Why one must use reweighting in Estimation Of Distribution Algorithms

Fabien Teytaud, Olivier Teytaud

Why one must use reweighting in Estimation Of Distribution Algorithms

F. Teytaud and O. Teytaud
TAO (Inria), LRI, UMR 8623(CNRS - Univ. Paris-Sud),
bat 490 Univ. Paris-Sud 91405 Orsay, France, fteytaud@lri.fr

ABSTRACT

We study the update of the distribution in Estimation of Distribution Algorithms, and show that a simple modification leads to unbiased estimates of the optimum. The simple modification (based on a proper reweighting of estimates) leads to a strongly improved behavior in front of premature convergence.

Categories and Subject Descriptors

G.1.6 [**Optimization**]: Unconstrained optimization; G.3 [**Probability and statistics**]: Multivariate statistics

General Terms

Algorithms

1. INTRODUCTION

The American election of 1936 is famous due to the error, in a very important part of American history [2], in the poll organized by the Literary Digest [1], in spite of a huge sampling - 2 millions of questionnaires. This error was due to the absence of reweighting: the readers of the Literary Digest were much more often Republicans than Democrats, and were therefore much more likely to vote for Landon; whereas Roosevelt finally got 61 % of votes, the Literary Digest predicted a comfortable win for Landon. The good result was predicted by George Gallup, with only a much smaller sample (50 000 people).

Estimation of Distribution Algorithms are similar to polls: they are based on samplings. However, to the best of our knowledge, the reweighting has not been experimented or analyzed in this context. This paper is based on this idea.

This paper considers the adaptation of the distribution in Estimation of Distribution Algorithms (EDA). EDA include e.g. UMDA [17], Compact Genetic Algorithm [11], Population-Based Incremental Learning [3], Relative Entropy [16], Cross-Entropy [4] and Estimation of Multivariate Normal Algorithm (EMNA) [12]. We will here focus on EMNA (cf Algorithm 1).

There are several advantages in EDA; simplicity, possible use of prior knowledge, compliance with mixed (continuous/discrete) spaces, applicability for multimodal optimization (with multimodal distributions calibrated by Expectation-Maximization algorithms); importantly, EDA are often nearly parameter free. However, an issue is premature convergence [23, 9, 14, 20]: for example, with a Gaussian EDA, if the initial point is too far from the target (formally, if the squared distance to the optimum divided by the initial variance is too large), then the EDA might diverge.

Algorithm 1 The EMNA algorithm.

Initialize $\sigma \in \mathbb{R}$, $x \in \mathbb{R}^N$.
while Halting criterion not fulfilled **do**
  for $l = 1..\lambda$ **do**
    $z_l = \sigma N_l(0, Id)$
    $x_l = x + z_l$
    $y_l = f(x_l)$
  end for
  Sort the individuals by increasing fitness; $y_{(1)} < y_{(2)} < \cdots < y_{(\lambda)}$.
  $z^{avg} = \frac{1}{\mu} \sum_{i=1}^{\mu} z_{(i)}$
  $\sigma = \sqrt{\frac{\sum_{i=1}^{\mu} ||z_{(i)} - z^{avg}||^2}{\mu \times N}}$
  $x = x + z^{avg}$
end while

2. VARIANCE REDUCTION

Random points have a lot of nice features, in particular the fact that it's easier to avoid bias with simple Monte-Carlo. How, several forms of variance reductions are possible without introducing any bias. These techniques are also termed importance sampling in recent literature, referring to either correcting an incorrect distribution (as in the present paper), or improving the variance by changing the distribution.

- Using symmetries of the problem provides huge improvements. This was very clearly shown with Buffon's needle. The problem in Buffon's needle consists in estimating the probability for a needle thrown at random on a table with parallel lines on it, to cross at least one of the lines (Buffon's needle can be used for approximating π, yet it's essentially a toy problem nowadays).

The interesting point is that this probability is more quickly estimated by launching something else than a needle: launch (several times) a cross, made of two needles joint at their middle, estimate the probability for this cross to meet at least one line, apply some algebra to the frequency, and you'll find a good estimate, much more precise than if you had simply launched a needle. Using such symmetries (using e.g. so-called antithetic variables) was exploited successfully in the field of EDA in [8], and in various other fields as well - e.g. [21] for Bayesian Networks. We will not take care of this specific techniques, which can be used simultaneously as (and independently of) the reweighting that we propose below.

- Stratification (related, but not equivalent to, quota sampling - in quota sampling the distribution in each "quota" is not specified whereas in stratification it is): choosing a partition of the domain, and sample independently each part of the partition; then, each point should be reweighted so that its weight is consistent with its real probability. Stratification can very strongly reduce the variance of the estimate. When you can't sample with the target distribution (as well as in EDA, when we use Gaussian sampling whereas we would like to sample uniformly - there's just no uniform distribution such that we are sure that for λ sufficiently large we will go sufficiently far away from the parent to cover the optimum), then reweighting is the key solution for removing the bias.

Discussions related to variance reduction can be found in [5, 10]. Some interesting historical and other related elements can be found in [22, 15].

