

On the ultimate convergence rates for isotropic algorithms and the best choices among various forms of isotropy

Sylvain Gelly, Jérémie Mary, Olivier Teytaud

▶ **To cite this version:**

Sylvain Gelly, Jérémie Mary, Olivier Teytaud. On the ultimate convergence rates for isotropic algorithms and the best choices among various forms of isotropy. Parallel Problem Solving from Nature, Sep 2006, Reykjavik. inria-00112816

HAL Id: inria-00112816
https://inria.hal.science/inria-00112816

Submitted on 9 Nov 2006

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

On the ultimate convergence rates for isotropic algorithms and the best choices among various forms of isotropy

Olivier Teytaud, Sylvain Gelly, Jérémie Mary

TAO (Inria), LRI, UMR 8623(CNRS - Univ. Paris-Sud),
bat 490 Univ. Paris-Sud 91405 Orsay, France, teytaud@lri.fr

Abstract.

```
@inProceedings{lbeda,
        author={S. Gelly and J. Mary and O. Teytaud},
        title={On the ultimate convergence rates for
isotropic algorithms and the best choices
among various forms of isotropy},
  booktitle = {$10^{th}$ International Conference on
Parallel Problem Solving from Nature (PPSN 2006)},
        year=2006}
```


In this paper, we show universal lower bounds for isotropic algorithms, that hold for any algorithm such that each new point is the sum of one already visited point plus one random isotropic direction multiplied by any step size (whenever the step size is chosen by an oracle with arbitrarily high computational power). The bound is $1 - O(1/d)$ for the constant in the linear convergence (i.e. the constant C such that the distance to the optimum after n steps is upper bounded by C^n), as already seen for some families of evolution strategies in [19, 12], in contrast with $1 - O(1)$ for the reverse case of a random step size and a direction chosen by an oracle with arbitrary high computational power. We then recall that isotropy does not uniquely determine the distribution of a sample on the sphere and show that the convergence rate in isotropic algorithms is improved by using stratified or antithetic isotropy instead of naive isotropy. We show at the end of the paper that beyond the mathematical proof, the result holds on experiments. We conclude that one should use antithetic-isotropy or stratified-isotropy, and never standard-isotropy.

1 Introduction : what is the price of isotropy

[3] has recalled that, empirically, all evolution strategies with a relevant choice of the step size exhibit a linear convergence rate. Such a linear convergence rate has been shown in various contexts (e.g. [1]), even for strongly irregular multi-modal

functions ([2]). Linearity is not so bad, but unfortunately [19, 12] showed that the constant in the linear convergence, for $(1+\lambda)$-ES and $1, \lambda$-ES in continuous domains, converges to 1 as $1 - O(1/d)$ as the dimension d increases ; this has been generalized in [17] to *all* comparison-based methods. On the other hand, mathematical programming methods, using the derivatives ([4, 7, 9, 16]), but also using only the fitness-values, reach a constant 0 in all dimensions and work in practice in huge dimension problems (see e.g. [18]).

So, we know that (i) comparison-based methods suffer from the $1-O(1/d)$ (ii) fitness-value-based methods do not. Where is the limit ? We here investigate the limit case for isotropic algorithms in two directions : (1) can isotropic algorithms avoid the $1-O(1/d)$ by using additional information such as a perfect line search with computational cost zero (2) can we do better than random independent sampling for isotropic algorithms ? The answer for (1) will be essentially no : naive isotropy leads to $1 - O(1/d)$. A more optimistic answer appears for (2) : yes, some nice samplings lead to better results than naive independent uniform samplings, namely : stratified isotropy, and antithetic isotropy.

The paper is organized as follows. Section 2 shows that a random step size forbids superlinear convergence, but allows a linear convergence with rate $\exp(-\Omega(1))$. Section 3 shows that a random independent direction forbids superlinear convergence and forbids a better constant than $1 - O(1/d)$, whatever may be the family of fitness functions and the algorithm, whatever may be its step-size rule or selection procedure provided that it uses isotropic random mutations. Section 4 then shows that isotropy does not necessarily imply naive independent identically distributed sampling, and that the convergence rate of $(1+\lambda) - ES$ on the sphere function is improved when using stratified sampling or antithetic sampling.

For the sake of clarity, without loss of generality we assume that the origin is the only optimum of the fitness (so the norm of a point is the distance to an optimum).

2 If the step-size is random

Consider an unconstrained optimization problem in $\mathbb{R}^d$. Consider any algorithm of the following form, based on at least one initial point for which the fitness has been computed (we assume that 0 has not been visited yet). Let's describe the n^{th} epoch of the algorithm :

- Consider X_n one of the previously visited points (points for which the fitness has been computed) ; you can choose it by any algorithm you want using any information you want ;
- Choose the direction $v \in \mathbb{R}^d$ with unit norm by any algorithm you want, using any information you want.
- Then, choose the step size σ in $[0, \infty[$; for the sake of simplicity of notations, we require that $\sigma \geq 0$, but if you prefer $\sigma \in \mathbb{R}$, you simply replace v by $-v$ with probability $1/2$;

– Evaluate the fitness at $X'_n = X_n + \sigma v$.

