Adaptive Noisy Optimization

Philippe Rolet, Olivier Teytaud

Adaptive Noisy Optimization

Philippe Rolet and Olivier Teytaud

TAO (Inria), Lri, Cnrs Umr 8623, Univ. Paris-Sud, F-91405 Orsay, France

Abstract. In this paper, adaptive noisy optimization on variants of the noisy sphere model is considered, i.e. optimization in which the *same* algorithm is able to adapt to several frameworks, including some for which no bound has never been derived. Incidentally, bounds derived by [16] for noise quickly decreasing to zero around the optimum are extended to the more general case of a positively lower-bounded noise thanks to a careful use of Bernstein bounds (using empirical estimates of the variance) instead of Chernoff-like variants.

1 Introduction

Noisy optimization is a critical part of optimization since many real-world applications are noisy. It is sometimes called "stochastic optimization" [17,13,6,14], but "stochastic optimization" now often refers to the optimization of deterministic fitness functions by stochastic algorithms. Therefore we will here use "noisy optimization". Noisy optimization often distinguishes between (i) cases in which the variance of the noise quickly decreases to zero around the optimum and (ii) cases in which the variance of the noise is lower bounded. In the literature, various theoretical analyses of complexity bounds can be found for (i), while works covering (ii) are scarce. This paper is concerned with an algorithm covering both frameworks. Various works [8,9,1] have investigated noisy optimization from a theoretical point of view, often with a rough mathematical analysis and sometimes with rigorous arguments (as e.g. [12,16]). In particular, some recent papers investigated the use of *bandit algorithms* [10], inspired from the multi-armed bandit framework (see e.g. [2]), that rely on concentration inequalities such as Hoeffding confidence bounds. The following work proposes a rigorous runtime analysis of noisy expensive optimization based on such a bandit algorithm, in frameworks that are not covered by previously published papers. Specifically, it is shown that the same algorithm can be optimal (within logarithmic factors) for several families of fitness functions simultaneously, extending results of [16] to a quite broader class of noisy fitnesses. The paper is organized as follows. Section 2 defines the framework and introduces some notations. Section 3 states lower bounds that have been derived for this framework in the extant literature, and briefly discusses possible expectations. Section 4 presents *races*, a central tool for our subsequent results. Section 5 introduces the algorithm and proves an upper bound on its runtime. Section 6 provides straightforward extensions of this analysis and pinpoints possible further work.

C. Di Chio et al. (Eds.): EvoApplications 2010, Part I, LNCS 6024, pp. 592–601, 2010.

2 Framework

The *black-box* optimization framework is described in Algorithm 1 (similar to [16], but for wider families of fitness functions): the algorithm can request fitness values at any point of the domain, and no other information on the fitness function is available. The paper is interested in *expensive* optimization: it is assumed that obtaining a fitness value is much more costly than running the optimization algorithm. Therefore, the complexity is measured by the number of requests to the fitness. Let X be a domain, and $f : X \times X \to [0, \infty[$ such that $\forall (x, x^*) \in X^2, f(x, x^*) > f(x^*, x^*)$. The class of fitness functions we consider is $\{x \mapsto f(x, t) | t \in X\}$, thus each fitness $f(., t)$ is parameterized by the (unknown) location of its optimum, t. The goal is to find the optimum t (also referred to as x^* in the sequel) of $f(., t)$, by observing noisy measurements of $f(., t)$ at requested points x_i. In the following, t is not handled stochastically, i.e. the lower bounds are not computed *in expectation* w.r.t. all the possible fitness functions yielded by different values of t. Rather, the worst case on t will be considered. For simplicity, we only deal with deterministic algorithms; the extension to stochastic algorithms is straightforward by including a random seed in the algorithm.

Noisy measurements of the fitness at point x are modeled by a random variable $Rand(x)$ satisfying

$$Rand(x) \in [0, 1], \quad \mathbb{E}[Rand(x)] = f(x, x^*) \tag{1}$$

This noise model fits, among others, the goal of finding for a given algorithm a set of parameters minimizing the probability of failure. It notably raises two issues: (i) since few assumptions on the distribution law are made, one cannot use the fact that the probability mass of $Rand(x)$ is centered on $f(x, x^*)$ (it would be the case in i.e. a gaussian noise model). (ii) it is not limited to values worse than those at the optimum as in previous analyses[12]. Importantly, while [1] emphasized that for many algorithms a residual error remains, ours is truly consistent (i.e. $\|x_n - x^*\| \to_\infty 0$) as shown in Theorem 1 of section 5.