3. WHY EMNA IS NOT (ALWAYS) CONSISTENT AND HOW TO MAKE IT CONSISTENT

In an update step of EMNA, we generate $x_1, \ldots, x_\lambda$, independently identically distributed (i.i.d). Typically the distribution is Gaussian, centered on x and with step-size σ. We consider this simple case, but the adaptation to other distributions is straightforward.

$y_1, \ldots, y_\lambda$ are the fitness values of the x_i; for some unknown fitness function f, $y_i = f(x_i)$. We then consider the sorted population; $x_{(1)}, \ldots, x_{(\lambda)}$ have fitness $y_{(1)} \leq y_{(2)} \leq \cdots \leq y_{(\lambda)}$ and are a permutation of $x_1, \ldots, x_n$. We will assume for the sake of clarity that all fitness values are distinct, but the result does not depend on this assumption.

Then we define

$$\hat{x} = \frac{1}{\mu} \sum_{i=1}^{\mu} x_{(i)}. \tag{1}$$

$\hat{x}$ is the estimate of the location of the optimum; it's the center of the next distribution. With the notation $\hat{\mathbb{E}}_\lambda$ (resp. $\hat{\mathbb{P}}_\lambda$) for the empirical averaging (resp. frequency) operator for the λ points $(x_1, y_1), \ldots, (x_\lambda, y_\lambda)$, this is equivalent to

$$\hat{x} = \hat{\mathbb{E}}_\lambda(x | y \leq y_{(\mu)}).$$

μ is usually equal to $\lambda/4$; we will consider, for simplicity, that $\lambda = 4\mu$, i.e. that λ can be divided by 4, but this assumption is not central; μ/λ might just converge to some limit in $]0, 1[$ as λ is large.

We now analyze conditions under which EDA are consistent.

THEOREM 1 (CONSISTENCY OF EDA). *Consider $x_1, x_2, x_3, x_4, \ldots$ i.i.d with density $d(x)$, and consider $y_1, y_2, y_3, y_4, \ldots$ such that the (x_i, y_i) are independently and identically distributed as (x, y) with $y = f(x)$. We assume that:*

1. *The repartition function $P(y \geq t)$ is continuous at $1/4$ and the $\frac{1}{4}$-quantile of y is well defined, i.e. there is one and only one t such that $P(y \leq t) = \frac{1}{4}$.*
2. *If $I \subset \mathbb{R}$ is compact then $x|y \in I$ has bounded range;*
3. *$\forall t > 0$, $\mathbb{E}_\lambda x | y < t = \arg\min f$.*

Then, with $\mu \to \infty$ and $\lambda = 4\mu$, $\hat{x} \to \arg\min f$.

Remarks:

- Condition 1 is stronger than necessary. The proof is simpler with this condition, which is enough to ensure that (i) there is a $\frac{1}{4}$ quantile (ii) there are not so many points very close to this quantile. A careful reduction of this assumption is beyond the scope of this paper.
- Condition 2 in particular holds if the function is coercive. With coercive functions, for $||x||$ sufficiently large, $f(x) > \sup I$.
- Condition 3 is the important condition for this paper: it looks like a simple technical assumption, but it does not hold in many important cases, as discussed later.

Proof: Assumption 1 implies that the $\frac{1}{4}$ quantile of y is well defined. The convergence of an empirical quantile (here $y_{(\mu)}$) to the quantile (and the existence of this quantile) holds almost surely as soon as this quantile is well defined (classical corollary of Kolmogorov-Smirnov's theorem). So assumption 1 implies

$$y_{(\mu)} \to t; \tag{2}$$

for the only t such that $P(y \leq t) = \frac{1}{4}$.

Consider some fixed $1 > \epsilon > 0$, assumption 1 implies that

$$\hat{\mathbb{P}}_\lambda(y \in]t - \epsilon, t + \epsilon[) \leq k(\epsilon) \tag{3}$$

for some function $k(.)$ such that $\lim_{\epsilon \to 0} k(\epsilon) = 0$.

Proof of Eq. 3: We need $k(\epsilon)$ such that $P(y \leq t + \epsilon) - P(y \leq t - \epsilon) \leq k(\epsilon)$; we can just set $k(\epsilon) = P(y \leq t + \epsilon) - P(y \leq t - \epsilon)$. By the continuity assumption, $k(\epsilon)$ goes to 0 as $\epsilon \to 0$. $\square$(proof of Eq. 3)

Let's define $I = [t - \epsilon, t + \epsilon]$. Then,

- for λ sufficiently large, by Eq. 3,

$$P(y \in I) \leq k(\epsilon); \tag{4}$$

- Thanks to assumption 2,

$$\hat{\mathbb{E}}_\lambda(x | y \in I) < k' \tag{5}$$

for some constant k' independent of ϵ.

