Non-convergence Analysis of Probabilistic Direct Search

Cunxin Huang* Zaikun Zhang†

August 15, 2024

Abstract

Direct-search methods are a major class in derivative-free optimization. The combination of direct search and randomization techniques leads to an efficient offspring, 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 the assumption for convergence is not met? In this paper, we analyze the non-convergence of the algorithm when polling directions are probabilistically ascent. Its analysis is basically related to the discussion on a random series. We further show the tightness of our non-convergence analysis in two perspectives. Our non-convergence theory completes the analytical framework for the probabilistic direct search, guiding the selection of the searching set in practice.

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

1 Introduction

This paper focuses on the probabilistic direct search method for solving the unconstrained optimization problem

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

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

†Department of Applied Mathematics, The Hong Kong Polytechnic University, Hong Kong, China (zaikun.zhang@polyu.edu.hk).

where $f : \mathbb{R}^n \to \mathbb{R}$ is a differentiable function. Direct search is a derivative-free optimization (DFO) method, which solves (1.1) without using derivatives and updates iterates based on simple comparisons of function values at finite sample points [15].

In the deterministic case, direct search methods need to select a searching set consisting of at least $n + 1$ directions for each iteration, which limits its capability. To overcome this, Gratton et al. [11] propose a framework of direct search based on probabilistic descent, referred to as probabilistic direct search. They show that given shrinking factor θ and expanding factor γ, the algorithm enjoys global convergence if the sequence of searching sets is p_0-probabilistically κ-descent with some positive κ and

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

Specially, if we choose each searching set to be a collection of m independent random directions following uniform distribution on the unit sphere, which is the typical choice in practice [11], then a sufficient condition for convergence is

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

This result not only provides more choices of searching sets for direct search, but also guides the analysis of the probabilistic trust-region model [12].

A natural question arises: what will happen if $m \leq \log_2(1 - \log\theta/\log\gamma)$? Furthermore, we would like to ask: is p_0-probabilistically κ-descent assumption essential for the convergence of probabilistic direct search? These two questions are both theoretically interesting and practically meaningful. Theoretically, answering these questions will not only verify whether p_0-probabilistically κ-descent assumption is essential, but also help fill gaps between the existing analysis and the behavior of the algorithm when the assumption is not met. Practically, it provides more detailed guidance in parameter selection.

To answer these two questions, we establish the *non-convergence theory* of probabilistic direct search. We prove that the algorithm will not converge if the searching set is p-probabilistically ascent (Definition 3.1) with $p > 1 - p_0$ and the objective function is smooth and convex. In particular, for the above-mentioned typical case of searching sets, the algorithm will not converge if $m < \log_2(1 - \log\theta/\log\gamma)$.

Moreover, on the one hand, we show that both our analysis and assumptions are tight by investigating two special cases, respectively. On the other hand, the essential role of the p_0-probabilistically κ-descent assumption in the convergence analysis of probabilistic direct search is further confirmed by our non-convergence analysis, where we answer the question raised above that the algorithm will not converge if $m < \log_2(1 - \log\theta/\log\gamma)$.

From a broader perspective, p_0-probabilistically κ-descent assumption belongs to submartingale-like assumptions, which are widely used in the convergence analysis of randomized version (some called probabilistic models) of optimization methods such as trust region [3, 28], line search [5, 7], and cubic regularization [7].

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 main ideas of this paper. Subsequently, in Section 4, we first show that the probabilistic direct search will not converge if $m < \log_2(1 - \log\theta/\log\gamma)$ by our non-convergence analysis. Additionally, we demonstrate the tightness of our non-convergence results by investigating two interesting cases. 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. DFO is a field that in recent decades arouses great interest in both academic research and practical applications [2, 9, 16]. Within the existing body of literature, DFO methods are broadly classified into two primary categories: direct-search methods and model-based methods. Detailed discussion on direct-search methods can be found in [15], and notable examples of direct search include the Nelder-Mead simplex method [18], the MADS methods [1, 17], and BFO [19, 20]. Contrary to direct-search methods using simple comparisons of function values, model-based methods construct local models through sampling under a trust-region [8] or line-search [4] framework. A wealth of classical literature on model-based methods can be referred to, such as [4, 8, 21, 22, 23, 24], with some well-known methods and software in this category including PDFO [25]. Recently, randomization techniques are introduced to both two categories and we refer to [3, 6, 7, 11, 12, 13].

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

2.1 Direct search based on sufficient decrease

To solve problem (1.1), we consider the following framework of direct search based on sufficient decrease.

Algorithm 2.1 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 $\rho : (0,\infty) \to (0,\infty)$. For $k = 0, 1, 2, \ldots,$ do the following.

1. Generate a set of nonzero vectors $\mathcal{D}_k \subset \mathbb{R}^n$ deterministically or stochastically.
2. Check whether there exists a $d \in \mathcal{D}_k$ such that

$$f(x_k) - f(x_k + \alpha_k d) \;>\; \rho(\alpha_k). \tag{2.1}$$

3. If d exists, set $x_{k+1} = x_k + \alpha_k d$, $\alpha_{k+1} = \gamma\alpha_k$; otherwise, set $x_{k+1} = x_k$, $\alpha_{k+1} = \theta\alpha_k$.

In Algorithm 2.1, inequality (2.1) is called the sufficient decrease condition, where we require that the forcing function ρ should be continuous, positive, nondecreasing, and satisfy $\rho(\alpha) = o(\alpha)$ when $\alpha \to 0^+$. The typical choice of the forcing function is $\rho(\alpha) = c\alpha^2$, where c is a positive constant. For simplicity, we declare the iteration existing a d satisfying (2.1) as a successful iteration; otherwise, an unsuccessful iteration. There are two points worth noting in Algorithm 2.1. Firstly, in the third step, we have not decided inside a successful iteration which direction d to move along if there are multiple choices. Roughly speaking, there are two typical strategies. One is to choose the direction that decreases the function value the most, which is called complete polling. The other is to choose the first direction that satisfies (2.1), which is called opportunistic polling. Secondly, we can set an upper bound $\alpha_{\max} \in (0,\infty]$ for step sizes and let $\alpha_{k+1} = \min\{\gamma\alpha_k, \alpha_{\max}\}$ for successful iterations.

We introduce the definition of cosine measure, which is a key concept in the later convergence analysis.

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

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

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 2.1. To avoid an ill-posed definition, from here and onwards, we assume by convention that $\mathrm{cm}\,(D, v) = 1$ when $v = 0$, which is the same as in [11].

If we assume there exists a $\kappa > 0$ such that $\mathrm{cm}(\mathcal{D}_k) \geq \kappa$ for each $k \geq 0$, then we can guarantee the convergence of Algorithm 2.1 under some technical assumptions [15], where the convergence means $\liminf_k \|\nabla f(x_k)\| = 0$.

2.2 Probabilistic direct search and its convergence

Let us first introduce the basic definitions and notations from probability theory that will be used throughout this paper.

We consider the probability space $(\Omega, \mathcal{F}, \mathbb{P})$ and denote the corresponding random variable of $\mathcal{D}_k$ in Algorithm 2.1 by $\mathfrak{D}_k$. Then we use $\{\mathfrak{D}_k\}$ to define a filtration $\mathbb{F} = \{\mathcal{F}_k\}$ in our probability space, where $\mathcal{F}_k = \sigma(\mathfrak{D}_0, \ldots, \mathfrak{D}_k)$. Here $\sigma(\mathfrak{D}_0, \ldots, \mathfrak{D}_k)$ represents the σ-algebra generated by random variables $\mathfrak{D}_0, \ldots, \mathfrak{D}_k$. Note that we define $\mathcal{F}_{-1}$ to be the trivial σ-algebra $\{\emptyset, \Omega\}$ which is also applied to the other filtration discussed throughout this paper for consistence of notations. Also, we denote the corresponding random variables of x_k, α_k, $\nabla f(x_k)$, d by X_k, A_k, G_k, $\mathfrak{d}$ respectively. Since all randomness come from $\{\mathfrak{D}_k\}$, we find that X_k, A_k, and G_k are all measurable to $\mathcal{F}_{k-1}$ for all $k \geq 0$. (In probability theory, $\{X_k\}$, $\{A_k\}$, and $\{G_k\}$ are called predictable processes with respect to the filtration $\mathbb{F}$.) We also denote the conditional expectation of a random variable W given a σ-algebra $\mathcal{G}$ by $\mathbb{E}(W \mid \mathcal{G})$ (see [10, Section 5.1] for more detailed definition). Then we define the conditional probability of an event $A \in \mathcal{G}$ given a σ-algebra $\mathcal{G}$ by $\mathbb{P}\left(A \mid \mathcal{G}\right) = \mathbb{E}\left(\mathbb{1}(A) \mid \mathcal{G}\right)$, where $\mathbb{1}(A)$ is the indicator function of the event A.

