2014 IEEE Congress on Evolutionary Computation (CEC)
July 6-11, 2014, Beijing, China

Identifying and Exploiting the Scale of a Search Space in Differential Evolution

James Montgomery
School of Engineering and ICT
University of Tasmania
Hobart, Australia
james.montgomery@utas.edu.au

Stephen Chen
School of Information Technology
York University
Toronto, Canada
sychen@yorku.ca

Yasser Gonzalez-Fernandez
School of Information Technology
York University
Toronto, Canada
ygf@yorku.ca

***Abstract*—Optimisation in multimodal landscapes involves two distinct tasks: identifying promising regions and location of the (local) optimum within each region. Progress towards the second task can interfere with the first by providing a misleading estimate of a region's value. Thresheld convergence is a generally applicable "meta"-heuristic designed to control an algorithm's rate of convergence and hence which mode of search it is using at a given time. Previous applications of thresheld convergence in differential evolution (DE) have shown considerable promise, but the question of which threshold values to use for a given (unknown) function landscape remains open. This work explores the use of clustering-based method to infer the distances between local optima in order to set a series of decreasing thresholds in a multi-start DE algorithm. Results indicate that on those problems where normal DE converges, the proposed strategy can lead to sizable improvements.**

I. Introduction

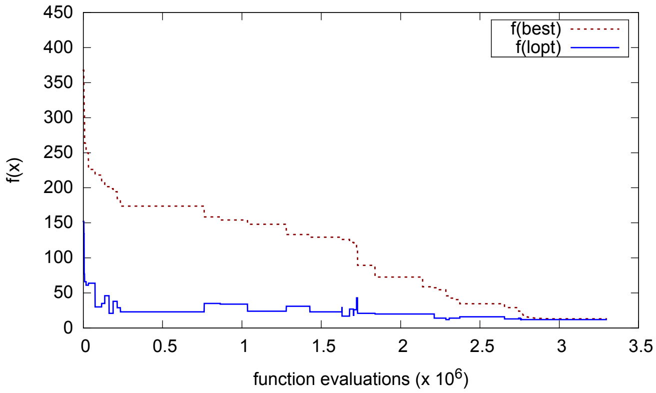
IT is a widely-held and oft-repeated belief that heuristic optimization algorithms must balance exploration and exploitation [1]. The role of exploration is to find the region with the best solutions, while the role of exploitation is to find the (local) optimum in a given region. However, it is unknown to what extent is is possible to genuinely balance these competing search behaviours, and if this balance involves separating initial exploration from later (exploitative) refinement of solutions, or an ongoing mixing of the two behaviours. In multimodal search landscapes, which most interesting real-world problems exhibit, it is clear that some search effort must be expended finding the most promising regions of the search space before those regions are explored/exploited in finer detail. Yet many population-based search heuristics do not gradually switch from exploration to exploitation but rather attempt to do both concurrently [2], [3]. Recent work suggests that this is not possible, and that any exploitation early in the search process makes exploration more difficult [3], [4].

The transition from global to local search is critical to the success of differential evolution (DE) [5]. A premature loss of diversity is a particular problem for the algorithm as its solution generation mechanism is based on the differences between solutions in solution space. While this reliance means the algorithm can be self-scaling—as the population converges the magnitude of exploratory moves also decreases—such convergence is typically irreversible, with the algorithm never again exploring widely. Indeed, DE can suffer from a cascading collapse in diversity, as when some solutions cluster in a good but small area they produce very small difference vectors that may 'recruit' other solutions to that same area [6], [7].

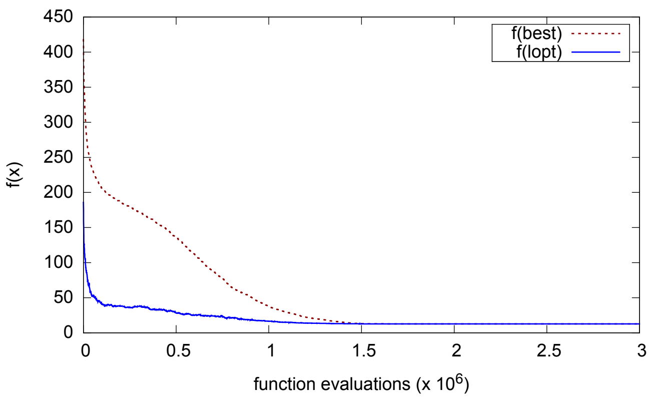
A recently proposed metaheuristic (in the sense that it is a general search principle rather than a specific search mechanism) is *thresheld convergence* (TC). The goal of thresheld convergence is to separate the processes of exploration and exploitation. Specifically, exploitation is "held" back by a threshold function. The allowed moves above the threshold size are more likely to be explorative (e.g., in new unexplored attraction basins), and the disallowed moves below the threshold size can potentially be exploitative (e.g., within a previously explored attraction basin). By preventing concurrent exploitation, the chances for the early discovery of local optima are reduced. With the subsequently decreased risk of premature convergence, it is expected that thresheld convergence can improve the performance of heuristic search techniques on multimodal fitness functions.

The first application of thresheld convergence in DE [3] was an efficient form of crowding that tested the distance between each new solution and the population member with which it would most likely cluster. Moves below a decaying threshold were disallowed. An improved version of DE with tresheld convergence (DE+TC) was introduced by Bolufé *et al.* [8]. That work introduced an adaptive threshold decay and 'pushed' new solutions outside the threshold boundary if they were too close, eliminating the wasted effort if a solution is merely discarded. Despite these improvements, the algorithm still worked best when the initial threshold was selected empirically, and the simple mechanism for adapting the threshold may not always be effective in selecting an appropriate threshold level. The present work investigates the use of clustering to estimate the distances between local optima and to set a sequence of diminishing thresholds for use by a multi-start version of DE+TC.