We assume that at each epoch σ has a non-increasing density on $[0, \infty[$. This constraint is verified by e.g. gaussian distributions (gaussian random variables have values in $]-\infty, \infty[$, but "gaussian steps + random isotropic direction" is equivalent to "absolute value of a gaussian step + random isotropic direction" and the absolute value of a gaussian step has decreasing density on $[0, \infty[)$. Provided that the constraint is verified for each epoch, whatever may be the algorithm for choosing the distribution, the results below will hold. The distribution can be bounded and we do not require it to be gaussian or any other particular form of distribution. This formalism includes many algorithms ; SA-ES for example are also included. What we only require is that each point is chosen by a random jump from a previously visited point (any previously tested point) with a distribution that might be restricted to a deterministic direction (possibly the exact direction to an optimum!), with density decreasing with the distance.

In all the paper, $[a]^+ = \max(a, 0)$. Then,

Theorem 1 (step-size does matter for super-linearity):

$$E\left([-\ln([||X'_n||/||X_n||]^+\right) \leq \int_{t>0} \min(1, \frac{2\exp(-t)}{1-\exp(-t)}) < \infty. \tag{1}$$

Moreover the variance is finite, and therefore this also implies that

$$\limsup \sqrt[n]{1/||X_n||} \leq \int_{t\geq 0} \min(1, \frac{2\exp(-t)}{1-\exp(-t)})dt. \tag{2}$$

Proof: The main tool is the equality

$$E[x]^+ = \int_{t\geq 0} P(x \geq t)$$

($E[x]^+ = \int_x \int_{t\geq 0} \mathbb{1}_{t\leq x} = \int_{t\geq 0} \int_x \mathbb{1}_{t\leq x} = \int_{t\geq 0} P(x \geq t)$).

We apply this to $x = -\ln(||X'_n||/||X_n||)$.

Let's then upper-bound $P(x \geq t)$ or $P(||X'_n||/||X_n|| \leq c)$ with $c = \exp(-t)$. We claim :

Lemma :

$$P(||X'_n||/||X_n|| \leq c) \leq \min(1, 2c/(1-c))$$

The proof of the lemma is as follows :

– rescale the problem so that X_n has norm 1 ;
– Let's $[A, B]$ be the segment in which X'_n verifies the inequalities $||X'_n||/||X_n|| \leq c$, with A closer to X_n. Then $[A, B]$ has size at most $2c$ (the diameter of the ball centered in 0);
– $||X_n - A|| \geq 1 - c$, so with f the density of σ, and a and b the distance to X_n of A and B respectively, $f(a)(1-c) \leq \int_0^a f(s)ds \leq 1$ because f is not increasing. Hence, the density of σ in $[a, b]$ is upper bounded by $1/(1-c)$.

The proof of the lemma is complete.

We can now finish the proof of the theorem.

$E[x]^+ \leq \int_{t>0} \min(1, 2\exp(-t)/(1-\exp(-t)))dt < \infty$ which is finite concludes the proof for the expectation (equation 1).

Ex^2 is finite as it is upper bounded by $\int_{t>0} \min(1, 2\exp(-\sqrt{t})/(1 - \exp(-\sqrt{t})))dt$. Therefore, as Ex^2 and $(Ex)^2$ are finite, the variance is finite (uniformly bounded, independently of n).

We now have to show equation 2. It can be seen as follows :

Let's define $z_n = \inf_{i\in[[1,n]]} \ln(||X'_i||)$. Then, by construction, eq. 3 and 4 hold :

$$z_n \geq \min(z_{n-1}, \ln(||X'_n||)) \tag{3}$$
$$\ln(||X'_n||) \geq \ln(||X'_n||/||X_n||) + \ln(||X_n||) \tag{4}$$

Equation 3 means that one of the followings holds

$$z_n \geq \ln(||X'_n||) \tag{5}$$
$$z_n \geq z_{n-1} \tag{6}$$

Equation 4 implies equation 7, and we can consider separately cases 5 and 6 :

- equation 7+5 lead to equation 8 ;
- equation 6 directly leads to equation 8 as $\min(0, \ln(||X'_n||/||X_n||)) \leq 0$.

Therefore in both cases equation 8 holds.

$$\ln(||X'_n||) \geq \ln(||X'_n||/||X_n||) + z_{n-1} \tag{7}$$
$$z_n \geq z_{n-1} + \min(0, \ln(||X'_n||/||X_n||)) \tag{8}$$

Dividing equation 8 by n leads to

$$z_n/n \geq \frac{1}{n}\sum_{i=1}^{n} \min(0, \ln(||X'_{n-1}||/||X_{n-1}||))$$

The average on the right-hand side is an average with finite variance ; therefore it converges almost surely to the expectation by Kolmogorov's strong law of large numbers. This provides the expected result.

