Non-convergence Analysis of Probabilistic Direct Search

Cunxin Huang* Zaikun Zhang†

February 15, 2026

Abstract

Direct-search methods are a major class in derivative-free optimization. The combination of direct search and randomization techniques leads to an efficient variant, namely probabilistic direct search. Its convergence analysis has been thoroughly explored in recent years under the probabilistic descent assumption. However, a natural question arises: how will this algorithm behave when assumptions for convergence are not met? In this paper, we analyze the non-convergence of the algorithm when the polling directions form probabilistic ascent sets. Its analysis is closely related to the discussion of a random series. We further show that our non-convergence analysis is tight. Our non-convergence theory completes the analytical framework for the probabilistic direct search, guiding the selection of the polling directions in practice.

Keywords: Derivative-free optimization, Direct search, Probabilistic method, Non-convergence analysis

1 Introduction

When will your algorithm fail to converge? This question is arguably as important as asking when it will converge, but is often not studied as much. A systematic investigation of this question may deepen our understanding about the behavior of the algorithm, guide its implementation in practice, and provide new perspectives on its convergence analysis. Our paper will address this question for the probabilistic direct search method [17] for the unconstrained optimization problem

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

where $f : \mathbb{R}^n \to \mathbb{R}$ is a smooth and convex function.

Direct search [21] is a class of derivative-free optimization (DFO) methods. They define iterates based on comparisons of function values sampled following a certain scheme without

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

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

explicitly building models for the objective or constraint functions. There are several types of direct search methods, and we will focus on the directional direct search based on sufficient decrease [14, Section 7.7] for solving (1.1). In the worst case, the deterministic version of this method needs to evaluate at least $n+1$ function values at each iteration, which becomes impractical when n is modestly large. To overcome this difficulty, Gratton et al. [17] propose a randomized version of this method, which we will refer to as the *probabilistic direct search*. They show that given shrinking factor θ and expanding factor γ, the algorithm enjoys global convergence if the sequence of polling direction sets is a sequence of p_0-probabilistic κ-descent sets with some positive κ and

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

In particular, if $\gamma > 1$ and we choose each polling direction set to be a collection of m independent random directions following the uniform distribution on the unit sphere, which is the typical choice in practice [17], then a sufficient condition for global convergence is

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

This result not only provides more choices of polling direction sets for direct search, but also guides the analysis of the probabilistic trust-region model [18].

A natural question arises: what will happen if $m \le \log_2(1-\log\theta/\log\gamma)$? Furthermore, we would like to ask: is the p_0-probabilistic κ-descent assumption essential for the convergence of probabilistic direct search? From a broader perspective, we are interested in whether "submartingale-like" assumptions are essential, which were first introduced in [3] and are widely used in the convergence analysis of stochastic oracles including randomized versions (some called probabilistic models) of optimization methods such as trust region [3, 37], line search [5, 8], and cubic regularization [8]. These questions are both theoretically interesting and practically meaningful, as the answers will provide a complete view of the behavior of probabilistic models and guide the selection of algorithmic parameters in practice.

In this paper, we answer the first two questions as a first step. We establish the *non-convergence theory* of probabilistic direct search and prove that the algorithm will not converge if the sequence of polling direction sets is a sequence of p-probabilistic ascent sets (Definition 3.1) with $p > 1-p_0$ and the objective function is smooth and convex. In particular, for the above-mentioned typical case, the algorithm will not be globally convergent if

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

It is still an open question whether the algorithm will converge when inequality (1.3) becomes an equality, although $\log_2(1-\log\theta/\log\gamma)$ is not an integer in most cases.

The remaining part of this paper is organized as follows. In Section 2, we provide a concise review of DFO and introduce the necessary concepts of probabilistic direct search. Section 3

establishes the non-convergence theory, forming the main ideas of this paper. Additionally, we show that the probabilistic direct search with typical polling direction sets will not converge if $m < \log_2(1 - \log\theta/\log\gamma)$ by our non-convergence result. Moreover, we construct an example to demonstrate that our probabilistic ascent assumption with $p > 1 - p_0$ cannot be weakened to $p \geq 1 - p_0$. A relaxation of the probabilistic ascent assumption is later discussed. We extend our non-convergence results to the nonsmooth case in Section 4. We summarize our findings and draw conclusions in Section 5.

2 Preliminaries

To put our research in context, we briefly review the landscape of DFO, which in recent decades has aroused great interest in both academic research and practical applications [2, 14, 22]. Within the existing body of literature, DFO methods are broadly classified into two primary categories: direct-search methods and model-based methods. Detailed discussions about direct-search methods can be found in [21], and notable examples of direct search include the Nelder-Mead simplex method [25], the MADS methods [1, 23], and BFO [26, 27]. In contrast to direct-search methods using simple comparisons of function values, model-based methods construct local models through sampling under a trust-region [13] or line-search [4] framework. A wealth of classical literature on model-based methods can be found in, for example, [4, 13, 29, 30, 31, 32], with some well-known methods and software in this category including Powell's methods and PDFO [33]. Recently, randomization techniques are introduced to both categories, and we refer to [3, 6, 7, 17, 18, 19].

In what follows, we review the framework of probabilistic direct search and introduce the necessary notations. Subsection 2.1 introduces the fundamental framework of direct search based on sufficient decrease, whereas Subsection 2.2 concentrates on the randomization techniques inherent in this framework along with the convergence theory.

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 called 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** [14, 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 [21, Algorithm 3.2]). As in [17], we omit such an option and focus on polling.

Remark 2.1. *To implement Algorithm 2.1, a polling strategy is needed to choose the direction d_k if there are multiple candidates satisfying (2.1). Two common strategies exist. One is to choose the direction that decreases the function value the most (pick the first in case of a tie), which is called complete polling. The other is to take the first direction fulfilling (2.1), which is known as*

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

opportunistic polling. We also need to set an order for evaluating $\{f(x_k + \alpha_k d) : d \in \mathcal{D}_k\}$ in the polling. A strategy suggested in [15, Section 4] is to decide the order by an oracle that can help us rank the decreases of f along the polling directions, the oracle in [15] being an approximate descent direction (called a descent indicator). 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 the set $\mathcal{D}$ with respect to a nonzero vector v, denoted by $\mathrm{cm}(\mathcal{D}, v)$, is defined as

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

where $\|\cdot\|$ is the Euclidean norm. In addition, the cosine measure of the set $\mathcal{D}$, 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 2.2. *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 $[-1, 1]$ (e.g., [17] defines $\mathrm{cm}(\cdot\,, 0) = 1$). We do not particularize this constant, because its value will not affect our non-convergence analysis. See Remark 3.2 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 [21, Theorem 3.11].

2.2 Probabilistic direct search and its convergence

Algorithm 2.2 presents the probabilistic direct search method, which was initially proposed in [17]. 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 in general, although the starting point and the initial step size are still chosen deterministically.

Algorithm 2.2 Probabilistic direct search based on sufficient decrease

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 [17], we adopt the notations summarized in Table 1. Additionally, we denote

$$G_k \;=\; \nabla f(X_k).$$

Table 1: Notations for random elements and their realizations

	Polling direction set	Iterate	Step size
Random element	$\mathfrak{D}_k$	X_k	A_k
Realization	$\mathcal{D}_k$	x_k	α_k

Similar to [17, Assumption 2.3], 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 investigation into Algorithm 2.2 heavily relies on the concept of σ-algebras and conditional probability with respect to them [16, Section 4.1]. For each $k \geq 0$, we define

$$\mathcal{F}_k \;=\; \sigma(\mathfrak{D}_0, X_1, \ldots, \mathfrak{D}_k, X_{k+1}), \tag{2.2}$$

which is the σ-algebra generated by $\mathfrak{D}_0, X_1, \ldots, \mathfrak{D}_k, X_{k+1}$. In addition, we define

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

Roughly speaking, $\mathcal{F}_k$ 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. Note that $\mathcal{F}_k$ does not involve X_0, which is deterministically chosen. Obviously, $\mathfrak{D}_k$ is $\mathcal{F}_k$ measurable and X_k is $\mathcal{F}_{k-1}$-measurable for each $k \geq 0$. If f is continuously differentiable, then G_k is also $\mathcal{F}_{k-1}$-measurable. In addition, A_k is $\mathcal{F}_{k-1}$-measurable by mathematical induction based on the recurrence

$$A_{k+1} \;=\; \gamma^{\mathbb{1}(X_{k+1} \neq X_k)} \theta^{\mathbb{1}(X_{k+1} = X_k)} A_k, \tag{2.3}$$

which holds because we have $\{A_{k+1} = \gamma A_k\} = \{X_{k+1} \neq X_k\}$ and $\{A_{k+1} = \theta A_k\} = \{X_{k+1} = X_k\}$ in Algorithm 2.2.

The global convergence theory of probabilistic direct search can be stated as follows.

Definition 2.2 ([17, 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 it satisfies

$$\mathbb{P}\left(\mathrm{cm}\left(\mathfrak{D}_k, -G_k\right) \geq \kappa \mid \mathcal{F}_{k-1}\right) \geq p \quad \text{for each } k \geq 0. \tag{2.4}$$

Theorem 2.1 ([17, 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 2.3. *The σ-algebra $\mathcal{F}_k$ defined in (2.2) will reduce to $\sigma(\mathfrak{D}_0, \ldots, \mathfrak{D}_{k-1})$ if we assume that X_k is measurable with respect to $\sigma(\mathfrak{D}_0, \ldots, \mathfrak{D}_{k-1})$. As clarified in Lemma C.1, this assumption is fulfilled by implementations of Algorithm 2.2 considered in [17]. However, such an assumption is not guaranteed if we allow the unspecified polling strategy in Algorithm 2.2 to involve randomness beyond the polling directions (see Example C.1). Therefore, we choose not to impose such an assumption. In this sense, Theorem 2.1 is indeed a slightly generalized version of [17, Theorem 3.4], but the proof remains essentially the same.*

Remark 2.4. *The probability in (2.4) is a probability with respect to a σ-algebra, which is a random variable (see [16, Section 4.1]). Following the convention in probability theory (e.g., [16, Page 179] and [20, Page 195]), the inequality in Definition 2.2 should be understood in the almost sure sense, that is,*

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

This is because the conditional probability $\mathbb{P}(\cdot \mid \mathcal{F}_{k-1})$, as a random variable, is only defined up to almost sure equivalence. Henceforth, all the equalities and inequalities should be understood in this way if they involve conditional probabilities or expectations with respect to a σ-algebra, and we will not repeat this point every time.