DE+TC is described in more detail in the next section. A cluster-based approach to inferring the distance between local optima is introduced in Section II, after which a multi-start DE+TC strategy with thresholds based on these inferred distances is described and evaluated in Section III. An improved version of the approach, whose restarts are less disruptive,



(a) From single run producing median quality final result



(b) Average over 51 runs

Fig. 1: Objective value of iteration (i.e., current global) best solution and its local optimum for 30D Rastrigin.

is examined in Section IV, followed by a discussion of its performance on difficult instances in Section V.

A. DE with Thresheld Convergence

As a metaheuristic there are many ways in which thresheld convergence may be integrated into DE. The approach used here was previously found [8] to be effective on the Black-box Optimisation Benchmark problems (see [9] for problem details). In the commonly used DE/rand/1/bin algorithm, a new candidate solution is generated in two steps: an intermediate solution is created by adding the weighted difference between two randomly selected population members to a third randomly selected member, called the *base*, then uniform crossover is performed between this intermediate solution and the *target* it may replace in the next generation. In DE+TC, if the distance between the candidate solution *new* and *base* is below the current threshold then the candidate is shifted away from *base* in the same direction (see Algorithm 1).

A key weakness of thresheld convergence in many earlier applications was the use of a scheduled (and problem-specific) threshold function to control the shift from exploration to exploitation. The best threshold function for each problem appears to depend on the size and spacing of the attraction

Algorithm 1 DE+TC solution modification rule

if $||new - base|| < threshold$ **then**
  $direction \leftarrow Normalize(new - base)$
  $new \leftarrow base + threshold \cdot direction$
end if

basins, which is often difficult to determine without specialised knowledge of a problem or by sampling the problem space. An innovation used by Bolufé *et al.* was an adaptive decay rule based on the following heuristic: if candidate solutions are being generated that are improvements on their parents then the threshold is likely set appropriately, whereas if no improvements are being made then it may be too high. The adaptive threshold update is defined by

$$threshold_i = \begin{cases} threshold_{i-1} & \text{if any replacements} \\ \beta \cdot threshold_{i-1} & \text{otherwise} \end{cases} \quad (1)$$

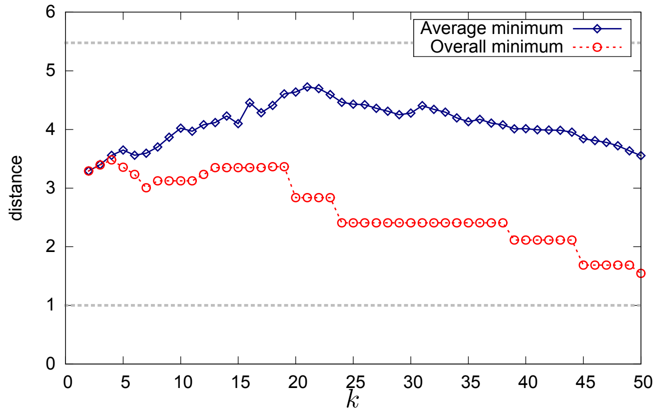
where i is the current generation and $\beta \in (0, 1]$ controls the threshold decay rate. The use of the adaptive threshold significantly reduced the algorithm's sensitivity to the initial threshold setting. Consequently, in all subsequent experiments described in this work that use DE+TC with an adaptive threshold, the initial threshold is set to 10% of the length of the main space diagonal and $\beta = 0.995$, which was found experimentally to work well.

II. Learning Inter-optima Distances in a Known Environment

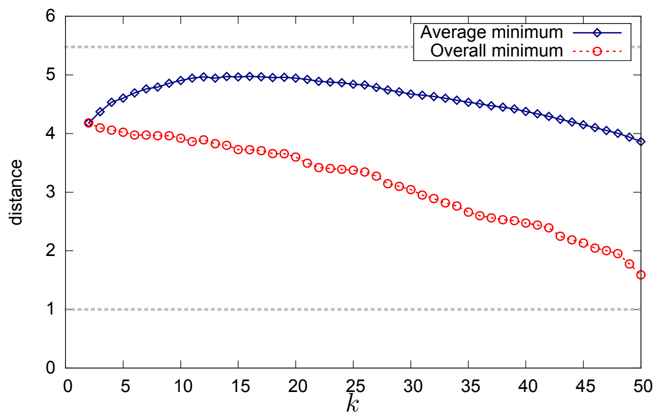
Previous work has shown that DE's search behaviour is characterised by an initial period of exploration followed by rapid convergence to a local optimum found during that exploratory phase (see, e.g., [3]) . Although population convergence is a necessary and useful feature of the algorithm, allowing it to reduce the scale of its exploratory moves, once the process begins it can lead to a cascading collapse in diversity. Identifying when the initial exploration phase will end can assist in implementing algorithm changes to promote population diversity when it is most needed.

The Rastrigin function [10] is a multimodal function constructed from the superposition of a cosine function with period 1 over the sphere function. It has many local optima whose positions can be found on integer values in the continuous search space, so it is possible to easily judge the quality of a solution's current basin of attraction and its relative quality within that basin. In this initial study, DE/rand/1/bin with $Cr = 0.9$, $F = 0.5$ and a population of 50 individuals was applied to Rastrigin in 30D and the quality of the generation best solution versus that of its local optimum measured over time.[1] The algorithm was allowed $10,000 \cdot D$ function evaluations

[1]Although Rastrigin is separable, which would suggest using a low value of Cr, as most real world problems are not, using a high value simulates DE's application to such functions more accurately. The value of $F = 0.5$ was found through an initial sensitivity analysis to produce good results on this function in 30D.