Instead of being required that $\{\mathrm{cm}(\mathcal{D}_k)\}$ shares a uniform positive lower bound, probabilistic direct search only needs to satisfy p_0-probabilistically κ-descent with the help of following definition.

Definition 2.2 ([11, Definition 3.1])**.** Consider Algorithm 2.1 with f being differentiable. The sequence $\{\mathfrak{D}_k\}$ is said to be p-probabilistically κ-descent if it satisfies

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

Remark 2.2. The inequality in Definition 2.2 should be understood in the almost sure sense, that is, for each $k \geq 0$,

$$\mathbb{P}\left(\mathbb{P}\left(\mathrm{cm}\left(\mathfrak{D}_k, -G_k\right) \geq \kappa \mid \mathcal{F}_{k-1}\right) \geq p\right) \;=\; 1.$$

This is because the conditional probability $\mathbb{P}(\cdot \mid \mathcal{F}_{k-1})$ is a random variable, which is only defined up to almost sure equivalence. Henceforth, all the 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.

In literature, we often assume the sequence $\{\mathfrak{D}_k\}$ shares uniform and positive lower and upper bounds on the length. Without loss of generality, we make a blanket assumption that each element of $\mathfrak{D}_k$ is a unit vector as follows.

Blanket Assumption. *Any realization of $\mathfrak{D}_k$ is a finite collection of unit vectors in $\mathbb{R}^n$.*

Using Definition 2.2, the convergence of probabilistic direct search is established as follows.

Theorem 2.1 ([11, Theorem 3.4])**.** *Consider Algorithm 2.1 with f being differentiable and lower bounded, and ∇f being Lipschitz continuous. If $\{\mathfrak{D}_k\}$ is p_0-probabilistically κ-descent with $p_0 = \log\theta/\log(\gamma^{-1}\theta)$ as κ being a positive constant, then*

$$\mathbb{P}\left(\liminf_{k\to\infty} \|G_k\| = 0\right) \;=\; 1. \tag{2.2}$$

In practice, $\mathfrak{D}_k$ is typically chosen to be m independent random vectors, uniformly distributed on the unit sphere in $\mathbb{R}^n$. The following corollary, derived from Theorem 2.1, provides guidance on selecting m to ensure convergence of the probabilistic direct search.

Corollary 2.1 ([11, Corollary B.4])**.** *Consider Algorithm 2.1 with f satisfying the assumptions in Theorem 2.1. Let $\mathfrak{D}_k = \{\mathfrak{d}_1, \ldots, \mathfrak{d}_m\}$, where $\{\mathfrak{d}_\ell\}_{\ell=1}^m$ are independent random vectors uniformly distributed on the unit sphere in $\mathbb{R}^n$. Then $\{G_k\}$ achieves (2.2) if*

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

3 Failure of global convergence

A natural question arises concerning the behavior of Algorithm 2.1 when $\{\mathfrak{D}_k\}$ fails to satisfy the p-probabilistically κ-descent property (Definition 2.2). More specifically, we are interested in the case that in Corollary 2.1

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

Thus, we give a simple test as follows. We choose the objective function $f(x) = x^\mathsf{T}x/2$ with $x \in \mathbb{R}^2$. We set the initial point $x_0 = (-10, 0)^\mathsf{T}$, the initial step size $\alpha_0 = 1$, the shrinking factor $\theta = 1/4$, the expanding factor $\gamma = 3/2$, and the forcing function $\rho(\alpha) = \alpha^2/10^3$.

We then generate $m = 2$ searching directions independently and uniformly distributed on the unit sphere in $\mathbb{R}^2$ in each iteration so that $m < \log_2(1 - \log\theta/\log\gamma)$. We run Algorithm 2.1 for 5000 times, in each of which we stop the algorithm if the step size is less than or equal to the machine epsilon ($\approx 2.22 \times 10^{-16}$). The results are shown in Figure 1, where the blue circle represents the initial point, the red pentagram represents the global minimizer, and each black dot represents the output of the algorithm for each run. We

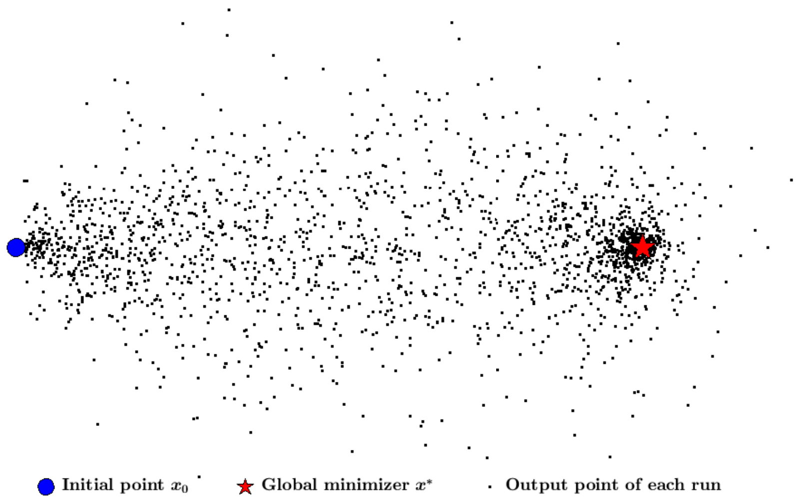

Figure 1: An illustrative example of failure of global convergence

observe that many runs of the algorithm fail to converge to the global minimizer, which motivates us to investigate the non-convergence behavior of Algorithm 2.1. In the remainder of this section, we will propose specific assumptions under which non-convergence will occur. First we introduce a new concept, denoted as "p-probabilistically ascent" as follows.

Definition 3.1 (p-probabilistically ascent)**.** Consider Algorithm 2.1 with f being differentiable. The sequence $\{\mathfrak{D}_k\}$ is said to be p-probabilistically ascent if it satisfies

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

Remark 3.1. It may be appealing to define p-probabilistically ascent as

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

for each $k \geq 0$. However, if we remove the indicator function $\mathbb{1}(G_k \neq 0)$, then the definition will enforce the algorithm never to find any stationary point almost surely since by our convention that $\mathrm{cm}\,(D, 0) = 1$, the left-hand side of the inequality (3.2) will collapse to 0 for any $\omega \in \Omega$ such that $G_k(\omega) = 0$. With the help of the indicator function, we can ensure that the inequality of the conditional probability is consistent when $G_k = 0$. We also see that $\{G_k \neq 0\} \in \mathcal{F}_{k-1}$ so that the inequality (3.1) makes sense and Definition 3.1 is well-defined.

We then define the indicator function

$$Y_k = \mathbb{1}(\{\mathrm{cm}\,(\mathfrak{D}_k, -G_k) \leq 0\} \cup \{G_k = 0\}). \tag{3.3}$$

We observe that if f is convex and smooth, then

$$A_{k+1} \;\leq\; \begin{cases} \theta A_k, & \text{if } Y_k = 1 \\ \gamma A_k, & \text{otherwise} \end{cases} \;=\; \gamma^{1-Y_k}\theta^{Y_k} A_k \tag{3.4}$$

since $\{Y_k = 1\}$ implies no descent directions in $\mathfrak{D}_k$. Moreover, we have the following lemma showing that $\{Y_k\}$ satisfies submartingale-like property.

