

Benchmarking a BI-Population CMA-ES on the BBOB-2009 Function Testbed

Nikolaus Hansen

► **To cite this version:**

Nikolaus Hansen. Benchmarking a BI-Population CMA-ES on the BBOB-2009 Function Testbed. ACM-GECCO Genetic and Evolutionary Computation Conference, Jul 2009, Montreal, Canada. 2009. <inria-00382093>

HAL Id: inria-00382093

<https://hal.inria.fr/inria-00382093>

Submitted on 7 May 2009

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.

Benchmarking a BI-Population CMA-ES on the BBOB-2009 Function Testbed

Nikolaus Hansen
Microsoft Research–INRIA Joint Centre
28 rue Jean Rostand
91893 Orsay Cedex, France
Nikolaus.Hansen@inria.fr

ABSTRACT

We propose a multistart CMA-ES with equal budgets for two interlaced restart strategies, one with an increasing population size and one with varying small population sizes. This BI-population CMA-ES is benchmarked on the BBOB-2009 noiseless function testbed and could solve 23, 22 and 20 functions out of 24 in search space dimensions 10, 20 and 40, respectively, within a budget of less than $10^6 D$ function evaluations per trial.

Categories and Subject Descriptors

G.1.6 [Numerical Analysis]: Optimization—*global optimization, unconstrained optimization*; F.2.1 [Analysis of Algorithms and Problem Complexity]: Numerical Algorithms and Problems

General Terms

Algorithms

Keywords

Benchmarking, Black-box optimization, Evolutionary computation, CMA-ES

1. INTRODUCTION

The *covariance matrix adaptation evolution strategy* (CMA-ES) is a stochastic, population-based search method in continuous search spaces, aiming at minimizing an objective function $f : \mathbb{R}^D \rightarrow \mathbb{R}$ in a black-box scenario. In this paper, the $(\mu/\mu_w, \lambda)$ -CMA-ES [3] is applied in a multistart strategy and benchmarked on 24 functions. Comprehensive results for the number of function evaluations to reach a target function value are given.

2. THE $(\mu/\mu_w, \lambda)$ -CMA-ES

In the standard $(\mu/\mu_w, \lambda)$ -CMA-ES [3, 6, 8], in each iteration step t , λ new solutions $\mathbf{x}_i \in \mathbb{R}^D$ are generated by sam-

pling a multi-variate normal distribution, $\mathcal{N}(\mathbf{0}, \mathbf{C}^t)$, with mean $\mathbf{0}$ and $n \times n$ covariance matrix \mathbf{C}^t , see Eq. (1). The μ best solutions are selected to update the distribution parameters for the next iteration $t + 1$. The complete algorithm is depicted in Table 1. We set $\mu = \lfloor \frac{\lambda}{2} \rfloor$, $w_i = \frac{\ln(\mu+1) - \ln i}{\sum_{j=1}^{\mu} (\ln(\mu+1) - \ln j)}$, $\mu_w^{-1} = \sum_{i=1}^{\mu} w_i^2$, and $d_\sigma = 1 + c_\sigma + 2 \max\left(0, \sqrt{\frac{\mu_w - 1}{D+1}} - 1\right)$ usually close to one, $\mathbf{C}^{t-\frac{1}{2}}$ is symmetric and satisfies $\mathbf{C}^{t-\frac{1}{2}} \mathbf{C}^{t-\frac{1}{2}} = (\mathbf{C}^t)^{-1}$. The remaining learning parameters have been slightly modified, without connection to the BBOB-2009 testbed, and are chosen as $c_\sigma = \frac{\mu_w + 2}{D + \mu_w + 5}$, $c_c = \frac{4 + \mu_w/D}{D + 4 + 2\mu_w/D}$, $c_1 = \frac{2}{(D+1.3)^2 + \mu_w}$ and $c_\mu = \min\left(1 - c_1, 2 \frac{\mu_w - 2 + 1/\mu_w}{(D+2)^2 + \mu_w}\right)$.

