Linear Convergence of Evolution Strategies with Derandomized Sampling Beyond Quasi-Convex Functions

Jérémie Decock, Olivier Teytaud

Linear Convergence of Evolution Strategies with Derandomized Sampling Beyond Quasi-Convex Functions

Jérémie Decock and Olivier Teytaud

TAO-INRIA, LRI, CNRS UMR 8623,
Université Paris-Sud, Orsay, France
{olivier.teytaud,jeremie.decock}@inria.fr

Abstract. We study the linear convergence of a simple pattern search method on non quasi-convex functions on continuous domains. Assumptions include an assumption on the sampling performed by the evolutionary algorithm (supposed to cover efficiently the neighborhood of the current search point), the conditioning of the objective function (so that the probability of improvement is not too low at each time step, given a correct step size), and the unicity of the optimum.

1 Introduction

Continuous evolutionary algorithms are well known for robust convergence. However, most proven results are for simple objective functions, e.g. sphere functions [1]. Results also include compositions with monotone functions (so that not only convex functions are covered), but the considered objective functions are nonetheless still almost always quasi-convex (i.e. sublevel sets are convex), as well as most derivative free optimization algorithms [4], whereas nearly all testbeds are based on more difficult functions [7, 11]. Extensions to non quasi-convex functions are still rare [12] and limited to convergence (i.e.: asymptotically we will find the optimum). We here extend such results to linear convergence (i.e. the precision after n iterations is $O(\exp(-\Omega(n)))$.There are works devoted to unimodal objective functions, without convexity assumptions [6], but such works are in the discrete domain and do not say anything for the linear convergence on continuous domains. All in all, only one of the six objective functions of Fig. 1 is covered by existing results, in terms of linear convergence.

In this paper, we prove linear convergence of a simple pattern search method with derandomized sampling on non quasi-convex families of functions. Section 2 presents the framework, and the assumptions under which our results hold. Section 3 is the mathematical analysis, under this set of assumptions. Section 4 presents the application to positive definite quadratic forms: it shows that the family of quadratic forms with conditioning bounded by some constant verifies our set of assumptions, and therefore that our evolution strategy with derandomized sampling has linear convergence rate on such objective functions. Incidentally, this section emphasizes the critical underlying assumptions for proving

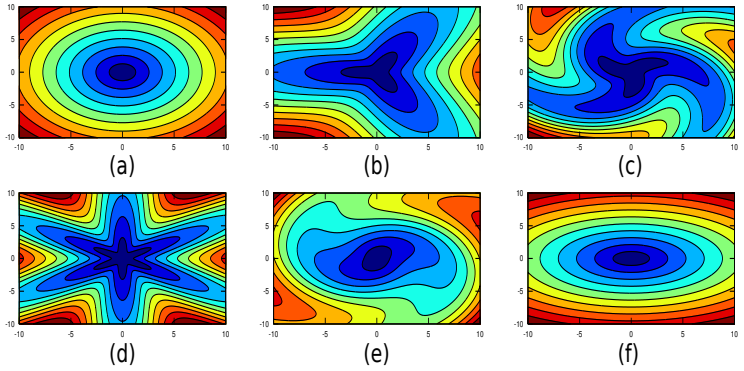

Fig. 1. Six graphical representations of easy objective functions; only the first one (sphere) is covered by existing linear convergence results. Even the sixth one (ellipsoids) is not included in published linear convergence results. We extend to all functions verifying Eqs. 1-6 (see these equations in text), including all functions presented here. We present assumptions under which our results hold in Section 2, the main result in Section 3, and we will show in details that the sixth case above (quadratic functions) is covered by the result in Section 4 (but case 1 is a special case of case 6, and cases 2, 3, 4, 5 can be tackled similarly). [12] provides other examples, with different but very related assumptions; their examples are also covered by our theorem.

the result, suggesting extensions to other families of fitness functions. Section 5 concludes and discusses limitations and further work.

2 A Simple Pattern Search Method

We consider an evolutionary algorithm as in Alg. 1. As the sampling is derandomized, we might indeed call this algorithm a pattern search method. We assume the followings.

The Objective Function

We assume that the function f has a unique minimum. Without loss of generality, we assume that the objective function verifies $f(0) = 0$ and that this is the minimum. The considered algorithms are invariant by transition or composition with monotone functions, so this does not reduce the generality of the analysis.

Conditioning

We assume that

$$K'||\boldsymbol{x}|| \leq f(\boldsymbol{x}) \leq K''||\boldsymbol{x}|| \tag{1}$$

for all $\boldsymbol{x}$ in $\mathbb{R}^d$ and for some constants $K' > 0$ and $K'' > 0$. We point out that, as we consider algorithms which are invariant under transformations of the

Algorithm 1 The Simple Evolution Strategy. In case there is no unicity for choosing $\boldsymbol{x}'$, any breaking tie solution is ok. (c) refers to the counting operation, which will be important in the proof. $[[1, k]]$ stands for the integer set $\{1, \ldots, k\}$.

