Quasi-random mutations for evolution strategies

Olivier Teytaud, Mohamed Jebalia, Anne Auger

Algorithms (X,sigma,eta) : quasi-random mutations for Evolution Strategies

O. Teytaud[1,2], M. Jebalia[1], A. Auger[3]

[1] Equipe TAO - INRIA Futurs, LRI, Bât. 490, Université Paris-Sud, 91405 Orsay Cedex. France
[2] Artelys, 12 rue du 4 septembre, 75002 Paris, France www.artelys.com
[3] CoLab, ETH Zentrum CAB F 84, Universitätstrasse 6, 8092 Zürich, Switzerland

Abstract. Randomization is an efficient tool for global optimization. We here define a method which keeps :

- the order 0 of evolutionary algorithms (no gradient) ;
- the stochastic aspect of evolutionary algorithms ;
- the efficiency of so-called "low-dispersion" points ;

and which ensures under mild assumptions global convergence with linear convergence rate. We use i) sampling on a ball instead of Gaussian sampling (in a way inspired by trust regions), ii) an original rule for step-size adaptation ; iii) quasi-monte-carlo sampling (low dispersion points) instead of Monte-Carlo sampling.

We prove in this framework linear convergence rates i) for global optimization and not only local optimization ; ii) under very mild assumptions on the regularity of the function (existence of derivatives is not required).

Though the main scope of this paper is theoretical, numerical experiments are made to backup the mathematical results.

1 Todo et commentaires sur les differentes modifs

1.1 Quelques commentaires divers

Les regles de typologie en anglais sont differentes du francais. Je suis loin d'etre une experte mais je sais qu'il ne faut pas d'espace avant les virgules ou point virgule.

J'ai essaye de mettre des labels qui ont un sens avec le nom de la section et utilise la convention classique:

labelsec:... pour une section

labelssec:... subsection, etc

Le package "showlabels" montre le nom des labels dans la marge (a enlever sans faute pour la version finale).

Il ne faut pas mettre de forme reduite a l'ecrit: pas de "Let's" mais "Let us".

1.2 A faire

Mohamed: (j'ai mis des labels TODO dans le texte sur certains trucs ou tu dois intervenir, cf aussi dessous pour les trucs un peu plus complique)

– definir **Interrupted local descent:** (Section 3.3)
Mohamed: c est fait, mais il faut changer la place ou c est defini ou le mettre en footnote lors de la definition de l algorithme
Anne 12 Septembre: c'est change de place.
Mohamed Le 12 Septembre: OK!

– Lemma 3, c'est quoi ce n_i le definir proprement et est ce que ca a besoin d'etre dans le theoreme?

Mohamed:

En fait n_i n a pas ete tres bien defini : la on veut parler de la i eme generation -qui est egalement la generation n - a laquelle il y a generation a l etape c d un nouvel individu qu on notera $(x_{(i)}, \sigma_{(i)}, \eta_{(i)})$ qui serq gqrde a la fin de la generation autrement dit qui va prendre la place de la plus pire des descentes locales. donc les indices i s il y en a une infinite vont former une sous suite de la suite n des generations, et un n_i correspond a la i eme generation de la suite des genration n dans laquelle il y a une descente locale interrompue et le lemme stipule, grosso-modo, qu il y aura au moins une de ces suites qui ne sera jamais interrompue.. et ca sera, je crois, la suite qui va converger dans la demonstration du theoreme.
Anne 12 Septembre: J'ai essaye de le definir dans le Lemme 3 Verfifier que c'est bon.
Mohamed Le 12 Septembre: Je viens de verfier, c est aussi OK!

– Theoreme et Lemme, donner les hypotheses dans les enonces. (We assume that Assumption A1 are satisfied.) Il faut egalement essayer de mettre les hypotheses minimales pour chaque Lemme / Theoreme, est ce qeu vraiement tout le paquet d'hyopthese est necessaire a chaque fois ou est ce qu'il faut couper les hypotheses en deux, trois. Ca aide vraiment a voir plus clair dans le cheminement d'avoir simplement les hypotheses necessaires.
Mohamed: Il manque peut etre de preciser quelles sont les hypotheses qui ont ete vraiment utilise dans un thm ou un lemme donne. De toute facon toutes les hypotheses sont dans la section assumptions.Je le fais si j ai le temps ou bien ulterieurement
Mohamed Le 12 Septembre:
Pour le lemme 1, les hypotheses qui ont ete utilisees dans la preuve sont:

- la definition d une descente locale (c est pas vraiment une hypothese mais une definition)
- les hypotheses sur $\Delta(B)$ donc Assumption A1/1 (puisque dans les hypotheses du lemme on a $x_n \in B$)
  Anne le 14 Septembre Je ne comprends pas pourquoi tu dis que l'on a besoin de l'hypothese A1.1 dans le lemme 1, $x_n \in B$ ne concerne pas Δ mais je ne comprends pas non plus pourquoi on a besoin de la minoration de η_n par Δ.

- Assumption Assumption A1/5 (puisque dans les hypotheses du lemme on a $x_n \in V$)

Anne le 14 Septembre pareil je comprends pas bien pourquoi on a besoin de ca non plus.