In practice, $\mathfrak{D}_k$ is typically chosen to be m independent random vectors uniformly distributed on the unit sphere in $\mathbb{R}^n$. Theorem 2.1 leads to Corollary 2.1 for this typical implementation.

Corollary 2.1 ([17, Corollary B.4]). *Consider Algorithm 2.2 with f satisfying the assumptions in Theorem 2.1. Let $\gamma > 1$ be a constant, $\{\mathfrak{D}_k\}$ be mutually independent, and each $\mathfrak{D}_k$ be a set of m independent random vectors uniformly distributed on the unit sphere in $\mathbb{R}^n$. Then $\mathbb{P}(\liminf_k \|G_k\| = 0) = 1$ if $m > \log_2(1 - \log\theta/\log\gamma)$.*

2.3 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". 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 $r > 0$. As in [35, 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 $\operatorname{gap}(\{a\}, B)$. 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 a convention, we define the summation and product over an empty index set as 0 and 1, respectively, which includes the cases of $i > j$ in $\sum_{k=i}^{j}$ and $\prod_{k=i}^{j}$.

3 Probabilistic ascent and non-convergence analysis

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? This section will address this question by introducing the concept of probabilistic ascent and establishing the non-convergence theory of probabilistic direct search. Before diving into the analysis, we first provide a numerical example in Subsection 3.1 to illustrate the failure of convergence of Algorithm 2.2 when the probabilistic descent condition does not hold. Then we introduce the concept of probabilistic ascent in Subsection 3.2. After that, we establish the non-convergence of probabilistic direct search via Markov's inequality in Subsection 3.4 and then via a Chernoff bound in Subsection 3.5. A weaker assumption will be proposed in Subsection 3.6 to broaden the non-convergence analysis.

3.1 Failure of global convergence: a numerical illustration

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 discussed in Corollary 2.1, with each $\mathfrak{D}_k$ being a set of m random vectors independently and uniformly distributed on the unit sphere.

For simplicity, we choose the objective function $f(x) = x^\mathsf{T} x$ with $x \in \mathbb{R}^2$. We set the forcing function $\rho(\alpha) = 10^{-3}\alpha^2$, the initial point $x_0 = (-10, 0)^\mathsf{T}$, the initial step size $\alpha_0 = 1$, the shrinking factor $\theta = 1/4$, and the expanding factor $\gamma = 3/2$. The polling sets are mutually independent, and each of them consists of $m = 2$ random vectors independently and uniformly distributed on the unit sphere in $\mathbb{R}^2$. 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 initial point, the pentagram represents the global minimizer, and each dot represents the best iterate (i.e., the one with the lowest function value) of the algorithm in a run. As we can see, 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 test, the results motivate us to conjecture that the algorithm fails to be globally convergent under this setting. We will confirm this conjecture in the subsequent analysis (see Corollary 3.1).

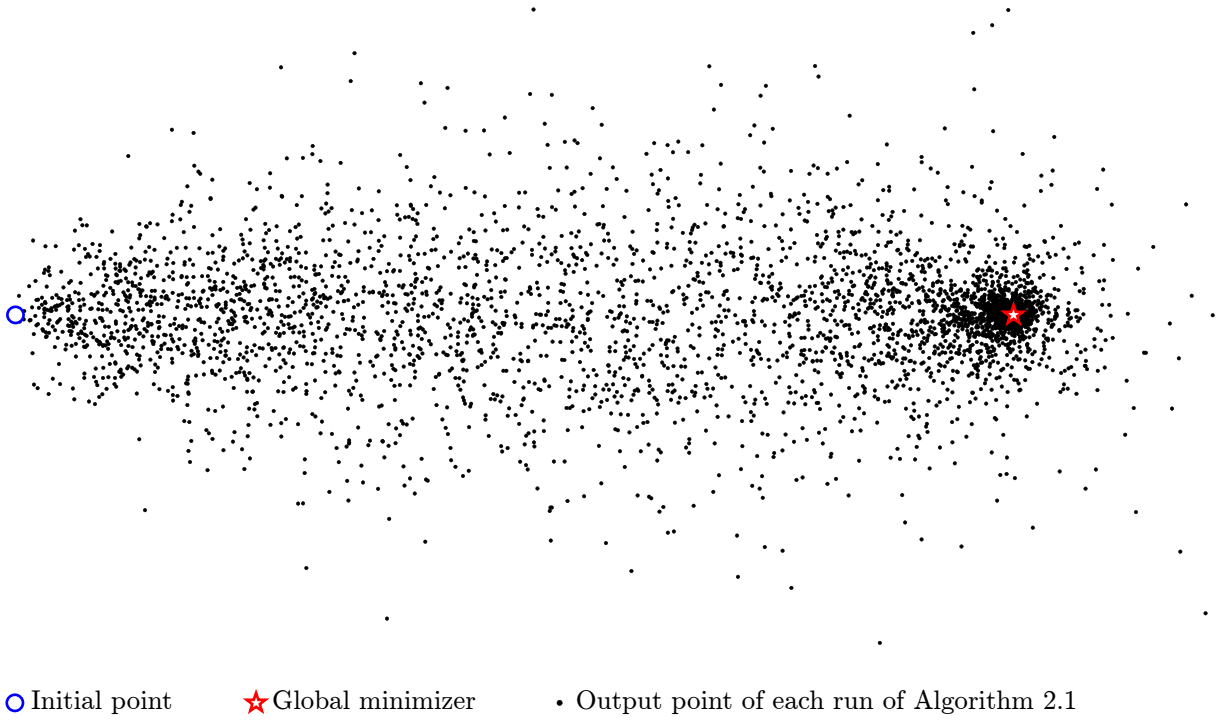

Figure 1: A test illustrating failure of convergence of Algorithm 2.2

3.2 Probabilistic ascent

Our non-convergence analysis relies on the concept of probabilistic ascent defined below. As mentioned in Remark 2.4, condition (3.1) in this definition should be understood in the almost sure sense.

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 it satisfies

$$\mathbb{P}\left(\mathrm{cm}\left(\mathfrak{D}_k, -G_k\right) \leq 0 \mid \mathcal{F}_{k-1}\right) \;\geq\; p\mathbb{1}\left(G_k \neq 0\right) \quad \text{for each } k \geq 0. \tag{3.1}$$

Proposition 3.1 shows that the sequence $\{\mathfrak{D}_k\}$ specified in Corollary 2.1 is a sequence of p-probabilistic ascent sets with $p = 2^{-m}$. The proof is given in Appendix B.

Proposition 3.1. *Consider Algorithm 2.2 with f being continuously differentiable on $\mathbb{R}^n$. Let $\{\mathfrak{D}_k\}$ be mutually independent, and each $\mathfrak{D}_k$ be a set of $m \geq 1$ independent random vectors uniformly distributed on the unit sphere in $\mathbb{R}^n$. Then $\{\mathfrak{D}_k\}$ is a sequence of p-probabilistic ascent sets with $p = 2^{-m}$.*

Remark 3.1. *It may be tempting to define p-probabilistic ascent as*

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

If we adopted this definition instead of Definition 3.1, then all the results requiring p-probabilistic ascent in this paper would still hold, since (3.2) is stronger than (3.1). However, in case one defines $\mathrm{cm}(\,\cdot\,, 0)$ to be positive (e.g., $\mathrm{cm}(\,\cdot\,, 0) = 1$ as in [17]), condition (3.2) with $p > 0$ will actually enforce

$$\mathbb{P}(G_k = 0) \;=\; 0 \quad \textit{for each } k \geq 0, \tag{3.3}$$

meaning that the algorithm almost never steps on a stationary point, which is a restriction that we do not want to impose. To see why (3.2) implies (3.3) when $p > 0$ and $\mathrm{cm}(\,\cdot\,, 0) > 0$, let us assume $\mathbb{P}(G_k = 0) > 0$. Then (3.2) and Lemma A.3 will lead to the contradiction that

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

Complementing Remark 3.1, Example 3.1 illustrates the difference between condition (3.1) in Definition 3.1 and condition (3.2). It also serves as an example to show that (3.3) is undesirable to impose when analyzing randomized algorithms like Algorithm 2.2, even though it is not uncommon to assume that algorithms never step on a stationary point in the deterministic case (e.g. [28, Section 1]).

Example 3.1. *Let $n = 1$, $f(x) = x^2$, and $x_0 = \alpha_0 = 1$. 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 values ± 1, each with probability $1/2$.*

By Proposition 3.1, $\{\mathfrak{D}_k\}$ is a sequence of $1/2$-probabilistic ascent sets as defined in Definition 3.1. However, whether $\{\mathfrak{D}_k\}$ satisfies condition (3.2) with $p = 1/2$ depends on the definition of $\mathrm{cm}(\cdot, 0)$. Suppose that we define $\mathrm{cm}(\cdot, 0) = 1$ following [17]. Then, as was pointed out in Remark 3.1, condition (3.2) with $p > 0$ necessitates (3.3), but we can check that

$$\mathbb{P}(G_1 = 0) = 1/2,$$

violating (3.3) for $k = 1$. Consequently, (3.2) cannot hold for any $p > 0$ if $\mathrm{cm}(\cdot, 0) = 1$.

In Example 3.1, condition (3.1) holds no matter how we define $\mathrm{cm}(\cdot, 0)$. Proposition 3.2 shows that such a condition is indeed always independent of $\mathrm{cm}(\cdot, 0)$. This proposition can be obtained by applying Lemma A.2 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 3.2. *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) $\mathbb{P}(\mathrm{cm}(\mathfrak{D}_k, -G_k) \leq 0 \mid \mathcal{F}_{k-1}) \geq p \mathbb{1}(G_k \neq 0)$.