Initialize $\boldsymbol{x} \in \mathbb{R}^d$
Parameters $k \in \mathbb{N}^*, \boldsymbol{\delta}_1, \ldots, \boldsymbol{\delta}_k \in \mathbb{R}^d, \sigma \in \mathbb{R}_+^*, k_1 \in \mathbb{N}^*, k_2 \in \mathbb{N}^*$
for $t = 1, 2, 3, \ldots$ **do**

// **just for archiving**
$\boldsymbol{X}_t \leftarrow \boldsymbol{x}$

// **mutations**
For $i \in [[1, k]]$, $\boldsymbol{x}_i \leftarrow \boldsymbol{x} + \sigma \boldsymbol{\delta}_i$

// **useful auxiliary variables**
$n \leftarrow$ number of $\boldsymbol{x}_i$ such that $f(\boldsymbol{x}_i) < f(\boldsymbol{x})$ (c)
$\boldsymbol{x}' \leftarrow \boldsymbol{x}_i$ with $i \in [[1, k]]$ such that $f(\boldsymbol{x}_i)$ is minimum

// **step-size adaptation**
if $n \leq k_1$ **then**
$\sigma \leftarrow \sigma / 2$
end if
if $n \geq k_2$ **then**
$\sigma \leftarrow 2\sigma$
end if

// **win: accepted mutation**
if $k_1 < n < k_2$ **then**
$\boldsymbol{x} \leftarrow \boldsymbol{x}'$
end if
end for

objective function by composition with monotonic functions, this assumption is not so strong as a constraint and quadratic positive definite forms with bounded condition number are in fact covered (their square root verifies Eq. 1).

Good Sampling

Here we use the derandomized sampling assumptions (Eqs. 2-6), which are crucial in our work. This sampling is deterministic, as in pattern search methods [4]. We assume that for some $0 < b < b' \leq 2b' \leq c' \leq c, 0 < \eta < 1$ and $\forall \boldsymbol{x} \in \mathbb{R}^d$,

$$\begin{aligned} \sigma \text{ too large: } & \sigma \geq b^{-1}||\boldsymbol{x}|| \\ & \Rightarrow \#\{i \in [[1,k]]; f(\boldsymbol{x} + \sigma\boldsymbol{\delta}_i) < f(\boldsymbol{x})\} \leq k_1 \end{aligned} \tag{2}$$

$$\begin{aligned} \sigma \text{ small enough: } & \sigma \leq b'^{-1}||\boldsymbol{x}|| \\ & \Rightarrow \#\{i \in [[1,k]]; f(\boldsymbol{x} + \sigma\boldsymbol{\delta}_i) < f(\boldsymbol{x})\} > k_1 \end{aligned} \tag{3}$$

$$\begin{aligned} \sigma \text{ large enough: } & \sigma \geq c'^{-1}||\boldsymbol{x}|| \\ & \Rightarrow \#\{i \in [[1,k]]; f(\boldsymbol{x} + \sigma\boldsymbol{\delta}_i) < f(\boldsymbol{x})\} < k_2 \end{aligned} \tag{4}$$

$$\begin{aligned} \sigma \text{ too small: } & \sigma \leq c^{-1}||\boldsymbol{x}|| \\ & \Rightarrow \#\{i \in [[1,k]]; f(\boldsymbol{x} + \sigma\boldsymbol{\delta}_i) < f(\boldsymbol{x})\} \geq k_2 \end{aligned} \tag{5}$$

$$\begin{aligned} \text{Perfect } \sigma\text{: } & b'^{-1}||\boldsymbol{x}|| \leq \sigma \leq c'^{-1}||\boldsymbol{x}|| \\ & \Rightarrow \exists i \in [[1,k]]; f(\boldsymbol{x} + \sigma\boldsymbol{\delta}_i) \leq \eta f(\boldsymbol{x}) \end{aligned} \tag{6}$$

Discussion on Assumptions

Assumptions 2, 3, 4, 5, 6 basically assume that the sampling is regular enough for the shape of the level sets. For example, the finite VC-dimension of ellipsoids ensure that, when the conditioning is bounded, quadratic functions verify the assumptions above (and therefore the theorem below) with arbitrarily high probability if $\boldsymbol{\delta}_1, \ldots, \boldsymbol{\delta}_k$ are randomly drawn and if k is large enough. Importantly, the critical assumption in the derandomization is that all iterations have the same $\boldsymbol{\delta}_1, \ldots, \boldsymbol{\delta}_k$. This will be developed in Section 4.