Remarque concernant la preuve du lemme2, il y a: "As $\sigma_k \leq \sigma\eta^k$...", c est pas plutot $\sigma_k = \sigma\eta^k$?????

Anne le 14 Septembre oui c'est sans doute plus clair si on met l'egalite, je l'ai fait.

je propose qu on modifie un petit la preuve du lemme 2 en ecrivant:

PROOF: As $\sigma_k \leq \sigma\eta^k$ (ou $\sigma_k = \sigma\eta^k$), so for any $p > 0$,

$$d(x_k, x_k + p) \leq \sigma\eta^k(1 - \eta^p)/(1 - \eta) \leq \sigma\eta^k/(1 - \eta)$$

so x_k is a Cauchy sequence in a Banach space so it converges. Let us call x_∞ it's limit.So, the following holds

$$d(x_k, x_\infty) \leq \sigma\eta^k/(1 - \eta)$$

jusqu'ici, on n a utilise que la definition de descente locale pour montrer que la suite x_k est convergente. Here we have two cases:

1. Assume that $f(x_k) \to f^*$. For k sufficiently large, $x_k \in V$, and so by hypothesis 5
$$f(x_k) \leq f^* + \alpha(\sigma\eta^k/(1 - \eta))^\beta$$
2. On the other hand, as $f(x_k)$ decreases, if it does not converge to f^*, then it is lower bounded by a value $> f^*$.

□

Anne le 14 Septembre Modifs faites dnas la preuve du lemme 2.

Donc dans la preuve du lemme 2,on a utilise:
les hypotheses Assumption A1/5 etAssumption A1/2
et des proprietes issues de la definition d une descente locale comme le fait que $f(x_k) \leq f(x_{k+1})$ donc $f(x_k)$ est une suite decroissante .

Donc pour les lemmes 1 et 2 les hypotheses utilises ont ete la defintion de descente locale avec la definition de B et les hypotheses Assumption A1/2 Assumption A1/5, c est a dire que dans ces deux lemmes on a montre des choses relatives a la suite de descentes x_k relatifs a la fonction f (qui verfie l hypothese Assumption A1/5) a optimise et a l ensemble B (qui verfie l hypothese Assumption A1/1) sans se soucier de l algorithme evolutionnaire proprement dit.

2 Introduction

Evolutionary algorithms (EAs) are zeroth-order stochastic optimization methods somehow inspired by the Darwinian theory of biological evolution: emergence of new species is the result of the interaction between natural selection and blind variations. Among the class of Evolutionary Algorithms, Evolution Strategies (ES) [**?**,**?**] are the most popular algorithms for solving continuous optimization problems, *i.e.* for optimizing real-valued function f defined on a subset of $\mathbb{R}^{dim}$ for some dimension dim. The common feature of EAs is to evolve a set of points of the search space: at each iteration, some points of the search space are randomly sampled, then evaluated (the f value of the points is computed) and last, some of them are selected. Those three steps are repeated until a stopping criterion is met.

Since the invention of ESs in the mid-sixties, researches to improve the performances of ESs focused on the so-called mutation operator [**?**,**?**,**?**]. This operator consists in sampling a gaussian random variable with a given step-size σ and a given covariance matrix C. The main issue has been the adaptation of the step-size parameter σ and of the covariance matrix C. The first step in this direction is the well-known one-fifth rule [**?**] based on the rate of successful mutations. Then Rechenberg [**?**] and Schwefel [**?**] proposed to self-adapt the parameters of the mutation operator, by mutating the step-size as well (this being usually achieved by multiplying the step-size by a log-normal random variable). For this technique, the so-called *mutative step-size adaptation*, a step size is associated to every individual in the population. This step-size undergoes variations and is used to mutate the object parameters of the individual. The individual is selected with its step-size and therefore the step-sizes automatically adapted. Intuitively unadapted step-sizes can not give successively good individuals.

In this paper, we use a similar concept for adapting the scale of the sampling at each generation but use a uniform sampling in a ball instead of the standard Gaussian distribution. The motivation is that with a ball we have a trust region-effect ([**?**]), *i.e.* the local operator can be trusted in this ball. Note that though this is not classical in the evolutionary computation community, Rudolph [**?**] already introduced –mainly for theoretical purposes– sampling of the unit ball instead of a Gaussian sampling. We also make use of a deterministic sampling, or *quasi random sampling*, where we moreover minimize the dispersion of the quasi-random points [**?**,**?**]. Quasi-random numbers have already proved to be successful in many areas one of which is the field of Monte Carlo methods allowing to speed up the convergence of those methods [**?**,**?**] but as far as we know low-dispersion points are new for the evolutionary computation community.

On a theoretical point of view, many papers deal with asymptotic properties of evolutionary algorithms [**?**,**?**] or their finite time convergence in discrete cases [**?**], but convergence rates are only given under strong assumption (unimodal functions and/or very convex functions and/or very smooth functions and/or only local convergence) [**?**,**?**,**?**,**?**,**?**,**?**,**?**]. In this paper we investigate the convergence of the new algorithm considered and we prove its convergence with order one. Compared to bundle-related methods (e.g. [**?**]), which ensure superlinear convergence, we here ensure global (non-convex) convergence and we are strictly of 0-order as we do not use sub-gradients.

