A Parameter Study for Differential Evolution

ROGER GÄMPERLE, SIBYLLE D. MÜLLER, PETROS KOUMOUTSAKOS
Institute of Computational Sciences, Department of Computer Science
Swiss Federal Institute of Technology Zürich (ETHZ)
CH - 8092 Zürich
SWITZERLAND

Abstract: - We assess the selection of strategy parameters for Differential Evolution on a set of test problems. The original algorithm is analyzed with respect to its performance depending on the choice of strategy parameters. Although empirical rules are provided in the literature [1], choosing the proper strategy parameters for Differential Evolution is more difficult than expected.

Key-Words: - Differential evolution, optimization, strategy parameters.

1 Introduction

Differential evolution (DE) belongs to the class of stochastic optimization algorithms which address the following search problem: Minimize an objective function which is a mapping from a parameter vector $\boldsymbol{x} \in \mathcal{R}^n$ ro $\mathcal{R}$. DE is characterized by self-organization, mutation, crossover, selection, and empirical selection of strategy parameters [1].

As stated in [1], strategy parameters for DE are not difficult to choose. However, the rules for choosing control parameters given in this reference are not general and therefore not suitable for practical applications. We try to devise more general rules by studying the performance of the strategy on a set of test problems. The sphere and Rosenbrock functions representing unimodal problems as well as the Rastrigin and modified Rosenbrock functions representing multimodal problems are tested.

Although there are only three strategy parameters, the following application of DE on several test functions shows that finding the global optimum is sensitive to the choice of these control parameters.

2 The DE Algorithm

One set of D optimization parameters, called an individual, is represented by a D-dimensional vector. A population consists of NP D-dimensional parameter vectors $x_{i,G},\ i = 1, 2, \ldots, NP$ for each generation G.

The three operations mutation, crossover, and selection are described in the following, see also [1] for a more detailed description:

Mutation:
For each target vector $x_{i,G}$, a mutant vector is generated according to $v_{i,G+1} = x_{r1,G} + F \cdot (x_{r2,G} - x_{r3,G})$with randomly chosen indexes $r_1, r_2, r_3 \in \{1, 2, \ldots, NP\}$. Note that indexes have to be different from each other and from the running index i. Therefore, the number of parameter vectors in a population must be at least four. F is a real and constant factor $\in [0, 2]$ that controls the amplification of the difference vector $(x_{r2,G} - x_{r3,G})$.

Note that the smaller the differences between parameters of parent r_2 and r_3, the smaller the difference vector and therefore the perturbation. That means if the population gets close to the optimum, the step length is automatically decreased. This is similar to the automatic step size control found in standard evolution strategies.

Crossover:
The target vector is mixed with the mutated vector using the following scheme to yield the trial vector $u_{i,G+1} = (u_{1i,G+1}, u_{2i,G+1},\ \ldots\ , u_{Di,G+1})$ where

$$u_{ji,G+1} = \begin{cases} v_{ji,G+1} & \text{if } (r(j) \leq CR) \text{ or } j = rn(i) \\ x_{ji,G} & \text{if } (r(j) > CR) \text{ and } j \neq rn(i) \end{cases}$$

for $j = 1, 2, \ldots, D$. $r(j) \in [0, 1]$ is the j^{th} evaluation of a uniform random number generator. CR is the crossover constant $\in [0, 1]$. $CR = 0$ means no crossover. $rn(i) \in (1, 2, \ldots, D)$ is a randomly chosen index which ensures that $u_{i,G+1}$ gets at least one element from $v_{i,G+1}$. Otherwise no new parent vector would be produced and the population would not alter.

Selection:
A "greedy" selection scheme is used: If and only if the trial vector yields a better cost function value compared to the parameter vector $x_{i,G}$, is it accepted as a new par-

ent vector for the following generation $G+1$. Otherwise, the target vector is retained to serve as a parent vector for generation $G+1$ once again.

There are three strategy parameters altogether: NP: Number of members in a population, F: Amplification factor of the difference vector, CR: Crossover constant.

