Noisy Optimization Complexity Under Locality Assumption

Jérémie Decock, Olivier Teytaud

Noisy Optimization Complexity Under Locality Assumption

Jérémie Decock
TAO team
Inria Saclay IDF, UMR CNRS 8623
Université Paris-Sud, Orsay, France
jeremie.decock@inria.fr

Olivier Teytaud
TAO team
Inria Saclay IDF, UMR CNRS 8623
Université Paris-Sud, Orsay, France
olivier.teytaud@inria.fr

ABSTRACT

In spite of various recent publications on the subject, there are still gaps between upper and lower bounds in evolutionary optimization for noisy objective function. In this paper we reduce the gap, and get tight bounds within logarithmic factors in the case of small noise and no long-distance influence on the objective function.

Categories and Subject Descriptors

G.1.6 [**Mathematics of Computing**]: Numerical Analysis—*Optimization*

General Terms

Theory

Keywords

Noisy optimization; black box complexity model; local sampling

1. INTRODUCTION

Noisy optimization is the optimization of stochastic objective functions, i.e. the objective value $f(\boldsymbol{x}, \omega)$ depends on $\boldsymbol{x}$ and on a random variable ω. Equivalently, $f(\boldsymbol{x})$ is a random variable distributed as $f(\boldsymbol{x}, \omega)$. Often (yet not always, as risk considerations can be involved), the goal is to optimize the expected value, i.e. $\mathbb{E}_\omega f(\boldsymbol{x}, \omega)$ where $\mathbb{E}_\omega$ is the expectation with respect to ω. We here work on derivative-free noisy optimization, on a continuous domain $D = [0, 1]^d$; we assume that the optimum is unique, and the detailed setting below will assume that getting rid of tricky local optima is less important than handling noise properly.

In all the paper, we use $\tilde{O}(f(n))$ as a short notation for $O(f(n)\log(n))$. Section 1.1 presents our framework, and in particular the family of functions on which we show lower bounds (all families including this family are also concerned by the lower bounds). Section 1.2 discusses the context of our results, in particular the critical "locality" assumption, discussing whether an algorithm uses points far from the optimum or not for improving its convergence rate.

Then, Section 2 gives the mathematical formulation and the main lemmas. Section 3 is the main result.

FOGA'13, January 16-20, 2013, Adelaide, Australia.

1.1 Framework

There are many convergence rates known in numerical optimization, depending on assumptions (derivative-free[1] or not, comparison-based [2] or not, global[3] or not). Robustness to noise, even early in the origins of evolutionary computation[4], is cited as a strength of these algorithms. In spite of this strong interest for noise in evolutionary computation, complexity in the noisy case is less clear, because details on the assumption have a big impact on the performance[5, 6, 7, 8, 9, 10]; a main distinction being adaptive noise[11] (with small noise variance around the optimum) and additive noise[12] (with lower-bounded variance around the optimum). Also, defining convergence rates is more difficult when either the algorithm or the objective function has a random part, because the distance of $\boldsymbol{x}_n$ to the optimum $\boldsymbol{x}^*$ is not a constant (n being the n^{th} objective function evaluation). Various convergence rates can be defined depending on the precise considered definition of convergence (we will not give too many details on this in this introduction, but our results will state precisely the definitions involved).

For example, the algorithm CLOP[13, 14] reaches a convergence $\|\boldsymbol{x}_n - \boldsymbol{x}^*\| = \tilde{O}(1/\sqrt{n})$on a wide range of symmetric objective functions, using regression; in a very structured case (when the family of functions is very well approximated by a known model), statistical tools (supervised machine learning) provide such fast rates, even on very flat functions.

An other example of algorithm for noisy optimization is R-EDA (Racing-based Estimation of Distribution Algorithm, [15, 16]). It considers a different definition of convergence (more on this later).R-EDA was proposed as a typical noisy optimization algorithm, easy to analyze and making a good approximation of real-world approaches[17]. The noisy optimization framework is described in Algorithm 1 and R-EDA is defined in Algorithm 2. For $\beta > 0$, the convergence rate is typically $\|\boldsymbol{x}_n - \boldsymbol{x}^*\| = \tilde{O}(1/n^{1/\beta})$ on the family of objective functions

$$F_{\beta,\gamma} = \{f_{\boldsymbol{x}^*,\beta,\gamma}(\boldsymbol{x}); \boldsymbol{x}^* \in D\}\,, \tag{1}$$

where $D = [0, 1]^d$ and

$$f_{\boldsymbol{x}^*,\beta,\gamma}(\boldsymbol{x}) = \mathcal{B}\left(\gamma\left(\frac{\|\boldsymbol{x} - \boldsymbol{x}^*\|}{\sqrt{d}}\right)^\beta + (1-\gamma)\right). \tag{2}$$

