Nikolaus Hansen and Stefan Kern

Evaluating the CMA Evolution Strategy on Multimodal Test Functions

Parallel Problem Solving from Nature, PPSN 2004

ERRATA:

Section 3.1. In "Additional bounds are implemented [. . .] by adding a quadratic penalty term. E.g., $f_{\text{Schwefel}}(\boldsymbol{x})+10^4\cdot\sum_{i=1}^{n}\theta(|x_i|-500)x_i^2$ is minimized, where $\theta(.)$ is the Heaviside function." the formula must read "$\ldots f_{\text{Schwefel}}(\boldsymbol{x}) + 10^4 \cdot \sum_{i=1}^{n} \theta(|x_i| - 500)^2$ is minimized. . . ".

Section 3.2. "The starting point $\boldsymbol{x}^{(0)}$ is sampled uniformly. . . " must read "The starting point $\langle\boldsymbol{x}\rangle_{\mathrm{w}}^{(0)}$ is sampled uniformly. . . "

Evaluating the CMA Evolution Strategy on Multimodal Test Functions

Nikolaus Hansen and Stefan Kern

Computational Science and Engineering Laboratory (CSE Lab),
Swiss Federal Institute of Technology (ETH) Zurich, Switzerland
{nikolaus.hansen, skern}@inf.ethz.ch,
http://www.icos.ethz.ch/cse/

Abstract. In this paper the performance of the CMA evolution strategy with rank-μ-update and weighted recombination is empirically investigated on eight multimodal test functions. In particular the effect of the population size λ on the performance is investigated. Increasing the population size remarkably improves the performance on six of the eight test functions. The optimal population size takes a wide range of values, but, with one exception, scales sub-linearly with the problem dimension. The global optimum can be located in all but one function. The performance for locating the global optimum scales between linear and cubic with the problem dimension. In a comparison to state-of-the-art global search strategies the CMA evolution strategy achieves superior performance on multimodal, non-separable test functions without intricate parameter tuning.

1 Introduction

The derandomized Evolution Strategy (ES) with Covariance Matrix Adaptation (CMA) [1] adapts the complete covariance matrix of the normal mutation (search) distribution. The CMA-ES exhibits several invariances. Hereunder are (a) invariance against order preserving (i.e. strictly monotonic) transformations of the objective function value; (b) invariance against angle preserving transformations of the search space (rotation, reflection, and translation) if the initial search point is transformed accordingly; (c) scale invariance if the initial scaling is chosen accordingly. Invariances are highly desirable: they imply uniform behavior on classes of functions and therefore generalizability of empirical results.

Originally designed for small population sizes, the CMA-ES was interpreted as a robust local search strategy [2]. It efficiently minimizes unimodal test functions [1] and in particular is superior on ill-conditioned and non-separable problems. It was successfully applied to a considerable number of real world problems.[1] In [3, 4] the CMA-ES was expanded by the so-called rank-μ-update. The

[1] See www.icos.ethz.ch/software/evolutionary_computation/cmaapplications.pdf for a list of references.

rank-μ-update exploits the information contained in large populations more effectively without affecting the performance[2] for small population sizes. It can reduce the time complexity of the strategy (i.e. the number *of generations* to reach a certain function value) from quadratic to linear [4]. A recent study [5] showed a surprisingly good performance of this CMA-ES on the multimodal Rastrigin function. Large populations and rank-μ-update were the prerequisites for this observation. Therefore, we empirically investigate the effect of the population size λ on the global search performance of the CMA-ES.

The remainder is organized as follows: In Sect. 2 we describe the CMA-ES using weighted recombination and rank-μ-update. In Sect. 3 test functions and methodology for the performance study are outlined. Section 4 examines the performance depending on the population size and compares the CMA-ES with other global search strategies. Sect. 5 gives a summary and conclusion.

2 The CMA-ES with Rank-μ-Update and Weighted Recombination

We thoroughly define the CMA-ES combining weighted recombination [1] and rank-μ-update of the covariance matrix [3, 4]. In this $(\mu_\mathrm{W}, \lambda)$-CMA-ES the λ individuals (candidate solutions) of generation $g+1$ are generated according to

$$\boldsymbol{x}_k^{(g+1)} \sim \mathcal{N}\Big(\langle \boldsymbol{x}\rangle_\mathrm{w}^{(g)}, {\sigma^{(g)}}^2 \boldsymbol{C}^{(g)}\Big), \quad k = 1, \ldots, \lambda, \tag{1}$$

where $\mathcal{N}(\boldsymbol{m}, \boldsymbol{C})$ denotes a normally distributed random vector with mean $\boldsymbol{m}$ and covariance matrix $\boldsymbol{C}$.[3]