Assumptions 6 and 1 use the fitness values; but they just have to hold for one of the fitness values obtained by replacing f with $g \circ f$ with g a monotone function.

3 Mathematical Analysis

Main Theorem: *Assume Eqs. 1-6. There exists a constant K, depending on $\eta, K', K'', \max_i ||\boldsymbol{\delta}_i||$ only such that for index t sufficiently large*

$$\ln(||\boldsymbol{X}_t||)/t \leq K < 0 \tag{7}$$

(with $\ln(0) = -\infty$*) where the sequence of* $\boldsymbol{X}_t$ *is defined as in Alg. 1.*

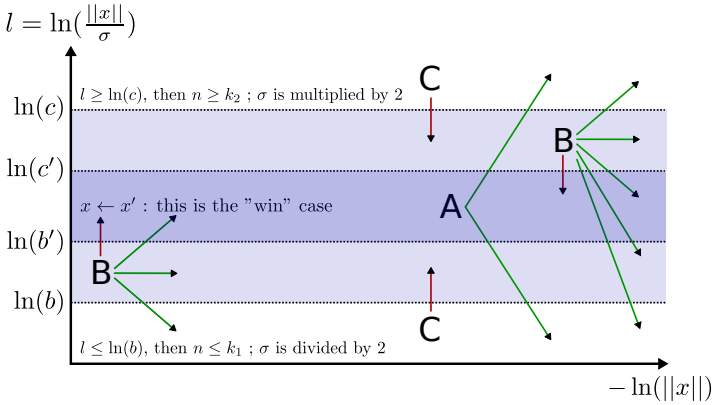

Fig. 2. The linear convergence proof in a nutshell. X-axis: $-\ln(||\boldsymbol{x}||)$. Y-axis: $l = \ln(\frac{||\boldsymbol{x}||}{\sigma})$. At each iteration, either **case A holds:** then the iteration is for sure an improvement by a factor at least η, or **case B holds:** the iteration can be an improvement or not; if not, the point is moved towards case A by $\ln(2)$ upwards or downwards, or **case C holds:** then $\boldsymbol{x}$ is moved upwards (if it is at the bottom) or downwards (if it is at the top). This ensures that after finitely many time steps we go back to case A unless there is a "win" by case B in the mean time. The crucial point for the proof is that each "win" is an improvement by least a controlled factor η, so that the slope of "win" arrows is bounded, so that there is linear convergence and not only an infinite sequence of "small" improvements.

Proof: First, we briefly explain and illustrate the proof, before the formal proof below. The proof is sketched in Fig. 2. At each iteration t, we are at some point in the figure; the x-axis is $-\ln(||\boldsymbol{x}||)$ (equivalent to $-\ln(f(\boldsymbol{x}))$, by Eq. 1), the y-axis is $l = \ln\left(\frac{||\boldsymbol{x}||}{\sigma}\right)$. The step-size adaptation ensures that if we are at the bottom ($l \leq \ln(b)$), we go upwards; if we are at the top ($l \geq \ln(c)$), we go downwards. Between $l = \ln(b)$ and $l = \ln(c)$, everything can happen; but if there's no "win" case in the mean time, we will arrive between $l = \ln(b')$ and $l = \ln(c')$, where a win is ensured. As steps are fast, this can not take too much time (if there is no "win", l increases by $\ln(2)$ or decreases by $\ln(2)$ in direction of the "forced win" range $[\ln(b'), \ln(c')]$). This will be formalized below. $c' \geq 2b'$ ensures that the algorithm can not jump from $l < \ln(b')$ to $l > \ln(c')$ or from $l > \ln(c')$ to $l < \ln(b')$. Therefore there is necessarily a "win" in the mean time. Eq. 6 ensures that wins provide a significant improvement.

We now write the proof formally. Consider an iteration of the algorithm, with n the number of mutations i with $f(\boldsymbol{x} + \sigma\boldsymbol{\delta}_i) < f(\boldsymbol{x})$ (as defined in Alg. 1, Eq. (c)).

