mathfrak{d}_k\}$, where $\mathfrak{d}_k$ is a random variable independent of $\mathcal{F}_{k-1}$ and takes either 1 or -1, each with probability $1/2$. Then $\mathbb{P}(X_k \to 0) = 1$ if and only if $\theta\gamma \geq 1$. When $\theta\gamma < 1$, there exist constants $C > 0$ and $\bar{\zeta} > 0$ such that*

$$\mathbb{P}\left(\inf_{k\geq 0} |X_k| \;\geq\; (1-\zeta)|x_0|\right) \;\geq\; C\zeta^{-\frac{\log 2}{\log \theta}} \quad \text{for } \zeta \in (0, \bar{\zeta}). \tag{3.1}$$

3.2 Probabilistic ascent

The concept of probabilistic ascent defined below will play a central role in our non-convergence analysis, mirroring the role of probabilistic descent in the convergence theory.

Definition 3.1 (p-probabilistic ascent)**.** Let $p \in [0, 1]$. Consider Algorithm 2.2 with f being continuously differentiable on $\mathbb{R}^n$. The sequence $\{\mathfrak{D}_k\}$ is said to be a sequence of p-probabilistic ascent sets if

$$\mathbb{P}\left(\mathrm{cm}(\mathfrak{D}_k, -G_k) \leq 0 \mid \mathcal{F}_{k-1}\right) \;\geq\; p\mathbb{1}\left(G_k \neq 0\right) \quad \text{for each } k \geq 0. \tag{3.2}$$

The event $\{\mathrm{cm}(\mathfrak{D}_k, -G_k) \leq 0\}$ means that $\mathfrak{D}_k$ does not contain any descent direction. Condition (3.2) requires that the probability of this event given the past is at least p whenever $G_k \neq 0$. One may ask whether we can omit the indicator $\mathbb{1}(G_k \neq 0)$ from condition (3.2). Readers interested in this subtlety can refer to Appendix B.

Proposition 3.1 shows that the polling direction sets used by the typical implementation of Algorithm 2.2 specified in Corollary 2.1 are probabilistic ascent. The proof is given in Appendix C.

Proposition 3.1. *Consider Algorithm 2.2 with f being continuously differentiable on $\mathbb{R}^n$. The sequence $\{\mathfrak{D}_k\}$ specified in Corollary 2.1 is a sequence of p-probabilistic ascent sets with $p = 2^{-m}$.*

Our analysis will heavily depend on the 0-1 process $\{Y_k\}$ with

$$Y_k = \mathbb{1}\Big(\min_{\mathfrak{d}\in\mathfrak{D}_k} \mathfrak{d}^\mathsf{T} G_k < 0\Big) \quad \text{for each } k \geq 0, \tag{3.3}$$

which provides us with an equivalent definition of probabilistic ascent as stated in Lemma 3.1. This equivalence is a simple consequence of Proposition B.1.

Lemma 3.1. *Consider Algorithm 2.2 with f being continuously differentiable on $\mathbb{R}^n$. For any $p \in [0, 1]$, $\{\mathfrak{D}_k\}$ is a sequence of p-probabilistic ascent sets if and only if the sequence $\{Y_k\}$ defined by (3.3) satisfies*

$$\mathbb{P}\left(Y_k = 0 \mid \mathcal{F}_{k-1}\right) \;\geq\; p \quad \textit{for each } k \geq 0. \tag{3.4}$$

Condition (3.4) is foundational to our analysis in Subsections 3.4 and 3.5. The properties of the sequence $\{Y_k\}$ needed in our analysis follow from (3.4) without relying on the specifics of Algorithm 2.2. Such properties include inequality (3.19) and the propositions in Subsection 3.5.1.

3.3 Key ingredients and framework of our analysis

Based on the sequence $\{Y_k\}$ defined in (3.3), we now introduce three additional sequences that will play major roles in our non-convergence study. We then present a lemma and build the framework of our analysis around it.

Let us define

$$\overline{Y}_k \;=\; \frac{1}{k}\sum_{\ell=0}^{k-1} Y_\ell, \quad U_k \;=\; \prod_{\ell=0}^{k-1} \gamma^{Y_\ell}\theta^{1-Y_\ell}, \quad E_k \;=\; \bigcap_{\ell=0}^{k-1}\{Y_\ell = 0\}, \tag{3.5}$$

with the convention that

$$\overline{Y}_0 \;=\; 0, \quad U_0 \;=\; 1, \quad E_0 \;=\; \Omega. \tag{3.6}$$

Note that $\{E_k\}$ is a nonincreasing sequence of events. In addition, since $0 < \theta < \gamma$, we have

$$E_k \;=\; \bigcap_{\ell=0}^{k-1}\{U_\ell = \theta^\ell\}. \tag{3.7}$$

We can check that Y_k is $\mathcal{F}_k$-measurable, and consequently, $\overline{Y}_k$, U_k, and E_k are $\mathcal{F}_{k-1}$-measurable.

Assuming that f is differentiable and convex,[1] Lemma 3.2 links the iterates $\{X_k\}$ with the sequences $\{Y_k\}$ and $\{U_k\}$. As will be detailed in the proof, the convexity of f provides a useful connection between Y_k and iteration k of Algorithm 2.2: if $Y_k = 0$, then the descent condition (2.1) cannot be satisfied, leading to $X_{k+1} = X_k$ and $A_{k+1} = \theta A_k$, which is essentially why the lemma holds.

Lemma 3.2. *Consider Algorithm 2.2 with f being differentiable and convex on $\mathbb{R}^n$. Then*

$$\sup_{k\geq 0} \|X_k - x_0\| \;\leq\; \alpha_0 \sum_{k=0}^{\infty} Y_k U_k \;\leq\; \alpha_0 \sum_{k=0}^{\infty} U_k. \tag{3.8}$$

Proof. For each $k \geq 0$, we note that

$$\|X_{k+1} - X_k\| \;\leq\; Y_k A_k. \tag{3.9}$$

Indeed, if $Y_k = 0$, then $\mathfrak{D}_k$ contains no descent direction, so that the descent condition (2.1) can never be satisfied due to the convexity of f, leading to $X_{k+1} = X_k$ and thus (3.9); when $Y_k = 1$, inequality (3.9) holds because of our blanket assumption that $\mathfrak{D}_k$ contains only unit vectors. Following a similar logic, we have

$$A_{k+1} \;\leq\; \gamma^{Y_k} \theta^{1-Y_k} A_k, \tag{3.10}$$

which is because $A_{k+1} = \theta A_k$ if $Y_k = 0$ and $A_{k+1} \leq \gamma A_k$ otherwise. Recalling $A_0 = \alpha_0$ and the definition of U_k in (3.5), we use (3.10) recursively and obtain

$$A_k \;\leq\; \alpha_0 \prod_{\ell=0}^{k-1} \gamma^{Y_\ell} \theta^{1-Y_\ell} \;=\; \alpha_0 U_k. \tag{3.11}$$

Since $X_0 = x_0$, by (3.9) and (3.11), we have

$$\|X_k - x_0\| \;\leq\; \sum_{\ell=0}^{k-1} \|X_{\ell+1} - X_\ell\| \;\leq\; \sum_{\ell=0}^{k-1} Y_\ell A_\ell \;\leq\; \alpha_0 \sum_{\ell=0}^{k-1} Y_\ell U_\ell \;\leq\; \alpha_0 \sum_{\ell=0}^{k-1} U_\ell, \tag{3.12}$$

where the last inequality is because $Y_\ell \leq 1$. Finally, we get (3.8) by taking the supremum over $k \geq 0$ in (3.12). □

Remark 3.1. *Lemma 3.2 remains valid if Algorithm 2.2 adopts the simple decrease condition (2.2) in place of the sufficient decrease condition (2.1), because the former still cannot be fulfilled if f is convex and $Y_k = 0$. Hence, our non-convergence theory based on Lemma 3.2 will also apply to the variant of Algorithm 2.2 with (2.2) replacing (2.1). This includes Theorems 3.1–3.3, 3.5, 4.1, and their corollaries.*

[1] Note that a real-valued differentiable convex function is continuously differentiable [48, Theorem 25.5].

Framework of our analysis. Roughly, the major step of our analysis is to show that

$$\mathbb{P}\left(\sup_{k\geq 0}\|X_k - x_0\| < \zeta\right) \;>\; 0 \tag{3.13}$$

for some $\zeta > 0$; once this is established, we will have $\{X_k\}$ bounded away from stationarity with positive probability if x_0 is "sufficiently non-stationary", e.g., if $\mathrm{gap}(x_0, \mathcal{S}(f)) \geq \zeta$. According to Lemma 3.2, we can establish (3.13) by proving

$$\mathbb{P}\left(\sum_{k=0}^{\infty} Y_k U_k < \frac{\zeta}{\alpha_0}\right) \;>\; 0, \tag{3.14}$$

or more strongly, $\mathbb{P}\big(\sum_{k=0}^{\infty} U_k < \zeta/\alpha_0\big) > 0$. Our main results Theorems 3.1–3.3 and 3.5 all follow this framework, directly or indirectly.

Remark 3.2. *According to* (3.12)*, inequality* (3.14) *indeed ensures* $\mathbb{P}\big(\sum_{k=0}^{\infty}\|X_{k+1}-X_k\| < \zeta\big) > 0$*, implying that*

$$\mathbb{P}\big(\{X_k\} \text{ converges to a point in } \mathcal{B}(x_0,\zeta)\big) \;>\; 0.$$

Hence, the analysis sketched above will actually guarantee that $\{X_k\}$ *converges to a non-stationary point (rather than diverges) with positive probability.*

Remark 3.3. *For any arbitrarily given set* $\mathcal{T}$*, inequality* (3.14) *can ensure that* $\{X_k\}$ *remains bounded away from* $\mathcal{T}$ *with positive probability as long as* x_0 *is sufficiently distant from* $\mathcal{T}$*, although the primary interest of our analysis is the special case with* $\mathcal{T} = \mathcal{S}(f)$*.*

3.4 Non-convergence analysis via Markov's inequality

We now conduct a non-convergence analysis of Algorithm 2.2 using Markov's inequality. It serves as a quick illustration of the framework presented at the end of Subsection 3.3. Its conclusion will be tightened by a refined analysis in Subsection 3.5 via a conditional Chernoff bound.

Theorem 3.1. *Consider Algorithm 2.2 with* f *being differentiable and convex on* $\mathbb{R}^n$*. If* $\{\mathfrak{D}_k\}$ *is a sequence of* p*-probabilistic ascent sets with* $p > (\gamma-1)/(\gamma-\theta)$*, then we have*

$$\mathbb{P}\left(\mathrm{gap}\left(\{X_k\}, \mathcal{S}(f)\right) > 0\right) \;>\; 0, \tag{3.15}$$

provided that $\mathrm{gap}(x_0, \mathcal{S}(f)) > \alpha_0/[1-\gamma(1-p)-\theta p]$.

Proof. Note that

$$\left\{\mathrm{gap}\left(\{X_k\}, \mathcal{S}(f)\right) > 0\right\} \;\supseteq\; \left\{\sup_{k\geq 0}\|X_k - x_0\| < \mathrm{gap}\left(x_0, \mathcal{S}(f)\right)\right\} \;\supseteq\; \left\{\sum_{k=0}^{\infty} U_k < \frac{\mathrm{gap}\left(x_0, \mathcal{S}(f)\right)}{\alpha_0}\right\},$$

where the last inclusion is due to Lemma 3.2. Therefore, it suffices to show that

$$\mathbb{P}\left(\sum_{k=0}^{\infty} U_k \geq \frac{\mathrm{gap}\left(x_0, \mathcal{S}(f)\right)}{\alpha_0}\right) \;<\; 1. \tag{3.16}$$

Define $\beta = 1/[1-\gamma(1-p)-\theta p]$. Recalling the assumption that $\mathrm{gap}(x_0, \mathcal{S}(f)) > \alpha_0\beta$ and Markov's inequality, we only need to prove that

$$\mathbb{E}\left(\sum_{k=0}^{\infty} U_k\right) \;\leq\; \beta. \tag{3.17}$$

Our assumption on p ensures $0 < \gamma(1-p)+\theta p < 1$, rendering $\beta = \sum_{k=0}^{\infty}[\gamma(1-p)+\theta p]^k$. Meanwhile, Tonelli's theorem [50, page 420] (also [22, Theorem 1.7.2]) yields $\mathbb{E}(\sum_{k=0}^{\infty} U_k) = \sum_{k=0}^{\infty}\mathbb{E}(U_k)$. Thus, the proof of (3.17) can be reduced to establishing

$$\mathbb{E}(U_k) \;\leq\; [\gamma(1-p)+\theta p]^k \quad \text{for each } k \geq 0. \tag{3.18}$$

The proof of (3.18) is standard. For each $k \geq 0$, using the tower property of conditional expectation and the definition of $\{U_k\}$ in (3.5), we have

$$\mathbb{E}(U_{k+1}) \;=\; \mathbb{E}\left(\mathbb{E}\left(\gamma^{Y_k}\theta^{1-Y_k}U_k \mid \mathcal{F}_{k-1}\right)\right) \;=\; \mathbb{E}\left(\mathbb{E}\left(\gamma^{Y_k}\theta^{1-Y_k} \mid \mathcal{F}_{k-1}\right)U_k\right),$$

where the last equality is because U_k is $\mathcal{F}_{k-1}$-measurable. By Lemma 3.1,

$$\mathbb{E}\left(\gamma^{Y_k}\theta^{1-Y_k} \mid \mathcal{F}_{k-1}\right) = \gamma\mathbb{P}(Y_k = 1 \mid \mathcal{F}_{k-1}) + \theta\mathbb{P}(Y_k = 0 \mid \mathcal{F}_{k-1}) \leq \gamma(1-p)+\theta p. \tag{3.19}$$

Hence, we have $\mathbb{E}(U_{k+1}) \leq [\gamma(1-p)+\theta p]\mathbb{E}(U_k)$, which implies (3.18) and concludes our proof. □

Remark 3.4. *Theorem 3.1 and its proof hold trivially if $\mathcal{S}(f) = \emptyset$, as we have $\mathrm{gap}(\,\cdot\,, \mathcal{S}(f)) = \infty$.*

Remark 3.5. *Estimating the probability in (3.16) by Markov's inequality can strengthen (3.15) to*

$$\mathbb{P}\left(\mathrm{gap}\left(\{X_k\}, \mathcal{S}(f)\right) > 0\right) \;\geq\; 1 - \frac{\alpha_0}{[1-\gamma(1-p)-\theta p]\,\mathrm{gap}(x_0, \mathcal{S}(f))}.$$

3.5 Non-convergence analysis via a conditional Chernoff bound

The analysis in the preceding subsection is based on the study of $\sum_{k=0}^{\infty} U_k$ by Markov's inequality, which is rather loose. This subsection conducts a refined analysis via a conditional Chernoff bound for $\{Y_k\}$. We establish the non-convergence of Algorithm 2.2 when $\{\mathfrak{D}_k\}$ is a sequence of p-probabilistic ascent sets with $p > p_*$ and x_0 is any non-stationary point, where p_* is defined by (1.4). Furthermore, we provide a lower bound for the probability that $\{X_k\}$ remains bounded away from stationarity and verify it numerically. This analysis enables us to demonstrate that the probabilistic descent assumption in Theorem 2.1 is necessary for the global convergence of Algorithm 2.2 with the typical implementation except for a boundary case.