(a) From single run producing median quality final result



(b) Average over 51 runs

Fig. 2: Overall and average minimum distance between cluster centroids for $k \in [2, 50]$ in 30D Rastrigin. Lines are a visual guide only.

(FEs), equating to 6000 generations. Fifty-one randomised trials were performed.

Fig. 1 shows the objective value of the generation best solution and the value of its current local optimum. Part (a) is from the trial which produced the median quality final result, while (b) shows the average across all trials, showing their general trends over time. Although this combination of settings (with Cr fixed at 0.9 and population size fixed at 50 *a priori*) was found earlier to produce the best results compared to using other values for F—i.e., the final result appears to be as good as the algorithm can do—the best local optimum is found after 10^5 FEs (2000 generations), and the quality of the current best solution's local optimum changes little after the first 25,000 FEs (500 generations). A little beyond half way through the search the current best solution has been locally optimised to be at its (suboptimal) local optimum. Thus, the global search has largely stopped after one third of the total FEs, and the algorithm has transitioned to a local search phase for the remaining two thirds of its run. Intuitively such search behaviour looks suboptimal, as the algorithm does not require the majority of its function evaluations merely to refine one or more known good locations.

In multimodal landscapes, DE populations tend to divide among the various attraction basins present, up until the point that the population collapses rapidly to a single local optimum. Given that most of the improvement in the best solution's local optimum has been achieved after approximately 10% of allowed FEs, and given that DE tends to improve solutions individually within their local basins of attraction until the population's cascading collapse, it is plausible that the population is scattered across local optima, with solutions of similar quality relative to their respective local optima. Examining the distances between clusters of solutions may reveal some of the search landscape's structure.

K-means clustering was applied to the DE population after 10% of the total FEs (i.e., 600 generations), with the value of k varied from 2 to 50 (i.e., each individual as its own 'cluster'). For each value of k, the minimum inter-cluster distance was measured as well as the average over the minimum inter-cluster distances per cluster. Fig. 2 depicts these values for (a) the trial that produced the median quality final result and (b) an average for each value of k over 51 trials (to indicate general trends).

The change in the minimum inter-cluster distance is indicative of DE's typical pattern of convergence: the population has a tendency to contract as the search progresses, especially in solution spaces with global structure. The variation in the average over each cluster's minimum inter-cluster distance suggests that individuals in the population have found a number of distinct local optima. Thus this average minimum distance between clusters (avg-min) may be considered an approximation of the minimum distance between *distinct* local optima (i.e., the nearest local optima different in every dimension). If the approximation is sufficiently close to the true value, then the distance between local optima that differ in only one dimension may be inferred. Assuming regularly-spaced optima and a separable problem that is identical in each dimension, the smallest distance between local optima may be approximated by $avg\text{-}min/\sqrt{D}$. These two assumptions will not be valid for many real-world problems, but they represent a starting point before more information about a problem is learned.

In 30D Rastrigin the true minimum distance between distinct local optima is $\sqrt{30} = 5.48$, indicated by the top dashed line in Fig. 2, while the true minimum distance between local optima that differ in one dimension is 1. Based on studies with Rastrigin and other problems, this work elects to take the median value of the avg-min as an approximation of the distance between distinct local optima. Although on average with Rastrigin this measure will underestimate the distance between distinct optima (the error is approximately 14%), it provides a straightforward way to select from the varying values of avg-min as k changes (which are often not as smooth as the trends in Fig. 2 suggest).

This approach to approximating inter-optima distances was tested on two variants of Rastrigin, one with each dimension's range doubled and another in which each dimensions range was doubled but the period of the cosine function halved. In the first case the technique produced indistinguishable values to normal Rastrigin (the distances were, indeed, unchanged), while in the second it produced values twice as large. This suggests that the use of k-means clustering in conjunction with normal DE is somewhat robust to changes in scale.

Given this insight into the distances between local optima, thresheld convergence may be applied with less direct user control and greater confidence.

III. A Simple Restart Strategy

Previous versions of DE with thresheld convergence (collectively called DE+TC hereafter) operated over the course of an entire run and attempted to guide the algorithm smoothly through different granularities of search. Early versions used a parameterised decaying threshold function, while more recent advancements introduced auto-adaptation based on the algorithm's performance. In the present work a restart strategy is implemented, with each restart having a progressively lower, but fixed threshold. Given the algorithm's observed behaviour on Rastrigin, each run is divided into six stages: an initial phase of normal DE lasting 10% of total FEs, followed by cluster analysis to determine threshold values; four stages of 20% total FEs of DE+TC with successively lower, but otherwise fixed thresholds; and a final 'local search' stage of normal DE with a reduced population size of 25, lasting 10% of total FEs. Each of the four middle stages represents an algorithm restart, although some information about previously discovered good solutions is maintained between each. At the first restart, the population best solution is maintained, while the 10 best solutions are retained on each subsequent restart. The final stage uses a reduced population to give each population member a greater number of opportunities to be updated and to focus the search on the most promising region found. Reducing DE's population size as the search progresses has previously been found to be effective [11].

The four threshold values are derived as follows. K-means clustering is performed on the population after 600 FEs, for values of $k \in [2, 50]$. The median value of the average minimum inter-cluster distance is taken as the initial threshold value α, while the final threshold value is $\beta = \alpha/\sqrt{D}$. The two intermediate threshold values are $(\alpha + \beta)/2$, followed by $2 \cdot \alpha$.