The recombination point $\langle \boldsymbol{x}\rangle_\mathrm{w}^{(g)} = \sum_{i=1}^{\mu} w_i \boldsymbol{x}_{i:\lambda}^{(g)}$ is the weighted mean of the selected individuals, $w_i > 0$ for all $i = 1 \ldots \mu$ and $\sum_{i=1}^{\mu} w_i = 1$. The index $i : \lambda$ denotes the i-th best individual. Setting all w_i to $1/\mu$ is equivalent to intermediate (multi-)recombination.The adaptation of the mutation parameters consists of two parts: (i) adaptation of the covariance matrix $\boldsymbol{C}^{(g)}$, and (ii) adaptation of the global step size $\sigma^{(g)}$. The covariance matrix $\boldsymbol{C}^{(g)}$ is adapted by the evolution path $\boldsymbol{p}_c^{(g+1)}$ and by the μ *weighted* difference vectors between the recent parents and $\langle \boldsymbol{x}\rangle_\mathrm{w}^{(g)}$:

$$\boldsymbol{p}_c^{(g+1)} = (1 - c_c) \cdot \boldsymbol{p}_c^{(g)} + H_\sigma^{(g+1)} \sqrt{c_c(2 - c_c)} \cdot \frac{\sqrt{\mu_\mathrm{eff}}}{\sigma^{(g)}} \Big(\langle \boldsymbol{x}\rangle_\mathrm{w}^{(g+1)} - \langle \boldsymbol{x}\rangle_\mathrm{w}^{(g)}\Big) \tag{2}$$

$$\begin{aligned} \boldsymbol{C}^{(g+1)} = {} & (1 - c_\mathrm{cov}) \cdot \boldsymbol{C}^{(g)} + c_\mathrm{cov} \frac{1}{\mu_\mathrm{cov}} \boldsymbol{p}_c^{(g+1)} \Big(\boldsymbol{p}_c^{(g+1)}\Big)^\mathrm{T} \\ & + c_\mathrm{cov} \cdot \left(1 - \frac{1}{\mu_\mathrm{cov}}\right) \sum_{i=1}^{\mu} \frac{w_i}{{\sigma^{(g)}}^2} \Big(\boldsymbol{x}_{i:\lambda}^{(g+1)} - \langle \boldsymbol{x}\rangle_\mathrm{w}^{(g)}\Big) \Big(\boldsymbol{x}_{i:\lambda}^{(g+1)} - \langle \boldsymbol{x}\rangle_\mathrm{w}^{(g)}\Big)^\mathrm{T} , \end{aligned} \tag{3}$$

[2] We define performance as the number of function evaluations needed to reach a certain function value.

[3] Note that $\mathcal{N}\Big(\langle \boldsymbol{x}\rangle_\mathrm{w}^{(g)}, {\sigma^{(g)}}^2 \boldsymbol{C}^{(g)}\Big) \sim \langle \boldsymbol{x}\rangle_\mathrm{w}^{(g)} + \sigma^{(g)} \boldsymbol{B}^{(g)} \boldsymbol{D}^{(g)} \mathcal{N}\Big(\boldsymbol{0}, \boldsymbol{I}\Big)$, see below.

where $H_\sigma^{(g+1)} = 1$ if $\frac{\|\boldsymbol{p}_\sigma^{(g+1)}\|}{\sqrt{1-(1-c_\sigma)^{2(g+1)}}} < (1.5 + \frac{1}{n-0.5})\mathrm{E}\big(\|\mathcal{N}(0,\boldsymbol{I})\|\big)$, and 0 otherwise. $\mu_{\mathrm{eff}} = 1/\sum_{i=1}^{\mu} w_i^2$ denotes the "variance effective selection mass" and $\mu_{\mathrm{eff}} = \mu$ if $w_i = 1/\mu$. The weights w_i are used for the summation term in (3), a matrix with rank $\min(\mu, n)$. Parameter $c_{\mathrm{cov}} \approx \min(1, 2\,\mu_{\mathrm{eff}}/n^2)$ determines the learning rate for the covariance matrix $\boldsymbol{C}$. The adaptation of the global step size $\sigma^{(g+1)}$ is based on a "conjugate" evolution path $\boldsymbol{p}_\sigma^{(g+1)}$:

$$\begin{aligned}\boldsymbol{p}_\sigma^{(g+1)} &= (1-c_\sigma)\cdot\boldsymbol{p}_\sigma^{(g)} \\ &\quad + \sqrt{c_\sigma(2-c_\sigma)}\cdot\boldsymbol{B}^{(g)}{\boldsymbol{D}^{(g)}}^{-1}{\boldsymbol{B}^{(g)}}^{\mathrm{T}}\cdot\frac{\sqrt{\mu_{\mathrm{eff}}}}{\sigma^{(g)}}\left(\langle\boldsymbol{x}\rangle_{\mathrm{w}}^{(g+1)} - \langle\boldsymbol{x}\rangle_{\mathrm{w}}^{(g)}\right).\end{aligned} \tag{4}$$

