

# Oblivious Randomized Direct Search for Real-Parameter Optimization

Jens Jägersküpper\*

Technische Universität Dortmund, Informatik 2, 44221 Dortmund, Germany  
JJ@Ls2.cs.uni-dortmund.de

**Abstract.** The focus is on black-box optimization of a function  $f: \mathbb{R}^N \rightarrow \mathbb{R}$  given as a black box, i. e. an oracle for  $f$ -evaluations. This is commonly called direct search, and in fact, most methods for direct search are heuristics. Theoretical results on the performance/behavior of such heuristics are still rare. One reason: Like classical optimization algorithms, also direct-search methods face the challenge of step-size control, and usually, the more sophisticated the step-size control, the harder the analysis. Obviously, when we want the search to actually converge to a stationary point (i. e., the distance from this point tends to zero) at a nearly constant rate, then step sizes must be adapted. In practice, however, obtaining an  $\varepsilon$ -approximation for a given  $\varepsilon > 0$  is often sufficient, and usually all  $N$  parameters are bounded, so that the maximum distance from the optimum is bounded. Thus, in such cases reasonable step sizes lie in a predetermined bounded interval. Considering the minimization of the distance from a fixed point as the objective, we address the question, for randomized heuristics that use isotropic sampling to generate new candidate solutions, whether we might get rid of step-size control – namely of the problems connected to it, like so-called premature convergence – by choosing step sizes randomly according to some properly pre-defined distribution over this interval. As this choice of step sizes is oblivious to the course of the optimization, we gain robustness against a loss of step-size control. Naturally, the question is: What is the price w. r. t. local convergence speed? As we shall see, merely a factor of order  $\ln(d/\varepsilon)$ , where  $d$  is the diameter of the the decision space, an  $N$ -dimensional interval region.

## 1 Introduction

Here optimization in high-dimensional Euclidean space  $\mathbb{R}^N$  is considered, and the crucial aspect is how the optimization time scales with  $N$ , the dimensionality of the search space. Furthermore, the optimization time depends on the *approximation error*  $\varepsilon$  – here defined as the Euclidean distance from the optimum point in  $\mathbb{R}^N$  – of the approximate solution to be found. That is, we consider the optimization time as a function of  $N$  as well as of  $\varepsilon$ . Unless stated differently, asymptotics (essentially  $O$  and  $\Omega$ ) are w. r. t.  $N \rightarrow \infty$ , though.

The scenario we consider is black-box optimization, i. e., the function  $f$  to optimized is given by a black box, namely an oracle for  $f$ -evaluations. In practice, in particular in various engineering disciplines, this is a very common situation:  $f$  is given

---

\* supported by the German Research Foundation (DFG) through the collaborative research center “Computational Intelligence” (SFB 531)

by simulations or even by real-world experiments. In such situations – unless simulations allow for algorithmic/automatic differentiation, which is rarely the case – there is no information about the gradient or the Hessian, so that classical optimization methods cannot be applied. In the beginning of black-box optimization, in order to make use of established first-order methods, usually gradient approximation by finite forward/symmetric differences was used, which costs  $N$  (or  $2N$ )  $f$ -evaluations per iteration. Nowadays, the focus lies on optimization methods that abandon gradient approximation, but try to find good solutions directly. Such methods are commonly called direct-search methods, and in fact, most of these are heuristics. Theoretical results on the performance and the behavior of such heuristics are still rare, cf. [1]. Among the first and most prominent direct-search heuristics are the pattern search by Hooke/Jeeves and the (downhill) simplex method by Nelder/Mead, cf. [2] for a comprehensive review. Surprisingly, also already in the 1960s randomized direct-search methods were proposed, one is the so-called *evolution strategy* by Rechenberg [3] and Schwefel. In this algorithm, in each iteration  $i$  a new candidate solution is generated by adding a so-called *Gaussian mutation vector*  $\mathbf{m} \in \mathbb{R}^N$  to the current candidate solution  $\mathbf{x}^{[i-1]}$ . Each component of  $\mathbf{m}$  is i. i. d. according to a zero-mean normal distribution with variance  $\sigma^2$ . If the so-called *mutant*  $\mathbf{y} := \mathbf{x}^{[i-1]} + \mathbf{m}$  improves upon  $\mathbf{x}^{[i-1]}$ , then  $\mathbf{x}^{[i]} := \mathbf{y}$ , otherwise  $\mathbf{x}^{[i]} := \mathbf{x}^{[i-1]}$ . Rechenberg and Schwefel focused on how to update  $\sigma$  adaptively to the course of the optimization, and they proposed different mechanisms how to adapt  $\sigma$  such that close-to-optimal (local) performance is achieved on the simple quadratic form  $\mathbf{x} \mapsto \sum_{k=1}^N x_k^2$ , which is commonly called SPHERE in the field of (meta)heuristics; for some  $\mathbf{x}^* \in \mathbb{R}^N$  let  $\text{SPHERE}_{\mathbf{x}^*}(\mathbf{x}) := \text{SPHERE}(\mathbf{x} - \mathbf{x}^*)$ . The simple heuristic just described is a so-called (1+1) Evolution Strategy. It fits the general framework of iterative methods considered in the following:

For a given initial candidate solution  $\mathbf{x}^{[0]} \in \mathbb{R}^N$  and  $i := 1$  DO

1. generate the displacement  $\mathbf{m}^{[i]} \in \mathbb{R}^N$  according to some distribution  $D^{[i]}$  over  $\mathbb{R}^N$
2. evaluate  $f$  at the sample  $\mathbf{y}^{[i]} := \mathbf{x}^{[i-1]} + \mathbf{m}^{[i]} \in \mathbb{R}^N$
3. decide whether to accept the new sample; if so,  $\mathbf{x}^{[i]} := \mathbf{y}^{[i]}$ , else  $\mathbf{x}^{[i]} := \mathbf{x}^{[i-1]}$
4.  $i := i + 1$  and GOTO 1 (unless stopping criterion met)

In many (meta)heuristics, the randomly chosen displacement vector  $\mathbf{m}$  follows a (multivariate) normal distribution. Actually, sampling in each iteration  $i$  a predefined number of search points each i. i. d. according to a (multivariate) normal distribution with mean  $\mathbf{x}^{[i-1]}$  was already proposed 1958 in [4]—without being specific about how to choose/adapt the variance, though. Since then randomized direct-search heuristics have become more and more popular, cf. [5].

