Direct Search Based on Probabilistic Descent

S. Gratton[*] C. W. Royer[†] L. N. Vicente[‡] Z. Zhang[§]

January 22, 2015

Abstract

Direct-search methods are a class of popular derivative-free algorithms characterized by evaluating the objective function using a step size and a number of (polling) directions. When applied to the minimization of smooth functions, the polling directions are typically taken from positive spanning sets which in turn must have at least $n+1$ vectors in an n-dimensional variable space. In addition, to ensure the global convergence of these algorithms, the positive spanning sets used throughout the iterations are required to be uniformly non-degenerate in the sense of having a positive (cosine) measure bounded away from zero.

However, recent numerical results indicated that randomly generating the polling directions without imposing the positive spanning property can improve the performance of these methods, especially when the number of directions is chosen considerably less than $n+1$.

In this paper, we analyze direct-search algorithms when the polling directions are probabilistic descent, meaning that with a certain probability at least one of them is of descent type. Such a framework enjoys almost-sure global convergence. More interestingly, we will show a global decaying rate of $1/\sqrt{k}$ for the gradient size, with overwhelmingly high probability, matching the corresponding rate for the deterministic versions of the gradient method or of direct search. Our analysis helps to understand numerical behavior and the choice of the number of polling directions.

Keywords: Derivative-free optimization, direct-search methods, polling, positive spanning sets, probabilistic descent, random directions.

1 Introduction

Minimizing a function without using derivatives has been the subject of recent intensive research, as it poses a number of mathematical and numerical challenges and it appears in various

[*] ENSEEIHT, INPT, rue Charles Camichel, B.P. 7122 31071, Toulouse Cedex 7, France (serge.gratton@enseeiht.fr).

[†] ENSEEIHT, INPT, rue Charles Camichel, B.P. 7122 31071, Toulouse Cedex 7, France (clement.royer@enseeiht.fr.) Support for this author comes from a doctoral grant of Université Toulouse III Paul Sabatier.

[‡] CMUC, Department of Mathematics, University of Coimbra, 3001-501 Coimbra, Portugal (lnv@mat.uc.pt, zhang@mat.uc.pt). Support for this research was provided by FCT under grants PTDC/MAT/116736/2010 and PEst-C/MAT/UI0324/2011 and by the Réseau Thématique de Recherche Avancée, Fondation de Coopération Sciences et Technologies pour l'Aéronautique et l'Espace, under the grant ADTAO.

[§] CERFACS-IRIT joint lab, 31000 Toulouse, France (zaikun.zhang@irit.fr). This author works within the framework of the project FILAOS funded by RTRA STAE.

applications of optimization [7]. In a simplified way, there are essentially two paradigms or main approaches to design an algorithm for Derivative-Free Optimization with rigorous (global) convergence properties. One possibility consists of building models based on sampling (typically by quadratic interpolation) for use in a trust-region algorithm, where the accuracy of the models ensures descent for small enough step sizes. The other main possibility is to make use of a number of directions to ensure descent when the size of the step is relatively small, and direct-search methods [7, 20] offer a well studied framework to use multiple directions in a single iteration.

If the objective function is smooth (continuously differentiable), it is well known that at least one of the directions of a positive spanning set (PSS) is descent (and by a PSS we mean a set of directions that spans $\mathbb{R}^n$ with non-negative coefficients). Based on this guarantee of descent, direct-search methods, at each iteration, evaluate the function along the directions in a PSS (a process called polling), using a step size parameter that is reduced when none of the directions leads to some form of decrease. There are a variety of globally convergent methods of this form, depending essentially on the choice of PSS and on the condition to declare an iteration successful (simple or sufficient decrease). Coordinate or compass search, for instance, uses the PSS formed by the $2n$ coordinate directions. Multiple PSSs can be used in direct search, in a finite number when accepting new iterates based on simple decrease (see pattern search [27] and generalized pattern search [3]), or even in an infinite number when accepting new iterates based on sufficient decrease (see generating set search [20]). In any of these cases, the PSSs in question are required to be uniformly non-degenerate in the sense of not being close to loose its defining property. More precisely, their cosine measure is required to be uniformly bounded away from zero, which in turn implies the existence, in any PSS used in a direct-search iteration, of a descent direction uniformly bounded away from being orthogonal to the negative gradient.

If the objective function is non-smooth, say only assumed locally Lipschitz continuous, the polling directions asymptotically used in direct-search run are required, in some normalized form, to densely cover the unit sphere. Mesh adaptive direct search [4] encompasses a process of dense generation that leads to global convergence. Such a process can be significantly simplified if the iterates are only accepted based on a sufficient decrease condition [29]. Anyhow, generating polling directions densely in the unit sphere naturally led to the use of randomly generated PSSs [4, 29].

However, one can take the issue of random generation one step further, and ask the question of whether one can randomly generate a set of directions at each direct-search iteration and use it for polling, without checking if it is a PSS. Moreover, one knows that a PSS must have at least $n+1$ elements in $\mathbb{R}^n$, and thus one even questions the need to generate so many directions and asks how many of them are appropriate to use. We were motivated to address this problem given the obvious connection to the trust-region methods based on probabilistic models recently studied in [5], as we will see later in our paper. Simultaneously, we were surprised by the numerical experiments reported [18] where generating the polling directions free of PSS rules (and possibly in a number less than $n+1$) was indeed beneficial.

Similarly to the notion of probabilistically fully linear model in [5], we then introduce in this paper the companion notion of a probabilistically descent set of directions, by requiring at least one of them to make an acute angle with the negative gradient with a certain probability, uniformly across all iterations. We then analyze what is direct search capable of delivering when the set of polling directions is probabilistically descent. Our algorithmic framework is extremely simple and it can be simplistically reduced here to: at each iteration k, generate a finite set D_k of polling directions that is probabilistically descent; if a $d_k \in D_k$ is found such that

$f(x_k+\alpha_k d_k) < f(x_k)-\rho(\alpha_k)$, where $\rho(\alpha_k) = \alpha_k^2/2$ and α_k is the step size, then $x_{k+1} = x_k+\alpha_k d_k$ and $\alpha_{k+1} = 2\alpha_k$, otherwise $x_{k+1} = x_k$ and $\alpha_{k+1} = \alpha_k/2$. We will then prove that such a scheme enjoys, with overwhelmingly high probability, a gradient decaying rate of $1/\sqrt{k}$ or, equivalently, that the number of iterations taken to reach a gradient of size ϵ is $\mathcal{O}(\epsilon^{-2})$. The rate is global since no assumption is made on the starting point. For this purpose, we first derive a bound, for all realizations of the algorithm, on the number of times the set of polling directions is descent. Such a bound is shown to be a fraction of k when k is larger than the inverse of the square of the minimum norm gradient up to the k-th iteration. When descent is probabilistically conditioned to the past, one can then apply a Chernoff type bound to prove that the probability of the minimum norm gradient decaying with a global rate of $1/\sqrt{k}$ approaches one exponentially. The analysis is carried out for a more general function ρ with the help of an auxiliary function φ which is a multiple of the identity when $\rho(\alpha_k) = \alpha_k^2/2$.

Although such a direct-search framework, where a set of polling directions is independently randomly generated at each iteration, shares similarities with other randomized algorithmic approaches, the differences are significant. Random search methods typically generate a single direction at each iteration, take steps independently of decrease conditions, and update the step size using some deterministic or probabilistic formula. The generated direction can be first premultiplied by an approximation to the directional derivative along the direction itself [25], and it was shown in [22] how to frame this technique using the concept of smoothing to lead to global rates of appropriate order. Methods using multiple points or directions at each iteration, such as evolution strategies (see [11] for a globally convergent version), typically impose some correlation among the direction sets from one iteration to another (as in covariance matrix adaptation [19]) and, again, take action and update step sizes independently of decrease conditions.

More related to our work is the contribution of [10], a derivative-free line search method that uses random searching directions, where at each iteration the step size is defined by a tolerant nonmonotone backtracking line search that always starts from the unitary step. The authors proved global convergence with probability one in [10], but no global rate was provided. In addition, global convergence in [10] is much easier to establish than in our paper, since the backtracking line search always starts from the unitary step, independently of the previous iterations, and consequently the step size is little coupled to the history of the computation.

The organization of the paper is the following. First, we recall in Section 2 the main elements of direct-search methods using PSSs and start a numerical illustration of the issues at stake in this paper. Then, we introduce in Section 3 the notion of probabilistic descent and show how it leads direct search to almost sure global convergence. The main result of our paper is presented in Section 4 where we show that direct search using probabilistic descent conditioned to the past attains the desired global rate with overwhelmingly high probability. In Section 5 we cover a number of related issues, among which how to derive in expectation the main result and what can still be achieved without conditioning to the past. Having already presented our findings, we then revisit the numerical illustrations with additional insight. Section 6 discusses how to extend our analysis to other algorithmic contexts, in particular to trust-region methods based on probabilistic models [5]. The paper is concluded with some final remarks in Section 7.

2 Direct search based on deterministic descent

In this paper we consider the minimization of a function $f : \mathbb{R}^n \to \mathbb{R}$, without imposing any constraints, and assuming that f is smooth, say continuously differentiable. As in [7], we consider first a quite general direct-search algorithmic framework but without a search step.

Algorithm 2.1 (Direct search) *Select $x_0 \in \mathbb{R}^n$, $\alpha_{\max} \in (0, \infty]$, $\alpha_0 \in (0, \alpha_{\max})$, $\theta \in (0, 1)$, $\gamma \in [1, \infty)$, and a forcing function $\rho : (0, \infty) \to (0, \infty)$.*

For each iteration $k = 0, 1, \ldots$

```
    Poll step          Choose a finite set $D_k$ of non-zero vectors. Start evaluating
                       $f$ at the polling points $\{x_k + \alpha_k d : d \in D_k\}$ following a
                       chosen order. If a poll point $x_k + \alpha_k d_k$ is found such that
                       $f(x_k + \alpha_k d_k) < f(x_k) - \rho(\alpha_k)$ then stop polling, set $x_{k+1} =$
                       $x_k + \alpha_k d_k$, and declare the iteration successful. Otherwise
                       declare the iteration unsuccessful and set $x_{k+1} = x_k$.

    Step size update   If the iteration was successful, set $\alpha_{k+1} = \min\{\gamma\alpha_k, \alpha_{\max}\}$.
                       Otherwise, set $\alpha_{k+1} = \theta\alpha_k$.
```

End for

An optional search step could be taken, before the poll one, by testing a finite number of points, looking for an x such that $f(x) < f(x_k) - \rho(\alpha_k)$. If such an x would be found, then one would define $x_{k+1} = x$, declare the iteration successful, and skip the poll step. However, ignoring such a search step allows us to focus on the polling mechanism of Algorithm 2.1, which is the essential part of the algorithm.

2.1 Deterministic descent and positive spanning

The choice of the direction sets D_k in Algorithm 2.1 is a major issue. To guarantee a successful iteration in a finite number of attempts, it is sufficient that at least one of the directions in D_k is a descent one. But global convergence to stationary points requires more as we cannot have all directions in D_k becoming arbitrarily close of being orthogonal to the negative gradient $-\nabla f(x_k) = -g_k$, and so one must have that

$$\forall k, \; \exists d_k \in D_k, \quad \frac{-d_k^\top g_k}{\|d_k\|\|g_k\|} \;\geq\; \kappa \;>\; 0, \tag{1}$$

for some κ that does not depend on k.