The paper is organized as follows: Section 3 presents our algorithm, Section 4 presents the theoretical results and Section 5 investigates numerically the theoretical results; section 6 comments upon the results obtained and conclude.

3 Definitions and properties

{qmca}

In this section we introduce the algorithm considered in this paper. As for the self-adaptive Evolution Strategies (SA-ES), a step-size is associated to each individual, moreover for reasons that will become clear in the sequel one individual is a triplet (x, σ, η) and not only (x, σ) as for the SA-ES. To create new points, the so-called *descent* operator is applied. It consists in choosing the best point among N neighbors of x (where the scale of the neighborhood is given by σ) and updating σ with η (see below). At each generation, new individuals are also randomly sampled. Finally individuals created from both sides are submitted to selection. After giving some definitions, we formally describe the *descent* operator and the algorithm:

3.1 General definitions

{ssec:def}

We consider the minimization of a real valued objective function f defined on X a subset of the real space $\mathbb{R}^{dim}$. We assume that the minimum of f is reached on X and denote $f^* = \min_{x \in X} f(x) \in \mathbb{R}$. Therefore $f := X \mapsto [f^*, \infty[$. Let opt denote the set of optima, *i.e.*

$$\mathrm{opt} = \{x \in X / f(x) = f^*\}.$$

Let $x \in \mathbb{R}^{dim}$ be a vector of $\mathbb{R}^{dim}$ and r a positive real number. We will denote $B(x, r)$ the closed ball of center x and radius r.
For a set E embedded in X we will denote $\overline{E}$ the complementary of X in $E \subset X$. $|E|$ will denote the cardinal of E.
The Euclidean distance on $\mathbb{R}^{dim}$ will be denoted $d(.,.)$, *i.e.* let $(x, y) \in \mathbb{R}^{dim} \times \mathbb{R}^{dim}$

$$d(x, y) = \sqrt{\sum_{i=1}^{dim} (x_i - y_i)^2}$$

3.2 Exploitation operator "descent"

{ssec:descent}

Let B be a set of N points of the search space, $B = \{B_1, \ldots, B_N\}$, we define *descent* as

$$descent(x, \sigma, \eta) = (x + \sigma B_\star, \eta\sigma, \eta)$$

where $\star = argmin_{j \in [[1,N]]} f(x + \sigma B_j)$ (any of the optimal in case of equality).

3.3 Algorithm

{ssec:algorithm}

The algorithm we investigate in the sequel is an evolutionary algorithm where a population P_n, where n is the iteration or generation index, is evolved. Each individual of the population is a triplet $(x, \sigma, \eta) \in \mathbb{R}^{dim} \times \mathbb{R}^+ \times \mathbb{R}^+$.

1. Sampling of N points $B = \{B_1, \ldots, B_N\}$ included in $B(0, 1)$.
2. Sampling of the initial population P_0 of (x, σ, η)
3. For n varying from 0 à $+\infty$
   (a) Creation of P^a_{n+1}, empty population.
   (b) **Descent step:** for each $(x, \sigma, \eta) \in P_n$, add $descent(x, \sigma, \eta)$ in P^a_{n+1} ; the population at the end of this step is P^b_{n+1}. [4]
   (c) **Random sampling step:** Random sampling of new individuals (x, σ, η) (see the Assumption subsection for the details), $P^{b'}_{n+1}$, the new population is

$$P^c_{n+1} = P^b_{n+1} \cup P^{b'}_{n+1}$$

   (d) **Selection step:** Selection of the best $|P_n|$ element of P^c_{n+1}, the population so generated is P_{n+1}.
   (e) Increase N by 1 and regenerate B, if at least one local descent is interrupted.

Local descent: We call *local descent* a sequence of successive points $((x_1, \sigma_1, \eta_1), \ldots, (x_n, \sigma_n, \eta_n))$ generated at step 3b, *i.e.*

$$\text{For } i > 1 \ (x_i, \sigma_i, \eta_i) = descent(x_{i-1}, \sigma_{i-1}, \eta_{i-1}).$$

Interrupted local descent: We will say that a local descent is interrupted if for some i (x_i, σ_i, η_i) is removed by the selection step.
Dispersion of B: We note $\Delta(B)$ (or Δ for short) the dispersion of B defined as

$$\Delta(B) = \sup_{x \in B(0,1)} \inf_{y \in B} \| x - y \| .$$

4 Results

{strois}

The convergence of the algorithm previously defined is analyzed in this Section.

4.1 Assumptions

We consider $V = f^{-1}([f^*, f^* + s[)$ for a given s, and assume that V is a neighborhood of $opt = f^{-1}(f^*)$.

{ass:algo}

Assumption A. *1. We require that Step 1 and 3e ensure that $0 \in B$, that Δ is non-increasing in N and that $\Delta \to 0$ as $N \to \infty$. For example, we might assume that each new B generated minimizes $\Delta(B)$ under the constraint $0 \in B$.*
2. We forbid $\eta \geq 1$ or $\eta \leq 0$; in all cases $\eta \in]0, 1[$.

[4] At the end of this step, we have $|P^b_{n+1}| = |P_n|$.