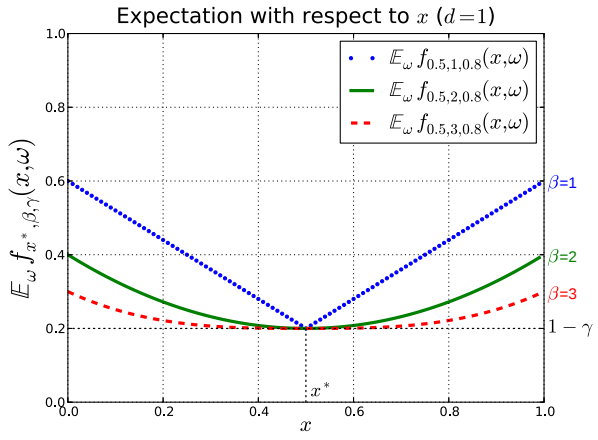

Figure 1: Expected values of $f_{\boldsymbol{x}^*,\beta,\gamma}$ with respect to x, for $d = 1$ and with the optimal point $\boldsymbol{x}^* = 0.5$, the noise level $\gamma = 0.8$ and β (the "flatness" of $\mathbb{E}f$ around $\boldsymbol{x}^*$) equals to 1, 2 and 3.

Here $\mathcal{B}(p)$ states for a Bernoulli random variable with parameter p (i.e. equal to 1 with probability p and 0 otherwise). Fig. 1 show the expected values of $f_{\boldsymbol{x}^*,\beta,\gamma}$ with respect to $\boldsymbol{x}$ for some set of parameters.

We will show in this paper that R-EDA is optimal within logarithmic factors for $F_{\beta,\gamma}$ under locality assumptions discussed below and for some values of γ (case with variance linearly decreasing as a function of expected fitness values).

Algorithm 2 R-EDA: algorithm for optimizing noisy fitness functions. *Bernstein* denotes a Bernstein race, as defined in Algorithm 3. The initial domain is $[\boldsymbol{x}_0^-, \boldsymbol{x}_0^+] \in \mathbb{R}^D$, δ is the confidence parameter. This algorithm goes back to [15, 16]. Please note that $\boldsymbol{x}_i^-$ and $\boldsymbol{x}_i^+$ are indexed by i, the iteration number, and not by the number of evaluations as in our convergence criteria.

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

Algorithm 3 Bernstein race between 3 points. Eq. 3 is Bernstein's inequality to compute the precision for empirical estimates (see e.g. [18, p124]); $\hat{\sigma}_i$ is the empirical estimate of the standard deviation of point $\boldsymbol{x}_i$'s associated random variable $F_t(\boldsymbol{x}_i)$ (it is 0 in the first iteration, which does not alter the algorithm's correctness); $\hat{f}(\boldsymbol{x})$ is the average of the fitness measurements at $\boldsymbol{x}$.

Bernstein$(\boldsymbol{x}_1, \boldsymbol{x}_2, \boldsymbol{x}_3, \delta')$
$T = 0$
repeat
  $T \leftarrow T + 1$
  Evaluate the fitness of points $\boldsymbol{x}_1, \boldsymbol{x}_2, \boldsymbol{x}_3$ 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*)

1.2 Symmetry assumptions and information theory

We pointed out above that the $\tilde{O}(1/\sqrt{n})$ is possible for algorithms (e.g. CLOP) using models which are very close to the real objective function (for example if we know that the fitness function is a Bernoulli as in Eq. 2). In these algorithms, the sampled point is not necessarily close to the optimum or to the current approximation of the optimum that the algorithm has. This is related to information theory: there are areas in which one gets more information than others, and points minimizing $\boldsymbol{x} \mapsto E_\omega f(\boldsymbol{x}, w)$ are not necessarily the most informative. Typically, when you have a relevant model of the objective function, you will learn more about the optimum by sampling maximum uncertainty points (i.e. points $\boldsymbol{x}$ such that $f(\boldsymbol{x}, \omega)$ has high variance), rather than by sampling points close to the optimum.

There are therefore two distinct strategies[1]:

- sampling close to the current estimation of the optimum;
- sampling maximum uncertainty areas.

As we have already said, getting knowledge on the objective function far from the current approximation of the optimum does not help for finding $\boldsymbol{x}^*$ if you have no model of the objective function (at least in a setting without tricky local optima). But, if you have a strong prior on the objective function, you can indeed sample only very far from the current estimation of the optimum, at locations where the objective function is less flat and from this sampling, you get knowledge on $\boldsymbol{x}^*$. Again, this leads to algorithms which sample much more far from the current approximation of the optimum than close to it.

[1]Interestingly there were debates on the mailing list dedicated to the BBOB noisy optimization testbeds because the designers of the testbeds assess the quality of optimization algorithms not from their estimation on the location of the optimum, but from the points they sample, which is clearly not fair for algorithms which are precisely based on estimating the optimum from points far from the optimum.