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

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

Remark 3.2. *Neither item (b) nor (c) in Proposition 3.2 relies on the definition of $\mathrm{cm}(\cdot, 0)$. Therefore, condition (3.1) 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 3.3. *Assuming $\mathbb{P}(G_k \neq 0) > 0$, by Proposition 3.2 and item (b) of Lemma A.4 (see also Remark A.1), we can rewrite the inequality in Definition 3.1 as*

$$\mathbb{P}_k(\mathrm{cm}\,(\mathfrak{D}_k, -G_k) \leq 0 \mid \mathcal{F}_{k-1}) \;\geq\; p \quad (\mathbb{P}_k\text{-a.s.}), \tag{3.4}$$

where $\mathbb{P}_k$ is the probability measure defined by $\mathbb{P}_k(E) = \mathbb{P}(E \mid G_k \neq 0)$ for all $E \in \mathcal{F}$, and $\mathbb{P}_k(\cdot \mid \mathcal{F}_{k-1})$ is the corresponding conditional probability with respect to $\mathcal{F}_{k-1}$. Inequality (3.4) leads to the following interpretation of probabilistic ascent in Definition 3.1: conditioned on $G_k \neq 0$, the probability of $\mathrm{cm}(\mathfrak{D}_k, -G_k) \leq 0$ is at least p regardless of $\mathcal{F}_{k-1}$. We choose not to use (3.4) in Definition 3.1 to avoid any assumption about $\mathbb{P}(G_k \neq 0)$.

Before ending this section, we refer interested readers to Appendix D, which discusses an alternative definition of probabilistic descent (see Definition 2.2) using a condition similar to (3.1).

3.3 Key ingredients of our analysis

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

$$Y_k = \mathbb{1}\left(\min_{\mathfrak{d}\in\mathfrak{D}_k} \mathfrak{d}^\top G_k < 0\right) \quad \text{for each } k \geq 0. \tag{3.5}$$

For each $k \geq 0$, we define

$$U_k = \prod_{\ell=0}^{k-1} \gamma^{Y_\ell}\theta^{1-Y_\ell}, \tag{3.6}$$

$$\overline{Y}_k = \frac{1}{k}\sum_{\ell=0}^{k-1} Y_\ell, \tag{3.7}$$

$$E_k = \bigcap_{\ell=0}^{k-1}\{Y_\ell = 0\}, \tag{3.8}$$

with the convention that

$$U_0 = 1, \quad \overline{Y}_0 = 0, \quad \text{and} \quad E_0 = \Omega. \tag{3.9}$$

Note that $\{E_k\}$ is a nonincreasing sequence of events. In addition, since $0 < \theta < 1 \leq \gamma$, we have

$$E_k = \bigcap_{\ell=0}^{k-1}\{U_\ell = \theta^\ell\}. \tag{3.10}$$

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

Assuming the convexity of f, Lemma 3.1 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. Note that a differentiable convex function is continuously differentiable [34, Theorem 25.5]. Hence, we can still use Definition 3.1 of probabilistic ascent if we assume f to be differentiable and convex.

Lemma 3.1. *Consider Algorithm 2.2 with f being differentiable and convex on $\mathbb{R}^n$. Then*

$$\sup_{k\geq 0}\|X_k - x_0\| \leq \alpha_0 \sum_{k=0}^{\infty} Y_k U_k \leq \alpha_0 \sum_{k=0}^{\infty} U_k. \tag{3.11}$$

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

$$\|X_{k+1} - X_k\| \leq Y_k A_k. \tag{3.12}$$

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 (3.12); when $Y_k = 1$,

inequality (3.12) holds because of our blanket assumption that $\mathfrak{D}_k$ contains only unit vectors. Following a similar logic, we have

$$A_{k+1} \;\leq\; \gamma^{Y_k}\theta^{1-Y_k}A_k, \tag{3.13}$$

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 (3.6) of U_k, we use (3.13) recursively and obtain

$$A_k \;\leq\; \alpha_0 \prod_{\ell=0}^{k-1} \gamma^{Y_\ell}\theta^{1-Y_\ell} \;=\; \alpha_0 U_k. \tag{3.14}$$

Since $X_0 = x_0$, by (3.12) and (3.14), we have

$$\|X_k - x_0\| \;\leq\; \sum_{\ell=0}^{k-1} \|X_{\ell+1} - X_\ell\| \;\leq\; \sum_{\ell=0}^{k-1} Y_\ell A_\ell \;\leq\; \alpha_0 \sum_{\ell=0}^{k-1} Y_\ell U_\ell \;\leq\; \alpha_0 \sum_{\ell=0}^{k-1} U_\ell, \tag{3.15}$$

where the last inequality is because $Y_\ell \leq 1$. Finally, we get (3.11) by taking the supremum over $k \geq 0$ in (3.15). □

Remark 3.4. *Lemma 3.1 plays a crucial role in our non-convergence analysis. Roughly speaking, the main idea of our analysis is to show that*

$$\mathbb{P}\left(\sup_{k\geq 0} \|X_k - x_0\| < \zeta\right) \;>\; 0 \tag{3.16}$$

for some $\zeta > 0$, so that $\{X_k\}$ is bounded away from $\mathcal{S}(f)$ with positive probability as long as x_0 is sufficiently far away from $\mathcal{S}(f)$, namely,

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

Lemma 3.1 reduces the work of establishing (3.16) to studying $\sum_{k=0}^{\infty} Y_k U_k$ or $\sum_{k=0}^{\infty} U_k$. Our main results Theorems 3.1, 3.2, 3.3, and 3.4 all follow this idea, directly or indirectly.

The sequence $\{Y_k\}$ also provides us with an equivalent definition of probabilistic ascent as stated in Lemma 3.2. This equivalence is a simple consequence of Proposition 3.2.

Lemma 3.2. *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 (3.5) satisfies*

$$\mathbb{P}\left(Y_k = 0 \mid \mathcal{F}_{k-1}\right) \;\geq\; p \quad \textit{for each } k \geq 0. \tag{3.17}$$

Condition (3.17) is foundational to our analysis in Subsections 3.4 and 3.5. The behaviour of the sequences $\{Y_k\}$ and $\{U_k\}$ needed in our analysis follows from (3.17) without relying on the specifics of Algorithm 2.2.

3.4 Non-convergence analysis via Markov's inequality

In this subsection, we use Markov's inequality to conduct the non-convergence analysis as a preliminary step. The main idea of the following Theorem 3.1 is that under suitable assumptions, the expectation of the series of step sizes is finite.

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*

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

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

Proof. We prove that $\mathbb{P}(\operatorname{gap}(\{X_k\}, \mathcal{S}(f)) = 0) < 1$. Note that

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

where the last inclusion is due to Lemma 3.1. Therefore, it suffices to show that

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

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

$$\mathbb{E}\left(\sum_{k=0}^{\infty} U_k \right) \;\leq\; \beta. \tag{3.18}$$

With our assumption on p ensuring $0 < \gamma(1-p)+\theta p < 1$, we have $\beta = \sum_{k=0}^{\infty}[\gamma(1-p)+\theta p]^k$. Meanwhile, Tonelli's theorem [36, page 420] (also [16, 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 (3.18) can be reduced to establishing

$$\mathbb{E}(U_k) \leq [\gamma(1-p)+\theta p]^k \quad \text{for each } k \geq 0. \tag{3.19}$$

The proof of (3.19) is standard. For each $k \geq 0$, using the tower property of conditional expectation and the definition of $\{U_k\}$ in (3.6), 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 3.2,

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

Hence, we have

$$\mathbb{E}(U_{k+1}) \;\leq\; [\gamma(1-p)+\theta p]\mathbb{E}\left(U_k\right),$$

which implies (3.19) and concludes our proof. □

Remark 3.5. *Theorem 3.1 holds trivially if $\mathcal{S}(f) = \emptyset$, because $\operatorname{gap}(\,\cdot\,, \mathcal{S}(f)) = \infty$ in this case.*

3.5 Non-convergence analysis via a Chernoff bound

In the preceding subsection, we establish the non-convergence results under the requirements that $p > (\gamma-1)/(\gamma-\theta)$ and $\mathrm{gap}(x_0, \mathcal{S}(f))$ being large enough. In this subsection, we will weaken the requirement on p to $p > p_*$ with

$$p_* \;=\; 1 - p_0 \;=\; \frac{\log\gamma}{\log(\theta^{-1}\gamma)}, \tag{3.20}$$

where p_0 is defined in the convergence theorem (Theorem 2.1), and we will relax the requirement on x_0. Moreover, our non-convergence results will not only hold for the stationarity measure $\mathrm{gap}(\,\cdot\,, \mathcal{S}(f))$, but also extend to any lower semicontinuous function.

3.5.1 Lemmas and key observations

We first present a few propositions regarding the 0-1 process $\{Y_k\}$ defined by (3.5) and its associated sequences $\{U_k\}$, $\{\overline{Y}_k\}$, and $\{E_k\}$ defined by (3.6)–(3.8). We emphasize that these propositions are purely consequences of condition (3.17) and independent of the algorithm.

Lemma 3.3 establishes a Chernoff-type bound for $\{Y_k\}$, which is essentially a generalization of [17, Lemma 4.5]. Lemma 3.4 shows that condition (3.17) 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 B since the arguments are straightforward.