3. *The generation method (step c) must generate 3-uples (x, σ, η) in an i.i.d manner ; the number of generated 3-tuples is upper bounded by a given constant G, and the density is lower bounded by $c > 0$ and upper bounded by $d < \infty$ on $V \times]0, 2\sup_{(a,b)\in V\times V} \parallel a - b \parallel [\times]0, 1[$, and x, σ and η are independent. Moreover, we generate at each step c at least one point (which can be removed in the selection step).*
4. *We keep, at step d, the $|P_n|$ best elements for the fitness. This selection depends on x only (not on σ and η) : in particular, $|P_{n+1}| = |P_n|$ and $\forall (x, \sigma_x, \eta_x) \in P^d_{n+1}, \forall (y, \sigma_y, \eta_y) \in P^c_{n+1} \setminus P^d_{n+1}\ f(x) \leq f(y)$.*
5. *We assume that if $x \in V$, the following holds :*

$$f^* + \alpha' d(x, opt)^\beta \leq f(x) \leq f^* + \alpha d(x, opt)^\beta$$

   with $\beta > 0$ and $0 < \alpha' \leq \alpha$.
6. *For $\epsilon > 0$ sufficiently small, the probability of generating (by random generation at Step 3c) an optimal point within ϵ is lower bounded by $K\epsilon^C$ and upper bounded by $K'\epsilon^C$ for some $C, K, K' > 0$ (consequence of assumption 5 and 3), i.e. $K\epsilon^C \leq P(f(x) \leq f^* + \epsilon) \leq K'\epsilon^C$.*

Comments: The fact that the coefficient β is the same on the left-hand and on the right-hand side in assumption 5 is, for us, the strongest assumption. Assumption 4 can be removed, with some technical modifications of the proof.

4.2 Preliminary results

We prove that if $\Delta(B)$ is sufficiently small in front of the constants of the problem and of η_n, and if the optimum is inside the initial ball, then linear convergence occurs.

Lemma 1 (Linear descent). *If $x_n \in V$ and $B(x_n, \sigma_n)$ intersects opt and*

$$\eta_n \geq \sqrt[\beta]{(\frac{\alpha}{\alpha'})}\Delta(B)$$

then $d(descent^k(x_n, \sigma_n, \eta_n), opt) \leq \eta_n^k \sigma_n$

PROOF: By induction, we show that $(c_k, r_k, \epsilon_k) = descent^k(x_n, \sigma_n, \eta_n)$ verifies : $opt \cap B(c_k, r_k)$ is non-empty. As the radius of the ball is upper-bounded by $\sigma_n \eta_n^k$, the result follows.

□

We now prove the following Lemma:

{lem:dicho}

Lemma 2. *Let $(x_k, \sigma_k, \eta) = descent^k(x, \sigma, \eta)$, then either* **P 1** *or* **P 2** *(but not both simultaneously) holds:*

P 1. *for k sufficiently large, $f(x_k) \leq f^* + \alpha(\sigma\eta^k/(1-\eta))^\beta$,*
P 2. *$f(x_k)$ is lower bounded by a constant $> f^*$.*

Interpretation: Some sequences converge quickly to the optimum and some sequences are lower bounded. There is no sequence converging slowly or sequence whose successive fitness accumulate around the optimum without converging to it.

PROOF: Assume that $f(x_k) \to f^*$. As $\sigma_k = \sigma\eta^k$, for any $p > 0$ we have

$$\begin{aligned} d(x_k, x_{k+p}) &\leq \sigma\eta^k(1 + \eta + \ldots + \eta^{p-1}) \\ &= \sigma\eta^k \frac{(1-\eta^p)}{(1-\eta)} \leq \frac{\sigma\eta^k}{(1-\eta)}. \end{aligned}$$

Then $(x_k)_{k\in\mathbb{N}}$ is a Cauchy sequence which therefore converges. Let x_∞ be its limit, from the previous equation, the following holds

$$d(x_k, x_\infty) \leq \sigma\eta^k/(1-\eta).$$

Only two situations can occur

1. Either $f(x_k) \to f^*$ and consequently for k sufficiently large, $x_k \in V$. With Assumption **A**.5 we have
$$f(x_k) \leq f^* + \alpha(\sigma\eta^k/(1-\eta))^\beta$$
which is the property **P 1.**.
2. Either $f(x_k)$ does not converge to f^* but as $f(x_k)$ decreases it is lower bounded by a value $> f^*$ which is the property **P 2.**

□

Satisfactory individual: The 3-uple (x, σ, η) is said ***satisfactory*** if the property **P 1.** defined in Lemma 2 holds.

{lem:lowbounddes}

Lemma 3. *Let $(n_i)_{i\in\mathbb{N}}$ be the subsequence of the index generation $n \in \mathbb{N}$ such that there exists an individual $(x_{(i)}, \sigma_{(i)}, \eta_{(i)})$ in P_n^c generated at Step c* **and** *selected at Step d. When $(x_{(i)}, \sigma_{(i)}, \eta_{(i)})$ is not unique, we choose it arbitrarily among possible points minimizing $f(x_{(i)})$.*

Assume that there are infinitely many interrupted local descent (which is equivalent to the fact that there are infinitely many i such that n_i is well defined). Then, for a given C, $P((x_{(i)}, \sigma_{(i)}, \eta_{(i)})$ satisfactory and non-interrupted$) \geq C > 0$ infinitely often.