The orthogonal matrix $\boldsymbol{B}^{(g)}$ and the diagonal matrix $\boldsymbol{D}^{(g)}$ are obtained through a principal component analysis of $\boldsymbol{C}^{(g)}$; $\boldsymbol{C}^{(g)} = \boldsymbol{B}^{(g)}{\boldsymbol{D}^{(g)}}^2{\boldsymbol{B}^{(g)}}^{\mathrm{T}}$ (cf. [1]). The global step size $\sigma^{(g+1)}$ obeys

$$\sigma^{(g+1)} = \sigma^{(g)}\cdot\exp\left(\frac{c_\sigma}{d_\sigma}\left(\frac{\|\boldsymbol{p}_\sigma^{(g+1)}\|}{\mathrm{E}\big(\|\mathcal{N}(0,\boldsymbol{I})\|\big)} - 1\right)\right), \tag{5}$$

where $\mathrm{E}\left(\|\mathcal{N}(0,\boldsymbol{I})\|\right) = \sqrt{2}\,\Gamma(\frac{n+1}{2})/\Gamma(\frac{n}{2}) \approx \sqrt{n}\,(1 - \frac{1}{4n} + \frac{1}{21n^2})$ is the expected length of $\boldsymbol{p}_{\boldsymbol{\sigma}}$ under random selection.

Initial values are $\boldsymbol{p}_\sigma^{(0)} = \boldsymbol{p}_c^{(0)} = \boldsymbol{0}$ and $\boldsymbol{C}^{(0)} = \boldsymbol{I}$, while $\boldsymbol{x}^{(0)}$ and $\sigma^{(0)}$ are problem dependent. Default strategy parameter values are

$$\lambda = 4 + \lfloor 3\cdot\ln(n)\rfloor, \quad \mu = \lfloor\lambda/2\rfloor, \quad w_{i=1\ldots\mu} = \frac{\ln(\mu+1)-\ln(i)}{\sum_{j=1}^{\mu}\ln(\mu+1)-\ln(j)}, \tag{6}$$

$$c_\sigma = \frac{\mu_{\mathrm{eff}}+2}{n+\mu_{\mathrm{eff}}+3},\; d_\sigma = 1 + 2\max\left(0, \sqrt{\frac{\mu_{\mathrm{eff}}-1}{n+1}} - 1\right) + c_\sigma,\; c_c = \frac{4}{n+4}, \tag{7}$$

$$\mu_{\mathrm{cov}} = \mu_{\mathrm{eff}},\; c_{\mathrm{cov}} = \frac{1}{\mu_{\mathrm{cov}}}\frac{2}{(n+\sqrt{2})^2} + \left(1 - \frac{1}{\mu_{\mathrm{cov}}}\right)\min\left(1, \frac{2\mu_{\mathrm{eff}}-1}{(n+2)^2+\mu_{\mathrm{eff}}}\right). \tag{8}$$

While $1/c_\sigma$, and $1/c_c$ can be interpreted as memory time constants, d_σ is a damping parameter. Parameters from (7) and (8) are not meant to be in the users choice. A profound discussion of the strategy parameters is given in [1].

We consider weighted recombination to be more natural than intermediate recombination, because the ranking of all $\lambda/2$ best individuals is fully regarded.[4] Nevertheless, to our experience weighted recombination, where $\mu \approx \lambda/2$, only slightly outperforms intermediate recombination, where $\mu \approx \lambda/4$.

3 Test Functions and Experimental Procedure

3.1 Test Functions

The unconstrained multimodal test problems are summarized in Table 1. All con-

[4] Even the mating success in nature is not well described by two possible outcomes.

Table 1. Test functions to be minimized and initialization regions

Name	Function	Init
Ackley	$f_{\text{Ackley}}(\boldsymbol{x}) = 20 - 20 \cdot \exp\left(-0.2\sqrt{\frac{1}{n}\sum_{i=1}^{n} x_i^2}\right) + \mathrm{e} - \exp\left(\frac{1}{n}\sum_{i=1}^{n}\cos(2\pi x_i)\right)$	$[1, 30]^n$
Bohachevsky	$f_{\text{Bohachevsky}}(\boldsymbol{x}) = \sum_{i=1}^{n-1}\left(x_i^2 + 2x_{i+1}^2 - 0.3\cos(3\pi x_i) - 0.4\cos(4\pi x_{i+1}) + 0.7\right)$	$[1, 15]^n$
Griewank	$f_{\text{Griewank}}(\boldsymbol{x}) = \frac{1}{4000}\sum_{i=1}^{n} x_i^2 - \prod_{i=1}^{n}\cos\left(\frac{x_i}{\sqrt{i}}\right) + 1$	$[10, 600]^n$
Rastrigin	$f_{\text{Rastrigin}}(\boldsymbol{x}) = 10n + \sum_{i=1}^{n}\left(x_i^2 - 10\cos(2\pi x_i)\right)$	$[1, 5]^n$
Scaled Rastrigin	$f_{\text{RastScaled}}(\boldsymbol{x}) = 10n + \sum_{i=1}^{n}\left((10^{\frac{i-1}{n-1}} x_i)^2 - 10\cos(2\pi 10^{\frac{i-1}{n-1}} x_i)\right)$	$[1, 5]^n$
Skew Rastrigin	$f_{\text{RastSkew}}(\boldsymbol{x}) = 10n + \sum_{i=1}^{n}\left(y_i^2 - 10\cos(2\pi y_i)\right)$, with $y_i = \begin{cases} 10 \cdot x_i & \text{if} x_i > 0, \\ x_i & \text{otherwise} \end{cases}$	$[1, 5]^n$
Schaffer	$f_{\text{Schaffer}}(\boldsymbol{x}) = \sum_{i=1}^{n-1}(x_i^2 + x_{i+1}^2)^{0.25} \cdot \left[\sin^2\left(50 \cdot (x_i^2 + x_{i+1}^2)^{0.1}\right) + 1.0\right]$	$[10, 100]^n$
Schwefel	$f_{\text{Schwefel}}(\boldsymbol{x}) = 418.9828872724339 \cdot n - \sum_{i=1}^{n} x_i \cdot \sin(\sqrt{\lvert x_i \rvert})$	$[-500, 300]^n$