3.5.1 Lemmas and observations

We now present several propositions about the 0-1 process $\{Y_k\}$ defined by (3.3) together with the associated sequences $\{\overline{Y}_k\}$, $\{U_k\}$, and $\{E_k\}$ given by (3.5). We emphasize that the propositions are purely consequences of condition (3.4) and independent of the algorithmic details of PDS, making them applicable in any context where (3.4) holds. Indeed, they require only a weaker version of (3.4) that replaces $\mathcal{F}_k$ with $\mathcal{F}_k^Y = \sigma(Y_0, \ldots, Y_k)$, which is a smaller (coarser) σ-algebra.

Because of the clear separation between the algorithmic details and the propositions in this subsection, which is done on purpose, readers who are primarily interested in the algorithmic aspects of our analysis can proceed directly to Subsection 3.5.2 and refer back to this subsection easily when needed.

Lemma 3.3 establishes a conditional Chernoff bound for $\{Y_k\}$, which is essentially a generalization of [24, Lemma 4.5]. Lemma 3.4 shows that condition (3.4) is preserved under conditioning on E_{k_0} with any given integer $k_0 \geq 0$, as long as we shift the indices of $\{Y_k\}$ and $\{\mathcal{F}_k\}$ by k_0. Both lemmas are proved in Appendix C since the arguments are straightforward.

Lemma 3.3. *If $0 < q < p \leq 1$, then condition (3.4) implies that*

$$\mathbb{P}\left(\overline{Y}_k \geq 1 - q \mid E_{k_0}\right) \;\leq\; \exp\left[-\frac{(p-q)^2}{2p}(k + k_0)\right] \quad \text{for all } k \geq 0 \text{ and } k_0 \geq 0. \tag{3.20}$$

Remark 3.6. *Noting the definition of E_k in (3.5), we can derive from condition (3.4) and the tower property of conditional expectations that*

$$\mathbb{P}(E_k) \;\geq\; p^k \quad \text{for each } k \geq 0.$$

Therefore, the conditional probability in Lemma 3.3 is well defined for any $p > 0$.

Lemma 3.4. *Suppose that $p > 0$. Given an integer $k_0 \geq 0$, define $\tilde{Y}_k = Y_{k_0+k}$ and $\tilde{\mathcal{F}}_k = \mathcal{F}_{k_0+k}$ for each k, and denote the probability measure $\mathbb{P}(\,\cdot \mid E_{k_0})$ by $\tilde{\mathbb{P}}(\cdot)$. Then condition (3.4) ensures that*

$$\tilde{\mathbb{P}}(\tilde{Y}_k = 0 \mid \tilde{\mathcal{F}}_{k-1}) \;\geq\; p \quad \text{for each } k \geq 0. \tag{3.21}$$

Proposition 3.2 is a key observation on the series $\sum_{k=k_0}^{\infty} U_k$, where $k_0 \geq 0$ is an integer. It shows that condition (3.4) with $p > p_*$ renders a lower bound for the cumulative distribution function of $\sum_{k=k_0}^{\infty} U_k$ conditioned on E_{k_0}. More importantly, this lower bound is a positive-valued function independent of k_0 after a suitable scaling.

Proposition 3.2. *If $p > p_*$, then condition (3.4) implies that there exists a function Υ satisfying*

$$\mathbb{P}\left(\sum_{k=k_0}^{\infty} U_k < \frac{\theta^{k_0}\zeta}{1-\theta} \,\middle|\, E_{k_0}\right) \;\geq\; \Upsilon(\zeta) \;>\; 0 \tag{3.22}$$

for all $\zeta > 1$ and $k_0 \geq 0$. Here, the function Υ is determined by p, θ, and γ.

Proof. Our proof has two steps. First, identify a function Υ fulfilling (3.22) for $\zeta > 1$ and $k_0 = 0$; second, prove that Υ still works when we relax k_0 to all nonnegative integers.

Step 1. Since $E_0 = \Omega$ as mentioned in (3.6), this step is to determine a positive number $\Upsilon(\zeta)$ for an arbitrarily given $\zeta > 1$ so that

$$\mathbb{P}(F) \;\geq\; \Upsilon(\zeta) \qquad \text{with} \qquad F \;=\; \left\{ \sum_{k=0}^{\infty} U_k < \frac{\zeta}{1-\theta} \right\}. \tag{3.23}$$

To this end, we consider the event E_l defined in (3.5) and note that

$$\mathbb{P}(F) \;\geq\; \mathbb{P}(F \cap E_l) \;=\; \mathbb{P}(F \mid E_l)\, \mathbb{P}(E_l) \tag{3.24}$$

for each $l \geq 0$. In the sequel, we will bound $\mathbb{P}(F \mid E_l)$ and $\mathbb{P}(E_l)$ from below, and select an l in such a way that (3.24) yields a desired lower bound for $\mathbb{P}(F)$.

Due to the definition of F in (3.23) and the fact that $E_l = \bigcap_{k=0}^{l-1} \{U_k = \theta^k\}$ mentioned in (3.7), it holds that

$$\mathbb{P}(F \mid E_l) \;=\; \mathbb{P}\left(\sum_{k=0}^{l-1} \theta^k + \sum_{k=l}^{\infty} U_k < \frac{\zeta}{1-\theta} \;\middle|\; E_l \right) \;\geq\; \mathbb{P}\left(\sum_{k=l}^{\infty} U_k < \frac{\zeta - 1}{1-\theta} \;\middle|\; E_l \right), \tag{3.25}$$

motivating us to bound $\sum_{k=l}^{\infty} U_k$ from above. To do this, we define $q = (p + p_*)/2$ and note that

$$\left\{ \sum_{k=l}^{\infty} U_k < \sum_{k=l}^{\infty} \left(\gamma^{1-q}\theta^q\right)^k \right\} \;\supseteq\; \bigcap_{k=l}^{\infty} \left\{ U_k^{1/k} < \gamma^{1-q}\theta^q \right\} \;=\; \bigcap_{k=l}^{\infty} \left\{ \overline{Y}_k < 1 - q \right\}, \tag{3.26}$$

where the last step is because $U_k^{1/k} = \gamma^{\overline{Y}_k}\theta^{1-\overline{Y}_k}$ by the definitions of $\overline{Y}_k$ and U_k in (3.5). Thus,

$$\begin{aligned} \mathbb{P}\left(\sum_{k=l}^{\infty} U_k < \sum_{k=l}^{\infty} \left(\gamma^{1-q}\theta^q\right)^k \;\middle|\; E_l \right) \;&\geq\; 1 - \mathbb{P}\left(\bigcup_{k=l}^{\infty} \left\{ \overline{Y}_k \geq 1 - q \right\} \;\middle|\; E_l \right) \\ &\geq\; 1 - \sum_{k=l}^{\infty} \exp\left[-\frac{(p-q)^2}{2p} (k + l) \right], \end{aligned} \tag{3.27}$$

which invokes Lemma 3.3 in the last step. Let l be the smallest nonnegative integer satisfying

$$\sum_{k=l}^{\infty} \left(\gamma^{1-q}\theta^q\right)^k \;\leq\; \frac{\zeta - 1}{1-\theta} \qquad \text{and} \qquad \sum_{k=l}^{\infty} \exp\left[-\frac{(p-q)^2}{2p} (k + l) \right] \;\leq\; \frac{1}{2}. \tag{3.28}$$

Such an l exists because $\gamma^{1-q}\theta^q < 1$ and $(p-q)^2/(2p) > 0$ (observe that $\gamma^{1-p_*}\theta^{p_*} = 1$ and recall that $0 \leq p_* < q < p$). The first inequality in (3.28) ensures that the right-hand side of (3.25) is no less than the left-hand side of (3.27), and the second inequality in (3.28) guarantees that the right-hand side of (3.27) is at least $1/2$. Therefore, we can join (3.25) and (3.27) to obtain

$$\mathbb{P}(F \mid E_l) \;\geq\; \frac{1}{2}.$$

Meanwhile, Remark 3.6 renders $\mathbb{P}(E_l) \geq p^l$. Hence, inequality (3.24) validates (3.23) if we set

$$\Upsilon(\zeta) \;=\; \frac{p^l}{2}, \tag{3.29}$$

which is legitimate because l is fully determined by ζ when p, θ, and γ are given. Thus, Υ is a function that completes the first step of the proof.

Step 2. Now, we prove that the function Υ found in the first step satisfies (3.22) for all $\zeta > 1$ and $k_0 \geq 0$. Fix an arbitrary $k_0 \geq 0$. Define $\tilde{\mathbb{P}}$, $\{\tilde{\mathcal{F}}_k\}$, and $\{\tilde{Y}_k\}$ as in Lemma 3.4. According to this lemma, condition (3.4) implies condition (3.21), which has exactly the same form as (3.4), with $\tilde{\mathbb{P}}$, $\{\tilde{Y}_k\}$, and $\{\tilde{\mathcal{F}}_k\}$ corresponding to $\mathbb{P}$, $\{Y_k\}$, and $\{\mathcal{F}_k\}$, respectively. Therefore, repeating the proof for (3.23), we can verify that Υ fulfills

$$\tilde{\mathbb{P}}(\tilde{F}) \;\geq\; \Upsilon(\zeta) \qquad \text{with} \qquad \tilde{F} \;=\; \left\{ \sum_{k=0}^{\infty} \tilde{U}_k < \frac{\zeta}{1-\theta} \right\} \tag{3.30}$$

for all $\zeta > 1$, where $\tilde{U}_k = \prod_{\ell=0}^{k-1} \gamma^{\tilde{Y}_\ell} \theta^{1-\tilde{Y}_\ell}$ for each $k \geq 0$. We will show that (3.30) is indeed (3.22). The definitions of $\{\tilde{Y}_k\}$, $\{U_k\}$, and E_{k_0} (see Lemma 3.4 and (3.5)) imply that

$$\tilde{U}_k \;=\; \prod_{\ell=k_0}^{k_0+k-1} \gamma^{Y_\ell} \theta^{1-Y_\ell} \;=\; U_{k_0}^{-1} U_{k_0+k} \;=\; \theta^{-k_0} U_{k_0+k} \tag{3.31}$$

when E_{k_0} occurs. Plugging $\tilde{\mathbb{P}}(\cdot) = \mathbb{P}(\,\cdot \mid E_{k_0})$ and (3.31) into (3.30), we have

$$\Upsilon(\zeta) \;\leq\; \mathbb{P}(\tilde{F} \mid E_{k_0}) \;=\; \mathbb{P}\left(\sum_{k=0}^{\infty} \theta^{-k_0} U_{k_0+k} < \frac{\zeta}{1-\theta} \,\middle|\, E_{k_0} \right) \;=\; \mathbb{P}\left(\sum_{k=k_0}^{\infty} U_k < \frac{\theta^{k_0}\zeta}{1-\theta} \,\middle|\, E_{k_0} \right),$$

which matches (3.22) as desired. This finishes our proof. □

Remark 3.7. *Given an integer $k_0 \geq 0$, condition (3.4) with $p > p_*$ indeed ensures the equivalence*

$$\mathbb{P}\left(\sum_{k=k_0}^{\infty} U_k < \frac{\theta^{k_0}\zeta}{1-\theta} \,\middle|\, E_{k_0} \right) > 0 \qquad \Longleftrightarrow \qquad \zeta > 1.$$

The implication from right to left is due to Proposition 3.2, while the reverse implication holds because $\sum_{k=k_0}^{\infty} U_k \geq \sum_{k=k_0}^{\infty} \theta^k = \theta^{k_0}/(1-\theta)$.

Proposition 3.2 leads to Proposition 3.3, a crucial observation on the cumulative distribution function of $\sum_{k=0}^{\infty} Y_k U_k$. When $\{Y_k\}$ fulfills condition (3.4) with $p > p_*$, this distribution function turns out to be positive everywhere on $(0, \infty)$, and its tail at 0^+ decays no faster than a power function with exponent $\log p / \log \theta$. This observation will help us establish the non-convergence result in Theorem 3.2 and derive a lower bound for the probability of non-convergence in Theorem 3.3.

Proposition 3.3. *For $\zeta > 0$, define*

$$\Phi(\zeta) \;=\; \mathbb{P}\left(\sum_{k=0}^{\infty} Y_k U_k < \zeta\right). \tag{3.32}$$

If $p > p_*$, then condition (3.4) implies that $\Phi(\zeta) > 0$ for all $\zeta > 0$, and that there exists a constant $C > 0$ such that

$$\Phi(\zeta) \;\geq\; C\,\zeta^{\frac{\log p}{\log \theta}} \quad \text{for } \zeta \in (0,1). \tag{3.33}$$

Proof. It suffices to prove (3.33), which will ensure the positivity of $\Phi(\zeta)$ for all $\zeta > 0$ because Φ is nondecreasing. Given a $\zeta \in (0,1)$, define

$$l \;=\; \left\lceil \frac{\log[\zeta(1-\theta)/2]}{\log \theta} \right\rceil. \tag{3.34}$$

Then $l \geq 0$. Recalling that $E_l = \bigcap_{k=0}^{l-1}\{Y_k = 0\}$ as defined in (3.5), we have

$$\left\{\sum_{k=0}^{\infty} Y_k U_k < \zeta\right\} \;\supseteq\; \left\{\sum_{k=l}^{\infty} Y_k U_k < \zeta\right\} \cap E_l \;\supseteq\; \left\{\sum_{k=l}^{\infty} U_k < \frac{2\theta^l}{1-\theta}\right\} \cap E_l, \tag{3.35}$$

where the last inclusion uses the inequality $Y_k \leq 1$ and the fact that $2\theta^l/(1-\theta) \leq \zeta$ by the definition (3.34) of l. Combining (3.35) with the definition of Φ in (3.32), we obtain

$$\Phi(\zeta) \;\geq\; \mathbb{P}\left(\sum_{k=l}^{\infty} U_k < \frac{2\theta^l}{1-\theta} \;\middle|\; E_l\right) \mathbb{P}(E_l) \;\geq\; \Upsilon(2) p^l,$$

where $\Upsilon(2)$ in the last step comes from Proposition 3.2 and p^l comes from Remark 3.6. Therefore,

$$\log\left[\Phi(\zeta)\zeta^{-\frac{\log p}{\log \theta}}\right] \;\geq\; \log[\Upsilon(2)] + l\log p - \left(\frac{\log p}{\log \theta}\right)\log \zeta \;=\; \log[\Upsilon(2)] + \left(l - \frac{\log \zeta}{\log \theta}\right)\log p.$$

Plugging the definition (3.34) of l into this inequality, we obtain by direct calculation that

$$\log\left[\Phi(\zeta)\zeta^{-\frac{\log p}{\log \theta}}\right] \;\geq\; \log[\Upsilon(2)] + \left(\frac{\log[(1-\theta)/2]}{\log \theta} - 1\right)\log p, \tag{3.36}$$

which implies (3.33), with C being the exponential of the right-hand side in (3.36). □

Note that the exponent in the lower bound in (3.33) is independent of γ. However, the lower bound itself does depend on γ through the constant C, which is increasing in $\Upsilon(2)$ and thus decreasing in γ by the proof of Proposition 3.2 (see (3.28) and (3.29) in particular).

3.5.2 Qualitative and quantitative non-convergence results

This subsection presents our main results on the non-convergence of Algorithm 2.2. We characterize the non-convergence of the algorithm qualitatively in Theorem 3.2 and quantitatively in Theorem 3.3, the latter providing a lower bound for the probability of non-convergence.