Now, consider the following equation for $\hat{x}$.

$$\left| \hat{\mathbb{E}}_\lambda x | y < y_{(\mu)} - \hat{E} x | y < t \right| \leq \hat{P}(y \in I) \left| \hat{\mathbb{E}}_\lambda(x | y \in I) - \hat{\mathbb{E}}_\lambda x | y < t \right| \tag{6}$$

Eqs. 4 and 5 plugged in Eq. 6 lead to:

$$\hat{\mathbb{E}}_\lambda x|y < y_{(\mu)} = \hat{\mathbb{E}}_\lambda x|y < y_t + O(k(\epsilon)). \quad (7)$$

The limit of Eq. 7 for $\epsilon \to 0$ is $\hat{E}x|y < y_\mu \to \mathbb{E}_\lambda x|y < t$, which in turn leads to

$$\hat{x} \to \mathbb{E}_\lambda x|y < t. \quad (8)$$

Eq. 8 and assumption 3 conclude the proof. □

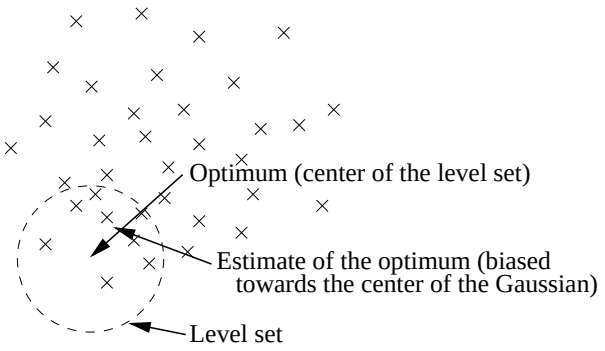

Figure 1: Illustration of the bias induced by selection in the case of unweighted points and non-uniform (e.g. Gaussian) distribution.

A possible solution consists in using a uniform distribution in a sufficiently large ball. However, Gaussian numbers are quite comfortable: they provide arbitrarily large values, whilst preserving small variance. How to have the best of both worlds ? The idea of reweighting consists in:

- sampling with your most comfortable distribution (here the Gaussian), with density $c(.)$;
- be aware of the target distribution, i.e. the distribution with which there would be no bias; let $t(.)$ be its density (here the uniform distribution);
- reweighting the points in order to correct the bias; the weight of a point x is the following ratio:

$$w(x) = \frac{t(x)}{c(x)}.$$

The implementation for EDA is quite straightforward: replace the weight 1 by this ratio between the target density and the density used for sampling. We will define the corresponding algorithm and experiment it in the next section. Interestingly, this ratio is nearly constant among selected points when they are close to the center of the optimum, but not at all when the selected points are far from the optimum, i.e. precisely in the cases in which EDA are sensitive to premature convergence.

4. EXAMPLES AND COUNTER-EXAMPLES

Let's discuss the assumption of this theorem. Assumptions 1 and 2 are weak assumptions, implies by e.g. smoothness and coercivity assumptions. Assumption 3 is seemingly quite natural but indeed does not hold in many cases. This section is devoted to examples and counter-examples for assumption 3. Thanks to reweighting we can come back to the averages for the uniform distribution; so we have to see cases in which the center of level sets is (resp. is not) the optimum. Essentially, we will see that in many cases the problem is not solved; essentially we can deal with local convergence. As discussed in section 6, reweighting makes sense in many cases; however, the proof as made in this paper considers only consistency of the estimation of the optimum.

4.1 Quadratic functions

Let's first consider quadratic functions.

- The standard case of a quadratic positive definite function (Fig. 2) is well handled by the theorem. The optimum is at the center of ellipsoids and therefore assumption 3 of the theorem clearly holds *for the uniform distribution* - for a Gaussian distribution, we have to reweight the points in order to ensure consistency. Typically, if the level sets are concentric ellipsoids, then $\mathbb{E}_\lambda x|y < t$ is equal to the optimum when the distribution is uniform, but certainly not if the distribution is Gaussian - in the case of a Gaussian distribution, the points are much more densely distributed close to the center of the distribution (see Figure 1); therefore, the estimate is not consistent, and increasing λ does not solve the problem.

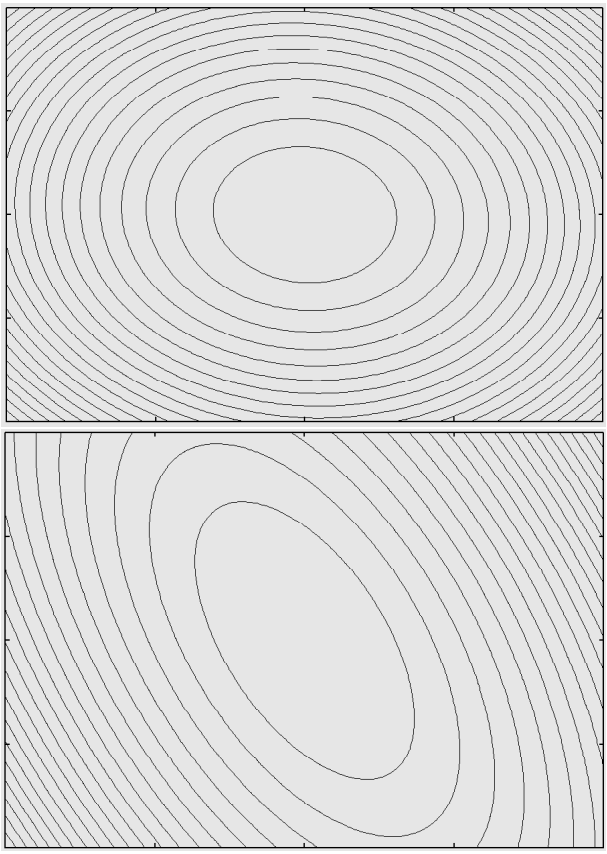

Figure 2: Standard cases of quadratic definite functions with or without bad conditioning. This case is handled by the theorem.