On the other hand, it is well known [9] (see also [7, 20]) that if D is a PSS, then

$$\forall v \in \mathbb{R}^n, \quad \exists d \in D, \quad \frac{d^\top v}{\|d\| \, \|v\|} \;>\; 0. \tag{2}$$

One can then see that condition (1) is easily verified if $\{D_k\}$ is chosen from a finite number of PSSs. When passing from a finite to an infinite number of PSSs, some uniform bound must be imposed in (2), which essentially amounts to say that the cosine measure of all the D_k is at least κ, where by the cosine measure [20] of a PSS D with non-zero vectors we mean the positive quantity

$$\mathrm{cm}(D) \;=\; \min_{v\in\mathbb{R}^n\setminus\{0\}} \max_{d\in D} \frac{d^\top v}{\|d\|\|v\|}.$$

One observes that imposing $\mathrm{cm}(D_k) \geq \kappa$ is much stronger than (1), and that it is the lack of knowledge of the negative gradient that leads us to the imposition of such a condition. Had we known $-g_k = -\nabla f(x_k)$, we would have just required $\mathrm{cm}(D_k, -g_k) \geq \kappa$, where $\mathrm{cm}(D, v)$ is the cosine measure of D given v, defined by:

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

To avoid an ill-posed definition when $g_k = 0$, we assume by convention that $\mathrm{cm}(D, v) = 1$ when $v = 0$.

Condition $\mathrm{cm}(D_k, -g_k) \geq \kappa$ is all one needs to guarantee a successful step for a sufficiently small step size α_k. For completeness we show such a result in Lemma 2.1 below, under Assumptions 2.1–2.3, which are assumed here and throughout the paper.

Assumption 2.1 *The objective function f is bounded from below and continuously differentiable in $\mathbb{R}^n$. ∇f is Lipschitz continuous in $\mathbb{R}^n$.*

Let then $f_{\text{low}} > -\infty$ be a lower bound of f, and $\nu > 0$ be a Lipschitz constant of ∇f.

Assumption 2.2 *The forcing function ρ is positive, non-decreasing, and $\rho(\alpha) = o(\alpha)$ when $\alpha \to 0^+$.*

The following function φ will make the algebra conveniently more concise. For each $t > 0$, let

$$\varphi(t) \;=\; \inf\left\{\alpha : \alpha > 0,\ \frac{\rho(\alpha)}{\alpha} + \frac{1}{2}\nu\alpha \;\geq\; t\right\}. \tag{3}$$

It is clear that φ a well defined non-decreasing function. Note that Assumption 2.2 ensures that $\varphi(t) > 0$ when $t > 0$. When $\rho(\alpha) = c\,\alpha^2/2$, one obtains $\varphi(t) = 2t/(c + \nu)$, which is essentially a multiple of the identity.

Assumption 2.3 *For each $k \geq 0$, D_k is a finite set of normalized vectors.*

Assumption 2.3 is made for the sake of simplicity, and it is actually enough to assume that the norms of all the polling directions are uniformly bounded away from zero and infinity.

Lemma 2.1 *The k-th iteration in Algorithm 2.1 is successful if*

$$\mathrm{cm}(D_k, -g_k) \;\geq\; \kappa \quad \textit{and} \quad \alpha_k \;<\; \varphi(\kappa\|g_k\|).$$

Proof. According to the definition of $\mathrm{cm}(D_k, -g_k)$, there exists $d_k^* \in D_k$ satisfying

$$d_k^{*\top} g_k \;=\; -\,\mathrm{cm}(D_k, -g_k)\|d_k^*\|\|g_k\| \;\leq\; -\kappa\|g_k\|.$$

Thus, by Taylor expansion,

$$f(x_k + \alpha_k d_k^*) - f(x_k) \;\le\; \alpha_k d_k^{*\top} g_k + \frac{1}{2}\nu\alpha_k^2 \;\le\; -\kappa\alpha_k\|g_k\| + \frac{1}{2}\nu\alpha_k^2. \tag{4}$$

Using the definition of φ, we obtain from $\alpha_k < \varphi(\kappa\|g_k\|)$ that

$$\frac{\rho(\alpha_k)}{\alpha_k} + \frac{1}{2}\nu\alpha_k \;<\; \kappa\|g_k\|.$$

Hence (4) implies $f(x_k + \alpha_k d_k^*) < f(x_k) - \rho(\alpha_k)$, and thus the k-th iteration is successful. □

2.2 A numerical illustration

A natural question that arises is whether one can gain by generating the vectors in D_k randomly hoping that (1) will then be satisfied frequently enough. For this purpose we ran Algorithm 2.1 in Matlab for different choices of the polling directions. The test problems were taken from `CUTEr` [17] with dimension $n = 40$. We set $\alpha_{\max} = \inf$, $\alpha_0 = 1$, $\theta = 0.5$, $\gamma = 2$, and $\rho(\alpha) = 10^{-3}\alpha^2$. The algorithm terminated when either α_k was below a tolerance of 10^{-10} or a budget of $2000n$ function evaluations was exhausted.

Table 1: Relative performance for different sets of polling directions ($n = 40$).

	$[I\ -I]$	$[Q\ -Q]$	$[Q_k\ -Q_k]$	$2n$	$n+1$	$n/2$	$n/4$	2	1
arglina	3.42	8.44	16.67	10.30	6.01	3.21	1.88	1.00	–
arglinb	20.50	10.35	11.38	7.38	2.81	2.35	1.85	1.00	2.04
broydn3d	4.33	6.55	11.22	6.54	3.59	2.04	1.28	1.00	–
dqrtic	7.16	9.37	19.50	9.10	4.56	2.77	1.70	1.00	–
engval1	10.53	20.89	23.96	11.90	6.48	3.55	2.08	1.00	2.08
freuroth	56.00	6.33	1.33	1.00	1.67	1.33	1.67	1.00	4.00
integreq	16.04	16.29	18.85	12.44	6.76	3.52	2.04	1.00	–
nondquar	6.90	30.23	17.36	7.56	4.23	2.76	1.87	1.00	–
sinquad	–	–	2.12	1.65	2.01	1.26	1.00	1.55	–
vardim	1.00	3.80	3.30	1.80	2.40	2.30	1.80	1.80	4.30

The first three columns of Table 1 correspond to the following PSS choices of the polling directions D_k: $[I\ -I]$ represents the columns of I and $-I$, where I is the identity matrix of size n; $[Q\ -Q]$ means the columns of Q and $-Q$, where Q is an orthogonal matrix obtained by the QR decomposition of a random vector, uniformly distributed on the unit sphere of $\mathbb{R}^n$, generated before the start of the algorithm and then fixed throughout the iterations; $[Q_k\ -Q_k]$ consists of the columns of Q_k and $-Q_k$, where Q_k is a matrix obtained in the same way as Q except that it is generated independently at each iteration. In the cases of $[I\ -I]$ and $[Q\ -Q]$, a cyclic polling procedure was applied to accelerate descent, by starting polling at the direction that led to previous success (if the last iteration was successful) or at the direction right after the last one used (if the last iteration was unsuccessful). In the remaining columns, D_k consists of m ($m = 2n, n+1, n/2, n/4, 2, 1$) independent random vectors uniformly distributed on the unit sphere in $\mathbb{R}^n$, independently generated at every iteration.

We counted the number of function evaluations taken to drive the function value below $f_{\text{low}} + \varepsilon[f(x_0) - f_{\text{low}}]$, where f_{low} is the true minimal value of the objective function f, and ε is a tolerance set to 10^{-3}. Given a test problem, we present in each row the ratio between the number of function evaluations taken by the corresponding version and the number of function evaluations taken by the best version. Because of the random nature of the computations, the number of function evaluations was obtained by averaging over ten independent runs, except for the first column. The symbol '–' indicates that the algorithm failed to solve the problem to the required precision at least once in the ten runs.

When D_k is randomly generated on the unit sphere (columns $m = 2n, n+1, n/2, n/4, 2, 1$ in Table 1), the vectors in D_k do not form a PSS when $m \leq n$, and are not guaranteed to do so when $m > n$. However, we can see clearly from Table 1 that randomly generating the polling directions in this way performed evidently better, despite the fact that their use is not covered by the classical theory of direct search. This evidence motivates us to establish the convergence theory of Algorithm 2.1 when the polling directions are randomly generated.

3 Probabilistic descent and global convergence

From now on, we suppose that the polling directions in Algorithm 2.1 are not defined deterministically but generated by a random process $\{\mathfrak{D}_k\}$. Due to the randomness of $\{\mathfrak{D}_k\}$, the iterates and stepsizes are also random processes (and we denote the random iterates by $\{X_k\}$). The realizations of $\{\mathfrak{D}_k\}$ and $\{X_k\}$ are $\{D_k\}$ and $\{x_k\}$, respectively. We notice, however, that the starting point x_0 (and thus the initial function value $f(x_0)$) and the initial stepsize α_0 are not random.

3.1 Probabilistic descent

The following concept of probabilistically descent sets of polling directions is critical to our analysis. We use it to describe the quality of the random polling directions in Algorithm 2.1. Recalling that $g_k = \nabla f(x_k)$, let G_k be the random variable corresponding to g_k.

Definition 3.1 *The sequence $\{\mathfrak{D}_k\}$ in Algorithm 2.1 is said to be p-*probabilistically *κ-*descent *if*

$$\mathbb{P}\left(\operatorname{cm}(\mathfrak{D}_0, -G_0) \;\geq\; \kappa\right) \;\geq\; p$$

and, for each $k \geq 1$,

$$\mathbb{P}\left(\operatorname{cm}(\mathfrak{D}_k, -G_k) \;\geq\; \kappa \mid \mathfrak{D}_0, \ldots, \mathfrak{D}_{k-1}\right) \;\geq\; p. \tag{5}$$

Definition 3.1 requires the cosine measure $\operatorname{cm}(\mathfrak{D}_k, -G_k)$ to be favorable in a probabilistic sense rather than deterministically. We will see that the role of probabilistically descent sets in the analysis of direct search (based on probabilistic descent) is similar to that of positive spanning sets in the theory of deterministic direct search.

Definition 3.1 is inspired by the definition of probabilistically fully linear models [5]. Inequality (5) involves the notion of conditional probability (see [26, Chapter II]) and says essentially that the probability of the event $\{\operatorname{cm}(\mathfrak{D}_k, -G_k) \geq \kappa\}$ is not smaller than p, no matter what happened with $\mathfrak{D}_0, \ldots, \mathfrak{D}_{k-1}$. It is stronger than assuming merely that $\mathbb{P}\left(\operatorname{cm}(\mathfrak{D}_k, -G_k) \geq \kappa\right) \geq p$.

The analysis in this paper would still hold even if a search step is included in the algorithm and possibly taken. The analysis would consider $\mathfrak{D}_k$ defined at all iterations even if the poll step is skipped.

3.2 Global convergence

The global convergence analysis of direct search based on probabilistic descent follows what was done in [5] for trust-region methods based on probabilistic models, and is no more than a reorganization of the arguments there now applied to direct search when no search step is taken.

It is well known that $\alpha_k \to 0$ in deterministic direct search (based on sufficient decrease as in Algorithm 2.1), no matter how the polling directions are defined [20]. A similar result can be derived by applying exactly the same argument to a realization of direct search now based on probabilistic descent.

Lemma 3.1 *For each realization of Algorithm 2.1,* $\lim_{k\to\infty} \alpha_k = 0$.

For each $k \geq 0$, we now define Y_k as the indicator function of the event

$$\{\text{the } k\text{-th iteration is successful}\}\,.$$

Furthermore, Z_k will be the indicator function of the event

$$\{\mathrm{cm}(\mathfrak{D}_k, -G_k) \geq \kappa\}\,, \tag{6}$$

where, again, κ is a positive constant and independent of the iteration counter. The Bernoulli processes $\{Y_k\}$ and $\{Z_k\}$ will play a major role in our analysis. Their realizations are denoted by $\{y_k\}$ and $\{z_k\}$. Notice, then, that Lemma 2.1 can be restated as follows: given a realization of Algorithm 2.1 and $k \geq 0$, if $\alpha_k < \varphi(\kappa\|g_k\|)$, then $y_k \geq z_k$.

Lemmas 2.1 and 3.1 lead to a critical observation presented below as Lemma 3.2. We observe that such a result holds without any assumption on the probabilistic behavior of $\{\mathfrak{D}_k\}$.