Note that the second equation implies:

$$Var[Rand(x)] \leq \mathbb{E}[Rand(x)]. \tag{2}$$

A simple example (from [16]) is $Rand(x) = 1$ with probability $f(x, x^*)$ and 0 otherwise. It is worth noticing that the algorithm considered for proving convergence and upper bound on convergence rates is invariant by addition of a constant. Therefore, our analysis is not restricted to $Rand(x) \in [0, 1]$ since Eq. 2 could be adapted to $Var[Rand(x)] \leq f(x) - \inf_u \inf Rand(u)$ (inf here stands for "essential infimum"). [16] were interested in the sphere function ($f(x, x^*) = ||x - x^*||$), a special case in which the variance promptly decreases to 0 around the optimum. In the sequel, wider classes of fitness functions will be studied:

Scaled sphere function. $f(x, x^*) = \lambda ||x - x^*||$ (case that might be handled similarly to [16]).

Algorithm 1. Noisy optimization framework. *Opt* is a deterministic optimization algorithm; it takes as input a sequence of visited points and their measured fitness values, and outputs a new point to be visited. Noisy fitness values are noted y_n^t since they depend on the fitness $f(.,t)$'s optimum t . The goal is to find points x of the domain such that $f(x,t)$ is as small as possible. Algorithm *Opt* is successful on target function $f(.,t)$ if $Loss(t, Opt)$ is small.

Parameter: N, number of fitness evaluations
for $n \in [[0, N-1]]$ **do**
  $x_{n+1} = Opt(x_1, \ldots, x_n, y_1^t, \ldots, y_n^t)$
  y_{n+1}^t is a draw of random variable $Rand(x_{n+1})$ (see Eqs 1-2)
end for
$Loss(t, Opt) = f(x_N, t)$

Scaled and translated sphere function: (noted S-T sphere from here on). $f(x, x^*) = \lambda||x - x^*|| + c$ (not covered by [16], and fundamentally harder since the variances does *not* decrease to 0 around the optimum).

Transformed sphere. $f(x, x^*) = g(||x - x^*||)$ for some increasing mapping g from $[0, \infty[$ onto a subset of $[0, 1]$.

We consider, in all these cases, a domain X whose diameter satisfies $\sup_{(x,y)\in X^2} ||x - y|| \leq 1$, and $x^* \in X$, so that these settings are well defined. Straightforward extensions are discussed in section 6. In the paper, $[[a, b]] = [a, b] \bigcap \mathbb{N}$. If $(a, b) \in (\mathbb{R}^D)^2$, then $[a, b] = \{x \in \mathbb{R}^D, \forall i \in [[1, D]], a_i \leq x_i \leq b_i\}$.

3 Lower Bounds

In this section we discuss the lower bounds for each of the three models described above. The goal of the rest of the paper will be to show that these bounds are reached within logarithmic factors. Sections 4 and 5 will describe an algorithm which has these guaranteed results on the models discussed above. Interestingly, the algorithm is the same for all models.

Scaled sphere function. In this setting, a direct consequence of [16] is that the distance between the optimum and its approximation is at best $O(1/n)$ after n iterations (for any algorithm). Theorem 3 (sec. 5) shows that this rate can be reached within logarithmic factors.

S-T sphere. Nothing has been poved in this case, to the best of our knowledge. No formal proofs on the matter will be given here; however, here are some intuitive ideas on the behavior of the lower bound. With the S-T sphere function, the variance can be lower bounded by some positive constant c: $\inf_x Var[Rand(x)] > c > 0$. Therefore, evaluating a point n times leads to a confidence interval on its mean whose length is roughly $\sqrt{c/n}$. As a consequence, the precision for an estimate of fitness with n evaluations cannot be less than $\Theta(1/\sqrt{n})$. Since the precision on the fitness space is linear as a function of the precision in the search space, it is reasonable to believe that $\|x_n^+ - x_n^-\| = \Theta(1/\sqrt{n})$

is the best achievable rate. This rate can be reached by our algorithm, as shown in section 5 (Theorem 2).