sidered functions have a high number of local optima, are scalable in the problem dimension, and have a minimal function value of 0. The known global minimum is located at $\boldsymbol{x} = \boldsymbol{0}$, except for the Schwefel function, where the global minimum within $[-500, 500]^n$ equals 420.96874636 in each coordinate. Additional bounds are implemented for f_{Schwefel} (in $[-500, 500]^n$) and f_{Ackley} (in $[-30, 30]^n$) by adding a quadratic penalty term. E.g., $f_{\text{Schwefel}}(\boldsymbol{x}) + 10^4 \cdot \sum_{i=1}^{n}\theta(\lvert x_i\rvert - 500)x_i^2$ is minimized, where $\theta(.)$ is the Heaviside function. The skew Rastrigin function was proposed by [6] to be deceptive for the CMA-ES.

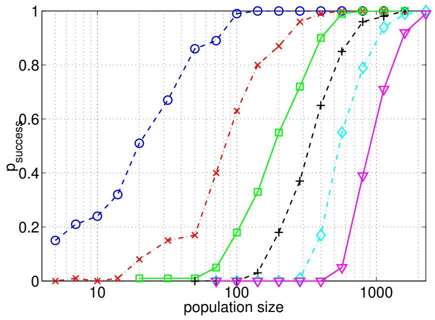
Besides f_{RastSkew} and f_{Schwefel}, the functions are point symmetrical around the global optimum. To avoid an easy exploitation of the symmetry, we suggest non-symmetrical initialization intervals, see Table 1. The Rastrigin functions and f_{Schwefel} are additively separable, while f_{Ackley} and $f_{\text{Bohachevsky}}$ are separable, in that the global optimum can be located by optimizing each variable independently. Recall, that both is not exploited by the CMA-ES, because *all results of the CMA-ES are invariant under orthogonal transformations (rotations) of the coordinate system, given accordingly transformed initial intervals.*

3.2 Experimental procedure

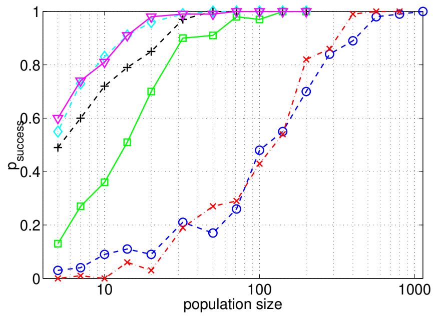
The performance of the CMA-ES is tested for dimensions $n = [2, 5, 10, 20, 40, 80]$.[5] All runs are performed with the default strategy parameter setting given in Sect. 2, except for the population size λ.[6] Starting from $\lambda = 5$, the population

[5] For simulations we used the MATLAB code `cmaes.m`, Version 2.24, available from http://www.icos.ethz.ch/software/evolutionary_computation/cma.

[6] Note that μ is chosen dependently on λ and further parameters depend on μ.



(a)



(b)

Fig. 1. Success rate to reach $f_{\mathrm{stop}} = 10^{-10}$ versus population size for (a) Rastrigin function (b) Griewank function for dimensions $n = 2$ ('$--\bigcirc--$'), $n = 5$ ('$-\cdot-\times-\cdot-$'), $n = 10$ ('—□—'), $n = 20$ ('$--+--$'), $n = 40$ ('$-\cdot-\diamondsuit-\cdot-$'), and $n = 80$ ('—▽—').

size is increased repeatedly in the sequence 5, 7, 10, 14, 20, 32, 50, $\lfloor 50\sqrt{2}\rfloor$, 100, $\lfloor 100\sqrt{2}\rfloor$, 200, $\ldots$, and for each setting 100 runs are conducted. The starting point $\boldsymbol{x}^{(0)}$ is sampled uniformly within the initialization intervals given in Table 1. The initial step size $\sigma^{(0)}$ is set to half of the initialization interval. Too small initial step sizes have a considerable impact on the performance on multimodal functions. Each run is stopped and regarded as successful, when the function value is smaller than $f_{\mathrm{stop}} = 10^{-10}$. Additionally, the run is stopped after 10^7 function evaluations, or when the condition number of the covariance matrix $\boldsymbol{C}$ exceeds 10^{14}, or by the option `TolX`, set to 10^{-15} (for f_{Schaffer} 10^{-30}).