Lemma 3.2 *For the stochastic processes* $\{G_k\}$ *and* $\{Z_k\}$*, where* $G_k = \nabla f(X_k)$ *and* Z_k *is the indicator of the event* (6)*, it holds that*

$$\left\{\liminf_{k\to\infty} \|G_k\| \;>\; 0\right\} \subset \left\{\sum_{k=0}^{\infty} [Z_k \ln\gamma + (1-Z_k)\ln\theta] \;=\; -\infty\right\}. \tag{7}$$

Proof. Consider a realization of Algorithm 2.1 for which $\liminf_{k\to\infty}\|g_k\|$ is not zero but a positive number ϵ. There exists a positive integer k_0 such that for each $k \geq k_0$ it holds $\|g_k\| \geq \epsilon/2$ and $\alpha_k < \varphi(\kappa\epsilon/2)$ (because $\alpha_k \to 0$ and $\varphi(\kappa\epsilon/2) > 0$), and consequently $\alpha_k < \varphi(\kappa\|g_k\|)$. Hence we can obtain from Lemma 2.1 that $y_k \geq z_k$. Additionally, we can assume that k_0 is large enough to ensure $\alpha_k \leq \gamma^{-1}\alpha_{\max}$ for each $k \geq k_0$. Then the stepsize update of Algorithm 2.1 gives us

$$\alpha_k \;=\; \alpha_{k_0}\prod_{l=k_0}^{k-1}\left(\gamma^{y_l}\theta^{1-y_l}\right) \;\geq\; \alpha_{k_0}\prod_{l=k_0}^{k-1}\left(\gamma^{z_l}\theta^{1-z_l}\right)$$

for all $k \geq k_0$. This leads to $\prod_{l=0}^{\infty}\left(\gamma^{z_l}\theta^{1-z_l}\right) = 0$, since $\alpha_{k_0} > 0$ and $\alpha_k \to 0$. Taking logarithms, we conclude that

$$\sum_{l=0}^{\infty}\left[z_l\ln\gamma + (1-z_l)\ln\theta\right] \;=\; -\infty,$$

which completes the proof. □

If $\{\mathfrak{D}_k\}$ is p_0-probabilistically κ-descent, with

$$p_0 \;=\; \frac{\ln\theta}{\ln(\gamma^{-1}\theta)}, \tag{8}$$

then similar to [5, Theorem 4.2], it can be checked that the random process

$$\left\{\sum_{l=0}^{k}\left[Z_l\ln\gamma + (1-Z_l)\ln\theta\right]\right\}$$

is a submartingale with bounded increments. Hence the event on the right-hand side of (7) has probability zero (see [5, Theorem 4.1]). Thereby we obtain the confirmation of global convergence of probabilistic direct search provided below.

Theorem 3.1 *If $\{\mathfrak{D}_k\}$ is p_0-probabilistically κ-descent in Algorithm 2.1, then*

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

We point out that this lim inf-type global convergence result is also implied by our global rate theory of Section 4, under a slightly stronger assumption (see Proposition 5.1).

4 Global rate for direct search based on probabilistic descent

For measuring the global rate of decay of the gradient, we consider the gradient $\tilde{g}_k$ with minimum norm among $g_0, g_1, \ldots, g_k$, and use $\tilde{G}_k$ to represent the corresponding random variable. Given a positive number ϵ, we define k_ϵ as the smallest integer k such that $\|g_k\| \le \epsilon$, and denote the corresponding random variable by K_ϵ. It is easy to check that $k_\epsilon \le k$ if and only if $\|\tilde{g}_k\| \le \epsilon$.

In the deterministic case, one either looks at the global decaying rate of $\|\tilde{g}_k\|$ or one counts the number of iterations (a worst case complexity bound) needed to drive the norm of the gradient below a given tolerance $\epsilon > 0$. When $\rho(\alpha)$ is a positive multiple of α^2, it can be shown [28] that the global rate is of the order of $1/\sqrt{k}$ and that the worst case complexity bound in number of iterations is of the order of ϵ^{-2}. Since the algorithms are now probabilistic, what interests us are lower bounds, as close as possible to one, on the probabilities $\mathbb{P}(\|\tilde{G}_k\| \le \mathcal{O}(1/\sqrt{k}))$ (global rate) and $\mathbb{P}(K_\epsilon \le \mathcal{O}(\epsilon^{-2}))$ (worst case complexity bound).

Our analysis is carried out in two major steps. In Subsection 4.1, we will concentrate on the intrinsic mechanisms of Algorithm 2.1, without assuming any probabilistic property about the polling directions $\{\mathfrak{D}_k\}$, establishing a bound on $\sum_{l=0}^{k-1} z_l$ for all realizations of the algorithm. In Subsection 4.2, we will see how this property can be used to relate $\mathbb{P}(\|\tilde{G}_k\| \le \epsilon)$ to the lower tail of the random variable $\sum_{l=0}^{k-1} Z_l$. We will concentrate on the probabilistic behavior of Algorithm 2.1, using a Chernoff bound for the lower tail of $\sum_{l=0}^{k-1} Z_l$ when $\{\mathfrak{D}_k\}$ is probabilistically descent, and then prove the desirable lower bounds on $\mathbb{P}(\|\tilde{G}_k\| \le \epsilon)$ and $\mathbb{P}(K_\epsilon \le k)$.

In order to achieve our goal, we henceforth make an additional assumption on the forcing function ρ as follows.

Assumption 4.1 *There exist constants $\bar{\theta}$ and $\bar{\gamma}$ satisfying $0 < \bar{\theta} < 1 \le \bar{\gamma}$ such that, for each $\alpha > 0$,*

$$\rho(\theta\alpha) \le \bar{\theta}\rho(\alpha), \quad \rho(\gamma\alpha) \le \bar{\gamma}\rho(\alpha).$$

Such an assumption is not restrictive as it holds in particular for the classical forcing functions of the form $\rho(\alpha) = c\,\alpha^q$, with $c > 0$ and $q > 1$.

4.1 Analysis of step size behavior and number of iterations with descent

The goal of this subsection is to establish a bound on $\sum_{l=0}^{k-1} z_l$ in terms of k and $\|\tilde{g}_k\|$ for an arbitrary realization of Algorithm 2.1, as presented in Lemma 4.2. To do this, we first show the boundedness of $\sum_{k=0}^{\infty} \rho(\alpha_k)$.

Lemma 4.1 *For each realization of Algorithm 2.1,*

$$\sum_{k=0}^{\infty} \rho(\alpha_k) \;\leq\; \frac{\bar{\gamma}}{1-\bar{\theta}} \left[\rho\left(\gamma^{-1}\alpha_0\right) + (f_0 - f_{\mathrm{low}})\right].$$

Proof. Consider a realization of Algorithm 2.1. We assume that there are infinitely many successful iterations as it is trivial to adapt the argument otherwise.

Let k_i be the index of the i-th successful iteration ($i \geq 1$). Define $k_0 = -1$ and $\alpha_{-1} = \gamma^{-1}\alpha_0$ for convenience. Let us rewrite $\sum_{k=0}^{\infty} \rho(\alpha_k)$ as

$$\sum_{k=0}^{\infty} \rho(\alpha_k) \;=\; \sum_{i=0}^{\infty} \sum_{k=k_i+1}^{k_{i+1}} \rho(\alpha_k), \tag{9}$$

and study first $\sum_{k=k_i+1}^{k_{i+1}} \rho(\alpha_k)$. According to Algorithm 2.1 and the definition of k_i, it holds

$$\begin{cases} \alpha_{k+1} \;\leq\; \gamma\alpha_k, & k = k_i, \\ \alpha_{k+1} \;=\; \theta\alpha_k, & k = k_i+1, \ldots, k_{i+1}-1, \end{cases}$$

which gives

$$\alpha_k \;\leq\; \gamma\theta^{k-k_i-1}\alpha_{k_i}, \quad k = k_i+1, \ldots, k_{i+1}.$$

Hence, by the monotonicity of ρ and Assumption 4.1, we have

$$\rho(\alpha_k) \;\leq\; \bar{\gamma}\bar{\theta}^{k-k_i-1}\rho(\alpha_{k_i}), \quad k = k_i+1, \ldots, k_{i+1}.$$

Thus

$$\sum_{k=k_i+1}^{k_{i+1}} \rho(\alpha_k) \;\leq\; \frac{\bar{\gamma}}{1-\bar{\theta}}\rho(\alpha_{k_i}). \tag{10}$$

Inequalities (9) and (10) imply

$$\sum_{k=0}^{\infty} \rho(\alpha_k) \;\leq\; \frac{\bar{\gamma}}{1-\bar{\theta}} \sum_{i=0}^{\infty} \rho(\alpha_{k_i}). \tag{11}$$

Inequality (11) is sufficient to conclude the proof because

$$\alpha_{k_0} \;=\; \gamma^{-1}\alpha_0 \quad \text{and} \quad \sum_{i=1}^{\infty} \rho(\alpha_{k_i}) \;\leq\; f(x_0) - f_{\mathrm{low}},$$

according to the definition of k_i. □

The bound on the sum of the series is denoted by

$$\beta \;=\; \frac{\bar{\gamma}}{1-\bar{\theta}} \left[\rho\left(\gamma^{-1}\alpha_0\right) + f(x_0) - f_{\mathrm{low}}\right]$$

and used next to bound the number of iterations with descent.

Lemma 4.2 *Given a realization of Algorithm 2.1 and a positive integer k,*

$$\sum_{l=0}^{k-1} z_l \;\le\; \frac{\beta}{\rho\left(\min\left\{\gamma^{-1}\alpha_0, \varphi\left(\kappa\left\|\tilde{g}_k\right\|\right)\right\}\right)} + p_0 k.$$

Proof. Consider a realization of Algorithm 2.1. For each $l \in \{0, 1, \ldots, k-1\}$, define

$$v_l \;=\; \begin{cases} 1 & \text{if } \alpha_l \;<\; \min\left\{\gamma^{-1}\alpha_0, \varphi(\kappa\|\tilde{g}_k\|)\right\}, \\ 0 & \text{otherwise.} \end{cases} \tag{12}$$

A key observation for proving the lemma is

$$z_l \;\le\; (1 - v_l) + v_l y_l. \tag{13}$$

When $v_l = 0$, inequality (13) is trivial; when $v_l = 1$, Lemma 2.1 implies that $y_l \ge z_l$ (since $\|\tilde{g}_k\| \le \|\tilde{g}_{k-1}\| \le \|\tilde{g}_l\|$), and hence inequality (13) holds. It suffices then to separately prove

$$\sum_{l=0}^{k-1}(1 - v_l) \;\le\; \frac{\beta}{\rho(\min\{\gamma^{-1}\alpha_0, \varphi(\kappa\|\tilde{g}_k\|)\})} \tag{14}$$

and

$$\sum_{l=0}^{k-1} v_l y_l \;\le\; p_0 k. \tag{15}$$

Because of Lemma 4.1, inequality (14) is justified by the fact that

$$1 - v_l \;\le\; \frac{\rho(\alpha_l)}{\rho(\min\{\gamma^{-1}\alpha_0, \varphi(\kappa\|\tilde{g}_k\|)\})},$$

which in turn is guaranteed by the definition (12) and the monotonicity of ρ.