Monotonically transformed sphere. If the transformation function g is an arbitrary monotonically increasing function, the problem can be made arbitrarily difficult. Therefore, we will only have to show that we guarantee convergence. This consistency will be proved in next section (Theorem 1).

4 Hoeffding/Bernstein Bounds; Their Application to Races

This section recalls some concentration inequalities necessary to analyze the complexity of the algorithm that will be used to prove upper bounds on convergence rates. These inequalities are aimed at quantifying the discrepancy between an average and an expectation. Here, we focus on bounded random variables. The well-known Hoeffding bounds [11] were the first to generalize bounds on binomial random variables to bounded random variables. For some of our purposes, an improved versions of these bounds accounting for the variance[5,3,4], known as Bernstein's bound, will be required. Writing a detailed survey of Hoeffding, Chernoff and Bernstein's bounds is beyond the scope of this paper. We will only present the Bernstein bound, within its application to *races*. A *race* between two or more random variables aims at distinguishing with high confidence random variables with better expectation from those with worse expectation.

Algorithm 2 presents a *Berstein race* for 3 random variables—it is called a Bernstein race because it makes use of the Bernstein confidence bound. The Bernstein race in this paper will be used to distinguish between points x_i of the domain X, based on random variables $Rand(x_i)$. At the end of the race, $3T$

Algorithm 2. Bernstein race between 3 points. Eq. 3 is Bernstein's inequality for estimating the precision for empirical estimates (see e.g. [7, p124]). $\hat{\sigma}_i$ is the empirical estimate of the standard deviation of point x_i's associated random variable $Rand(x_i)$ (it is 0 in the first iteration, which does not alter the algorithm's correctness). $\hat{f}(x)$ is the average of the fitness measurements at x.

$Bernstein(a_1, a_2, a_3, \delta')$
$T = 0$
repeat
  $T \leftarrow T + 1$
  Evaluate the fitness of points x_1, x_2, x_3 once, *i.e.* evaluate the noisy fitness at each of these points.
  Evaluate the precision:

$$\epsilon_{(T)} = 3 \log \left(\frac{3\pi^2 T^2}{6\delta'} \right) / T + \max_i \hat{\sigma}_i \sqrt{2 \log \left(\frac{3\pi^2 T^2}{6\delta'} \right) / T}. \quad (3)$$

until Two points (*good* ,*bad*) satisfy $\hat{f}(bad) - \hat{f}(good) \geq 2\epsilon$ — **return** $(good, bad)$

evaluations have been performed, therefore T is referred to as the halting time in the sequel. The reason why δ' is used in Alg. 2 as the confidence parameter instead of δ will appear later on. Let us define Δ as

$$\begin{aligned}\Delta = &\sup\{\mathbb{E}Rand(x_1), \mathbb{E}Rand(x_2), \mathbb{E}Rand(x_3)\} \\ &- \inf\{\mathbb{E}Rand(x_1), \mathbb{E}Rand(x_2), \mathbb{E}Rand(x_3)\}.\end{aligned}$$

It is known[15] that if $\Delta > 0$,

- with probability $1 - \delta'$, the Bernstein race is consistent:

$$\mathbb{E}Rand(good) < \mathbb{E}Rand(bad). \tag{4}$$

- the Bernstein race halts almost surely, and with probability at least $1 - \delta'$, the the halting time T verifies

$$T \leq K \log\left(\frac{1}{\delta' \Delta}\right) / \Delta^2 \quad \text{where } K \text{ is a universal constant.} \tag{5}$$

- if, in addition,

$$\Delta \geq C \sup\{\mathbb{E}Rand(x_1), \mathbb{E}Rand(x_2), \mathbb{E}Rand(x_3)\}, \tag{6}$$

  then the Bernstein race halts almost surely, and with probability at least $1 - \delta'$, the halting time T verifies

$$T \leq K' \log\left(\frac{1}{\delta' \Delta}\right) / \Delta \quad \text{where } K' \text{ depends on } C \text{ only.} \tag{7}$$

The interested reader is referred to [15] and references therein for more.