A. Initial Results

Hereafter the techniques developed will be demonstrated with respect to Rastrigin and evaluated on the benchmark problems from the CEC2013 Single Objective Real Parameter Optimisation Competition, later referred to simply as the "CEC2013 benchmarks". The CEC2013 problems are defined in [12] and summarised in Table I. Set 2 is of particular interest, as these problems are multimodal and likely to have some degree of global structure—it is such problems that thresheld convergence was designed to suit. For both Rastrigin and the CEC2013 benchmarks 51 randomised trials were run for each algorithm tested.

TABLE I: CEC2013 Benchmarks

Category	f	Name
Unimodal Functions	1	Sphere Function
	2	Rotated High Conditioned Elliptic Function
	3	Rotated Bent Cigar Function
	4	Rotated Discus Function
	5	Different Powers Function
Basic Multimodal Functions	6	Rotated Rosenbrock's Function
	7	Rotated Schaffers F7 Function
	8	Rotated Ackley's Function
	9	Rotated Weierstrass Function
	10	Rotated Griewank's Function
	11	Rastrigin's Function
	12	Rotated Rastrigin's Function
	13	Non-Continuous Rotated Rastrigin's Function
	14	Schwefel's Function
	15	Rotated Schwefel's Function
	16	Rotated Katsuura Function
	17	Lunacek bi-Rastrigin Function
	18	Rotated Lunacek bi-Rastrigin Function
	19	Expanded Griewank's plus Rosenbrock's Function
	20	Expanded Scaffer's F6 Function
Composition Functions	21	Composition Function 1 (n=5,Rotated)
	22	Composition Function 2 (n=3,Unrotated)
	23	Composition Function 3 (n=3,Rotated)
	24	Composition Function 4 (n=3,Rotated)
	25	Composition Function 5 (n=3,Rotated)
	26	Composition Function 6 (n=5,Rotated)
	27	Composition Function 7 (n=5,Rotated)
	28	Composition Function 8 (n=5,Rotated)

Fig. 3 shows the average objective values of the generation best solution (i.e., population best solution) and the value of its current local optimum each generation for (a) DE+TC and (b) the restart strategy introduced here. Comparing these with Fig. 1(b) shows that standard DE+TC does prolong the period during which the quality of the best solution is not locally optimised, but the effect is very small, while DE+TC with restarts and set thresholds further prolongs the period of exploration. The average quality of the final result for DE, DE+TC and DE+TC with restarts were 12.82, 12.93 and 11.3, respectively. Comparing the outcomes with t-tests, the slight difference between DE and DE+TC is not statistically significant, while the observed improvement when restarts are added is statistically significant at the 5% level.

DE, DE+TC with adaptive threshold reduction, and DE+TC with restarts were applied to the CEC2013 benchmarks, with results presented in Table II. The relative difference (%-diff) is calculated as $(a-b)/\arg\max\{a, b\}$, where a is the average final result of DE and b is the average final result of the comparison algorithm. Hence, positive values indicate that a new technique performs better than standard DE. Differences between algorithm's distributions of final results were, in general, found to be normal or close to normal, so t-tests were conducted to compare them.

DE+TC with an adaptive threshold produces statistically significant improvements on four functions from set 2, with generally equivalent results to DE on most other functions. In comparison, the addition of thresholds selected by analysis of the search space yielded statistically significant and

TABLE II: Results of DE, DE+TC (with adaptive threshold decay) and the first version of multi-start DE+TC with uniform restarts applied to the CEC2013 benchmarks. Bold values indicate differences that are statistically significant at the 1% level.