Now consider inequality (15). If $v_l = 0$ for all $l \in \{0, 1, \ldots, k-1\}$, then (15) holds. Consider then that $v_l = 1$ for some $l \in \{0, 1, \ldots, k-1\}$. Let $\bar{l}$ be the largest one of such integers. Then

$$\sum_{l=0}^{k-1} v_l y_l \;=\; \sum_{l=0}^{\bar{l}} v_l y_l. \tag{16}$$

Let us estimate the sum on the right-hand side. For each $l \in \left\{0, 1, \ldots, \bar{l}\right\}$, Algorithm 2.1 together with the definitions of v_l and y_l give

$$\begin{cases} \alpha_{l+1} \;=\; \min\{\gamma\alpha_l, \alpha_{\max}\} \;=\; \gamma\alpha_l & \text{if } v_l y_l = 1, \\ \alpha_{l+1} \;\ge\; \theta\alpha_l & \text{if } v_l y_l = 0, \end{cases}$$

which implies

$$\alpha_{\bar{l}+1} \;\ge\; \alpha_0 \prod_{l=0}^{\bar{l}} \left(\gamma^{v_l y_l}\theta^{1 - v_l y_l}\right). \tag{17}$$

On the other hand, since $v_{\bar{l}} = 1$, we have $\alpha_{\bar{l}} \leq \gamma^{-1}\alpha_0$, and hence $\alpha_{\bar{l}+1} \leq \alpha_0$. Consequently, by taking logarithms, one can obtain from inequality (17) that

$$0 \;\geq\; \ln(\gamma\theta^{-1}) \sum_{l=0}^{\bar{l}} v_l y_l + (\bar{l}+1)\ln\theta,$$

which leads to

$$\sum_{l=0}^{\bar{l}} v_l y_l \;\leq\; \frac{\ln\theta}{\ln(\gamma^{-1}\theta)}(\bar{l}+1) \;=\; p_0(\bar{l}+1) \;\leq\; p_0 k \tag{18}$$

since $\ln(\gamma^{-1}\theta) < 0$. Inequality (15) is then obtained by combining inequalities (16) and (18). □

Lemma 4.2 allows us to generalize the worst case complexity of deterministic direct search [28] for more general forcing functions when $\gamma > 1$. As we prove such a result we gain also momentum for the derivation of the global rates for direct search based on probabilistic descent.

Proposition 4.1 *Assume that $\gamma > 1$ and consider a realization of Algorithm 2.1. If, for each $k \geq 0$,*

$$\mathrm{cm}(D_k, -g_k) \;\geq\; \kappa \tag{19}$$

and

$$\epsilon \;\leq\; \frac{\gamma}{\kappa\alpha_0}\rho(\gamma^{-1}\alpha_0) + \frac{\nu\alpha_0}{2\kappa\gamma}, \tag{20}$$

then

$$k_\epsilon \;\leq\; \frac{\beta}{(1-p_0)\rho[\varphi(\kappa\epsilon)]}.$$

Proof. By the definition of k_ϵ, we have $\|\tilde{g}_{k_\epsilon - 1}\| \geq \epsilon$. Therefore, from Lemma 4.2 (which holds also with $\|\tilde{g}_k\|$ replaced by $\|\tilde{g}_{k-1}\|$) and the monotonicity of ρ and φ, we obtain

$$\sum_{l=0}^{k_\epsilon - 1} z_l \;\leq\; \frac{\beta}{\rho\left(\min\{\gamma^{-1}\alpha_0, \varphi(\kappa\epsilon)\}\right)} + p_0 k_\epsilon. \tag{21}$$

According to (19), $z_k = 1$ for each $k \geq 0$. By the definition of φ, inequality (20) implies

$$\varphi(\kappa\epsilon) \;\leq\; \varphi\left[\frac{\rho(\gamma^{-1}\alpha_0)}{\gamma^{-1}\alpha_0} + \frac{\nu}{2}\gamma^{-1}\alpha_0\right] \;\leq\; \gamma^{-1}\alpha_0. \tag{22}$$

Hence (21) reduces to

$$k_\epsilon \;\leq\; \frac{\beta}{\rho\left[\varphi(\kappa\epsilon)\right]} + p_0 k_\epsilon.$$

Since $\gamma > 1$, one has, from (8), $p_0 < 1$, and the proof is completed. □

4.2 Global rate (with conditioning to the past)

In this subsection, we will study the probability $\mathbb{P}(\|\tilde{G}_k\| \le \epsilon)$ (and equivalently, $\mathbb{P}(K_\epsilon \le k)$) with the help of Lemma 4.2. First we present a universal lower bound for this probability, which holds without any assumptions on the probabilistic behavior of $\{\mathfrak{D}_k\}$ (Lemma 4.3). Using this bound, we prove that $\mathbb{P}(\|\tilde{G}_k\| \le \mathcal{O}(1/\sqrt{k}))$ and $\mathbb{P}(K_\epsilon \le \mathcal{O}(\epsilon^{-2}))$ are overwhelmingly high when $\rho(\alpha) = c\,\alpha^2/2$ (Corollaries 4.1 and 4.2), which will be given as special cases of the results for general forcing functions (Theorems 4.1 and 4.2).

Lemma 4.3 *If*

$$\epsilon \;\le\; \frac{\gamma}{\kappa\alpha_0}\rho(\gamma^{-1}\alpha_0) + \frac{\nu\alpha_0}{2\kappa\gamma}, \tag{23}$$

then

$$\mathbb{P}\left(\|\tilde{G}_k\| \le \epsilon\right) \;\ge\; 1 - \pi_k\!\left(\frac{\beta}{k\rho[\varphi(\kappa\epsilon)]} + p_0\right), \tag{24}$$

where $\pi_k(\lambda) = \mathbb{P}\left(\sum_{l=0}^{k-1} Z_l \le \lambda k\right)$.

Proof. According to Lemma 4.2 and the monotonicity of ρ and φ, we have

$$\left\{\|\tilde{G}_k\| \ge \epsilon\right\} \;\subset\; \left\{\sum_{l=0}^{k-1} Z_l \le \frac{\beta}{\rho(\min\{\gamma^{-1}\alpha_0, \varphi(\kappa\epsilon)\})} + p_0 k\right\}. \tag{25}$$

Again, as in (22), by the definition of φ, inequality (23) implies $\varphi(\kappa\epsilon) \le \gamma^{-1}\alpha_0$. Thus we rewrite (25) as

$$\left\{\|\tilde{G}_k\| \ge \epsilon\right\} \;\subset\; \left\{\sum_{l=0}^{k-1} Z_l \le \frac{\beta}{\rho[\varphi(\kappa\epsilon)]} + p_0 k\right\},$$

which gives us inequality (24) according to the definition of π_k. □

This lemma enables us to lower bound $\mathbb{P}(\|\tilde{G}_k\| \le \epsilon)$ by just focusing on the function π_k, which is a classical object in probability theory. Various lower bounds can then be established under different assumptions on $\{\mathfrak{D}_k\}$.

Given the assumption that $\{\mathfrak{D}_k\}$ is p-probabilistically κ-descent, Definition 3.1 implies that

$$\mathbb{P}(Z_0 = 1) \;\ge\; p \quad \text{and} \quad \mathbb{P}\left(Z_k = 1 \mid Z_0, \ldots, Z_{k-1}\right) \;\ge\; p \quad (k \ge 1). \tag{26}$$

It is known that the lower tail of $\sum_{l=0}^{k-1} Z_l$ obeys a Chernoff type bound, even when conditioning to the past replaces the more traditional assumption of independence of the Z_k's (see, for instance, [15, Problem 1.7] and [13, Lemma 1.18]). We present such a bound in Lemma 4.4 below and give a proof in Appendix A, where it can be seen the role of the concept of probabilistic descent.

Lemma 4.4 *Suppose that* $\{\mathfrak{D}_k\}$ *is* p*-probabilistically* κ*-descent and* $\lambda \in (0, p)$*. Then*

$$\pi_k(\lambda) \;\le\; \exp\left[-\frac{(p-\lambda)^2}{2p}k\right]. \tag{27}$$

Now we are ready to present the main results of this section. In these results we will assume that $\{\mathfrak{D}_k\}$ is p-probabilistically κ-descent with $p > p_0$, which cannot be fulfilled unless $\gamma > 1$ (since $p_0 = 1$ if $\gamma = 1$).

Theorem 4.1 *Suppose that $\{\mathfrak{D}_k\}$ is p-probabilistically κ-descent with $p > p_0$ and that*

$$k \;\geq\; \frac{(1+\delta)\beta}{(p-p_0)\rho[\varphi(\kappa\epsilon)]}, \tag{28}$$

for some positive number δ, and where ϵ satisfies

$$\epsilon \;\leq\; \frac{\gamma}{\kappa\alpha_0}\rho(\gamma^{-1}\alpha_0) + \frac{\nu\alpha_0}{2\kappa\gamma}. \tag{29}$$

Then

$$\mathbb{P}\left(\|\tilde{G}_k\| \leq \epsilon\right) \;\geq\; 1 - \exp\left[-\frac{(p-p_0)^2\delta^2}{2p(1+\delta)^2}k\right]. \tag{30}$$

Proof. According to Lemma 4.3 and the monotonicity of π_k,

$$\mathbb{P}\left(\|\tilde{G}_k\| \leq \epsilon\right) \;\geq\; 1 - \pi_k\left(\frac{\beta}{k\rho[\varphi(\kappa\epsilon)]} + p_0\right) \;\geq\; 1 - \pi_k\left(\frac{p-p_0}{1+\delta} + p_0\right).$$

Then inequality (30) follows directly from Lemma 4.4. □

Theorem 4.1 reveals that, when ϵ is an arbitrary number but small enough to verify (29) and the iteration counter satisfies (28) for some positive number δ, the norm of the gradient is below ϵ with overwhelmingly high probability. Note that ϵ is related to k through (28). In fact, if $\rho \circ \varphi$ is invertible (which is true when the forcing function is a multiple of α^q with $q > 1$), then, for any positive integer k, sufficiently large to satisfy (28) and (29) all together, one can set

$$\epsilon \;=\; \frac{1}{\kappa}(\rho\circ\varphi)^{-1}\left(\frac{(1+\delta)\beta}{p-p_0}\frac{1}{k}\right), \tag{31}$$

and what we have in (30) can then be read as (since ϵ and k satisfy the assumptions of Theorem 4.1)

$$\mathbb{P}\left(\|\tilde{G}_k\| \leq \kappa^{-1}(\rho\circ\varphi)^{-1}(\mathcal{O}(1/k))\right) \;\geq\; 1 - \exp(-Ck),$$

with C a positive constant.

A particular case is when the forcing function is a multiple of the square of the step size

$$\rho(\alpha) \;=\; \frac{1}{2}c\,\alpha^2,$$

where it is easy to check that

$$\varphi(t) \;=\; \frac{2t}{c+\nu}.$$

One can then obtain the global rate $\|\tilde{G}_k\| \leq \mathcal{O}(1/\sqrt{k})$ with overwhelmingly high probability, matching [28] for deterministic direct search, as it is shown below in Corollary 4.1, which is just an application of Theorem 4.1 for this particular forcing function. For simplicity, we will set $\delta = 1$.

Corollary 4.1 *Suppose that $\{\mathfrak{D}_k\}$ is p-probabilistically κ-descent with $p > p_0$, $\rho(\alpha) = c\,\alpha^2/2$, and*

$$k \;\geq\; \frac{4\gamma^2\beta}{c(p-p_0)\alpha_0^2}. \tag{32}$$

Then

$$\mathbb{P}\left(\|\tilde{G}_k\| \leq \left(\frac{\beta^{\frac{1}{2}}(c+\nu)}{c^{\frac{1}{2}}(p-p_0)^{\frac{1}{2}}\kappa}\right)\frac{1}{\sqrt{k}}\right) \;\geq\; 1 - \exp\left[-\frac{(p-p_0)^2}{8p}k\right]. \tag{33}$$

Proof. As in (31), with $\delta = 1$, let

$$\epsilon \;=\; \frac{1}{\kappa}(\rho\circ\varphi)^{-1}\left(\frac{2\beta}{p-p_0}\frac{1}{k}\right). \tag{34}$$

Then, by straightforward calculations, we have

$$\epsilon \;=\; \left(\frac{\beta^{\frac{1}{2}}(c+\nu)}{c^{\frac{1}{2}}(p-p_0)^{\frac{1}{2}}\kappa}\right)\frac{1}{\sqrt{k}}. \tag{35}$$

Moreover, inequality (32) gives us

$$\epsilon \;\leq\; \frac{\beta^{\frac{1}{2}}(c+\nu)}{c^{\frac{1}{2}}(p-p_0)^{\frac{1}{2}}\kappa}\left(\frac{4\gamma^2\beta}{c(p-p_0)\alpha_0^2}\right)^{-\frac{1}{2}} \;=\; \frac{(c+\nu)\alpha_0}{2\kappa\gamma}. \tag{36}$$

Definition (34) and inequality (36) guarantee that k and ϵ satisfy (28) and (29) for $\rho(\alpha) = c\,\alpha^2/2$ and $\delta = 1$. Hence we can plug (35) into (30) yielding (33). □

Based on Lemmas 4.3 and 4.4 (or directly on Theorem 4.1), one can lower bound $\mathbb{P}(K_\epsilon \leq k)$ and arrive at a worst case complexity result.