5 Upper Bounds for Noisy Optimization

Algorithm 3 (based on the Bernstein race discussed above) will be used for proving our upper bounds. This algorithm was proposed in [16], with a weaker version of races. In the present work, the race was improved so that it can deal with more general settings than those of [16]. Informally, the algorithm (adapted from [16] for the case of variance not decreasing to zero around the optimum) is as follows. The domain is a hyper-rectangle $[x_0^-, x_0^+]$ of $\mathbb{R}^D$. Define the *axes* of a hyper-rectangle as the lines parallel to any edge of the hyper-rectangle, and containing the center of the hyper-rectangle. At iteration n, the algorithm considers the axis on which the hyper-rectangle is the largest (any rule for breaking ties is allowed). Three points are placed along this axis, one at the center of the hyper-rectangle, and the two others at the two intersections between the axis and the hyper-rectangle's frontier.

Then, the algorithm uses the Bernstein race for selecting a point $good_n$ and a point bad_n among these three points, such that the $good_n$ point is closer to

Algorithm 3. Algorithm for optimizing noisy fitness functions. *Bernstein* denotes a Bernstein race, as defined in Algorithm 2. The initial domain is $[x_0^-, x_0^+] \in \mathbb{R}^d$. δ is the confidence parameter.

$n \leftarrow 0$
while True **do**
  $c = \arg\max_i (x_n^+)_i - (x_n^-)_i$ // Pick the coordinate with highest uncertainty
  $\delta_n^{\max} = (x_n^+)_c - (x_n^-)_c$
  for $i \in [[1, 3]]$ **do**
    ${x'}^i_n \leftarrow \frac{1}{2}(x_n^- + x_n^+)$. // Consider the middle point
    $({x'}^i_n)_c \leftarrow (x_n^-)_c + \frac{i-1}{2}(x_n^+ - x_n^-)_c$. //except that the c^{th} coordinate may take // 3 different values
  end for
  $(good_n, bad_n) = Bernstein({x'}^1_n, {x'}^2_n, {x'}^3_n, \frac{6\delta}{\pi^2(n+1)^2})$. // a good and a bad point
  Let H_n be the halfspace $\{x \in \mathbb{R}^D; ||x - good_n|| \leq ||x - bad_n||\}$.
  Split the domain: $[x_{n+1}^-, x_{n+1}^+] = H_n \cap [x_n^-, x_n^+]$.
  $n \leftarrow n + 1$
end while

the optimum than the bad_n point. The Bernstein race described in section 4 by algorithm 2 guarantees this with confidence $1 - \delta'$.[1]

In the *transformed sphere* models under analysis, $\mathbb{E}\ Rand(x)$ is increasing as a function of $\|x - x^*\|$, thus the optimum is in the hyper-rectangle $H = \{x \in \mathbb{R}^D; \|x - good_n\| \leq \|x - bad_n\|$ with probability $1 - \delta$. The first lemma for our proofs is given below:

Lemma 1. *Let $\delta > 0$, and let fitness f be an increasing transformation of the sphere function $x \mapsto \|x - x^*\|$.*[2] *Let $Rand(x)$ be the noisy answer to an evaluation of f as defined above. If, in algorithm 3, the Bernstein race halts at all steps until iteration n, then:*

$$\left(\frac{3}{4}\right)^n ||x_0^+ - x_0^-|| \leq ||x_n^+ - x_n^-|| \leq \left(\frac{3}{4}\right)^{\lfloor n/D \rfloor} ||x_0^+ - x_0^-||, \tag{8}$$

$$\text{and} \quad (\forall i < n, \mathbb{E}Rand(good_i) \leq \mathbb{E}Rand(bad_i)) \Rightarrow x^* \in [x_n^-, x_n^+], \tag{9}$$

and for some constant K depending on the dimension only,

$$x^* \in [x_n^-, x_n^+] \Rightarrow \exists (good_n, bad_n) \in \{{x'}^1_n, {x'}^2_n, {x'}^3_n\}^2, \tag{10}$$

$$\|x^* - bad_n\| \geq \|x^* - good_n\| + K||x_n^+ - x_n^-|| \tag{11}$$

[1] Note that this particular kind of race is not interested in knowing how good remaining points (other than $good_n$ and bad_n) are. It might be that in our case the third point is even closer to the optimum, but the point of this race is not to determine the closest point, it is to provide two points such that one is closer than the other.