Other Variants of DE:

There are several variants of DE which can be classified using the notation [1] $DE/x/y/z$ where x specifies the vector to be mutated, y is the number of difference vectors used, and z denotes the crossover scheme.

x can be 'rand' (randomly chosen population vector) or 'best' (the best vector from the current population). Since we use only one difference vector, y is one in the described scheme. The current variant for z is 'bin' which means crossover due to independent binomial experiments. Using this notation, the basic DE-strategy can be written as: $DE/rand/1/bin$. Another possibility is the method $DE/best/2/bin$, where we have the following mutant vector $v_{i,G+1} = x_{best,G} + F \cdot (x_{r1,G} - x_{r2,G} + x_{r3,G} - x_{r4,G})$.

3 Performance on Test Functions

For each test at least 20 runs were made, and a maximum of $G_{max} = 10^4$ generations were allowed unless otherwise stated. For all functions, the termination criterion is $|f_{opt} - f_{act}| < \epsilon$ where f_{opt} is the optimum function value, f_{act} is the actual function value, and ϵ is the precision value. The *I*nitial *P*arameter *R*ange *IPR* and a termination criterion were defined. Initial parameter values were drawn randomly from the *IPR*. For an initial normal distribution with variance σ^2 and expectation value ξ, we write $IPR_\sigma = [\xi \pm \sigma]$. For a uniform distribution within the limits $-\xi$ and ξ, we write $IPR = [-\xi; \xi]$.

3.1 Sphere Function

The D-dimensional sphere is defined as $f_{sphere}(\boldsymbol{x}) = \sum_{i=1}^{D} (x_i - 1)^2$.

2-dimensional Sphere:

We set $\epsilon = 10^{-10}$ and $IPR_\sigma = [-5 \pm 1]$. Note that the IPR_σ is defined such that it did not center around the minimum $(1; 1)$.

For the first test, we set $F = 0.4$, $CR = 0.5$, and $NP = 15$. The result was that the vectors did not reach the minimum but got stuck on their way to the minimum, meaning that they got closer to each other and the difference vector for the perturbation decreased. Note that if we choose an IPR that centers around the minimum, the vectors do not get stuck, since in this case the population approaches the minimum from all directions and does not converge before arriving at the minimum.

To avoid premature "convergence", the step size needs to be increased. Therefore, in the next test a higher amplification factor, $F = 0.6$, was chosen. With other settings unchanged, the minimum was found in all test runs.

5-dimensional Sphere:

CR was set to a relatively high value, $CR = 0.9$. That means that on the average 90% of the elements of the trial vector were identical to those of the mutant vector, which implied a high diversity. The other parameters were chosen as follows: $F = 0.6$, $NP = 15$, $\epsilon = 10^{-10}$, $IPR_\sigma = [0 \pm 1]$. The minimum was not found with these settings. It seems that due to the high crossover constant the path length increased without a significantly higher speed to approach the minimum. With each generation the individuals got closer to each other and converged before they reached the minimum.

On the other hand, if CR is chosen too small, DE needs more generations to find the minimum or it might even not find the minimum at all. If, for instance, the crossover parameter is set to $CR = 0$, only one element of a vector is altered. This means that a population member can move only parallel to the axes. The restriction to certain moving directions decreases the convergence speed. Moreover, there are functions where the minimum cannot be reached with only vertical or horizontal steps.

From the first test runs for the 2-dimensional sphere it became clear that F must not be smaller than a certain value in order to find the minimum. The following tests showed that this value highly depends on the cost function. Hence, no exact rule for the lower bound of F can be given for each function, but in most cases it should be at least larger than $F = 0.4$. Nevertheless, it is interesting how the number of function evaluations changes if the amplification factor F is increased. The strategy parameters were set as follows: $\epsilon = 10^{-10}$, $NP = 15$, and $CR = 0.5$, $IPR_\sigma = [0 \pm 1]$. A smaller amplification factor results in a smaller number of function evaluations $\overline{nfe}$ and in an increasing risk not to find the minimum.

Next, we study the influence of the crossover constant CR on $\overline{nfe}$. We already know that, at least for the sphere, a too high crossover constant has the effect that the minimum is not found. The same settings for

the 5-dimensional sphere were used as before but with $F = 0.6$ and CR between 0.0 and 0.85 in steps of 0.05. In this case the solution was found even if CR is set to zero. On the average fewer function evaluations were used for a higher crossover constant. But again, the risk that the minimum is not found increases for a larger CR.