- The degenerated case (Figure 3) in which the level sets are infinite rectangles, with a line of optima, is not

directly handled; we have only considered the case in which the optimum is a point and the level sets are bounded (assumption 2). We conjecture that the result should be nearly preserved but have no proof of it.

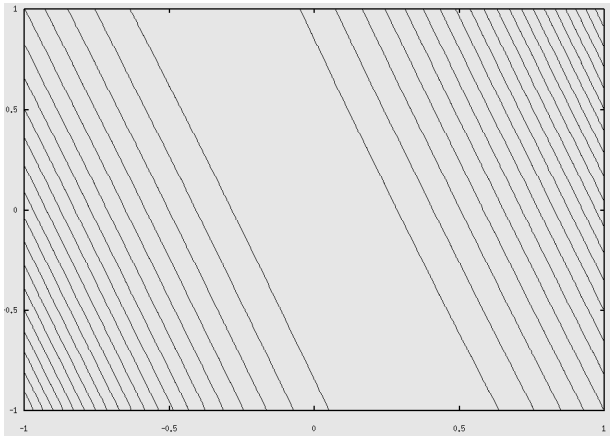

Figure 3: Degenerated quadratic function; one eigenvalue is null.

- The case of a saddle point (Figure 4) is not well handled also - the level sets are not ellipsoids at all, and we have to estimate level sets and not just the location of the optimum with some $\hat{x}$ (see however section 6, first point).

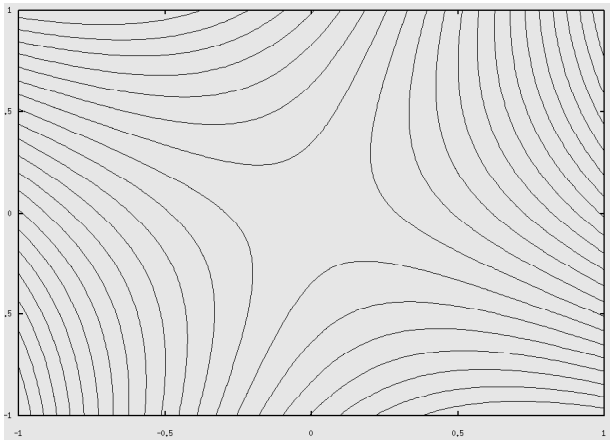

Figure 4: Quadratic function with saddle point. This is a multimodal case. This is not handled by the theorem; we do not have to estimate the position of the optimum but the level set.

4.2 Ellipsoid level sets of non-quadratic functions

We have discussed quadratic functions; this is not equivalent to ellipsoid level sets. We point out that whenever the sampling is made thanks to the uniform distribution, the fact that all level sets are ellipsoids is not enough for ensuring that the center of a level set is the optimum. Consider for example the following fitness function:

$$f^{-1}(t) =$$

$$\{cos(0.9t) + t * cos(u) + sin(0.9t) + t * sin(u);$$

$$u \in [0, 2\pi]\} \text{ for } t \geq 0.$$

This fitness (see level sets in Fig. 5) has optimum in $(1, 0)$, which is the center of almost no level set for the uniform distribution.

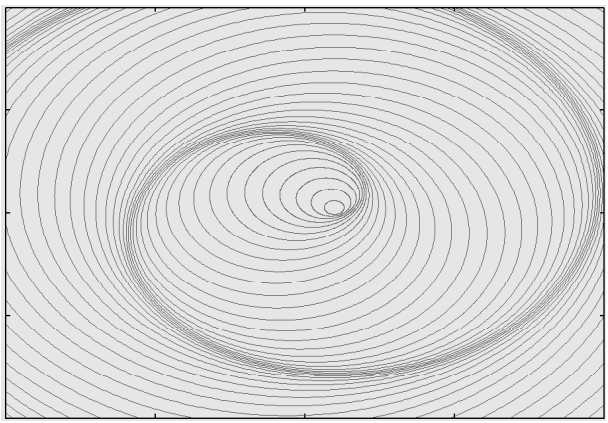

Figure 5: Level sets of a fitness function which does not satisfy assumption 3 whenever the level sets are ellipsoids.

For a fitness as on Fig. 5, correcting the sampling bias in order to "recover" the uniform sampling is not enough; i.e. the reweighting technique that we propose below is not proved consistent in that case. However, in spite of the lack of proof in the general case, we guess that our technique also avoids premature convergence in such cases. To the best of our knowledge, proofs of consistency in such non-quasi-convex functions are still very rare (see however [25]).

4.3 Other cases; why we try to get to much with the consistency theorem

We have shown several counter-examples. However, these counter-examples are cases in which $\hat{x}$ does not converge to the optimum as $\lambda \to \infty$, in one generation; this is not necessary for the convergence of the EDA. Therefore we require too much. For example, the two very similar Figures 6 and 7 have probably very similar behaviors; but only the first one is handled, due to the dissymmetry of case 7.