Lemma 3.3. *If $0 < q < p \leq 1$, then condition (3.17) implies that*

$$\mathbb{P}\left(1 - \overline{Y}_k \leq q \mid E_{k_0}\right) \;\leq\; \exp\left[-\frac{(p-q)^2}{2p}(k+k_0)\right] \quad \textit{for all } k \geq 0 \textit{ and } k_0 \geq 0. \tag{3.21}$$

Remark 3.6. *Noting the definition (3.8) of E_k, we can derive from condition (3.17) and the tower property of conditional expectations that*

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

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 $\tilde{\mathbb{P}}$. Then condition (3.17) implies that*

$$\tilde{\mathbb{P}}(\tilde{Y}_k = 0 \mid \tilde{\mathcal{F}}_{k-1}) \;\geq\; p \quad \textit{for each } k \geq 0. \tag{3.22}$$

Proposition 3.3 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 (3.17) 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 3.3. *If $p > p_*$, then condition (3.17) implies that there exists a function Υ satisfying*

$$\mathbb{P}\left(\sum_{k=k_0}^{\infty} U_k < \frac{\theta^{k_0}\zeta}{1-\theta} \,\middle|\, E_{k_0}\right) \;\geq\; \Upsilon(\zeta) \;>\; 0 \tag{3.23}$$

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 (3.23) 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 (3.9), this step is to find a positive value $\Upsilon(\zeta)$ for an arbitrarily given $\zeta > 1$ so that

$$\mathbb{P}(F) \;\geq\; \Upsilon(\zeta) \qquad \text{with} \qquad F \;=\; \left\{\sum_{k=0}^{\infty} U_k < \frac{\zeta}{1-\theta}\right\}. \tag{3.24}$$

To this end, we consider the event E_m defined in (3.8) and note that

$$\mathbb{P}(F) \;\geq\; \mathbb{P}(F \cap E_m) \;=\; \mathbb{P}(F \mid E_m)\,\mathbb{P}(E_m) \tag{3.25}$$

for each $m \geq 0$. In the sequel, we will bound $\mathbb{P}(F \mid E_m)$ and $\mathbb{P}(E_m)$ from below, and select an m in order that (3.25) yields a desired lower bound for $\mathbb{P}(F)$.

Due to the definition of F in (3.24) and the fact that $E_m = \bigcap_{k=0}^{m-1}\{U_k = \theta^k\}$ mentioned in (3.10), it holds that

$$\mathbb{P}(F \mid E_m) \;=\; \mathbb{P}\left(\sum_{k=0}^{m-1}\theta^k + \sum_{k=m}^{\infty} U_k < \frac{\zeta}{1-\theta} \,\middle|\, E_m\right) \;\geq\; \mathbb{P}\left(\sum_{k=m}^{\infty} U_k < \frac{\zeta-1}{1-\theta} \,\middle|\, E_m\right), \tag{3.26}$$

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

$$\left\{\sum_{k=m}^{\infty} U_k < \sum_{k=m}^{\infty}\left(\gamma^{1-q}\theta^q\right)^k\right\} \;\supseteq\; \bigcap_{k=m}^{\infty}\left\{U_k^{1/k} < \gamma^{1-q}\theta^q\right\} \;=\; \bigcap_{k=m}^{\infty}\left\{1-\overline{Y}_k > q\right\}, \tag{3.27}$$

where the equality is because $U_k^{1/k} = \gamma^{\overline{Y}_k}\theta^{1-\overline{Y}_k}$ by definitions (3.6)–(3.7) of U_k and $\overline{Y}_k$. Thus,

$$\begin{aligned}\mathbb{P}\left(\sum_{k=m}^{\infty} U_k < \sum_{k=m}^{\infty}\left(\gamma^{1-q}\theta^q\right)^k \,\middle|\, E_m\right) \;&\geq\; 1 - \mathbb{P}\left(\bigcup_{k=m}^{\infty}\left\{1-\overline{Y}_k \leq q\right\} \,\middle|\, E_m\right) \\ &\geq\; 1 - \sum_{k=m}^{\infty}\exp\left[-\frac{(p-q)^2}{2p}(k+m)\right],\end{aligned} \tag{3.28}$$

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

$$\sum_{k=m}^{\infty}\left(\gamma^{1-q}\theta^q\right)^k \;\leq\; \frac{\zeta-1}{1-\theta} \qquad \text{and} \qquad \sum_{k=m}^{\infty}\exp\left[-\frac{(p-q)^2}{2p}(k+m)\right] \;\leq\; \frac{1}{2}. \tag{3.29}$$

Such an m 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 $0 \le p_* < q < p$). The first inequality in (3.29) ensures that the right-hand side of (3.26) is no less than the left-hand side of (3.28), and the second inequality in (3.29) guarantees that the right-hand side of (3.28) is at least $1/2$. Therefore, we can join (3.26) and (3.28) to obtain

$$\mathbb{P}(F \mid E_m) \;\ge\; \frac{1}{2}.$$

Meanwhile, we have $\mathbb{P}(E_m) \ge p^m$ by Remark 3.6. Hence, inequality (3.25) implies (3.24) with

$$\Upsilon(\zeta) \;=\; \frac{p^m}{2}.$$

Given p, θ, and γ, the integer m is fully determined by ζ and so is $\Upsilon(\zeta)$, defining a function Υ that is sufficient for the first step of the proof.

Step 2. Now, we prove that the function Υ found in the last step satisfies (3.23) for all $\zeta > 1$ and $k_0 \ge 0$. Fix an arbitrary $k_0 \ge 0$. Define $\tilde{\mathbb{P}}$, $\{\tilde{\mathcal{F}}_k\}$, and $\{\tilde{Y}_k\}$ as in Lemma 3.4. According to this lemma, condition (3.17) implies condition (3.22), which has exactly the same form as (3.17), 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 (3.24), we can verify that Υ fulfills

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

for all $\zeta > 1$, where $\tilde{U}_k = \prod_{\ell=0}^{k-1} \gamma^{\tilde{Y}_\ell} \theta^{1-\tilde{Y}_\ell}$ for each $k \ge 0$. We will show that (3.30) ensures (3.23). The definitions of $\{\tilde{Y}_k\}$, $\{U_k\}$, and E_{k_0} (see Lemma 3.4, (3.6), and (3.8), respectively) imply that

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

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

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

for all $\zeta > 1$, which matches (3.23) as desired. This finishes our proof. □

Remark 3.7. *Given an integer $k_0 \ge 0$, condition (3.17) with $p > p_*$ indeed ensures the following 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 3.3, while the reverse implication holds because $\sum_{k=k_0}^{\infty} U_k \ge \sum_{k=k_0}^{\infty} \theta^k = \theta^{k_0}/(1-\theta)$.

Proposition 3.3 leads to Proposition 3.4, a crucial observation on the cumulative distribution function of $\sum_{k=0}^{\infty} Y_k U_k$. When $\{Y_k\}$ fulfills condition (3.17) 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 3.4. *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 (3.17) implies that there exists a constant $C > 0$ such that

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

Proof. Given a $\zeta \in (0, 1)$, define

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

Then $m \geq 0$. Recalling that $E_m = \bigcap_{k=0}^{m-1} \{Y_k = 0\}$ as defined in (3.8), we have

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

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

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

where $\Upsilon(2)$ in the last step comes from Proposition 3.3 and p^m comes from Remark 3.6. Therefore,

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

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

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

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

3.5.2 Qualitative and quantitative non-convergence results

This subsection presents our main results on the non-convergence of Algorithm 2.2. Under a probabilistic ascent assumption on the polling direction sets, we characterize the non-convergence 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. Moreover, we show the tightness of our probabilistic ascent assumption by Example 3.2.

It is worth mentioning that we will use a lower semicontinuous function μ to measure the distance of a given point to optimality, with a point $x \in \mathbb{R}^n$ being optimal if and only if $\mu(x) = 0$. 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 2.1 stays away from the optimal set with positive probability under a probabilistic ascent assumption, provided that the algorithm is initialized at a non-optimal 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*

$$\mathbb{P}\left(\inf_{k\geq 0} \mu(X_k) > 0\right) \; > \; 0 \tag{3.36}$$

for any function $\mu : \mathbb{R}^n \to (-\infty, \infty]$ that is lower semicontinuous, provided that $\mu(x_0) > 0$. In particular, the conclusion holds if μ is $f(\cdot) - \inf f$, $\operatorname{gap}(\cdot, \mathcal{S}(f))$, or $\|\nabla f(\cdot)\|$.

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,

$$\left\{\inf_{k\geq 0} \mu(X_k) > 0\right\} \; \supseteq \; \Big\{\{X_k\} \subseteq \{x : \mu(x) > \varepsilon\}\Big\} \; \supseteq \; \Big\{\{X_k\} \subseteq \mathcal{B}(x_0, \delta)\Big\}. \tag{3.37}$$

Meanwhile, Lemma 3.1 implies that

$$\Big\{\{X_k\} \subseteq \mathcal{B}(x_0, \delta)\Big\} \; \supseteq \; \left\{\sup_{k\geq 0} \|X_k - x_0\| < \delta\right\} \; \supseteq \; \left\{\sum_{k=0}^{\infty} Y_k U_k < \delta/\alpha_0\right\}. \tag{3.38}$$

The last event in (3.38) has a positive probability by Proposition 3.4, because $\{Y_k\}$ satisfies condition (3.17) according to Lemma 3.2. Therefore, (3.37) and (3.38) yield (3.36). □

Remark 3.8. *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, while the one in Theorem 3.1 can only be $\operatorname{gap}(\cdot, \mathcal{S}(f))$. Third, even when $\mu(x) = \operatorname{gap}(x, \mathcal{S}(f))$, the condition $\mu(x_0) > 0$ in Theorem 3.2 is weaker than $\operatorname{gap}(x_0, \mathcal{S}(f)) > \alpha_0/[1 - \gamma + p(\gamma - \theta)]$ 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 initial point is non-optimal.