The dependence of $\overline{nfe}$ on the strategy parameter NP was tested using again the 5-dimensional sphere cost function with the same settings as above, except $CR = 0.4$, and $F = 0.6$. This time we also compare the behavior of DE/best/2/bin with the behavior of DE/rand/1/bin. $\overline{nfe}$ increases linearly as a function of NP and the number of function evaluations for one member of a population remains constant (figure omitted due to space limitations). Nevertheless, the total $\overline{nfe}$ increases due to NP. Note that if there are only 5 vectors the minimum is not found in every test run. It seems that in this case the diversity of the population is too small.

Comparing to DE/rand/1/bin, DE/best/2/bin needs almost the same number of function evaluations to find the minimum except for a small population where DE/best/2/bin converged slower than DE/rand/1/bin. However, in this case it still finds the minimum in 95% of the runs whereas DE/rand/1/bin was successful in only 25%. Since the perturbation vector of DE/best/2/bin is on the average larger than for DE/rand/1/bin, a smaller amplification factor yields better results with DE/best/2/bin even for larger populations.

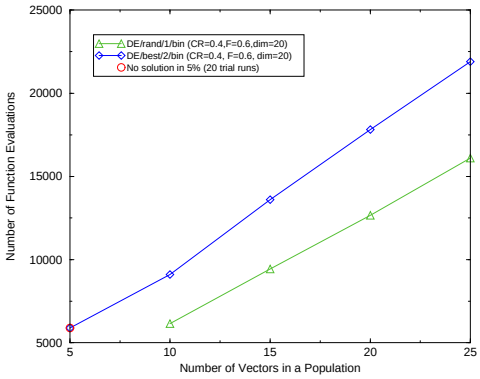

Figure 1: 20-dimensional sphere: $\overline{nfe}$ as a function of the parameter NP. Note that for $NP = 5$, DE/rand/1/bin does not find the minimum.

20-dimensional Sphere:
Finally the 20-dimensional sphere cost function was used for several test runs with varying strategy parameter NP. Fig. 1 shows that DE/best/2/bin takes many more function evaluations than DE/rand/1/bin for the same size of a population. However, DE/best/2/bin proves to be the better variant for convergence using a small population size. DE/rand/1/bin did not find the global minimum in all 20 test runs for $NP = 5$ whereas DE/best/2/bin found it with probability 95%.

Summary for the Sphere function:
Tab. 1 shows the best results for the sphere function. The amplification factor and the crossover constant for the best solutions did not change significantly if we increased the dimension or the size of the population. This means that we can keep the values of F and CR for different settings of D and NP.

We conclude for the sphere function: The amplification factor F should not be smaller than a certain value to avoid that the population converges before arriving at the minimum. On the other hand, the amplification factor F should not be chosen too large because the number of function evaluations increases as F increases. The crossover constant CR should not be too large to avoid that the perturbations get too high and the convergence speed decreases. However, a small CR decreases diversity and might cause the strategy to get stuck. For the same size of the population DE/best/2/bin and DE/rand/1/bin perform similarly. However, for a small population it is more likely to find the minimum using DE/best/2/bin instead of DE/rand/1/bin due to the improved diversity of the trial vectors. The best results were achieved with DE/best/2/bin.

DE-strategy	D	NP	CR	F	$\overline{nfe}$
best/1/bin	2	8	0.4	0.45	306 ± 46
best/1/bin	5	8	0.4	0.45	834 ± 235
best/2/bin	20	10	0.4	0.45	4634 ± 639

Table 1: Best results for the sphere function.

3.2 Rosenbrock's Function

The D-dimensional Rosenbrock's function is defined as $f_{rosen}(\boldsymbol{x}) = \sum_{i=1}^{D-1}[100 \cdot (x_i^2 - x_{i+1})^2 + (x_i - 1)^2]$

2-dimensional Rosenbrock's function:
An $IPR = [-10; 10]$ was used and after some trials we fixed the strategy parameters to $F = 0.9$, $CR = 0.9$, and $NP = 15$.