Algorithm 1 Noisy optimization framework. This is not an optimization algorithm; this just explains the framework of noisy optimization with Bernoulli random variables (the fitness function outputs 1 if random ($=\omega_n$) is less than $\mathbb{E}f_{\boldsymbol{x}^*,\beta,\gamma}(\boldsymbol{x}_{\boldsymbol{x}^*,n,\omega,\omega'})$). *Optimize* is an optimization algorithm taking as input a sequence of visited points, their binary noisy fitness values and an internal noise. It outputs a new point to be visited, looking for points $\boldsymbol{x}$ of the domain such that $f_{\boldsymbol{x}^*,\beta,\gamma}(\boldsymbol{x})$ is as small as possible.

Input:
ω the uniform noise of f,
ω' a random seed of the algorithm,
$\boldsymbol{x}^*$ the optimal point,
β and γ two fixed parameters of f.
Output:
$\boldsymbol{x}_{\boldsymbol{x}^*,n,\omega,\omega'}$ the estimation of the optimum.

for all $n = 1, 2, 3, \dots$ **do**
  $\boldsymbol{x}_{\boldsymbol{x}^*,n,\omega,\omega'} = \text{Optimize}(\boldsymbol{x}_{\boldsymbol{x}^*,1,\omega,\omega'},\ \dots,\ \boldsymbol{x}_{\boldsymbol{x}^*,n-1,\omega,\omega'},\ y_1,\ \dots,\ y_{n-1},\ \omega')$
  if $\omega_n \leq \mathbb{E}f_{\boldsymbol{x}^*,\beta,\gamma}(\boldsymbol{x}_{\boldsymbol{x}^*,n,\omega,\omega'})$ **then**
    $y_n = 1$
  else
    $y_n = 0$
  end if
end for
return $\boldsymbol{x}_{\boldsymbol{x}^*,n,\omega,\omega'}$

These algorithms have two distinct steps:

- Exploration: choosing a point $\boldsymbol{x}_n$ for which they want to sample $f(\boldsymbol{x}_n)$.
- Recommendation: providing an estimate $\tilde{\boldsymbol{x}}^*_n$ of $\arg\min \mathbb{E}f$.

Nevertheless, various compromises are possible (for not relying too much on the model) between strategies relying only on maximum uncertainty (strongly trusting a model) and strategies relying only on sampling close to the estimation of the optimum; [19] is based on methods for choosing to which extent the model should be trusted.

So far, we have discussed the distinction between algorithms which samples close to the estimation of the optimum and those which samples at maximum uncertainty areas. But in this paper we will only consider the fastest theoretical convergence rates for algorithms which samples close to the estimation of the optimum. Formally speaking, we assume the following *locality assumption* for some $0 < \delta < 1/2$ and some constant $C(d) > 0$ depending on d only,

$$\forall f \in F, \forall i \leq n, \|\boldsymbol{x}_i - \boldsymbol{x}^*\| \leq \frac{C(d)}{i^\alpha}, \tag{4}$$

with probability at least $1 - \delta/2$; $\alpha > 0$ large implies that there is a fast convergence. This equation depends on n; in fact, the whole work would make sense with n replaced by ∞, but we can show our results for any n sufficiently large, so we keep this assumption under this form (the main theorem will assume this for all n, but results in it are derived for n sufficiently large).

Actually, convergence rates could be formalized differently. For example, we might consider that the rate is α if

$$\forall f \in F, \forall n, \|\boldsymbol{x}_{k_n} - \boldsymbol{x}^*\| \leq \frac{C(d)}{{k_n}^\alpha}. \tag{5}$$

for some increasing sequence k_n. This is a weaker assumption, because only $\boldsymbol{x}_i$ such that $\exists n; i = k_n$ have a fast rate; other points can be sampled anywhere in the domain without modifying the measure α of the convergence rate. For instance, the following algorithm satisfies Eq. 5 but not Eq. 4, for $0 < \alpha < 1/2$:

- define $k_n = n^2$;
- if i different from k_n for all n, then do *exploration* ($\boldsymbol{x}_i$ is uniformly drawn on the domain);
- otherwise, do *exploitation* ($\boldsymbol{x}_i$ is the maximum likelihood estimate of $\boldsymbol{x}^*$ given the $\boldsymbol{x}_i$ and their fitness evaluations).

Indeed, this algorithm is quite good for the objective function model that we have chosen (see Eq. 2); but it does not satisfy Eq.4 as some points are sampled far from the optimum. In fact the algorithm above can even be optimized by choosing $\boldsymbol{x}_i$, for exploration, at locations where it is most likely to help finding the optimum (this is active learning); see e.g. [20, 21]. Also, sometimes, the $\boldsymbol{x}_i$ for exploitation are computed, but not evaluated.

When designing a testbed for noisy optimization, this issue makes sense. Many optimization algorithms for noisy settings distinguish $\boldsymbol{x}_i$'s which are supposed to be good approximations of the optimum and $\boldsymbol{x}_i$'s which are sampled for gathering information about the optimum. If the testbed makes no difference between the two kinds of points, it implicitly assumes that sampling far from the optimum for gathering information is unlikely to be a good method. Rates reachable with no constraints are a well established part of the state of the art[22]; we here focus on rates which can be attained when focusing on Eq. 4 as a criterion for convergence. Importantly, this criterion is also relevant when all fitness values sampled are important; e.g. when improving, online, a production unit.