The theorem is proved.□

3 If the direction is random

This section generalizes [12] to any algorithm in which each newly visited point is equal to an old one plus a vector whose direction is uniform in the sphere (whenever the distance depends on the direction, i.e. is not chosen independently

of the direction, even if it is optimal, and whenever the algorithm computes the gradient, the Hessian or anything else).

Consider an unconstrained optimization problem in $\mathbb{R}^d$. Consider any algorithm of the following form, based on at least one initial point for which the fitness has been computed :

- Consider X_n one of the previously visited points (points for which the fitness has been computed) ; you can choose this point, among previously visited points, by any algorithm you want using any information you want, even knowing the position of the optimum ;
- Choose the direction $v \in \mathbb{R}^d$ randomly in the unit sphere ;
- Choose the step size $\sigma > 0$ by any algorithm you want, using any information you want ; it can be stochastic as well ; it can depend on v, e.g. it can minimize the distance between $X_n + \sigma v$ and the optimum ;
- Evaluate the fitness at $X'_n = X_n + \sigma v$.

As the previously stated theorem, this result applies to a wide range of evolution strategies. We only require that each new visited point is chosen by a random jump from a previously visited point.

Then, the following holds :

Theorem 2 (direction does matter for convergence rates):
Assume $d > 1$. Then,
$E[-\ln(||X'_n||/||X_n||)]$ is finite and decreases as $O(1/d)$.

Proof : The main tool is the equality

$$E[x]^+ = \int_{t \geq 0} P(x \geq t) \tag{9}$$

We apply this to $x = -\ln(||X'_n||/||X_n||)$.

We have to upper bound $P(x \geq t)$.

Lemma: *The probability of having an angle between v and $-X_n$ lower than α is $\frac{1}{2} - \frac{1}{2}F_\beta(cos^2(\alpha); \frac{1}{2}, (d-1)/2)$ (using the β distribution) for $\alpha < \pi/2$ and $d > 1$.*

This lemma is a lemma for us, but it's a theorem itself, and it can be found in [8] (with also many related results that could be related to evolution strategies). A simple geometry argument shows that $||X'_n||/||X_n|| < c$ occurs with probability at most

$$\frac{1}{2}(1 - F_\beta(1 - c^2; \frac{1}{2}, (d-1)/2) \tag{10}$$

where $F_\beta(x; \beta_1, \beta_2) = \int_0^x \frac{\Gamma(\beta_1+\beta_2)t^{\beta_1-1}(1-t)^{\beta_2-1}}{\Gamma(\beta_1)\Gamma(\beta_2)}dt$,

(note that the probability in equation 10 is reached if the step size σ is chosen by minimization of $||X_n + \sigma v||$)

Therefore, $P(||X'_n||/||X_n|| < c) \leq \frac{1}{2} - \frac{1}{2}\int_0^{1-c^2} \frac{\Gamma(d/2)(t)^{-\frac{1}{2}}(1-t)^{d-1}}{\Gamma(1/2)\Gamma((d-1)/2)}$, and

$$P(-\ln(||X'_n||/||X_n|| > u)) \leq \frac{1}{2}\int_{1-exp(-2u)}^{1} \frac{\Gamma(d/2)(t)^{-\frac{1}{2}}(1-t)^{(d-3)/2}}{\Gamma(1/2)\Gamma((d-1)/2)}dt \tag{11}$$

We now compute $E = E[-\ln(||X'_n||/||X_n||)]^+$, thanks to equations 9 and 11. $E \leq \int_0^\infty \frac{1}{2} \int_{1-exp(-2u)}^1 \frac{\Gamma(d/2)(t)^{-\frac{1}{2}}(1-t)^{(d-3)/2}}{\Gamma(1/2)\Gamma((d-1)/2)} dtdu$ with equality if σ minimizes $||X_n + \sigma v||$.

$\frac{\Gamma(d/2)}{\Gamma((d-1)/2)} = (1 + o(1))\sqrt{(d-1)/2}$ ([10, 13] for a proof and more details on the $o(1)$), we see that this expectation is $E = \frac{1}{2}\left(\sqrt{\frac{d-1}{2\pi}}\right)(1 + o(1)) \int_0^\infty \int_{1-\exp(-2u)}^1 \frac{(1-t)^{(d-3)/2}}{\sqrt{t}} dtdu.$

with $f(t) = \frac{(1-t)^{(d-3)/2}}{\sqrt{t}}$:

$$E = \frac{1}{2}\left(\sqrt{\frac{d-1}{2\pi}}\right)(1 + o(1)) \int_0^\infty \int_0^1 \mathbb{1}_{t \geq 1-\exp(-2u)} f(t) dtdu$$

$$E = \frac{1}{2}\left(\sqrt{\frac{d-1}{2\pi}}\right)(1 + o(1)) \int_0^1 \int_0^\infty \mathbb{1}_{t \geq 1-\exp(-2u)} f(t) dudt$$