Test runs for the 2-dimensional Rosenbrock function and a uniform distribution in the region $[-2 <$

$x < -1; 1 < y < 2]$ revealed that DE/rand/1/bin performs worse than DE/best/2/bin. The parameters were set to $NP = 15$, $F = 0.9$, $CR = 0.9$. Note that since the perturbation vector for DE/best/2/bin consists of two difference vectors the step size is larger than for DE/rand/1/bin. For this reason one can expect that $\overline{nfe}$ is lower when choosing a smaller amplification factor F. And indeed, $\overline{nfe}$ gets smaller with $F = 0.6$ in our tests.

For all the following test runs, $IPR_\sigma = [0 \pm 0.1]$ and $\epsilon = 10^{-10}$.

5-dimensional Rosenbrock's function:
For the 5-dimensional Rosenbrock function, we used $NP = 15$, $CR = 0.9$, and F was varied. With increasing F the mean number of function evaluations became larger. For $F \leq 0.7$ the global minimum was not found (at least not within $G_{max} \cdot NP = 150\,000$ function evaluations). There is the same tendency as for the sphere. However, the lower bound for F was at $F = 0.7$, compared with $F = 0.45$ for the sphere. This shows the difficulty when choosing F for a real problem. We do not want a high amplification factor since this means a high number of function evaluations. On the other hand, if we choose a small F the risk increases that the minimum is not found at all.

Next, CR is varied and the amplification factor is set to $F = 0.9$. $\overline{nfe}$ decreases for growing CR. If CR is large the diversity of the population is relatively high. Here, a high diversity improved the convergence speed. Note that for this test function the global minimum was not found for $CR = 0$ since for Rosenbrock's function it is not possible to find the global minimum with only steps that are parallel to the axes. As opposed to the sphere function, $CR = 1$ yields a solution.

We compare DE/rand/1/bin with DE/best/2/bin using $F = 0.9$ and $CR = 0.9$. The results were similar to the ones for the sphere function. Here, DE/best/2/bin is worse if the population size is high (about three or more times D). For a small population (about one to two times D) DE/best/2/bin is better than DE/rand/1/bin. Since for DE/best/2/bin the perturbation vector consists of two difference vectors, the diversity is higher compared to DE/rand/1/bin. Hence, we get about the same diversity like for DE/rand/1/bin with a smaller population size. The smaller the population size, the smaller $\overline{nfe}$ since for each member of a population and for each generation a function evaluation is needed.

Summary for Rosenbrock's function:
Tab. 2 shows the best results for Rosenbrock's function. Note that all of them were achieved with DE/best/2/bin. The results can be summarized as follows: For Rosenbrock's function the behavior of $\overline{nfe}$ as a function of F, CR, and NP is similar to the behavior for the sphere function. However, the values of the parameters are not the same, e.g. the amplification factor must not be lower than about $F = 0.7$ for Rosenbrock's function whereas for the sphere function $F = 0.45$ is still possible.

DE-strategy	D	NP	CR	F	$\overline{nfe}$
best/2/bin	2	10	0.9	0.6	627 ± 80
best/2/bin	5	10	0.9	0.6	3496 ± 761
best/2/bin	20	15	0.9	0.6	111961 ± 22677

Table 2: Best results for Rosenbrock's function.

3.3 Rastrigin's Function

The multimodal D-dimensional Rastrigin function

$$f_{rast}(\boldsymbol{x}) = (D \cdot 10) + \left[\sum_{i=1}^{D} \left(x_i^2 - 10\cos(2\pi x_i)\right)\right]$$

is tested using $IPR = [-600; 600]$ and $\epsilon = 10^{-6}$.

In the first test run with $CR = 0.5$, $NP = 15$, and $F = 0.5$, the global minimum was always found and the average number of function evaluations was $\overline{nfe} = 938 \pm 70$ when using DE/best/2/bin. For the same settings DE/rand/1/bin found the global optimum only in 95% of the runs with a $\overline{nfe} = 1179 \pm 91$. As long as the step size is high enough (larger than the smallest distance between two adjoining local minima) it is possible to advance to the global minimum. If the step size was too small the population could not escape a local minimum.