5. EXPERIMENTAL RESULTS

This section is devoted to experimenting the convergence of the weighted version of EMNA in a framework in which the baseline version that not converge, namely an initial point far from the optimum and a small initial step-size. In this section we use Algorithm 2 (to be compared with Algorithm 1, the baseline EMNA).

5.1 Comparison on the sphere function with poor initialization: reweighting avoids premature convergence

First, algorithm 2 is now compared to Algorithm 1. The results are averaged over 11 runs in various dimensions. The initial point is $(1, 1, \ldots, 1)$, the number of generations is $25\lfloor D^{3/2} \rfloor$ and the initial step-size is $\sigma = 0.1$. The population size is $\lambda = 10D\lfloor\sqrt{D}\rfloor$ in Fig. 8, $\lambda = 10 \times D^2$ in Fig.

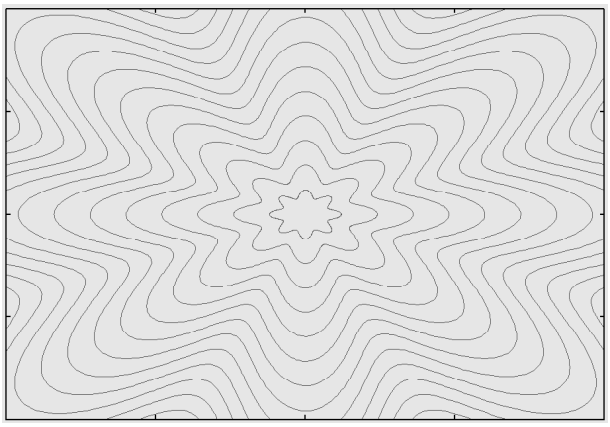

Figure 6: Here, the t level set is $\{t\times(1+0.2*cos(8u))\times cos(u), t\times(1+0.2*cos(8u))sin(u); u\in[0,2\pi]\}$. Thanks to symmetries, the assumption 3 is verified: the center of the level set is the optimum.

Algorithm 2 The EMNA algorithm with weighted averages.

Initialize $\sigma \in \mathbb{R}$, $y \in \mathbb{R}^N$.
while Halting criterion not fulfilled **do**
  for $l = 1..\lambda$ **do**
    $z_l = \sigma N_l(0, Id)$
    $y_l = y + z_l$
    $f_l = f(y_l)$
  end for
  Let $w(i) = 1/density(x_i)$ with $density$ the density of the distribution used for generating the offspring.
  Sort the indices by increasing fitness; $f_{(1)} < f_{(2)} < \cdots < f_{(\lambda)}$.
  $z^{avg} = \frac{1}{\sum_i^\mu w_{(i)}} \sum_{i=1}^{\mu} w_{(i)} z_{(i)}$
  $\sigma = \sqrt{\frac{\sum_{i=1}^{\mu} w(i)||z_{(i)} - z^{avg}||^2}{\sum_{i=1}^{\mu} w(i)\times N}}$
  $y = y + z^{avg}$
end while

9, $\lambda = 10 \times D^3$ in Fig. 10, with D the dimension. The "scores" are $D\log(x)/n$, where n is the number of generations, x is the estimate of the optimum at the end of the run - this is the approximation of the normalized asymptotic convergence rate. 0 means a premature convergence, and a significantly negative number means a linear convergence; we have linear convergence in all cases with reweighting, and never without, as shown in Figure 10.

5.2 EMNA with reweighting avoids premature convergence even without centering assumption

We claimed earlier that assumption 3 could probably be relaxed, and in particular that even if fitness functions as in Figure 11 (all but the first plot) do not verify the assumption 3, it should lead to a linear convergence as for the "spherical" fitness function (first plot, in the same figure). Figure 12 shows that high values of x make the problem more difficult,

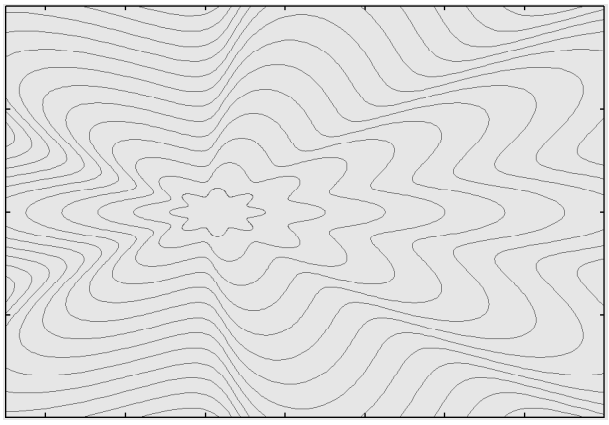

Figure 7: Here, the t level set is $\{\frac{1}{10}t + t\times(1+0.2*cos(8u))\times cos(u), t\times(1+0.2*cos(8u))sin(u); u\in[0,2\pi]\}$. Due to dissymmetry, the center of the level set if not the optimum and the condition of the theorem are not met. However, it is clear that the algorithm will converge in this case as well as we don't have to be exactly centered on the optimum. This will be discussed on a similar problem in section 5, see in particular Fig. 11.

but that the convergence holds anyway (Fig 12).