Then, $E = (1 + o(1))\frac{1}{4}\sqrt{\frac{d-1}{2\pi}} \int_0^1 \frac{t^{(d-3)/2}}{\sqrt{1-t}}(-\ln(t))dt$

We now just have to show that the integral decreases quickly to 0 as a function of d to get our lower bound. Just split the integral in $\int_0^{\frac{1}{2}}$ and $\int_{\frac{1}{2}}^1$:

$$E \leq K \int_0^{\frac{1}{2}} t^{\frac{d-4}{2}} \underbrace{\sqrt{-t\ln(t)^2/(1-t)}}_{bounded} + K \int_{\frac{1}{2}}^1 t^{\frac{(d-3)}{2}} \underbrace{\sqrt{-\frac{1}{2}\ln(t)^2/(1-t)}}_{=\Theta(\sqrt{1-t})}$$

The first summand is exponentially decreasing to 0 as $d \to \infty$. The second is $\Theta(1/d)$. This concludes the proof.□

4 Isotropic $(1 + \lambda)$-ES and a comparison among isotropic samplings

We have shown that with independent isotropic mutations, even with perfect step size chosen a posteriori, we have a linear convergence rate with constant $1-O(1/d)$. We can study more carefully $(1+\lambda)$-ES with perfect step size on the sphere, in order to show the superiority of unusual isotropy. $(1+\lambda)$-ES are λ-fully-parallel ; they are probably a good choice for complex functions on which more deterministic or more structured approaches would fail, and if you have a set of λ processors for parallelizing the fitness-evaluations. Therefore, it is worth studying it. We show here that you must choose $(1 + \lambda)$-ES with *stratified* isotropic or *antithetic* isotropic sampling instead of $(1 + \lambda)$ *standard* isotropic sampling. We show that, at least on the sphere, it is better in all cases. The proofs below show that the convergence rate is better, but also that the distribution of the progress rate itself ($\frac{||X_{n+1}||}{||X_n||}$) is shifted in the good direction. At least for the sphere with step size equal to the distance to the optimum, we show that *all probabilities*

of a given progress-rate are improved. Formally: for any c, $P(\frac{||X_{n+1}||}{||X_n||} < c)$ is greater or equal to its value in the naive case, with only equality in non-standard cases. We have postulated isotropy : this means that the probability of having one point in each given infinitesimal spherical cap is the same in any direction. This is uniformity on the unit sphere. But isotropy does not mean that all the offspring must be independent and identically distributed. We can consider independence and identical distribution (this is the naive usual case), but we can also consider independent non-identically distributed individuals (this is stratification, a.k.a. jittering, and this does not forbid overall uniformity as we will see below) and we can consider non-independently distributed individuals (this is antithetic sampling, and it is also compatible with uniformity).

Some preliminary elements will be necessary for both cases. $(1 + \lambda)$-ES has a population reduced at one individual X_n at epoch n and it generates λ directions randomly on the sphere. Then, for each direction, a step-size determines a point, and the best of these λ points is selected as the new population. Let v a vector toward the optimum (so in the good direction). Let's note γ_i the angle between the i^{th} point and v. We assume that the step size is the distance to the optimum. If $\gamma_i \geq \frac{\pi}{3}$ then the new point will not be better than X_n. Hence, we can consider $\theta_i = \min(\gamma_i, \frac{\pi}{3})$. Let $\theta = min_i \theta_i$. θ is a random variable. As we assume that the step size is the distance to the optimum, the norm of X_{n+1} is exactly $2 * |\sin(\theta/2)|||X_n||$. In the sequel, we note for short $ssin(x) = 2\sin(x/2)||X_n||$; the norm of X_{n+1} is exactly $|ssin(\theta)|$. Then $log(||X_{n+1}||) = \log(|ssin(\min_{i \in [[1,\lambda]]} |\theta_i|)|)$. Therefore, we will have to study this quantity in sections below. For sake of clarity we assume that $||X_n|| = 1$ (without loss of generality).

4.1 Stratification works

Let's consider a stratified sampling instead of a standard random independent sampling of the unit sphere for the choice of directions. We will consider the following simple sampling schema : (1) split the unit sphere in λ regions of same area ; (2) instead of drawing λ points independently uniformly in the sphere, draw 1 point in each of the λ regions. Such a stratification is also called *jittered* sampling (see e.g. [5]). In some cases, we define stratifications according to an auxiliary variable : let $v(.)$ a function (any function, there's no hypothesis on it) from the sphere to $[[0, \lambda - 1]]$. The i^{th} generated point ($i \in [[0, \lambda - 1]]$) is uniformly independently distributed in $v^{-1}(i)$. We note $\pi_k(x)$ the k^{th} coordinate of x : $x = (\pi_0(x), \pi_1(x), \pi_2(x), \ldots, \pi_{d-1}(x))$.

Let's see some examples of stratification :

1. for $\lambda = d$, we can split the unit sphere according to $v(x) = \arg\max_{i \in [[0,d-1]]} |\pi_i(x)|$. We will see below that for a good anticorrelation, this is probably not a very good choice.
2. for $\lambda = 2d$, we can split the unit sphere according to $v(x) = \arg\max_{i \in [[0,2d-1]]} (-1)^i \pi_{\lfloor \frac{i}{2} \rfloor}(x)$.