2.1 BI-Population Multistart Scheme

The $(\mu/\mu_w, \lambda)$ -CMA-ES with the default population size $\lambda_{\text{def}} = 4 + \lfloor 3 \ln D \rfloor$ is a robust and fast *local* search method [9]. With a large(r) population size a more global search can be accomplished successfully [6, 7]. After a first single run with default population size, we apply two interlaced multistart regimes, each equipped with a function evaluation budget accounting for the so far conducted function evaluations. Depending on which budget value is smaller, a complete run of either one or the other strategy is launched. The first and last restart are conducted under the first regime.

Under the first regime, we restart with **increasing population size**, where before each restart the population size λ is increased by a factor of two [1]. At most nine restarts are conducted, *i.e.*, the largest population size is $\lambda = 2^9 \lambda_{\text{def}} = 512 \lambda_{\text{def}}$. The initial $\sigma^0 = 2$ (*i.e.*, 1/5 of the domain width). The budget is loaded from the first restart, *i.e.*, the first single run with population size λ_{def} is disregarded.

Second, a multistart regime with **small population size** is applied, where the population size λ is set to

$$\lambda_s = \left\lfloor \lambda_{\text{def}} \left(\frac{1}{2} \frac{\lambda_\ell}{\lambda_{\text{def}}} \right)^{\mathcal{U}[0,1]^2} \right\rfloor,$$

where λ_ℓ is the latest population size from the first regime with increasing (large) λ . Here $\mathcal{U}[0,1]$ denote independent uniformly distributed numbers in $[0,1]$ and $\lambda_s \in [\lambda_{\text{def}}, \lambda/2]$. The initial step-size is set to $\sigma^0 = 2 \times 10^{-2 \mathcal{U}[0,1]}$. A maximum number of function evaluations of half of the recent large budget is enforced, but probably of minor relevance. The second multistart regime is launched, if and only if its recent budget is smaller than the one for the first regime with increasing populations.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. To copy otherwise, to republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee.

GECCO'09, July 8–12, 2009, Montréal Québec, Canada.
Copyright 2009 ACM 978-1-60558-505-5/09/07 ...\$5.00.

Table 1: Update equations for the state variables in the $(\mu/\mu_w, \lambda)$ -CMA-ES with iteration index $t = 0, 1, 2, \dots$, where $\mathbf{p}_\sigma^{t=0} = \mathbf{p}_c^{t=0} = \mathbf{0}$ and $\mathbf{C}^{t=0} = \mathbf{I}$. Here, $\mathbf{x}_{i:\lambda}$ is the i -th best of the solutions $\mathbf{x}_1, \dots, \mathbf{x}_\lambda$ and $h_\sigma = 1$ if $\|\mathbf{p}_\sigma^{t+1}\| < \sqrt{1 - (1 - c_\sigma)^{2(t+1)}} \left(1.4 + \frac{2}{D+1}\right) \mathbb{E}\|\mathcal{N}(\mathbf{0}, \mathbf{I})\|$ and zero otherwise. Further symbols and constants and $\mathbf{m}^{t=0}$ and $\sigma^{t=0}$ are given in the text. The chosen ordering of equations allows to remove the time index in all variables but \mathbf{m}^t

Given $t \in \mathbb{N}$, $\mathbf{m}^t \in \mathbb{R}^D$, $\sigma^t \in \mathbb{R}_+$, $\mathbf{C}^t \in \mathbb{R}^{D \times D}$ positive definite, $\mathbf{p}_\sigma^t \in \mathbb{R}^D$, and $\mathbf{p}_c^t \in \mathbb{R}^D$

$$\mathbf{x}_i \sim \mathbf{m}^t + \sigma^t \times \mathcal{N}_i(\mathbf{0}, \mathbf{C}^t) \quad \text{is normally distributed for } i = 1, \dots, \lambda \text{ and evaluated on } f \quad (1)$$