Lemma 3.1. *Consider Algorithm 2.1 with f being differentiable and define Y_k as (3.3). If $\{\mathfrak{D}_k\}$ is p-probabilistically ascent, then we have*

$$\mathbb{P}\left(Y_k = 1 \mid \mathcal{F}^Y_{k-1}\right) \;\geq\; p,$$

where $\mathcal{F}^Y_{k-1} = \sigma(Y_0, \ldots, Y_{k-1})$.

Proof. Using the definition of Y_k, we have

$$\begin{aligned} \mathbb{P}\,(Y_k = 1 \mid \mathcal{F}_{k-1}) \;&=\; \mathbb{P}\,(\{\mathrm{cm}\,(\mathfrak{D}_k, -G_k) \leq 0\} \cup \{G_k = 0\} \mid \mathcal{F}_{k-1}) \\ &=\; \mathbb{P}\,(\mathrm{cm}\,(\mathfrak{D}_k, -G_k) \leq 0 \mid \mathcal{F}_{k-1}) + \mathbb{P}\,(G_k = 0 \mid \mathcal{F}_{k-1}) \\ &\geq\; p\mathbb{1}(G_k \neq 0) + \mathbb{1}(G_k = 0) \;\geq\; p, \end{aligned}$$

where the second equality is due to the fact that $\{\mathrm{cm}\,(\mathfrak{D}_k, -G_k) \leq 0\}$ and $\{G_k = 0\}$ are disjoint events, and the last inequality is because $\{G_k = 0\} \in \mathcal{F}_{k-1}$ so that

$$\mathbb{P}\,(G_k = 0 \mid \mathcal{F}_{k-1}) \;=\; \mathbb{E}\,(\mathbb{1}(G_k = 0) \mid \mathcal{F}_{k-1}) \;=\; \mathbb{1}(G_k = 0).$$

Recalling $\mathcal{F}_k = \sigma(\mathfrak{D}_0, \ldots, \mathfrak{D}_k)$, we have $Y_k \in \mathcal{F}_k$ so that $\mathcal{F}^Y_{k-1} \subseteq \mathcal{F}_{k-1}$, which implies the desired inequality. □

In the following, we will first use the definition of p-probabilistically ascent to establish the non-convergence of probabilistic direct search by Markov's inequality in Section 3.1 and Chernoff bound in Section 3.2, respectively. A weaker assumption will be proposed in Section 3.3 to broaden non-convergence analysis.

3.1 Non-convergence analysis by Markov's inequality

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

Theorem 3.1. *Consider Algorithm 2.1 with f being convex, differentiable, and having an optimal solution set $\mathcal{S}$. If $\{\mathfrak{D}_k\}$ is p-probabilistically ascent with $p > (\gamma - 1)/(\gamma - \theta)$, then we have*

$$\mathbb{P}\left(\liminf_{k\to\infty} \operatorname{dist}(X_k, \mathcal{S}) = 0\right) \;<\; 1$$

provided $\operatorname{dist}(x_0, \mathcal{S}) > \alpha_0/(1 - \gamma + p(\gamma - \theta))$.

Proof. Denote $\alpha_0/(1 - \gamma + p(\gamma - \theta))$ by q. Then it is sufficient to prove

$$\mathbb{P}\left(\sum_{k=0}^{\infty} A_k \geq \operatorname{dist}(x_0, \mathcal{S})\right) \;\leq\; \frac{q}{\operatorname{dist}(x_0, \mathcal{S})}$$

since $\mathbb{P}(\liminf_k \operatorname{dist}(X_k, \mathcal{S}) = 0) \leq \mathbb{P}(\sum_{k=0}^{\infty} A_k \geq \operatorname{dist}(x_0, \mathcal{S}))$. Using Markov's inequality, it suffices to prove

$$\mathbb{E}\left(\sum_{k=0}^{\infty} A_k\right) \;\leq\; q.$$

Since A_k is positive for each k, it is equivalent to prove $\sum_{k=0}^{\infty} \mathbb{E}(A_k) \leq q$ by Tonelli's theorem [26, Tonelli's Theorem, Page 420]. After defining Y_k as (3.3) and using (3.4), we establish the inequality for expectations of step sizes by

$$\begin{aligned} \mathbb{E}(A_{k+1}) &= \mathbb{E}\left(\mathbb{E}\left(A_{k+1} \mid \mathcal{F}_{k-1}\right)\right) \\ &\leq \mathbb{E}\left(A_k \mathbb{E}\left(\gamma^{1-Y_k}\theta^{Y_k} \mid \mathcal{F}_{k-1}\right)\right) \\ &\leq (\gamma(1-p) + \theta p)\mathbb{E}\left(A_k\right), \end{aligned} \tag{3.5}$$

where the first equality is due to the tower property of expectations, and the last inequality is due to the assumption of $\{\mathfrak{D}_k\}$. Iteratively using (3.5), we have $\mathbb{E}(A_k) \leq (\gamma(1-p)+\theta p)^k \alpha_0$ for each $k \geq 0$. Since $\gamma(1-p) + \theta p > 0$, we conclude our proof. □

3.2 Non-convergence analysis by Chernoff bound

In the preceding section, the requirement that $p > (\gamma - 1)/(\gamma - \theta)$ appeared overly rigid. Hence, in this section, we aim to weaken it to $p > p_*$ with

$$p_* \;=\; 1 - p_0 \;=\; \frac{\log \gamma}{\log(\theta^{-1}\gamma)}, \tag{3.6}$$

where p_0 is defined in the convergence theorem (Theorem 2.1).

We observe that inequality (3.4) provides us a natural upper bound for each A_k:

$$A_k \le \alpha_0 \prod_{\ell=0}^{k-1} \gamma^{1-Y_\ell} \theta^{Y_\ell}. \tag{3.7}$$

It will turn out that the 0-1 process $\{Y_k\}$ plays an essential role, and we will analyze it in the following. Hence, given any 0-1 process $\{Y_k\}$, we define the filtration $\mathbb{F}^Y = \{\mathcal{F}_k^Y\}$, where $\mathcal{F}_k^Y = \sigma(Y_0, \ldots, Y_k)$. We define another two stochastic processes $\{U_k\}_{k\ge1}$ and $\{\overline{Y}_k\}_{k\ge1}$ as follows.

$$U_k = U_k(Y_0, \ldots, Y_{k-1}) = \prod_{\ell=0}^{k-1} \gamma^{1-Y_\ell} \theta^{Y_\ell} \quad \text{for each } k \ge 1, \tag{3.8}$$

and

$$\overline{Y}_k = \overline{Y}_k(Y_0, \ldots, Y_{k-1}) = \frac{1}{k} \sum_{\ell=0}^{k-1} Y_\ell \quad \text{for each } k \ge 1. \tag{3.9}$$

For clarity and convenience of notations, we will keep these notations till the end of this paper. Recalling Lemma 3.1, we have $\{Y_k\}$ satisfies the submartingale-like property

$$\mathbb{P}\left(Y_k = 1 \mid \mathcal{F}_{k-1}^Y\right) \ge p \quad \text{for each } k \ge 0 \tag{3.10}$$

if $\{\mathfrak{D}_k\}$ is p-probabilistically ascent. Following lemmas will be devoted to the analysis of the behavior of $\{U_k\}_{k\ge1}$ and $\{\overline{Y}_k\}_{k\ge1}$ when inequality (3.10) holds. Before this, Lemma 3.2 will show that the inequality concerning the conditional probability with respect to a σ-algebra will be preserved when translated to the conditional probability with respect to a nonzero measure event.