Define $l = \ln\left(\frac{||\boldsymbol{x}||}{\sigma}\right)$. Eqs. 2-6 can be rephrased as follows:

$$l \leq \ln(b) \Rightarrow \#\{i \in [[1,k]]; f(\boldsymbol{x} + \sigma\boldsymbol{\delta}_i) < f(\boldsymbol{x})\} \leq k_1 \quad (8)$$

$$l \geq \ln(b') \Rightarrow \#\{i \in [[1,k]]; f(\boldsymbol{x} + \sigma\boldsymbol{\delta}_i) < f(\boldsymbol{x})\} > k_1 \quad (9)$$

$$l \leq \ln(c') \Rightarrow \#\{i \in [[1,k]]; f(\boldsymbol{x} + \sigma\boldsymbol{\delta}_i) < f(\boldsymbol{x})\} < k_2 \quad (10)$$

$$l \geq \ln(c) \Rightarrow \#\{i \in [[1,k]]; f(\boldsymbol{x} + \sigma\boldsymbol{\delta}_i) < f(\boldsymbol{x})\} \geq k_2 \quad (11)$$

$$\ln(b') \leq l \leq \ln(c') \Rightarrow \exists i \in [[1,k]]; f(\boldsymbol{x} + \sigma\boldsymbol{\delta}_i) \leq \eta f(\boldsymbol{x}) \quad (12)$$

Define $\boldsymbol{x}'$ as in Alg. 1. We get the following behavior:

- Forced increase: if $l \leq \ln(b)$, then $n \leq k_1$; σ is divided by 2, and l is increased by $\ln(2)$ (Eq. 8). This is a case C in Fig. 2.
- Forced decrease: if $l \geq \ln(c)$, then $n \geq k_2$; σ is multiplied by 2, and l is decreased by $\ln(2)$ (Eq. 11). This is a case C in Fig. 2.
- Forced win: if $\ln(b') \leq l \leq \ln(c')$, then $\boldsymbol{x} \leftarrow \boldsymbol{x}'$; this is the "sure win" case (Eq. 12); l can be increased (at most by $\max_i ||\boldsymbol{\delta}_i||$) or decreased (by $\Delta = \ln(||\boldsymbol{x}||/||\boldsymbol{x}'||)$). This is a case A in Fig. 2.

Importantly, these 3 cases do not cover all possible cases; $\ln(c') < l < \ln(c)$ and $\ln(b) < l < \ln(b')$ are not covered in items above. These two remaining cases are termed case B in Fig. 2.

Step 1: Showing that there are infinitely many wins.

The two first lines above (case $l \leq \ln(b)$ and case $l \geq \ln(c)$) ensure that if l is too low or too high, it eventually comes back to the range $[\ln(b'), \ln(c')]$ (where a win necessarily occurs), unless there is a win in the mean time (in the range $[\ln(b), \ln(c)]$ where wins are not sure but are possible). Importantly, l can increase or decrease by $\ln(2)$ at most; so the algorithm can not jump from less than $\ln(b')$ to more than $\ln(c')$. This ensures that infinitely often we have a win $\boldsymbol{x} \leftarrow \boldsymbol{x}'$. But we want linear convergence. Therefore we must consider how many steps there are before we come back to a "win", and how large are improvements in case of "win".

Step 2: showing that "wins" are big enough.

In all cases of "win", i.e. $k_1 < n < k_2$, with $\Delta = \ln(||\boldsymbol{x}||/||\boldsymbol{x}'||)$, we know that $f(\boldsymbol{x}') \leq \eta f(\boldsymbol{x})$ and $f(\boldsymbol{x}') \leq K''||\boldsymbol{x}'|| \leq \frac{K''}{K'}\frac{||\boldsymbol{x}'||}{||\boldsymbol{x}||} f(\boldsymbol{x})$ so that $\ln(f(\boldsymbol{x}))$ is decreased by at least

$$\max(\ln(1/\eta), \ln(K'/K'') + \Delta). \quad (13)$$

After a "win", with $l' = \ln\left(\frac{||\boldsymbol{x}'||}{\sigma}\right)$,

- if $l' \leq \ln(b')$, then the number of iterations before the next win is at most $z = 1 + \ln(\frac{c}{b})\Delta/\ln(2)$, because $l' \geq \ln(b) - \Delta \geq \ln(b') - \ln(b'/b) - \Delta \geq \ln(b') - \ln(c/b) - \Delta$ and forced increase are by steps of at least $\ln(2)$.

- if $l' \geq \ln(c')$, then the number of iterations before the next win is at most $z = 1 + \ln(\frac{c}{b}) \max_i \ln(||\boldsymbol{\delta}_i||)/\ln(2)$, because $l' \leq \ln(c) - max_i \ln(\boldsymbol{\delta}_i) \leq \ln(c') - \ln(c'/c) - \max_i \ln(\boldsymbol{\delta}_i) \leq \ln(c') - \ln(c/b) - \max_i \ln(\boldsymbol{\delta}_i)$ and forced decreases are by steps of at least $\ln(2)$.
- less than in both cases above, otherwise.

In both cases, Eq. 13 divided by z is lower bounded by some positive constant

$$\begin{aligned} & ProgressRate \\ &= \text{Eq. 13 divided by } z \\ &= \frac{\max(\ln(1/\eta), \ln(K'/K'') + \Delta_i)}{\min(1 + \ln(c/b)\Delta_i/\ln(2), 1 + \ln(c/b)\max_j \ln(||\boldsymbol{\delta}_j||)/\ln(2))}. \end{aligned} \tag{14}$$

Step 3: summing iterations.