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*

$$\Psi(\zeta) \;=\; \mathbb{P}\left(\inf_{k\geq 0} \mu(X_k) \;\geq\; (1-\zeta)\mu(x_0)\right) \tag{3.39}$$

satisfies

$$\Psi(\zeta) \;\geq\; C\,\zeta^{\frac{\log p}{\log \theta}} \quad \textit{for} \;\; \zeta \in (0, \overline{\zeta}). \tag{3.40}$$

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

$$|\mu(x) - \mu(x_0)| \;\leq\; L\|x - x_0\| \quad \text{for all } x \in \mathcal{B}(x_0, \delta). \tag{3.41}$$

For all $\zeta \in (0,\, L\delta)$, combining (3.41) with Lemma 3.1 renders

$$\left\{\inf_{k\geq 0} \mu(X_k) \;\geq\; \mu(x_0) - \zeta\right\} \;\supseteq\; \Big\{\{X_k\} \subseteq \mathcal{B}(x_0, \zeta/L)\Big\} \;\supseteq\; \left\{\sum_{k=0}^{\infty} Y_k U_k < \zeta/(L\alpha_0)\right\}.$$

Consequently, the definition of Φ in (3.32) and that of Ψ in (3.39) yield

$$\Psi(\zeta) \;\geq\; \Phi(\zeta/(L\alpha_0)).$$

Thus, Proposition 3.4 implies the desired lower bound for $\Psi(\zeta)$. □

Recall Corollary 2.1, which states that Algorithm 2.2 will converge with probability 1 if $m > \log_2(1 - \log\theta/\log\gamma)$. The following corollary shows 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 mutually independent, and each $\mathfrak{D}_k$ be a set of $m \geq 1$ independent random vectors uniformly distributed on the unit sphere in $\mathbb{R}^n$. If $\gamma = 1$ or*

$$m \;<\; \log_2\left(1 - \frac{\log\theta}{\log\gamma}\right), \tag{3.42}$$

then (3.36) 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 (3.40) 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 (3.20) if $\gamma = 1$ or m satisfies (3.42). □

Remark 3.9. *Comparing Corollaries 2.1 and 3.1, we observe that their requirements on the algorithmic parameters θ, γ, and m are nearly the complements of each other. The only gap is the marginal case with $m = \log_2(1 - \log\theta/\log\gamma)$, which is not a concern unless $\log_2(1 - \log\theta/\log\gamma)$ is an integer.*

Note that Theorem 3.2 requires $\{\mathfrak{D}_k\}$ to be a sequence of p-probabilistic ascent sets with $p > p_*$. We use Example 3.2 to show that such a requirement cannot be relaxed to $p \geq p_*$. In this example, $\{\mathfrak{D}_k\}$ is a sequence of p-probabilistic ascent sets with $p = p_*$, but Algorithm 2.2 converges with probability 1. Note that this 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., any coordinate vector). Then we set $\mathfrak{D}_k = \{\xi_k \mathfrak{d}_k\}$, where ξ_k is a random variable that is independent of $\mathcal{F}_{k-1}$ and equals 1 and -1 with probability p_* and $1-p_*$, respectively. 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 3.2. Meanwhile, one can check that $\{\mathfrak{D}_k\}$ is a sequence of p_0-probabilistic 1-descent sets (note that $p_0 = 1 - p_*$), implying that $\mathbb{P}\left(\liminf_k \|G_k\| = 0\right) = 1$ according to [17] (see also Theorem 2.1).

Remark 3.10. *Consider Algorithm 2.2 with $\gamma = \theta^{-1}$, which renders $p_* = 1/2$. Then Example 3.1 is indeed a one-dimensional special case of Example 3.2.*

3.5.3 Numerical verification of the quantitative non-convergence result

In this subsection, we demonstrate the quantitative non-convergence result in Theorem 3.3 numerically. As an example, we will focus on the case with $\mu(x) = f(x) - \inf f$, which reduces the function Ψ defined in (3.39) to

$$\Psi(\zeta) \;=\; \mathbb{P}\left(\inf_{k\geq 0} f(X_k) \;\geq\; f(x_0) - \zeta\right).$$

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.

The experiment is set up in the same way as in Subsection 3.1 except for the algorithmic parameters θ, γ, and m. To ensure the representativeness of the results, we randomly sample five values of the triple (θ, γ, m) as follows.

(a) Sample p_* and θ uniformly from the intervals $(0, 0.45)$ and $(0.25, 0.75)$, respectively.

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

This sampling scheme ensures that inequality (3.42) holds. Hence, Φ satisfies inequality (3.40) 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, each of which is terminated when the step size drops below the machine epsilon or the number of function evaluations reaches 10^3. The best (lowest) function value found in each run is denoted by f_{best}. Then we define

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

which is our estimation of $\Psi(\zeta)$.

Figure 2 plots $\log_{10}[\hat{\Psi}(\zeta)]/(\log p/\log\theta)$ against $\log_{10}\zeta$, with ζ varying between 10^{-3} and 10^{-1}. Each curve 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.

Across all the samples, the curves are almost parallel to the reference line, which is consistent with the rate in Theorem 3.3. Indeed, the almost perfect parallelism motivates us to conjecture that the rate in the theorem is tight, which is an interesting topic for future research.

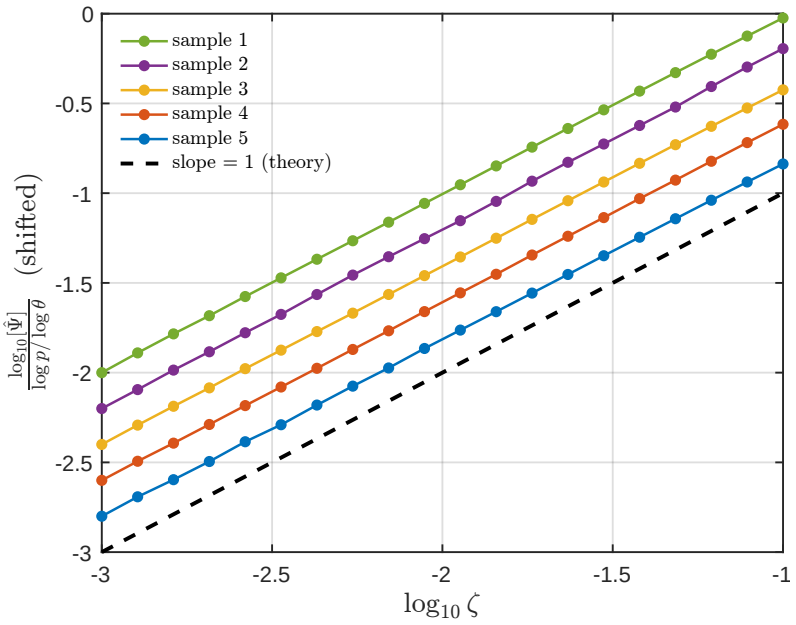

Figure 2: Curves of $\log_{10}[\hat{\Psi}(\zeta)]/(\log p/\log\theta)$ versus $\log_{10}\zeta$ for five random samples of (θ, γ, m). The curves are vertically shifted for clarity. The dashed line is a reference line with slope 1.

3.6 Non-convergence under a weaker assumption

Example 3.2 shows that we cannot weaken our assumption in Theorem 3.2 by replacing $p > p_*$ with $p \geq p_*$. However, this subsection will show that it is indeed possible to relax the definition of probabilistic ascent to obtain a weaker assumption 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

$$\mathbb{P}\left(\liminf_{k\to\infty}\left\{\mathbb{P}\left(\mathrm{cm}\left(\mathfrak{D}_k, -G_k\right) \leq 0 \mid \mathcal{F}_{k-1}\right) \geq p\mathbb{1}(G_k \neq 0)\right\}\right) \;>\; 0. \tag{3.43}$$

According to Proposition 3.2, condition (3.43) holds if and only if the sequence $\{Y_k\}$ defined in (3.5) satisfies

$$\mathbb{P}\Big(\liminf_{k\to\infty}\left\{\mathbb{P}\left(Y_k = 0 \mid \mathcal{F}_{k-1}\right) \geq p\right\}\Big) \;>\; 0. \tag{3.44}$$

Remark 3.11. *Condition* (3.44) *means that the event*

$$\{\mathbb{P}(Y_k = 0 \mid \mathcal{F}_{k-1}) \geq p \text{ for all sufficiently large } k\} \tag{3.45}$$

occurs with positive probability. This is weaker than

$$\mathbb{P}\left(\bigcap_{k=0}^{\infty}\left\{\mathbb{P}\left(Y_k = 0 \mid \mathcal{F}_{k-1}\right) \geq p\right\}\right) \;>\; 0, \tag{3.46}$$

which means that the event $\{\mathbb{P}(Y_k = 0 \mid \mathcal{F}_{k-1}) \geq p \text{ for each } k\}$ *has a positive probability. Condition* (3.44) *is also weaker than*

$$\sum_{k=0}^{\infty}\mathbb{P}\left(\left\{\mathbb{P}\left(Y_k = 0 \mid \mathcal{F}_{k-1}\right) < p\right\}\right) \;<\; \infty, \tag{3.47}$$

since (3.47) *implies that the event* (3.45) *occurs a.s. by the Borel–Cantelli Lemma [16, Theorem 2.3.1].*

Remark 3.12. *As stated in Lemma 3.2,* $\{\mathfrak{D}_k\}$ *is a sequence of p-probabilistic ascent sets if and only if the sequence* $\{Y_k\}$ *satisfies*

$$\mathbb{P}\left(Y_k = 0 \mid \mathcal{F}_{k-1}\right) \;\geq\; p \quad \text{for each } k \geq 0, \tag{3.48}$$

which is stronger than condition (3.44)*. Therefore, condition* (3.43) *can be regarded as a relaxation of p-probabilistic ascent defined in Definition 3.1. In addition, condition* (3.48) *implies both* (3.46) *and* (3.47)*, either of which in turn implies* (3.44) *as discussed in Remark 3.11.*

Before we show the non-convergence result under assumption (3.43), we need to introduce Lemma 3.6, which will be proved based on Lemma 3.5, a strong law of large numbers for martingales.