Theorem 4.2 *Suppose that $\{\mathfrak{D}_k\}$ is p-probabilistically κ-descent with $p > p_0$ and*

$$\epsilon \;\leq\; \frac{\gamma}{\kappa\alpha_0}\rho(\gamma^{-1}\alpha_0) + \frac{\nu\alpha_0}{2\kappa\gamma}.$$

Then, for each $\delta > 0$,

$$\mathbb{P}\left(K_\epsilon \leq \left\lceil\frac{(1+\delta)\beta}{(p-p_0)\rho[\varphi(\kappa\epsilon)]}\right\rceil\right) \;\geq\; 1 - \exp\left[-\frac{\beta(p-p_0)\delta^2}{2p(1+\delta)\rho[\varphi(\kappa\epsilon)]}\right]. \tag{37}$$

Proof. Letting

$$k \;=\; \left\lceil\frac{(1+\delta)\beta}{(p-p_0)\rho[\varphi(\kappa\epsilon)]}\right\rceil$$

we have $\mathbb{P}(K_\epsilon \leq k) = \mathbb{P}(\|\tilde{G}_k\| \leq \epsilon)$ and then inequality (37) follows from Theorem 4.1 as

$$k \;\geq\; \frac{(1+\delta)\beta}{(p-p_0)\rho[\varphi(\kappa\epsilon)]}.$$

□

Since $\rho(\alpha) = o(\alpha)$ (from the definition of the forcing function ρ) and $\varphi(\kappa\epsilon) \leq 2\nu^{-1}\kappa\epsilon$ (from definition (3) of φ), it holds that the lower bound in (37) goes to one faster than $1 - \exp(-\epsilon^{-1})$. Hence, we conclude from Theorem 4.2 that direct search based on probabilistic descent exhibits a worst case complexity bound in number of iterations of the order of $1/\rho[\varphi(\kappa\epsilon)]$ with overwhelmingly high probability, matching Proposition 4.1 for deterministic direct search. To see this effect more clearly, we present in Corollary 4.2 the particularization of Theorem 4.2 when $\rho(\alpha) = c\,\alpha^2/2$, taking $\delta = 1$ as in Corollary 4.1. The worst case complexity bound is then of the order of $1/\epsilon^2$ with overwhelmingly high probability, matching [28] for deterministic direct search. The same matching happens when $\rho(\alpha)$ is a power of α with exponent q, where the bound is $\mathcal{O}(\epsilon^{-\frac{q}{\min\{q-1,1\}}})$.

Corollary 4.2 *Suppose that $\{\mathfrak{D}_k\}$ is p-probabilistically κ-descent with $p > p_0$, $\rho(\alpha) = c\,\alpha^2/2$, and*

$$\epsilon \leq \frac{(c+\nu)\alpha_0}{2\kappa\gamma}.$$

Then

$$\mathbb{P}\left(K_\epsilon \leq \left\lceil \frac{\beta(c+\nu)^2}{c(p-p_0)\kappa^2}\epsilon^{-2}\right\rceil\right) \geq 1 - \exp\left[-\frac{\beta(p-p_0)(c+\nu)^2}{8cp\kappa^2}\epsilon^{-2}\right]. \tag{38}$$

It is important to understand how the worst case complexity bound (38) depends on the dimension n of the problem. For this purpose, we first need to make explicit the dependence on n of the constant in $K_\epsilon \leq \lceil \beta(c+\nu)^2/[c(p-p_0)\kappa^2]\epsilon^{-2}\rceil$ which can only come from p and κ and is related to the choice of $\mathfrak{D}_k$.

One can choose $\mathfrak{D}_k$ as m directions uniformly independently distributed on the unit sphere, with m independent of n, in which case p is a constant larger than p_0 and $\kappa = \tau/\sqrt{n}$ for some constant $\tau > 0$ (both p and τ are totally determined by γ and θ without dependence on m or n; see Corollary B.1 in Appendix B and the remarks after it). In such a case, from Corollary 4.2,

$$\mathbb{P}\left(K_\epsilon \leq \left\lceil \frac{\beta(c+\nu)^2}{c(p-p_0)\tau^2}\left(n\epsilon^{-2}\right)\right\rceil\right) \geq 1 - \exp\left[-\frac{\beta(p-p_0)(c+\nu)^2}{8cp\kappa^2}\epsilon^{-2}\right].$$

To derive a worst case complexity bound in terms of the number of function evaluations, one just needs then to see that each iteration of Algorithm 2.1 costs at most m function evaluations. Thus, if K_ϵ^f represents the number of function evaluations within K_ϵ iterations, we obtain

$$\mathbb{P}\left(K_\epsilon^f \leq \left\lceil \frac{\beta(c+\nu)^2}{c(p-p_0)\tau^2}\left(n\epsilon^{-2}\right)\right\rceil m\right) \geq 1 - \exp\left[-\frac{\beta(p-p_0)(c+\nu)^2}{8cp\kappa^2}\epsilon^{-2}\right]. \tag{39}$$

The worst case complexity bound is then $\mathcal{O}(mn\epsilon^{-2})$ with overwhelmingly high probability, which is clearly better than the corresponding bound $\mathcal{O}(n^2\epsilon^{-2})$ for deterministic direct search [28] if m is chosen an order of magnitude smaller than n.

4.3 High probability iteration complexity

Given a confidence level P, the following theorem presents an explicit bound for the number of iterations which can guarantee that $\|\tilde{G}_k\| \leq \epsilon$ holds with probability at least P. Bounds of this type are interesting in practice and have been considered in the theoretical analysis of probabilistic algorithms (see, for instance, [23, 24]).

Theorem 4.3 *Suppose that $\{\mathfrak{D}_k\}$ is p-probabilistically κ-descent with $p > p_0$. Then for any*

$$\epsilon \;\le\; \frac{\gamma}{\kappa\alpha_0}\rho(\gamma^{-1}\alpha_0) + \frac{\nu\alpha_0}{2\kappa\gamma}$$

and $P \in (0,1)$, it holds $\mathbb{P}(\|\tilde{G}_k\| \le \epsilon) \ge P$ whenever

$$k \;\ge\; \frac{3\beta}{2(p-p_0)\rho[\varphi(\kappa\epsilon)]} - \frac{3p\ln(1-P)}{(p-p_0)^2}. \tag{40}$$

Proof. By Theorem 4.1, $\mathbb{P}(\|\tilde{G}_k\| \le \epsilon) \ge P$ is achieved when

$$k \;\ge\; \max\left\{\frac{(1+\delta)\beta}{(p-p_0)\rho[\varphi(\kappa\epsilon)]}, -\frac{2p(1+\delta)^2\ln(1-P)}{\delta^2(p-p_0)^2}\right\} \tag{41}$$

for some positive number δ. Hence it suffices to show that the right-hand side of (40) is bigger than that of (41) for properly chosen δ. For simplicity, denote

$$c_1 \;=\; \frac{\beta}{(p-p_0)\rho[\varphi(\kappa\epsilon)]}, \quad c_2 \;=\; -\frac{2p\ln(1-P)}{(p-p_0)^2}.$$

Let us consider the positive number δ such that

$$(1+\delta)c_1 \;=\; \frac{(1+\delta)^2}{\delta^2}c_2.$$

It is easy to check that

$$\delta \;=\; \frac{1}{2c_1}\left(c_2 + \sqrt{c_2^2 + 4c_1c_2}\right) \;\le\; \frac{1}{2} + \frac{3c_2}{2c_1}.$$

Thus

$$\max\left\{(1+\delta)c_1, \frac{(1+\delta)^2}{\delta^2}c_2\right\} \;=\; (1+\delta)c_1 \;\le\; \frac{3}{2}(c_1+c_2),$$

which completes the proof. □

5 Discussion and extensions

5.1 Establishing global convergence from global rate

It is interesting to notice that Theorem 4.1 implies a form of global convergence of direct search based on probabilistic descent, as we mentioned at the end of Subsection 3.2.

Proposition 5.1 *Suppose that $\{\mathfrak{D}_k\}$ is p-probabilistically κ-descent with $p > p_0$. Then*

$$\mathbb{P}\left(\inf_{k\ge 0}\|G_k\| = 0\right) \;=\; 1.$$

Proof. We prove the result by contradiction. Suppose that

$$\mathbb{P}\left(\inf_{k\geq 0}\|G_k\| > 0\right) \;>\; 0.$$

Then there exists a positive constant $\epsilon > 0$ satisfying (29) and

$$\mathbb{P}\left(\inf_{k\geq 0}\|G_k\| \geq \epsilon\right) \;>\; 0. \tag{42}$$

But Theorem 4.1 implies that

$$\lim_{l\to\infty}\mathbb{P}\left(\|\tilde{G}_l\| \geq \epsilon\right) \;=\; 0,$$

which then contradicts (42), because $\mathbb{P}(\|\tilde{G}_l\| \geq \epsilon) \geq \mathbb{P}\left(\inf_{k\geq 0}\|G_k\| \geq \epsilon\right)$ for each $l \geq 0$. □

If we assume for all realizations of Algorithm 2.1 that the iterates never arrive at a stationary point in a finite number of iterations, then the events $\{\liminf_{k\to\infty}\|G_k\| = 0\}$ and $\{\inf_{k\geq 0}\|G_k\| = 0\}$ are identical. Thus, in such a case, Proposition 5.1 reveals that

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

if $\{\mathfrak{D}_k\}$ is p-probabilistically κ-descent with $p > p_0$.

5.2 The behavior of expected minimum norm gradient

In this subsection we study how $\mathbb{E}(\|\tilde{G}_k\|)$ behaves with the iteration counter k. For simplicity, we consider the special case of the forcing function $\rho(\alpha) = c\,\alpha^2/2$.

Proposition 5.2 *Suppose $\{\mathfrak{D}_k\}$ is p-probabilistically κ-descent with $p > p_0$, $\rho(\alpha) = c\,\alpha^2/2$, and*

$$k \;\geq\; \frac{4\gamma^2\beta}{c(p-p_0)\alpha_0^2}.$$

Then

$$\mathbb{E}\left(\|\tilde{G}_k\|\right) \;\leq\; c_3 k^{-\frac{1}{2}} + \|g_0\|\exp\left(-c_4 k\right),$$

where

$$c_3 \;=\; \frac{\beta^{\frac{1}{2}}(c+\nu)}{c^{\frac{1}{2}}(p-p_0)^{\frac{1}{2}}\kappa}, \qquad c_4 \;=\; \frac{(p-p_0)^2}{8p}.$$

Proof. Let us define a random variable H_k as

$$H_k \;=\; \begin{cases} c_3 k^{-\frac{1}{2}} & \text{if } \|\tilde{G}_k\| \leq c_3 k^{-\frac{1}{2}}, \\ \|g_0\| & \text{otherwise.} \end{cases}$$

Then $\|\tilde{G}_k\| \leq H_k$, and hence

$$\mathbb{E}\left(\|\tilde{G}_k\|\right) \;\leq\; \mathbb{E}\left(H_k\right) \;\leq\; c_3 k^{-\frac{1}{2}} + \|g_0\|\mathbb{P}\left(\|\tilde{G}_k\| > c_3 k^{-\frac{1}{2}}\right).$$

Therefore it suffices to notice

$$\mathbb{P}\left(\|\tilde{G}_k\| > c_3 k^{-\frac{1}{2}}\right) \;\leq\; \exp(-c_4 k),$$

which is a straightforward application of Corollary 4.1. □

In deterministic direct search, when a forcing function $\rho(\alpha) = c\,\alpha^2/2$ is used, $\|\tilde{g}_k\|$ decays with $\mathcal{O}(k^{-\frac{1}{2}})$ when k tends to infinity [28]. Proposition 5.2 shows that $\mathbb{E}(\|\tilde{G}_k\|)$ behaves in a similar way in direct search based on probabilistic search.