Eq. 14 is the progress rate between two wins, after normalization by the number of steps between these two wins. Hence if $t > n_0$,

$$\begin{aligned} \ln(f(\boldsymbol{X}_t)) \leq \ln(f(\boldsymbol{X}_1)) - (t - n_0) \times \\ \sum_i \frac{\max(\ln(1/\eta), \ln(K'/K'') + \Delta_i)}{\min(1 + \Delta_i/\ln(2), 1 + \max_j \ln(||\boldsymbol{\delta}_j||)/\ln(2))} \end{aligned} \tag{15}$$

where the summation is for i index of an iteration t with a "win", and n_0 is the number of initial iterations before a "win" (i.e. n_0 depends on the initial conditions but it is finite).

Eq. 15 yields the expected result. □

This result would be void if there was no algorithm and no space of functions for which assumptions 1-6 hold. Therefore, next Section is devoted to showing that for the important case of families of quadratic functions with bounded conditioning, assumptions 1-6 hold, and therefore the theorem above holds.

4 Application to Quadratic Functions

This section shows an example of application of the theorem above. The main strength of our results is that it covers many families of functions; yet, Eqs. 1-6 are not so readable. We show in this section that a simple family of fitness functions verify all the assumptions.

We consider the application to positive definite quadratic forms with bounded conditioning, i.e. we consider $f \in F$ with F the set of quadratic positive definite objective functions f such that

$$\frac{maxEigenValue(Hessian(f))}{minEigenValue(Hessian(f))} < c_{max} < \infty. \tag{16}$$

Notably, thanks to the use of VC-dimension, the approach is indeed quite generic and can be applied to all families of functions obtained by rotation/translation from fitness functions in Fig. 1.

Instead of working on Q directly, with $\boldsymbol{x} \mapsto Q(\boldsymbol{x} - \boldsymbol{x}^*)$ a quadratic form with Q positive definite with optimum in 0, we work on $\boldsymbol{x} \mapsto \sqrt{Q(\boldsymbol{x} - \boldsymbol{x}^*)}$, so that Eq. 1 is verified; as considered algorithms are invariants by composition with monotone functions, this does not change the result.

We assume that $\boldsymbol{\delta}_1, \boldsymbol{\delta}_2, \ldots, \boldsymbol{\delta}_k$ are independently uniformly randomly drawn in the unit ball $B(0,1)$. From now on, we note $p = p_{\boldsymbol{x},\sigma,f}$ the probability that $f(\boldsymbol{x} + \sigma\boldsymbol{\delta}_i)$ is in $E = f^{-1}([0, f(\boldsymbol{x})[)$, and $\hat{p} = \hat{p}_{\boldsymbol{x},\sigma,f}$ the frequency $\frac{1}{k}\sum_{i=1}^{k} \mathbf{1}_{\boldsymbol{x}+\sigma\boldsymbol{\delta}_i \in E}$. We will often drop the indices for short.

The assumptions in Section 2 essentially mean that frequencies are close to expectations for $\boldsymbol{x} + \sigma\boldsymbol{\delta}_i \in f^{-1}([0, f(\boldsymbol{x})[)$ and $\boldsymbol{x} + \sigma\boldsymbol{\delta}_i \in f^{-1}([0, \eta f(\boldsymbol{x})[)$, independently of $\boldsymbol{x}$, σ, f. This is typically a case in which VC-dimension [14] can help.

The purpose of this section is to show Eqs. 1-6, for a given family of functions, namely the family F defined above; by proving Eqs. 1-6, we show the following.

Corollary: *Assume that the $\boldsymbol{\delta}_i$ are uniformly randomly drawn in the unit ball $B(0,1)$. Assume that F is the set of quadratic functions with minimum in 0 ($f(0) = 0$) which verify Eq. 16 for some $c_{max} < \infty$. Then, almost surely on the sequence $\boldsymbol{\delta}_1, \boldsymbol{\delta}_2, \ldots, \boldsymbol{\delta}_k$, for k large enough and some parameters k_1 and k_2 of Alg. 1, then Eqs. 1-6 hold, and therefore for some $K < 0$, for all $t > 0$,*

$$\ln(||\boldsymbol{X}_t||)/t \leq K \tag{17}$$

with $\ln(0) = -\infty$ and where the sequence of $\boldsymbol{X}_t$ is defined as in Alg. 1.

Proof: We use the main theorem above for proving Eq. 17, so we just have to prove that Eqs. 1-6 hold.