func.	DE		DE+TC		%-diff	t-test	Multi-start DE+TC (uniform restarts)		%-diff	t-test
	mean	std dev	mean	std dev			mean	std dev		
1	0.00E+00	0.00E+00	8.08E-07	2.02E-07	**-100%**	0	0.00E+00	0.00E+00	0%	—
2	1.53E+05	6.82E+04	1.87E+05	8.56E+04	-18%	0.017	2.76E+05	1.46E+05	**-45%**	0.000
3	1.00E+07	1.27E+07	8.88E+04	5.72E+05	**99%**	0.000	1.87E+06	6.18E+06	**81%**	0.000
4	4.35E+02	4.03E+02	5.83E+02	4.54E+02	-25%	0.044	4.27E+02	2.44E+02	2%	0.448
5	1.18E+00	8.31E+00	9.48E-07	3.26E-07	100%	0.159	2.63E-05	1.44E-04	100%	0.159
6	2.03E+01	1.55E+01	9.34E+00	6.92E+00	**54%**	0.000	2.25E+01	2.01E+01	-10%	0.265
7	5.96E+00	5.85E+00	1.10E+00	2.15E+00	**82%**	0.000	3.07E+00	4.20E+00	**49%**	0.003
8	2.09E+01	4.85E-02	2.09E+01	5.65E-02	0%	0.397	2.09E+01	4.78E-02	0%	0.269
9	1.60E+01	5.87E+00	1.52E+01	3.87E+00	5%	0.204	1.81E+01	7.03E+00	-12%	0.052
10	1.31E-01	8.82E-02	7.43E-03	7.44E-03	**94%**	0.000	1.44E-02	1.47E-02	**89%**	0.000
11	1.55E+01	4.57E+00	1.30E+01	3.87E+00	**16%**	0.002	1.20E+01	4.99E+00	**23%**	0.000
12	7.64E+01	6.31E+01	9.47E+01	6.47E+01	-19%	0.077	1.24E+02	5.91E+01	**-39%**	0.000
13	1.41E+02	4.49E+01	1.47E+02	3.85E+01	-4%	0.256	1.48E+02	4.28E+01	-5%	0.203
14	9.72E+02	3.07E+02	1.05E+03	3.53E+02	-8%	0.110	1.13E+03	3.06E+02	**-14%**	0.005
15	7.21E+03	3.04E+02	7.21E+03	2.64E+02	0%	0.464	7.18E+03	2.71E+02	0%	0.291
16	2.52E+00	2.86E-01	2.45E+00	2.60E-01	3%	0.110	2.45E+00	3.32E-01	3%	0.128
17	5.54E+01	1.03E+01	5.79E+01	9.09E+00	-4%	0.095	6.13E+01	1.00E+01	**-10%**	0.002
18	1.93E+02	1.33E+01	1.93E+02	1.00E+01	0%	0.412	1.99E+02	8.81E+00	**-3%**	0.006
19	3.35E+00	2.11E+00	4.37E+00	3.26E+00	-23%	0.032	3.59E+00	2.34E+00	-7%	0.289
20	1.18E+01	4.20E-01	1.20E+01	3.55E-01	-1%	0.011	1.18E+01	3.47E-01	1%	0.145
21	2.84E+02	7.41E+01	2.83E+02	6.97E+01	0%	0.477	3.11E+02	7.92E+01	-9%	0.036
22	8.11E+02	3.01E+02	8.43E+02	2.63E+02	-4%	0.288	9.12E+02	3.14E+02	-11%	0.052
23	6.96E+03	5.42E+02	7.03E+03	4.00E+02	-1%	0.229	7.14E+03	2.77E+02	-3%	0.016
24	2.32E+02	1.00E+01	2.21E+02	1.49E+01	**5%**	0.000	2.35E+02	9.05E+00	-1%	0.081
25	2.56E+02	7.76E+00	2.48E+02	7.91E+00	**3%**	0.000	2.55E+02	7.55E+00	1%	0.183
26	2.53E+02	6.41E+01	2.60E+02	6.41E+01	-3%	0.301	2.37E+02	5.69E+01	7%	0.085
27	6.29E+02	7.60E+01	5.64E+02	9.51E+01	**10%**	0.000	6.04E+02	7.87E+01	4%	0.052
28	3.00E+02	2.76E-13	3.00E+02	5.38E-03	**0%**	0.000	3.00E+02	7.17E-04	**0%**	0.000

large improvements in three functions of set 2 and worked more poorly than DE+TC with adaptive threshold reduction on many other problems. However, it is worth noting that initial experiments (not included here) with higher initial thresholds produced very poor outcomes, indicating that the current adaptive rule (i.e., to decrease the threshold if no improvement in a generation) is not universally effective in identifying "appropriate" threshold values given an arbitrary initial threshold; direct observation of the search space can eliminate the need for selecting an initial threshold manually and yield better results if a poor choice were made. Moreover, inspection of the final population spread in a number of trials of the multi-start DE+TC revealed that the population had not converged, suggesting that the restart strategy employed was setting the search back too far. A more targeted restart strategy is described next.

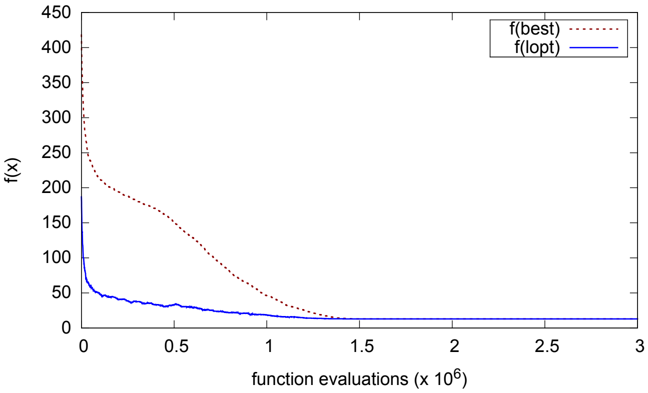
IV. Focused Restarts

A revised version of the multi-start DE+TC was developed in which the first restart uses uniform random initialisation of the new population as before, but each subsequent restart initialises solutions in a manner similar to UMDA [13], [14]. New solutions are initialised dimension-wise using a Gaussian $\mathcal{N}(\mu, \sigma)$ where μ is equal to the previous population's mean position and σ is equal to s multiplied by the previous population's standard deviation, where s is a parameter. Empirical testing (of $s \in \{0.5, 1, 2\}$) suggested that $s = 0.5$ produces the best results across the benchmark set.

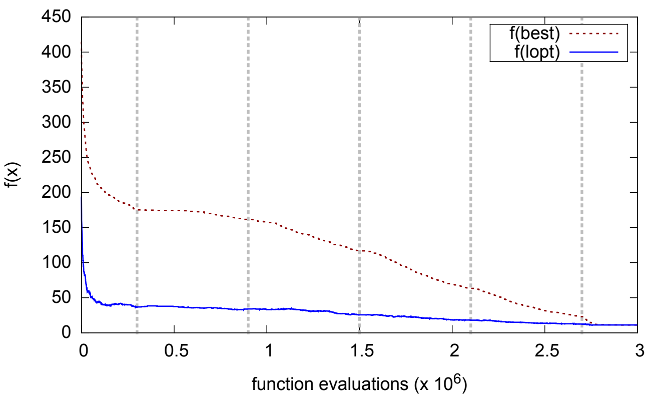
Table III shows the average result achieved by the multi-start DE with UMDA restarts with relative differences compared to DE. This version of the approach achieves statistically significant improvements on a number of problems from set 2 of the benchmarks, with more consistent results than DE+TC with adaptive threshold reduction. However, it achieves largely equivalent results on the rest of the set (none of the cases where its average performance is worse were statistically significant). This suggests that either the approach to deriving thresholds does not work with these problems or DE is a poor fit for these problems, issues explored next.