Interpretation : Lemma 3 states that if infinitely many new local descent occur, then infinitely many of these new descents have a lower bounded probability of being uninterrupted. Lemma 3 will be used in the main Theorem to get a contradiction : if infinitely many new descents are started, by Lemma 3 (almost surely) infinitely many of them are non-interrupted, so there are more and more non-interrupted sequences, so, as the population is bounded after a finite time there is no more room for a new descent (see the Theorem for more details).

PROOF:

1. Assume that n_i is well defined for all $i \in \mathbb{N}$. Note that this implies that Δ decreases to 0 (by hypothesis 1).

2. Note w_n the worst fitness among P_n^b. By construction w_n is non-increasing. As it is lower-bounded, it converges.
3. Let us show that it almost surely converges to f^*. The proof is as follows :
   - Assume, in order to get a contradiction, that w_n is lower bounded by some $f^* + \epsilon$ where $\epsilon = 1/2^k$ for some integer $k > 0$.
   - Then with Assumption **A**. 6, infinitely many new points (generated in steps 2c) are generated with fitness $< f^* + \epsilon$.
   - The number of points in P_n^b with fitness $\geq f^* + \epsilon$ is decreased of one at each generation of points with fitness $< f^* + \epsilon$. As this occurs infinitely often, after a finite time (almost surely), w_n must decrease below $f^* + \epsilon$. This is true for any $\epsilon = 1/2^k$ with probability 1; by countable intersection, it is true with probability 1 for all $\epsilon = 1/2^k$.
   - Therefore w_n decreases to $f^* + \epsilon$.
4. Note that $f(x_{(i)}) \leq w_{n_i}$ (because if $f(x_{(i)}) \geq w_{n_i}$ then by construction, $x_{(i)}$ would not be selected). Therefore, the fitness of $x_{(i)}$ converges to f^*.
5. Let us show that the event

$$\{(x_{(i)}, \sigma_{(i)}, \eta_{(i)}) \text{ satisfactory and } \eta_{(i)} \leq 0.9\}$$

   occurs with probability at least $1 - D$ for some $D < 1$ if i is sufficiently large.
   - The event $\{(x_{(i)}, \sigma_{(i)}, \eta_{(i)})$ satisfactory and $\eta_{(i)} \leq 0.9\}$ in particular holds if the assumptions of Lemma 1 and $\eta \leq 0.9$ are verified. This is the case whenever $\sigma \geq d(opt, \overline{V})$ and $0.9 \geq \eta \geq \Delta \sqrt[\beta]{\alpha/\alpha'}$ and if $f(x) < f^* + \alpha' d(opt, \overline{V})^\beta$.
   - The latter inequality holds if i is sufficiently large, as $f(x_{(i)})$ converges to f^*.
   - Other inequalities occur independently with probability lower bounded by a constant > 0, provided that Δ is sufficiently small.
   - The probability of these three inequalities simultaneously is lower-bounded by a positive constant $1 - D$ ($D < 1$), provided that Δ is sufficiently small. Δ goes to 0 (point 1 above) and therefore Δ is sufficiently small if i is sufficiently large.
6. Note E_i' the event that $\sigma \geq d(opt, \overline{V})$ and $0.9 \geq \eta \geq \Delta \sqrt[\beta]{\alpha/\alpha'}$ and $f(x) < f^* + \alpha' d(opt, \overline{V})^\beta$. We have shown above that $P(\neg E_i') \geq 1 - D$.
7. Note E_i the event $\{(x_{(i)}, \sigma_{(i)}, \eta_{(i)})$ verifies E_i' and is never interrupted $\}$ in the sense that its successive sons generated in step (b) are never eliminated in step (d).
8. By Lemma 2, if E_i' occurs, then the k^{th} iterate of the local descent (from $(x_{(i)}, \sigma_{(i)}, \eta_{(i)})$) has fitness bounded above by $\alpha^C(\sigma_{(i)}\eta_{(i)}^k/(1 - \eta_{(i)}))^\beta$.
9. Therefore, conditionally to E_i', the probability of interruption of the k^{th} iterate is upper bounded by $K'\alpha^C(\sigma_{(i)}\eta_{(i)}^k/(1 - \eta_{(i)}))^{\beta C}$.
10. So $P(\neg E_i | E_i')$ is upper bounded by the $\sum_{k=0}^{\infty} K'\alpha^C(\sigma_{(i)}\eta_{(i)}^k/(1 - \eta_{(i)}))^{\beta C}$.
11. Now, recall that $P(\neg E_i) = P(\neg E_i | E_i')P(E_i') + P(\neg E_i')$ (as E_i implies E_i'), and therefore $P(\neg E_i) \leq P(\neg E_i | E_i') + P(\neg E_i')$.
12. Then, combining points 11, 10 and 6 above, $P(\neg E_i) \leq 1 - D + \sum_{k=0}^{\infty} K'\alpha^C(\sigma_{(i)}\eta_{(i)}^k/(1 - \eta_{(i)}))^{\beta C}$.
13. σ having a density lower-bounded by a constant > 0 in the neighbourhood of 0, $P(E_i)$ is infinitely often larger than a given $W > 0$ (for example, $W = 1 - D/2$).