Lemma 3.5 ([10]). *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 (3.44) implies that*

$$\mathbb{P}\left(\sum_{k=0}^{\infty} U_k < \infty\right) > 0. \tag{3.49}$$

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 in (3.6) and $\overline{Y}_k$ in (3.7), we have

$$\log\left(U_k^{1/k}\right) = \log\left(\gamma^{\overline{Y}_k}\theta^{1-\overline{Y}_k}\right) = \log\theta + \overline{Y}_k \log(\theta^{-1}\gamma) = [(p_* - 1) + \overline{Y}_k]\log(\theta^{-1}\gamma), \tag{3.50}$$

where the last step uses the fact that $p_* = (\log\gamma)/\log(\theta^{-1}\gamma)$. Since $\log(\theta^{-1}\gamma) > 0$, equality (3.50) indicates that

$$\left\{\limsup_{k\to\infty} \overline{Y}_k < 1 - p_*\right\} \subseteq \left\{\sum_{k=0}^{\infty} U_k < \infty\right\}.$$

Therefore, by our assumption that $p > p_*$, inequality (3.49) can be established by proving

$$\mathbb{P}\left(\limsup_{k\to\infty} \overline{Y}_k \leq 1 - p\right) > 0. \tag{3.51}$$

To this end, let us define

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

Then $\mathbb{E}(Y_k + P_k - 1 \mid \mathcal{F}_{k-1}) = 0$ for each $k \geq 0$, implying that $\left\{\sum_{\ell=0}^{k-1}(Y_\ell + P_\ell - 1)\right\}$ is a martingale with respect to $\{\mathcal{F}_k\}$. 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 we define $\overline{P}_k = k^{-1}\sum_{\ell=0}^{k-1} P_\ell$. Hence, we have

$$\limsup_{k\to\infty} \overline{Y}_k = 1 - \liminf_{k\to\infty} \overline{P}_k \quad \text{a.s.}$$

Consequently,

$$\begin{aligned}
\mathbb{P}\left(\limsup_{k\to\infty} \overline{Y}_k \leq 1 - p\right) &= \mathbb{P}\left(\liminf_{k\to\infty} \overline{P}_k \geq p\right) \\
&\geq \mathbb{P}\left(\liminf_{k\to\infty} P_k \geq p\right) \\
&\geq \mathbb{P}\left(\liminf_{k\to\infty} \{P_k \geq p\}\right).
\end{aligned} \tag{3.52}$$

The two inequalities in (3.52) can be verified by noticing that

$$\liminf_{k\to\infty} \overline{P}_k \;\geq\; \liminf_{k\to\infty} P_k \quad \text{and} \quad \left\{\liminf_{k\to\infty} P_k \geq p\right\} \;\supseteq\; \liminf_{k\to\infty} \{P_k \geq p\}\,.$$

Finally, the last probability in (3.52) is positive by condition (3.44). The proof is complete. □

Remark 3.13. *In comparison with Lemma 3.6, condition* (3.17) *with* $p > p_*$ *implies*

$$\mathbb{P}\left(\sum_{k=0}^{\infty} U_k < \infty\right) \;=\; 1. \tag{3.53}$$

The proof is similar to that of Lemma 3.6. The major difference is that (3.17) *directly leads to*

$$\mathbb{P}\left(\liminf_{k\to\infty}\{P_k \geq p\}\right) = 1. \tag{3.54}$$

Combining (3.54) *and* (3.52)*, we see that the probability in* (3.51) *equals 1, implying* (3.53)*.*

Now, we are ready to present the non-convergence result under the weaker assumption (3.43). Its proof is similar to that of Theorem 3.2 with the help of Lemma 3.6.

Theorem 3.4. *Consider Algorithm 2.2 with* f *being differentiable and convex on* $\mathbb{R}^n$*. If* $\{\mathfrak{D}_k\}$ *satisfies* (3.43) *with* $p > p_*$*, then there exists a positive constant* ζ *such that*

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

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

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

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

Then we have

$$\mathbb{P}\left(\{X_k\} \subseteq \mathcal{B}(x_0, \zeta)\right) \;\geq\; \mathbb{P}\left(\sup_{k\geq 0} \|X_k - x_0\| < \zeta\right) \;\geq\; \mathbb{P}\left(\sum_{k=0}^{\infty} U_k < \frac{\zeta}{\alpha_0}\right) \;>\; 0,$$

where the second inequality uses Lemma 3.1. Therefore, when $\operatorname{gap}(x_0, \mathcal{S}(f)) > \zeta$, we have

$$\mathbb{P}\left(\operatorname{gap}(\{X_k\}, \mathcal{S}(f)) > 0\right) \;\geq\; \mathbb{P}\left(\{X_k\} \subseteq \mathcal{B}(x_0, \zeta)\right) \;>\; 0. \qquad \square$$

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 results in Theorems 3.1–3.3 still hold if we generalize Definition 3.1 of probabilistic ascent to Definition 4.1 as follows. Theorem 3.4 can be similarly extended.

Definition 4.1. 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

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

where $f^\circ(\,\cdot\,;\mathfrak{d})$ is the generalized directional derivative of f along the direction $\mathfrak{d}$, and $\partial_{\mathrm{C}} f(\cdot)$ is the Clarke subdifferential of f (see [11, Definitions 1.1 and 1.3] and [12, Section 2.1]).

Remark 4.1. *Condition (4.1) reduces to (3.1) when f is continuously differentiable on $\mathbb{R}^n$, since in that case we have $f^\circ(X_k;\mathfrak{d}) = \mathfrak{d}^{\mathsf{T}} G_k$ and $\partial_{\mathrm{C}} f(X_k) = \{G_k\}$ (see [11, Proposition 1.13]).*

For later usage, Proposition 4.1 verifies the measurability of two random variables, which correspond to $\mathbb{1}(G_k \neq 0)$ and $\mathbb{1}(\min_{\mathfrak{d}\in\mathfrak{D}_k} \mathfrak{d}^{\mathsf{T}} G_k < 0)$ in the smooth case.

Proposition 4.1. *Consider Algorithm 2.2 with f being locally Lipschitz continuous on $\mathbb{R}^n$. Then we have the following for each $k \geq 0$.*

(a) $\mathbb{1}(0 \notin \partial_{\mathrm{C}} f(X_k))$ *is $\mathcal{F}_{k-1}$-measurable.*

(b) $\mathbb{1}(\min_{\mathfrak{d}\in\mathfrak{D}_k} f^\circ(X_k;\mathfrak{d}) < 0)$ *is $\mathcal{F}_k$-measurable.*

Proof. Item (a) holds because $\{x \in \mathbb{R}^n : 0 \notin \partial_{\mathrm{C}} f(x)\}$ is open by [12, Proposition 2.1.5]. For item (b), it suffices to note that $\mathbb{1}(f^\circ(X_k;\mathfrak{d}) < 0)$ is $\mathcal{F}_k$-measurable for each $\mathfrak{d} \in \mathfrak{D}_k$, which is true since $(x, d) \mapsto f^\circ(x; d)$ is upper semicontinuous [12, Proposition 2.1.1 (b)] and hence Borel. □

Proposition 4.2 generalizes Proposition 3.1, namely the probabilistic ascent of the sequence $\{\mathfrak{D}_k\}$ specified in Corollary 2.1. The proof is given in Appendix B.

Proposition 4.2. *With probabilistic ascent defined in Definition 4.1, Proposition 3.1 still holds even if we remove the differentiability assumption about f.*

To generalize Lemma 3.2, namely the equivalence between probabilistic ascent and condition (3.17), we shift the definition of Y_k from (3.5) to

$$Y_k \;=\; \mathbb{1}\left(\min_{\mathfrak{d}\in\mathfrak{D}_k} f^\circ(X_k;\mathfrak{d}) < 0 \right). \tag{4.2}$$

Lemma 4.1. *With probabilistic ascent defined in Definition 4.1 and Y_k defined in (4.2), Lemma 3.2 still holds when f is locally Lipschitz continuous rather than continuously differentiable.*

Proof. We only need to prove the equivalence between (4.1) and

$$\mathbb{P}\Big(\min_{\mathfrak{d}\in\mathfrak{D}_k} f^\circ(X_k;\mathfrak{d}) \geq 0 \;\Big|\; \mathcal{F}_{k-1} \Big) \;\geq\; p \quad \text{for each } k \geq 0. \tag{4.3}$$

Recalling that $f^\circ(X_k;\mathfrak{d}) = \max\{g^\mathsf{T}\mathfrak{d} : g \in \partial_{\mathrm{C}} f(X_k)\}$ [11, Proposition 1.4], we have

$$\{0 \in \partial_{\mathrm{C}} f(X_k)\} \;\subseteq\; \Big\{ \min_{\mathfrak{d}\in\mathfrak{D}_k} f^\circ(X_k;\mathfrak{d}) \geq 0 \Big\}.$$

Therefore, conditions (4.1) and (4.3) are equivalent according to Lemma A.2. □

Remark 4.2. *Since conditions* (4.1) *and* (4.3) *are equivalent, Definition 4.1 can be stated without the indicator function* $\mathbb{1}(0 \notin \partial_{\mathrm{C}} f(X_k))$. *However, we prefer the current form because it is consistent with Definition 3.1, where the* $\mathbb{1}(G_k \neq 0)$ *term is necessary according to Remark 3.1.*

Now, we can extend 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 even 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 3.3 and 3.4 still hold under the new settings. Lemma 3.1 remains true as it does not rely on the differentiability of f. Lemma 3.2 holds according to Lemma 4.1. Lemmas 3.3 and 3.4, Propositions 3.3 and 3.4 are valid because they only depend on Lemma 3.2 and the $\mathcal{F}_k$-measurability of Y_k, which is guaranteed by item (b) of Proposition 4.1. The proof is complete. □

Remark 4.3. *For Theorem 4.1, we can take the lower semicontinuous function* μ *in* (3.36) *to be* $f(\cdot) - \inf f$, $\operatorname{gap}(\,\cdot\,, \mathcal{S}(f))$, *and* $\operatorname{gap}(0, \partial_{\mathrm{C}} f(\cdot))$. *It is clear that* $f(\cdot) - \inf f$ *and* $\operatorname{gap}(\,\cdot\,, \mathcal{S}(f))$ *are lower semicontinuous. The lower semicontinuity of* $\operatorname{gap}(0, \partial_{\mathrm{C}} f(\cdot))$ *is also basic, but we provide a proof in Lemma A.5 for completeness.*

Theorem 3.1 also holds in the nonsmooth case since it is a weaker version of Theorem 3.2. We can also extend Theorem 3.4 to the nonsmooth case. Specifically, Theorem 3.4 holds without the differentiability assumption about f as long as we replace condition (3.43) with

$$\mathbb{P}\left(\liminf_{k\to\infty} \left\{ \mathbb{P}\Big(\min_{\mathfrak{d}\in\mathfrak{D}_k} f^\circ(X_k;\mathfrak{d}) \geq 0 \;\Big|\; \mathcal{F}_{k-1} \Big) \geq p\mathbb{1}\left(0 \notin \partial_{\mathrm{C}} f(X_k)\right) \right\} \right) \;>\; 0. \tag{4.4}$$

Similar to Lemma 4.1, condition (4.4) is equivalent to (3.44) with $\{Y_k\}$ defined by (4.2). This leads to Lemma 3.6 and then Theorem 3.4.