For higher values of F the chance of finding the global minimum increases since due to the larger perturbation vector the population is able to escape local minima. If we consider the number of the successful test runs, a large CR does not differ much from a small CR. For a small F (e.g. $F = 0.3$), the best results were achieved with DE/best/2/bin and the worst results with a large CR. The most important steering parameter turned out to be the size of a population. If we choose $NP = 40$ the global minimum was found in almost every test run. Even for relatively small amplification factors the global minimum was still found with large populations.

3.4 Modified Rosenbrock Function

The optimization of the multimodal 2-dimensional modified Rosenbrock function

$$f_{modros}(\boldsymbol{x}) = 100.0 \cdot (x_2 - x_1^2)^2 + (1.0 - x_1)^2 + a(\boldsymbol{x})$$

where

$$a(\boldsymbol{x}) = 74 - 400 \cdot exp\Big(10(-(x_1+1.0)^2 - (x_2+1.0)^2)\Big)$$

is initialized with $IPR = [-2; 2]$ and it is $\epsilon = 10^{-6}$. A smaller size of the population is worse with respect to the global optimization behavior. However, a large population increases $\overline{nfe}$. An amplification factor $F = 0.9$ performed better than $F = 0.5$. DE/best/2/bin was better than DE/rand/1/bin. If just one member of the population gets to the area where the cost values are below the cost value for the local minimum it will become the best vector. For DE/best/2/bin the perturbation vector is added to the best vector of the current population. This means that the population has the tendency to move to the global minimum.

4 Choice of Strategy Parameters

However, the application of DE on several test functions showed that the capability of finding the global minimum and a fast convergence rate are very sensitive to the choice of the control variables NP, F, and CR. Some rules of thumb for their choice are given below.

Population Size NP:
According to our experience a reasonable choice for the population size is between $NP = 3 \cdot D$ and $NP = 8 \cdot D$. Note, that NP must be at least 4 for DE/rand/1/bin and 5 for DE/best/2/bin respectively to ensure that DE will have enough mutually different vectors. The larger the population, the larger the probability of finding the global minimum for multimodal functions. A larger population implies a larger number of function evaluations, however.

Amplification Factor F:
F should not be smaller than a certain value to prevent premature convergence. This value depends on the cost function and on the other strategy parameters, e.g. for the sphere function $F = 0.5$ ($CR = 0.5$, $NP = 15$) and for Rosenbrock's function $F = 0.75$ ($CR = 0.9$, $NP = 15$). A larger F increases the probability for escaping a local optimum (see Fig. 2). However, for $F > 1$ the convergence speed decreases. It is more difficult to converge for a population when the perturbation is larger than the distance between two members. A good initial choice for the amplification factor is $F = 0.6$. If one suspects that with this setting only a local optimum is found, then F should be increased.

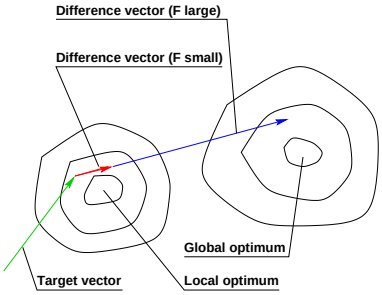

Figure 2: If F is chosen too small it gets more difficult to escape local optima.

Crossover Constant CR:
A large CR often speeds up convergence. However, from a certain value upwards the convergence speed may decrease or the population may converge prematurely. This value depends on the cost function and is located in the region $CR = 0.9 \ldots 1.0$. A good choice for the crossover constant is a value between $CR = 0.3$ and $CR = 0.9$.

DE Variants:
DE/best/2/bin seems to be better than DE/rand/1/bin with respect to both convergence speed and global optimization. For DE/best/2/bin the perturbation vector is added to the best vector of the current population. This means that the population has the tendency to move to the local minimum. When the members of the population converge to a local minimum the difference vectors decrease and so does the chance to escape a local minimum.