To sum up, the locality assumption (Eq.4) used to obtain our main result means that the algorithm has a given rate if and only if all its search points follow this rate; it is not allowed, for instance, to have one point out of two which

is close to the optimum, and another far away in order to get some information which, for some reason, would help for the convergence. In other words, this assumption means that we consider rates that can be reached without relying on long distance correlations between fitness values. This is by no mean a negligible technical detail; as we have already said, there are fast algorithm which rely on sampling far from the optimum; these fast algorithms, however, can only be fast when there is a strong structure on the objective functions so that sampling far away can provide significant improvements on the convergence rate. This paper is devoted to showing bounds on rates for algorithms which do not use such sampling "far" from the current estimate of the optimum.

The results can therefore be viewed with two different complementary conclusions:

- either as the proof that fast rates (faster than the limits obtained in this paper) can only be obtained by sampling also far from the optimum;
- or as the proof that fast rates (faster than the limits obtained in this paper) are only possible for algorithm which assume some strong "flatness" of the objective function around the optimum, and these algorithms will not be so fast if we test them on other objective functions.

Under the locality assumption (Eq. 4), we show that for the family $F_{\beta,\gamma}$ of objective functions ($\gamma > 0$), α is necessarily less than or equal to $1/\beta$ - if the function is very flat around the optimum (β large), the convergence of algorithms sampling close to the optimum (Eq. 4) is necessarily slow ($\alpha \leq 1/\beta$ in Eq. 4). On the other hand, for $\beta = 2$, local algorithms (like R-EDA) have the same order as the rate reached by machine learning methods, and can even be better for $\beta = 1$, reaching $\tilde{O}(1/n)$ when $\beta = 1$ instead of $\tilde{O}(1/\sqrt{n})$ by max-uncertainty sampling.

2. MODELS AND LEMMAS

Algorithm 1 describes our framework. As introduced in section 1.2 (Eq. 4), we assume, for some $0 < \delta < 1/2$, the locality assumption

$$\forall f \in F, \forall i \leq n, \|\boldsymbol{x}_i - \boldsymbol{x}^*\| \leq \frac{C(d)}{i^\alpha}$$

with probability at least $1 - \delta/2$. This contains two important elements:

- there is an $\tilde{O}(1/n^\alpha)$ convergence to the optimum;
- there is no sampling far from the current estimate of the optimum.

This implies that the algorithm converges and does not sample far from its limit.

As already stated in Eq. 1 and 2, we also consider the family of functions

$$F = F_{\beta,\gamma} = \{f_{\boldsymbol{x}^*,\beta,\gamma}(\boldsymbol{x}) \ ; \ \boldsymbol{x}^* \in D\},$$

where

$$f_{\boldsymbol{x}^*,\beta,\gamma}(\boldsymbol{x}) = \mathcal{B}\left(\gamma\left(\frac{\|\boldsymbol{x}_n - \boldsymbol{x}^*\|}{\sqrt{d}}\right)^\beta + (1-\gamma)\right),$$

that is to say the random variable ω is uniform in $[0,1]$ and $f(\boldsymbol{x},\omega) = 1$ if and only if $\omega \leq \gamma\left(\frac{\|\boldsymbol{x}_n - \boldsymbol{x}^*\|}{\sqrt{d}}\right)^\beta + (1-\gamma)$ ($f(\boldsymbol{x},\omega) = 0$ otherwise).

Equation 2 and the locality assumption (Eq. 4) imply

$$\underbrace{\mathbb{E}f(\boldsymbol{x}^*)}_{1-\gamma} \ \leq \ \mathbb{E}f(\boldsymbol{x}_n) \ \leq \ \underbrace{\mathbb{E}f(\boldsymbol{x}^*)}_{1-\gamma} + \frac{\gamma}{d^{\beta/2}}\frac{C(d)^\beta}{n^{\alpha\beta}},$$

with probability at least $1 - \delta/2$ and where $\mathbb{E}f(\boldsymbol{x})$ is a short notation for $\mathbb{E}_\omega f(\boldsymbol{x},\omega)$.

The n^{th} function evaluation y_n is, by definition, at $\boldsymbol{x}_{\boldsymbol{x}^*,n,\omega,\omega'}$ which depends on $\boldsymbol{x}^*$ (which specifies the objective function), ω (which is the sequence of random variables ω of the stochasticity of the objective function), ω' (which is the sequence of random choices within the optimization algorithm, which might be stochastic). It also depends on β and γ, but these will be considered as fixed.

For short, we will note $X_{n,\Omega}$, with $\Omega = (\omega,\omega')$ the set of all the $\boldsymbol{x}_{\boldsymbol{x}^*,n,\omega,\omega'}$ for all $\boldsymbol{x}^*$, that is to say $X_{n,\Omega}$ is the set of all points which can be chose by the optimization algorithm *Optimize* for the nth point to sample if we consider a given noise and internal randomness.