Lemma 3.2. *Let $\{Y_k\}$ be a 0-1 process satisfying*

$$\mathbb{P}\left(Y_k = 1 \mid \mathcal{F}_{k-1}^Y\right) \ge p \quad \textit{for each } k \ge 0.$$

Then for each $k \ge 0$, we have

$$\mathbb{P}\left(Y_k = 1 \mid E\right) \ge p \quad \textit{for all } E \in \mathcal{F}_{k-1}^Y \textit{ such that } \mathbb{P}(E) > 0.$$

Lemma 3.3 establishes an upper bound for the conditional probability of the random variable $\overline{Y}_k$ being smaller than some given real number via the Chernoff bound. This forms a pivotal step towards demonstrating non-convergence.

Lemma 3.3. *Let $\{Y_k\}$ be a 0-1 process satisfying*

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

Then for any $q < p$, we have

$$\mathbb{P}\left(\overline{Y}_k \leq q \mid E_{k_0}\right) \;\leq\; \exp\left[-\frac{(q-p)^2}{2p}(k+k_0)\right] \quad \textit{for all } k \geq 0 \textit{ and } k_0 \geq 0, \tag{3.11}$$

where $E_{k_0} = \bigcap_{\ell=0}^{k_0-1}\{Y_\ell = 1\}$.

Proof. We only consider the nontrivial case where $k > k_0 \geq 1$ in inequality (3.11). By the definition of $\overline{Y}_k$ and Markov's inequality,

$$\begin{aligned}
\mathbb{P}\left(\overline{Y}_k \leq q \mid E_{k_0}\right) \;&=\; \mathbb{P}\left(\exp\left(-t\sum_{\ell=0}^{k-1} Y_\ell\right) \geq \mathrm{e}^{-tkq} \;\middle|\; E_{k_0}\right)\\
&\leq\; \mathrm{e}^{tkq}\;\mathbb{E}\left(\prod_{\ell=0}^{k-1} \mathrm{e}^{-tY_\ell} \;\middle|\; E_{k_0}\right) \qquad (3.12)\\
&=\; \mathrm{e}^{tkq-tk_0}\;\mathbb{E}\left(\prod_{\ell=k_0}^{k-1} \mathrm{e}^{-tY_\ell} \;\middle|\; E_{k_0}\right),
\end{aligned}$$

where t is an arbitrary positive number. Then we pay attention to $\mathbb{E}(\prod_{\ell=k_0}^{k-1} \mathrm{e}^{-tY_\ell} \mid E_{k_0})$. We use the tower property of expectations and get

$$\mathbb{E}\left(\prod_{\ell=k_0}^{k-1} \mathrm{e}^{-tY_\ell} \;\middle|\; \mathcal{F}_{k_0-1}^Y\right) \;=\; \mathbb{E}\left(\mathbb{E}\left(\mathrm{e}^{-tY_{k-1}} \mid \mathcal{F}_{k-2}^Y\right)\prod_{\ell=k_0}^{k-2} \mathrm{e}^{-tY_\ell} \;\middle|\; \mathcal{F}_{k_0-1}^Y\right), \tag{3.13}$$

where we set $\prod_{\ell=k_0}^{k-2} \mathrm{e}^{-tY_\ell} = 1$ when $k = k_0 + 1$ by convention. By the assumption on the conditional probability of $\{Y_k\}$, we have

$$\mathbb{E}\left(\mathrm{e}^{-tY_{k-1}} \mid \mathcal{F}_{k-2}^Y\right) \;\leq\; p\mathrm{e}^{-t} + 1 - p \;\leq\; \exp(p\mathrm{e}^{-t} - p), \tag{3.14}$$

where the last inequality is due to $x + 1 \leq \mathrm{e}^x$ for all x. By equality (3.13) and inequality (3.14), we have

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

Observing that $\mathbb{P}(E_{k_0}) \geq p^{k_0} > 0$, we have

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

by Lemma 3.2. Hence, we can further rewrite inequality (3.12) as

$$\mathbb{P}\left(\overline{Y}_k \le q \mid E_{k_0}\right) \;\le\; \exp[p(k-k_0)(\mathrm{e}^{-t}-1)+tkq-tk_0]. \tag{3.15}$$

Since inequality (3.15) holds for all $t>0$, we select $t=\log(p/q)$. Then we have

$$p(k-k_0)(\mathrm{e}^{-t}-1)+tkq-tk_0 \;=\; -(\frac{k}{2\xi}+\frac{k_0}{2\xi^2})(q-p)^2+(\frac{k_0}{p}-k_0)(q-p) \quad (q<\xi<p),$$

where the equality comes from Taylor expansion of the function

$$f(q) \;=\; (k-k_0)(q-p)+(k_0-kq)\log(\frac{q}{p})$$

at the point p. Therefore, one can show that

$$p(k-k_0)(\mathrm{e}^{-t}-1)+tkq \;\le\; -\frac{(q-p)^2}{2p}(k+k_0),$$

and conclude from inequality (3.15) that

$$\mathbb{P}\left(\overline{Y}_k \le q \mid E_{k_0}\right) \;\le\; \exp\left[-\frac{(q-p)^2}{2p}(k+k_0)\right].$$

□

Lemma 3.3 allows us to establish the following lemma, revealing the behavior of the series of $\{U_k\}_{k\ge1}$ when $\{Y_k\}$ fulfills inequality (3.10) with $p>p_*$.

Lemma 3.4. *Let $\{Y_k\}$ be a 0-1 process satisfying*

$$\mathbb{P}\left(Y_k=1 \mid \mathcal{F}^Y_{k-1}\right) \;\ge\; p \quad \text{for each } k\ge0,$$

where $p>p_*$. Then we have

$$\mathbb{P}\left(\sum_{k=1}^{\infty}U_k<\zeta\right) \;>\; 0 \quad\iff\quad \zeta>\frac{\theta}{1-\theta}.$$

Proof. The proof from left to right is shown by $U_k\ge\theta^k$. We focus on the proof from right to left. Let us fix one $\zeta>\theta/(1-\theta)$. For each $k_0\ge1$, we have

$$\mathbb{P}\left(\sum_{k=1}^{\infty}U_k<\zeta\right) \;\ge\; \mathbb{P}\left(\left\{\sum_{k=1}^{\infty}U_k<\zeta\right\}\bigcap E_{k_0}\right),$$

where E_{k_0} is defined in Lemma 3.3. After observing E_{k_0} is not a null set for each $k_0\ge1$, it suffices to prove that there exists a $k_0\ge1$ such that

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

We observe that

$$\mathbb{P}\left(\sum_{k=1}^{\infty} U_k < \zeta \;\middle|\; E_{k_0}\right) \;\geq\; \mathbb{P}\left(\sum_{k=k_0+1}^{\infty} U_k < \zeta - \frac{\theta}{1-\theta} \;\middle|\; E_{k_0}\right), \tag{3.16}$$

then we only need to prove the right hand side of inequality (3.16) is positive. For any $q > p_*$, we have $\gamma^{1-q}\theta^q < 1$, which leads to the convergence of the series $\sum_{k=k_0}^{\infty} \left(\gamma^{1-q}\theta^q\right)^k$. Then, for any $q > p_*$, there always exists an N such that, for any $k_0 \geq N$,

$$\sum_{k=k_0+1}^{\infty} \left(\gamma^{1-q}\theta^q\right)^k \;\leq\; \zeta - \frac{\theta}{1-\theta}. \tag{3.17}$$

Then we can conclude from equality (3.16) and inequality (3.17) that it suffices to prove there exists a $q > p_*$ such that the following inequality holds for all sufficient large k_0

$$\mathbb{P}\left(\sum_{k=k_0+1}^{\infty} U_k < \sum_{k=k_0+1}^{\infty} \left(\gamma^{1-q}\theta^q\right)^k \;\middle|\; E_{k_0}\right) \;>\; 0.$$