4 Simulation Results

The success rate to reach f_{stop} depends strongly on the population size, see Fig. 1, where exemplary results are shown for (a) $f_{\mathrm{Rastrigin}}$ and (b) f_{Griewank}. The Rastrigin function (Fig. 1a) represents the typical picture. The graphs have a sigmoidal shape and larger dimensions require larger population sizes to approach 100% success rate. We observe two exceptions from this behavior. First, on f_{RastSkew} the success rates are low with any population size, and, except for very low dimensions, f_{RastSkew} is not solvable for the CMA-ES. The second exception is shown in Fig. 1b: on f_{Griewank} smaller dimensions require larger population sizes, but success rates of 100% can be achieved in all dimensions.

Figures 2a and 3a show the scaling of the (minimal) population size w.r.t. the problem dimension to achieve success rates of (at least) $5, 25, 50, 75, 95\%$. Graphs for $f_{\mathrm{RastScaled}}$ are almost identical with $f_{\mathrm{Rastrigin}}$ and therefore omitted. Except for f_{RastSkew} (not shown), larger success rates require larger population sizes. The figures are sorted from (i.a) to (vi.a) by increasing slopes that indicate the scaling. The steepest slope (f_{Schwefel}) is slightly above linear. For all other functions the slope is sub-linear.

Figures 2b and 3b show performance versus population size. Performance is measured as mean number of function evaluations for successful runs, divided by