$$\mathbf{m}^{t+1} = \sum_{i=1}^{\mu} w_i \mathbf{x}_{i:\lambda} \quad \text{where } f(\mathbf{x}_{1:\lambda}) \leq \dots \leq f(\mathbf{x}_{\mu:\lambda}) \leq f(\mathbf{x}_{\mu+1:\lambda}) \dots \quad (2)$$

$$\mathbf{p}_\sigma^{t+1} = (1 - c_\sigma) \mathbf{p}_\sigma^t + \sqrt{c_\sigma(2 - c_\sigma)\mu_w} \mathbf{C}^{t-\frac{1}{2}} \frac{\mathbf{m}^{t+1} - \mathbf{m}^t}{\sigma^t} \quad (3)$$

$$\mathbf{p}_c^{t+1} = (1 - c_c) \mathbf{p}_c^t + h_\sigma \sqrt{c_c(2 - c_c)\mu_w} \frac{\mathbf{m}^{t+1} - \mathbf{m}^t}{\sigma^t} \quad (4)$$

$$\mathbf{C}^{t+1} = (1 - c_1 - c_\mu + (1 - h_\sigma) c_1 c_c (2 - c_c)) \mathbf{C}^t + c_1 \mathbf{p}_c^{t+1} \mathbf{p}_c^{t+1 \top} + c_\mu \sum_{i=1}^{\mu} w_i \frac{\mathbf{x}_{i:\lambda} - \mathbf{m}^t}{\sigma^t} \times \frac{(\mathbf{x}_{i:\lambda} - \mathbf{m}^t)^\top}{\sigma^t} \quad (5)$$

$$\sigma^{t+1} = \sigma^t \times \exp\left(\frac{c_\sigma}{d_\sigma} \left(\frac{\|\mathbf{p}_\sigma^{t+1}\|}{\mathbb{E}\|\mathcal{N}(\mathbf{0}, \mathbf{I})\|} - 1\right)\right) \quad (6)$$

2.2 Initial and Termination Criteria

The initial mean \mathbf{m}^0 is sampled uniformly distributed in $[-4, 4]^D$. A single run of the $(\mu/\mu_w, \lambda)$ -CMA-ES is terminated, when the final target function value is reached or one of the following termination conditions is satisfied.

MaxIter = $100 + 50(D + 3)^2/\sqrt{\lambda}$ is the maximal number of iterations in each run of CMA-ES

TolHistFun = 10^{-12} : the range of the best function values during the last $10 + \lceil 30D/\lambda \rceil$ iterations is smaller than **TolHistFun**.

EqualFunVals: in more than $1/3^{\text{rd}}$ of the last D iterations the objective function value of the best and the k -th best solution are identical, that is $f(\mathbf{x}_{1:\lambda}) = f(\mathbf{x}_{k:\lambda})$, where $k = 1 + \lfloor 0.1 + \lambda/4 \rfloor$.

TolX = 10^{-12} : all components of \mathbf{p}_c^t and all square roots of diagonal components of \mathbf{C}^t , multiplied by σ^t/σ^0 , are smaller than **TolX**.

TolUpSigma = 10^{20} : $\sigma^t/\sigma^0 > \text{TolUpSigma} \sqrt{l^t}$, where l^t is the largest eigenvalue of \mathbf{C}^t , indicates a mismatch between σ increase and decrease of all eigenvalues in \mathbf{C} . In this, rather untypical, case the progression of the strategy is usually very low and a restart is indicated.

Stagnation: the median of the 20 newest values is not smaller than the median of the 20 oldest values, respectively, in the two arrays containing the best function values and the median function values of the last $\lceil 0.2t + 120 + 30D/\lambda \rceil$ iterations.

ConditionCov: the condition number of \mathbf{C}^t exceeds 10^{14} .