Hence the expected result : E_i, having probability $\geq W > 0$ for any i (conditionally to the past and current epochs of the algorithm), occurs almost surely infinitely often.

□

4.3 Almost sure convergence with order one

We now investigate the global convergence properties of our algorithm. The delicate part is that it is not enough to have the fact that after a finite number of iterations we are close to the optimum and therefore convergence holds. Indeed, there is always a risk that a local descent is interrupted. Therefore we are going to formalize in the proof below the fact that with probability 1, under minimal assumptions, there is a non-interrupted local descent that converges linearly. We emphasize the fact that this proof could be applied for other operators as well. The only requirement is to have enough fast convergence for the local operator. The heart of the proof can be outlined as follows:

- any non-satisfactory local descent will be interrupted (consequence of Lemma 2 and of Assumption 6) by a new local descent; each new local descent has a probability lower bounded by a constant > 0 of being satisfactory ; so, there are infinitely many satisfactory local descent (this is step 1 of the proof below) as long as none of them is satisfactory and non-interrupted ; so we always have a satisfactory local descent among the future populations ;
- these local descents have a probability of being interrupted which decreases so quickly (by Lemma 3), that after some time they are no more interrupted (this is the step 2 of the proof) ;
- hence, the convergence is linear (step 3) and moreover N is bounded (step 3).

The detailed proof comes after the Theorem:

Theorem 1. *We have almost sure convergence at least linear of the error to the optimal error, i.e.* $inf_{(x,\sigma,\eta)\in P_n}\ (f(x) - f^*) \leq A/B^n$ *for some* $A > 0$ *and* $B > 1$. *Moreover,* N *is almost surely bounded.*

PROOF:

1. **Step 1 : Let us show that with probability** 1**, there exists infinitely many values of** n **such that there exists** (x, σ, η) **satisfactory in** P_n^d**.**
   Let us make the hypothesis H1 (to get a contradiction), that for any $n > n_0$, there is no (x, σ, η) in P_n^d such that $f(descent^k(x, \sigma, \eta)) \to f^*$ for $k \to \infty$ (independently of any interruption ; we consider the theoretical sequence of $descent^k(.)$ as $k \to \infty$).
   Moreover, let us assume (one again in order to get a contradiction), the hypothesis H2: there exists n_1 such that for any $n > n_1$, the step (c) of generation of points does not provide any point better than the worst point resulting from step (b).
   Then, if $n > n_1$, $P_n^d = \{descent^{n-n_1}(x, \sigma, \eta) | (x, \sigma, \eta) \in P_{n_1}^d\}$; moreover, N, B and Δ become constant. The $f(descent^{n-n_1}(x, \sigma, \eta))$ are lower bounded by a given $f^* + \epsilon$, for a given $\epsilon > 0$. This is proved by the application of:

– H1 (which states that none of the local descents converges) and
– Lemma 2 (which states that if local descents do not converge to f^* then they are lower bounded).

to the finit set of local descents from $P^d_{n_1}$.
Then for each n, at step 3c, the probability of generating a new point (x_n, σ_n, η_n) better than the local descents is lower bounded by some $P*$, where $P*$ is provided by Assumption **A**.6.
So, such a generation necessarily occurs, with probability 1.
So, we have a contradiction with H2. So, under hypothesis H1, H2 does not hold, infinitely often, a new point (x, σ, η) generated at step c is added to P^d_n.
We have assumed H1, and proved that H2 does not hold. Let us now look for a contradiction, so that we can prove that H1 does not hold.
N increases for each n such that the followings holds : "a point generated at step (c) is integrated to P^d_n". As this occurs infinitely often (as $H2$ is false), $\Delta \to 0$.
Consider the probability of generating (x, σ, η) satisfactory.

$$\Pi = P((x, \sigma, \eta) \text{ satisfactory}|s) \geq$$
$$\underbrace{P(x \in V|s)}_{\Pi_1} \times \underbrace{P(\sigma \geq sup_{z \in V} d(z, opt)|s)}_{\Pi_2}$$
$$\times \underbrace{P(\eta \geq \sqrt[\beta]{\alpha/\alpha'}\Delta(B)|s)}_{\Pi_3}$$

where $P(E|s)$ is the probability of an event E conditionnally to the fact that the point (x, σ, η) coming from the generation step (c) is selected and is the best selected point.
Π_1 is asymptotically lower bounded by a constant > 0 (and indeed converges to 1), Π_2 is lower bounded by a positive constant thanks to Assumption **A**.3, and Π_3 is lower bounded by a positive constant when Δ is sufficiently small, what occurs as $\Delta \to 0$.
The probability of getting a (x, σ, η) satisfactory and non-interrupted is thus lower-bounded for each step n during which a new point is generated at step c. Consequently this event occurs necessarily for infinitely many values of n, with probability 1.
So with probability 1, we have contradiction with hypothesis H1. So we can claim that there exists infinitely many values of n such that there exists some (x, σ, η) in P^d_n such that $descent^k(x, \sigma, \eta) \to opt$ if $k \to \infty$.