From the definitions of U_k and $\overline{Y}_k$, we can have

$$\begin{aligned}
\mathbb{P}\left(\sum_{k=k_0+1}^{\infty} U_k < \sum_{k=k_0+1}^{\infty} \left(\gamma^{1-q}\theta^q\right)^k \;\middle|\; E_{k_0}\right) \;&\geq\; \mathbb{P}\left(\bigcap_{k=k_0+1}^{\infty} \left\{U_k < \left(\gamma^{1-q}\theta^q\right)^k\right\} \;\middle|\; E_{k_0}\right) \\
&=\; \mathbb{P}\left(\bigcap_{k=k_0+1}^{\infty} \left\{\overline{Y}_k > q\right\} \;\middle|\; E_{k_0}\right) \\
&\geq\; 1 - \sum_{k=k_0+1}^{\infty} \mathbb{P}(\overline{Y}_k \leq q \mid E_{k_0}),
\end{aligned} \tag{3.18}$$

where the last inequality is due to the subadditivity of probability. According to Lemma 3.3 and inequality (3.18), it suffices to prove there exists a $q \in (p_*, p)$ such that for all sufficient large k_0 the following inequality holds

$$\sum_{k=k_0+1}^{\infty} \exp\left[-\frac{(q-p)^2}{2p}(k+k_0)\right] < 1,$$

which is true by the convergence of the series. □

Lemma 3.4 provides a qualitative result that the probability is nonzero provided that $\zeta > \theta/(1-\theta)$. Actually, based on the same framework of the proof of Lemma 3.4, we can obtain lower bounds of the probability with more careful analysis, which are shown in the following proposition, whose proof is provided in Appendix A for readers who are interested in the details.

Proposition 3.1. *Let* $\{Y_k\}$ *be a 0-1 process satisfying*

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

where $p > p_*$. *Then we have*

$$\mathbb{P}\left(\sum_{k=1}^{\infty} U_k < t\right) \;\geq\; \left(t - \frac{\theta}{1-\theta}\right)^{C_1} \qquad \textit{for all } t \in (\theta/(1-\theta), C_2), \tag{3.19}$$

and

$$\mathbb{P}\left(\sum_{k=1}^{\infty} U_k < t\right) \;\geq\; 1 - t^{-C_3} \qquad \textit{for all } t \in (C_4, \infty), \tag{3.20}$$

where C_1, C_2, C_3, *and* C_4 *are positive constants depending on* p, θ, *and* γ.

Building on Lemma 3.4, we establish the following non-convergence theorem for probabilistic direct search based on the assumption that $\{\mathfrak{D}_k\}$ is p-probabilistically ascent with $p > p_*$.

Theorem 3.2. *Consider Algorithm 2.1 with* f *being convex, differentiable, and having an optimal solution set* $\mathcal{S}$. *If* $\{\mathfrak{D}_k\}$ *is* p*-probabilistically ascent with* $p > p_*$, *then we have*

$$\mathbb{P}\left(\liminf_{k\to\infty} \operatorname{dist}(X_k, \mathcal{S}) = 0\right) \;<\; 1,$$

provided $\operatorname{dist}(x_0, \mathcal{S}) > \alpha_0/(1-\theta)$.

Proof. Define Y_k as (3.3). Then we can use inequality (3.7) and establish

$$\sum_{k=1}^{\infty} A_k \;\leq\; \alpha_0 \sum_{k=1}^{\infty} U_k.$$

By the assumption of $\{\mathfrak{D}_k\}$, we find that $\{Y_k\}$ satisfies the conditions in Corollary 3.4. Let $\zeta = \operatorname{dist}(x_0, \mathcal{S})/\alpha_0 - 1$. Since $\operatorname{dist}(x_0, \mathcal{S}) > \alpha_0/(1-\theta)$, we have

$$\begin{aligned} \mathbb{P}\left(\liminf_{k\to\infty} \operatorname{dist}(X_k, \mathcal{S}) = 0\right) \;&\leq\; \mathbb{P}\left(\sum_{k=0}^{\infty} A_k \geq \operatorname{dist}(x_0, \mathcal{S})\right) \\ &\leq\; \mathbb{P}\left(\sum_{k=1}^{\infty} U_k \geq \zeta\right) \;<\; 1, \end{aligned} \tag{3.21}$$

where the last inequality is due to Corollary 3.4. $\square$

If we assume ∇f or f is Lipschitz continuous respectively, using the same argument as in the proof of Theorem 3.2, we have the following corollaries of non-convergence provided that the norm of the initial gradient or function value is large enough.

Corollary 3.1. *Consider Algorithm 2.1 with f being convex, differentiable, and ∇f being L-Lipschitz continuous. If $\{\mathfrak{D}_k\}$ is p-probabilistically ascent with $p > p_*$, then we have*

$$\mathbb{P}\left(\liminf_{k\to\infty} \|G_k\| = 0\right) \;<\; 1,$$

provided $\|g_0\| > \alpha_0 L/(1-\theta)$.

Corollary 3.2. *Consider Algorithm 2.1 with f being convex, differentiable, and L-Lipschitz continuous. If $\{\mathfrak{D}_k\}$ is p-probabilistically ascent with $p > p_*$, then we have*

$$\mathbb{P}\left(\lim_{k\to\infty} f(X_k) = \inf_{x\in\mathbb{R}} f(x)\right) \;<\; 1,$$

provided $f(x_0) > \alpha_0 L/(1-\theta) + \inf_{x\in\mathbb{R}} f(x)$.

Proofs of both corollaries are almost the same with that in Theorem 3.2. It is noted that the result in Corollary 3.2 is about lim instead of lim inf since Algorithm 2.1 is monotone about $\{f(X_k)\}$.

3.3 Non-convergence under a weaker assumption

In this section, we will introduce a weaker assumption than the p-probabilistically ascent assumption with $p > p_*$. This less stringent assumption offers a broader perspective on the circumstances under which the algorithm may fail to converge.

The fundamental concept here is to explore the condition under which the series $\sum_k U_k$ converges with probability 1. To accomplish this, we need to prove Lemma 3.6 first, whose proof relies on the Azuma-Hoeffding inequality, content of which is shown as follows.

Lemma 3.5 (Azuma-Hoeffding Inequality [27, Theorem 2.2.6])**.** *Suppose $\{W_k\}$ is a martingale, and*

$$|W_k - W_{k-1}| \;\le\; c_k \;\; \textit{a.s.} \quad \textit{for each } k \ge 1.$$

Then for each $k \ge 1$ and all $\varepsilon > 0$,

$$\mathbb{P}\left(|W_k - W_0| \ge \varepsilon\right) \;\le\; 2\exp\left(\frac{-\varepsilon^2}{2\sum_{\ell=1}^{k} c_\ell^2}\right).$$

With the help of Lemma 3.5, we present as follows Lemma 3.6, which can be regarded as a non-i.i.d. version of the Law of Large Numbers.

Lemma 3.6. *Suppose $\{W_k\}$ is a uniformly bounded stochastic process. Then we have*

$$\mathbb{P}\left(\lim_{k\to\infty}\frac{1}{k}\sum_{\ell=0}^{k-1}(W_\ell - V_\ell) = 0\right) \;=\; 1,$$

where $V_\ell = \mathbb{E}\left(W_\ell \mid \mathcal{F}_{\ell-1}^W\right)$ with $\mathcal{F}_{\ell-1}^W = \sigma(W_0,\ldots,W_{\ell-1})$ for each $\ell \geq 0$.