NoEffectAxis: \mathbf{m}^t remains numerically constant when adding $0.1\sigma^t \sqrt{l^t} \mathbf{v}^t$, where l^t is the $1+(t \bmod D)$ -largest eigenvalue of \mathbf{C}^t and \mathbf{v}^t is the corresponding normalized eigenvector.

NoEffectCoor: any element of \mathbf{m}^t remains numerically constant when adding $0.2\sigma^t \mathbf{l}^t$, where elements of \mathbf{l}^t are the square root of the diagonal elements of \mathbf{C}^t . Condition **NoEffectCoor** was never satisfied.

Most criteria are standard part of our production codes of $(\mu/\mu_w, \lambda)$ -CMA-ES (see also next section). Restarts are launched until the final target function value or the largest, final population size is reached (see above). In neither case more than $10^6 D$ function evaluations were conducted.

3. PARAMETER TUNING

No thorough parameter study has been done. We have experimented with restarts from a so-far best found solution point but had comparatively little success. The parameters for the first multistart scheme are taken from [1], those for the second are ad-hoc settings. We reckon that even smaller population sizes λ_s could be useful. The maximum number of iterations **MaxIter** has been set to prevent excessive long runs and is chosen such that most functions should be solvable within this limit. Most other termination criteria are standard, while **TolUpSigma** and **Stagnation** have been only recently added to the set of standard termination criteria. The former indicates a problem in acquiring the functions topography and seems only effective up to $D = 10$. The latter is of major relevance for noisy functions. The same D -dependent parameter setting is used on all functions and therefore the crafting effort [4] computes to **CrE** = 0.

4. CPU TIMING EXPERIMENT

For the timing experiment the complete algorithm was run on f_8 and restarted until at least 30 seconds had passed (according to Figure 2 in [4]). These experiments have been conducted with an Intel dual core T5600 processor with 1.8 GHz under Linux 2.6.27-11 using Matlab R2008a. The results are shown in the following table.

| D | 2 | 3 | 5 | 10 | 20 | 40 | 80 |
|--------------------------|-----|-----|-----|-----|-----|-----|-----|
| seconds $\times 10^{-4}$ | 2.8 | 2.4 | 2.0 | 1.8 | 1.8 | 2.0 | 6.0 |

Up to 10-D, the necessary CPU time even reduces with increasing dimension, presumably due to a larger number of initialization procedures for the restarts until 30 seconds have passed.

Equations (1) and (3) require a decomposition of C^t . An eigendecomposition with time complexity $\propto D^3$ is applied and for computational efficiency reasons only conducted until after

$$\frac{(c_1 + c_\mu)^{-1}}{10D} \quad (7)$$

iterations have passed. Therefore, a slightly outdated decomposition is used in case. This policy results in a quadratic scaling of the internal time complexity with the dimension. For larger dimension, a computational burden between 10^{-8} and $10^{-7} \times D^2$ seconds per function evaluation is the typical outcome of timing experiments (for $D = 80$ the table reveals $9 \times 10^{-8} D^2$ seconds).

5. RESULTS AND DISCUSSION

Results from experiments according to [4] on the benchmark functions given in [2, 5] are presented in Figures 1 and 2 and in Table 2.

The number of solved functions amounts to 24, 24, 24, 23, 22, 20 out of 24 for dimension 2, 3, 5, 10, 20, 40. Two functions, f_3 and f_4 , seem to become practically unsolvable with increasing dimension. The scaling of the running time (expected number of function evaluations, ERT) with the problem dimension is linear for f_1 , f_5 and f_{12} and clearly sub-quadratic for most unimodal functions. For the multi-modal functions the scaling is typically quadratic, in some cases worse, but never better. Running times to reach the final target function value in 20-D range between D and somewhat above $3 \times 10^5 D$. They are typically above $300D$ and below $30\,000D$.