The probably most apparent rule to decide (in Instruction 3) whether the sample  $\mathbf{y}$  is to be accepted to become the next candidate solution is so-called *elitist selection*, namely, in the case of minimization,  $\mathbf{x}^{[i]} := \mathbf{y}^{[i]}$  if and only if  $f(\mathbf{y}^{[i]}) \leq f(\mathbf{x}^{[i-1]})$ . This rule is commonly used, and it might be one reason why SPHERE is so attractive as a starting point for a theoretical analysis: In this scenario the approximation error in the search space (namely the Euclidean distance from the optimum) is reduced if and only if there is an improvement w. r. t. the  $f$ -value. Apparently, this makes the reasoning easier. (We focus on the approximation error in the search space here.) Moreover,

when considering a fixed distribution  $D$  in Instruction 1 for the sampling, elitist selection in combination with  $f := \text{SPHERE}$  results in maximum expected reduction of the approximation error because negative gains (i. e.,  $\mathbf{y}$  is further away from the optimum) are zeroed out, whereas positive gains (i. e.,  $\mathbf{y}$  is closer to the optimum) are accepted. Thus, this combination can somewhat be considered a best-case scenario.

One reason for choosing the normal distribution to generate new search points seems to be that this distribution has maximum (differential) entropy. Another reason is the following invariance property: An  $N$ -dimensional Gaussian mutation (the  $N$  components are i. i. d. according to a zero-mean normal distribution with variance  $\sigma^2$ ) is *isotropically distributed* over  $\mathbb{R}^N$ , i. e., its distribution is spherically symmetric, more precisely, invariant w. r. t. orthonormal transformations. The nice property of an isotropically distributed vector is that its (possibly) random length is independent of its random direction and that the direction is uniformly random:

**Proposition 1.** *Let the vector  $\mathbf{u}$  be uniformly distributed over the unit hyper-sphere  $\{\mathbf{y} \in \mathbb{R}^N \mid |\mathbf{y}| = 1\}$ . A vector  $\mathbf{x}$  is isotropically distributed over  $\mathbb{R}^N$  if and only if there exists a non-negative random variable  $\ell$  (independent of  $\mathbf{u}$ ) such that the distribution of  $\mathbf{x}$  equals the one of  $\ell \cdot \mathbf{u}$ .*

A formal proof can be found in [6, Sec. 2.1] for instance. The random length of a Gaussian mutation (a vector that is distributed according to an isotropic multivariate normal distribution) follows a scaled  $\chi$ -distribution.

In black-box optimization, when we do not know anything about  $f$ , using isotropic distributions (centered at the current iterate) to sample new candidate solutions seems reasonable because of the invariance properties. When we restrict the class of algorithms covered by our framework given above by requiring (in each iteration  $i$ ) the distributions  $D^{[i]}$  in Instruction 1 to be isotropic, then we can ask for an upper bound on the expected reduction of the approximation error in one step. Therefore, one may think of the best-case scenario in which SPHERE is minimized and elitist selection is used (in which positive gains (reduction of the distance from the optimum) are accepted, whereas negative gains are zeroed out). Let  $\mathbf{x}^* \in \mathbb{R}^N$  denote the optimum and let  $d^{[i]}$  be defined as the distance of  $\mathbf{x}^{[i]}$  from the optimum after the  $i$ th iteration. Furthermore, for a given distance  $d^{[i-1]}$ , let  $\Delta^{[i]}: \mathbb{R}^N \rightarrow \mathbb{R}$  denote the random variable defined as  $\text{dist}(\mathbf{x}^{[i-1]} + \mathbf{m}^{[i]}, \mathbf{x}^*) - d^{[i-1]}$  which is induced by the distribution  $D^{[i]}$  used to sample  $\mathbf{m}^{[i]}$  in the  $i$ th iteration. For  $\text{SPHERE}_{\mathbf{x}^*}$ , elitist selection corresponds to the indicator variable  $\mathbb{1}_{\{\Delta^{[i]} \geq 0\}}$  (which resolves to “1” if  $\Delta^{[i]} \geq 0$ , otherwise to “0”), so that the random variable  $\Delta_+^{[i]} := \Delta^{[i]} \cdot \mathbb{1}_{\{\Delta^{[i]} \geq 0\}}$  corresponds to the spatial gain towards the optimum  $\mathbf{x}^*$  in the  $i$ th iteration. Note that the distribution of  $\Delta_+^{[i]}$  has an atom at zero with a weight equal to the probability that  $\mathbf{y}^{[i]} = \mathbf{x}^{[i-1]} + \mathbf{m}^{[i]}$  is such that it is discarded in Instruction 3.

As shown in [7], for any isotropic distribution  $D^{[i]}$  over  $\mathbb{R}^N$  the expected spatial gain towards a predefined point (for instance  $\mathbf{x}^*$ ) is bounded above by

$$\mathbb{E} \left[ \Delta_+^{[i]} \right] < d^{[i-1]} \cdot 0.52 / (N - 1) \quad \text{for } N \geq 4. \quad (1)$$

Thus, if in each iteration  $i$  the isotropic distribution  $D^{[i]}$  was the best possible, then we would observe linear convergence (w. r. t. the distance from the optimum) at an expected

rate larger (i. e. worse) than  $1 - 0.52/(N - 1)$ . By substituting  $(N - 1)/0.52$  for  $n$  in the well-known inequality  $(1 - 1/n)^{n-1} > 1/e$ , we easily get that the total expected gain after (the first)  $k$  iterations is less than halve the initial approximation error unless  $k > \ln 2 / (0.52 / (N - 1.52)) > 1.33N - 2.03$ . (Due to the best-case assumption on the  $D^{[i]}$ , the factors by which the approximation error is reduced in  $k$  sequent steps are in fact i. i. d., so that we can indeed take the expectation of the one-step factor to the  $k$ th power to obtain the expectation of the factor which corresponds to the total reduction in the  $k$  steps.) Yet this does not tell us much anyway: The randomness is in the total gain rather than in the number of iterations. Instead, we would like to know a lower bound on the expected number of steps necessary to actually halve the approximation error. The local/one-step result from Equation (1) can indeed be transformed into the following lower-bound result on the runtime [8, Thm. 13]:

**Theorem 2.** *For any heuristic that fits our framework: When the  $D^{[i]}$  are isotropic distributions, then the expected number of iterations necessary to halve the approximation error (defined as the distance from a fixed point in  $\mathbb{R}^N$ ) is bounded from below by  $0.5 / (0.52 / (N - 1)) > 0.96N - 1 = \Omega(N)$ .*

Note that this theorem holds for *any* adaptation mechanisms which determines for each iteration  $i$  according to what isotropic distribution  $D^{[i]}$  to sample  $\mathbf{m}^{[i]}$ . Interestingly, as shown in [9], even if in each iteration  $i$  the point  $\mathbf{x}^{[i]}$  was magically chosen from the line  $\{\mathbf{x}^{[i-1]} + \alpha \cdot \mathbf{m}^{[i]} \mid \alpha \in \mathbb{R}\}$  such that the distance of  $\mathbf{x}^{[i]}$  from the optimum is minimum (a “perfect” line search along a uniformly random direction), we would observe linear convergence at an expected rate larger (i. e. worse) than  $1 - 1/N$ .

Now, talking about linear convergence at an expected rate makes sense only if the steps resemble each other up to a rescaling of the situation, which is in fact the case when assuming that in each iteration  $D^{[i]}$  was chosen as the best isotropic distribution, namely the one that maximizes the expected gain. When considering a concrete heuristic, namely a concrete adaptation mechanism to determine the  $D^{[i]}$ , then – because of the black-box scenario – it seems that the  $D^{[i]}$  just cannot be chosen such that a steady convergence is observed. Rather the expected reduction of the approximation error will vary from step to step. This is particularly true for step-size adaptations that aim at maximizing the local convergence speed, i. e., they try to choose in each iteration  $i$  the distribution  $D^{[i]}$  such that expected one-step gain is maximum. To get around this non-steadiness – and to preclude detrimental effects of a possible loss of step-size control like premature convergence – one may ask the following

**Question:** Is there a distribution  $D^*$  over  $\mathbb{R}^N$  such that using  $D^{[i]} := D^*$  in each iteration  $i$  results in a virtually steady convergence (i. e. at a virtually constant expected rate) to the optimum  $\mathbf{x}^*$  when minimizing  $\text{SPHERE}_{\mathbf{x}^*}$ ?

It is quite easy to see that such a distribution cannot exist if the approximation error is supposed to become arbitrarily small. Such a  $D^*$  might exist, however, when we merely aim at an  $\varepsilon$ -approximation and know an upper bound  $d_{\max}$  on the approximation errors that can occur.

In fact, we shall see in Section 3 that for the latter situation there is an isotropic distribution such that the algorithm in our framework that uses this distribution in each

iteration and elitist selection converges linearly at an expected rate smaller (i. e. better) than  $1 - 1/O(N \cdot \ln(d_{\max}/\varepsilon))$ . This algorithm will be called *oblivious randomized direct search (ORDS)*. As we shall moreover see, as long as the approximation error, namely the distance from the optimum  $\mathbf{x}^*$  of  $\text{SPHERE}_{\mathbf{x}^*}$ , is in the interval  $[2\varepsilon, d_{\max}]$ , the expected number of iterations that ORDS needs to halve the approximation error is bounded above by  $O(N \cdot \ln(d_{\max}/\varepsilon))$ . This is off from the general lower bound for isotropic sampling (Theorem 2) merely by a factor of order  $\ln(d_{\max}/\varepsilon)$ . This is a remarkable property – especially when considered together with the results of related work to be discussed in the next section.

## 2 Related Work

A question similar to the one posed above has already been investigated in [10] (although with a different motivation). The scenario investigated therein is the minimization of a unimodal one-dimensional function over the interval  $(-1, 1]$ . The search wraps around when the interval is left; e. g., when the point  $x = 1.5$  is sampled,  $f(-0.5)$  is computed. In this scenario, the approximation error is bounded above by 1. The authors propose using the following distribution with the density  $f_{RH} : \mathbb{R} \rightarrow \mathbb{R}_{\geq 0}$  defined by  $f_{RH}(m) := 1/|2 \cdot p \cdot m|$  for  $m \in [\varepsilon, 1] \cup [-1, -\varepsilon]$  and  $f_{RH}(m) := 0$  otherwise, where  $p := \ln(1/\varepsilon)$  for normalization and  $\varepsilon \in (0, 1)$  is the predefined smallest step length. Note that  $\int_{d/2}^d f_{RH}(m) dm = (\ln(d) - \ln(d/2))/(2p) = \ln(2)/(2p)$ , which is independent of  $d$ . Thus, the probability to halve the distance from the optimum is at least  $\ln(2)/(2p)$  in a step – independently of the distance  $d$  from the optimum. Obviously,  $\varepsilon \leq d/2$  is required for actual independence of  $d$ . Concerning the expected number of steps to halve the approximation error, the authors conclude that “the expected waiting time (and this is clearly an upper bound) is thus  $2p/\ln 2$ . The number of steps required to get within  $\delta$  of the optimum is therefore  $O(p \cdot \ln(1/\delta))$ .” Apparently,  $\delta \geq 2\varepsilon$  seems to be assumed there. As long as the approximation error is at least  $\delta + \varepsilon$ , the expected factor by which the distance from the optimum is reduced equals  $1 - \alpha/\ln(1/\varepsilon)$  for some  $\alpha > 0$  almost constant. Obviously, the smaller the minimum step length  $\varepsilon$ , the larger (i. e. worse) the expected convergence rate. Taking the minor technical issue discussed above into account, for  $\delta = 2\varepsilon$  we obtain a bound of  $O(\ln^2(1/\varepsilon))$  to find an  $2\varepsilon$ -approximation when using a minimal step length of  $\varepsilon$ . Compared to binary search, this is off by a factor of order  $\ln(1/\varepsilon)$ . Dietzfelbinger/Rowe/Wegener/Woelfel [11] focus on whether this  $\ln(1/\varepsilon)$ -factor is inherent to the usage of a fixed distribution, i. e., whether any fixed distribution according to which the samples are drawn needs  $\Omega(\ln^2(1/\varepsilon))$  iterations (in expectation) to obtain an  $\varepsilon$ -approximation. Actually, they consider a discrete version of the problem, where a *blind search* on the integers  $0, \dots, n$  is performed using a fixed distribution  $\mu$  over  $\{1, \dots, n\}$  for the sampling. Namely, the search starts at a (uniformly) random position in  $\{0, \dots, n\}$ . In each iteration the new position is given by the current position minus a number chosen according to  $\mu$  – given that this new position is non-negative; otherwise the search stays at its position. Dietzfelbinger et al. prove that for the distribution defined by  $\mu(m) := 1/(m \cdot H_n)$  for  $m \in \{1, \dots, n\}$  and  $\mu(m) := 0$  otherwise (where  $H_n = \sum_{i=1}^n 1/i$  is the  $n$ th Harmonic number; for normalization) the expected number of iterations to reach position