It is worth stressing that our results allow us to use a lower semicontinuous function μ to measure the distance of a given point to optimality, examples of such an optimality measure include $f(\cdot) - \inf f$, $\operatorname{gap}(\,\cdot\,, \mathcal{S}(f))$, and $\|\nabla f(\cdot)\|$.

Theorem 3.2 is our qualitative non-convergence result, stating that Algorithm 2.2 stays away from the optimal set with positive probability under a probabilistic ascent assumption, provided that the algorithm is initialized at a non-stationary point.

Theorem 3.2. *Consider Algorithm 2.2 with f being differentiable and convex on $\mathbb{R}^n$. Suppose that $\{\mathfrak{D}_k\}$ is a sequence of p-probabilistic ascent sets with $p > p_*$. Then we have*

$$\mathbb{P}\left(\inf_{k\geq 0} \mu(X_k) > 0\right) \;>\; 0 \tag{3.37}$$

for any function $\mu : \mathbb{R}^n \to (-\infty, \infty]$ that is lower semicontinuous with $\mu(x_0) > 0$. In particular, the conclusion holds if μ is $f(\cdot) - \inf f$, $\operatorname{gap}(\,\cdot\,, \mathcal{S}(f))$, or $\|\nabla f(\cdot)\|$ and x_0 is not stationary.

Proof. Take a positive constant $\varepsilon < \mu(x_0)$. By the lower semicontinuity of μ, there exists a $\delta > 0$ such that $\{x : \mu(x) > \varepsilon\} \supseteq \mathcal{B}(x_0, \delta)$. Hence,

$$\left\{\inf_{k\geq 0} \mu(X_k) > 0\right\} \;\supseteq\; \Big\{\{X_k\} \subseteq \{x : \mu(x) > \varepsilon\}\Big\} \;\supseteq\; \Big\{\{X_k\} \subseteq \mathcal{B}(x_0, \delta)\Big\}. \tag{3.38}$$

Meanwhile, Lemma 3.2 implies that

$$\Big\{\{X_k\} \subseteq \mathcal{B}(x_0, \delta)\Big\} \;\supseteq\; \left\{\sup_{k\geq 0} \|X_k - x_0\| < \delta\right\} \;\supseteq\; \left\{\sum_{k=0}^{\infty} Y_k U_k < \frac{\delta}{\alpha_0}\right\}. \tag{3.39}$$

The last event in (3.39) has a positive probability by Proposition 3.3, because $\{Y_k\}$ satisfies condition (3.4) according to Lemma 3.1. Therefore, (3.38) and (3.39) yield (3.37). □

Remark 3.8. *Theorem 3.2 holds even if the lower semicontinuous function μ has no relation to stationarity, although we are primarily interested in the case where $\mu(x) = 0$ if and only if x is a stationary point. The same remark applies to Theorem 3.3 below. See also Remark 3.3.*

Theorem 3.2 is stronger than Theorem 3.1 in three aspects. First, Theorem 3.2 has a weaker requirement on p since $p_* = (\log\gamma)/\log(\theta^{-1}\gamma) < (\gamma-1)/(\gamma-\theta)$. Second, the optimality measure in Theorem 3.2 can be any lower semicontinuous function μ, whereas Theorem 3.1 applies only to $\operatorname{gap}(\,\cdot\,, \mathcal{S}(f))$. Third, even when $\mu(x) = \operatorname{gap}(x, \mathcal{S}(f))$, the condition $\mu(x_0) > 0$ in Theorem 3.2 is weaker than $\operatorname{gap}(x_0, \mathcal{S}(f)) > \alpha_0/[1 - \gamma(1-p) - \theta p]$ in Theorem 3.1.

Theorem 3.3 is our quantitative non-convergence result, which estimates the probability that the optimality measure in Theorem 3.2 remains close to its initial value. This provides a lower bound for the non-convergence probability of Algorithm 2.2 if its starting point is not stationary.

Theorem 3.3. *Under the settings of Theorem 3.2, if we assume further that μ is locally Lipschitz continuous at x_0, then there exist constants $C > 0$ and $\overline{\zeta} > 0$ such that the function*

$$\Psi(\zeta) \;=\; \mathbb{P}\Big(\inf_{k\geq 0} \mu(X_k) \;\geq\; (1-\zeta)\,\mu(x_0)\Big) \tag{3.40}$$

satisfies

$$\Psi(\zeta) \;\geq\; C\,\zeta^{\frac{\log p}{\log \theta}} \quad \textit{for} \;\; \zeta \in (0, \overline{\zeta}). \tag{3.41}$$

Proof. By our assumption on μ, there exist constants $L > 0$ and $\delta > 0$ such that

$$|\mu(x) - \mu(x_0)| \;\leq\; L\|x - x_0\| \quad \text{for all } x \in \mathcal{B}(x_0, \delta). \tag{3.42}$$

For all $\zeta \in (0,\, L\delta/\mu(x_0))$, combining (3.42) with Lemma 3.2 renders

$$\left\{\inf_{k\geq 0} \mu(X_k) \;\geq\; (1-\zeta)\,\mu(x_0)\right\} \;\supseteq\; \left\{\sup_{k\geq 0} \|X_k - x_0\| < \frac{\zeta\mu(x_0)}{L}\right\} \;\supseteq\; \left\{\sum_{k=0}^{\infty} Y_k U_k < \frac{\zeta\mu(x_0)}{L\alpha_0}\right\}.$$

Consequently, the definition of Φ in (3.32) and that of Ψ in (3.40) yield

$$\Psi(\zeta) \;\geq\; \Phi\!\left(\frac{\zeta\mu(x_0)}{L\alpha_0}\right).$$

Hence, Lemma 3.1 and Proposition 3.3 ensure the existence of C and $\overline{\zeta}$ that validate (3.41). □

Recall Corollary 2.1, which states that Algorithm 2.2 will converge with probability 1 if $m > \log_2(1 - \log\theta/\log\gamma)$. We can now show the non-convergence side.

Corollary 3.1. *Consider Algorithm 2.2 with f being differentiable and convex on $\mathbb{R}^n$. Let $\{\mathfrak{D}_k\}$ be defined as in Corollary 2.1. If $\gamma = 1$ or*

$$m \;<\; \log_2\left(1 - \frac{\log\theta}{\log\gamma}\right), \tag{3.43}$$

then (3.37) holds for any function $\mu : \mathbb{R}^n \to (-\infty, \infty]$ that is lower semicontinuous with $\mu(x_0) > 0$. If we further assume that μ is locally Lipschitz continuous at x_0, then there exist constants $C > 0$ and $\overline{\zeta} > 0$ such that (3.41) holds with $p = 2^{-m}$.

Proof. Proposition 3.1 ensures that $\{\mathfrak{D}_k\}$ is a sequence of p-probabilistic ascent sets with $p = 2^{-m}$. According to Theorems 3.2 and 3.3, it suffices to show that $2^{-m} > p_*$, which is guaranteed by the definition of p_* in (1.4) if $\gamma = 1$ or m satisfies (3.43). □

Remark 3.9. *Comparing Corollaries 2.1 and 3.1, we observe that their requirements on the algorithmic parameters θ, γ, and m are almost the negations of each other. The only gap is the boundary situation with $m = \log_2(1 - \log\theta/\log\gamma)$, which is a concern only if $\log_2(1 - \log\theta/\log\gamma)$ happens to be an integer. Neither the convergence theory in [24] nor our non-convergence theory covers this situation, leaving an interesting topic for future research. The one-dimensional case is however a trivial exception, because $\{\mathfrak{D}_k\}$ is a sequence of p_0-probabilistic 1-descent sets if $m = \log_2(1 - \log\theta/\log\gamma)$, making Algorithm 2.2 converge according to Theorem 2.1.*

For the typical implementation of Algorithm 2.2 specified in Corollary 2.1, we are now able to present the equivalence between the global convergence of the algorithm and the probabilistic descent property of $\{\mathfrak{D}_k\}$, excluding the situation where $\log_2(1 - \log\theta/\log\gamma)$ is an integer.

Theorem 3.4. *Consider Algorithm 2.2 with f being differentiable, convex, and bounded below on $\mathbb{R}^n$, and with ∇f being Lipschitz continuous on $\mathbb{R}^n$. Let $\{\mathfrak{D}_k\}$ be defined as in Corollary 2.1. Suppose that $x_0 \notin \mathcal{S}(f)$, $\gamma > 1$, and $\log_2(1 - \log\theta/\log\gamma)$ is not an integer. Then the following statements are equivalent.*

(a) $\mathbb{P}(\liminf_k \|G_k\| = 0) = 1$.

(b) $m > \log_2(1 - \log\theta/\log\gamma)$.

(c) *$\{\mathfrak{D}_k\}$ is a sequence of p_0-probabilistic κ-descent sets for some $\kappa > 0$.*

Proof. The implication (a) $\Rightarrow$ (b) is guaranteed by Corollary 3.1 and the fact that m cannot equal $\log_2(1 - \log\theta/\log\gamma)$. The implication (b) $\Rightarrow$ (c) is because of [24, Corollary B.4], and (c) $\Rightarrow$ (a) is due to Theorem 2.1. □

For the sequence $\{\mathfrak{D}_k\}$ in Corollary 2.1, if the inequality $\mathbb{P}(\mathrm{cm}(\mathfrak{D}_k, -G_k) \geq \kappa \mid \mathcal{F}_{k-1}) \geq p_0$ in Definition 2.2 holds for one single $k \geq 0$, then it holds for all k (see [24, proof of Corollary B.3]). This fact is essential for the implication (a) $\Rightarrow$ (c) in Theorem 3.4. A general $\{\mathfrak{D}_k\}$, however, may ensure the convergence of Algorithm 2.2 while satisfying the aforementioned inequality only for a set of k (e.g., for k large enough), without being a sequence of p_0-probabilistic κ-descent sets in the sense of Definition 2.2. The convergence analysis of Algorithm 2.2 in such a scenario is beyond the scope of this paper, but Subsection 3.7 will discuss its non-convergence under a similar setting.

3.5.3 Tightness of the ascent probability

Theorem 3.2 requires $\{\mathfrak{D}_k\}$ to be a sequence of p-probabilistic ascent sets with $p > p_*$. Such a requirement cannot be relaxed to $p \geq p_*$, which can be seen from the one-dimensional case mentioned in Remark 3.9. Indeed, if $n = 1$ and $m = \log_2(1 - \log\theta/\log\gamma)$, then $\{\mathfrak{D}_k\}$ is both a sequence of p_*-probabilistic ascent sets and a sequence of p_0-probabilistic 1-descent sets, the latter ensuring the convergence of Algorithm 2.2 by Theorem 2.1. Example 3.2 extends this idea to general dimensions. Note that the example defines $\{\mathfrak{D}_k\}$ using gradient information, even though practical implementations of Algorithm 2.2 are supposed to be derivative-free.

Example 3.2. *Consider Algorithm 2.2 with f being continuously differentiable and bounded below on $\mathbb{R}^n$, and ∇f being Lipschitz continuous on $\mathbb{R}^n$. For each $k \geq 0$, define*

$$\mathfrak{d}_k \;=\; \begin{cases} G_k/\|G_k\|, & \text{if } G_k \neq 0, \\ d, & \text{otherwise}, \end{cases}$$

where d is a fixed unit vector (e.g., a coordinate vector). Then we set $\mathfrak{D}_k = \{\xi_k \mathfrak{d}_k\}$, where ξ_k is a random variable independent of $\mathcal{F}_{k-1}$, taking 1 and -1 with probability p_* and p_0, respectively (recall that $p_* + p_0 = 1$). Note that

$$\mathbb{P}\Big(\min_{\mathfrak{d}\in\mathfrak{D}_k} \mathfrak{d}^{\mathsf{T}} G_k \geq 0 \;\Big|\; \mathcal{F}_{k-1}\Big) \;=\; \mathbb{P}\big(\xi_k \mathfrak{d}_k^{\mathsf{T}} G_k \geq 0 \mid \mathcal{F}_{k-1}\big) \;\geq\; \mathbb{P}(\xi_k = 1 \mid \mathcal{F}_{k-1}) \;=\; p_*.$$

Hence, $\{\mathfrak{D}_k\}$ is a sequence of p_*-probabilistic ascent sets according to Proposition B.1. Similarly, one can check that $\{\mathfrak{D}_k\}$ is a sequence of p_0-probabilistic 1-descent sets according to Definition 2.2. Therefore, Theorem 2.1 ensures that $\mathbb{P}\left(\liminf_k \|G_k\| = 0\right) = 1$.

Complementing Example 3.2, Proposition 3.4 clarifies why the analysis framework described in Subsection 3.3 is incapable of handling the case with $\{\mathfrak{D}_k\}$ being a sequence of p_*-probabilistic ascent sets: in this case, it can happen that $\sum_{k=0}^{\infty} Y_k U_k$ almost surely diverges to infinity, trivializing the bounds in Lemma 3.2. We will prove this proposition in Appendix C.

Proposition 3.4. *If the sequence $\{Y_k\}$ satisfies*

$$\mathbb{P}(Y_k = 0 \mid \mathcal{F}_{k-1}) \;=\; p_* \quad \text{for each } k \geq 0, \tag{3.44}$$

then $\mathbb{P}\left(\sum_{k=0}^{\infty} Y_k U_k = \infty\right) = 1$.

3.6 The motivating examples revisited

3.6.1 Numerical verification of the quantitative non-convergence result

In this subsection, we demonstrate the quantitative non-convergence result in Theorem 3.3 numerically by a refined version of the experiment in Subsection 3.1.1.

As an example, we will focus on the case with $\mu(x) = f(x) - \inf f$, which reduces the function Ψ defined in (3.40) to

$$\Psi(\zeta) \;=\; \mathbb{P}\Big(\inf_{k\geq 0} f(X_k) \;\geq\; (1-\zeta)\, f(x_0) + \zeta \inf f\Big).$$

Theorem 3.3 shows that the tail of Ψ at 0^+ decays at a rate no faster than $\zeta^{\log p/\log\theta}$. Geometrically speaking, if we plot $\Psi(\zeta)$ against ζ on a log-log scale, the slope of the curve at 0^+ should be no more than $\log p/\log\theta$, which will be illustrated numerically by the following experiment.

We set up the objective function and the algorithm in the same way as in Subsection 3.1.1 except for the algorithmic parameters θ, γ, and m. To ensure the representativeness of the results, instead of fixing $(\theta, \gamma, m) = (1/4, 3/2, 2)$ as in Subsection 3.1.1, we randomly sample five values of the triple (θ, γ, m) by the following scheme.

(a) Sample p_* and θ uniformly from the intervals $(0, 9/20)$ and $(1/4, 3/4)$, respectively.

(b) Set $\gamma = \theta^{p_*/(p_*-1)}$ and $m = \lfloor -\log_2 p_* - \mathrm{eps} \rfloor$, where eps is the machine epsilon.

This scheme ensures that inequality (3.43) holds. Hence, the function Ψ satisfies inequality (3.41) in Theorem 3.3 (see Corollary 3.1).

Given a sample of (θ, γ, m), we perform $N = 10^7$ independent runs of Algorithm 2.2. The best (lowest) function value found in each run is denoted by f_{best}. Then we define

$$\hat{\Psi}(\zeta) \;=\; \frac{1}{N} \cdot \big[\,\text{number of runs with } f_{\mathrm{best}} \;\geq\; (1-\zeta)\, f(x_0) + \zeta \inf f\,\big],$$

which will be used as an empirical estimator for $\Psi(\zeta)$. Note that $\inf f = 0$ in our experiment.