3. for $\lambda = 2^d$, we can split the unit sphere according to the auxiliary variable $v(x) = (sign(\pi_0(x)), sign(\pi_1(x)), sign(\pi_2(x)), \ldots, sign(\pi_{d-1}(x)))$.
4. for $\lambda = d + 1$, we can also split the unit sphere according to the faces of a regular simplex centered on 0.
5. for $\lambda = 2$, we can split the unit sphere with respect to any hyperplane including 0.
6. for $\lambda = d!$, we can split the unit sphere with respect to the ranking of the d coordinates.
7. for $\lambda = 2^d d!$, we can split the unit sphere with respect to the ranking of the absolute values of the d coordinates and the sign of each coordinate.

However, any stratification in λ parts $S_1, \ldots, S_\lambda$ of equal measure works (and indeed, various other stratifications also do the job). We here consider stratification randomly rotated at each generation (uniformly among rotations) and with each stratum measurable and having non-empty interior.

Theorem 3 (stratification works). *For the sphere function $x \mapsto ||x||^2$ with step size the distance to the optimum, the expected convergence rate* $\exp(E(-\log(||X_{n+1}||/||X_n||)))$ *for $(1 + \lambda)$-ES increases when using stratification.*

Proof : Consider the probability of $|ssin(\theta)| > c$ for some $c > 0$. $P_{naive} = P(|ssin(\theta)| > c) = P(|ssin(\theta_i)| > c)^\lambda$ if naive sampling. Consider the same probability in the case of stratification. $P_{strat} = P(|ssin(\theta)| > c) = \Pi_{i \in [[1,\lambda]]} P(|ssin(\theta_i)| > c)$. where θ_i is drawn in the i^{th} stratum.

Let's introduce some notations. Note P_i the probability that $|ssin(v)| > c$ and that $v \in S_i$, where v is a random unit vector uniformly distributed on the sphere. Note $P(S_i)$ the probability that $v \in S_i$. Then $\Pi_i \frac{P_i}{\sum_j P_j} \leq (1/\lambda)^\lambda$ (by concavity of the logarithm). The equality is only reached if all the P_i are equal.

This implies that $\Pi_i \frac{P_i}{\sum_j P_j} \leq \Pi_i P(S_i)$, what leads to $\Pi_{i \in [[1,\lambda]]} \frac{P_i}{P(S_i)} \leq (\sum_i P_i)^\lambda$. This is exactly $P_{strat} \leq P_{naive}$. This is true for any value of c. Using $E \max(X, 0) = \int_{t \geq 0} P(X > t)$ for any real-valued random variable X, this implies with $X = -\log |ssin(\theta)|$ that $E - \log(|ssin(\theta)|)$ can be worse than naive when using stratification. Indeed, it is strictly better (larger) as soon as the P_i are not all equal for at least one value of c. This is in particular the case for c small, which leads to $P_i < 1$ only for one value of i. □

Remark. We have assumed above that the step size was the distance to the optimum. Indeed, the result is very similar with other step-size-rules, provided that the probability of reaching $||X_{n+1}|| < c$ is not the same for all strata for at least an open set of values of c.

We present in figure 1 experiments on three stratifications (1 to 3 in the list above).

4.2 Antithetic variables work

The principle of antithetic variables is as follows (in the case of k antithetic variables): (1) instead of generating λ individuals, generate only λ/k individuals $x_0, \ldots, x_{\lambda/k-1}$ (assuming that k divides λ); (2) define $x_{i+a\lambda/k}$, for

$a \in [[1, 2, \ldots, k-1]]$, as $x_{i+a\lambda/k} = f_a(x_i)$ where the f_i's are (possibly random) functions. A more restricted but sufficient framework is as follows : choose a fixed set S of λ/k individuals, and choose as set of points $rot_1(S)$, $rot_2(S)$,..., $rot_k(S)$ (of overall size λ) where the r_i are independent uniform rotations in $\mathbb{R}^d$. The limit case $k = 1$ (which is indeed the best one) is defining one set S of λ individuals, and using $rot(S)$ with rot a random rotation.

We first consider here a set S of 3 points on the sphere, which are $(1, 0, 0, \ldots, 0)$, $(\cos(2\pi/3), \sin(2\pi/3), \ldots, 0)$, $(\cos(4\pi/3), \sin(4\pi/3), 0, \ldots, 0)$ (the optimal and natural spherical code for $n = 3$). The angle between two of these points is $2\pi/3$.