Step 1: using VC-dimension for approximating expectations by frequencies. Thanks to the finiteness of the VC-dimension of quadratic forms (see e.g. [5]), we know that for all $\epsilon > 0$, almost surely in $\boldsymbol{\delta}_1, \boldsymbol{\delta}_2, \ldots, \boldsymbol{\delta}_k$, for all $\delta > 0$ and k sufficiently large, with probability at least $1 - \delta$,

$$\sup_{\boldsymbol{x},f,\sigma>0} |\hat{p}_{\boldsymbol{x},\sigma,f} - p_{\boldsymbol{x},\sigma,f}| \leq \epsilon/2 \tag{18}$$

where $\boldsymbol{x}$ ranges over the domain, f ranges over F.

For short, we will often drop the indices, so that Eq. 18 becomes Eq. 19:

$$\sup_{\boldsymbol{x},f,\sigma>0} |\hat{p} - p| \leq \epsilon/2 \tag{19}$$

The important point here is that this result is a uniform result (uniform on $f \in F$); this is not just a simple law of large numbers, it is a uniform law of large numbers, so that it is not a mistake if there is a supremum on $\boldsymbol{x}, \sigma, f$. Almost surely, the supremum is bounded; it is not only bounded almost surely for each $\boldsymbol{x}, \sigma, f$ separately, and this is the key concept in this proof.

Step 2: showing that $\boldsymbol{\sigma}$ small leads to high acceptance rate and $\boldsymbol{\sigma}$ high leads to small acceptance rate. Thanks to the bounded conditioning (Eq. 16), there exists $\epsilon > 0$ s.t.

$$s < \frac{1}{2} s' \tag{20}$$

$$\text{with } s = \sup \left\{ \frac{\sigma}{||\boldsymbol{x}||}; \sigma, \boldsymbol{x}, f \text{ s.t. } p \geq \frac{\epsilon}{2} \right\}$$

$$\text{and } s' = \inf \left\{ \frac{\sigma}{||\boldsymbol{x}||}; \sigma, \boldsymbol{x}, f \text{ s.t. } p < \frac{1}{2} - \frac{\epsilon}{2} \right\}$$

because $s \to 0$ as $\epsilon \to 0$ and $s' \to \infty$ as $\epsilon \to 0$.

Eq. 20 implies

$$\sup \left\{ \frac{\sigma}{||\boldsymbol{x}||}; \sigma, \boldsymbol{x}, f \text{ s.t. } \hat{p} \geq \epsilon \right\} \leq \sup \left\{ \frac{\sigma}{||\boldsymbol{x}||}; \sigma, \boldsymbol{x}, f \text{ s.t. } \hat{p} \geq \frac{\epsilon}{2} \right\} \tag{21}$$

$$\text{and } \frac{1}{2} s' \leq \frac{1}{2} \inf \left\{ \frac{\sigma}{||\boldsymbol{x}||}; \sigma, \boldsymbol{x}, f \text{ s.t. } \hat{p} < \frac{1}{2} - \epsilon \right\}. \tag{22}$$

So, Eqs. 21, 22 and 19, with k large enough, imply

$$\sup \left\{ \frac{\sigma}{||\boldsymbol{x}||}; \sigma, \boldsymbol{x}, f \text{ s.t. } \hat{p} \geq \epsilon \right\} < \frac{1}{2} \inf \left\{ \frac{\sigma}{||\boldsymbol{x}||}; \sigma, \boldsymbol{x}, f \text{ s.t. } \hat{p} < \frac{1}{2} - \epsilon \right\}. \tag{23}$$

Eq. 23 provide k_1, k_2, c' and b' as follows for Eqs. 4 and 3:

$$\begin{aligned}
\frac{1}{b'} &= 2 \sup \left\{ \frac{\sigma}{||\boldsymbol{x}||}; \sigma, \boldsymbol{x}, f \text{ s.t. } \hat{p} \geq \epsilon \right\} \\
\frac{1}{c'} &= \frac{1}{2} \inf \left\{ \frac{\sigma}{||\boldsymbol{x}||}; \sigma, \boldsymbol{x}, f \text{ s.t. } \hat{p} < \frac{1}{2} - \epsilon \right\} \\
k_1 &= \lfloor \epsilon k \rfloor \\
k_2 &= \left\lceil (\frac{1}{2} - \epsilon) k \right\rceil
\end{aligned}$$

$\epsilon < \frac{1}{10}$ (due to step 1), and Eqs. above imply $c' \geq 2b'$.