Hence, we can conclude that our non-convergence theory for probabilistic direct search is still valid in the nonsmooth case. Indeed, it is also possible to extend the convergence theory in [17] to the nonsmooth case similarly, but it is beyond the scope of this paper.

5 Conclusion

We establish the non-convergence theory for probabilistic direct search. The proof technique is mainly based on the analysis of the probability that the series of step sizes converges. Specifically, the series of step sizes in Algorithm 2.2 converges with positive probability if the sequence of polling direction sets is a sequence of p-probabilistic ascent sets with $p > \log\gamma/\log(\theta^{-1}\gamma)$, where θ and γ are shrinking and expanding factors of the step size, respectively. More importantly, in the typical case where we choose uniform distribution on the unit sphere in $\mathbb{R}^n$ as the distribution of polling directions, the series of step sizes converges with probability 1 if the number of polling directions in each iteration is strictly less than $\log_2(1 - \log\theta/\log\gamma)$, which is almost the counterpart of the convergence result. A weaker assumption is proposed to replace the probabilistic ascent assumption while the nonzero probability of the convergence of the series of step sizes is still guaranteed. The final part demonstrates the tightness of our non-convergence analysis by explaining the reason why our analysis techniques cannot cover the case of p-probabilistic ascent with $p = p_*$ instead of $p > p_*$ and providing a concrete example showing that probabilistic direct search converges when $\{\mathfrak{D}_k\}$ is a sequence of p_*-probabilistic ascent sets. Finally, we extend our results to the nonsmooth case.

References

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

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

[3] 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.

[4] 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.

[5] 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.

[6] J. Blanchet, C. Cartis, M. Menickelly, and K. Scheinberg. Convergence rate analysis of a stochastic trust region method via supermartingales. *INFORMS J. Optim.*, 1:92–119, 2019.

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

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

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

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

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

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

[13] 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.

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

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

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

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

[18] 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.

[19] 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.

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

[21] 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.

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

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

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

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

[26] 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.

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

[28] 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.

[29] 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.

[30] M. J. D. Powell. UOBYQA: unconstrained optimization by quadratic approximation. Technical Report DAMTP 2000/NA14, Department of Applied Mathematics and Theoretical Physics, Cambridge University, Cambridge, 2000.

[31] M. J. D. Powell. The NEWUOA software for unconstrained optimization without derivatives. Technical Report DAMTP 2004/NA05, Department of Applied Mathematics and Theoretical Physics, Cambridge University, Cambridge, 2004.

[32] 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.

[33] T. M. Ragonneau and Z. Zhang. PDFO: a cross-platform package for Powell's derivative-free optimization solvers. arXiv:2302.13246, 2023.

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

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

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

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

A Basic lemmas

Lemma A.1 is a straightforward consequence of [16, 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 elaborates on the equivalence among several probability inequalities. It is useful for interpreting the conditions in Definition 3.1 and 4.1.

Lemma A.2. *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:*

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

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 $E^{\mathrm{c}} \in \mathcal{G}$, Lemma A.1 yields $\mathbb{P}(F \cap E^{\mathrm{c}} \mid \mathcal{G}) = \mathbb{P}(F \mid \mathcal{G})\mathbb{1}(E^{\mathrm{c}}) \ \geq \ p\mathbb{1}(E^{\mathrm{c}})$.
(b) $\Rightarrow$ (c): Since $E \cup F = E \cup (F \cap E^{\mathrm{c}})$ and $E \in \mathcal{G}$, we have

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

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

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

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

Lemma A.3 presents a basic connection between the conditional probability with respect to a σ-algebra and that with respect to an event.

Lemma A.3. *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 $\mathbb{P}(E \mid \hat{F}) < p$. □

In the following, for 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 any 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 use $\mathbb{E}_F$ to denote the expectation under $\mathbb{P}_F$, and $\mathbb{E}_F(\,\cdot \mid \mathcal{G})$ to denote the corresponding conditional expectation with respect to a σ-algebra $\mathcal{G}$. It is well known that

$$\mathbb{E}(X \mathbb{1}(F)) \;=\; \mathbb{E}_F(X) \mathbb{P}(F) \tag{A.2}$$

for any random variable X (see, e.g., [20, Section 8.1]). Consequently,

$$\mathbb{E}(X \mathbb{1}(F)) \;=\; \mathbb{E}\left(\mathbb{E}_F(X \mid \mathcal{G}) \mathbb{1}(F)\right), \tag{A.3}$$

which can be obtained by multiplying both sides of the equality $\mathbb{E}_F(X) = \mathbb{E}_F(\mathbb{E}_F(X \mid \mathcal{G}))$ with $\mathbb{P}(F)$.

Lemma A.4. *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

$$\mathbb{E}(\mathbb{E}_F(X \mid \mathcal{G})\, \mathbb{P}(F \mid \mathcal{G})\, \mathbb{1}(E)) \;=\; \mathbb{E}(X \mathbb{1}(F)\, \mathbb{1}(E)) \tag{A.4}$$

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.4) 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.3) 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.4) holds.

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

$$Z\mathbb{P}(F \mid \mathcal{G}) \ \geq \ p\mathbb{P}(F \mid \mathcal{G}) \quad (\mathbb{P}\text{-a.s.}) \qquad \Longleftrightarrow \qquad Z \ \geq \ p \quad (\mathbb{P}_F\text{-a.s.}). \tag{A.5}$$

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

$$\big\{Z\mathbb{P}(F \mid \mathcal{G}) \ \geq \ p\mathbb{P}(F \mid \mathcal{G})\big\} \ = \ \big\{Z \geq p \ \text{ or } \ \mathbb{P}(F \mid \mathcal{G}) = 0\big\} \ = \ \big\{W = 0\big\} \tag{A.6}$$

and that

$$\mathbb{P}_F(Z < p)\,\mathbb{P}(F) \ = \ \mathbb{P}\left(\{Z < p\} \cap F\right) \ = \ \mathbb{E}\big(\mathbb{1}(Z < p)\mathbb{P}(F \mid \mathcal{G})\big) \ = \ \mathbb{E}(W). \tag{A.7}$$

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.6), while the second comes from (A.7) and the fact that $\mathbb{P}(F) > 0$. Hence, (A.5) holds. The proof is complete. □

Remark A.1. *Item (a) of Lemma A.4 is a generalization of equality (A.2). In light of (a), item (b) shows that we can cancel out $\mathbb{P}(F \mid \mathcal{G})$ 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. When $X = \mathbb{1}(E)$ for an event E, item (b) reduces to*

$$\mathbb{P}(EF \mid \mathcal{G}) \ \geq \ p\mathbb{P}(F \mid \mathcal{G}) \quad (\mathbb{P}\text{-}a.s.) \qquad \Longleftrightarrow \qquad \mathbb{P}_F(E \mid \mathcal{G}) \ \geq \ p \quad (\mathbb{P}_F\text{-}a.s.). \tag{A.8}$$

Note that $\mathbb{P}(F \mid \mathcal{G}) = \mathbb{1}(F)$ if $F \in \mathcal{G}$, as is the case in Remark 3.3 and the proof of Lemma 3.4.

Lemma A.5. *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 [34, 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. □

B Proofs of Proposition 3.1, Lemma 3.3, Lemma 3.4, and Proposition 4.2

To prove Proposition 3.1, we need Lemma B.1 as follows, particularly its item (b).

Lemma B.1. *Let X and Y be random vectors. Consider a measurable function h with $\mathbb{E}(|h(X, Y)|) < \infty$, 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 [16, Theorem 2.1.12], which considers random variables rather than random vectors. We omit its proof, which is a straightforward extension of that for [16, 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

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

for all $E \in \mathcal{G}$. The right-hand side of (B.1) 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

$$\mathbb{E}(h(X,Y)\mathbb{1}(E)) \;=\; \mathbb{E}(H(Y)\mathbb{1}(E)). \tag{B.2}$$

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 $\hat{H}$ and obtain

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

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

$$\hat{H}(\hat{y}) \;=\; \mathbb{E}(h(X,y)z) \;=\; H(y)z. \tag{B.4}$$

Plugging (B.4) and the definitions of $\hat{h}$ into (B.3), we obtain $\mathbb{E}(h(X,Y)Z) = \mathbb{E}(H(Y)Z)$, which is (B.2). This completes the proof. □

Remark B.1. *Taking expectation on both sides of the equality in item (b) of Lemma B.1, we can recover item (a) by the tower property of conditional expectation. We also note that item (b) is a generalization of [16, Example 4.1.7] (see also [9, page 148]), where $\mathcal{G} = \sigma(Y)$.*

Now we are ready to prove Proposition 3.1.