Theorem 4 (antithetism works). *For the sphere function $x \mapsto ||x||^2$ with step size equal to the distance to the optimum, the expected convergence rate $\exp(E(-\log(||X_{n+1}||/||X_n||)))$ for $(1 + \lambda)$-ES increases when using antithetic sampling with the spherical code of 3 points.*

Proof : As previously, without loss of generality we can assume $||X_n|| = 1$. We consider $\exp(E(-\log(||X_{n+1}||)))$. As above, we show that for any c,

$$P(||X_1|| > c \text{ with antithetic variables}) \leq P(||X_{n+1}|| > c) \tag{12}$$

Using $E \max(x, 0) = \int_{t \geq 0} P(x \geq t)$, this is sufficient for the expected result. The inequality on expectations is strict as soon as it is strict in a neighborhood of some c. The probability $P(||X_{n+1}|| > c)$, in both cases, antithetic variables or not, is by independence the power $\frac{\lambda}{3}$ of the result for $\lambda = 3$. Therefore, it is sufficient to show the result for $\lambda = 3$. Yet another reduction holds on c: $c > 1$ always leads to a probability 0 as the step-size will be 0 if the direction does not permit improvement. Therefore, we can restrict our attention to $c < 1$.

So, we have to prove equation 12 in the case $c < 1$, $\lambda = 3$. In the antithetic case the candidates for X_{n+1} are $X_n + y_i$ where $y_0 = rot(x_0), y_1 = rot(x_1), y_2 = rot(x_2)$. In the naive case these candidates y_0, y_1, y_2 are randomly drawn in the sphere. We note $\gamma = \min(|angle(-y_i, X_n)|)$ (the y_i realizing this minimum verifies $X_{n+1} = X_n + y_i$ if $||X_n + y_i|| < ||X_n||$). Let θ the angle such that $\gamma \leq \theta \Rightarrow ||X_{n+1}|| < c$

In the antithetic case the the spherical caps s_i located at $-y_i$, and of angle θ are disjoint because $c < 1$ so $\theta < \frac{\pi}{3}$. But in the naive one they can overlap with non zero probability. As $P(||X_{n+1}|| < c) = P(X_n \in \cup_i s_i)$, this shows equation 12, which concludes the proof. □

The proof can be extended to show that $k = 1$ leads to a better convergence rate than $k > 1$, at least if we consider the optimal set S of λ points. But we unfortunately not succeeded in showing the same results for explicit larger numbers of antithetic variables in this framework. We only conjecture that randomly drawing rotations of explicit good spherical codes ([6]) on the sphere leads to similar results. However, we proved the following

Theorem 5 (arbitrarily large good antithetic variables exist). For any $\lambda \geq 2$, there exists a finite subset s of the unit sphere in $\mathbb{R}^d$ with cardinal λ such that the convergence rate of $(1 + \lambda)$-ES is faster with a sampling by random permutation of s than with uniform independent identically distributed sampling, with step size equal to the distance to the optimum.

Proof: We consider the sphere problem with optimum in zero and X_n of norm 1.

Let s a sample of λ random points (uniform, independent) on the unit sphere. Let $f(s) = E_{rot}(\ln ||X_{n+1}||)$ (as above rot is a random linear transformation with $rot \times rot' = 1$). If s is reduced to a single element we reach a maximum for f (as the probability of $\ln(X_{n+1}) < c$ is lower than for any set with at least two points).

$f(s)$ is therefore a continuous function, with some values larger than $E_s f(s)$. Therefore, the variance of $f(s)$ is non-zero. Therefore, thanks to this non-zero variance, there exists s' such that $f(s') < E_s f(s)$.

$E_s f(s)$ is the progress rate when using naive sampling and $f(s')$ is the progress rate when using an antithetic sampling by rotation of s'. So, this precisely means that there exists good values of s leading to an antithetic sampling that works better than the naive approach. □

We have stated the result for $(1 + \lambda)$-ES with λ antithetic variables, but the same holds for λ/k antithetic variables with the same proof. This does not explicitly provided a set s', but it provides a way of optimizing it by numerical optimization of $E \ln(X_{n+1})$ that can be optimized once for all for any fixed value of λ. Despite the lack of theoretical proof, we of course conjecture that standard spherical codes are a good solution. This will be verified in experiments (figure 1, plots 4,5,6). However, we see that it works in simulations for moderate numbers of antithetic variables placed according to standard spherical codes. But for $k = 2^d$ antithetic variables at the vertices of an hypercube, it does not work when dimension increases, i.e. hypercube sampling is not a good sampling. Note that the spherical codes $\lambda = 2d$ (generalized octahedron, also termed biorthogonal spherical code) and $\lambda = d + 1$ (simplex), which are nice and optimal for various points of view, seem to scale with dimension. Their benefit in terms of the reduction of the number of function evaluations behaves well when d increases. Of course, more experimental works remain to be done.

5 Conclusion

We have shown that (i) superlinear methods require a fine decision about the stepsize, with at most a very little randomization; (ii) if we accept linear convergence rates and keep the randomization of the step size, we however need, in order to break the curse of dimensionality (i.e. keeping a convergence rate far from 1), a fine decision about the direction, with at most a very little randomization. This shows the price of isotropy, which is only a choice when less randomized techniques can not work. In a second part, we have shown that isotropy can be improved; the naive isotropic method can be very easily replaced by a non i.i.d sampling, thanks to stratification (jittering) or antithetic variables. Moreover, it really works on experiments.