Proof. Define $H_k = W_k - V_k$ and $S_k = \sum_{\ell=0}^{k-1} H_\ell$ for each $k \geq 0$, where we let $S_0 = 0$. Then we find that $\{S_k\}$ is a martingale. Further, since $\{W_k\}$ is uniformly bounded, there exists a constant c such that

$$|S_k - S_{k-1}| \;\leq\; c \;\; a.s. \quad \text{for each } k \geq 1.$$

Then by Azuma-Hoeffding Inequality (Lemma 3.5), we have, for each $k \geq 0$ and each $n \geq 1$,

$$\mathbb{P}\left(|S_k| > k/n\right) \;\leq\; 2\exp\left(-\frac{k}{2c^2(n+1)^2}\right),$$

which implies that $\sum_{k=0}^{\infty}\mathbb{P}\left(|S_k| > k/n\right) < \infty$. Then by Borel-Cantelli Lemma, we have, for each $n \geq 1$,

$$\mathbb{P}\left(\limsup_{k\to\infty}\left\{\left|\frac{S_k}{k}\right| > \frac{1}{n}\right\}\right) \;=\; 0.$$

Noticing that

$$\limsup_{k\to\infty}\left\{\left|\frac{S_k}{k}\right| > \frac{1}{n}\right\} \;=\; \left\{\limsup_{k\to\infty}\left|\frac{S_k}{k}\right| > \frac{1}{n}\right\}, \tag{3.22}$$

we have $\mathbb{P}(\lim_k S_k/k = 0) = 1$ by letting $n \to \infty$. □

Remark 3.2. We simply proof the equality (3.22). For a given sequence of functions $\{f_k\}$ and a real number y, we need to show that

$$\limsup_{k\to\infty}\{f_k > y\} \;=\; \left\{\limsup_{k\to\infty} f_k > y\right\}. \tag{3.23}$$

By the equivalent definition of lim sup of a set sequence, we have the left-hand side of the equality (3.23) is equivalent to

$$\{x : \text{there exists a subsequence } k(i) \text{ such that } f_{k(i)}(x) > y\},$$

which is equivalent to the right-hand side of the equality (3.23).

The next lemma answers the question of when the series $\sum_k U_k$ converges with probability 1 by proposing a lim inf-type assumption on $\{Y_k\}$.

Lemma 3.7. *Let* $\{Y_k\}$ *be a 0-1 process satisfying*

$$\mathbb{P}\left(\liminf_{k\to\infty} \mathbb{P}\left(Y_k = 1 \mid \mathcal{F}_{k-1}^Y\right) > p_*\right) \;>\; 0. \tag{3.24}$$

Then we have

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

Proof. Recall the definitions of U_k in (3.8) and $\overline{Y}_k$ in (3.9). Then by the root test of series, it suffices to prove that

$$\mathbb{P}\left(\liminf_{k\to\infty} \overline{Y}_k > p_*\right) \;>\; 0.$$

Let us define $P_k = \mathbb{P}\left(Y_k = 1 \mid \mathcal{F}_{k-1}^Y\right)$. By Lemma 3.6, we have

$$\lim_{k\to\infty} (\overline{Y}_k - \frac{1}{k}\sum_{\ell=0}^{k-1} P_\ell) = 0 \quad \text{a.s.}$$

Then, by computation rules of $\liminf$, we have

$$\liminf_{k\to\infty} \overline{Y}_k = \liminf_{k\to\infty} \frac{1}{k}\sum_{\ell=0}^{k-1} P_\ell \quad \text{a.s.}$$

Thus, we only need to prove that

$$\mathbb{P}\left(\liminf_{k\to\infty} \frac{1}{k}\sum_{\ell=0}^{k-1} P_\ell > p_*\right) \;>\; 0.$$

Recalling our assumption of $\{Y_k\}$ that $\mathbb{P}\left(\liminf_k P_k > p_*\right) \;>\; 0$, we conclude our proof by noticing

$$\liminf_{k\to\infty} \frac{1}{k}\sum_{\ell=0}^{k-1} P_\ell \;\geq\; \liminf_{k\to\infty} P_k.$$

□

Similar to how we demonstrated Theorem 3.2, we also establish the corresponding non-convergence theorem for probabilistic direct search using Lemma 3.7.

Theorem 3.3. *Consider Algorithm 2.1 with* f *being convex, differentiable, and having an optimal solution set* $\mathcal{S}$. *If* $\{\mathfrak{D}_k\}$ *satisfies*

$$\mathbb{P}\left(\liminf_{k\to\infty} \mathbb{P}\left(\{\mathrm{cm}\left(\mathfrak{D}_k, -G_k\right) \leq 0\} \cup \{G_k = 0\} \mid \mathcal{F}_{k-1}\right) > p_*\right) \;>\; 0, \tag{3.25}$$

then there exists a constant ζ *such that*

$$\mathbb{P}\left(\liminf_{k\to\infty} \mathrm{dist}(X_k, \mathcal{S}) = 0\right) \;<\; 1$$

provided $\mathrm{dist}(x_0, \mathcal{S}) > \zeta$.

Proof. We can duplicate the proof of Theorem 3.2 until finding a constant to establish inequality (3.21). After defining Y_k as (3.3), we have the series $\sum_{k=1}^{\infty} U_k$ is finite with nonzero probability by Lemma 3.7. By the subadditivity of probability, there exists a constant ζ_0 such that

$$\mathbb{P}\left(\sum_{k=1}^{\infty} U_k \le \zeta_0\right) > 0.$$

Thus, we conclude that, for any objective function which is convex and has an optimal solution set $\mathcal{S}$, if $\operatorname{dist}(x_0, \mathcal{S}) > \zeta = \alpha_0(1+\zeta_0)$, then

$$\mathbb{P}\left(\liminf_{k\to\infty} \operatorname{dist}(X_k, \mathcal{S}) = 0\right) \;\le\; \mathbb{P}\left(\sum_{k=1}^{\infty} U_k > \zeta_0\right) \;<\; 1.$$

□

Remark 3.3. It is not hard to find that, $\{\mathfrak{D}_k\}$ that is p-probabilistically ascent property with $p > p_*$ will automatically satisfy the inequality (3.25), meaning that (3.25) provides a weaker assumption. More specifically, recalling that the conditional probability $P_k = \mathbb{P}(\{\operatorname{cm}(\mathfrak{D}_k, -G_k) \le 0\} \cup \{G_k = 0\} \mid \mathcal{F}_{k-1})$ is a random variable, we find that p-probabilistically ascent with $p > p_*$ requires all P_k to be larger than p_* almost surely for each $k \ge 1$, while (3.25) only requires the lim inf of P_k to be larger than p_* with positive probability.

4 Tightness of the non-convergence results

In this section, we mainly focus on demonstrating that our non-convergence analysis for probabilistic direct search is tight. First, we show that our analysis nearly encompasses the counterpart of Corollary 2.1 except one particular case. We then explain the reason behind the diminished effectiveness of our analysis at this particular case. Furthermore, we provide an example that probabilistic direct search does converge in this case.

Recall what has been established in Corollary 2.1, which states that Algorithm 2.1 will converge with probability 1 if

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

Then the following corollary shows the non-convergence side.

Corollary 4.1. *Let $\mathfrak{D}_k = \{\mathfrak{d}_1, \dots, \mathfrak{d}_m\}$, where $\{\mathfrak{d}_\ell\}_{\ell=1}^m$ are i.i.d. random vectors uniformly distributed on the unit sphere in $\mathbb{R}^n$. If*

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

then for any convex and differentiable function that has an optimal solution set $\mathcal{S}$ satisfying $\operatorname{dist}(x_0, \mathcal{S}) > \alpha_0/(1-\theta)$*, we have*

$$\mathbb{P}\left(\liminf_{k\to\infty} \operatorname{dist}(X_k, \mathcal{S}) = 0\right) \; < \; 1.$$

Proof. By Theorem 3.2, we only need to prove that $\{\mathfrak{D}_k\}$ is p-probabilistically ascent with $p > p_*$. From the construction of $\{\mathfrak{D}_k\}$, we have

$$\begin{aligned} \mathbb{P}\left(\operatorname{cm}(\mathfrak{D}_k, -G_k) \le 0 \mid \mathcal{F}_{k-1}\right) \; &= \; \mathbb{P}\left(\operatorname{cm}(\mathfrak{D}_k, -G_k) \le 0\right) \mathbb{1}(G_k \neq 0) \\ &= \; \left[\mathbb{P}\left(\mathfrak{d}_\ell^\top G_k \ge 0\right)\right]^m \mathbb{1}(G_k \neq 0) \; = \; 2^{-m}\mathbb{1}(G_k \neq 0), \end{aligned}$$

which means that $\{\mathfrak{D}_k\}$ is 2^{-m}-probabilistically ascent. Recalling the definition of p_*, it is not hard to see that $2^{-m} > p_*$ if and only if $m < \log_2\left(1 - \log\theta/\log\gamma\right)$. □

Remark 4.1. Combining Corollaries 2.1 and 4.1, we observe that the non-convergence region of the algorithmic parameters θ, γ, and m is nearly the complement of the convergence region, except the special case $m = \log_2(1 - \log\theta/\log\gamma)$. But this special case is not a concern, since in most cases $\log_2(1 - \log\theta/\log\gamma)$ is not an integer.