The failure on f_3 for larger dimensions is unexpected and caused by the introduced deformation of the Rastrigin function (see [2, 5]). We suspect that a local minimum with a larger attraction basin has been generated, while this seems not to be the case for f_{15} .

Functions f_4 and f_{24} had been designed to be deceptive for evolution strategies. Nevertheless, f_{24} can be solved, but only with a very large budget of $3 \times 10^5 D^2$ function evaluations, also due to a small success probability.

6. SUCCESSFUL POPULATION SIZE

We investigate the population sizes of the final successful runs whenever at least one restart was executed. In Table 3 minimal, median (the larger in case of even data) and maximal population size are given. For the functions not listed, no restarts were necessary in 20-D (with one exception with a single restart in one trial on f_9). On all multi-modal functions f_{15-24} restarts are applied. Functions 20 and 24 require a population size above 1000. Functions 19, 21 and 23 are solved with the largest range of different population sizes.

Table 4 tabulates minimal, median (the larger in case of even data) and maximal initial step-size σ^0 of the final successful runs, whenever $\sigma^0 < 2$ in at least one case. Only for functions 23 and 24, the smaller initial step-size appears to be beneficial, while for f_{22} the data are not conclusive. The

Table 3: Final population sizes in 20-D, where $\lambda_{\text{def}} = 12$, when at least one restart was executed

| f | min | med | max |
|-----|------|------|------|
| 7 | 96 | 96 | 96 |
| 13 | 24 | 48 | 96 |
| 15 | 200 | 384 | 768 |
| 16 | 56 | 115 | 384 |
| 17 | 48 | 96 | 192 |
| 18 | 96 | 192 | 192 |
| 19 | 236 | 6144 | 6144 |
| 20 | 3072 | 6144 | 6144 |
| 21 | 12 | 101 | 1678 |
| 22 | 14 | 45 | 202 |
| 23 | 114 | 381 | 1441 |
| 24 | 2137 | 4456 | 4675 |

Table 4: Initial step-size σ^0 of successful restarts in 20-D for functions, where $\sigma^0 < 2$ was successful at least once

| f | min | med | max |
|-----|-------|-------|-------|
| 15 | 0.04 | 2 | 2 |
| 16 | 0.066 | 2 | 2 |
| 17 | 1.06 | 2 | 2 |
| 19 | 0.054 | 2 | 2 |
| 21 | 0.044 | 1.66 | 2 |
| 22 | 0.024 | 0.4 | 0.6 |
| 23 | 0.02 | 0.032 | 0.1 |
| 24 | 0.036 | 0.068 | 0.166 |

multi-modal functions f_{17} , f_{18} and f_{20} were never solved with an initially small step-size.

7. CONCLUSION

The BI-population CMA-ES performs satisfactorily on many functions of the BBOB-2009 testbed and exhibits a reasonable scaling behavior: between linear and quadratic on unimodal functions, between quadratic and cubic on multi-modal functions. Yet, it can be considerably outperformed at least (a) on functions that are smooth, “regular” and only moderately ill-conditioned (f_1 , f_5 , f_8 , f_9), (b) on separable functions (in particular f_3 and f_4) and (c) on the multi-modal functions f_{21} and f_{22} . The former two cases are intrinsic and connected to invariance properties of the algorithm, namely (a) invariance to order-preserving transformations of the function value and (b) rotational invariance. Case (c) might be successfully addressed by an improved restart schedule.

Acknowledgments

The author would like to acknowledge the great and hard work of the BBOB team with particular kudos to Raymond Ros, Steffen Finck and Anne Auger, and Anne Auger and Marc Schoenauer for their kind and persistent support.

8. REFERENCES

- [1] A. Auger and N. Hansen. A restart CMA evolution strategy with increasing population size. In *Proceedings of the IEEE Congress on Evolutionary Computation (CEC 2005)*, pages 1769–1776. IEEE Press, 2005.