zero is  $O(\ln^2 n)$ . Their main result is, however, that the expected number of steps is  $\Omega(\ln^2 n)$  for *any* distribution  $\mu$ , i. e., losing a factor of order  $\ln n$  compared to binary search is inherent to blind search on the integers using a fixed distribution.

### 3 “Oblivious Randomized Direct Search” and its Analysis

In the present paper, we focus on direct search in  $N$ -dimensional Euclidean space. Namely, we consider the minimization of  $\text{SPHERE}_{\mathbf{x}^*}$ , i. e., the minimization of the (squared) distance from the unique optimum  $\mathbf{x}^* \in \mathbb{R}^N$ . Note the following obvious, but important observation: As  $\mathbf{x}^*$  is not known, any candidate for the distribution  $D^*$  that might satisfy the property we ask for in the question at the end of Section 1 must necessarily be isotropic! Since any isotropic distribution can be decomposed according to Proposition 1, we are actually looking for some length distribution  $L^*$  such that the distribution  $D^* \sim L^* \cdot U$  over  $\mathbb{R}^N$  has the desired property, where  $U$  is uniformly distributed upon the unit hyper-sphere (uniformly random direction, independent of  $L^*$ ). And since we consider isotropic distributions, we can restrict ourselves to the distance from the optimum  $\mathbf{x}^*$ . Now, assume that the current candidate solution  $\mathbf{x}$  is located at distance  $d$  from  $\mathbf{x}^*$ . Then  $p_{d,\ell,\alpha} := \text{P}\{\text{dist}(\mathbf{x} + \ell \cdot U, \mathbf{x}^*) \leq \alpha \cdot d\}$  equals the probability that adding an isotropically distributed vector with a fixed length  $\ell$  to  $\mathbf{x}$  generates a point such that the approximation error is reduced by (at least) the factor  $\alpha \in (0, 1)$ . Note that  $p_{d',\ell',\alpha} = p_{d,\ell,\alpha}$  whenever  $\ell'/d' = \ell/d$  because of scale invariance. Now assume that the length  $\ell$  is not fixed, but independently chosen according to some probability distribution with density  $\mu$ . Then the probability to reduce the approximation error by at least the factor  $\alpha \in (0, 1)$  equals

$$p_{d,\mu,\alpha} := \int_{(1-\alpha)d}^{(1+\alpha)d} p_{d,\ell,\alpha} \cdot \mu(\ell) \, d\ell.$$

The integral limits are due to the following fact: For the hyper-sphere with radius  $\ell$  centered at  $\mathbf{x}$  to intersect with the hyper-ball with radius  $\alpha \cdot d$  centered at  $\mathbf{x}^*$ , the radius  $\ell$  must be in the interval  $[d - \alpha d, d + \alpha d]$ . If  $\ell$  is smaller than  $d - \alpha d$  or larger than  $d + \alpha d$ , the sphere and the ball do not intersect. By substituting  $d \cdot x$  for  $\ell$  and using  $p_{d,d \cdot x,\alpha} = p_{1,x,\alpha}$ , we obtain

$$p_{d,\mu,\alpha} = \int_{(1-\alpha)}^{(1+\alpha)} p_{1,x,\alpha} \cdot \mu(x \cdot d) \cdot d \, dx.$$

Thus, if  $\mu$  was such that  $\mu(x \cdot d) \cdot d$  is independent of  $d$ , i. e.,  $\mu(x \cdot d) \cdot d = \mu(x \cdot d') \cdot d'$ , then  $p_{d,\mu,\alpha}$  would indeed be independent of  $d$ . As a consequence, we choose  $\mu(\ell)$  as  $\beta/\ell$  for some constant  $\beta > 0$ . Then  $\mu(x \cdot d) \cdot d = \beta/x$ , so that

$$p_{d,\mu,\alpha} = \beta \cdot \int_{(1-\alpha)}^{(1+\alpha)} \frac{p_{1,x,\alpha}}{x} \, dx,$$

which seems independent of  $d$ . However, as already pointed out in the discussion of related work in Section 2,  $p_{d,\mu,\alpha}$  is actually independent of  $d$  only if the support of  $\mu$  covers  $[d - \alpha d, d + \alpha d]$ . Note that  $\mu$  must have bounded support  $[a, b]$  with  $0 < a < b$

in our case since neither  $\int_0^a 1/x \, dx$  nor  $\int_b^\infty 1/x \, dx$  are finite. Actually, when we choose  $[a, b]$  as the support for  $\mu$ , then  $1/\beta$  equals  $\int_a^b 1/x \, dx = \ln b - \ln a = \ln(b/a)$  for normalization. Later we will focus on how to choose the support  $[a, b] \subset \mathbb{R}_{>0}$  of

$$\mu: \mathbb{R} \rightarrow \mathbb{R} \text{ with } \mu(\ell) := \begin{cases} \frac{1}{\ell \cdot \ln(b/a)} & \text{for } \ell \in [a, b] \\ 0 & \text{for } \ell \notin [a, b]. \end{cases} \quad (2)$$