To obtain our main result, we first need a combinatorial lemma as follows:

LEMMA 1. *The cardinality of $X_{n,\Omega}$ is at most 2^N, where N is the cardinality of*

$$\left\{1 \leq i \leq n \ ; \ \underbrace{\mathbb{E}f(\boldsymbol{x}^*)}_{1-\gamma} \leq \omega_i \leq \underbrace{\mathbb{E}f(\boldsymbol{x}^*)}_{1-\gamma} + \frac{\gamma}{d^{\beta/2}}\frac{C(d)^\beta}{i^{\alpha\beta}}\right\}.$$

PROOF. $\boldsymbol{x}_n$ is deterministic as a function of Ω and of the fitness values; so the possible values of $\boldsymbol{x}_{\boldsymbol{x}^*,n,\omega,\omega'}$ only depend on the 2^N possible values of the y_i for i in the set above. □

We also need the following

LEMMA 2. *Let N be the cardinality of*

$$\left\{1 \leq i \leq n \ ; \ \mathbb{E}f(\boldsymbol{x}^*) \leq \omega_i \leq \mathbb{E}f(\boldsymbol{x}^*) + \frac{\gamma}{d^{\beta/2}}\frac{C(d)^\beta}{i^{\alpha\beta}}\right\}.$$

Then, N has expectation at most

$$z = \frac{\gamma}{d^{\beta/2}}C(d)^\beta \sum_{i=1}^{n} i^{-\alpha\beta} \tag{6}$$

and variance also at most z.

which is an immediate consequence of the definition of N.

Lemma 2 and Chebyshev's inequality[23, 24, 25] ensure the following lemma:

LEMMA 3. *Consider $\delta \in [0,1]$. $N \leq z + \sqrt{z}\,(\delta/2)^{-1/2}$ with probability at least $1 - \delta/2$.*

Lemmas 1 and 3 together imply that the cardinality of $X_{n,\Omega}$ is at most

$$2^{z + \frac{\sqrt{z}}{\sqrt{\delta/2}}} \tag{7}$$

with probability at least $1 - \delta/2$.

3. MAIN RESULTS

THEOREM 1. *Assume that F is as proposed in Eq. 2 for some $1 > \gamma > 0$ and $\beta > 0$. Assume that Eq. 4 (locality assumption) holds for all $n \geq 1, d \geq 1$, and for some $C(d)$, $\delta < 1$, $\alpha > 0$, i.e.*

$$\forall f \in F, \forall i \leq n, \|\boldsymbol{x}_i - \boldsymbol{x}^*\| \leq \frac{C(d)}{i^\alpha},$$

with probability at least $1 - \delta/2$. Then $\alpha \leq 1/\beta$.

PROOF. Let us show that $\alpha\beta \leq 1$. In order to do so, let us assume, in order to get a contradiction, that $\alpha\beta > 1$; then, knowing convergence of Riemann series for $\alpha\beta > 1$

$$\sum_{i=1}^{n} \frac{1}{i^{\alpha\beta}} \quad < \quad \frac{\alpha\beta}{\alpha\beta - 1}$$

equation 6 leads to:

$$z \leq \frac{\gamma C(d)^\beta}{d^{\beta/2}} \frac{\alpha\beta}{\alpha\beta - 1} \text{ if } \alpha\beta > 1 \tag{8}$$

Consider any optimization algorithm (stochastic or not). Eq. 8 implies the finiteness of z, and therefore by Eq. 7 the finiteness of $X_{n,\Omega}$, bounded above by a constant C independent of n, with probability at least $1 - \delta/2$.

We will here use sets of points with lower bounded distance to each other; such sets are classical in statistics[26], and are now also used for building lower bounds based on information theory[27, 28].

Consider R a set of cardinality C' such that

$$\frac{C'}{C} > \frac{1 - \frac{\delta}{2}}{1 - \delta} \tag{9}$$

and such that two distinct elements of R are at distance greater than 2ϵ, with $\epsilon = C(d)/n^\alpha$, from each other; such a set certainly exists if n is large enough. A nice property of this set is that if the optimum $\boldsymbol{x}^*$ is uniformly drawn in R, then it can only be found with probability $1 - \delta/2$ and with precision $C(d)/n^\alpha$ if $X_{n,\Omega}$ contains one point close to r for a proportion at least $1 - \delta/2$ of points $r \in R$.