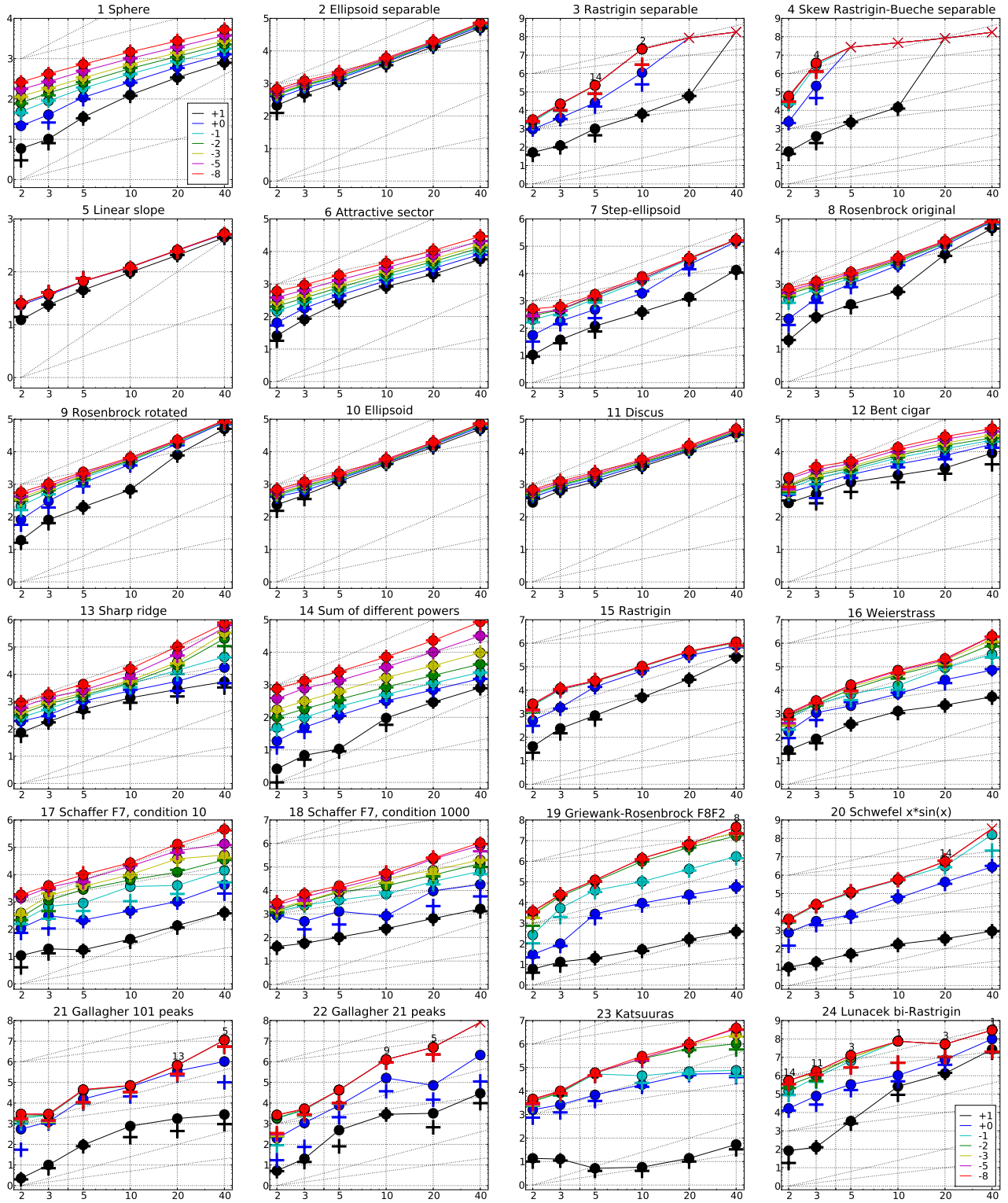


Figure 1: Expected Running Time (ERT, \bullet) to reach $f_{\text{opt}} + \Delta f$ and median number of function evaluations of successful trials (+), shown for $\Delta f = 10, 1, 10^{-1}, 10^{-2}, 10^{-3}, 10^{-5}, 10^{-8}$ (the exponent is given in the legend of f_1 and f_{24}) versus dimension in log-log presentation. The $\text{ERT}(\Delta f)$ equals to $\#FES(\Delta f)$ divided by the number of successful trials, where a trial is successful if $f_{\text{opt}} + \Delta f$ was surpassed during the trial. The $\#FES(\Delta f)$ are the total number of function evaluations while $f_{\text{opt}} + \Delta f$ was not surpassed during the trial from all respective trials (successful and unsuccessful), and f_{opt} denotes the optimal function value. Crosses (\times) indicate the total number of function evaluations $\#FES(-\infty)$. Numbers above ERT-symbols indicate the number of successful trials. Annotated numbers on the ordinate are decimal logarithms. Additional grid lines show linear and quadratic scaling.