Note the similarity between  $\mu$  and the distributions in the two related papers discussed in Section 2. In the remainder, we focus on the following iterative method:

**ORDS**<sub>[a,b]</sub> (Oblivious Randomized Direct Search) is the method in our framework that uses elitist selection in Instruction 3 and in each iteration the isotropic distribution  $D^* \sim \mu \cdot U$  in Instruction 1, where  $U$  is uniformly distributed upon the unit hyper-sphere and  $\mu$  as in Equation (2) with support  $[a, b]$ .

Unfortunately, for the analysis of ORDS the reciprocal of the probability to halve the approximation error in a single step does not result in a reasonable upper bound on the expected waiting time until the approximation error is halved. The reason is that, for  $N \geq 4$ , the probability to halve the approximation error in an iteration  $i$  is exponentially small in  $N$ , namely smaller than  $2^{-N} \cdot 0.43\sqrt{N-1}$  for any isotropic distribution  $D^{[i]}$  [12, Lemma 3]. Instead, we will explicitly calculate a lower bound on the expected one-step gain in the following – and with it an upper bound on the expected convergence rate. For a start, we follow [12, p. 329]:

“Consider the hyper-plane  $H$  that contains the current candidate solution  $\mathbf{x}$  ( $\neq \mathbf{x}^*$ ) and is orthogonal to the line passing through  $\mathbf{x}$  and  $\mathbf{x}^*$ . Assume that the isotropically distributed vector  $\mathbf{m}$  happens to have the length  $\ell > 0$ . Then  $\mathbf{y} = \mathbf{x} + \mathbf{m}$  is uniformly distributed upon the hyper-sphere centered at  $\mathbf{x}$  with radius  $\ell$ . The random variable

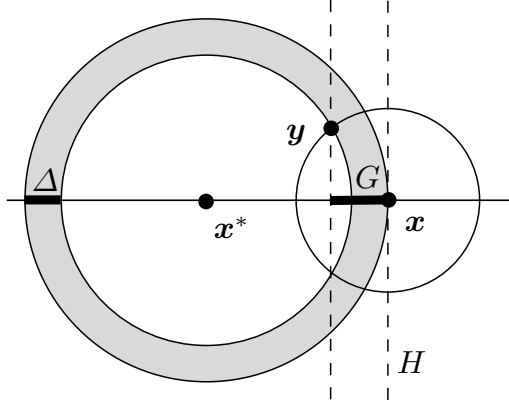
$$G_\ell(\mathbf{y}) := \begin{cases} \text{dist}(\mathbf{y}, H) & \text{if } \mathbf{y} \text{ lies in the half-space w. r. t. } H \text{ containing } \mathbf{x}^* \\ -\text{dist}(\mathbf{y}, H) & \text{otherwise} \end{cases}$$

corresponds to the *signed distance* of the sample  $\mathbf{x} + \mathbf{m}$  from the hyper-plane  $H$  (under the condition that  $|\mathbf{m}| = \ell$ ). Obviously, the support of  $G_\ell$  is  $[-\ell, \ell]$ . For  $N \geq 4$  the density of  $G_\ell$  at some  $g \in [-\ell, \ell]$  equals  $(1 - (g/\ell)^2)^{(N-3)/2} / (\ell \cdot \Psi)$ , where  $\Psi := \int_{-1}^1 (1 - x^2)^{(N-3)/2} \, dx$  (for normalization); cf. [8]. Actually, for a given distance of  $d = \text{dist}(\mathbf{x}, \mathbf{x}^*)$  from the optimum we are interested in the random variable

$$\Delta_{d,\ell}(\mathbf{y}) := d - \text{dist}(\mathbf{y}, \mathbf{x}^*)$$

which corresponds to the spatial gain towards the optimum  $\mathbf{x}^*$  (under the condition  $|\mathbf{m}| = \ell$ ). The support of the random variable  $\Delta_{d,\ell}$  is  $[-\ell, \min\{\ell, 2d - \ell\}]$ . The interrelation between  $\Delta_{d,\ell}$  and  $G_\ell$  is depicted in Figure 1. Simple geometry reveals (for any  $\mathbf{y}$  with distance  $\ell$  from  $\mathbf{x}^*$ ) the interrelation

$$G_\ell(\mathbf{y}) = \Delta_{d,\ell}(\mathbf{y}) + \frac{\ell^2 - (\Delta_{d,\ell}(\mathbf{y}))^2}{2d}.$$



**Fig. 1.** Interrelation of the random variables  $G_\ell$  and  $\Delta_{d,\ell}$

As a consequence for the present situation,  $G_\ell(\mathbf{y}) \geq \Delta_{d,\ell}(\mathbf{y}) \geq G_\ell(\mathbf{y}) - \ell^2/(2d)$ , where  $\mathbf{y}$  is a point from the hyper-sphere with radius  $\ell$  centered at  $\mathbf{x}$ . In particular,  $\Delta_{d,\ell}(\mathbf{y}) = 0$  corresponds to  $G_\ell(\mathbf{y}) = \ell^2/(2d)$ , so that

$$\mathbb{E}[\Delta_{d,\ell}^+] := \mathbb{E}[\Delta_{d,\ell} \cdot \mathbf{1}_{\{\Delta_{d,\ell} \geq 0\}}] \geq \mathbb{E}[G_\ell \cdot \mathbf{1}_{\{G_\ell \geq \ell^2/(2d)\}}] - \ell^2/(2d). \quad (3)$$