2. **Step 2 : Let us show that finitely many points (x, σ, η) generated in (c) are selected in (d).**
Note $(x_{(i)}, \sigma_{(i)}, \eta_{(i)})$ the sequence of 3-uples generated at step c and selected in P^d_n (not removed by the selection step) and satisfactory (ie, are in the first case of Lemma 2) and are the best (from the point of view of the fitness) among the (x, σ, η) generated in step c and incorporated in $P_{n,d}$.
Let us do, in order to get a contradiction, the hypothesis that this sequence is infinite (which is equivalent to assuming that there are infinitely many 3-uples generated in step (c) selected in step (d)).

Then, for i large enough $P((x_{(i)}, \sigma_{(i)}, \eta_{(i)})$ verifies Lemma 1 and is not interrupted$)$ is infinitely often lower bounded by a positive constant (Lemma 3).
So, this occurs, almost surely, infinitely often. As the number of non-interrupted local descents is bounded above by the population size, there is contradiction.

3. **Conclusion :**
By step 1, we know that with probability 1, infinitely many 3-uples (x, σ, η) satisfactory are in some P_n^d.
By step 2, we know that these 3-uples can only a finite number of times come from random generations (as only a finite number of points can come from step (c) and be included to P_n^d). So, finitely many local descents are interrupted (each interruption is the integration in (d) of a point coming from step (c)).
So after a finite time, no more local descent is interrupted; N **is now constant** (and so, does not go to infinity) and the satisfactory local descent (whose existence is almost sure thanks to step 1) **goes to the optimum, with linear convergence** thanks to Lemma 2.

□

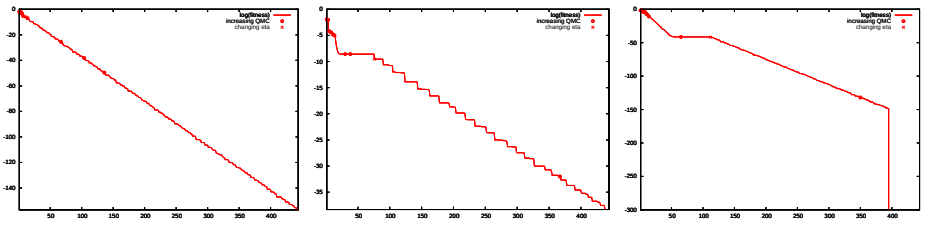
5 Practical experiments

{pratique}

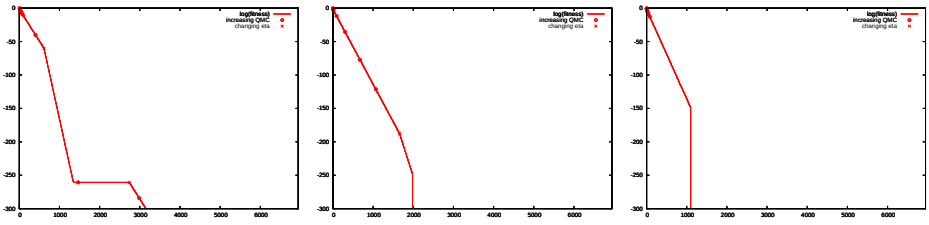
We have experimented our method on different simple objective functions $f_{L_p}(x) = \sqrt[p]{\sum x_i^p}$ satisfying the assumptions we made for the convergence. Figure 1 shows the linear convergence of the method. We observe the changes of convergence rates due to the changes of η associated to the best point in the population and the increases of N leading to a N-points quasi-random sampling. The choice of B for a given value of N has been performed by optimizing the disrepancy of the points. This part of the procedure is time-consuming when N increases. Note that such sets of points in the ball could of course be evaluated off line. Very efficient and fast algorithms exists for quasi-monte-carlo generation in the sense of standard discrepancy, but as far as we know no equivalent algorithms exist for the optimization of Δ. Interestingly, experiments with random sampling once per increase of N leads to similar results (note that the result about linear convergence remains theoretically true) but the case with one new sampling at each 3c step leads to much worse results. This suggests that quasi-random mutations (at least, stabilizing the random part by keeping the same B until N increases) are not only of theoretical interest (for proving our results of linear convergence on a very large family of fitness functions) but also of practical interest. Note that on the other hand, we need random points for the almost sure convergence and we did not proceed to any quasi-randomization of this random part - in this work globalization remains the work of random.

These results are naive results coming from an Octave implementation. A more optimized implementation, based on EO classes in C++, is in progress. First in dimension 2, for norm L_p with $p = 1$, $p = 3$, $p = 5$; "increasing QMC" denotes epochs at which $N \leftarrow N + 1$:

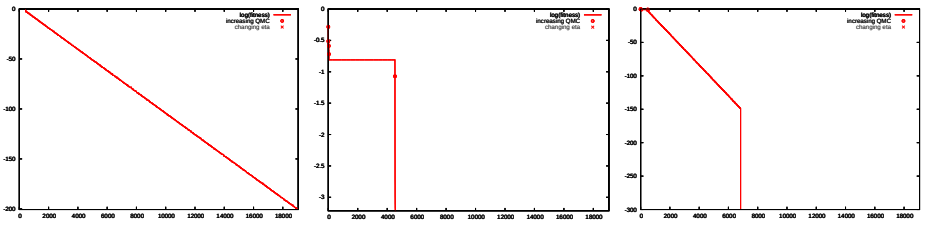
Figure 2 presents the histogram of the distribution of $\log(f_{L_5})$ after $500 \times (d/3)^2$ fitness-evaluations.



dim=2



dim=5



dim=10. In the second case, convergence did not occur yet, the log(fitness) stays at -3.