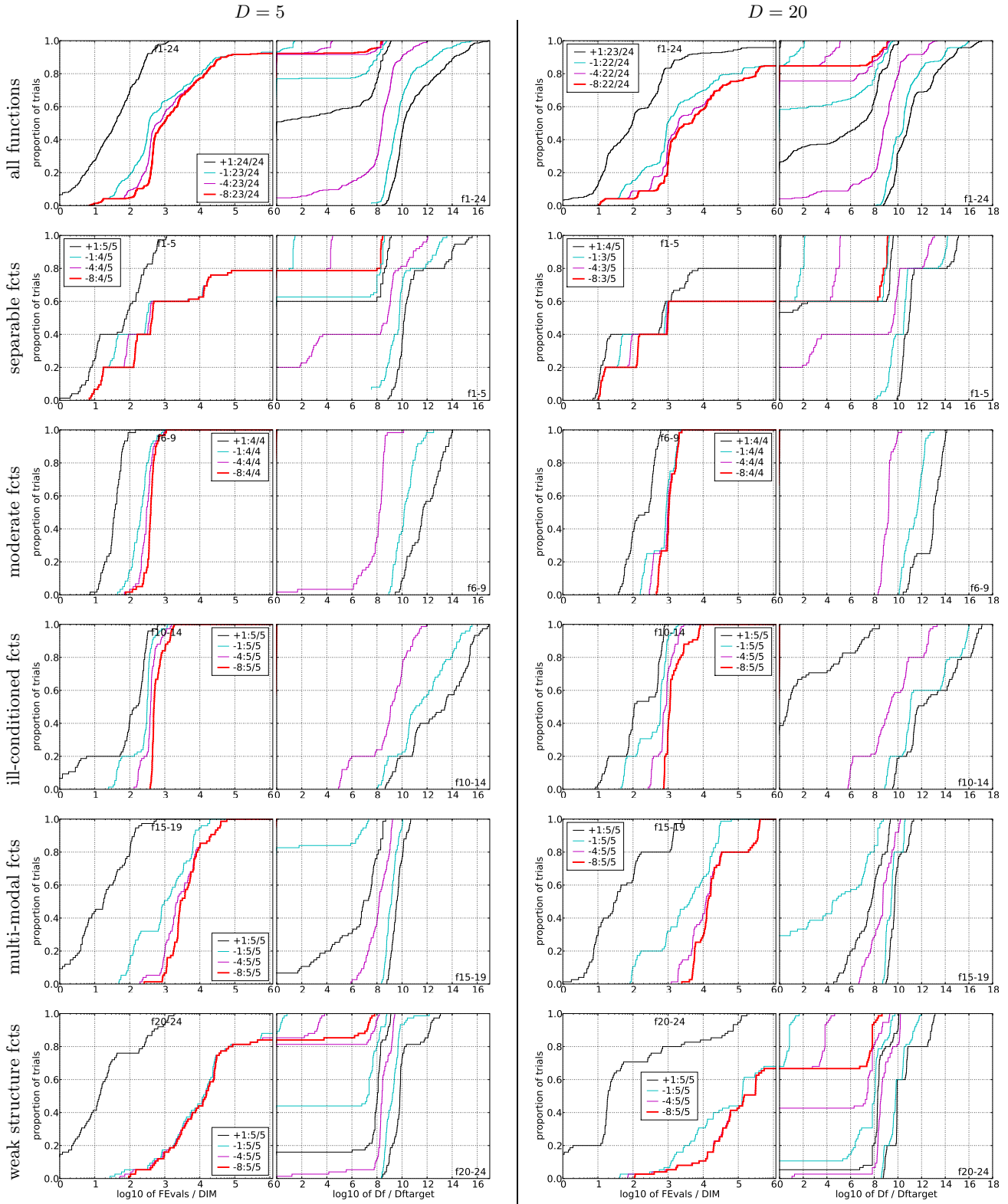


Figure 2: Empirical cumulative distribution functions (ECDFs), plotting the fraction of trials versus running time (left subplots) or versus Δf (right subplots). The thick red line represents the best achieved results. Left subplots: ECDF of the running time (number of function evaluations), divided by search space dimension D , to fall below $f_{\text{opt}} + \Delta f$ with $\Delta f = 10^k$, where k is the first value in the legend. Right subplots: ECDF of the best achieved Δf divided by 10^k (upper left lines in continuation of the left subplot), and best achieved Δf divided by 10^{-8} for running times of $D, 10D, 100D \dots$ function evaluations (from right to left cycling black-cyan-magenta). Top row: all results from all functions; second row: separable functions; third row: misc. moderate functions; fourth row: ill-conditioned functions; fifth row: multi-modal functions with adequate structure; last row: multi-modal functions with weak structure. The legends indicate the number of functions that were solved in at least one trial. FEvals denotes number of function evaluations, D and DIM denote search space dimension, and Δf and Df denote the difference to the optimal function value.

- [2] S. Finck, N. Hansen, R. Ros, and A. Auger. Real-parameter black-box optimization benchmarking 2009: Presentation of the noiseless functions. Technical Report 2009/20, Research Center PPE, 2009.