[2] The transformed sphere covers models of the S-T sphere, and of the scaled sphere.

Due to length constraints, the proof of this lemma is not given here. A very similar lemma is used for the case of a variance decreasing to zero around the optimum, in [16]. □

A consequence of this lemma is the following convergence guarantee:

Theorem 1 (Consistency of Algo. 3 for the transformed sphere). *In the transformed sphere model, Algo. 3 ensures* $x_n^- \to x^*$ *and* $x_n^+ \to x^*$ *with probability at least* $1 - \delta$.

Proof
Eq. 9 of the previous lemma implies that $\|x_n^+ - x_n^-\| \to 0$. We will now show that with probability $1 - \delta$, $x^* \in [x_n^-, x_n^+]$ by establishing the left-hand side of Eq. 4 by induction. This will be sufficient to prove theorem 1.

- The induction hypothesis $\mathcal{H}(n)$ is as follows:

  With probability at least $1 - \sum_{k=1}^{n+1} \frac{6\delta}{\pi^2 k^2}, \forall i < n, \mathbb{E}Rand(good_i) \leq \mathbb{E}Rand(bad_i)$.

- $\mathcal{H}(0) : x^* \in [x_0^-, x_0^+]$ by definition.
- Let us assume $\mathcal{H}(n-1)$ for $n > 0$. For clarity, the statement $\forall i < n, \mathbb{E}Rand(good_i) \leq \mathbb{E}Rand(bad_i)$ is written $G(n)$.

$$\begin{aligned} P(G(n)) &= P(G(n-1), \mathbb{E}Rand(good_n) \leq \mathbb{E}Rand(bad_n)) \\ &= P(\mathbb{E}Rand(good_n) \leq \mathbb{E}Rand(bad_n) \mid G(n-1)) P(G(n-1)) \\ &= (1 - \sum_{k=1}^{n} \frac{6\delta}{\pi^2 k^2})(1 - \frac{6\delta}{\pi^2 (n+1)^2}) \qquad (12) \\ &\geq 1 - \sum_{k=1}^{n+1} \frac{6\delta}{\pi^2 k^2} \end{aligned}$$

  which proves $\mathcal{H}(n)$. The first term of eq. 12 is the application of $\mathcal{H}(n-1)$. The second term is a property of the Bernstein race described in Algo. 2, and used in Algo. 3.

It only remains to observe that $\sum_{i=1}^{\infty} (6\delta/(\pi i)^2) = \delta$ to conclude. □

The number of iterations is of course log-linear ($\log(\|x_n^- - x^*\|)/n$ is upper bounded by a negative constant), but the number of evaluations per iteration might be arbitrary large. More precise (faster) results, for the other (simpler) models will now be considered.

Theorem 2 (Hoeffding rates for the S-T sphere model). *Consider the S-T sphere model, and a fixed dimension* D. *The number of evaluations requested by Algo. 3 for reaching precision* ϵ *with probability at least* $1 - \delta$ *is* $O(\frac{\log(\log(1/\epsilon)/\delta)}{\epsilon^2})$.

Proof. Eq. 11 ensures that

$$\begin{aligned}\Delta_n =& \sup\{\mathbb{E}Rand({x'}^1_n), \mathbb{E}Rand({x'}^2_n), \mathbb{E}Rand({x'}^3_n)\} \\ & - \inf\{\mathbb{E}Rand({x'}^1_n), \mathbb{E}Rand({x'}^2_n), \mathbb{E}Rand({x'}^3_n)\} \\ & \geq \|x^* - bad_n\| - \|x^* - good_n\|\end{aligned}$$

verifies $\Delta_n = \Omega(\|x_n^+ - x_n^-\|)$. Therefore, applying the concentration inequality 5, the number of evaluations in the n^{th} iteration is at most

$$O\left(\log\left(\frac{6\delta}{\pi^2(n+1)^2}\right)/\|x_n^- - x_n^+\|^2\right). \tag{13}$$