Consider $f_{\boldsymbol{x}^*} = f_{\boldsymbol{x}^*,\beta,\gamma}$ with $\boldsymbol{x}^*$ uniformly distributed in R. Then:

$$\begin{aligned}
& P(\|\boldsymbol{x}_n - \boldsymbol{x}^*\| \leq \epsilon) \\
\leq \; & \mathbb{E}_\Omega P_{\boldsymbol{x}^*}(\boldsymbol{x}^* \in \text{Enl}(X_{n,\Omega}, \epsilon)) \\
\leq \; & P(\#X_{n,\Omega} \leq C) P_{\boldsymbol{x}^*}(\boldsymbol{x}^* \in \text{Enl}(X_{n,\Omega}, \epsilon) | \#X_{n,\Omega} \leq C) \\
& + P(\#X_{n,\Omega} > C) \\
\leq \; & (1 - \frac{\delta}{2}) \frac{C}{C'} + \frac{\delta}{2} \\
< \; & 1 - \frac{\delta}{2}
\end{aligned}$$

where $\text{Enl}(U, \epsilon)$ is the ϵ-enlargement of U defined as:

$$\text{Enl}(U, \epsilon) = \left\{ \boldsymbol{x}; \exists \boldsymbol{x}' \in U, \|\boldsymbol{x} - \boldsymbol{x}'\| \leq \epsilon \right\}.$$

This contradicts Eq. 4.

This concludes the proof of $\alpha\beta \leq 1$. □

4. CONCLUSION

Our results are based on the *locality assumption* (Eq. 4). They show tight results in some cases. The locality assumption is somewhat natural, as most evolution strategies (not all, but almost all) verify this assumption; for example, [29], one of the main evolutionary optimization convergence results, shows a linear convergence, with no sampling far away; [30], showing faster rates with surrogate models, also verifies this assumption; and polynomial rates in noisy optimization as in [14, 31] have the same property as well as many experimentally known rates [32]. Nonetheless, other assumptions leading to similar results (e.g. assumptions ensuring that points far from the optimum cannot help too much) are worth being investigated for clarifying the overall picture.

Basically, our results are about optimization tricks as follows: an optimization algorithm uses the *far sampling trick* if there is a clear distinction between the approximation of the optimum that they propose and the search points that they use for exploring the fitness function.

Our results can be seen either:

- As proofs that fast rates (faster than those in the tables below) are possible only if you assume that points far away from the optimum do bring information, by statistical model estimation, which are relevant for improving the rate (so that the "far sampling trick" can work).
- As proofs that algorithms which, like most evolutionary algorithms (but not all), do not sample far away from the optimum, can not be optimal when the function is "flat" enough around the optimum (there are algorithms and families of functions for which better rates are possible - which does not mean that algorithms which do not want to use the far sampling trick are necessarily bad algorithms).

The hot discussions on the BBOB mailing list, around the fact that the test beds should distinguish the search points used for approximating the optimum and the search points used for gathering information by sampling far away, suggest that this paper comes at the right moment for noisy optimization formal analysis.

Table 1 summarizes the state of the art; new result from this paper are in bold, and we emphasize cases in which a gap is known. We see that our results show the tightness in the case $\gamma = 1$ (small noise; the variance goes to zero around the optimum), and reduce the gap in the case $\gamma < 1$ (large noise). Our results also show that for fast rates, sampling far from the optimum is necessary; e.g. if $\beta = 4$, $\alpha = 1/2$ is possible only with sampling far from the optimum. Though, this is only possible if there are long range dependencies on the fitness function, so that such points far from the optimum can be used.

Further work

The locality assumption might be or might not be, depending on the application, a good idea. From many discussions around that, we believe that there is room for works like this one, in which a locality assumption prevents the use of information far from the optimum, reliable only when strong assumptions on the model are available. It is also a model which shows that some rates imply sampling far from the

"flatness" β	Proved rate for R-EDA in [16] ("flatness" on an envelope of the fitness function; the fitness function does not have to be flat around $\boldsymbol{x}^*$)	Lower bound in [16]	R-EDA experimental rate in [14] (on functions with invariances)	This paper (lower bound under locality assumption)	Rate for learnable cases
Framework $\gamma = 1$ (small noise)					
1	$\alpha \geq 1$	$\alpha \leq 1$	$\alpha = 1$	$\alpha \leq 1$	$\alpha = 1/2$
2	$\alpha \geq 1/2$	$\alpha \leq 1$	$\alpha = 1/2$	$\alpha \leq \mathbf{1/2}$	$\alpha = 1/2$
4	$\alpha \geq 1/4$	$\alpha \leq 1$	$\alpha = 1/4$	$\alpha \leq \mathbf{1/4}$	$\alpha = 1/2$
Framework $\gamma < 1$ (large noise)					
1	$\alpha \geq 1/2$	$\alpha \leq 1$	$\alpha = 1/2$	$\alpha \leq 1$	$\alpha = 1/2$
2	$\alpha \geq 1/4$	$\alpha \leq 1$	$\alpha = 1/4$	$\alpha \leq \mathbf{1/2}$	$\alpha = 1/2$
4	$\alpha \geq 1/8$	$\alpha \leq 1$	$\alpha = 1/8$	$\alpha \leq \mathbf{1/4}$	$\alpha = 1/2$