5.3 Looking closely at the evolution of σ with poor initialization

We will now investigate the evolution of σ in a restricted setting, namely only one offspring, in order to see if increasing the number of points solves the problem. Figure 13 clearly shows that increasing λ is useless for avoiding premature convergence when there's no reweighting, whereas Figure 14 shows that when weighting is used, premature convergence is consistently avoided when λ is sufficiently large. This shows the consistency of the weighted approach and the weakness of the unweighted approach.

6. DISCUSSION

We have proved a theorem of consistency of the estimation of optima in EDA thanks to reweighting. This proof has the following limitations:

- We have only shown that the algorithm is consistent for estimating the optimum when we know the distribution for which the optimum is the expected value in a level set (typically, uniform distribution if the optimum is at the center of the level sets). This shows for a Gaussian EDA, reweighting ensures that ellipsoid level sets are consistently estimated. This does not say anything for the approximate level sets for other distributions. However, this is only technical and the principle of reweighting can be extended far beyond this. This will be the subject of an extension of this work; the goal would be the proof of a property of type "with reweighting, level sets are consistently estimated for some metric".
- The consistency is also shown asymptotically, and experimentally the avoidance of premature convergence requires big samples. It is likely that other tricks from

Dimension	Score without reweighting	Score with reweighting	P-value
3	-0.00888172	-0.561429	0.00225104
4	-0.0107948	-1.35278	8.472e-10
6	-0.0110491	-1.35565	3.2428e-07
7	-0.012486	-2.43707	0
8	-0.0146252	-2.20972	0
9	-0.0145236	-2.51588	0
11	-0.0128873	-2.3909	0
12	-0.0131643	-1.931	1.46267e-09
13	-0.0152632	-1.39598	2.42622e-05
14	-0.0159825	-1.14553	0.000411108
15	-0.0165677	-1.46499	2.33928e-05
16	-0.0176763	-0.904438	0.00345763

Figure 8: Normalized convergence rates for various dimensionalities, $\lambda = 10D\lfloor\sqrt{D}\rfloor$ and p-value of significance. See text for the detailed experimental setup; the fitness is the sphere function and there is a poor initialization. Results are highly significant.

Dimension	Score without reweighting	Score with reweighting	P-value
3	-0.0104643	-1.78895	0
4	-0.0144236	-2.18543	0
6	-0.0128907	-2.51126	0
7	-0.0135233	-2.63108	0
8	-0.0162723	-2.71583	0
9	-0.016863	-2.81322	0
11	-0.0147054	-2.96843	0
12	-0.0165426	-3.04377	0
13	-0.0177544	-3.09284	0
14	-0.0189371	-3.14667	0
15	-0.0198078	-3.23661	0
16	-0.0204953	-3.31655	0

Figure 9: Normalized convergence rates for various dimensionalities, $\lambda = 10D^2$ and p-value of significance. See text for the detailed experimental setup. Sphere function, poor initialization (see text). Results are highly significant.

statistics could be used for improving the variance of estimates, e.g. bootstrap [7, 6], confidence regions for M-statistics [28]. This is not done in this paper. Interestingly, bootstrap can be used for estimating confidence intervals as well, and can be used in difficult cases (i.e. beyond the simple conditional expectation studied in this paper).

- Other classical tricks for improving samplings consist in using quasi-random sequences in the sampling [18, 19]; quasi-random sampling is an active area of research with strong recent improvements in large dimension [13, 24]. Such improvements have already been tested for mutations in evolutionary algorithms [27, 26]. Also, antithetic variables are easy to use in EDA, see e.g. [8]

- In this paper, we considered reweighting for correcting a bias (typically, Gaussian distribution instead of uniform distribution for estimating the mass center of an ellipsoidal level set). Other forms of weights can be used. In the case of integration, it has been pointed out that sometimes reweighting points or using them in a more complicated manner than simple weighted averages is more important than well distributing them [13]. For example, one can weight points proportionally to the probability of their Voronoi cell. With such a technique, [29] points out that in dimension 1 and 2 respectively, the integration error can be reduced from $1/\sqrt{n}$ (random points) or $\log(n)^d/n$ (quasi-random points) to $O(1/n^4)$ and $O(1/n^2)$ on twice differentiable functions. Unfortunately, such results only hold in small dimensions, and [29] points out that possibly, in high dimension, naive Monte-Carlo methods have some form of optimality among various possible uses of (unbiased) random i.i.d samplings.

Dimension	Score without reweighting	Score with reweighting	P-value
3	-0.0121133	-2.12388	0
4	-0.0150117	-2.28472	0
5	-0.0114129	-2.39826	0
6	-0.0135526	-2.54386	0
7	-0.0152274	-2.65001	0
8	-0.0165787	-2.73935	0
9	-0.0181852	-2.83664	0
10	-0.0144745	-2.80969	0
11	-0.0160375	-2.99463	0
12	-0.0173464	-3.06692	0
13	-0.0186479	-3.13602	0
14	-0.0195956	-3.20573	0
15	-0.020884	-3.2708	0
16	-0.022177	-3.33649	0

Figure 10: Normalized convergence rates for various dimensionalities, $\lambda = 10D^3$ and p-value of significance. Results are averaged over 11 runs; the fitness is the sphere function and the initialization is poor; see text for more details.