Now, let us consider the number $N(\epsilon)$ of iterations before a precision ϵ is reached. Eq. 8 shows that there is a constant $k < 1$ such that

$$\epsilon \leq \|x_n^+ - x_n^-\| \leq Ck^{N(\epsilon)} \tag{14}$$

Injecting Eq. 14 in Eq. 13 shows that the cost (the number of evaluations) in the last call to the Bernstein race is

$$Bound_{last}(\epsilon) = O\left(-\log\left(\frac{6\delta}{\pi^2(N(\epsilon)+1)^2}\right)/\epsilon^2\right). \tag{15}$$

Since $N(\epsilon) = O(\log(1/\epsilon))$, $Bound_{last} = O(\log(\log(1/\epsilon')/\delta))$. For a fixed dimension D, there exists $k' > 1$ such that the cost of the $(N(\epsilon) - i)^{th}$ iteration is at most

$$O(\lceil Bound_{last}/(k')^i \rceil) \tag{16}$$

because the algorithm ensures that after D iterations, $\|x_n^+ - x_n^-\|$ decreases by at least 3/4.

The sum of the costs for $N(\epsilon)$ iterations is therefore the sum of $O(Bound_{last}/(k')^i)$ for $i \in [[0, N(\epsilon) - 1]]$, that is $O(Bound_{last}/(1 - k')) = O(Bound_{last})$ (plus $O(N(\epsilon))$ for the rounding associated to the $\lceil ... \rceil$ in Eq. 16).

The overall cost is therefore $O(Bound_{last} + \log(1/\epsilon))$. This yields the expected result. □

Theorem 3 (Bernstein rates for the scaled sphere model). *Consider the scaled sphere model, and a fixed dimension D. Then, the number of evaluations requested for reaching precision ϵ with probability at least $1-\delta$ is $O(\frac{\log(\log(1/\epsilon)/\delta)}{\epsilon})$.*

Proof. The proof follows the lines of the proof of Theorem 2, except for one point. As well as for Theorem 2, we use the fact that for the scaled sphere model (and in fact also for the S-T sphere model), Eq. 11 holds, which implies (with $\Delta_n = \sup_i \mathbb{E}Rand({x'}^i_n) - \inf_i \mathbb{E}Rand({x'}^i_n)$)

$$\Delta_n = \Omega(\|x_n^+ - x_n^-\|). \tag{17}$$

However, for the scaled sphere model, we can also claim

$$\sup_i \mathbb{E}Rand({x'}^i_n) = O(\|x_n^+ - x_n^-\|). \tag{18}$$

Eqs. 17 and 18 lead to Eq. 6.

Furthermore, Eq. 6 implies that Eq. 15 can be replaced by

$$Bound_{last}(\epsilon) = O\left(-\log\left(\frac{6\delta}{\pi^2(N(\epsilon)+1)^2}\right)/\epsilon\right). \tag{19}$$

The summation as in the proof of Theorem 2 now leads to an overall cost $O(\frac{\log(\log(1/\epsilon)/\delta)}{\epsilon})$. □

6 Discussion

We considered the optimization of noisy fitness functions, where the fitness in x is randomized, with values in $[0,1]$, and expected value $f(x,x^*)$ where x^* is the optimum. The following models were studied: (i) Sphere function: $f(x,x^*) = \|x-x^*\|$; (ii) Scaled sphere function: $f(x,x^*) = \lambda\|x-x^*\|$; (iii) S-T sphere function: $f(x,x^*) = \lambda\|x-x^*\|+c$; (iv) Transformed sphere: $f(x,x^*) = g(\|x-x^*\|)$. The first case only was in the state of the art. The same algorithm (using Bernstein's inequality) ensures that with probability $1-\delta$, the optimum x^* is in a set of diameter δ_n (after n fitness evaluations), which provably decreases as shown in Table 1. There are some straightforward extensions, the main one being that convergence rates only depends on $f(x,x^*)$ for x close to x^*: all f such that $f(x,x^*) = \Theta(||x-x^*||)$ lead to the same asymptotic rate as the scaled sphere; and all f such that $f(x,x^*) - c = \Theta(||x-x^*||)$ lead to the same asymptotic rate as the scaled and translated sphere function. Therefore, it is likely that the proposed approach is much more general than variants of the sphere model as formally considered here. It has been shown in [16] that some links exist between the rates for $f(x,x^*) = ||x-x^*||$ and $f(x,x^*) = ||x-x^*||^p$; these links will not be developed here. The main further works are: (i) formalizing the lower bound for the case of the scaled and translated sphere function; (ii) experiment real-world algorithms or adapted version of real-world algorithms (as e.g. [10]) on these fitness functions.