Table 1: This table summarizes the state of the art in terms of α for which $||\tilde{\boldsymbol{x}}_n - \boldsymbol{x}^*|| = \tilde{O}(1/n^\alpha)$ is possible. Rates for E-EDA include functions which are not differentiable, and are just upper bounded by flat functions around $\boldsymbol{x}^*$ with β coefficient; see [14] for more details. Experimental rates for R-EDA are for functions with strong invariances/symmetries; see [14] for details. The last column presents results with $\alpha = 1/2$, corresponding to cases in which using statistical model estimation is possible: the limit case of infinite differentiability in [33, 34, 22], is also reached by quadratic logistic regression under parametric assumptions on the objective function[14]; assumptions are not directly comparable to those of the other columns. Lower bounds on the complexity (upper bounds on α) from this paper are under the additional assumption of local sampling.

optimum. Fast optimization algorithms might, as CLOP, be a compromise between sampling close to the optimum and sampling on areas of maximum uncertainty. Further investigations on intermediate models might be a good idea.

There is still a gap between the upper and the lower bound, for algorithms having the locality assumption, in the case $\gamma < 1$ (large noise), which is an immediate further work.

We consider noisy optimization in the case of local convergence; clearly, the global convergence case can also be interesting[35].

We did not compute exactly constants C and C'. Maybe it is possible to obtain more information on the constant in the convergence using detailed computations of C and C'.

Acknowledgments

We are grateful to the Dagstuhl seminar on Evolutionary Algorithms, to the Montefiore institute in University of Belgium in which author #2 had interesting discussions around noisy optimization. We are grateful to various members of the Tao team for interesting discussions, as well as discussions in the BBOB mailing list. We are grateful to European project MASH, FP7 program.

5. REFERENCES

[1] A. Conn, K. Scheinberg, and L. Toint, "Recent progress in unconstrained nonlinear optimization without derivatives," 1997. [Online]. Available: citeseer.ist.psu.edu/conn97recent.html

[2] B. Doerr and C. Winzen, "Towards a complexity theory of randomized search heuristics: Ranking-based black-box complexity," in *CSR*, ser. Lecture Notes in Computer Science, A. S. Kulikov and N. K. Vereshchagin, Eds., vol. 6651. Springer, 2011, pp. 15–28.

[3] S. Grünewälder, J.-Y. Audibert, M. Opper, and J. Shawe-Taylor, "Regret Bounds for Gaussian Process Bandit Problems," in *JMLR Workshop and Conference Proceedings : AISTATS 2010*, vol. 9, Chia Laguna Resort, Sardinia, Italie, 2010, pp. 273–280. [Online]. Available: http://hal-enpc.archives-ouvertes.fr/hal-00654517

[4] H.-P. Schwefel, *Numerical Optimization of Computer Models*. New-York: John Wiley & Sons, 1981, 1995 – 2^{nd} edition.

[5] H.-G. Beyer, "Mutate large, but inherit small ! On the analysis of mutations in $(1, \lambda)$-ES with noisy fitness data," in *Proc. of the 5^{th} Conference on Parallel Problems Solving from Nature*, T. Bäck, G. Eiben, M. Schoenauer, and H.-P. Schwefel, Eds. Springer Verlag, 1998, pp. 109–118.

[6] J. Fitzpatrick and J. Grefenstette, "Genetic algorithms in noisy environments, in machine learning: Special issue on genetic algorithms, p. langley, ed. dordrecht: Kluwer academic publishers, vol. 3, pp. 101 120," 1988.

[7] D. V. Arnold and H.-G. Beyer, "Local performance of the (1+1)-ES in a noisy environment," *IEEE Transactions on Evolutionary Computation*, vol. 6, no. 1, pp. 30–41, 2002.

[8] D. V. Arnold and H. georg Beyer, "Evolution strategies with cumulative step length adaptation on the noisy parabolic ridge," Tech. Rep., 2006.

[9] U. Hammel and T. Bäck, "Evolution strategies on noisy functions: How to improve convergence properties," in *Parallel Problem Solving From Nature*, ser. LNCS, Y. Davidor, H.-P. Schwefel, and R. Männer, Eds., vol. 866. Jerusalem: springer, 9–14Oct. 1994, pp. 159–168.

[10] J. M. Fitzpatrick and J. J. Grefenstette, "Genetic algorithms in noisy environments," *Machine Learning*, vol. 3, pp. 101–120, 1988.

[11] M. Jebalia and A. Auger, "On multiplicative noise models for stochastic search," in *Parallel Problem Solving From Nature*, dortmund Allemagne, 2008.

[Online]. Available: http://hal.inria.fr/inria-00287725/en/

[12] O. Teytaud and A. Auger, "On the adaptation of the noise level for stochastic optimization," in *IEEE Congress on Evolutionary Computation*, Singapour, 2007. [Online]. Available: http://hal.inria.fr/inria-00173224/en/

[13] R. Coulom, "Clop: Confident local optimization for noisy black-box parameter tuning," in *ACG*, ser. Lecture Notes in Computer Science, H. J. van den Herik and A. Plaat, Eds., vol. 7168. Springer, 2011, pp. 146–157.