The case $m = \log_2(1 - \log\theta/\log\gamma)$ mentioned above is a specific instance of $\{\mathfrak{D}_k\}$ satisfying p-probabilistically ascent with $p = p_*$. Next, we discuss on the reason why our analysis does not work under the assumption of p_*-probabilistically ascent. To begin with, we provide the definition of the recurrent value of a random walk as follows.

Definition 4.1 (Recurrent value of a random walk [10, Section 4.2])**.** Let $\{S_k\}$ be a random walk in $\mathbb{R}^n$. The number $x \in \mathbb{R}^n$ is said to be a recurrent value for $\{S_k\}$ if, for every $\varepsilon > 0$,

$$\mathbb{P}\left(\|S_k - x\|_\infty < \varepsilon \text{ i.o.}\right) \; = \; 1.$$

Definition 4.2. A random walk $\{S_k\}$ is said to be recurrent if the set of its recurrent values is nonempty.

Then we present the well-known Chung-Fuchs Theorem below, which provides a sufficient condition for a one-dimensional random walk to be recurrent.

Theorem 4.1 (Chung-Fuchs Theorem [10, Theorem 4.2.7])**.** *Let $\{S_k\}$ be a one-dimensional random walk. If $S_k/k \to 0$ in probability, then $\{S_k\}$ is recurrent.*

In Theorems 3.2 and 3.3, our proofs depend on the analysis of the probability that the series $\sum_k U_k$ converges. However, the subsequent discussion reveals that if we allow $p = p_*$ in inequality (3.10), the series $\sum_k U_k$ can even diverge with probability 1, which is the scenario we aim to avoid.

Lemma 4.1. *Let $\{Y_k\}$ be a 0-1 process satisfying*

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

where p is a constant between 0 and 1. Then $Y_0, Y_1, \ldots$ are i.i.d.

Proposition 4.1. *Let $\{Y_k\}$ be a 0-1 process satisfying*

$$\mathbb{P}\left(Y_k = 1 \mid \mathcal{F}_{k-1}^Y\right) \;=\; p_* \quad \textit{for each } k \geq 0,$$

where $p_* = \log\gamma / \log(\theta^{-1}\gamma)$. Then we have

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

Proof. When $\gamma = 1$, it is trivial. When $\gamma > 1$, by the assumption of $\{Y_k\}$ and Lemma 4.1, we know that $Y_0, Y_1, \ldots$ are i.i.d. random variables. Let $I_k = Y_k \log\theta + (1 - Y_k)\log\gamma$. Then we find that $I_0, I_1, \ldots$ are i.i.d., and satisfy that, for every $k \geq 0$,

$$\mathbb{P}\left(I_k = \log\theta\right) \;=\; 1 - \mathbb{P}\left(I_k = \log\gamma\right) \;=\; p_*,$$

which implies $\mathbb{E}I_k = 0$. Define $S_k = \sum_{\ell=0}^{k-1} I_\ell$. Since $I_0, I_1, \ldots$ are i.i.d., by definition $\{S_k\}$ is indeed a one-dimensional random walk. By the i.i.d. property, $\{I_k\}$ enjoys the law of large numbers so that $S_k/k \to 0$ in probability. Then Theorem 4.1 tells us the random walk $\{S_k\}$ is recurrent. Suppose that δ is one of the recurrent values of $\{S_k\}$. Then for any given $\varepsilon > 0$, with probability 1, there exists a subsequence of $\{S_k\}$ denoted by $\{S_{k(i)}\}$ such that $S_{k(i)} > \delta - \varepsilon$ for all $i \geq 0$. Then we have with probability 1,

$$\begin{aligned}\sum_{k=1}^{\infty} \exp\left(S_k\right) \;&\geq\; \sum_{i=1}^{\infty} \exp\left(S_{k(i)}\right) \\ &>\; \exp\left(\delta - \varepsilon\right) \sum_{i=1}^{\infty} 1 \;=\; \infty.\end{aligned}$$

We finish our proof by recalling that $U_k = \exp(S_k)$ for all $k \geq 0$. □

Having explained why the previous analysis fails, in the subsequent part we construct a $\{\mathfrak{D}_k\}$ that is p_*-probabilistically ascent but leads to the convergence of probabilistic direct search with probability 1. Before presenting the example, we introduce the following lemma, which is useful for the proof.

Lemma 4.2 ([14, Theorem 3.1]). *Consider Algorithm 2.1 with the complete polling strategy, and with f being differentiable and lower bounded, and ∇f being Lipschitz continuous. Define the series*

$$S(\kappa) \;=\; \sum_{k=1}^{\infty} \prod_{\ell=0}^{k-1} \gamma^{1-Y_\ell(\kappa)} \theta^{Y_\ell(\kappa)}, \tag{4.1}$$

where $Y_\ell(\kappa) = \mathbb{1}(\mathrm{cm}(\mathfrak{D}_\ell, -G_\ell) \le \kappa)$. If there exists a $\kappa > 0$ such that $\mathbb{P}(S(\kappa) = \infty) = 1$, then

$$\mathbb{P}\left(\liminf_{k\to\infty} \|G_k\| = 0\right) \;=\; 1.$$

We now provide a concrete example to demonstrate that probabilistic direct search converges when $\{\mathfrak{D}_k\}$ is p_*-probabilistically ascent.

Theorem 4.2. *We assume that $f : \mathbb{R}^n \to \mathbb{R}$ is continuously differentiable, and ∇f is L-Lipschitz continuous. Consider Algorithm 2.1 with $\mathfrak{D}_k = \{\mathfrak{d}_k\}$ such that, for each $k \ge 0$,*

$$\begin{aligned}
\mathbb{P}\left(\mathfrak{d}_k = \frac{G_k}{\|G_k\|} \middle| \mathcal{F}_{k-1}\right) \;&=\; p_* \mathbb{1}(G_k \ne 0),\\
\mathbb{P}\left(\mathfrak{d}_k = -\frac{G_k}{\|G_k\|} \middle| \mathcal{F}_{k-1}\right) \;&=\; (1-p_*) \mathbb{1}(G_k \ne 0),\\
\mathbb{P}\left(\mathfrak{d}_k = v_0 \mid \mathcal{F}_{k-1}\right) \;&=\; \mathbb{1}(G_k = 0),
\end{aligned}$$

where v_0 is a fixed unit vector. Then we have

$$\mathbb{P}\left(\liminf_{k\to\infty} \|G_k\| = 0\right) \;=\; 1.$$

Proof. Applying Lemma 4.2, it is sufficient to prove whether complete polling assumption holds and whether there exists a positive κ such that the series $S(\kappa)$ diverges with probability 1, where $S(\kappa)$ is defined in the equation (4.1). By the construction of $\mathfrak{D}_k$ that it only contains one direction in each iteration, the complete polling assumption holds automatically. Let us fix $\kappa = 1/2$ from now on and rewrite the series $S(1/2)$ as

$$S(\frac{1}{2}) \;=\; \sum_{k=1}^{\infty} \exp\left(W_k\right)$$

where $W_k = \sum_{\ell=0}^{k-1} (Y_\ell(1/2) \log\theta + (1 - Y_\ell(1/2)) \log\gamma)$. We find that it is sufficient to prove that

$$\mathbb{P}\left(\limsup_{k\to\infty} W_k > -\infty\right) \;=\; 1.$$

Based on our assumption on $\mathfrak{d}_k$, we know that $\{W_k\}$ is a submartingale with bounded increments. Then we finish our proof by applying Doob's decomposition theorem ([10, Theorem 5.2.10]) and martingale bounded increments theorem ([10, Theorem 5.3.1]). □

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. More specifically, the series of step sizes in Algorithm 2.1 converges with a nonzero probability if the set of polling directions are p-probabilistically ascent with $p > \log\gamma / \log(\theta^{-1}\gamma)$, where θ and γ are shrinking and expanding factors of the step size, respectively. Moreover, a weaker assumption is proposed to substitute the "p-probabilistically ascent" assumption while the nonzero probability of the convergence of series of step sizes is still guaranteed. The final part demonstrates the tightness of our non-convergence analysis by observing that there is almost no gap between the convergence and non-convergence theory under the typical choice of searching sets. Besides, we explain the reason for the failure of our analysis tools when allowing p-probabilistically ascent with $p = p_*$ instead of $p > p*$. Finally, we provide a concrete example showing that probabilistic direct search converges when $\{\mathfrak{D}_k\}$ is p_*-probabilistically ascent.