We think that the results can be extended essentially by taking into account that we do not have to ensure that the center of the level set is the optimum, but only that we reduce the size of the sampling whilst keeping a good density around the optimum. We should not loose in this process the following strong points of the approach presented in this paper:

- We don't need any convexity or quasi-convexity assumption. This is quite important as it is quite natural to take care of non quasi-convex fitness functions; to the best of our knowledge, only this paper and [25] have considered non-quasi-convex fitness functions.

- The results have this advantage that the proposed modification is quite simple, and can be used in many cases: it is just a reweighting. Also, this is not only theory for theory (even if theory for theory is interesting in many cases); we can implement the modification and have immediate very clear improvements.

Acknowledgements

REMOVED FOR DOUBLE BLIND

7. REFERENCES

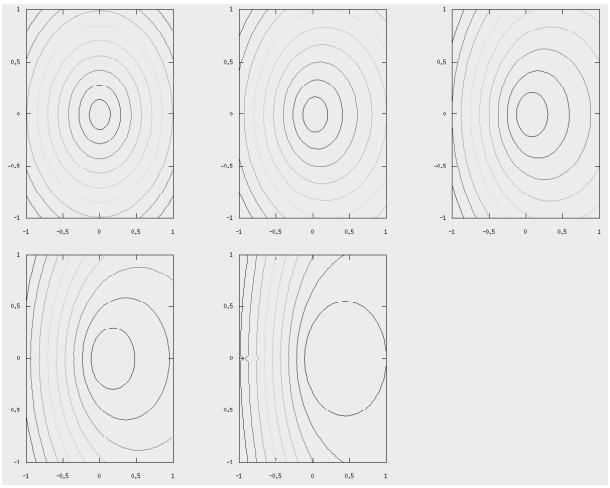

Figure 11: Fitness function with level set $f^{-1}(t) = \{(t(x+cos(u), tsin(u); u \in [0, 2\pi]\}$ for $x = 0, 0.2, 0.4, 0.6, 0.8$ respectively. With $x \neq 0$, the third assumption of the theorem does not hold; however, results in Fig. 12 clearly show that EMNA with reweighting solves this fitness function even with poor initialization.

Value of x	Score without reweighting	Score with reweighting	P-value
0	-0.00867799	-1.08108	8.83104e-10
0.1	-0.00847504	-1.27573	5.35127e-14
0.2	-0.00803038	-0.334781	0.0117148
0.3	-0.00904768	-0.343002	0.0113305
0.4	-0.00885484	-0.33109	0.00766513
0.5	-0.00822882	-0.416994	0.00179155
0.6	-0.00768837	-0.301457	0.00930038
0.7	-0.00729352	-0.159433	0.0337745
0.8	-0.00619166	-0.302029	0.00596901
0.9	-0.00563185	-0.144462	0.0337107

Figure 12: This result shows that the weighted version of EMNA solves the fitness function presented in Fig. 11 and are much faster than the non-weighted version, in spite of the fact that the third assumption of the theorem is not verified.

Dim. \ λ	10	100	1000	10000	100000	1000000
2	07	00	00	00	00	00
3	06	01	00	00	00	00
4	03	02	00	00	00	00
5	10	06	00	00	00	00
6	03	06	00	00	00	00
7	04	04	00	00	00	00
8	02	05	00	00	00	00
9	10	07	00	00	00	00
10	03	07	00	00	00	00
11	02	01	00	00	00	00
12	03	07	00	00	00	00
13	01	05	00	00	00	00
14	02	05	00	00	00	00
15	04	02	00	00	00	00
16	02	03	00	00	00	00
17	00	02	00	00	00	00

Figure 13: Without reweighting: for various values of λ, we consider an initial point $(10, 10, \ldots, 10)$ with initial $\sigma = 1$. We check the probability (percentage) that σ increases (therefore leading to convergence). We see that this probability does not increase but *decreases* to 0, consistently with theory. On the other hand, we can see on Figure 14 that with reweighting the probability converges to 1 as λ becomes large. All probabilities are estimated on 30 runs.

[1] Literary digest, 1936.

[2] K. Andersen. *The Creation of a Democratic Majority: 1928-1936*.

[3] S. Baluja. Population-based incremental learning: A method for integrating genetic search based function optimization and competitive learning,. Technical Report CMU-CS-94-163, School of Computer Science, Carnegie Mellon University, Pittsburgh, PA, 1994.

[4] Y. Cai, X. Sun, H. Xu, and P. Jia. Cross entropy and adaptive variance scaling in continuous eda. In *GECCO '07: Proceedings of the 9th annual conference on Genetic and evolutionary computation*, pages 609–616, New York, NY, USA, 2007. ACM.

[5] D. R. Cox and W. L. Smith. *Queues*. Methuen/Wiley, London, 1961.

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

[7] B. Efron and R. Tibshirani. *An introduction to the bootstrap*, volume 57 of *Monographs on Statistic and Applied Probability*. Chapman & Hall, 1993.