Figure 2 plots $\log_{10}[\hat{\Psi}(\zeta)]/(\log p/\log\theta)$ against $\log_{10}\zeta$, with ζ varying in $[10^{-3}, 10^{-1}]$. Each curve in the figure corresponds to a sample of (θ, γ, m). Since we are concerned with the slopes rather than the intercepts, the curves are vertically shifted by small constants to separate them visually. As a reference, the figure includes a black dashed line with slope 1.

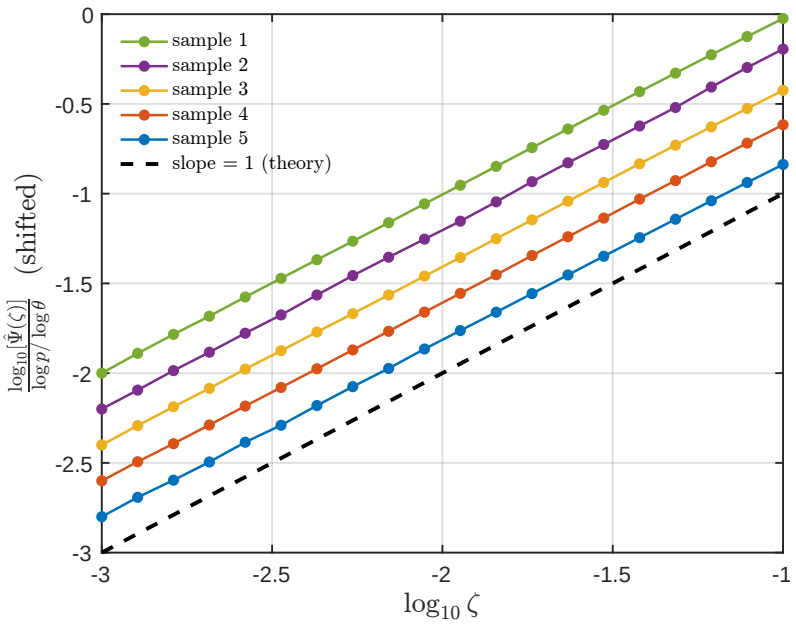


Figure 2: Curves of $\log_{10}[\hat{\Psi}(\zeta)]/(\log p/\log\theta)$ versus $\log_{10}\zeta$ for five random samples of (θ, γ, m). The curves are vertically shifted for clarity. The dashed line is a reference line with slope 1.

Admittedly, we have no guarantee about the quality of the estimator $\hat{\Psi}$ for Ψ, and the interval $[10^{-3}, 10^{-1}]$ is not necessarily a subset of the interval $(0, \overline{\zeta})$ mentioned in Theorem 3.3. Nevertheless, across all the samples, the curves in Figure 2 are almost perfectly parallel to the reference line, which is consistent with the rate in Theorem 3.3. Indeed, the theorem only indicates that the slopes of the curves are no more than that of the reference line. The *surprisingly perfect* parallelism strongly motivates us to conjecture that the rate in the theorem is tight, which is an interesting topic for future research.

3.6.2 Explanations about Example 3.1

Now we explain Example 3.1. Recall that $n = 1$ and note that the $\{\mathfrak{D}_k\}$ in this example is a particular case of the $\{\mathfrak{D}_k\}$ specified in Corollaries 2.1 and 3.1 with $m = 1$.

When $\theta\gamma \geq 1$, we have $\gamma > 1$ and $\log_2(1 - \log\theta/\log\gamma) \leq 1 = m$. According to Corollary 2.1 and Remark 3.9, it holds that $\mathbb{P}(\liminf_k |X_k| = 0) = 1$. Note that $\liminf_k |X_k| = 0$ is equivalent to $X_k \to 0$ in this example due to the monotonicity of Algorithm 2.2.

When $\theta\gamma < 1$, we have $\gamma = 1$ or $\log_2(1 - \log\theta/\log\gamma) > 1 = m$. Therefore, Corollary 3.1 ensures $\mathbb{P}(\inf_{k\geq 0}|X_k| > 0) > 0$ and yields the lower bound in (3.1).

3.7 Non-convergence under a weaker condition

Example 3.2 shows that we cannot weaken our assumption in Theorem 3.2 by replacing $p > p_*$ with $p \geq p_*$. However, it is indeed possible to relax the definition of probabilistic ascent to obtain a weaker condition that renders a weaker non-convergence result compared with Theorem 3.2.

Consider Algorithm 2.2 with f being differentiable and convex on $\mathbb{R}^n$. In place of probabilistic ascent, this subsection assumes that $\{\mathfrak{D}_k\}$ satisfies

$$\mathbb{P}\Big(\liminf_{k\to\infty}\left\{\mathbb{P}\left(\mathrm{cm}\left(\mathfrak{D}_k, -G_k\right) \leq 0 \mid \mathcal{F}_{k-1}\right) \geq p\mathbb{1}(G_k \neq 0)\right\}\Big) \;>\; 0. \tag{3.45}$$

According to Proposition B.1, condition (3.45) holds if and only if the sequence $\{Y_k\}$ defined in (3.3) fulfills

$$\mathbb{P}\Big(\liminf_{k\to\infty}\left\{\mathbb{P}\left(Y_k = 0 \mid \mathcal{F}_{k-1}\right) \geq p\right\}\Big) \;>\; 0. \tag{3.46}$$

Remark 3.10. *Condition* (3.46) *means that the event*

$$\{\mathbb{P}(Y_k = 0 \mid \mathcal{F}_{k-1}) \geq p \text{ for all sufficiently large } k\} \tag{3.47}$$

occurs with positive probability. This is weaker than

$$\mathbb{P}\left(\bigcap_{k=0}^{\infty}\left\{\mathbb{P}\left(Y_k = 0 \mid \mathcal{F}_{k-1}\right) \geq p\right\}\right) \;>\; 0, \tag{3.48}$$

which means that the event $\{\mathbb{P}(Y_k = 0 \mid \mathcal{F}_{k-1}) \geq p \text{ for each } k\}$ *happens with positive probability. Condition* (3.46) *is also weaker than*

$$\sum_{k=0}^{\infty}\mathbb{P}\left(\left\{\mathbb{P}\left(Y_k = 0 \mid \mathcal{F}_{k-1}\right) < p\right\}\right) \;<\; \infty, \tag{3.49}$$

which implies that the event (3.47) *occurs a.s. by the Borel–Cantelli Lemma [22, Theorem 2.3.1].*

As stated in Lemma 3.1, $\{\mathfrak{D}_k\}$ is a sequence of p-probabilistic ascent sets if and only if the sequence $\{Y_k\}$ satisfies $\mathbb{P}\left(Y_k = 0 \mid \mathcal{F}_{k-1}\right) \geq p$ for each $k \geq 0$, a condition stronger than (3.46). Indeed, such a condition ensures both (3.48) and (3.49), either of which in turn implies (3.46) as discussed in Remark 3.10. Therefore, condition (3.45) can be regarded as a relaxation of p-probabilistic ascent defined in Definition 3.1.

Before showing the non-convergence result under condition (3.45), we introduce Lemma 3.6, which will be proved based on Lemma 3.5, a strong law of large numbers for martingales.

Lemma 3.5 ([14])**.** *Let* $\{W_k\}$ *be a martingale. If there exists an* $\alpha \geq 1$ *such that*

$$\sum_{k=1}^{\infty}\mathbb{E}\left(|W_k - W_{k-1}|^{2\alpha}\right)/k^{1+\alpha} < \infty,$$

then we have $W_k/k \to 0$ *a.s. In particular,* $W_k/k \to 0$ *a.s. if* $\{W_k\}$ *has bounded increments.*

Lemma 3.6. *If $p > p_*$, then condition (3.46) implies that*

$$\mathbb{P}\left(\sum_{k=0}^{\infty} U_k < \infty\right) \;>\; 0. \tag{3.50}$$

Proof. By the root test, the series $\sum_{k=0}^{\infty} U_k$ converges if $\limsup_k U_k^{1/k} < 1$. Recalling the definitions of U_k and $\overline{Y}_k$ in (3.5) as well as the fact that $p_* = (\log\gamma)/\log(\theta^{-1}\gamma)$, we have

$$\log\left(U_k^{1/k}\right) \;=\; \log\left(\gamma^{\overline{Y}_k}\theta^{1-\overline{Y}_k}\right) \;=\; \log\theta + \overline{Y}_k\log(\theta^{-1}\gamma) \;=\; [(p_*-1)+\overline{Y}_k]\log(\theta^{-1}\gamma). \tag{3.51}$$

Hence, it holds that

$$\left\{\sum_{k=0}^{\infty} U_k < \infty\right\} \;\supseteq\; \left\{\limsup_{k\to\infty}\log\left(U_k^{1/k}\right) < 0\right\} \;=\; \left\{\limsup_{k\to\infty}\overline{Y}_k < 1-p_*\right\}, \tag{3.52}$$

where the last step uses equality (3.51) and $\log(\theta^{-1}\gamma) > 0$. Therefore, by our assumption that $p > p_*$, inequality (3.50) can be established by proving

$$\mathbb{P}\left(\limsup_{k\to\infty}\overline{Y}_k \le 1-p\right) \;>\; 0. \tag{3.53}$$

To prove (3.53), let us define

$$P_k \;=\; \mathbb{P}\left(Y_k = 0 \mid \mathcal{F}_{k-1}\right) \quad \text{for each } k \ge 0.$$

Then $\mathbb{E}(Y_k + P_k - 1 \mid \mathcal{F}_{k-1}) = 0$ for each k, so $\left\{\sum_{\ell=0}^{k-1}(Y_\ell + P_\ell - 1)\right\}$ is a martingale with respect to $\{\mathcal{F}_{k-1}\}$. In addition, this martingale has bounded increments. Thus, Lemma 3.5 leads to

$$\lim_{k\to\infty}\left(\overline{Y}_k + \overline{P}_k - 1\right) = 0 \quad \text{a.s.},$$

where $\overline{P}_k = k^{-1}\sum_{\ell=0}^{k-1} P_\ell$. Hence,[2] we have $\limsup_k \overline{Y}_k + \liminf_k \overline{P}_k = 1$ a.s., implying

$$\mathbb{P}\left(\limsup_{k\to\infty}\overline{Y}_k \;\le\; 1-p\right) \;=\; \mathbb{P}\left(\liminf_{k\to\infty}\overline{P}_k \;\ge\; p\right) \;\ge\; \mathbb{P}\left(\liminf_{k\to\infty}\{P_k \ge p\}\right). \tag{3.54}$$

The right-hand side of (3.54) is precisely the probability in condition (3.46) and hence is positive. Therefore, inequality (3.53) is justified and the proof is complete. □

Remark 3.11. *In comparison with (3.50), condition (3.4) with $p > p_*$ implies*

$$\mathbb{P}\left(\sum_{k=0}^{\infty} U_k < \infty\right) \;=\; 1. \tag{3.55}$$

The proof is similar to that of Lemma 3.6. The major difference is that (3.4) ensures that the right-hand side of (3.54) equals 1, and hence (3.55) holds according to (3.52). In addition, by Proposition 3.2, condition (3.4) with $p > p_*$ implies that $\mathbb{P}(\sum_{k=0}^{\infty} U_k < s) > 0$ for all $s > 1/(1-\theta)$, whereas (3.50) only ensures that $\mathbb{P}(\sum_{k=0}^{\infty} U_k < s) > 0$ for some sufficiently large s.

[2] Recall that $\limsup_k a_k + \liminf_k b_k = \lim_k(a_k + b_k)$ for bounded real sequences $\{a_k\}$ and $\{b_k\}$ when the limit on the right-hand side exists.

Now, we are ready to present the non-convergence result under the weaker condition (3.45). Its proof is similar to that of Theorem 3.2 with the help of Lemma 3.6.

Theorem 3.5. *Consider Algorithm 2.2 with f being differentiable and convex on $\mathbb{R}^n$. If $\{\mathfrak{D}_k\}$ satisfies (3.45) with $p > p_*$, then there exists a positive constant ζ such that*

$$\mathbb{P}\left(\operatorname{gap}(\{X_k\}, \mathcal{S}(f)) > 0\right) \;>\; 0, \tag{3.56}$$

provided that $\operatorname{gap}(x_0, \mathcal{S}(f)) \geq \zeta$.

Proof. By Lemma 3.6, there exists a positive constant ζ such that

$$\mathbb{P}\left(\sum_{k=0}^{\infty} U_k < \frac{\zeta}{\alpha_0}\right) \;>\; 0. \tag{3.57}$$

Meanwhile, when $\operatorname{gap}(x_0, \mathcal{S}(f)) \geq \zeta$, we have

$$\{\operatorname{gap}(\{X_k\}, \mathcal{S}(f)) > 0\} \;\supseteq\; \left\{\sup_{k\geq 0} \|X_k - x_0\| < \zeta\right\} \;\supseteq\; \left\{\sum_{k=0}^{\infty} U_k < \frac{\zeta}{\alpha_0}\right\},$$

where the last inclusion is due to Lemma 3.2. Therefore, (3.56) holds according to (3.57). □

4 Extension to the nonsmooth case

In this section, we extend our non-convergence results to the nonsmooth case, assuming that $f : \mathbb{R}^n \to \mathbb{R}$ is only convex but not necessarily differentiable. We will show that the non-convergence theorems in Section 3 still hold if we generalize Definition 3.1 of probabilistic ascent to Definition 4.1 below.

We use $f^\circ(\,\cdot\,; d)$ to denote the Clarke generalized directional derivative of f in a direction $d \in \mathbb{R}^n$, and $\partial_{\mathrm{C}} f$ to denote the Clarke subdifferential of f (see [15, Definitions 1.1 and 1.3]). When f is convex, they reduce to the usual (one-sided) directional derivative and the subdifferential, respectively [16, Proposition 2.2.7], but we do not need convexity until Theorem 4.1.

Definition 4.1. Let $p \in [0, 1]$. Consider Algorithm 2.2 with f being locally Lipschitz continuous on $\mathbb{R}^n$. The sequence $\{\mathfrak{D}_k\}$ is said to be a sequence of p-probabilistic ascent sets if it satisfies

$$\mathbb{P}\Big(\min_{\mathfrak{d}\in\mathfrak{D}_k} f^\circ(X_k; \mathfrak{d}) \geq 0 \;\Big|\; \mathcal{F}_{k-1}\Big) \;\geq\; p \quad \text{for each } k \geq 0. \tag{4.1}$$

Remark 4.1. *When f is continuously differentiable on $\mathbb{R}^n$, Definition 4.1 is equivalent to Definition 3.1. See Proposition B.1 and Remark B.2 for details.*

Proposition 4.1 extends Proposition 3.1 to the nonsmooth case. See Appendix C for its proof.