Step 3: showing that $\boldsymbol{k}$ large enough and $\boldsymbol{\sigma}$ well chosen leads to at least one mutation with significant improvement. Similarly, k large enough yield

$$\begin{aligned}
b^{-1} &= \sup \left\{ \frac{\sigma}{||\boldsymbol{x}||}; \sigma, \boldsymbol{x}, f \text{ s.t. } \hat{p} > k_1/k \right\}, \\
c^{-1} &= \inf \left\{ \frac{\sigma}{||\boldsymbol{x}||}; \sigma, \boldsymbol{x}, f \text{ s.t. } \hat{p} < k_2/k \right\},
\end{aligned}$$

which provide Eqs. 5 and 2 with $b < c$ thanks to $\epsilon < \frac{1}{10}$ (ϵ was chosen with $\epsilon < \frac{1}{10}$ in step 1). Eqs. 2-5 then imply $b < b'$ and $c' < c$.

We now have to ensure Eq. 6. Equations Eq. 1-5 are proven above for k sufficiently large; from now on, we note $q = q_{\boldsymbol{x},\sigma,f}$ the probability that $f(\boldsymbol{x}+\sigma\boldsymbol{\delta}_1)$ is in $E' = f^{-1}([0, \eta f(\boldsymbol{x})[)$, and $\hat{q} = \hat{q}_{\boldsymbol{x},\sigma,f}$ the frequency $\frac{1}{k}\sum_{i=1}^{k} \mathbf{1}_{\boldsymbol{x}+\sigma\boldsymbol{\delta}_i \in E'}$. For showing Eq. 6, let us assume

$$b^{-1} \leq \frac{\sigma}{||\boldsymbol{x}||} \leq c^{-1};$$

this implies $q > \epsilon_0$ for some $\epsilon_0 > 0$; so for k sufficiently large for ensuring $\sup_{\sigma,\boldsymbol{x},f} |q_{\boldsymbol{x},\sigma,f} - \hat{q}_{\boldsymbol{x},\sigma,f}| \leq \epsilon_0/2$, by VC-dimension, we get $q' \geq \epsilon_0/2 > 0$, which implies that at least one $\boldsymbol{\delta}_i$ verifies $\boldsymbol{x} + \boldsymbol{\delta}_i \in E'$. This is exactly Eq. 6.

Step 4: concluding. We have shown Eqs. 1-6 for square roots of positive definite quadratic normal forms with bounded conditioning. Therefore, the main theorem can be applied and leads to Eq. 17. □

5 Discussion and Conclusion

This work provides, to the best of our knowledge, the first proof of linear convergence of evolutionary algorithms (here, the Simple Evolution Strategy in Alg. 1) in continuous domains on non quasi-convex functions. Indeed, even the application to quadratic positive definite forms is new. This proof is for derandomized samplings only, which means that the mutations $\boldsymbol{\delta}_i$, before multiplication by the step-size which obviously varies, are constant. A main missing point for an application is the evaluation of the convergence rate as a function of condition numbers (see extensions below) and the extension to randomized algorithms preferred by many practitioners.

In Section 5.1 we discuss extensions of this paper that we plane to develop in the near future, and in Section 5.2 deeper (harder to get rid of) limitations.

5.1 Extensions

Two properties are used for applying our main theorem to quadratic functions with a bound on condition numbers:

- VC-dimension of level sets. VC-dimension is a classical easy tool for showing that a family of functions verify a property such as Eq. 19 for arbitrarily small $\epsilon > 0$, provided that k is large enough.
- Eq. 20, also crucial in the proof, is directly a consequence of bounded conditioning (assumption formalized in Eq. 16).

With these two assumptions, we can show Eqs. 1-6, and then the theorem can be applied. This is enough for objective functions with level sets having simple graphical representations with rotations/translations.

However, we do not need assumptions so strong as finite VC-dimension for showing Eqs. 1-6. Glivenko-Cantelli results are enough; and for this, finiteness of the bracketing covering numbers, for example, is enough [13]; this is the most natural extension of this work. In particular, there are results showing the finiteness of bracketing covering numbers for families of Hölderian spaces of functions; this is a nice path for applying results from this paper to wide families of functions.

Assumptions in [2] are slightly different from assumptions in this paper; their main assumption are

- the frontier of any level set $f^{-1}(r)$ has a bounded curvature.
- for some $C_{min} \in \mathbb{R}$ and $C_{max} \in \mathbb{R}$, with $\boldsymbol{x}^*$ the (assumed unique) optimum of the objective function f and $f(\boldsymbol{x}^*) = 0$, for any $r \in \mathbb{R}$, we have

$$B(\boldsymbol{x}^*, C_{min} r) \subset f^{-1}(r) \subset B(\boldsymbol{x}^*, C_{max} r).$$

The second assumption is equivalent to our conditioning assumption, but the first one is not directly equivalent to our derandomized sampling assumptions. Refining the assumptions might be possible by combining their assumptions and our assumptions.