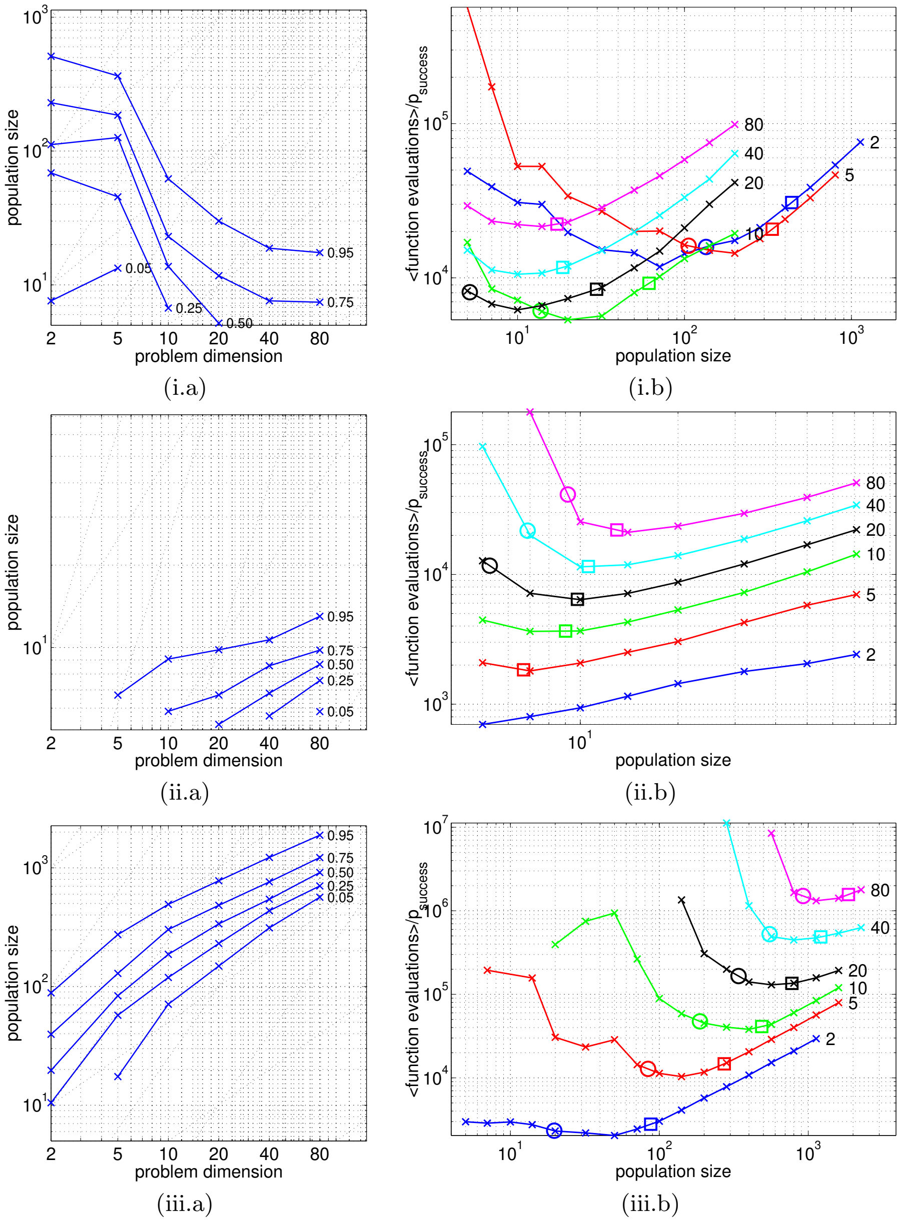


Fig. 2. (i) Griewank function, (ii) Ackley function, (iii) Rastrigin function. (a) population size to reach success rates of 0.05, 0.25, 0.5, 0.75, and 0.95 versus problem dimension n. The sloping grid indicates linear and quadratic dependency. (b) average number of function evaluations to reach $f_{\text{stop}} = 10^{-10}$ divided by the success rate, versus population size for problem dimensions $n = 2, 5, 10, 20, 40, 80$. The symbols $\bigcirc$ and $\square$ indicate success rates of 50% and 95%, respectively. Missing points on the left side of a graph indicate that no run (out of 100 runs) was successful.

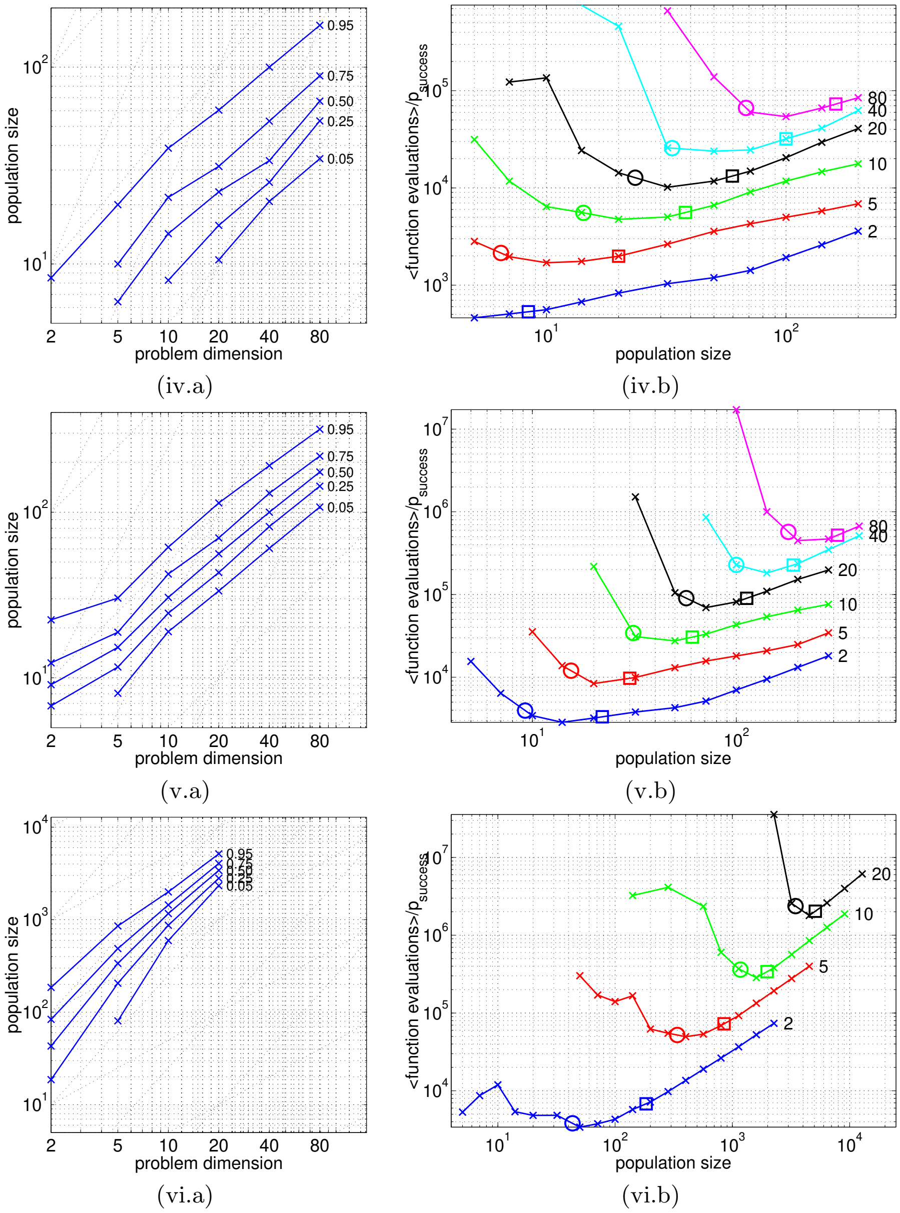


Fig. 3. (iv) Bohachevsky function, (v) Schaffer function, (vi) Schwefel function. (a) population size to reach success rates of 0.05, 0.25, 0.5, 0.75, and 0.95 versus problem dimension n. The sloping grid indicates linear and quadratic dependency. (b) average number of function evaluations to reach $f_{\text{stop}} = 10^{-10}$ divided by the success rate, versus population size for problem dimensions $n = 2, 5, 10, 20, 40, 80$. The symbols $\bigcirc$ and $\square$ indicate success rates of 50% and 95%, respectively. Missing points on the left side of a graph indicate that no run (out of 100 runs) was successful.