[14] R. Coulom, P. Rolet, N. Sokolovska, and O. Teytaud, "Handling expensive optimization with large noise," in *FOGA*, H.-G. Beyer and W. B. Langdon, Eds. ACM, 2011, pp. 61–68.

[15] P. Rolet and O. Teytaud, "Bandit-based estimation of distribution algorithms for noisy optimization: Rigorous runtime analysis," in *Proceedings of Lion4 (accepted); presented in TRSH 2009 in Birmingham*, 2009.

[16] ——, "Adaptive Noisy Optimization," in *EvoStar 2010*, Istambul, Turquie, Feb. 2010. [Online]. Available: http://hal.inria.fr/inria-00459017

[17] V. Heidrich-Meisner and C. Igel, "Uncertainty handling cma-es for reinforcement learning," in *GECCO*, F. Rothlauf, Ed. ACM, 2009, pp. 1211–1218.

[18] L. Devroye, L. Györfi, and G. Lugosi, *A probabilistic Theory of Pattern Recognition*. Springer, 1997.

[19] R. Coulom, P. Rolet, N. Sokolovska, and O. Teytaud, "Handling expensive optimization with large noise," in *FOGA*, H.-G. Beyer and W. B. Langdon, Eds. ACM, 2011, pp. 61–68.

[20] D. R. Jones, M. Schonlau, and W. J. Welch, "Efficient global optimization of expensive black-box functions," *J. of Global Optimization*, vol. 13, no. 4, pp. 455–492, 1998.

[21] J. Villemonteix, E. Vazquez, and E. Walter, "An informational approach to the global optimization of expensive-to-evaluate functions," *Journal of Global Optimization*, p. 26 pages, 09 2008. [Online]. Available: dx.doi.org/10.1007/s10898-008-9354-2 http://hal-supelec.archives-ouvertes.fr/hal-00354262/en/

[22] V. Fabian, "Stochastic approximation of minima with improved asymptotic speed," *Ann. Math. Statist.*, vol. 38, no. 1, pp. 191–200, 1967.

[23] L. Bienaymé, "Considérations à l'appui de la découverte de laplace," *Comptes Rendus de l'Académie des Sciences*, vol. 37, pp. 309–324, 1853.

[24] P. Chebyshev, "Sur les valeurs limites des integrales," *Math Pure Appl*, vol. 19, p. 157£160, 1874.

[25] A. Markov, "On certain applications of algebraic continued fractions," Ph.D. dissertation, St Petersburg, 2002.

[26] A. V. D. Vaart and J. Wellner, *Weak Convergence and Empirical Processes*. Springer series in statistics, 1996.

[27] O. Teytaud and S. Gelly, "General lower bounds for evolutionary algorithms," in 10^{th} *International Conference on Parallel Problem Solving from Nature (PPSN 2006)*, 2006.

[28] H. Fournier and O. Teytaud, "Lower bounds for comparison based evolution strategies using vc-dimension and sign patterns," *Algorithmica*, vol. 59, no. 3, pp. 387–408, 2011.

[29] A. Auger, "Convergence results for (1,λ)-SA-ES using the theory of φ-irreducible Markov chains," *Theoretical Computer Science*, vol. 334, no. 1-3, pp. 35–69, 2005.

[30] A. Auger, M. Schoenauer, and O. Teytaud, "Local and global order 3/2 convergence of a surrogate evolutionnary algorithm," in *Gecco*, 2005, p. 8 p.

[31] D. V. Arnold and H.-G. Beyer, "Efficiency and mutation strength adaptation of the (mu/mui,lambda)-es in a noisy environment," in *Parallel Problem Solving from Nature*, ser. LNCS, M. S. et al., Ed., vol. 1917. springer, 2000, pp. 39–48.

[32] H.-G. Beyer, *The Theory of Evolution Strategies*, ser. Natural Computing Series. Springer, Heideberg, 2001.

[33] H. Chen, "Lower rate of convergence for locating a maximum of a function," *Ann. Statist.*, vol. 16, no. 3, pp. 1330–1334, 1988.

[34] J. Kiefer and J. Wolfowitz, "Stochastic estimation of the maximum of a regression function," *Annals of Mathematical Statistics*, vol. 23, no. 3, p. 462£466, 1952.

[35] E. Vazquez, J. Villemonteix, M. Sidorkiewicz, and E. Walter, "Global optimization based on noisy evaluations: an empirical study of two statistical approaches," *Journal of Global Optimization*, p. 17 pages, 2008. [Online]. Available: dx.doi.org/10.1007/s10898-008-9313-y http://hal-supelec.archives-ouvertes.fr/hal-00354656/en/

[36] H.-G. Beyer and W. B. Langdon, Eds., *Foundations of Genetic Algorithms, 11th International Workshop, FOGA 2011, Schwarzenberg, Austria, January 5-8, 2011, Proceedings*. ACM, 2011.