5.3 Global rate (without conditioning to the past)

Assuming that $\{\mathfrak{D}_k\}$ is p-probabilistically κ-descent as defined in Definition 3.1, we obtained the Chernoff-type bound (27) for π_k, and then established the worst case complexity of Algorithm 2.1 when $p > p_0$. As mentioned before, inequality (5) in Definition 3.1 is stronger than the one without conditioning to the past. A natural question is then to see what can be obtained if we weaken this requirement by not conditioning to the past. It turns out that we can still establish some lower bound for $\mathbb{P}(\|\tilde{G}_k\| \leq \epsilon)$, but much weaker than inequality (30) obtained in Theorem 4.1 and in particular not approaching one.

Proposition 5.3 *If*

$$\mathbb{P}\left(\mathrm{cm}(\mathfrak{D}_k, -G_k) \geq \kappa\right) \;\geq\; p \tag{43}$$

for each $k \geq 0$ *and*

$$\epsilon \;\leq\; \frac{\gamma}{\kappa\alpha_0}\rho(\gamma^{-1}\alpha_0) + \frac{\nu\alpha_0}{2\kappa\gamma},$$

then

$$\mathbb{P}\left(\|\tilde{G}_k\| \leq \epsilon\right) \;\geq\; \frac{p-p_0}{1-p_0} - \frac{\beta}{(1-p_0)k\rho[\varphi(\kappa\epsilon)]}.$$

Proof. Due to Lemma 4.3, it suffices to show that

$$\pi_k\left(\frac{\beta}{k\rho[\varphi(\kappa\epsilon)]} + p_0\right) \;\leq\; \frac{(1-p) + \beta/(k\rho[\varphi(\kappa\epsilon)])}{1-p_0}. \tag{44}$$

Let us study $\pi_k(\lambda)$. According to (43) and to the definition of Z_k,

$$\mathbb{P}\left(Z_k = 1\right) \;\geq\; p$$

for each $k \geq 0$. Hence

$$\pi_k(\lambda) \;=\; \mathbb{P}\left(\sum_{l=0}^{k-1} Z_l \;\leq\; \lambda k\right) \;\leq\; \frac{\mathbb{E}\left(\sum_{l=0}^{k-1}(1-Z_l)\right)}{k-\lambda k} \;\leq\; \frac{k-pk}{k-\lambda k} \;=\; \frac{1-p}{1-\lambda}.$$

Thus

$$\pi_k\left(\frac{\beta}{k\rho[\varphi(\kappa\epsilon)]} + p_0\right) \;\leq\; \frac{1-p}{(1-p_0) - \beta/(k\rho[\varphi(\kappa\epsilon)])}. \tag{45}$$

If $\beta/(k\rho[\varphi(\kappa\epsilon)]) \leq p - p_0$, then (44) follows from (45) and the fact that $a/b \leq (a+c)/(b+c)$ for $0 < a \leq b$ and $c \geq 0$. If $\beta/(k\rho[\varphi(\kappa\epsilon)]) > p - p_0$, then (44) is trivial since $\pi_k(\lambda) \leq 1$. □

5.4 A more detailed look at the numerical experiments

An understanding of what is at stake in the global analysis of direct search based on probabilistic descent allows us to revisit the numerical experiments of Subsection 2.2 with additional insight. Remember that we tested Algorithm 2.1 by choosing $\mathfrak{D}_k$ as m independent random vectors uniformly distributed on the unit sphere in $\mathbb{R}^n$. In such a case, the almost-sure global convergence of Algorithm 2.1 is guaranteed as long as $m > \log_2[1-(\ln\theta)/(\ln\gamma)]$, as it can be concluded from Theorem 3.1 (see Corollary B.1 in Appendix B). For example, when $\gamma = 2$ and $\theta = 0.5$, the algorithm converges with probability 1 when $m \geq 2$, even if such values of m are much smaller than the number of elements of the positive spanning sets with smallest cardinality in $\mathbb{R}^n$, which is $n+1$ (41 in the case tested in Subsection 2.2).

We point out here that our theory is applicable only if $\gamma > 1$. Moreover, for deterministic direct search based on positive spanning sets (PSSs), setting $\gamma = 1$ tends to lead to better numerical performance (see, for instance, [8]). In this sense, the experiment in Subsection 2.2 is biased in favor of direct search based on probabilistic descent. To be fairer, we designed a new experiment by keeping $\gamma = 1$ when the sets of polling directions are guaranteed PSSs (which is true for the versions corresponding to $[I\ \ -I]$, $[Q\ \ -Q]$, and $[Q_k\ \ -Q_k]$), while setting $\gamma > 1$ for direct search based on probabilistic descent. All the other parameters were selected as in Subsection 2.2.

In the case of direct search based on probabilistic descent, we pick now $\gamma = 2$ and $\gamma = 1.1$ as illustrations, and as for the cardinality m of $\mathfrak{D}_k$ we simply take the smallest integers satisfying $m > \log_2[1-(\ln\theta)/(\ln\gamma)]$, which are 2 and 4 respectively. Table 2 presents the results of the redesigned experiment with $n = 40$. Table 3 shows what happened for $n = 100$. The data is organized in the same way as in Table 1.

We can see from the tables that direct search based on probabilistic descent still outperforms (for these problems) the direct-search versions using PSSs, even though the difference is not so considerable as in Table 1. We note that such an effect is even more visible when the dimension is higher ($n = 100$), which is somehow in agreement with the fact that (39) reveals a worst case complexity in function evaluations of $\mathcal{O}(mn\epsilon^{-2})$, which is more favorable than $\mathcal{O}(n^2\epsilon^{-2})$ for deterministic direct search based on PSSs when m is significantly smaller than n.

Table 2: Relative performance for different sets of polling directions ($n = 40$).

	$[I\ \ -I]$	$[Q\ \ -Q]$	$[Q_k\ \ -Q_k]$	2 ($\gamma = 2$)	4 ($\gamma = 1.1$)
arglina	1.00	3.17	37.19	5.86	6.73
arglinb	34.12	5.34	32.56	1.00	2.02
broydn3d	1.00	1.91	5.96	2.04	3.47
dqrtic	1.18	1.36	28.32	1.00	1.48
engval1	1.05	1.00	16.44	2.29	2.89
freuroth	17.74	7.39	7.48	1.35	1.00
integreq	1.54	1.49	5.36	1.00	1.34
nondquar	1.00	2.82	8.02	1.37	1.73
sinquad	–	1.26	–	1.00	–
vardim	20.31	11.02	2.97	1.00	1.84

Table 3: Relative performance for different sets of polling directions ($n = 100$).

	$[I \ -I]$	$[Q \ -Q]$	$[Q_k \ -Q_k]$	2 ($\gamma = 2$)	4 ($\gamma = 1.1$)
arglina	1.00	3.86	105.50	5.86	7.58
arglinb	138.28	107.32	106.23	1.00	1.99
broydn3d	1.00	2.57	12.07	1.92	3.21
dqrtic	3.01	3.25	–	1.00	1.46
engval1	1.04	1.00	43.00	2.06	2.84
freuroth	31.94	17.72	12.42	1.36	1.00
integreq	1.83	1.66	13.46	1.00	1.22
nondquar	1.18	2.83	23.15	1.00	1.17
sinquad	–	–	–	–	–
vardim	112.22	19.72	8.04	1.00	2.36

6 Other algorithmic contexts

Our analysis is wide enough to be carried out to other algorithmic contexts where no derivatives are used and some randomization is applied to probabilistically induce descent. In particular, it will also render a global rate of $1/\sqrt{k}$ with overwhelmingly high probability for trust-region methods based on probabilistic models [5]. We recall that Lemma 4.2 is the key result leading to the global rate results in Subsection 4.2. Lemma 4.2 was based only on the two following elements about a realization of Algorithm 2.1:

1. if the k-th iteration is successful, then $f(x_k) - f(x_{k+1}) \geq \rho(\alpha_k)$ (in which case α_k is increased), and

2. if $\mathrm{cm}(D_k, -g_k) \geq \kappa$ and $\alpha_k < \varphi(\kappa\|g_k\|)$, then the k-th iteration is successful (Lemma 2.1).

One can easily identify similar elements in a realization of the trust-region algorithm proposed in [5, Algorithm 3.1]. In fact, using the notations in [5] and denoting

$$\mathcal{K} \;=\; \{k \in \mathbb{N} : \rho_k \geq \eta_1 \text{ and } \|g_k\| \geq \eta_2 \delta_k\}\,, \tag{46}$$

one can find positive constants μ_1 and μ_2 such that:

1. if $k \in \mathcal{K}$, then $f(x_k) - f(x_{k+1}) \geq \mu_1 \delta_k^2$ and δ_k is increased (by [5, Algorithm 3.1]), and

2. if m_k is $(\kappa_{eg}, \kappa_{ef})$-fully linear and $\delta_k < \mu_2\|g_k\|$, then $k \in \mathcal{K}$ (by [5, Lemma 3.2]).

One sees that δ_k and $\mathcal{K}$ play the same roles as α_k and the set of successful iterations in our paper. Based on these facts, one can first follow Lemma 4.1 to obtain a uniform upper bound μ_3 for $\sum_{k=0}^{\infty} \delta_k^2$, and then prove that $(\kappa_{eg}, \kappa_{ef})$-fully linear models appear at most

$$\frac{\mu_3}{\min\left\{\gamma^{-2}\delta_0^2, \mu_2^2\|\tilde{g}_k\|^2\right\}} + \frac{k}{2} \tag{47}$$

times in the first k iterations, a conclusion similar to Lemma 4.2. It is then straightforward to mimic the analysis in Subsection 4.2, and prove a $1/\sqrt{k}$ rate for the gradient (with overwhelmingly high probability) if the random models are probabilistically fully linear (see [5, Definition 3.2]).

7 Concluding remarks

We introduced a new proof technique for establishing global rates or worst case complexity bounds for randomized algorithms where the choice of the new iterate depends on a decrease condition and where the quality of the object (directions, models) used for descent is favorable with a certain probability. The proof technique separates the counting of the number of iterations that descent holds from the probabilistic properties of such a number.

In the theory of direct search for smooth functions, the polling directions are typically required to be uniformly non-degenerated positive spanning sets. However, numerical observation suggests that the performance of direct search can be improved by instead randomly generating the polling directions. To understand this phenomenon, we proposed the concept of probabilistically descent sets of directions and studied direct search based upon it. An argument inspired by [5] shows that direct search (Algorithm 2.1) converges almost surely if the sets of polling directions are p_0-probabilistically κ-descent for some $\kappa > 0$, where $p_0 = (\ln\theta)/[\ln(\gamma^{-1}\theta)]$, γ and θ being the expanding and contracting parameters of the step size. But more insightfully, we established in this paper the global rate and worst case complexity of direct search when the sets of polling directions are p-probabilistically κ-descent for some $p > p_0$ and $\kappa > 0$. It was proved that in such a situation direct search enjoys the global rate and worst case complexity of deterministic direct search with overwhelmingly high probability. In particular, when the forcing function is $\rho(\alpha) = c\,\alpha^2/2$ ($c > 0$), the norm of the gradient is driven under $\mathcal{O}(1/\sqrt{k})$ in k iterations with a probability that tends to 1 exponentially when $k \to \infty$, or equivalently, the norm is below ϵ in $\mathcal{O}(\epsilon^{-2})$ iterations with a probability that tends to 1 exponentially when $\epsilon \to 0$. Based on these conclusions, we also showed that the expected minimum norm gradient decays with a rate of $1/\sqrt{k}$, which matches the behavior of the gradient norm in deterministic direct search or steepest descent.