[8] S. Gelly, J. Mary, and O. Teytaud. On the ultimate convergence rates for isotropic algorithms and the best choices among various forms of isotropy. In 10^{th} *International Conference on Parallel Problem Solving from Nature (PPSN 2006)*, 2006.

[9] J. Grahl, P. A. Bosman, and F. Rothlauf. The correlation-triggered adaptive variance scaling idea. In *GECCO '06: Proceedings of the 8th annual conference on Genetic and evolutionary computation*, pages 397–404, New York, NY, USA, 2006. ACM.

[10] D. Gross and C. L. Harris. *Fundamentals of Queueing Theory*. Wiley & Sons, New York, 1974.

[11] G. R. Harik, F. G. Lobo, and D. E. Goldberg. The compact genetic algorithm. *IEEE Trans. on Evolutionary Computation*, 3(4):287, November 1999.

[12] P. Larranaga and J. A. Lozano. *Estimation of Distribution Algorithms. A New Tool for Evolutionary Computation*. Kluwer Academic Publishers, 2001.

[13] P. L'Ecuyer and C. Lemieux. *Recent Advances in Randomized Quasi-Monte Carlo Methods*, pages 419 – 474. Kluwer Academic, 2002.

[14] J. Liu and H.-F. Teng. Model learning and variance control in continuous edas using pca. In *ICICIC '08: Proceedings of the 2008 3rd International Conference on Innovative Computing Information and Control*, page 555, Washington, DC, USA, 2008. IEEE Computer Society.

[15] B. J. Morgan. *Elements of simulation*. Chapman & Hall, Ltd., London, UK, UK, 1984.

[16] H. Mühlenbein and R. Höns. The estimation of

Dim. \ λ	10	100	1000	10000	100000	1000000
2	03	30	92	100	98	99
3	03	48	96	93	98	99
4	02	48	90	96	97	98
5	03	49	91	94	96	98
6	00	47	85	94	97	97
7	01	42	90	89	95	97
8	01	50	86	91	92	99
9	00	37	72	91	96	95
10	01	43	79	80	92	97
11	01	35	76	87	93	90
12	00	34	71	85	88	94
13	00	26	69	89	89	96
14	00	20	63	80	90	89
15	00	17	62	82	89	90
16	00	21	75	80	81	90
17	00	19	70	79	86	93

Figure 14: As Fig. 13 but with reweighting. The probability of non-premature convergence is close to 1 for λ large.

distributions and the minimum relative entropy principle. *Evolutionary Computation*, 13(1):1–27, 2005.

[17] H. Mühlenbein and T. Mahnig. Evolutionary computation and Wright's equation. *Theoretical Computer Science*, 287(1):145–165, 2002.

[18] H. Niederreiter. *Random Number Generation and Quasi-Monte Carlo Methods*. 1992.

[19] A. B. Owen. Quasi-Monte Carlo sampling. In H. W. Jensen, editor, *Monte Carlo Ray Tracing: Siggraph 2003 Course 44*, pages 69–88. SIGGRAPH, 2003.

[20] P. Posík. Preventing premature convergence in a simple eda via global step size setting. In *PPSN*, pages 549–558, 2008.

[21] A. Salmerón and S. Moral. Importance sampling in bayesian networks using antithetic variables. In *ECSQARU '01: Proceedings of the 6th European Conference on Symbolic and Quantitative Approaches to Reasoning with Uncertainty*, pages 168–179, London, UK, 2001. Springer-Verlag.

[22] L. W. Schruben and B. H. Margolin. Pseudorandom number assignment in statistically designed simulation and distribution sampling experiments. *Journal of the American Statistical Association*, 73:504–525, 1978.

[23] J. L. Shapiro. Drift and scaling in estimation of distribution algorithms. *Evolutionary Computation*, 13(1), 2005.

[24] I. Sloan and H. Woźniakowski. When are quasi-Monte Carlo algorithms efficient for high dimensional integrals? *Journal of Complexity*, 14(1):1–33, 1998.

[25] O. Teytaud. Conditionning, halting criteria and choosing lambda. In *EA07*, Tours France, 2007. G.: Mathematics of Computing/G.1: NUMERICAL ANALYSIS, G.: Mathematics of Computing/G.1: NUMERICAL ANALYSIS/G.1.6: Optimization.

[26] O. Teytaud. When does quasi-random work?. In G. Rudolph, T. Jansen, S. M. Lucas, C. Poloni, and N. Beume, editors, *PPSN*, volume 5199 of *Lecture Notes in Computer Science*, pages 325–336. Springer, 2008.

[27] O. Teytaud and S. Gelly. DCMA: yet another derandomization in covariance-matrix-adaptation. In *GECCO '07: Proceedings of the 9th annual conference on Genetic and evolutionary computation*, pages 955–963, New York, NY, USA, 2007. ACM.

[28] A. van der Vaart and J. Wellner. *Weak Convergence and Empirical Processes, With Applications to Statistics*. Springer Series in Statistics. Springer-Verlag New York, Inc., 1996.

[29] S. Yakowitz, J. E. Krimmel, and F. Szidarovszky. Weighted monte carlo integration. *SIAM Journal on Numerical Analysis*, 15(6):1289–1300, 1978.