Note that  $\mathbb{E}[\Delta_{d,\ell}^+]$  is the expected one-step gain towards the optimum  $\mathbf{x}^*$  at distance  $d$ , given that the length of the isotropic distribution happens to be  $\ell$ , when minimizing  $\text{SPHERE}_{\mathbf{x}^*}$  using elitist selection (a best-case scenario). As the integral behind  $\mathbb{E}[\Delta_{d,\ell}^+]$  seems to not have an algebraically closed form, we will use the lower bound  $\mathbb{E}[G_\ell \cdot \mathbf{1}_{\{G_\ell \geq \ell^2/(2d)\}}] - \ell^2/(2d)$ , which can be easily calculated. Therefore recall that the density of  $G_\ell$  at  $g \in [-\ell, \ell]$  equals  $(1 - (g/\ell)^2)^{(N-3)/2}/(\ell \cdot \Psi)$ . Utilizing that  $(1 - x^2)^{(N-1)/2}/(1 - N)$  is an anti-derivative of the function  $x \cdot (1 - x^2)^{(N-3)/2}$ , for  $g \in [-\ell, \ell]$  and  $N \geq 4$

$$\begin{aligned} \mathbb{E}[G_\ell \cdot \mathbf{1}_{\{G_\ell \geq g\}}] &= \frac{1}{\Psi \cdot \ell} \cdot \int_g^\ell x \cdot (1 - (x/\ell)^2)^{(N-1)/2} dx \\ &= \frac{\ell^2}{\Psi \cdot \ell} \cdot \left[ \frac{-1}{N-1} \cdot (1 - (x/\ell)^2)^{(N-1)/2} \right]_g^\ell \\ &= \frac{\ell}{\Psi \cdot (N-1)} \cdot (1 - (g/\ell)^2)^{(N-1)/2}. \end{aligned}$$

As  $\Psi = \sqrt{\pi} \cdot \Gamma((N-1)/2)/\Gamma(N/2)$ , here  $\Gamma$  denotes the well-known gamma function,

$$\frac{1}{\Psi \cdot (N-1)} \geq \sqrt{(N-2)/2\pi}/(N-1) \geq 1/\sqrt{2\pi(N+1)} > 0.3989/\sqrt{N+1}$$

where we use  $\sqrt{N-2}/(N-1) \geq 1/\sqrt{N+1}$  for  $N \geq 3$ . Thus, for  $g := \ell^2/(2d)$

$$\mathbb{E}[G_\ell \cdot \mathbf{1}_{\{G_\ell \geq \ell^2/(2d)\}}] > (1 - (\ell/(2d))^2)^{(N-1)/2} \cdot \ell \cdot 0.3989/\sqrt{N+1}.$$

For  $\ell \leq d/\sqrt{N}$ , we have  $(1 - (\ell/(2d))^2)^{(N-1)/2} \geq (1 - 0.25/N)^{(N-1)/2} \geq e^{-1/8} > 0.8824$ , so that (using  $0.8824 \cdot 0.3989 > 0.35$ )

$$\mathbb{E}[G_\ell \cdot \mathbf{1}_{\{G_\ell \geq \ell^2/(2d)\}}] > 0.35 \ell/\sqrt{N+1} \quad \text{for } \ell \leq d/\sqrt{N} \text{ and } N \geq 4. \quad (4)$$

Together with Equation (3), we thus obtain the following lower bound on the expected spatial gain towards the optimum in the search space:

$$\mathbb{E} \left[ \Delta_{d,\ell}^+ \right] > \frac{0.35 \ell}{\sqrt{N+1}} - \frac{\ell^2}{2d} \text{ for } \ell \leq d/\sqrt{N} \text{ and } N \geq 4. \quad (5)$$

(In particular, for a length  $\ell$  of  $0.35d/\sqrt{N+1}$  this lower bound on the expected gain resolves to  $0.06125d/(N+1) = \Omega(d/N)$ , which is off from the general upper bound in Equation (1) by a factor of less than  $8.5 + 17/(N-1)$  only. As a consequence, using the lower bound from Equation (5) is safe – in particular when we focus on the asymptotic order of the gain as  $N$  grows.)

Now, recall from Equation (2) the distribution  $\mu$  with support  $[a, b]$  according to which the length  $\ell$  of the isotropically distributed vector is chosen. Then, given that  $a \leq 0.1d/\sqrt{N+1}$  and  $b \geq 0.7d/\sqrt{N+1}$ , we obtain (for  $N \geq 4$ )

$$\begin{aligned} \mathbb{E} \left[ \Delta_{d,\mu}^+ \right] &= \int_0^\infty \mathbb{E} \left[ \Delta_{d,\ell}^+ \right] \cdot \mu(\ell) \, d\ell \\ &> \int_{0.1d/\sqrt{N+1}}^{0.7d/\sqrt{N+1}} \left( \frac{0.35 \ell}{\sqrt{N+1}} - \frac{\ell^2}{2d} \right) \cdot \frac{1}{\ell \cdot \ln(b/a)} \, d\ell \\ &= \frac{0.09d}{(N+1) \cdot \ln(b/a)}, \end{aligned}$$

where we use that  $\Delta_{d,\ell}^+$  (and thus its expectation) is non-negative anyway, so that integration limits can be chosen, and the bound from Equation (5). Note that this lower bound on the expected one-step gain is off from the general upper bound in Equation (1) by a factor of less than  $\ln(b/a) \cdot 5.8 \cdot (N+1)/(N-1) = O(\ln(b/a))$ . All in all, we have just proved the following.