Proposition 4.1. *Consider Algorithm 2.2 with f being locally Lipschitz continuous on $\mathbb{R}^n$. The sequence $\{\mathfrak{D}_k\}$ specified in Proposition 3.1 is a sequence of p-probabilistic ascent sets as defined in Definition 4.1 with $p = 2^{-m}$.*

The 0-1 process $\{Y_k\}$ defined in (3.3) is fundamental to our non-convergence analysis in the smooth case. We now extend the definition of $\{Y_k\}$ from (3.3) to

$$Y_k \;=\; \mathbb{1}\Big(\min_{\mathfrak{d}\in\mathfrak{D}_k} f^\circ(X_k;\mathfrak{d}) < 0\Big) \quad \text{for each } k \geq 0. \tag{4.2}$$

As in the smooth case, Y_k is $\mathcal{F}_k$-measurable for each $k \geq 0$, because the mapping $(x, d) \mapsto f^\circ(x; d)$ is upper semicontinuous [16, Proposition 2.1.1 (b)] and hence Borel measurable. It is clear that $\{\mathfrak{D}_k\}$ is a sequence of p-probabilistic ascent sets if and only if $\mathbb{P}(Y_k = 0 \mid \mathcal{F}_{k-1}) \geq p$ for each $k \geq 0$. In other words, Lemma 3.1 remains valid under the new definition of $\{Y_k\}$.

Now, we extend our non-convergence results in Theorems 3.2 and 3.3 to the nonsmooth case.

Theorem 4.1. *With probabilistic ascent defined in Definition 4.1, Theorems 3.2 and 3.3 still hold if we remove the differentiability assumption about f and replace $\|\nabla f(\cdot)\|$ with $\operatorname{gap}(0, \partial_{\mathrm{C}} f(\cdot))$.*

Proof. We only need to verify that Lemmas 3.1–3.4 and Propositions 3.2–3.3 remain valid in the new setting. Lemma 3.1 holds as mentioned above. For Lemma 3.2, we note that the convexity of f makes it impossible to satisfy the descent condition in Algorithm 2.2 when $Y_k = 0$, leading to inequalities (3.9)–(3.10), which in turn validate Lemma 3.2. Lemmas 3.3–3.4 and Propositions 3.2–3.3 are still true because they only rely on Lemma 3.1 and the $\mathcal{F}_k$-measurability of Y_k. The proof is complete. □

Remark 4.2. *Theorem 4.1 allows the lower semicontinuous function μ in (3.37) to be $f(\cdot) - \inf f$, $\operatorname{gap}(\cdot, \mathcal{S}(f))$, and $\operatorname{gap}(0, \partial_{\mathrm{C}} f(\cdot))$. The lower semicontinuity of $f(\cdot) - \inf f$ and $\operatorname{gap}(\cdot, \mathcal{S}(f))$ are trivial. That of $\operatorname{gap}(0, \partial_{\mathrm{C}} f(\cdot))$ is also basic, but we present it as Lemma A.7 for completeness.*

Theorem 3.1 also holds in the nonsmooth case since it is weaker than Theorem 3.2. As for Theorem 3.5, we can extend it to the nonsmooth case by replacing condition (3.45) with

$$\mathbb{P}\left(\liminf_{k\to\infty} \left\{\mathbb{P}\Big(\min_{\mathfrak{d}\in\mathfrak{D}_k} f^\circ(X_k;\mathfrak{d}) \geq 0 \;\Big|\; \mathcal{F}_{k-1}\Big) \geq p\right\}\right) \;>\; 0, \tag{4.3}$$

which is equivalent to condition (3.46) after we switch the definition of $\{Y_k\}$ to (4.2). Lemma 3.6 still holds, leading to an analogue of Theorem 3.5 with condition (3.45) changed to (4.3).

Hence, we have extended our non-convergence theory of PDS to the nonsmooth case.

5 Conclusions and perspectives

We have established the non-convergence theory of PDS (Algorithm 2.2). For convex objectives, if the polling direction sets satisfy the p-probabilistic ascent condition (Definition 3.1) with

$$p \;>\; p_* \;=\; \frac{\log \gamma}{\log(\theta^{-1}\gamma)},$$

then the iterates of PDS remain bounded away from the solution set with positive probability unless the starting point is a solution (Theorem 3.2). More significantly, we provide a lower

bound on this probability (Theorem 3.3), and our numerical experiments suggest that the bound is sharp.

For the typical implementation of PDS, where each polling set consists of m i.i.d. random directions uniformly distributed on the unit sphere with no dependence on existing polling directions or iterates, the above result implies that the algorithm is not globally convergent for convex objectives if $\gamma = 1$ or $m < \log_2(1 - \log\theta/\log\gamma)$, which is the opposite of the known convergence condition unless $m = \log_2(1 - \log\theta/\log\gamma)$. This confirms that the probabilistic descent condition in the existing convergence theory is essential for the typical implementation (Theorem 3.4).

It is also noteworthy that the condition $p > p_*$ for non-convergence cannot be relaxed to $p \geq p_*$ (Example 3.2), although a weaker non-convergence condition does exist (Theorem 3.5). In addition, our non-convergence theory covers the nonsmooth case by replacing the probabilistic ascent condition with a natural generalization based on the Clarke subdifferential (Section 4).

Two technical problems remain open. The first one is whether the aforementioned typical implementation of PDS converges in the borderline scenario with $m = \log_2(1 - \log\theta/\log\gamma)$. The second is the sharpness of the lower bound on the non-convergence probability in Theorem 3.3.

The most intriguing direction for future research, however, extends far beyond PDS. As mentioned in Section 1, we are interested in developing analogous non-convergence theories for other randomized methods, including probabilistic trust region [6, 24, 55], line search [9, 11], cubic regularization [11], and subspace methods [10, 47]. Similar to the probabilistic descent condition (2.5), the convergence analysis of these methods relies on submartingale-like assumptions in the form of

$$\mathbb{P}\Big(F_k \mid \mathcal{G}_{k-1}\Big) \;\geq\; P_0 \quad \text{for each } k \geq 0,$$

where F_k is an event favorable to the success of the algorithm at iteration k, $\mathcal{G}_{k-1}$ is a σ-algebra representing the history of the algorithm up to iteration $k-1$, and P_0 is a positive constant. How essential are these assumptions for the convergence? Answering this question would deepen our understanding of these methods and of submartingale-like assumptions in general. We are currently working on the corresponding non-convergence theories, with the aim of providing a comprehensive answer to this question. We hope that systematic non-convergence analysis of this kind, in contrast to the construction of particular non-convergence examples, will receive more attention in the study of optimization algorithms.

Defined according to the step size contracting-expanding scheme of PDS, the series

$$\sum_{k=0}^{\infty}\prod_{\ell=0}^{k-1}\gamma^{Y_\ell}\theta^{1-Y_\ell} \quad \text{and} \quad \sum_{k=0}^{\infty}Y_k\prod_{\ell=0}^{k-1}\gamma^{Y_\ell}\theta^{1-Y_\ell}$$

have played a vital role in this paper, which denotes $\prod_{\ell=0}^{k-1}\gamma^{Y_\ell}\theta^{1-Y_\ell}$ by U_k. They control the distance between the iterates and the starting point, and their boundedness is closely related to the non-convergence of PDS. The above-mentioned randomized methods all contain contracting-expanding schemes for certain algorithmic parameters. Hence, we expect that similar series will be

instrumental in their non-convergence analysis, where the tools and results developed here about these series will be a useful guide. This is particularly true of the conditional Chernoff bound and its consequences in Subsection 3.5.1, which are independent of the algorithmic details of PDS. Moreover, these series can indeed provide a new way of establishing the global convergence of a class of randomized methods including PDS, which is another topic we are currently studying but is beyond the scope of this paper. Finally, the stochastic processes defined by these series appear to have interesting probabilistic structures. Further study of them may reveal additional connections between probability theory and numerical optimization. This paper represents only a small step in this direction.

Acknowledgments

This paper is part of the PhD thesis of Cunxin Huang, co-supervised by Zaikun Zhang and Professor Xiaojun Chen from The Hong Kong Polytechnic University. Both authors are grateful to Professor Chen for her help during Huang's PhD study.

Declarations

Funding. This work was supported by the National Key R&D Program of China (2023YFA1009300), the General Program of the National Natural Science Foundation of China (12571335), and the Hong Kong PhD Fellowship Scheme (PF21-56718).
Conflict of interest. The authors declare that they have no conflict of interest.

References

[1] C. Audet. Convergence results for generalized pattern search algorithms are tight. *Optim. Eng.*, 5:101–122, 2004.

[2] C. Audet and J. E. Dennis, Jr. Analysis of generalized pattern searches. *SIAM J. Optim.*, 13:889–903, 2002.

[3] C. Audet and J. E. Dennis, Jr. Mesh adaptive direct search algorithms for constrained optimization. *SIAM J. Optim.*, 17:188–217, 2006.

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

[5] C. Audet, S. Le Digabel, V. R. Montplaisir, and C. Tribes. Algorithm 1027: NOMAD version 4: Nonlinear optimization with the MADS algorithm. *ACM Trans. Math. Software*, 48:1–22, 2022.

[6] A. S. Bandeira, K. Scheinberg, and L. N. Vicente. Convergence of trust-region methods based on probabilistic models. *SIAM J. Optim.*, 24:1238–1264, 2014.

[7] I. Ben Gharbia and J. Gilbert. Nonconvergence of the plain Newton-min algorithm for linear complementarity problems with a P-matrix. *Math. Program.*, 134:349–364, 2012.

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

[9] A. S. Berahas, L. Cao, and K. Scheinberg. Global convergence rate analysis of a generic line search algorithm with noise. *SIAM J. Optim.*, 31:1489–1518, 2021.

[10] C. Cartis and L. Roberts. Scalable subspace methods for derivative-free nonlinear least-squares optimization. *Math. Program.*, 199:461–524, 2023.

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

[12] E. Çinlar. *Probability and Stochastics*. Springer, New York, 2011.

[13] C. Chen, B. He, Y. Ye, and X. Yuan. The direct extension of ADMM for multi-block convex minimization problems is not necessarily convergent. *Math. Program.*, 155:57–79, 2016.

[14] Y. S. Chow. On a strong law of large numbers for martingales. *Ann. Math. Statist.*, 38:610–610, 1967.

[15] F. H. Clarke. Generalized gradients and applications. *Trans. Amer. Math. Soc.*, 205:247–262, 1975.

[16] F. H. Clarke. *Optimization and Nonsmooth Analysis*, volume 5 of *Classics Appl. Math.* SIAM, Philadelphia, 1990.

[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.*, 20:387–415, 2009.

[18] A. R. Conn, K. Scheinberg, and L. N. Vicente. *Introduction to Derivative-Free Optimization*, volume 8 of *MOS-SIAM Ser. Optim.* SIAM, Philadelphia, 2009.

[19] A. L. Custódio, K. Scheinberg, and L. N. Vicente. Methodologies and software for derivative-free optimization. In T. Terlaky, M. F. Anjos, and S. Ahmed, editors, *Advances and Trends in Optimization with Engineering Applications*, pages 495–506. SIAM, Philadelphia, 2017.

[20] A. L. Custódio and L. N. Vicente. Using sampling and simplex derivatives in pattern search methods. *SIAM J. Optim.*, 18:537–555, 2007.

[21] Y. H. Dai. A perfect example for the BFGS method. *Math. Program.*, 138:501–530, 2013.

[22] R. Durrett. *Probability: Theory and Examples*. Camb. Ser. Stat. Probab. Math. Cambridge University Press, Cambridge, fifth edition, 2019.

[23] K. J. Dzahini, F. Rinaldi, C. W. Royer, and D. Zeffiro. Direct-search methods in the year 2025: Theoretical guarantees and algorithmic paradigms. *EURO J. Comput. Optim.*, 13:1–24, 2025.

[24] S. Gratton, C. W. Royer, L. N. Vicente, and Z. Zhang. Direct search based on probabilistic descent. *SIAM J. Optim.*, 25:1515–1541, 2015.

[25] S. Gratton, C. W. Royer, L. N. Vicente, and Z. Zhang. Complexity and global rates of trust-region methods based on probabilistic models. *IMA J. Numer. Anal.*, 38:1579–1597, 2018.

[26] S. Gratton, C. W. Royer, L. N. Vicente, and Z. Zhang. Direct search based on probabilistic feasible descent for bound and linearly constrained problems. *Comput. Optim. Appl.*, 72:525–559, 2019.

[27] M. Hong, S. Zeng, J. Zhang, and H. Sun. On the divergence of decentralized nonconvex optimization. *SIAM J. Optim.*, 32:2879–2908, 2022.

[28] A. Klenke. *Probability Theory: A Comprehensive Course*. Springer Nature Switzerland AG, Gewerbestrasse, third edition, 2020.

[29] T. G. Kolda, R. M. Lewis, and V. Torczon. Optimization by direct search: New perspectives on some classical and modern methods. *SIAM Rev.*, 45:385–482, 2003.

[30] J. Larson, M. Menickelly, and S. M. Wild. Derivative-free optimization methods. *Acta Numer.*, 28:287–404, 2019.

[31] S. Le Digabel. Algorithm 909: NOMAD: Nonlinear optimization with the MADS algorithm. *ACM Trans. Math. Software*, 37:44:1–44:15, 2011.

[32] P. Lévy. *Théorie de l'addition des variables aléatoires*. Gauthier-Villars, Paris, 1937.

[33] R. M. Lewis and V. Torczon. Pattern search algorithms for bound constrained minimization. *SIAM J. Optim.*, 9:1082–1099, 1999.

[34] W. Mascarenhas. The divergence of the BFGS and Gauss Newton methods. *Math. Program.*, 147:253–276, 2014.

[35] W. Mulzer. Five proofs of Chernoff's bound with applications. *Bull. Eur. Assoc. Theor. Comput. Sci.*, 1:1–18, 2018.

[36] J. A. Nelder and R. Mead. A simplex method for function minimization. *Comput. J.*, 7:308–313, 1965.

[37] M. Porcelli and Ph. L. Toint. BFO, a trainable derivative-free brute force optimizer for nonlinear bound-constrained optimization and equilibrium computations with continuous and discrete variables. *ACM Trans. Math. Software*, 44:6:1–6:25, 2017.

[38] M. Porcelli and Ph. L. Toint. Global and local information in structured derivative free optimization with BFO. *arXiv:2001.04801*, 2020.

[39] M. J. D. Powell. On search directions for minimization algorithms. *Math. Program.*, 4:193–201, 1973.

[40] M. J. D. Powell. Convergence properties of a class of minimization algorithms. In O. L. Mangasarian, R. R. Meyer, and S. M. Robinson, editors, *Nonlinear Programming 2: Proceedings of the Special Interest Group on Mathematical Programming Symposium Conducted by the Computer Sciences Department at the University of Wisconsin-Madison, April 15–17, 1974*, pages 1–27. Academic Press, 1975.

[41] M. J. D. Powell. A direct search optimization method that models the objective and constraint functions by linear interpolation. In S. Gomez and J.-P. Hennart, editors, *Advances in Optimization and Numerical Analysis*, pages 51–67. Kluwer Academic, Dordrecht, 1994.

[42] M. J. D. Powell. UOBYQA: unconstrained optimization by quadratic approximation. *Math. Program., Ser. B*, 92:555–582, 2002.

[43] M. J. D. Powell. The NEWUOA software for unconstrained optimization without derivatives. In G. Di Pillo and M. Roma, editors, *Large-scale Nonlinear Optimization*, pages 255–297. Springer, Boston, 2006.

[44] M. J. D. Powell. The BOBYQA algorithm for bound constrained optimization without derivatives. Technical Report DAMTP 2009/NA06, Department of Applied Mathematics and Theoretical Physics, Cambridge University, Cambridge, 2009.

[45] F. Ramponi. Consistency of the scenario approach. *SIAM J. Optim.*, 28:135–162, 2018.