Since for DE/best/2/bin two difference vectors are added to the target vector, the amplification factor F should be generally smaller than for DE/rand/1/bin. If we have a population of six individuals, 360 different perturbation vectors are possible for DE/best/2/bin, but only 30 for DE/rand/1/bin. This means that with DE/best/2/bin a lot more different trial vectors can be generated as with DE/rand/1/bin. In this way a larger parameter range can be covered from generation to generation.

5 Comparison with other Evolution Strategies

DE is compared with an evolution strategy with Covariance Matrix Adaptation (CMA-ES) [2] and without CMA (ES). See [2] for the strategy parameters of the two evolution algorithms. The precision value was ($\epsilon = 10^{-10}$) and results are summarized in Table 3.

	DE	ES	CMA-ES
Sph 5D	$8.3 \cdot 10^2 \pm 2.3 \cdot 10^2$	$8 \cdot 10^2$	$7 \cdot 10^2$
Sph 20D	$4.6 \cdot 10^3 \pm 6.4 \cdot 10^2$	$3 \cdot 10^3$	$3 \cdot 10^3$
Ros 5D	$3.5 \cdot 10^3 \pm 0.8 \cdot 10^3$	$2 \cdot 10^5$	$3 \cdot 10^3$
Ros 20D	$1.1 \cdot 10^5 \pm 0.23 \cdot 10^5$	10^6	$3 \cdot 10^4$

Table 3: Comparison of different optimization strategies on unimodal test functions.

The start point for ES and CMA-ES was $x = 0$. For the sphere function, $IPR_\sigma = [0 \pm 1]$. and for Rosenbrock's function, $IPR_\sigma = [0 \pm 0.1]$.

The results show that the Differential Evolution strategy performs similar to the CMA-ES and ES for the sphere case. However, DE proved to be the better variant compared to the simple ES for Rosenbrock's function.

The 2-dimensional modified Rosenbrock function and Rastrigin's function serve us to compare the frequency of finding the global minimum of multimodal functions. Fifty test runs were done with DE. The settings of the strategy parameters for DE, ES, CMA-ES, and Random search were as follows:

- *DE for Rosenbrock's function*: $G_{max} = 3000$, $IPR = [-2; 2]$, $CR = 0.9$, $F = 0.9$, and $NP = 20$.
- *DE for Rastrigin's function*: $G_{max} = 3000$, $IPR = [-600; 600]$, $CR = 0.6$, $F = 0.4$, and $NP = 15$.
- *ES/CMA-ES*: $\mu = 2$ parents, $\lambda = 10$ children, $50\,000$ function evaluations, and initial global step size $\delta = 10^{-2}$.
- *Random search*: 50000 function evaluations.

DE with a precision $\epsilon = 10^{-6}$ needed only 938 ± 70 function evaluations to reach the goal for Rastrigin's function and 2292 ± 293 for Rosenbrock's function respectively, as shown in Table 4.

DE performed much better than the other strategies with regard to the global optimization performance.

	Rastrigin 2D	Mod. Ros. 2D
DE/best/2/bin	100%	50 %
ES	37 %	10%
CMA-ES	33 %	7%
Random search	0.1%	100%

Table 4: Comparison of different optimization strategies with respect to the global optimization performance.

However, we should note that these good results were only achieved by trying several different settings for the strategy parameters. The convergence speed and the global optimization behavior depended a lot on the parameter settings.

6 Summary and Conclusions

To find appropriate strategy parameters, DE was applied to several unimodal and multimodal test functions. DE's global optimization performance and convergence speed compare well with ES and CMA-ES when good strategy parameters are chosen. However, it turned out that the performance of DE is very sensitive to the choice of the strategy parameters.

Bibliography

[1] Storn, R., "Differential Evolution - A Simple and Efficient Heuristic Strategy for Global Optimization over Continuous Spaces". *Journal of Global Optimization*, vol. 11, Dordrecht, pp. 341-359, 1997.

[2] Hansen, N., Ostermeier, A., "Convergence Properties of Evolution Strategies with the Derandomized Covariance Matrix Adaptation: The $(\mu/\mu_I, \lambda)$-CMA-ES," *Proceedings of the 5th European Congress on Intelligent Techniques and Soft Computing (EUFIT'97)*, pp. 650-654, 1997.