V. Discussion

Figs. 4 and 5 show the population spread for DE (blue) and multi-start DE+TC with UMDA initialisation (orange) for selected functions from set 2 where the multi-start DE+TC worked well (in 4) or poorly (in 5). Considering the cases where the proposed technique worked well, for those functions where DE normally converges the addition of TC extended the period of exploration, while the final phase allowed population convergence and hence, refinement of the final solution's quality. In the case of function 20, where the algorithm normally does not converge completely, the proposed approach, using UMDA reinitialisation, was able to focus its local search



(a) DE+TC with adaptive threshold reduction



(b) Multi-start DE+TC with uniform restarts

Fig. 3: Objective value of iteration (i.e., current global) best solution and its local optimum for 30D Rastrigin.

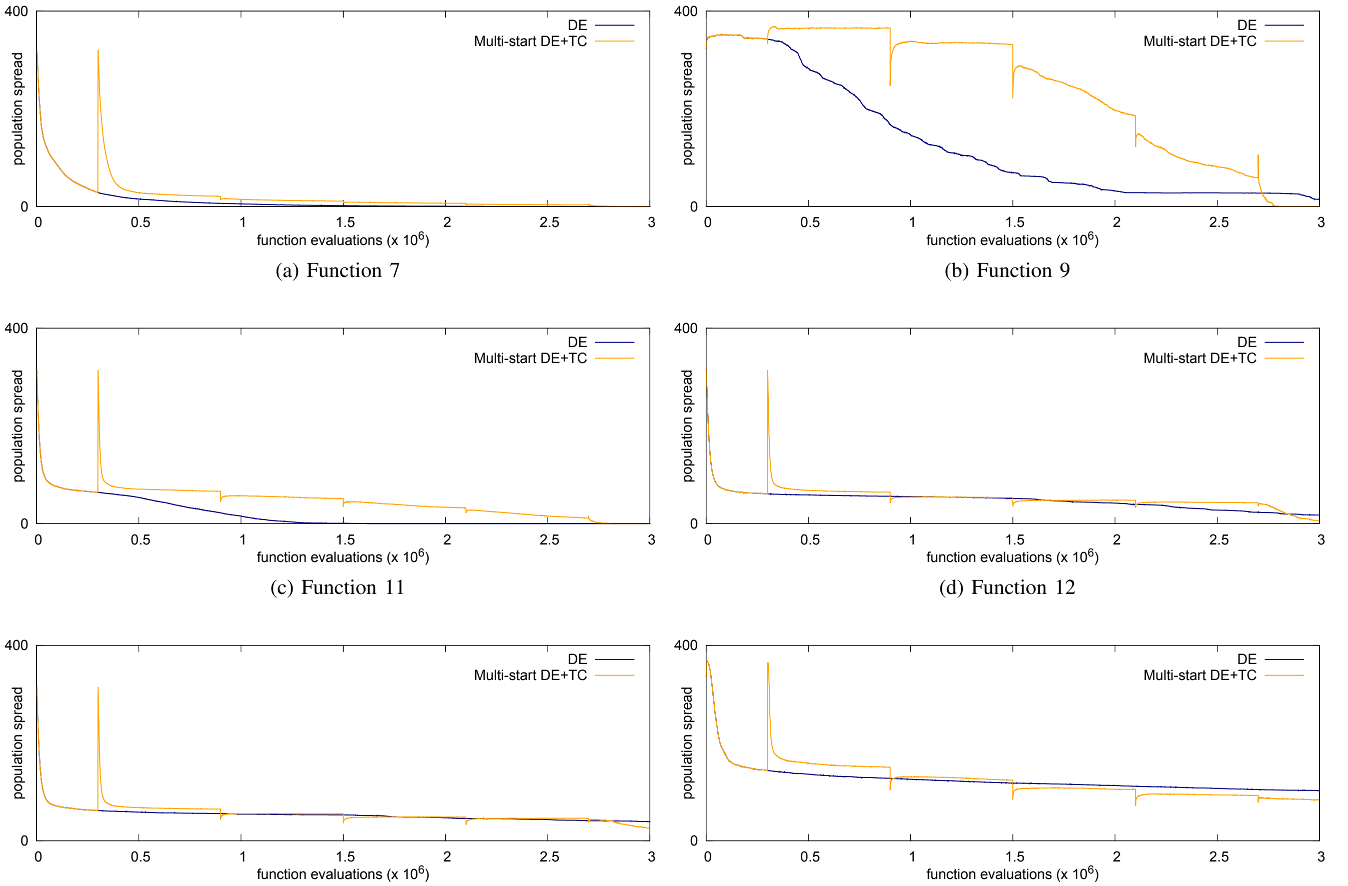
TABLE III: Results of multi-start DE+TC with UMDA-based restarts applied to the CEC2013 benchmarks. Bold values indicate differences that are statistically significant at the 1% level.