[46] L. M. Rios and N. V. Sahinidis. Derivative-free optimization: a review of algorithms and comparison of software implementations. *J. Global Optim.*, 56:1247–1293, 2013.

[47] L. Roberts and C. W. Royer. Direct search based on probabilistic descent in reduced spaces. *SIAM J. Optim.*, 33:3057–3082, 2023.

[48] R. T. Rockafellar. *Convex Analysis*. Princeton University Press, New Jersey, 1970.

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

[50] H. L. Royden and P. M. Fitzpatrick. *Real Analysis*. Prentice Hall, Upper Saddle River, NJ, fourth edition, 2010.

[51] A. N. Shiryaev. *Probability*. Springer-Verlag, New York, 2nd edition, 1996.

[52] J. Thompson. Examples of non-convergence of conjugate descent algorithms with exact line-searches. *Math. Program.*, 12:356–360, 1977.

[53] V. Torczon. On the convergence of pattern search algorithms. *SIAM J. Optim.*, 7:1–25, 1997.

[54] L. N. Vicente. Worst case complexity of direct search. *EURO J. Comput. Optim.*, 1:143–153, 2013.

[55] X. Wang and Y. Yuan. Stochastic trust region methods with trust region radius depending on probabilistic models. *J. Comput. Math.*, 40:294–334, 2022.

[56] Y. Yuan. An example of non-convergence of trust region algorithms. In Y. Yuan, editor, *Advances in Nonlinear Programming*, pages 205–215. Kluwer Academic Publishers, Dordrecht, 1998.

[57] Y. Zhang, C. Chen, N. Shi, R. Sun, and Z.-Q. Luo. Adam can converge without any modification on update rules. In S. Koyejo, S. Mohamed, A. Agarwal, A. Oh, D. Belgrave, and K. Cho, editors, *Advances in Neural Information Processing Systems 35 (NeurIPS 2022)*, pages 28386–28399. Curran Associates, Inc., 2022.

Appendices

A Basic lemmas

This section collects some basic lemmas needed in our analysis. They are entirely independent of our algorithmic discussion. We begin with Lemma A.1, a direct consequence of [22, Theorem 4.1.14].

Lemma A.1. *If E and F are events, and $\mathcal{G}$ is a σ-algebra with $F \in \mathcal{G}$, then $\mathbb{P}(EF \mid \mathcal{G}) = \mathbb{P}(E \mid \mathcal{G})\mathbb{1}(F)$.*

Lemma A.2 presents a basic connection between the conditional probability with respect to a σ-algebra and that with respect to an event.

Lemma A.2. *Let E be an event and $\mathcal{G}$ be a σ-algebra. Then $\mathbb{P}(E \mid \mathcal{G}) \geq p$ if and only if $\mathbb{P}(E \mid F) \geq p$ for all $F \in \mathcal{G}$ with $\mathbb{P}(F) > 0$.*

Proof. For all $F \in \mathcal{G}$ with $\mathbb{P}(F) > 0$, the law of total probability and Lemma A.1 yield

$$\mathbb{P}(E \mid F) \;=\; \frac{\mathbb{E}\big(\mathbb{P}(EF \mid \mathcal{G})\big)}{\mathbb{P}(F)} \;=\; \frac{\mathbb{E}\big(\mathbb{P}(E \mid \mathcal{G})\mathbb{1}(F)\big)}{\mathbb{P}(F)}. \tag{A.1}$$

If $\mathbb{P}(E \mid \mathcal{G}) \geq p$ a.s., then (A.1) yields $\mathbb{P}(E \mid F) \geq p$. If $\mathbb{P}(E \mid F) \geq p$ for all $F \in \mathcal{G}$ with $\mathbb{P}(F) > 0$, then the event $\hat{F} = \{\mathbb{P}(E \mid \mathcal{G}) < p\} \in \mathcal{G}$ must have probability zero, or else (A.1) implies $\mathbb{P}(E \mid \hat{F}) < p$. □

Lemma A.3 elaborates on the equivalence among several probability inequalities. It is useful for interpreting the conditions in Definitions 3.1 and 4.1.

Lemma A.3. *Let $p \in [0,1]$ be a constant, E and F be events, and $\mathcal{G}$ be a σ-algebra with $E \in \mathcal{G}$. Then the following three inequalities are equivalent to each other:*

$$\mathbb{P}(F \mid \mathcal{G}) \;\geq\; p\mathbb{1}(E^c), \qquad \mathbb{P}(F \cap E^c \mid \mathcal{G}) \;\geq\; p\mathbb{1}(E^c), \qquad \textit{and} \qquad \mathbb{P}(E \cup F \mid \mathcal{G}) \;\geq\; p.$$

In particular, if $E \subseteq F$, then they are all equivalent to $\mathbb{P}(F \mid \mathcal{G}) \geq p$.

Proof. We refer to the three inequalities as (a), (b), and (c), in left-to-right order.
(a) $\Rightarrow$ (b): Since $E^c \in \mathcal{G}$, Lemma A.1 yields $\mathbb{P}(F \cap E^c \mid \mathcal{G}) = \mathbb{P}(F \mid \mathcal{G})\mathbb{1}(E^c) \;\geq\; p\mathbb{1}(E^c)$.
(b) $\Rightarrow$ (c): Since $E \cup F = E \cup (F \cap E^c)$ and $E \in \mathcal{G}$, we have

$$\mathbb{P}(E \cup F \mid \mathcal{G}) \;=\; \mathbb{P}(E \mid \mathcal{G}) + \mathbb{P}(F \cap E^c \mid \mathcal{G}) \;=\; \mathbb{1}(E) + \mathbb{P}(F \cap E^c \mid \mathcal{G}) \;\geq\; \mathbb{1}(E) + p\mathbb{1}(E^c) \;\geq\; p.$$

(c) $\Rightarrow$ (a): Since $(E \cup F) \cap E^c = F \cap E^c$ and $E^c \in \mathcal{G}$, Lemma A.1 leads to

$$\mathbb{P}(F \mid \mathcal{G}) \;\geq\; \mathbb{P}((E \cup F) \cap E^c \mid \mathcal{G}) \;=\; \mathbb{P}(E \cup F \mid \mathcal{G})\mathbb{1}(E^c) \;\geq\; p\mathbb{1}(E^c).$$

(c) reduces to $\mathbb{P}(F \mid \mathcal{G}) \geq p$ when $E \subseteq F$. □

Lemma A.4 and the subsequent Corollary A.1 are helpful in the proofs of Propositions 3.1 and 4.1.

Lemma A.4. *Let X and Y be random vectors. Consider a Borel measurable function h such that $\mathbb{E}(|h(X,Y)|) < \infty$ and define $H(y) = \mathbb{E}(h(X,y))$.*

(a) *If X is independent of Y, then $\mathbb{E}(h(X,Y)) = \mathbb{E}(H(Y))$.*

(b) *If X is independent of Y and a σ-algebra $\mathcal{G}$, then* $\mathbb{E}(h(X,Y) \mid \mathcal{G}) = \mathbb{E}(H(Y) \mid \mathcal{G})$.

Proof. Item (a) is a generalization of [22, Theorem 2.1.12], which considers random variables rather than random vectors. We omit its proof, which is a straightforward extension of that for [22, Theorem 2.1.12].

Now we prove item (b) based on (a). Since $\mathbb{E}(H(Y) \mid \mathcal{G})$ is $\mathcal{G}$-measurable, the definition of conditional expectation tells us that we only need to verify

$$\mathbb{E}(h(X,Y)\mathbb{1}(E)) \;=\; \mathbb{E}(\mathbb{E}(H(Y) \mid \mathcal{G})\mathbb{1}(E)) \tag{A.2}$$

for all $E \in \mathcal{G}$. The right-hand side of (A.2) equals $\mathbb{E}(H(Y)\mathbb{1}(E))$ due to the fact that $E \in \mathcal{G}$ and the tower property of conditional expectation. Hence, we only need to check

$$\mathbb{E}(h(X,Y)\mathbb{1}(E)) \;=\; \mathbb{E}(H(Y)\mathbb{1}(E)). \tag{A.3}$$

Denote $\mathbb{1}(E)$ by Z and define $\hat{Y} = (Y, Z)$. Then X is independent of $\hat{Y}$ by our assumption. Define

$$\hat{h}(x,\hat{y}) = h(x,y)z \quad \text{and} \quad \hat{H}(\hat{y}) = \mathbb{E}(\hat{h}(X,\hat{y})),$$

where $\hat{y} = (y,z)$, with y and z having the same dimensions as Y and Z, respectively. Then we can apply item (a) to $\hat{h}$ and $\hat{H}$, obtaining

$$\mathbb{E}(\hat{h}(X,\hat{Y})) \;=\; \mathbb{E}(\hat{H}(\hat{Y})). \tag{A.4}$$

In addition, by the definition of $\hat{H}$ and H, we have

$$\hat{H}(\hat{y}) \;=\; \mathbb{E}(h(X,y)z) \;=\; H(y)z. \tag{A.5}$$

Plugging (A.5) and the definitions of $\hat{h}$ into (A.4), we obtain $\mathbb{E}(h(X,Y)Z) = \mathbb{E}(H(Y)Z)$, which is (A.3). This completes the proof. □

Remark A.1. *Taking expectation on both sides of the equality in item (b) of Lemma A.4, we can recover item (a) by the tower property of conditional expectation. We also note that item (b) is a generalization of [22, Example 4.1.7] (see also [12, page 148]), where $\mathcal{G} = \sigma(Y)$.*

Recalling that the probability of an event is the expectation of its indicator function, we obtain Corollary A.1 from item (b) of Lemma A.4.

Corollary A.1. *Let X and Y be random vectors, $\mathcal{G}$ be a σ-algebra, and h be a Borel measurable function. If X is independent of Y and $\mathcal{G}$, then $\mathbb{P}(h(X,Y) \geq 0 \mid \mathcal{G}) = \mathbb{E}(P(Y) \mid \mathcal{G})$ with $P(y) = \mathbb{P}(h(X,y) \geq 0)$.*

Now we present Lemma A.5, which will help us prove Lemma 3.4. As a preparation, given an event F with $\mathbb{P}(F) > 0$, we let $\mathbb{P}_F$ be the probability measure defined by $\mathbb{P}_F(E) = \mathbb{P}(E \mid F)$ for every event E, and $\mathbb{P}_F(\cdot \mid \mathcal{G})$ be the corresponding conditional probability with respect to a σ-algebra $\mathcal{G}$. Moreover, we let $\mathbb{E}_F$ denote the expectation under $\mathbb{P}_F$, and $\mathbb{E}_F(\cdot \mid \mathcal{G})$ denote the corresponding conditional expectation. It is well known that

$$\mathbb{E}(X\mathbb{1}(F)) \;=\; \mathbb{E}_F(X)\mathbb{P}(F) \tag{A.6}$$

for any random variable X (see, e.g., [28, Section 8.1]). Consequently, we have

$$\mathbb{E}(X\mathbb{1}(F)) \;=\; \mathbb{E}\left(\mathbb{E}_F(X \mid \mathcal{G})\mathbb{1}(F)\right). \tag{A.7}$$

To see (A.7), multiply both sides of the equality $\mathbb{E}_F(X) = \mathbb{E}_F(\mathbb{E}_F(X \mid \mathcal{G}))$ by $\mathbb{P}(F)$ and then apply (A.6).

Lemma A.5. *Let X be a random variable, F be an event with $\mathbb{P}(F) > 0$, and $\mathcal{G}$ be a σ-algebra.*

(a) *It holds that $\mathbb{E}(X\mathbb{1}(F) \mid \mathcal{G}) = \mathbb{E}_F(X \mid \mathcal{G})\,\mathbb{P}(F \mid \mathcal{G})$.*

(b) *For any $p \in [0,1]$, we have the following equivalence:*

$$\mathbb{E}(X\mathbb{1}(F) \mid \mathcal{G}) \;\geq\; p\mathbb{P}(F \mid \mathcal{G}) \quad (\mathbb{P}\text{-}a.s.) \qquad \Longleftrightarrow \qquad \mathbb{E}_F(X \mid \mathcal{G}) \;\geq\; p \quad (\mathbb{P}_F\text{-}a.s.).$$

Proof. (a) Since $\mathbb{E}_F(X \mid \mathcal{G})\mathbb{P}(F \mid \mathcal{G})$ is $\mathcal{G}$-measurable, the definition of conditional expectation tells us that we only need to verify

$$\mathbb{E}(\mathbb{E}_F(X \mid \mathcal{G})\,\mathbb{P}(F \mid \mathcal{G})\,\mathbb{1}(E)) \;=\; \mathbb{E}(X\mathbb{1}(F)\,\mathbb{1}(E)) \tag{A.8}$$

for all $E \in \mathcal{G}$. Denote $Y = \mathbb{E}_F(X \mid \mathcal{G})\mathbb{1}(E)$. Then the left-hand side of (A.8) equals

$$\mathbb{E}(Y\,\mathbb{P}(F \mid \mathcal{G})) \;=\; \mathbb{E}(Y\,\mathbb{E}(\mathbb{1}(F) \mid \mathcal{G})) \;=\; \mathbb{E}(\mathbb{E}(Y\mathbb{1}(F) \mid \mathcal{G})) \;=\; \mathbb{E}(Y\mathbb{1}(F)).$$

To calculate $\mathbb{E}(Y\mathbb{1}(F))$, we first note that $Y = \mathbb{E}_F(X\mathbb{1}(E) \mid \mathcal{G})$ and then apply (A.7) to the random variable $X\mathbb{1}(E)$, obtaining

$$\mathbb{E}(Y\mathbb{1}(F)) \;=\; \mathbb{E}\left(\mathbb{E}_F(X\mathbb{1}(E) \mid \mathcal{G})\,\mathbb{1}(F)\right) \;=\; \mathbb{E}\left([X\mathbb{1}(E)]\mathbb{1}(F)\right).$$

Therefore, equality (A.8) holds.