Table 1. Precision δ_n (diameter of the region in which might be the optimum) as a function of the number n of fitness evaluations. The $\tilde{O}(.)$ means that logarithmic factors are present. Dependencies in δ can be found in detailed results; the dependency in the dimension can be computed from the proofs, but we guess they are not optimal. The constants depend on λ and on the dimension; c has no impact on the constant for the scaled and translated sphere function.

Model	Precision $\|x_n^- - x_n^+\|$
Sphere function	$\tilde{O}(1/n)$
Scaled sphere function	$\tilde{O}(1/n)$
Scaled and translated sphere function	$\tilde{O}(1/\sqrt{n})$
Transformed sphere	$o(1)$

References

1. Arnold, D.V., Beyer, H.-G.: Efficiency and mutation strength adaptation of the (mu/mui,lambda)-es in a noisy environment. In: Deb, K., Rudolph, G., Lutton, E., Merelo, J.J., Schoenauer, M., Schwefel, H.-P., Yao, X. (eds.) PPSN 2000. LNCS, vol. 1917, pp. 39–48. Springer, Heidelberg (2000)
2. Auer, P.: Using confidence bounds for exploitation-exploration trade-offs. The Journal of Machine Learning Research 3, 397–422 (2003)
3. Bernstein, S.: On a modification of chebyshev's inequality and of the error formula of laplace. Original publication: Ann. Sci. Inst. Sav. Ukraine, Sect. Math. 1 3(1), 38–49 (1924)
4. Bernstein, S.: The Theory of Probabilities. Gastehizdat Publishing House, Moscow (1946)
5. Chernoff, H.: A measure of asymptotic efficiency for tests of a hypothesis based on the sum of observations. Annals of Math. Stat. 23, 493–509 (1952)
6. Denton, B.: Review of "stochastic optimization: Algorithms and applications" by stanislav uryasev and panos m. Interfaces 33(1), 100–102 (2003)
7. Devroye, L., Györfi, L., Lugosi, G.: A probabilistic Theory of Pattern Recognition. Springer, Heidelberg (1997)
8. Fitzpatrick, J.M., Grefenstette, J.J.: Genetic algorithms in noisy environments. Machine Learning 3, 101–120 (1988)
9. Hammel, U., Bäck, T.: Evolution strategies on noisy functions: How to improve convergence properties. In: Davidor, Y., Männer, R., Schwefel, H.-P. (eds.) PPSN 1994. LNCS, vol. 866, pp. 159–168. Springer, Heidelberg (1994)
10. Heidrich-Meisner, V., Igel, C.: Hoeffding and bernstein races for selecting policies in evolutionary direct policy search. In: ICML 2009: Proceedings of the 26th Annual International Conference on Machine Learning, pp. 401–408. ACM, New York (2009)
11. Hoeffding, W.: Probability inequalities for sums of bounded random variables. Journal of the American Statistical Association 58, 13–30 (1963)
12. Jebalia, M., Auger, A.: On multiplicative noise models for stochastic search. In: Rudolph, G., Jansen, T., Lucas, S., Poloni, C., Beume, N. (eds.) PPSN 2008. LNCS, vol. 5199, pp. 52–61. Springer, Heidelberg (2008)
13. Kall, P.: Stochastic Linear Programming. Springer, Berlin (1976)
14. Marti, K.: Stochastic Optimization Methods. Springer, Heidelberg (2005)
15. Mnih, V., Szepesvári, C., Audibert, J.-Y.: Empirical Bernstein stopping. In: ICML 2008: Proceedings of the 25th international conference on Machine learning, pp. 672–679. ACM, New York (2008)
16. Rolet, P., Teytaud, O.: Bandit-based estimation of distribution algorithms for noisy optimization: Rigorous runtime analysis. Submitted to Lion4; presented in TRSH 2009 in Birmingham (2009)
17. Sengupta, J.K.: Stochastic Programming. Methods and Applications. North-Holland, Amsterdam (1972)