Fig. 1. Fitness value in logarithmic scale vs number of generations for $f_{L_p}(x) = \sqrt[p]{\sum x_i^p}$ with respectively from left to right, $p = 1, 3, 5$. Due to numerical precisions, $\log(f_{L_p})$ can be equal to $-\infty$. A cross indicates when a new η is chosen. A circle indicates when N is increased by 1. The random generation for x is uniform on $[-1, 1]^d$, η is uniform on $[0, 1]$, 10σ is the absolute value of a standard Gaussian, the population size is 5, the number of random generations at step 3c is 25 and N is initialized to 1. A cross indicates when a new η is chosen. A circle indicates when N is increased by 1. It may be observed that N quickly stabilizes.

{fig:conv}

6 Discussions - Conclusions

{conclusion}

We have designed a new algorithm using a representation (x, σ, η) instead of (x, σ). This algorithm takes into account different areas of applied mathematics:

- quasi-random points (low-dispersion points, [**?**]);
- trust-regions ([**?**]);
- adaptive step-size coming from evolution strategies [**?**,**?**];
- random diversification of the population for global optimization.

A very important remark is that as for classical ES, the algorithm considered here only use the information given by the fitness through the ranking of individuals. Therefore everything is invariant with respect to monotonic transformation of the fitness. In particular all the results holds for $x \mapsto g(f(x))$ where f satisfies the assumptions required for our Theorems and g is a strictly increasing function. This implies notably that convexity is not required for the convergence.

Compared to state-of-the art theoretical results for convergence of adaptive evolution strategies [**?**], our assumptions are here weaker. Indeed in [**?**] asymptotic linear convergence is proved for any $x \mapsto g(f(x))$ where g is monotonic and f is the sphere function. The main points here are i) use of (x, σ, η) instead of (x, σ) ; ii) generation of points on a close ball, instead of Gaussian sampling, so that this algorithm can ensure (under some conditions which are asymptotically satisfied with probability 1) that the fact that the optimum lies in $B(x, \sigma)$ is preserved from parents to children ; iii) use of quasi-random sequences ensuring that Δ goes to 0 as $N \to \infty$.

Experiments confirm the theoretical study but are very preliminary. In fact, we implemented the precise Algorithm, where each generation at step 3c has to be generated independently with the same distribution at each epoch, whereas intuition suggests that better heuristics for new generations should dramatically reduce the time before reaching linear convergence; such implementations, and the corresponding proofs are yet to be done. Note that even in dimension 10, our very simple implementation, thanks to linear convergence, could reach the limit of the machine precision. These results are not at all results due to multiple attempts and empirical calibration of the parameters; we simply implemented the algorithm in a naive manner, without any heuristic added; our results are the most immediate consequences of theory above.

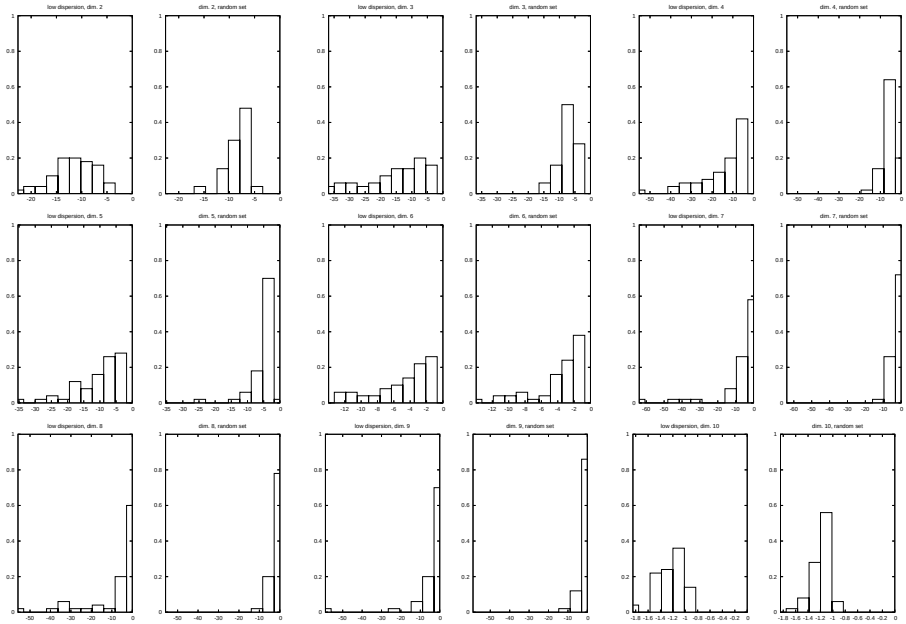


Fig. 2. Histogram of the distribution of $\log(f_{L_5})$ after $500 \times (d/3)^2$ fitness-evaluations for the dimension indicated at the top of the graphs. For each couple of graph, on the left with low-dispersion points resulting from gradient-based optimization on $\Delta(B)$; on the right, with random points

{fig:histo}