Condition numbers are classical for estimating the difficulty of local convergence; a nice condition number for difficult optimization should generalize some classical condition number from the literature, and include non-differentiable functions as well. [12] did a first step for that; in particular, isotropic algorithms do not solve functions with infinite condition number (for the definition of [12]), whereas covariance-based algorithms [10, 8] do.

5.2 Limitations

In this paper, we work on an evolutionary algorithm for which mutations $\boldsymbol{\delta}_i$'s are randomly drawn once and for all (the same mutation vectors $\boldsymbol{\delta}_1, \ldots, \boldsymbol{\delta}_k$ for all iterations of the algorithms). This makes the proof much easier. We believe that the proof can be extended to the case in which the mutations are randomly drawn at each iteration, as in most usual cases; yet, the adaptation is not straightforward; we must study the frequency (over iterations) at which assumptions 2-6 hold, and the consequences of bad cases on Eq. 15. For this paper, we just assume that the $\boldsymbol{\delta}_i$'s are randomly drawn once and for all iterations; equivalently, they could be quasi-randomized.

Cumulative adaptation [9] is not considered in our analysis; this is a considerably harder step for generalizing our results, because then the simple separation between 5 cases (see Fig. 2) is the idea that clearly divides the proof between step-size adaptation and progress rate.

This work covers quadratic functions, but the rates are not independent of the conditioning, so complementary results are necessary for algorithms evolving a covariance matrix, such as [10, 3, 8]. Maybe ergodic Markov chains are a better tool for showing such results [1].

We work under assumptions which imply a very large k. More precisely, using VC-dimension or bracketing numbers, it is possible to get explicit bounds on k, but these numbers will be far above the usual values for k. Obtaining results for limited values of k is a classical challenge in machine learning, and for the moment only huge values of k are applicable when using VC-dimension assumptions. Seemingly, weaker assumptions are enough, such as Glivenko-Cantelli properties [13]. For this paper, VC-dimension is easier to use and sufficient for our purpose.

Acknowledgements We are grateful to Rémi Bergasse [2] for interesting discussions.

References

1. A. Auger. Convergence results for (1,λ)-SA-ES using the theory of φ-irreducible Markov chains. *Theoretical Computer Science*, 334(1-3):35–69, 2005.
2. R. Bergasse. Stratégies d'évolution dérandomisées. Technical report, Ecole Normale Supérieure de Lyon, 2007.
3. H.-G. Beyer and B. Sendhoff. Covariance matrix adaptation revisited - the CMSA evolution strategy. In G. Rudolph, T. Jansen, S. M. Lucas, C. Poloni, and N. Beume, editors, *Proceedings of PPSN*, pages 123–132, 2008.
4. A. Conn, K. Scheinberg, and L. Toint. Recent progress in unconstrained nonlinear optimization without derivatives, 1997.
5. L. Devroye, L. Györfi, and G. Lugosi. *A probabilistic Theory of Pattern Recognition*. Springer, 1997.
6. S. Droste, T. Jansen, and I. Wegener. On the optimization of unimodal functions with the (1+1) evolutionary algorithm. In A. Eiben, T. Bck, M. Schoenauer, and H.-P. Schwefel, editors, *Parallel Problem Solving from Nature - PPSN V*, volume 1498 of *Lecture Notes in Computer Science*, pages 13–22. Springer Berlin / Heidelberg, 1998. 10.1007/BFb0056845.
7. N. I. M. Gould, D. Orban, and P. L. Toint. Cuter and sifdec: A constrained and unconstrained testing environment, revisited. *ACM Trans. Math. Softw.*, 29(4):373–394, 2003.
8. N. Hansen and A. Ostermeier. Completely derandomized self-adaptation in evolution strategies. *Evolutionary Computation*, 9(2):159–195, 2001.
9. A. Ostermeier, A. Gawelczyk, and N. Hansen. Step-size adaptation based on non-local use of selection information. In *Parallel Problem Solving from Nature PPSN III*, pages 189–198. Springer, 1994.
10. H.-P. Schwefel. *Numerical Optimization of Computer Models*. John Wiley & Sons, New-York, 1981. 1995 – 2^{nd} edition.
11. T.Back, F.Hoffmeister, and H.Schewefel. A survey of evolution strategies. Technical report, Dpt. of Computer Science XI, University of Dortmund, D-4600 , Dortmund 50, Germany, 1991.
12. O. Teytaud. Conditionning, halting criteria and choosing lambda. In *EA07*, Tours France, 2007.
13. A. V. D. Vaart and J. Wellner. *Weak Convergence and Empirical Processes*. Springer series in statistics, 1996.
14. V. Vapnik and A. Chervonenkis. On the uniform convergence of relative frequencies of events to their probabilities. In *Theory of probability and its applications, 16:264-280*, 1971.