A more precise output of this paper is then a worst case complexity bound of $\mathcal{O}(mn\epsilon^{-2})$, in terms of functions evaluations, for this class of methods when m is the number of (independently uniformly distributed) random directions used for polling. As said above, such a bound does not hold deterministically but under a probability approaching 1 exponentially when $\epsilon \to 0$, but this is still clearly better than $\mathcal{O}(n^2\epsilon^{-2})$ (known for deterministic direct search) in a serial environment where m is chosen significantly smaller than n.

Open issues

When we were finishing this paper, Professor Nick Trefethen brought to our attention the possibility of setting $\mathfrak{D}_k = \{\mathfrak{d}, -\mathfrak{d}\}$, where $\mathfrak{d}$ is a random vector independent of the previous iterations and uniformly distributed on the unit sphere of $\mathbb{R}^n$. Given a unit vector $v \in \mathbb{R}^n$ and a number $\kappa \in [0,1]$, it is easy to see that the event $\{\mathrm{cm}(\mathfrak{D}_k, v) \geq \kappa\}$ is the union of $\{\mathfrak{d}^\top v \geq \kappa\}$ and $\{-\mathfrak{d}^\top v \geq \kappa\}$, whose intersection has probability zero, and therefore

$$\mathbb{P}\big(\,\mathrm{cm}(\mathfrak{D}_k, v) \geq \kappa\big) \;=\; \mathbb{P}\big(\mathfrak{d}^\top v \geq \kappa\big) + \mathbb{P}\big(-\mathfrak{d}^\top v \geq \kappa\big) \;=\; 2\varrho,$$

ϱ being the probability of $\{\mathfrak{d}^\top v \geq \kappa\}$. By the same argument as in the proof of Proposition B.1, one then sees that $\{\mathfrak{D}_k\}$ defined in this way is 2ϱ-probabilistically κ-descent. Given any constants γ and θ satisfying $0 < \theta < 1 < \gamma$, we can pick $\kappa > 0$ sufficiently small so that $2\varrho > (\ln\theta)/[\ln(\gamma^{-1}\theta)]$, and then Algorithm 2.1 conforms to the theory presented in Subsection 3.2 and Section 4. Moreover, the set $\{\mathfrak{d}, -\mathfrak{d}\}$ turns out to be *optimal* among all the sets $\mathfrak{D}$ consisting of 2 random vectors uniformly distributed on the unit sphere, in the sense that it maximizes the probability $\mathbb{P}\big(\operatorname{cm}(\mathfrak{D}, v) \geq \kappa\big)$ for each $\kappa \in [0, 1]$. In fact, if $\mathfrak{D} = \{\mathfrak{d}_1, \mathfrak{d}_2\}$ with $\mathfrak{d}_1$ and $\mathfrak{d}_2$ uniformly distributed on the unit sphere, then

$$\mathbb{P}\big(\operatorname{cm}(\mathfrak{D}, v) \geq \kappa\big) \;=\; 2\varrho - \mathbb{P}\big(\{\mathfrak{d}_1^\top v \geq \kappa\} \cap \{\mathfrak{d}_2^\top v \geq \kappa\}\big) \;\leq\; 2\varrho,$$

and the maximal value 2ϱ is attained when $\mathfrak{D} = \{\mathfrak{d}, -\mathfrak{d}\}$ as already discussed. We tested the set $\{\mathfrak{d}, -\mathfrak{d}\}$ numerically (with $\gamma = 2$ and $\theta = 1/2$), and it performed even better than the set of 2 independent vectors uniformly distributed on the unit sphere (yet the difference was not substantial), which illustrates again our theory of direct search based on probabilistic descent. It remains to know how to define $\mathfrak{D}$ so that it maximizes $\mathbb{P}\big(\operatorname{cm}(\mathfrak{D}, v) \geq \kappa\big)$ for each $\kappa \in [0, 1]$, provided that $\mathfrak{D}$ consists of $m > 2$ random vectors uniformly distributed on the unit sphere, but this is out of the scope of the paper.

A number of other issues remain also to be investigated related to how known properties of deterministic direct search extend to probabilistic descent. For instance, one knows that in some convex instances [12] the worst case complexity bound of deterministic direct search can be improved to the order of $1/\epsilon$ (corresponding to a global rate of $1/k$). One also knows that some deterministic direct-search methods exhibit an r-linear rate of local convergence [14]. Finally, extending our results to the presence of constraints and/or non-smoothness may be also of interest.

A Proof of Lemma 4.4

Proof. The result can be proved by standard techniques of large deviations. Let t be an arbitrary positive number. By Markov's Inequality,

$$\pi_k(\lambda) \;=\; \mathbb{P}\left(\exp\left(-t\sum_{l=0}^{k-1} Z_l\right) \geq \exp(-t\lambda k)\right) \;\leq\; \exp(t\lambda k)\,\mathbb{E}\left(\prod_{l=0}^{k-1} \mathrm{e}^{-tZ_l}\right). \tag{48}$$

Now let us study $\mathbb{E}(\prod_{l=0}^{k-1} \mathrm{e}^{-tZ_l})$. By Properties $\mathbf{G}^*$ and $\mathbf{K}^*$ of Shiryaev [26, page 216], we have

$$\mathbb{E}\left(\prod_{l=0}^{k-1} \mathrm{e}^{-tZ_l}\right) \;=\; \mathbb{E}\left(\mathbb{E}\left(\mathrm{e}^{-tZ_{k-1}} \mid Z_0, Z_1, \ldots, Z_{k-2}\right)\prod_{l=0}^{k-2} \mathrm{e}^{-tZ_l}\right). \tag{49}$$

According to (26) and the fact that the function $r\mathrm{e}^{-t} + (1 - r)$ is monotonically decreasing in r, it holds (with $\bar{p} = \mathbb{P}\left(Z_{k-1} = 1 \mid Z_0, Z_1, \ldots, Z_{k-2}\right) \geq p$)

$$\mathbb{E}\left(\mathrm{e}^{-tZ_{k-1}} \mid Z_0, Z_1, \ldots, Z_{k-2}\right) \;=\; \bar{p}\mathrm{e}^{-t} + (1 - \bar{p}) \;\leq\; p\mathrm{e}^{-t} + (1 - p) \;\leq\; \exp\big(p\mathrm{e}^{-t} - p\big),$$

which implies, from equality (49), that

$$\mathbb{E}\left(\prod_{l=0}^{k-1} \mathrm{e}^{-tZ_l}\right) \leq \exp\big(p\mathrm{e}^{-t} - p\big)\,\mathbb{E}\left(\prod_{l=0}^{k-2} \mathrm{e}^{-tZ_l}\right).$$

By recursively iterating the above estimation, we finally arrive at

$$\mathbb{E}\left(\prod_{l=0}^{k-1} \mathrm{e}^{-tZ_l}\right) \;\le\; \exp\bigl[k(p\mathrm{e}^{-t} - p)\bigr] .$$

Inequality (48) can then be rewritten as

$$\pi_k(\lambda) \;\le\; \exp\bigl[k(t\lambda + p\mathrm{e}^{-t} - p)\bigr] , \tag{50}$$

which holds for all $t > 0$. Let us select $t = \ln(\lambda^{-1}p)$. Then we have

$$t\lambda + p\mathrm{e}^{-t} - p \;=\; \lambda \ln(\lambda^{-1}p) + \lambda - p \;=\; -\frac{1}{2\xi}(\lambda - p)^2 \qquad (\lambda < \xi < p),$$

the second equality coming from Taylor expansion of the function $\lambda \mapsto \lambda \ln(\lambda^{-1}p) + \lambda - p$ at the point p. Thus, we conclude from inequality (50) that

$$\pi_k(\lambda) \;\le\; \exp\left[-\frac{(\lambda - p)^2}{2p} k\right] .$$

□

B A practical implementation of probabilistic descent sets

In the numerical experiments of Subsection 2.2, we chose $\mathfrak{D}_k$ as m independent random vectors uniformly distributed on the unit sphere in $\mathbb{R}^n$. Now we prove that the polling directions defined in this way are probabilistic descent as defined in Definition 3.1. Moreover, we will present practical estimations for p and κ (see Proposition B.1 below).

We assume throughout this section that the polling sets are mutually independent and that for each $k \ge 0$,

$$\mathfrak{D}_k \;=\; \{\mathfrak{d}_1, \ldots, \mathfrak{d}_m\} ,$$

where $\mathfrak{d}_1$, …, $\mathfrak{d}_m$ are independent random vectors uniformly distributed on the unit sphere. In computations, $\mathfrak{d}_1$, …, $\mathfrak{d}_m$ can be obtained by normalizing independent random vectors from the n-dimensional standard normal distribution [21].

We need two lemmas to deal with the probabilities involved in Definition 3.1. The first one is a corollary of Fubini's Theorem [6, Page 42]. A similar conclusion can be found in [16, Example 5.1.5].

Lemma B.1 ([6, page 148]) *If U and V are independent random variables, then for a random variable defined by $h(U, V)$, where h is a non-negative function, it holds*

$$\mathbb{E}\left(h(U, V) \mid V\right) \;=\; \bar{h}(V),$$

where

$$\bar{h}(v) \;=\; \mathbb{E}\left(h(U, v)\right) .$$

The second one is a bound on a probability that will be used later.

Lemma B.2 *Given $v \in \mathbb{R}^n$ and $\tau \in [0, \sqrt{n}]$, it holds*

$$\mathbb{P}\left(\operatorname{cm}(\mathfrak{D}_k, v) \geq \frac{\tau}{\sqrt{n}}\right) \;\geq\; 1 - \left(\frac{1}{2} + \frac{\tau}{\sqrt{2\pi}}\right)^m . \tag{51}$$

Proof. If $v = 0$, then $\operatorname{cm}(\mathfrak{D}_k, v) \equiv 1$ by definition, and there is nothing to prove. Hence we suppose that v is nonzero (and, without loss of generality, normalized). When $n = 1$, inequality (51) is also trivial. Therefore we assume $n \geq 2$. According to the implementation of $\mathfrak{D}_k$, we have

$$\mathbb{P}\left(\operatorname{cm}(\mathfrak{D}_k, v) \geq \frac{\tau}{\sqrt{n}}\right) \;=\; 1 - \left[1 - \mathbb{P}\left(\mathfrak{d}^\top v \geq \frac{\tau}{\sqrt{n}}\right)\right]^m ,$$

with $\mathfrak{d}$ being a random vector uniformly distributed on the unit sphere. To establish the desired lower bound for this probability, we study the function

$$\varrho(\kappa) \;=\; \mathbb{P}\left(\mathfrak{d}^\top v \geq \kappa\right), \quad \kappa \in [0, 1],$$

and it suffices to prove that

$$\varrho(\kappa) \;\geq\; \frac{1}{2} - \kappa\sqrt{\frac{n}{2\pi}}. \tag{52}$$

Since the distribution of $\mathfrak{d}$ is uniform on the unit sphere, $\varrho(\kappa)$ is proportional to the area A of the spherical cap

$$\left\{d \in \mathbb{R}^n : \|d\| = 1 \text{ and } d^\top v \geq \kappa\right\}$$

of unit radius and height

$$h \;=\; 1 - \kappa.$$

Recalling the area formula for spherical caps, we have

$$A \;=\; \frac{1}{2} A_n \operatorname{I}\left(2h - h^2, \,\frac{n-1}{2}, \,\frac{1}{2}\right) \;=\; \frac{1}{2} A_n \operatorname{I}\left(1 - \kappa^2, \,\frac{n-1}{2}, \,\frac{1}{2}\right),$$

where A_n is the area of the unit sphere in $\mathbb{R}^n$, and I is the regularized incomplete Beta function [1] defined by

$$\operatorname{I}(u, a, b) \;=\; \frac{1}{\operatorname{B}(a, b)} \int_0^u t^{a-1}(1-t)^{b-1} dt, \tag{53}$$

with B being the Beta function. Hence

$$\varrho(\kappa) \;=\; \frac{1}{2} \operatorname{I}\left(1 - \kappa^2, \,\frac{n-1}{2}, \,\frac{1}{2}\right). \tag{54}$$