The main limit of this work is its restriction to isotropic methods. A second limit is that we have considered the second order of sampling *inside* each epoch, but not *between* successive epochs. In particular, Gauss-Seidel or generalized

versions of Gauss-Seidel ([14, 15]) are not concerned; we have not considered correlations between directions chosen at successive epochs; for example, it would be natural, at epoch $n+1$, to have directions orthogonal to, or very different from, the chosen direction at epoch n. This is beyond the simple framework here, in particular because of the optimal step size, and will be the subject of a further work.

The restriction to 3 antithetic variables in theorem 4 simplifies the theorem; this hypothesis should be relaxed in a future work. Theorem 5 shows that good point sets exist for any number of antithetic variables, theorem 4 explicitly exhibits 3 antithetic variables that work and that are equal to the optimal spherical code for $n = 3$, but figure 1 (figs. 4,5,6) suggests that more generally octahedron-sampling or simplex-sampling (which are very good spherical codes, see e.g. [6]) are very efficient, and in particular that the improvement remains strong when dimension increases. Are spherical codes ([6]) the best choice, as intuition suggests, and are there significant improvements for a number $n = \lambda/k$ of antithetic variables large in front of d ? This is directly related to the speed-up of parallelization.

Acknowledgements

This work was supported in part by the Pascal Network of Excellence. We thank A. Auger and J. Jägersküpper for fruitful discussions.

References

1. A. Auger. Convergence results for (1,λ)-SA-ES using the theory of φ-irreducible markov chains. *Theoretical Computer Science*, 2005. in press.
2. A. Auger, M. Jebalia, and O. Teytaud. Xse: quasi-random mutations for evolution strategies. In *Proceedings of Evolutionary Algorithms, 12 pages*, 2005.
3. H.-G. Beyer. *The Theory of Evolutions Strategies*. Springer, Heidelberg, 2001.
4. C. G. Broyden. The convergence of a class of double-rank minimization algorithms 2, the new algorithm. j. of the inst. for math. and applications, 6:222-231, 1970.
5. B. Chazelle. *The discrepancy method: randomness and complexity*. Cambridge University Press, New York, NY, USA, 2000.
6. J. H. Conway and N. J. Sloane. *Sphere packings, lattices and groups*. 1998.
7. R. Fletcher. A new approach to variable-metric algorithms. computer journal, 13:317-322, 1970.
8. G. Frahm and M. Junker. Generalized elliptical distributions: Models and estimation. Technical Report 0037, 2003.
9. D. Goldfarb. A family of variable-metric algorithms derived by variational means. mathematics of computation, 24:23-26, 1970.
10. U. Haagerup. The best constants in the khintchine inequality. *Studia Math.*, 70:231–283, 1982.
11. N. Hansen and A. Ostermeier. Completely derandomized self-adaptation in evolution strategies. *Evolutionary Computation*, 11(1), 2003.
12. J. Jagerskupper. In between progress rate and stochastic convergence. *Dagstuhl's seminar*, 2006.

13. Literka. A remarkable monotonic property of the gamma function. Technical report, 2005.
14. R. Salomon. Resampling and its avoidance in genetic algorithms. In V. W. Porto, N. Saravanan, D. Waagen, and A. E. Eiben, editors, *Evolutionary Programming VII*, pages 335–344, Berlin, 1998. Springer.
15. R. Salomon. The deterministic genetic algorithm: Implementation details and some results, 1999.
16. D. F. Shanno. Conditioning of quasi-newton methods for function minimization. mathematics of computation, 24:647-656, 1970.
17. O. Teytaud and S. Gelly. General lower bounds for evolutionary algorithms, ppsn 2006.
18. Z. Wang, K. Droegemeier, L. White, and I. M. Navon. Application of a new adjoint newton algorithm to the 3-d arps storm scale model using simulated data. *Monthly Weather Review, 125, No. 10, 2460-2478*, 1997.
19. C. Witt and J. Jägersküpper. Rigorous runtime analysis of a (mu+1) es for the sphere function. In *Proceedings of the 2005 Conference on Genetic and Evolutionary Computation*, pages 849–856, 2005.