func.	Multi-start DE+TC (UMDA restarts)		%-diff	*t*-test
	mean	std dev		
1	0.00E+00	0.00E+00	0%	—
2	2.45E+05	1.09E+05	**-38%**	0.000
3	2.28E+06	5.13E+06	**77%**	0.000
4	3.66E+02	2.15E+02	16%	0.059
5	4.75E-06	2.26E-05	100%	0.159
6	1.49E+01	1.26E+01	27%	0.127
7	1.16E+00	1.71E+00	**81%**	0.000
8	2.10E+01	3.53E-02	0%	0.115
9	1.17E+01	3.57E+00	**27%**	0.006
10	1.35E-02	1.25E-02	**90%**	0.000
11	9.43E+00	2.37E+00	**39%**	0.000
12	4.12E+01	3.56E+01	**46%**	0.005
13	1.09E+02	5.94E+01	**23%**	0.003
14	1.06E+03	3.29E+02	-8%	0.048
15	7.14E+03	3.48E+02	1%	0.102
16	2.56E+00	2.55E-01	-2%	0.164
17	6.05E+01	1.19E+01	-8%	0.034
18	1.90E+02	9.21E+00	1%	0.099
19	4.06E+00	2.60E+00	-18%	0.138
20	1.16E+01	4.63E-01	**2%**	0.002
21	2.79E+02	6.65E+01	1%	0.500
22	9.24E+02	2.28E+02	-12%	0.031
23	6.89E+03	2.98E+02	1%	0.481
24	2.31E+02	1.10E+01	1%	0.034
25	2.52E+02	7.49E+00	2%	0.017
26	2.15E+02	4.13E+01	**15%**	0.000
27	5.81E+02	7.44E+01	**8%**	0.001
28	3.00E+02	4.74E-04	**0%**	0.000

in the final generations. Of note is that using $\sigma = 0.5 \cdot \mathit{sample\,std\,dev}$ actually reduces population spread at each restart, which is followed by a rapid period of expansion due to the thresheld convergence mechanism. Previously, Ali [15] found that during its earliest iterations, normal DE generates a large number of moves outside the search bounds, which are essentially wasted. Conceivably this can also occur when the population has contracted to a good region, as outward looking moves cannot find improved positions and are discarded. Since the rapid expansion observed here can only occur if new solutions at the threshold are improvements on the randomly reinitialised ones, it is plausible that the contraction imposed by the UMDA reinitialisation allows these outward looking moves to be successful again.

Fig. 5 shows convergence plots for functions where the new technique was ineffective. Notably, for functions 8, 15 and 16, normal DE does not converge, and the proposed approach performs largely equivalently. These functions exhibit a range of landscapes with some similarities: function 8 (rotated Ackley's) consists of a bumpy plain with a very small globally optimal region, while 15 (rotated Schwefel's) is deliberately deceptive, and 16 (rotated Katsuura) also has a very large number of local optima. The lack of convergence in DE on these problems may be in part because the functions are rotated (the DE/rand/1/bin variant used here is not rotationally-invariant), but it also suggests that the algorithm itself is ill-suited to these landscapes. DE's behaviour on function 8 is particularly illustrative: after a short time each population member has been improved within one of the many small local optima that litter the landscape, but thereafter the search flounders, unable to move any solution because no *improving* move can be generated. Not only does this mislead the clustering mechanism's estimate of the distance between local optima, but the thresholds have little to no impact on such a scattered population in any case. As thresheld convergence is designed to combat "premature convergence", which really means a loss in a population-based heuristic's ability to discover new regions to be locally optimised before its function evaluation limit is reached, if the population does not exhibit convergence then TC will not impact its behaviour.

Fig. 5(d) is for function 17 (rotated Lunacek bi-Rastrigin) where the multi-start DE+TC appeared to actively hinder the search (the equivalent plot for function 19 (expanded Griewank's plus Rosenbrock's function) shows a similar pattern). In both functions 17 and 19 diversity is maintained above the normal low level of DE, suggesting that the thresholds



(a) Function 7 (b) Function 9

(c) Function 11 (d) Function 12

(e) Function 13 (f) Function 20

Fig. 4: Mean distance between individuals and population centroid for selected CEC2013 benchmarks where multi-start DE+TC outperformed normal DE.

should have been lower than those selected by the clustering technique. Further work is required to (1) explore alternative mechanisms to identify the distances between optima, which may not use the underlying algorithm to sample points in the search space, and (2) determine suitable algorithmic changes to allow DE to continue exploration in landscapes where its search stalls without converging.

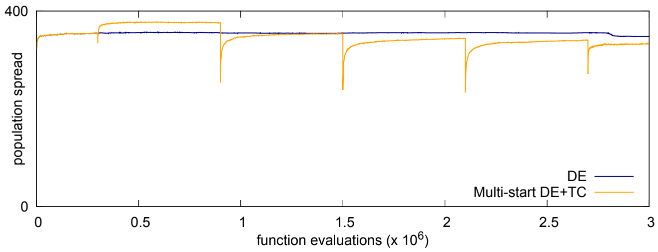
A. *Multi-start TC in other algorithms*

As in previous studies with thresheld convergence, its demonstration as a general technique has involved parallel studies in both differential evolution and particle swarm optimization (PSO). While previous studies have found roughly similar (performance-based) effects of thresheld convergence on both DE and PSO (see, e.g. [4], [8]), the current parallel study [16] has lead to highly divergent results. The most noticeable difference is in the effectiveness of clustering to measure the distance among local optima in the search space. Due to its use of attraction vectors, communicating particles in PSO are quickly and directly drawn towards the best positions of their neighbours. This leads to clear and distinct clusters, and these clusters can then lead to quite accurate estimates on the distances among attraction basins in the search space. Conversely, the population in DE appears to converge in a more homogeneous fashion, so the resulting information from clustering appears to be less useful.