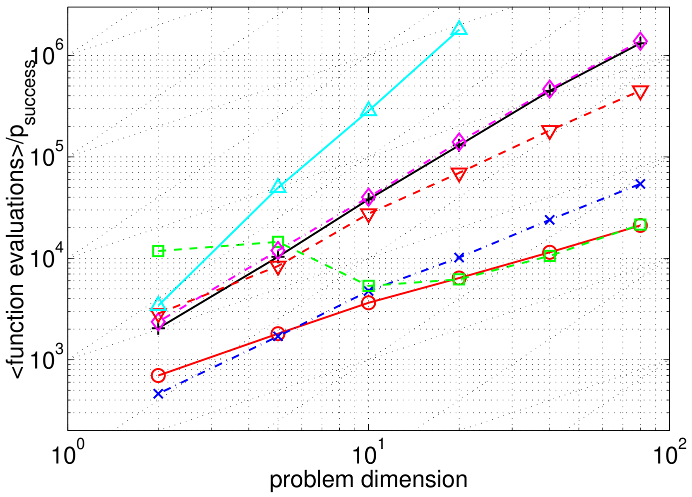


Fig. 4. Mean number of function evaluations to reach $f_{\rm stop}$ versus problem dimension n for CMA-ES on Ackley ('—◯—'), Bohachevsky ('−·−×−·−'), Griewank ('−−□−−'), Rastrigin ('—+—'), Scaled Rastrigin ('−·−◇−·−'), Schaffer ('−−▽−−'), and Schwefel ('—△—') function.

the success rate. This performance measure assumes the same expected number of function evaluations for successful and for unsuccessful runs. The best performance is usually achieved for success rates between 50% and 95%. The impact of a smaller than optimal population size can be high. With increasing population size the performance decreases at most linearly.

Figure 4 shows the scaleup of the performance with optimal population size. The scaleup with n, put in order, appears to be at most linear for $f_{\rm Ackley}$ and $f_{\rm Griewank}$, between linear and quadratic for $f_{\rm Bohachevsky}$, $f_{\rm Schaffer}$, $f_{\rm Rastrigin}$, and $f_{\rm RastScaled}$, and slightly below cubic for $f_{\rm Schwefel}$.

Table 2 compares the performance of the CMA-ES with optimal population size to the performance of Differential Evolution (DE) [7], the Robust Evolution Strategy (RES) [8], and a Local Optima Smoothing (LOS) [9] restart BFGS algorithm. For each function the results with the best parameter tuning were taken from the respective publication and additional experiments were performed with DE and BFGS.[7] Only on the additively separable functions $f_{\rm Rastrigin}$ and $f_{\rm Schwefel}$, DE outperforms the CMA-ES by a factor of five to 50. Otherwise, the CMA-ES outperforms DE by a factor of at least three, while DE even fails to find the global optimum on the non-separable $f_{\rm Rastrigin}(\boldsymbol{Ax})$, where an orthogonal transformation $\boldsymbol{A}$ of the search space is applied, and performes much worse

[7] In [9] the results for LOS are stated as numbers of necessary local searches to hit the global optimum. The average number of function evaluations for a local search using BFGS is simulated with MATLAB's `fminunc` with `MaxFunEvals` $= 500n$ and `TolX` $=$ `TolFun` $= 10^{-3}$ (0.9 for $f_{\rm Rastrigin}$). A *random* restart strategy performs much worse than LOS.

Table 2. Average number of function evaluations to reach f_{stop} of CMA-ES versus DE, RES, and LOS on Griewank, Ackley, Rastrigin, and Schwefel function. If f_{stop} is not reached the *best function value/number of function evaluations* are stated. The initialization regions do not apply to CMA-ES for which the more difficult intervals from Table 1 are used. Results are taken from [7–9], except for the cases marked with * obtained using the DE code available from http://www.icsi.berkeley.edu/∼storn/code.html. The matrix $\boldsymbol{A} = [\boldsymbol{o}_1, \ldots, \boldsymbol{o}_n]^T$ implements an angle-preserving (i.e. orthogonal) linear transformation of $\boldsymbol{x}$ (cf. [1]), chosen anew for each run.

Function	f_{stop}	init	n	CMA-ES	DE	RES	LOS
$f_{\text{Ackley}}(\boldsymbol{x})$	1e-3	$[-30, 30]^n$	20	**2667**	.	.	6.0e4
			30	**3701**	12481	1.1e5	9.3e4
			100	**11900**	36801	.	.
$f_{\text{Griewank}}(\boldsymbol{x})$	1e-3	$[-600, 600]^n$	20	**3111**	8691	.	.
			30	**4455**	11410 *	*8.5e-3/2e5*	.
			100	**12796**	31796	.	.
$f_{\text{Rastrigin}}(\boldsymbol{x})$	0.9	$[-5.12, 5.12]^n$	20	68586	**12971**	.	9.2e4
		DE: $[-600, 600]^n$	30	147416	**20150** *	1.0e5	2.3e5
			100	1010989	**73620**	.	.
$f_{\text{Rastrigin}}(\boldsymbol{Ax})$	0.9	$[-5.12, 5.12]^n$	30	**152000**	*171/1.25e6* *	.	.
			100	**1011556**	*944/1.25e6* *	.	.
$f_{\text{Schwefel}}(\boldsymbol{x})$	1e-3	$[-500, 500]^n$	5	43810	**2567** *	.	7.4e4
			10	240899	**5522** *	.	5.6e5