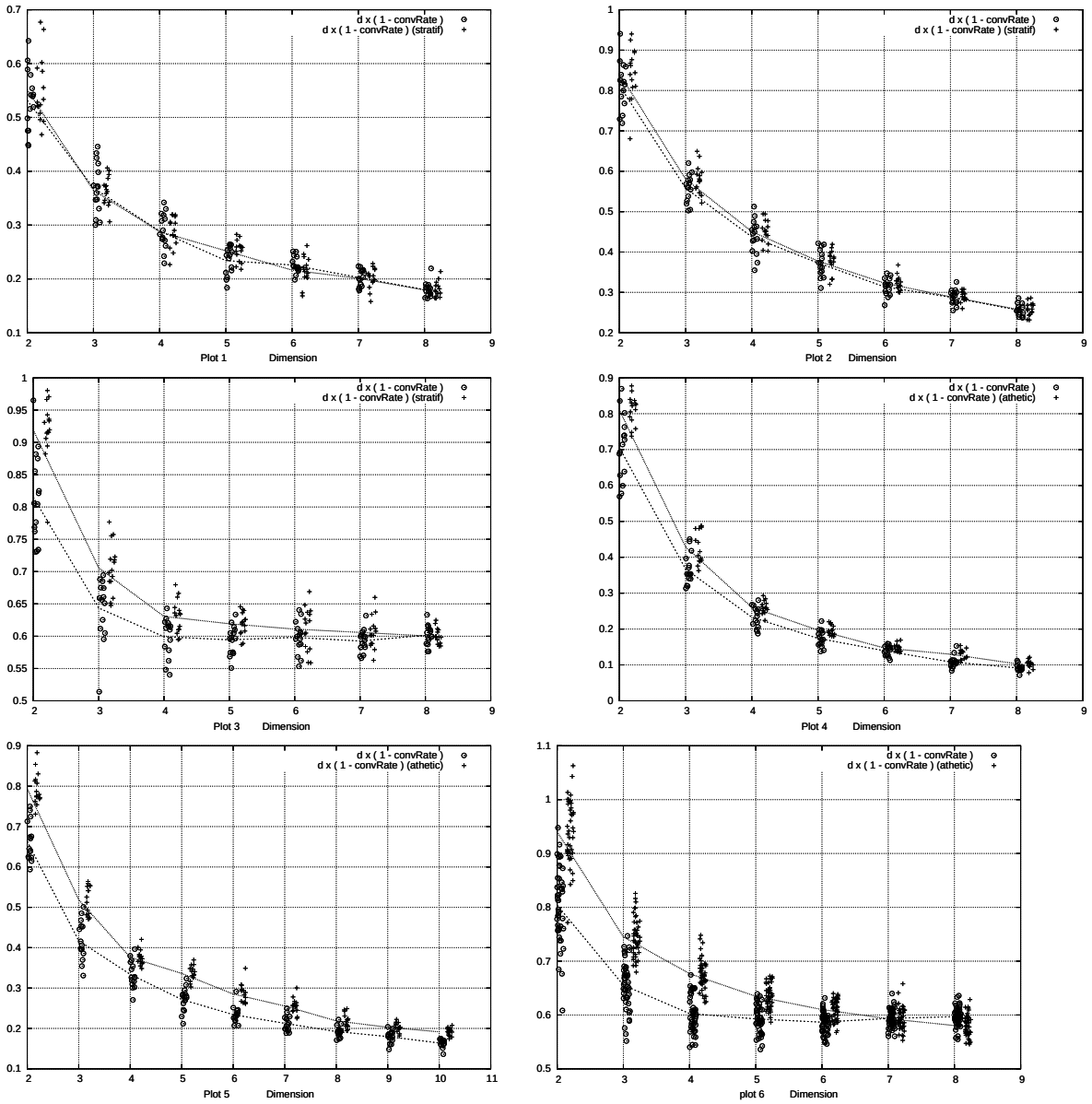


Fig. 1. *Antithetic variables look better.* Plots 1,2,3: with ρ the average progress rate $\sqrt[n\lambda]{||X_n||/||X_0||}$ on the sphere, we plot $d(1-\rho)$ in two cases (i) independent uniform sampling (ii) stratified sampling. Each point corresponds to one run. $n = 100$ for each run. The step size is equal to the optimal one. The three plots respectively deal with $\lambda = d$, $\lambda = 2d$ and $\lambda = 2^d$. The improvement in terms of number of fitness-evaluations is the ratio between the log(.) of the convergence rates. For dimension 2, the difference in terms of number of function-evaluations is close to 20 % but quickly decreases. Plots 4,5,6: with ρ the average progress rate $\sqrt[n\lambda]{||X_n||/||X_0||}$ on the sphere, we plot $d(1-\rho)$ in two cases (i) independent uniform sampling (ii) antithetic sampling with $\lambda = 3$ (plot 4) or $\lambda = d$ with an antithetic sampling by random rotation of a regular simplex (plot 5) or $\lambda = 2^d$ with an antithetic sampling by random rotation of $\{-1, 1\}^d$ (plot 6). n and the step size are as for previous plots. For dimension 2 to 6, the difference in terms of number of function-evaluations for a given precision are between 12 % and 18 % for $\lambda = 2^d$ and remain close to 20 % for the octahedron $\lambda = 2d$ for any value of d. We also experiments the direct inclusion of quasi-random numbers in the Covariance-Matrix-Adaptation algorithm ([11]); the resulting algorithm, termed DCMA, in which the only difference with CMA is that the random-generator is replaced by a quasi-random generator, is more stable and faster than the classical CMA; results are presented in http://www.lri.fr/~ teytaud/resultsDCMA.pdf.