When $n = 2$, by plugging (53) into (54), calculating the integral, and noticing $\operatorname{B}(\frac{1}{2}, \frac{1}{2}) = \pi$, we have

$$\varrho(\kappa) \;=\; \frac{1}{2} \operatorname{I}\left(1 - \kappa^2, \,\frac{1}{2}, \,\frac{1}{2}\right) \;=\; \frac{1}{\pi} \arcsin \sqrt{1 - \kappa^2} \;=\; \frac{1}{2} - \frac{1}{\pi} \arcsin \kappa \;\geq\; \frac{1}{2} - \frac{\kappa}{2},$$

and therefore inequality (52) is true. To prove (52) in the situation of $n \geq 3$, we examine definition (53) and find that

$$\begin{aligned}
\mathrm{I}\,(u,a,b) &= 1-\frac{1}{\mathrm{B}(a,b)}\int_u^1 t^{a-1}(1-t)^{b-1}dt \\
&\geq 1-\frac{1}{\mathrm{B}(a,b)}\int_u^1 (1-t)^{b-1}dt \\
&= 1-\frac{(1-u)^b}{b\,\mathrm{B}(a,b)}
\end{aligned}$$

when $a \geq 1$. Hence, using equation (54), we obtain

$$\varrho(\kappa) \;\geq\; \frac{1}{2}-\frac{\kappa}{\mathrm{B}(\frac{n-1}{2},\frac{1}{2})}$$

when $n \geq 3$. Thus we can arrive at inequality (52) as long as

$$\mathrm{B}\left(\frac{n-1}{2},\frac{1}{2}\right) \;\geq\; \sqrt{\frac{2\pi}{n}}. \tag{55}$$

Inequality (55) is justified by the facts

$$\mathrm{B}\left(\frac{n-1}{2},\frac{1}{2}\right) \;=\; \frac{\Gamma(\frac{n-1}{2})\,\Gamma(\frac{1}{2})}{\Gamma(\frac{n}{2})} \;=\; \frac{\Gamma(\frac{n-1}{2})}{\Gamma(\frac{n}{2})}\sqrt{\pi},$$

and

$$\Gamma\left(\frac{n}{2}\right) \;\leq\; \left[\Gamma\left(\frac{n-1}{2}\right)\Gamma\left(\frac{n+1}{2}\right)\right]^{\frac{1}{2}} \;=\; \sqrt{\frac{n-1}{2}}\,\Gamma\left(\frac{n-1}{2}\right),$$

the second of which is because Γ is log-convex, meaning that $\ln\Gamma$ is convex [2, Theorem 2.1]. □

Now we present the main result of this section. It concludes under the assumptions in this appendix that $\{\mathfrak{D}_k\}$ is probabilistically descent and it introduces easy-to-use estimations for p and κ.

Proposition B.1 *Given $\tau \in [0,\sqrt{n}]$, $\{\mathfrak{D}_k\}$ is p-probabilistically $(\tau/\sqrt{n})$-descent with*

$$p \;\leq\; 1-\left(\frac{1}{2}+\frac{\tau}{\sqrt{2\pi}}\right)^m. \tag{56}$$

Proof. According to Definition 3.1, we need to prove that

$$\mathbb{P}\left(\mathrm{cm}\,(\mathfrak{D}_0,-G_0) \;\geq\; \frac{\tau}{\sqrt{n}}\right) \;\geq\; p, \tag{57}$$

and

$$\mathbb{P}\left(\mathrm{cm}\,(\mathfrak{D}_k,-G_k) \;\geq\; \frac{\tau}{\sqrt{n}}\;\middle|\;\mathfrak{D}_0,\dots,\mathfrak{D}_{k-1}\right) \;\geq\; p \tag{58}$$

for each $k \geq 1$, with p satisfying (56). Inequality (57) follows directly from Lemma B.2, since $G_0 \equiv \nabla f(x_0)$. In the following we will show how to obtain (58) from Lemmas B.1 and B.2.

Let us fix $k \geq 1$. For any sets D_0, ..., D_{k-1}, D_k of polling directions, define

$$h(D_0,\ldots,D_{k-1},D_k) \;=\; \begin{cases} 1 & \text{if } \mathrm{cm}(D_k,-g_k) \geq \frac{\tau}{\sqrt{n}}, \\ 0 & \text{else,} \end{cases}$$

where $g_k \equiv g_k(D_0,\ldots,D_{k-1})$ is totally defined given $D_0,\ldots,D_{k-1}$. Then h is a well defined function of D_0, ..., D_{k-1}, D_k. Moreover,

$$\mathbb{P}\left(\mathrm{cm}(\mathfrak{D}_k,-G_k) \geq \frac{\tau}{\sqrt{n}} \;\middle|\; \mathfrak{D}_0,\ldots,\mathfrak{D}_{k-1}\right) \;=\; \mathbb{E}\left(h(\mathfrak{D}_0,\ldots,\mathfrak{D}_{k-1},\mathfrak{D}_k) \;\middle|\; \mathfrak{D}_0,\ldots,\mathfrak{D}_{k-1}\right).$$

Due to the independence of the sets of polling directions, we know from Lemma B.1 (with $U = \mathfrak{D}_k$ and $V = (\mathfrak{D}_0,\ldots,\mathfrak{D}_{k-1})$) that

$$\mathbb{E}\left(h(\mathfrak{D}_0,\ldots,\mathfrak{D}_{k-1},\mathfrak{D}_k) \;\middle|\; \mathfrak{D}_0,\ldots,\mathfrak{D}_{k-1}\right) \;=\; \bar{h}(\mathfrak{D}_0,\ldots,\mathfrak{D}_{k-1}),$$

with

$$\bar{h}(D_0,\ldots,D_{k-1}) \;=\; \mathbb{E}\left(h(D_0,\ldots,D_{k-1},\mathfrak{D}_k)\right) \;=\; \mathbb{P}\left(\mathrm{cm}(\mathfrak{D}_k,-g_k) \;\geq\; \frac{\tau}{\sqrt{n}}\right) \;\geq\; p,$$

the last inequality coming from Lemma B.2. Hence (58) holds, and the proof is completed. □

Proposition B.1 enables us to derive a bound for m that ensures global convergence and a global rate to Algorithm 2.1.

Corollary B.1 *If*

$$m \;>\; \log_2\left(1 - \frac{\ln\theta}{\ln\gamma}\right), \tag{59}$$

then $\{\mathfrak{D}_k\}$ is p-probabilistic $\tau/\sqrt{n}$-descent for some constants $p > p_0$ and $\tau > 0$ that are totally determined by γ and θ.

Proof. Let m_0 be the minimal integer that satisfies

$$m_0 \;>\; \log_2\left(1 - \frac{\ln\theta}{\ln\gamma}\right). \tag{60}$$

Then $m \geq m_0$. Given inequality (60), we have

$$1 - \left(\frac{1}{2}\right)^{m_0} \;>\; 1 - \left(1 - \frac{\ln\theta}{\ln\gamma}\right)^{-1} \;=\; p_0.$$

Thus, there exists a sufficiently small positive constant τ such that

$$1 - \left(\frac{1}{2} + \frac{\tau}{\sqrt{2\pi}}\right)^{m_0} \;>\; p_0.$$

Let

$$p \;=\; 1 - \left(\frac{1}{2} + \frac{\tau}{\sqrt{2\pi}}\right)^{m_0}. \tag{61}$$

Then

$$p \leq 1 - \left(\frac{1}{2} + \frac{\tau}{\sqrt{2\pi}}\right)^m, \tag{62}$$

and it is easy to check that both τ and p can be totally determined by γ and θ. The proof is concluded by applying Proposition B.1. □

We notice that the constants p and τ in Corollary B.1 are totally determined by γ and θ without dependence on m or n, and that the bound (59) for m is also totally determined by γ and θ. These observations are important for us to understand how the problem dimension influences the global rate and worst case complexity bound (see the discussion at the end of Subsection 4.2).

According to Corollary B.1, when m satisfies the bound (59), all of our theory is applicable to Algorithm 2.1 under the assumptions in this appendix. For example, when $\gamma = 2$ and $\theta = 0.5$, taking $m = 2$ can guarantee that Algorithm 2.1 enjoys the global convergence and global rate by us established, no matter how large the problem is (see Subsection 5.4).

References

[1] M. Abramowitz and I. Stegun. *Handbook of Mathematical Functions with Formulas, Graphs, and Mathematical Tables.* United States Department of Commerce, National Bureau of Standards, Washington, tenth edition, 1972.

[2] E. Artin. *The Gamma Function.* Holt, Rinehart and Winston, New York, 1964. Translated to English by M. Butler.

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

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

[5] 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.

[6] E. Çınlar. *Probability and Stochastics.* Graduate Texts in Mathematics. Springer, New York, 2011.

[7] A. R. Conn, K. Scheinberg, and L. N. Vicente. *Introduction to Derivative-Free Optimization.* MPS-SIAM Series on Optimization. SIAM, Philadelphia, 2009.

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

[9] C. Davis. Theory of positive linear dependence. *Amer. J. Math.*, 76:733–746, 1954.

[10] M. A. Diniz-Ehrhardt, J. M. Martínez, and M. Raydan. A derivative-free nonmonotone line-search technique for unconstrained optimization. *J. Comput. Appl. Math.*, 219:383–397, 2008.

[11] Y. Diouane, S. Gratton, and L. N. Vicente. Globally convergent evolution strategies. *Math. Program.*, (to appear).

[12] M. Dodangeh and L. N. Vicente. Worst case complexity of direct search under convexity. *Math. Program.*, (to appear).

[13] B. Doerr. Analyzing randomized search heuristics: Tools from probability theory. In A. Auger and B. Doerr, editors, *Theory of Randomized Search Heuristics: Foundations and Recent Developments*, volume 1 of *Series on Theoretical Computer Science*, pages 1–20. World Scientific, Singapore, 2011.

[14] E. D. Dolan, R. M. Lewis, and V. Torczon. On the local convergence of pattern search. *SIAM J. Optim.*, 14:567–583, 2003.

[15] D. P. Dubhashi and A. Panconesi. *Concentration of Measure for the Analysis of Randomized Algorithms.* Cambridge University Press, Cambridge, 2009.

[16] R. Durrett. *Probability: Theory and Examples.* Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, Cambridge, fourth edition, 2010.

[17] N. I. M. Gould, D. Orban, and P. L. Toint. CUTEr, a Constrained and Unconstrained Testing Environment, revisited. *ACM Trans. Math. Software*, 29:373–394, 2003.

[18] S. Gratton and L. N. Vicente. A merit function approach for direct search. *SIAM J. Optim.*, 24:1980–1998, 2014.

[19] N. Hansen and A. Ostermeier. Adapting arbitrary normal mutation distributions in evolution strategies: The covariance matrix adaptation. In *Proceedings of the 1996 IEEE International Conference on Evolutionary Computation*, pages 312–317, 1996.

[20] 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.

[21] M. Muller. A note on a method for generating points uniformly on n-dimensional spheres. *Commun. ACM*, 2:19–20, 1959.

[22] Y. Nesterov. Random gradient-free minimization of convex functions. Technical Report 2011/1, CORE, 2011.

[23] Y. Nesterov. Efficiency of coordinate descent methods on huge-scale optimization problems. *SIAM J. Optim.*, 22:341–362, 2012.

[24] P. P. Richtárik and M. Takáč. Iteration complexity of randomized block-coordinate descent methods for minimizing a composite function. *Math. Program.*, 144:1–38, 2014.

[25] B. T. Polyak. *Introduction to Optimization.* Optimization Software – Inc., New York, 1987.

[26] A. N. Shiryaev. *Probability.* Graduate Texts on Mathematics. Springer-Verlag, New York, 1995.

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

[28] L. N. Vicente. Worst Case Complexity of Direct Search. *EURO Journal on Computational Optimization*, 1:143–153, 2013.

[29] L. N. Vicente and A. L. Custódio. Analysis of direct searches for discontinuous functions. *Math. Program.*, 133:299–325, 2012.