- [3] N. Hansen. The CMA evolution strategy: a comparing review. In J. Lozano, P. Larranaga, I. Inza, and E. Bengoetxea, editors, *Towards a new evolutionary computation. Advances on estimation of distribution algorithms*, pages 75–102. Springer, 2006.
- [4] N. Hansen, A. Auger, S. Finck, and R. Ros. Real-parameter black-box optimization benchmarking 2009: Experimental setup. Technical Report RR-6828, INRIA, 2009.
- [5] N. Hansen, S. Finck, R. Ros, and A. Auger. Real-parameter black-box optimization benchmarking 2009: Noiseless functions definitions. Technical Report RR-6829, INRIA, 2009.
- [6] N. Hansen and S. Kern. Evaluating the CMA evolution strategy on multimodal test functions. In X. Yao et al., editors, *Parallel Problem Solving from Nature - PPSN VIII, LNCS 3242*, pages 282–291. Springer, 2004.
- [7] N. Hansen, S. D. Müller, and P. Koumoutsakos. Reducing the time complexity of the derandomized evolution strategy with covariance matrix adaptation (CMA-ES). *Evolutionary Computation*, 11(1):1–18, 2003.
- [8] N. Hansen, A. Niederberger, L. Guzzella, and P. Koumoutsakos. A method for handling uncertainty in evolutionary optimization with an application to feedback control of combustion. *IEEE Transactions on Evolutionary Computation*, 13(1):180–197, 2009.
- [9] N. Hansen and A. Ostermeier. Completely derandomized self-adaptation in evolution strategies. *Evolutionary Computation*, 9(2):159–195, 2001.