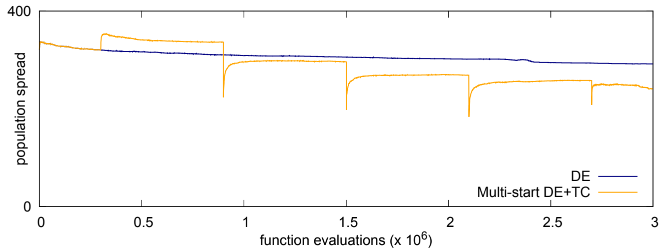
Another interesting divergence is that the best relative performance of adding the current implementation of thresheld convergence to DE occurs on function 10 where a relative improvement of 90% is achieved. Conversely, in the PSO study [16], function 10 led to some of the worst performances for the current implementation of thresheld convergence with negative effects of up to 90%. Going forward, we hope to use deeper analysis of this anomaly to better understand the search processes of DE and PSO in multimodal fitness landscapes and to perhaps build a hybrid method that leverages the advantages of each method (e.g., using the clustering effects of PSO to identify the scale of the search space, and then using the homogeneous search properties of DE to exploit this scale).

VI. Concluding Remarks

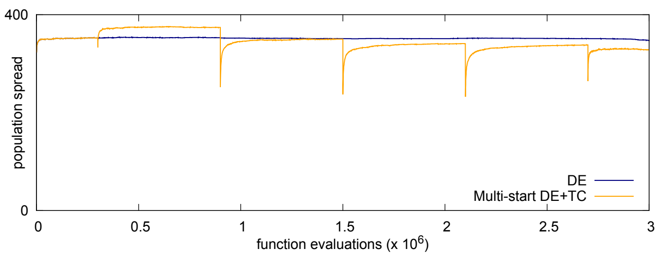
Thresheld convergence is a generally applicable technique for directly influencing the rate at which population-based



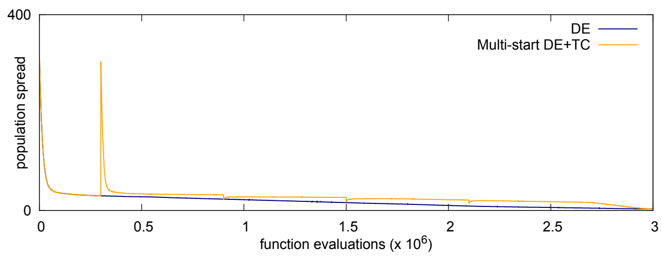
(a) Function 8



(b) Function 15



(c) Function 16



(d) Function 17

Fig. 5: Mean distance between individuals and population centroid for selected CEC2013 benchmarks where multi-start DE+TC performed (a–c) equavilently to DE or (d) more poorly than DE.

heuristics such as DE converge. Previous work has demonstrated its efficacy across a range of algorithms and problem benchmarks. This paper examined the use of a clustering-based approach to estimate the distances between optima and set thresholds appropriately in a multi-start DE+TC algorithm, with improvements over normal DE and a variant of DE+TC with adaptive threshold reduction on a number of multimodal problems. Future work will investigate alternative ways of estimating inter-optima distances and in identifying cases where the underlying DE algorithm is failing to converge (and hence when an alternative to TC is required to drive the search forward).

References

[1] M. Črepinšek, S.-H. Liu, and M. Mernik, "Exploration and exploitation in evolutionary algorithms: A survey," *ACM Computing Surveys*, vol. 45, no. 3, pp. 35:1–35:33, 2013.

[2] S. Chen and J. Montgomery, "A simple strategy to maintain diversity and reduce crowding in particle swarm optimization," in *AI2011*, 2011, pp. 281–290.

[3] J. Montgomery and S. Chen, "A simple strategy for maintaining diversity and reducing crowding in differential evolution," in *IEEE CEC*, 2012, pp. 2692–2699.

[4] S. Chen and J. Montgomery, "Particle swarm optimization with thresheld convergence," in *IEEE CEC*, 2013, pp. 510–516.

[5] K. Price, R. Storn, and J. Lampinen, *Differential Evolution: A Practical Approach to Global Optimization*. Berlin: Springer, 2005.

[6] J. Montgomery, "Differential evolution: Difference vectors and movement in solution space," in *IEEE CEC*, 2009, pp. 2833–2840.

[7] ——, "Crossover and the different faces of differential evolution searches," in *IEEE CEC*, 2010, pp. 1804–1811.

[8] A. Bolufé-Röhler, S. Estévez-Velarde, A. Piad-Morffis, S. Chen, and J. Montgomery, "Differential evolution with thresheld convergence," in *IEEE CEC*, 2013, pp. 40–47.

[9] N. Hansen, S. Finck, R. Ros, and A. Auger, "Real-parameter black-box optimization benchmarking 2009: Noiseless functions definitions," INRIA, Technical Report RR-6829, 2009.

[10] L. A. Rastrigin, *Systems of Extremal Control*. Nauka, 1974.

[11] J. Brest and M. Sepesy Maučec, "Population size reduction for the differential evolution algorithm," *Applied Intelligence*, vol. 29, no. 3, pp. 228–247, 2008.

[12] J. Liang, B. Qu, P. Suganthan, and A. Hernández-Díaz, "Problem definitions and evaluation criteria for the cec 2013 special session on real-parameter optimization," Computational Intelligence Laboratory, Zhengzhou University, Technical Report 201212, 2013.

[13] P. Larrañaga, R. Etxeberria, J. A. Lozano, and J. M. Peña, "Optimization by learning and simulation of bayesian and gaussian networks," University of the Basque Country, Spain, Technical Report EHU-KZAA-IK-4/99, 1999.

[14] H. Mühlenbein, "The equation for response to selection and its use for prediction," vol. 5, no. 3, pp. 303–346, 1997.

[15] M. M. Ali, "Differential evolution with preferential crossover," *Eur. J. Oper. Res.*, vol. 181, no. 3, pp. 1137–1147, 2007.

[16] Y. Gonzalez-Fernandez and S. Chen, "Identifying and exploiting the scale of a search space in particle swarm optimization," in press.