(b) Denote $Z = \mathbb{E}_F(X \mid \mathcal{G})$. According to (a), we only need to prove the equivalence

$$Z\mathbb{P}(F \mid \mathcal{G}) \;\geq\; p\mathbb{P}(F \mid \mathcal{G}) \quad (\mathbb{P}\text{-a.s.}) \qquad \Longleftrightarrow \qquad Z \;\geq\; p \quad (\mathbb{P}_F\text{-a.s.}). \tag{A.9}$$

To this end, defining the nonnegative random variable $W = \mathbb{1}(Z < p)\,\mathbb{P}(F \mid \mathcal{G})$, we observe that

$$\big\{Z\mathbb{P}(F \mid \mathcal{G}) \;\geq\; p\mathbb{P}(F \mid \mathcal{G})\big\} \;=\; \big\{Z \geq p \text{ or } \mathbb{P}(F \mid \mathcal{G}) = 0\big\} \;=\; \big\{W = 0\big\}, \tag{A.10}$$

and that

$$\mathbb{P}_F(Z < p)\,\mathbb{P}(F) \;=\; \mathbb{P}\left(\{Z < p\} \cap F\right) \;=\; \mathbb{E}\big(\mathbb{1}(Z < p)\mathbb{P}(F \mid \mathcal{G})\big) \;=\; \mathbb{E}(W). \tag{A.11}$$

Therefore, we have the following two equivalences:

$$Z\mathbb{P}(F \mid \mathcal{G}) \;\geq\; p\mathbb{P}(F \mid \mathcal{G}) \quad (\mathbb{P}\text{-a.s.}) \qquad \Longleftrightarrow \qquad W \;=\; 0 \quad (\mathbb{P}\text{-a.s.}) \qquad \Longleftrightarrow \qquad \mathbb{P}_F(Z < p) \;=\; 0,$$

where the first one is due to (A.10), while the second comes from (A.11) and the fact that $\mathbb{P}(F) > 0$. Hence, (A.9) holds. The proof is complete. □

Remark A.2. *Item (a) of Lemma A.5 generalizes equality (A.6). In light of (a), item (b) shows that we can cancel $\mathbb{P}(F \mid \mathcal{G})$ from both sides of the almost sure inequality $\mathbb{E}(X\mathbb{1}(F) \mid \mathcal{G}) \geq p\mathbb{P}(F \mid \mathcal{G})$, switching from $\mathbb{P}$ to $\mathbb{P}_F$ for the almost-sureness. For an event E, item (b) with $X = \mathbb{1}(E)$ leads to the equivalence*

$$\mathbb{P}(EF \mid \mathcal{G}) \;\geq\; p\mathbb{P}(F \mid \mathcal{G}) \quad (\mathbb{P}\text{-}a.s.) \qquad \Longleftrightarrow \qquad \mathbb{P}_F(E \mid \mathcal{G}) \;\geq\; p \quad (\mathbb{P}_F\text{-}a.s.). \tag{A.12}$$

Note that $\mathbb{P}(F \mid \mathcal{G}) = \mathbb{1}(F)$ if $F \in \mathcal{G}$, as is the case in the proof of Lemma 3.4.

Lemma A.6 presents an elementary inequality needed in the proof of Lemma 3.3.

Lemma A.6. *Suppose that $k > k_0 \geq 0$ and $0 < q < p \leq 1$. Then*

$$\inf_{t>0} \; t(kq - k_0) + p(k - k_0)(\mathrm{e}^{-t} - 1) \;\leq\; -\frac{(q-p)^2}{2p}(k + k_0).$$

Proof. Considering $t = \log(p/q)$, we only need to prove

$$(kq - k_0)\log(p/q) + (k - k_0)(q - p) \;\leq\; -\frac{(q-p)^2}{2p}(k + k_0). \tag{A.13}$$

Regard the left-hand side of (A.13) as a function of q and denote it by $\varphi(q)$. Then

$$\varphi(p) = 0, \quad \varphi'(p) \;=\; \frac{k_0}{p} - k_0 \;\geq\; 0, \quad \text{and} \quad \varphi''(q) \;=\; -\frac{k}{q} - \frac{k_0}{q^2}.$$

By the Taylor expansion of $\varphi(q)$ at the point p, there exists a $\xi \in (q,\, p)$ such that

$$\varphi(q) \;=\; \varphi'(p)(q-p) + \frac{1}{2}\varphi''(\xi)(q-p)^2 \;\leq\; -\frac{(q-p)^2}{2}\left(\frac{k}{\xi} + \frac{k_0}{\xi^2}\right) \;\leq\; -\frac{(q-p)^2}{2p}(k + k_0). \qquad \square$$

Lemma A.7 shows that $\mathrm{gap}(0, \partial_{\mathrm{C}} f(\cdot))$ is lower semicontinuous if f is convex as mentioned in Remark 4.2.

Lemma A.7. *Let $f : \mathbb{R}^n \to \mathbb{R}$ be convex. Then $\mu(x) = \mathrm{gap}(0, \partial_{\mathrm{C}} f(x))$ is lower semicontinuous for $x \in \mathbb{R}^n$.*

Proof. Fix an $x \in \mathbb{R}^n$ and an $\varepsilon > 0$. By [48, Corollary 24.5.1], there exists a $\delta > 0$ such that

$$\partial_{\mathrm{C}} f(y) \subseteq \partial_{\mathrm{C}} f(x) + \mathcal{B}(0, \varepsilon) \quad \text{for all } y \in \mathcal{B}(x, \delta).$$

This implies that

$$\mathrm{gap}(0, \partial_{\mathrm{C}} f(y)) \;\geq\; \mathrm{gap}(0, \partial_{\mathrm{C}} f(x)) - \varepsilon \quad \text{for all } y \in \mathcal{B}(x, \delta).$$

Hence, μ is lower semicontinuous. $\square$

B Discussions about the definition of probabilistic ascent

This section discusses Definition 3.1 of probabilistic ascent, especially the role of the indicator $\mathbb{1}(G_k \neq 0)$. As Remark B.1 will clarify, this indicator ensures that the concept of probabilistic ascent and the subsequent theory are invariant to the value of $\mathrm{cm}(\,\cdot\,, 0)$, which is purposefully unspecified in Definition 2.1. Removing this indicator from Definition 3.1 would make the theory rely on $\mathrm{cm}(\,\cdot\,, 0)$, leading to an undesirable restriction if one defines $\mathrm{cm}(\,\cdot\,, 0) = 1$ following [24], as will be explained in Remark B.3.

We begin with Proposition B.1, which provides two equivalent reformulations for the inequality in condition (3.2) of Definition 3.1. This proposition can be proved by applying Lemma A.3 to the events $E = \{G_k = 0\}$ and $F = \{\mathrm{cm}(\mathfrak{D}_k, -G_k) \leq 0\}$ while noting that $E \cup F = \{\min_{\mathfrak{d} \in \mathfrak{D}_k} \mathfrak{d}^{\mathsf{T}} G_k \;\geq\; 0\}$.

Proposition B.1. *Let $p \in [0, 1]$. Consider Algorithm 2.2 with f being continuously differentiable on $\mathbb{R}^n$. For each $k \geq 0$, the following inequalities are equivalent to each other.*

(a) $\mathbb{P}\left(\mathrm{cm}(\mathfrak{D}_k, -G_k) \leq 0 \mid \mathcal{F}_{k-1}\right) \geq p\mathbb{1}(G_k \neq 0)$.

(b) $\mathbb{P}\left(\{\mathrm{cm}(\mathfrak{D}_k, -G_k) \leq 0\} \cap \{G_k \neq 0\} \mid \mathcal{F}_{k-1}\right) \geq p\mathbb{1}(G_k \neq 0)$.

(c) $\mathbb{P}\left(\min_{\mathfrak{d} \in \mathfrak{D}_k} \mathfrak{d}^{\mathsf{T}} G_k \geq 0 \mid \mathcal{F}_{k-1}\right) \geq p$.

Remark B.1. *Neither item (b) nor (c) in Proposition B.1 relies on the value of* $\mathrm{cm}(\,\cdot\,,0)$. *Therefore, condition (3.2) based on (a) is independent of* $\mathrm{cm}(\,\cdot\,,0)$. *Consequently, no matter how we define* $\mathrm{cm}(\,\cdot\,,0)$, *Definition 3.1 of probabilistic ascent is invariant, and the results in this paper hold without any modification.*

Remark B.2. *Since items (a) and (c) in Proposition B.1 are equivalent, condition (3.2) is equivalent to*

$$\mathbb{P}\Big(\min_{\mathfrak{d}\in\mathfrak{D}_k} \mathfrak{d}^\mathsf{T} G_k \geq 0 \;\Big|\; \mathcal{F}_{k-1}\Big) \;\geq\; p \quad \textit{for each } k \geq 0. \tag{B.1}$$

This is exactly condition (4.1) in the continuously differentiable case, where we have $f^\circ(X_k;\mathfrak{d}) = \mathfrak{d}^\mathsf{T} G_k$.

What if we dropped the indicator $\mathbb{1}(G_k \neq 0)$ from the definition of probabilistic ascent, adopting an alternative definition that requires

$$\mathbb{P}\left(\mathrm{cm}\left(\mathfrak{D}_k, -G_k\right) \leq 0 \mid \mathcal{F}_{k-1}\right) \;\geq\; p \quad \text{for each } k \geq 0 \tag{B.2}$$

in place of condition (3.2)? Above all, every result requiring p-probabilistic ascent in this paper would still hold, since (B.2) is stronger than (3.2). However, as we will explain in Remark B.3, the theory built on this alternative definition would rely on the value of $\mathrm{cm}(\,\cdot\,,0)$, which would imply an undesirable limitation if one defines $\mathrm{cm}(\,\cdot\,,0) = 1$ as in [24].

Remark B.3. *Suppose that we define* $\mathrm{cm}(\,\cdot\,,0) = 1$ *following [24]. Then condition (B.2) cannot be satisfied for any* $p > 0$ *unless*

$$\mathbb{P}(G_k = 0) \;=\; 0 \quad \textit{for each } k \geq 0. \tag{B.3}$$

Condition (B.3) means that Algorithm 2.2 almost never steps on a stationary point, which is a restriction that we do not want to impose. To see why (B.2) with $p > 0$ *necessitates (B.3), let us assume that* $\mathbb{P}(G_k = 0) > 0$ *for a certain* $k \geq 0$. *Then (B.2) and Lemma A.2 will lead to the contradiction that*

$$p \;\leq\; \mathbb{P}(\mathrm{cm}\left(\mathfrak{D}_k, -G_k\right) \leq 0 \mid G_k = 0) \;=\; \mathbb{P}(\mathrm{cm}\left(\mathfrak{D}_k, 0\right) \leq 0 \mid G_k = 0) \;=\; 0,$$

where the last step is because $\mathrm{cm}(\mathfrak{D}_k, 0) = 1$.

To illustrate Remark B.3, let us recall Example 3.1, choosing $x_0 = \alpha_0$ in particular. Then we have

$$\mathbb{P}(G_1 = 0) = \mathbb{P}(\mathfrak{d}_0 = -1) = \frac{1}{2},$$

violating (B.3) for $k = 1$. Consequently, Remark B.3 shows that condition (B.2) with $p > 0$ cannot hold if one defines $\mathrm{cm}(\,\cdot\,,0) = 1$ as in [24], making the theory based on (B.2) inapplicable to such a simple example, even though $\{\mathfrak{D}_k\}$ fits our definition of 1/2-probabilistic ascent according to Proposition 3.1. This example also clarifies why (B.3) is undesirable to impose when analyzing randomized algorithms like Algorithm 2.2, although it is not uncommon to assume that algorithms never step on stationary points in the deterministic case (e.g., [40, Section 1]).

Before ending this section, we propose Definition B.1 as an alternative to Definition 2.2 for probabilistic descent. In this definition, $\mathbb{1}(G_k \neq 0)$ plays a role like in Definition 3.1.

Definition B.1 (Alternative definition of probabilistic descent)**.** Identical to Definition 2.2 except that we replace condition (2.5) with

$$\mathbb{P}\left(\mathrm{cm}\left(\mathfrak{D}_k, -G_k\right) \geq \kappa \mid \mathcal{F}_{k-1}\right) \;\geq\; p\mathbb{1}(G_k \neq 0) \quad \text{for each } k \geq 0. \tag{B.4}$$

Similar to Definition 3.1, Definition B.1 is invariant to the value of $\mathrm{cm}(\,\cdot\,,0)$. If $\mathrm{cm}(\,\cdot\,,0)=1$ as in [24], it is equivalent to Definition 2.2, because we have $\{G_k=0\}\subseteq\{\mathrm{cm}(\mathfrak{D}_k,-G_k)\geq\kappa\}$ in this case, which ensures the equivalence by Lemma A.3. However, if one chooses to define $\mathrm{cm}(\,\cdot\,,0)<\kappa$ (e.g., $\mathrm{cm}(\,\cdot\,,0)=0$ may be appealing for symmetry), the argument in Remark B.3 can be adapted to show that Definition 2.2 also necessitates (B.3) when $p>0$, whereas Definition B.1 is free from this restriction.

C Proofs of Propositions 3.1, 3.4, 4.1, and Lemmas 3.3–3.4

This section proves several propositions and lemmas that appeared in the main text. They will be proved in the order of their appearance.

Proposition 3.1 can be established with the help of Corollary A.1.

Proof of Proposition 3.1. Fix an arbitrary $k\geq 0$. Due to Proposition B.1, it suffices to prove that

$$\mathbb{P}\left(\min_{\mathfrak{d}\in\mathfrak{D}_k}\mathfrak{d}^{\mathsf{T}}G_k\geq 0\,\middle|\,\mathcal{F}_{k-1}\right)\;\geq\;2^{-m}. \tag{C.1}$$

By our assumption, $\mathfrak{D}_k$ is independent of G_k and $\mathcal{F}_{k-1}$. Hence, Corollary A.1 implies that the conditional probability in (C.1) equals $\mathbb{E}\left(P(G_k)\mid\mathcal{F}_{k-1}\right)$ with

$$P(g)\;=\;\mathbb{P}\left(\min_{\mathfrak{d}\in\mathfrak{D}_k}\mathfrak{d}^{\mathsf{T}}g\geq 0\right)\;=\;\mathbb{P}\left(\mathfrak{d}^{\mathsf{T}}g\geq 0\text{ for each }\mathfrak{d}\in\mathfrak{D}_k\right). \tag{C.2}$$

Since $\mathfrak{D}_k$ consists of m independent random vectors uniformly distributed on the unit sphere, the right-hand side of (C.2) is at least 2^{-m} (it is exactly 2^{-m} if $g\neq 0$ and is 1 if $g=0$). Therefore, (C.1) holds. □

Indeed, $\mathbb{E}(P(G_k)\mid\mathcal{F}_{k-1})$ in the above proof equals $P(G_k)$ since G_k is $\mathcal{F}_{k-1}$-measurable, but we do not need this fact to establish the desired inequality.

Lemma 3.3 can be obtained using the moment method for deriving Chernoff bounds [35].

Proof of Lemma 3.3. The inequality in (3.20) holds trivially when $k\leq k_0$, because (3.5) implies that $E_{k_0}\subseteq\{\overline{Y}_k=0\}$ in this case, rendering $\mathbb{P}(\overline{Y}_k\geq 1-q\mid E_{k_0})=0$ since $q<1$. Let us focus on the nontrivial case where $k>k_0\geq 0$.

Fixing an arbitrary $t>0$, we first make two claims: one is

$$\mathbb{P}\left(1-\overline{Y}_k\leq q\mid E_{k_0}\right)\;\leq\;\mathrm{e}^{t(kq-k_0)}\,\mathbb{E}\left(\prod_{\ell=k_0}^{k-1}\mathrm{e}^{-t(1-Y_\ell)}\,\middle|\,E_{k_0}\right), \tag{C.3}$$

and the other is

$$\mathbb{E}\left(\prod_{\ell=k_0}^{k-1}\mathrm{e}^{-t(1-Y_\ell)}\,\middle|\,E_{k_0}\right)\;\leq\;\exp[p(k-k_0)(\mathrm{e}^{-t}-1)]. \tag{C.4}$$

Once inequalities (C.3) and (C.4) are proved, we will have

$$\mathbb{P}\left(1-\overline{Y}_k\leq q\mid E_{k_0}\right)\;\leq\;\exp[t(kq-k_0)+p(k-k_0)(\mathrm{e}^{-t}-1)],$$

which will render (3.20) according to Lemma A.6. We now prove the two claims by standard techniques.