Proof of Proposition 3.1. It suffices to prove that

$$\mathbb{P}\left(\{\mathrm{cm}\left(\mathfrak{D}_k, -G_k\right) \le 0\} \cap \{G_k \neq 0\} \mid \mathcal{F}_{k-1}\right) \;\ge\; 2^{-m}\mathbb{1}\left(G_k \neq 0\right). \tag{B.5}$$

Notice that the left-hand side of (B.5) can be rewritten as

$$\mathbb{E}\left(\mathbb{1}(\mathrm{cm}\left(\mathfrak{D}_k, -G_k\right) \le 0)\mathbb{1}\left(G_k \neq 0\right) \mid \mathcal{F}_{k-1}\right). \tag{B.6}$$

By item (b) of Lemma B.1, the conditional expectation (B.6) equals $\mathbb{E}\left(H(G_k) \mid \mathcal{F}_{k-1}\right)$ with

$$H(g) \;=\; \mathbb{E}(\mathbb{1}(\mathrm{cm}\left(\mathfrak{D}_k, -g\right) \le 0)\mathbb{1}\left(g \neq 0\right)) \;=\; \begin{cases} 2^{-m}, & \text{if } g \neq 0, \\ 0, & \text{if } g = 0, \end{cases}$$

where the last equality holds because $\mathfrak{D}_k$ consists of m independent random vectors uniformly distributed on the unit sphere. We then complete the proof by observing that

$$\mathbb{E}\left(H(G_k) \mid \mathcal{F}_{k-1}\right) \;=\; \mathbb{E}\left(2^{-m}\mathbb{1}(G_k \neq 0) \mid \mathcal{F}_{k-1}\right) \;=\; 2^{-m}\mathbb{1}\left(G_k \neq 0\right),$$

where the last equality is because G_k is $\mathcal{F}_{k-1}$-measurable. □

To prove Lemma 3.3, we first present Lemma B.2.

Lemma B.2. *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

$$(kq - k_0)\log(p/q) + (k - k_0)(q - p) \;\leq\; -\frac{(q-p)^2}{2p}(k + k_0). \tag{B.7}$$

Regard the left-hand side of (B.7) 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

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

Now we prove Lemma 3.3 using the moment method for deriving Chernoff bounds [24].

Proof of Lemma 3.3. The inequality in (3.21) holds trivially when $k = k_0$, because $1 - \overline{Y}_k = 1 > q$ when E_{k_0} happens, implying that the conditional probability in (3.21) is zero. 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

$$\mathbb{P}\left(1 - \overline{Y}_k \leq q \mid E_{k_0}\right) \;\leq\; \mathrm{e}^{t(kq - k_0)}\, \mathbb{E}\left(\prod_{\ell=k_0}^{k-1} \mathrm{e}^{-t(1-Y_\ell)} \,\middle|\, E_{k_0}\right), \tag{B.8}$$

and the other is

$$\mathbb{E}\left(\prod_{\ell=k_0}^{k-1} \mathrm{e}^{-t(1-Y_\ell)} \,\middle|\, E_{k_0}\right) \;\leq\; \exp[p(k - k_0)(\mathrm{e}^{-t} - 1)]. \tag{B.9}$$

Once inequalities (B.8) and (B.9) are proved, we will have

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

and then the proof will be completed by Lemma B.2. We now prove the two claims by standard techniques.

For (B.8), by definition (3.7) of $\overline{Y}_k$, definition (3.8) of E_{k_0}, and Markov's inequality, we have

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

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 (B.9), we use the tower property of conditional expectation to get

$$\mathbb{E}\left(\prod_{\ell=k_0}^{k-1} \mathrm{e}^{-t(1-Y_\ell)} \,\middle|\, \mathcal{F}_{k_0-1}\right) \;=\; \mathbb{E}\left(\mathbb{E}\left(\mathrm{e}^{-t(1-Y_{k-1})} \mid \mathcal{F}_{k-2}\right) \prod_{\ell=k_0}^{k-2} \mathrm{e}^{-t(1-Y_\ell)} \,\middle|\, \mathcal{F}_{k_0-1}\right), \tag{B.10}$$

where $\prod_{\ell=k_0}^{k-2} \mathrm{e}^{-t(1-Y_\ell)} = 1$ when $k = k_0 + 1$. By condition (3.17), we have

$$\mathbb{E}\left(\mathrm{e}^{-t(1-Y_{k-1})} \mid \mathcal{F}_{k-2}\right) \;\le\; p\mathrm{e}^{-t} + (1-p) \;\le\; \exp(p\mathrm{e}^{-t} - p), \tag{B.11}$$

where the last inequality is because $x + 1 \le \mathrm{e}^x$ for all x. By equality (B.10) and inequality (B.11), we have

$$\begin{aligned}\mathbb{E}\left(\prod_{\ell=k_0}^{k-1} \mathrm{e}^{-t(1-Y_\ell)} \,\middle|\, \mathcal{F}_{k_0-1}\right) \;&\le\; \exp[p(\mathrm{e}^{-t}-1)]\;\mathbb{E}\left(\prod_{\ell=k_0}^{k-2} \mathrm{e}^{-t(1-Y_\ell)} \,\middle|\, \mathcal{F}_{k_0-1}\right)\\ &\le\; \exp[p(k-k_0)(\mathrm{e}^{-t}-1)],\end{aligned} \tag{B.12}$$

where the second inequality follows from the recursive application of the first one. Since $\mathbb{P}(E_{k_0}) > 0$ by Remark 3.6, inequality (B.12) implies (B.9) by Lemma A.3. □

Lemma 3.4 is a straightforward consequence of Lemma A.4, or, more precisely, Remark A.1.

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

Now we prove Proposition 4.2. It is similar to the proof of Proposition 3.1.

Proof of Proposition 4.2. First, the function $h(\mathcal{D}, x) = \mathbb{1}(\min_{d\in\mathcal{D}} f^\circ(x;d) \ge 0)\mathbb{1}(0 \notin \partial_{\mathrm{C}} f(x))$ is Borel by the same argument as in the proof of Proposition 4.1. Then similar to the proof of Proposition 3.1, by item (b) of Lemma B.1 and item (a) of Proposition 4.1, we only need to show that

$$H(x) \;=\; \mathbb{E}\left(h(\mathfrak{D}_k, x)\right) \;\ge\; \begin{cases} 2^{-m}, & \text{if } 0 \notin \partial_{\mathrm{C}} f(x),\\ 0, & \text{if } 0 \in \partial_{\mathrm{C}} f(x).\end{cases}$$

It suffices to prove that when $0 \notin \partial_{\mathrm{C}} f(x)$, we have

$$\mathbb{P}\left(f^\circ(x;\mathfrak{d}) \ge 0\right) \;\ge\; \frac{1}{2},$$

where $\mathfrak{d}$ is uniformly distributed on the unit sphere in $\mathbb{R}^n$, which is true since $\{d : f^\circ(x;d) \ge 0\}$ contains a half-space $\{v : g^\mathsf{T} v \ge 0\}$ with g being any element in $\partial_{\mathrm{C}} f(x)$. □

C (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 C.1, the measurability can fail for certain implementations of Algorithm 2.2. For the concept of measurability, we refer to [16, Section 1.2].

Lemma C.1 establishes the measurability of the iterates for certain implementations of Algorithm 2.2, covering [17, Algorithm 2.1]. The proof is elementary, but it clarifies the role of the polling strategy in the measurability.

Lemma C.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,

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

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 the 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 (C.1). Consequently, A_{k+1} is $\mathcal{F}_k^{\mathfrak{D}}$-measurable by the recurrence relation (2.3) and the induction hypothesis. The induction finishes for complete polling.

The case of opportunistic polling can be handled similarly. In this case, equation (C.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 C.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 C.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* $\mathfrak{D}_k = \{\mathfrak{d}_k^1, \dots, \mathfrak{d}_k^m\}$ *with* $\mathfrak{d}_k^1, \dots, \mathfrak{d}_k^m$ *being random vectors.*

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

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

(d) *Use opportunistic polling.*

Since X_k *depends on* π_{k-1}*, we cannot guarantee its* $\mathcal{F}_{k-1}^{\mathfrak{D}}$*-measurability if* π_{k-1} *is not* $\mathcal{F}_{k-1}^{\mathfrak{D}}$*-measurable, or informally, if* π_{k-1} *contains randomness beyond* $\mathcal{F}_{k-1}^{\mathfrak{D}}$*. Similar to [15, Section 4], we can define* π_{k-1} *by ranking the directions in* $\mathfrak{D}_{k-1}$ *according to a stochastic oracle independent of the polling directions. Or we simply pick the sequences* $\{\pi_k\}$ *and* $\{\mathfrak{D}_k\}$ *independently. In these cases,* X_k *can be measurable with respect to* $\sigma(\mathfrak{D}_0, \pi_0, \dots, \mathfrak{D}_{k-1}, \pi_{k-1})$*, but not with respect to* $\mathcal{F}_{k-1}^{\mathfrak{D}}$.

D Discussions about the definition of probabilistic descent

Comparing Definition 2.2 of probabilistic descent with Definition 3.1 of probabilistic ascent, one may ask why the latter involves $\mathbb{1}(G_k \neq 0)$ whereas the former does not. To answer this question, we propose an alternative definition of probabilistic descent in Definition D.1, with $\mathbb{1}(G_k \neq 0)$ playing a role like in Definition 3.1.

Definition D.1 (Alternative definition of probabilistic descent)**.** Identical to Definition 2.2 except that we replace condition (2.4) with

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

Definition D.1 is equivalent to Definition 2.2 if $\operatorname{cm}(\,\cdot\,, 0) \geq \kappa$ (e.g., [17] defines $\operatorname{cm}(\,\cdot\,, 0) = 1$). Indeed, we have $\{G_k = 0\} \subseteq \{\operatorname{cm}(\mathfrak{D}_k, -G_k) \geq \kappa\}$ in this case, ensuring the equivalence by Lemma A.2.

Definition D.1 has the advantage that it is invariant no matter how we choose the value of $\operatorname{cm}(\,\cdot\,, 0)$, because the inequality in condition (D.1) is equivalent to

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

according to Lemma A.2. In contrast, Definition 2.2 does rely on this value, as can be illustrated by an example similar to Example 3.1. In case one defines $\operatorname{cm}(\,\cdot\,, 0) < \kappa$ (e.g., $\operatorname{cm}(\,\cdot\,, 0) = 0$ may be appealing for symmetry), Definition 2.2 will be more restrictive than Definition D.1 for the same reason explained in Remark 3.1. Nevertheless, this does not affect [17], which imposes $\operatorname{cm}(\,\cdot\,, 0) = 1$ as mentioned before.