**Lemma 3.** *Let  $ORDS_{[a,b]}$  minimize  $\text{SPHERE}_{\mathbf{x}^*}$  in  $\mathbb{R}^N$ ,  $N \geq 4$ . If the approximation error  $d$  (distance from the optimum  $\mathbf{x}^*$ ) and  $a$  and  $b$  are such that  $a \leq 0.1d/\sqrt{N+1}$  as well as  $b \geq 0.7d/\sqrt{N+1}$ , then the expected factor by which the approximation error is reduced in a step is smaller (i. e. better) than  $1 - \frac{0.09}{(N+1) \cdot \ln(b/a)}$ .*

When we aim at an  $\varepsilon$ -approximation, then  $a := 0.1\varepsilon/\sqrt{N+1}$  must be chosen for the preceding lemma to apply. (Actually,  $a := \varepsilon/\sqrt{N}$  should work; the factor 0.1 is due to the application of the bound in Equation (5) and the rough estimation of the integral's value.) For the choice of  $b$  recall the discussion that led to the definition of  $\mu$  in Equation (2). Hence, we choose  $b$  as twice the maximum possible approximation error. In the SPHERE scenario, this is twice the initial approximation error. If the initial approximation error is not known – like in the following setting – an upper bound can be used. (The length of a diagonal in the  $N$ -dimensional interval region  $[0, 1]^N$  is  $\sqrt{N}$ .)

**Theorem 4.** *Let  $ORDS_{[a,b]}$  minimize  $\text{SPHERE}_{\mathbf{x}^*}$  in  $\mathbb{R}^N$ ,  $N \geq 4$ . Assume that the optimum point  $\mathbf{x}^*$  as well as the initial search point lie in the set  $[0, 1]^N$ . Then choosing  $a := 0.1\varepsilon/\sqrt{N+1}$  and  $b := 2\sqrt{N+1}$  ensures linear convergence at an expected rate smaller (i. e. better) than  $1 - \frac{0.09}{(N+1) \cdot \ln(20(N+1)/\varepsilon)} = 1 - 1/O(N \ln(N/\varepsilon))$  until the approximation error drops below  $\varepsilon + a < \varepsilon \cdot (1 + 0.1/\sqrt{N})$ .*

Now that we know an upper bound on the expected factor by which the approximation error is reduced in each step (unless the approximation error drops below  $\varepsilon + a$ ) we would like to turn this into an upper bound on the expected number of steps to reduce the approximation error by a predefined amount. The following lemma, which can be found in [8] including a full proof, will enable us to do so.

**Lemma 5.** *Let  $X_1, X_2, \dots$  denote random variables with bounded support and  $S$  the random variable defined by  $S := \min\{t \mid X_1 + \dots + X_t \geq g\}$  for a predefined  $g > 0$ . Given that  $S$  is a stopping time, i. e., the event  $\{S = k\}$  depends solely on  $X_1, \dots, X_k$ , if  $E[S] < \infty$  and  $E[X_i \mid S \geq i] \geq \phi > 0$  for all  $i$ , then  $E[S] \leq E[X_1 + \dots + X_S]/\phi$ .*

*Proof.* Note that the  $X_i$  need not be independent and that, since the  $X_i$  are bounded, the precondition  $E[S] < \infty$  implies  $E[X_1 + \dots + X_S] < \infty$ . Then

$$E[X_1 + \dots + X_S] = \sum_{i=1}^{\infty} P\{S \geq i\} \cdot E[X_i \mid S \geq i] \geq \sum_{i=1}^{\infty} P\{S \geq i\} \cdot \phi = E[S] \cdot \phi$$

where the first equation is the major part of the proof of Wald's equation.  $\square$

We concentrate on the expected number of steps to halve the approximation error, and thus, for the application of Lemma 5 we let  $X_i$  denote the spatial gain towards the optimum in the  $i$ th iteration and choose  $g := d^{[0]}/2$  and  $\ell := \frac{0.09}{(N+1) \cdot \ln(20(N+1)/\varepsilon)} d^{[0]}/2$ , where we use that  $0 \leq X_i \leq d^{[0]}$  in our scenario, and that the condition  $\{S \geq i\}$  merely means that the approximation error has not been halved within the first  $i - 1$  iterations, i. e.,  $d^{[i-1]} > d^{[0]}/2$ . Finally, we use the trivial bound  $E[X_1 + \dots + X_S] \leq d^{[0]}$  (which actually costs us a factor of nearly 2) and note that  $E[S] < \infty$  since in each iteration the success/stopping region is hit with positive probability. All in all, the application of the previous lemma yields the following upper bound on the expected number of steps to halve the approximation error.

**Corollary 6.** *Consider the settings from Theorem 4. Then, unless the approximation error is smaller than  $2(\varepsilon + a) < 2\varepsilon(1 + 0.1/\sqrt{N})$ , the expected number of steps to halve the approximation error is at most  $d^{[0]}/\frac{0.09 \cdot d^{[0]}/2}{(N+1) \cdot \ln(20(N+1)/\varepsilon)}$ , which is smaller than  $22.3 \cdot (N + 1) \cdot (\ln(N + 1) + 3 - \ln \varepsilon) = O(N \cdot \ln(N/\varepsilon))$ .*

This upper bound is off from the general lower bound when using isotropic samples (Theorem 2) – which covers perfect adaptation – by less than a factor of  $23.2 \ln(N/\varepsilon)$  for large  $N$ . As the location of the optimum  $\mathbf{x}^* \in [0, 1]^N$  is not known, random initialization is the most appropriate choice, finally yielding the main result:

**Theorem 7.** *Let  $ORDS_{[a,b]}$  minimize  $\text{SPHERE}_{\mathbf{x}^*}$  in  $\mathbb{R}^N$ ,  $N \geq 4$ , for some  $\mathbf{x}^* \in [0, 1]^N$ , where  $a := 0.1\varepsilon/\sqrt{N+1}$  and  $b := 2\sqrt{N+1}$ . Then, unless the approximation error is smaller than  $\varepsilon' := \varepsilon + a < \varepsilon \cdot (1 + 0.1/\sqrt{N})$ , the search converges linearly at an expected rate smaller than  $1 - \frac{0.09}{(N+1) \cdot \ln(20(N+1)/\varepsilon)}$ , and moreover, when the initial candidate solution is sampled uniformly at random from  $[0, 1]^N$ , the expected number of steps to obtain an  $\varepsilon'$ -approximation is  $O(N \cdot \ln^2(N/\varepsilon))$ .*