For (C.3), by the definitions of $\overline{Y}_k$ and E_{k_0} in (3.5) as well as Markov's inequality, we have

$$\begin{aligned}\mathbb{P}\left(1-\overline{Y}_k \le q \mid E_{k_0}\right) &= \mathbb{P}\left(\exp\Big[-t\sum_{\ell=0}^{k-1}(1-Y_\ell)\Big] \ge \mathrm{e}^{-tkq} \,\Big|\, E_{k_0}\right) \\ &\le \mathrm{e}^{tkq}\,\mathbb{E}\left(\prod_{\ell=0}^{k-1}\mathrm{e}^{-t(1-Y_\ell)} \,\Big|\, E_{k_0}\right) = \mathrm{e}^{t(kq-k_0)}\,\mathbb{E}\left(\prod_{\ell=k_0}^{k-1}\mathrm{e}^{-t(1-Y_\ell)} \,\Big|\, E_{k_0}\right),\end{aligned}$$

where the last equality is because $\prod_{\ell=0}^{k_0-1}\mathrm{e}^{-t(1-Y_\ell)} = \mathrm{e}^{-tk_0}$ when E_{k_0} happens.

For (C.4), we use the tower property of conditional expectation to get

$$\mathbb{E}\left(\prod_{\ell=k_0}^{k-1}\mathrm{e}^{-t(1-Y_\ell)} \,\Big|\, \mathcal{F}_{k_0-1}\right) = \mathbb{E}\left(\mathbb{E}\left(\mathrm{e}^{-t(1-Y_{k-1})} \mid \mathcal{F}_{k-2}\right)\prod_{\ell=k_0}^{k-2}\mathrm{e}^{-t(1-Y_\ell)} \,\Big|\, \mathcal{F}_{k_0-1}\right), \tag{C.5}$$

with $\prod_{\ell=k_0}^{k-2}\mathrm{e}^{-t(1-Y_\ell)} = 1$ when $k = k_0+1$. By condition (3.4), we have

$$\mathbb{E}\left(\mathrm{e}^{-t(1-Y_{k-1})} \mid \mathcal{F}_{k-2}\right) \le p\mathrm{e}^{-t} + (1-p) \le \exp(p\mathrm{e}^{-t} - p), \tag{C.6}$$

where the last inequality is because $x+1 \le \mathrm{e}^x$ for all x. By equality (C.5) and inequality (C.6), we have

$$\begin{aligned}\mathbb{E}\left(\prod_{\ell=k_0}^{k-1}\mathrm{e}^{-t(1-Y_\ell)} \,\Big|\, \mathcal{F}_{k_0-1}\right) &\le \exp[p(\mathrm{e}^{-t}-1)]\,\mathbb{E}\left(\prod_{\ell=k_0}^{k-2}\mathrm{e}^{-t(1-Y_\ell)} \,\Big|\, \mathcal{F}_{k_0-1}\right) \\ &\le \exp[p(k-k_0)(\mathrm{e}^{-t}-1)],\end{aligned} \tag{C.7}$$

the second inequality following from the recursive application of the first one. Since $E_{k_0} \in \mathcal{F}_{k_0-1}$ and $\mathbb{P}(E_{k_0}) > 0$ by Remark 3.6, inequality (C.7) implies (C.4) by Lemma A.2. The proof is complete. □

Lemma 3.4 is a straightforward consequence of Lemma A.5, or, more precisely, Remark A.2.

Proof of Lemma 3.4. Since $p > 0$, the probability measure $\mathbb{P}(\,\cdot \mid E_{k_0})$ is well defined according to Remark 3.6. Fix an integer $k \ge 0$. Then $E_{k_0} \in \mathcal{F}_{k_0-1} \subseteq \mathcal{F}_{k_0+k-1} = \tilde{\mathcal{F}}_{k-1}$ by the definitions of $\{\mathcal{F}_k\}$ and $\{\tilde{\mathcal{F}}_k\}$. Thus, condition (3.4) and Lemma A.1 yield

$$\mathbb{P}(\{\tilde{Y}_k = 0\} \cap E_{k_0} \mid \tilde{\mathcal{F}}_{k-1}) = \mathbb{P}(Y_{k_0+k} = 0 \mid \mathcal{F}_{k_0+k-1})\,\mathbb{1}(E_{k_0}) \ge p\mathbb{1}(E_{k_0}).$$

Hence, recalling that $\tilde{\mathbb{P}}(\cdot) = \mathbb{P}(\,\cdot \mid E_{k_0})$, we have $\tilde{\mathbb{P}}(\tilde{Y}_k = 0 \mid \tilde{\mathcal{F}}_{k-1}) \ge p$ according to Remark A.2. □

Proposition 3.4 can be proved using Lévy's Conditional Borel–Cantelli Lemma [32, Corollaire 68] (see also [22, Theorem 4.3.4] and [28, Exercise 11.2.7]).

Proof of Proposition 3.4. Since $Y_kU_k \ge 0$ for each $k \ge 0$, it suffices to prove that

$$\mathbb{P}\left(\limsup_{k\to\infty} Y_kU_k > 0\right) = 1. \tag{C.8}$$

To this end, we first note that

$$\mathbb{P}\left(\limsup_{k\to\infty} U_k > 0\right) = 1. \tag{C.9}$$

Indeed, since $\log U_k = \sum_{\ell=0}^{k-1}[Y_\ell\log\gamma + (1-Y_\ell)\log\theta]$, condition (3.44) and the definition of p_* in (1.4) ensure that $\{\log U_k\}$ is a martingale with respect to $\{\mathcal{F}_{k-1}\}$, and this martingale has bounded increments. Hence, [22, Theorem 4.3.1] indicates that $\limsup_k(\log U_k) > -\infty$ a.s., which is equivalent to (C.9).

To finish the proof of (C.8), we will demonstrate that

$$\mathbb{P}\left(Y_k U_k \geq \varepsilon \text{ i.o.}\right) \geq \mathbb{P}\left(U_k \geq \varepsilon \text{ i.o.}\right) \quad \text{for all } \varepsilon > 0. \tag{C.10}$$

Once (C.10) is established, plugging $\varepsilon = \ell^{-1}$ into it and then taking limit as $\ell \to \infty$ will render $\mathbb{P}(\limsup_k Y_k U_k > 0) \geq \mathbb{P}(\limsup_k U_k > 0)$, which will lead to (C.8) in light of (C.9).

Now we prove (C.10) for an arbitrary $\varepsilon > 0$. For each $k \geq 0$, since $\{Y_k U_k \geq \varepsilon\} = \{Y_k = 1\} \cap \{U_k \geq \varepsilon\}$ and $\{U_k \geq \varepsilon\} \in \mathcal{F}_{k-1}$, Lemma A.1 and condition (3.44) yield

$$\mathbb{P}(Y_k U_k \geq \varepsilon \mid \mathcal{F}_{k-1}) = \mathbb{P}(Y_k = 1 \mid \mathcal{F}_{k-1})\, \mathbb{1}(U_k \geq \varepsilon) = (1 - p_*)\, \mathbb{1}(U_k \geq \varepsilon). \tag{C.11}$$

Recalling that $0 < \theta < 1 \leq \gamma$, we have $p_* < 1$ according to (1.4). Hence, (C.11) implies that

$$\left\{\sum_{k=0}^{\infty} \mathbb{P}(Y_k U_k \geq \varepsilon \mid \mathcal{F}_{k-1}) = \infty\right\} \supseteq \left\{\sum_{k=0}^{\infty} \mathbb{1}(U_k \geq \varepsilon) = \infty\right\} = \{U_k \geq \varepsilon \text{ i.o.}\}. \tag{C.12}$$

Meanwhile, the left-hand side of (C.12) has the same probability as the event $\{Y_k U_k \geq \varepsilon \text{ i.o.}\}$ by Lévy's Conditional Borel–Cantelli Lemma. Therefore, (C.10) holds and the proof is complete. □

The proof of Proposition 4.1 is similar to that of Proposition 3.1.

Proof of Proposition 4.1. Similar to the proof of Proposition 3.1, Corollary A.1 implies that the conditional probability in condition (4.1) equals $\mathbb{E}\left(P(X_k) \mid \mathcal{F}_{k-1}\right)$ with

$$P(x) = \mathbb{P}\left(\min_{\mathfrak{d} \in \mathfrak{D}_k} f^\circ(x; \mathfrak{d}) \geq 0\right). \tag{C.13}$$

For any $x \in \mathbb{R}^n$, picking a $g \in \partial_{\mathrm{C}} f(x)$, we have $P(x) \geq \mathbb{P}\left(\min_{\mathfrak{d} \in \mathfrak{D}_k} \mathfrak{d}^{\mathsf{T}} g \geq 0\right)$, which is at least 2^{-m} as in the proof of Proposition 3.1. Hence, (4.1) holds with $p = 2^{-m}$ and the proof is complete. □

D (Non-)Measurability of iterates with respect to polling directions

In this section, we discuss when the iterates of Algorithm 2.2 are measurable with respect to the polling directions, and when they are not. Often omitted in literature, this type of discussion is essential for the mathematical rigour of our analysis. Indeed, as we will see in Example D.1, the measurability can fail for some implementations of Algorithm 2.2. For the concept of measurability, we refer to [22, Section 1.2].

Lemma D.1 establishes the measurability of the iterates for certain implementations of Algorithm 2.2, covering [24, Algorithm 2.1]. The proof is elementary, but it clarifies the role of the polling strategy in the measurability.

Lemma D.1. *Let m be a positive integer and f be continuous on $\mathbb{R}^n$. Consider Algorithm 2.2 with the following configuration for each $k \geq 0$.*

(a) *Generate $\mathfrak{D}_k = \{\mathfrak{d}_k^1, \ldots, \mathfrak{d}_k^m\}$ with $\mathfrak{d}_k^1, \ldots, \mathfrak{d}_k^m$ being random vectors.*

(b) *Set the order of function evaluations as $f(X_k + A_k \mathfrak{d}_k^1), \ldots, f(X_k + A_k \mathfrak{d}_k^m)$ before polling.*

(c) *Use either opportunistic polling or complete polling.*

Let $\mathcal{F}_k^{\mathfrak{D}} = \sigma(\mathfrak{D}_0, \ldots, \mathfrak{D}_k)$ for each $k \geq 0$ and $\mathcal{F}_{-1}^{\mathfrak{D}} = \{\emptyset, \Omega\}$. Then X_k is $\mathcal{F}_{k-1}^{\mathfrak{D}}$-measurable for each $k \geq 0$.

Proof. We will prove by induction that X_k and A_k are both $\mathcal{F}_{k-1}^{\mathfrak{D}}$-measurable for each $k \geq 0$. The base case $k = 0$ holds trivially since X_0 and A_0 are not random. Assuming that X_k and A_k are $\mathcal{F}_{k-1}^{\mathfrak{D}}$-measurable, let us prove that X_{k+1} and A_{k+1} are both $\mathcal{F}_k^{\mathfrak{D}}$-measurable. Before starting, note that the induction hypothesis implies that X_k and A_k are $\mathcal{F}_k^{\mathfrak{D}}$-measurable since $\mathcal{F}_{k-1}^{\mathfrak{D}} \subseteq \mathcal{F}_k^{\mathfrak{D}}$. Define $\mathfrak{d}_k^0 = 0$ and

$$V^i \;=\; f(X_k + A_k \mathfrak{d}_k^i), \quad i = 0, 1, \ldots, m.$$

Then each V^i is $\mathcal{F}_k^{\mathfrak{D}}$-measurable since f is continuous. $\rho(A_k)$ is also $\mathcal{F}_k^{\mathfrak{D}}$-measurable as ρ is monotone.

Now, we consider the case of complete polling. In this case,

$$X_{k+1} \;=\; X_k + A_k \sum_{i=1}^{m} \mathfrak{d}_k^i W^i, \tag{D.1}$$

where W^i $(i = 1, \ldots, m)$ is the indicator defined by

$$W^i \;=\; \mathbb{1}\big(i \text{ is the smallest integer such that } V^i = \min\{V^1, \ldots, V^m\}, \text{ and } V^0 - V^i > \rho(A_k)\big)\,.$$

Note that at most one of $W^1, \ldots, W^m$ is 1, and they are all 0 if complete polling fails. Moreover,

$$W^i \;=\; \left[\prod_{j=1}^{i-1} \mathbb{1}\big(V^i < V^j\big) \prod_{j=i+1}^{m} \mathbb{1}\big(V^i \leq V^j\big)\right] \mathbb{1}\big(V^0 - V^i > \rho(A_k)\big)\,,$$

which is $\mathcal{F}_k^{\mathfrak{D}}$-measurable due to the $\mathcal{F}_k^{\mathfrak{D}}$-measurability of $V^0, \ldots, V^m$ and $\rho(A_k)$. Therefore, X_{k+1} is $\mathcal{F}_k^{\mathfrak{D}}$-measurable according to (D.1). Consequently, A_{k+1} is $\mathcal{F}_k^{\mathfrak{D}}$-measurable by the recurrence relation (2.4) and the induction hypothesis. The induction finishes for complete polling.

The case of opportunistic polling can be handled similarly. In this case, equation (D.1) holds with

$$\begin{aligned} W^i \;&=\; \mathbb{1}\big(i \text{ is the smallest integer such that } V^0 - V^i > \rho(A_k)\big) \\ &=\; \left[\prod_{j=1}^{i-1} \mathbb{1}\big(V^0 - V^j \leq \rho(A_k)\big)\right] \mathbb{1}\big(V^0 - V^i > \rho(A_k)\big)\,, \end{aligned}$$

which is $\mathcal{F}_k^{\mathfrak{D}}$-measurable. Everything else is the same as complete polling. □

However, if the polling in Algorithm 2.2 involves randomness beyond the polling directions, then X_k may not be $\mathcal{F}_{k-1}^{\mathfrak{D}}$-measurable. This is illustrated by Example D.1. For this reason, our analysis uses $\mathcal{F}_k = \sigma(\mathfrak{D}_0, X_1, \ldots, \mathfrak{D}_k, X_{k+1})$ rather than $\mathcal{F}_k^{\mathfrak{D}}$ as the filtration.

Example D.1. *Let m be a positive integer and f be continuous on $\mathbb{R}^n$. Consider Algorithm 2.2 with the following configuration for each $k \geq 0$.*

(a) *Generate $\mathfrak{D}_k = \{\mathfrak{d}_k^1, \ldots, \mathfrak{d}_k^m\}$ with $\mathfrak{d}_k^1, \ldots, \mathfrak{d}_k^m$ being random vectors.*

(b) *Pick a random permutation π_k of $\{1, \ldots, m\}$.*

(c) *Set the order of function evaluations as $f(X_k + A_k \mathfrak{d}_k^{\pi_k(1)}), \ldots, f(X_k + A_k \mathfrak{d}_k^{\pi_k(m)})$ before polling.*

(d) *Use opportunistic polling.*

Since the calculation of X_k involves π_{k-1}, we cannot guarantee the $\mathcal{F}_{k-1}^{\mathfrak{D}}$-measurability of X_k if π_{k-1} is not $\mathcal{F}_{k-1}^{\mathfrak{D}}$-measurable, or informally, if π_{k-1} contains randomness beyond $\mathcal{F}_{k-1}^{\mathfrak{D}}$. For example, similar to [20, Section 4], we can define π_{k-1} by ranking the directions in $\mathfrak{D}_{k-1}$ according to a stochastic oracle independent of $\mathfrak{D}_{k-1}$. Then X_k is measurable with respect to $\sigma(\mathfrak{D}_0, \pi_0, \ldots, \mathfrak{D}_{k-1}, \pi_{k-1})$, but not necessarily with respect to $\mathcal{F}_{k-1}^{\mathfrak{D}}$.