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

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

[8] 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.

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

[10] R. Durrett. *Probability: Theory and Examples.* Camb. Ser. Stat. Probab. Math. Cambridge University Press, Cambridge, fourth edition, 2010.

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

[12] 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.

[13] 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.

[14] C. Huang and Z. Zhang. Revisiting the convergence analysis of direct search and derivative-dree trust region. `https://github.com/OptHuang/RCDS_article`, 2024.

[15] 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.

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

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

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

[19] 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.

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

[21] 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.

[22] 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.

[23] 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.

[24] 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.

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

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

[27] R. Vershynin. *High-Dimensional Probability: An Introduction with Applications in Data Science*. Cambridge University Press, Cambridge, 2018.

[28] X. Wang and Y. Yuan. Stochastic trust region methods with trust region radius depending on probabilistic models. *arXiv:1904.03342*, 2019.

A Proof of Proposition 3.1

Proof. Fixing $q = (p_* + p)/2$, we denote $\gamma^{1-q}\theta^q$ and $(q-p)^2/(2p)$ by λ and ν respectively. We also denote $\mathbb{P}(\sum_{k=1}^{\infty} U_k < t)$ by $F(t)$ for simplicity throughout the following proof.

Proof of Inequality (3.19).

Fix a $t \in (\theta/(1-\theta), \infty)$ and define $\delta = t - \theta/(1-\theta)$. We claim that

$$F(t) \;\geq\; p^{k_t}\left[1 - \frac{\mathrm{e}^{-\nu k_t}}{1-\mathrm{e}^{-\nu}}\right], \tag{A.1}$$

with

$$k_t \;=\; \left\lceil \frac{1}{\log\lambda}\left[\log(1-\lambda) + \log\delta\right]\right\rceil.$$

If the claim is true, it is sufficient to prove that there exists a positive constant C_1 such that for all δ small enough,

$$p^{k_t}\left[1 - \frac{\mathrm{e}^{-\nu k_t}}{1-\mathrm{e}^{-\nu}}\right] \;\geq\; \delta^{C_1},$$

which is equivalent to

$$C_1 \;\geq\; \frac{\log p}{\log\delta}\left\lceil \frac{1}{\log\lambda}\left[\log(1-\lambda) + \log\delta\right]\right\rceil + \frac{1}{\log\delta}\log\left[1 - \frac{\mathrm{e}^{-\nu k_t}}{1-\mathrm{e}^{-\nu}}\right]. \tag{A.2}$$

By letting $t \downarrow \theta/(1-\theta)$, we observe that the limit of the right-hand side of inequality (A.2) is $\log p/\log \lambda > 0$, which implies the existence of C_1.

Now we focus on proving the inequality (A.2). Defining E_{k_0} as in Lemma 3.3, we have for each $k_0 \geq 0$,

$$F(t) \;\geq\; \mathbb{P}(E_{k_0})\,\mathbb{P}\left(\sum_{k=1}^{\infty} U_k < t \,\middle|\, E_{k_0}\right),$$

According to the definition of E_{k_0}, we have

$$\begin{aligned} F(t) \;&\geq\; p^{k_0}\mathbb{P}\left(\sum_{k=k_0+1}^{\infty} U_k < t - \sum_{k=1}^{k_0}\theta^k \,\middle|\, E_{k_0}\right) \\ &\geq\; p^{k_0}\mathbb{P}\left(\sum_{k=k_0+1}^{\infty} U_k < \delta \,\middle|\, E_{k_0}\right). \end{aligned} \tag{A.3}$$

If we further require k_0 to satisfy $\sum_{k=k_0+1}^{\infty}\lambda^k \leq \delta$, then we have

$$\begin{aligned} \mathbb{P}\left(\sum_{k=k_0+1}^{\infty} U_k < \delta \,\middle|\, E_{k_0}\right) \;&\geq\; \mathbb{P}\left(\sum_{k=k_0+1}^{\infty} U_k < \sum_{k=k_0+1}^{\infty}\lambda^k \,\middle|\, E_{k_0}\right) \\ &\geq\; \mathbb{P}\left(\bigcap_{k=k_0+1}^{\infty}\left\{U_k < \lambda^k\right\} \,\middle|\, E_{k_0}\right) \\ &\geq\; 1 - \sum_{k=k_0+1}^{\infty}\mathbb{P}\left(\overline{Y}_k \leq q \mid E_{k_0}\right). \end{aligned} \tag{A.4}$$

Combining inequalities (A.3) and (A.4) with Lemma 3.3, we have

$$F(t) \;\geq\; p^{k_0+1}\left[1 - \frac{\mathrm{e}^{-\nu(k_0+1)}}{1-\mathrm{e}^{-\nu}}\right]$$

for all $k_0 \geq [\log(1-\lambda) + \log\delta]/\log\lambda$. We conclude our proof by choosing $k_0 = k_t$.

Proof of Inequality (3.20).

Fix a $t > \theta/(1-\theta)$. We claim that

$$F(t) \;\geq\; 1 - \frac{\mathrm{e}^{-\nu(k_t+1)}}{1-\mathrm{e}^{-\nu}}, \tag{A.5}$$

with

$$k_t \;=\; \left\lfloor \frac{1}{\log\gamma}\log\left[(\gamma-1)\left(t - \frac{\lambda}{1-\lambda}\right)\right] - 1 \right\rfloor.$$

If the claim is true, then it is sufficient to prove that there exists a positive constant C_3 such that for all t large enough,

$$\frac{\mathrm{e}^{-\nu(k_t+1)}}{1-\mathrm{e}^{-\nu}} \;\leq\; t^{-C_3},$$

which is equivalent to

$$C_3 \;\leq\; \frac{\nu}{\log t}\left[\frac{\log(\gamma-1)}{\log\gamma}+\frac{\log(t-\frac{\lambda}{1-\lambda})}{\log\gamma}\right]+\frac{\log(1-\mathrm{e}^{-\nu})}{\log t}. \tag{A.6}$$

By letting $t \to \infty$, we observe that the limit of the right-hand side of inequality (A.6) is $\nu/\log\gamma > 0$, which implies the existence of C_3.

Now we focus on proving the inequality (A.6). For each $k_0 \geq 0$, we have

$$F(t) \;\geq\; \mathbb{P}\left(\sum_{k=k_0+1}^{\infty} U_k < t - \sum_{k=1}^{k_0}\gamma^k\right).$$

If we further require k_0 to satisfy $\sum_{k=1}^{\infty}\lambda^k \leq t - \sum_{k=1}^{k_0}\gamma^k$, then we have then we have

$$\begin{aligned} F(t) \;&\geq\; \mathbb{P}\left(\sum_{k=k_0+1}^{\infty} U_k < \sum_{k=k_0+1}^{\infty}\lambda^k\right) \\ &\geq\; 1-\frac{\mathrm{e}^{-\nu(k_0+1)}}{1-\mathrm{e}^{-\nu}}, \end{aligned}$$

where we omit the same steps before the last inequality as in the inequality (A.4). We finish our proof by choosing $k_0 = k_t$ to satisfy the requirement. □