*Proof.* Obviously, the (expected) initial approximation error is bounded above by  $\sqrt{N}$ . As a consequence, the approximation error must be halved at most  $\ln(\sqrt{N}/\varepsilon)/\ln 2$  times in expectation (w.r.t. the initialization) to obtain an  $\varepsilon$ -approximation. As the random initialization and the sampling of  $\mu$  are independent, we can multiply by the expected number of steps to halve the approximation error to obtain an upper bound on the expected runtime of  $22.3 \cdot (N+1) \cdot \ln(20(N+1)/\varepsilon) \cdot \ln(\sqrt{N}/\varepsilon)/\ln 2$ , which is bounded above by  $32.2 \cdot (N+1) \cdot \ln^2((N+1)/\varepsilon) + O(N \cdot \ln(N/\varepsilon))$ .  $\square$

## 4 Discussion and Conclusion

The choice of the unit hyper-cube  $[0, 1]^N$  as the decision space was somewhat arbitrary, of course. For any bounded  $N$ -dimensional interval region we can choose  $b$  as twice its diameter  $d$ . Then the expected number of steps to halve the approximation error on SPHERE is  $O(N \cdot \ln(d/\varepsilon))$ , which is larger than the general lower bound (when using isotropic sampling to generate candidate solutions) by a factor of order  $\ln(d/\varepsilon)$ . Actually, this is the factor that ORDS loses in the best-case scenario. In practice, the optimization scenario is often not best-case, but the function to be optimized may be multi-modal for instance. Then (usual) step-size controls result in the convergence to the stationary point that is closest to the (random) starting point. On the one hand, they (usually) accelerate the convergence to the nearest local optimum, but on the other hand, the step sizes rapidly become too small to escape the local optimum region. This is bad in particular when there are many local optima, so that a large number of restarts is necessary to initialize within one of the “good” local-optimum regions. Also ORDS may converge to the nearest local optimum, but as it does so without favoring smaller and smaller step-sizes, and because of the heavy tailed distribution of the step-lengths, ORDS simultaneously searches more globally, i. e., the chance of escaping the local optimum region is preserved. This can also be considered as an implicit mechanism for automated restarts. In particular, after such a “restart”, i. e., after a long step that made the search leave the current local optimum region into another one, the step-sizes need *not* be re-adjusted in ORDS as it would be necessary for (usual) step-size controls. Preliminary experimental investigations of ORDS support this hypothesis on its behavior, and also the simulations presented in [10] (the first work discussed in Section 2) indicate that the concept behind ORDS can work well. (The authors consider a bunch of multi-modal test functions, but also “a difficult real-world application, from medical image interpretation.”)

Thus, besides the interesting theoretical aspects of ORDS and its runtime analysis presented here, the distribution used in ORDS to sample new candidate solutions may indeed be used in more complex algorithms. For instance, it could be used within the CMA-ES (Covariance Matrix Adaptation Evolution Strategy, cf. [13]) instead of multivariate normal distributions. This would supersede the complex and expensive step-size adaptation (by the so-called cumulative step-size adaptation (CSA)), but the learning and the continuous adaptation of the inverse Hessian (similar to a quasi-Newton approach) would be retained, which particularly helps with the optimization of ill-conditioned problems. An experimental investigation of such an algorithm with a (sta-

tistical) comparison to other direct-search heuristics shall help us assess the potential for practical optimization.

**Acknowledgment.** The author thanks Ingo Wegener (for posing the underlying question and giving some initial thoughts about the subject of this paper) and the reviewers who provided detailed and helpful comments.

## References

1. Wegener, I.: Towards a theory of randomized search heuristics. In: Proc. 28th Int'l Symposium on Mathematical Foundations of Computer Science (MFCS). Volume 2747 of LNCS., Springer (2003) 125–141
2. Kolda, T.G., Lewis, R.M., Torczon, V.: Optimization by direct search: New perspectives on some classical and modern methods. *SIAM Review* **45**(3) (2004) 385–482
3. Rechenberg, I.: Cybernetic solution path of an experimental problem. Royal Aircraft Establishment (1965)
4. Brooks, S.H.: A discussion of random methods for seeking maxima. *Operations Research* **6**(2) (1958) 244–251
5. Wegener, I.: Randomized search heuristics as an alternative to exact optimization. In: Logic versus Approximation. Volume 3075 of LNCS., Springer (2004) 138–149
6. Fang, K.T., Kotz, S., Ng, K.W.: Symmetric multivariate and related distributions. Volume 36 of Monographs on statistics and applied probability. Chapman & Hall, London (1990)
7. Jägersküpper, J.: Analysis of a simple evolutionary algorithm for minimization in Euclidean spaces. In: Proc. 30th Int'l Colloquium on Automata, Languages and Programming (ICALP). Volume 2719 of LNCS., Springer (2003) 1068–79
8. Jägersküpper, J.: Algorithmic analysis of a basic evolutionary algorithm for continuous optimization. *Theoretical Computer Science* **379**(3) (2007) 329–347
9. Jägersküpper, J.: Lower bounds for hit-and-run direct search. In: Proc. 4th Int'l Symposium on Stochastic Algorithms: Foundations and Applications (SAGA). Volume 4665 of LNCS., Springer (2007) 118–129
10. Rowe, J.E., Hidovic, D.: An evolution strategy using a continuous version of the Gray-code neighbourhood distribution. In: Proc. Genetic and Evolutionary Computation Conference (GECCO). Volume 3102 of LNCS., Springer (2004) 725–736
11. Dietzfelbinger, M., Rowe, J.E., Wegener, I., Woelfel, P.: Tight bounds for blind search on the integers. In: Proc. 25th Annual Symposium on Theoretical Aspects of Computer Science (STACS). Volume 8001 of Dagstuhl Seminar Proceedings., IBFI Schloss Dagstuhl, Germany (2008) 241–252
12. Jägersküpper, J.: Lower bounds for randomized direct search with isotropic sampling. *Operations Research Letters* **36**(3) (2008) 327–332
13. Hansen, N., Ostermeier, A.: Completely derandomized self-adaptation in evolution strategies. *Evolutionary Computation* **9**(2) (2001) 159–195