on $f_{\text{Schwefel}}(\boldsymbol{Ax})$ (not shown).[8] This supports our hypothesis that DE strongly exploits the separability of the function. The RES too exploits separability by sampling Cauchy distributions which strongly favor steps in coordinate directions. Even so, on the separable $f_{\text{Rastrigin}}$ RES outperforms CMA-ES only by a factor of 1.5 , while on f_{Ackley} and f_{Griewank} it performs worse by a factor of 30 or more. The LOS performs between a factor 1.5 (on $f_{\text{Rastrigin}}$) and a factor of 25 (on f_{Ackley}) worse than the CMA-ES.[9]

5 Summary and Conclusion

The CMA-ES with rank-μ-update is investigated on a suit of eight highly multimodal test functions for problem dimensions between 2 and 80. Tuning (that is increasing) the population size considerably improves the performance on six of the functions, compared to the performance with default population size.

[8] On $f_{\text{Rastrigin}}$, the parent number equals 20 in DE. Increasing the parent number improves the performance on $f_{\text{Rastrigin}}(\boldsymbol{Ax})$. However, even with 500 parents for $n = 10$, the minimum function value reached does not drop below 2 after 10^7 function evaluations. Choosing the recombination parameter CR = 1, DE becomes invariant against orthogonal transformations, but performs even worse on $f_{\text{Rastrigin}}$.

[9] Two parameters of LOS, r and K, see [9], are chosen to be optimal for each entry. In particular r has a considerable impact on the performance.

On seven of the eight functions, the CMA-ES can precisely locate the global optimum. If the local optima can be interpreted as pertubations of an underlying unimodal function, the CMA-ES with a large population size can "detect" the global topology. Then, the global optimum is located within $300n$ and $500n^2$ function evaluations. A strong asymmetry of the underlying function jeopardizes a successful detection and can lead to a failure (as on f_{RastSkew}). The optimal population size usually scales sub-linearly with the problem dimension n, but significantly depends on the test function considered.

The results were compared with other global search strategies, stated to achieve superior results in earlier investigations. Surprisingly, the CMA-ES outperforms these global searchers, typically by a factor of three, with the following exception. Only if the function is *additively separable*, Differential Evolution strongly outperforms the CMA-ES. If the search space is rotated, the performance of the CMA-ES is unchanged, however Differential Evolution massively degrades in performance or even fails to locate the global optimum with a reasonable probablity. For the CMA-ES the population size was tuned, while for the compared algorithms up to three parameters had to be tuned to the given objective function. In our opinion, tuning the population size in the CMA-ES is comparatively unproblematic. The results suggest that a CMA-ES restart strategy with a successively increased population size (by a factor of three, initialized with the default population size) constitutes a highly competitive, quasi parameter free global optimization algorithm for non-separable objective functions.

References

1. Hansen, N., Ostermeier, A.: Completely derandomized self-adaptation in evolution strategies. Evol. Comput. **9** (2001) 159–195
2. Hansen, N., Ostermeier, A.: Adapting arbitrary normal mutation distributions in evolution strategies: The covariance matrix adaptation. In: Proceedings of the 1996 IEEE Conference on Evolutionary Computation (ICEC '96). (1996) 312–317
3. Müller, S.D., Hansen, N., Koumoutsakos, P.: Increasing the serial and the parallel performance of the CMA-evolution stategy with large populations. In: Parallel Problem Solving from Nature (PPSN). (2002)
4. Hansen, N., Müller, S.D., Koumoutsakos, P.: Reducing the time complexity of the derandomized evolution strategy with covariance matrix adaptation (CMA-ES). Evol. Comput. **11** (2003) 1–18
5. Kern, S., Müller, S.D., Hansen, N., Büche, D., Ocenasek, J., Koumoutsakos, P.: Learning probability distributions in continuous evolutionary algorithms – a comparative review. Natural Computing **3** (2004) 77–112
6. Büche, D.: Personal communication (2003)
7. Storn, R., Price, K.: Differential evolution: A simple and efficient heuristic for global optimization over continuous spaces. J. Glob. Opt. **11** (1997) 341–359
8. Ohkura, K., Matsumura, Y., Ueda, K.: Robust evolution strategies. Lect. Notes Comput. Sc. **1585** (1999) 10–17
9. Addis, B., Locatelli, M., Schoen, F.: Local optima smoothing for global optimization. Technical report dsi 5–2003, Dipartimento di Sistemi e Informatica, Università degli Studi di Firenze (2003)