# Multimodal Optimization by Covariance Matrix Self-Adaptation Evolution Strategy with Repelling Subpopulations

*Ali Ahrari, Kalyanmoy Deb, Mike Preuss*

COIN Report Number 2016024

## Important Notice

This is the pre-print version of the published paper [1] in "*Evolutionary Computation*". Major revisions were performed in the subsequent stages. Please use the published version for citation, which is available at http://www.mitpressjournals.org/toc/evco/0/ja . If you are interested in this work but do not have access to the published version, please contact the corresponding author.

[1] Ahrari, A., Deb, K., & Preuss, M. (2016). Multimodal Optimization by Covariance Matrix Self-Adaptation Evolution Strategy with Repelling Subpopulations. *Evolutionary Computation*, in press, doi:10.1162/EVCO_a_00182

# Multimodal Optimization by Covariance Matrix Self-Adaptation Evolution Strategy with Repelling Subpopulations

**Ali Ahrari** ahrarial@msu.edu
Department of Mechanical Engineering, Michigan State University, East Lansing, MI 48825, USA

**Kalyanmoy Deb** kdeb@egr.msu.edu
Department of Electrical and Computer Engineering, Michigan State University, East Lansing, MI 48825, USA

**Mike Preuss** mike.preuss@cs.tu-dortmund.de
Department of Computer Science, TU Dortmund University, Germany

**Abstract**

During the recent decades, many niching methods have been proposed and verified on some available test problems. However, they often rely on some particular assumptions associated with distribution of minima, or shape and size of the basins, which can seldom be made in practical optimization problems. A novel niching technique based on repelling subpopulations is developed in this study which does not make any of these assumptions. This technique is incorporated into the Covariance Matrix Self-Adaptation Evolution Strategy (CMA-ES), a potent global optimization method to enable parallel convergence to different minima. The resultant method, called the covariance matrix self-adaptation with repelling subpopulations (RS-CMSA), is assessed and compared to several state-of-the-art niching methods on a standard test suite for multimodal optimization. An organized procedure for parameter setting is followed which assumes a rough estimation of the desired/expected number of minima is available. Performance sensitivity to the accuracy of this estimation is also studied by introducing the concept of robust mean peak ratio. Based on the numerical results, RS-CMSA emerged as the most successful method when robustness and efficiency is considered at the same time.

**Keywords**

Niching, Adaptive normalized niche radius, Mahalanobis distance metric, Robust mean peak ratio, CEC2013 test suite.

## 1 Introduction

In recent decades, stochastic optimization by using meta-heuristic methods has gained wide application in different fields of applied sciences, where the problem landscape may exhibit challenging features such as multimodality, discontinuity, ill-condition and correlation. On the other hand, practical considerations have evolved flexibility of optimization methods in order to solve problems in which more than one objective is pursued (multi-objective optimization), uncertainties exist (robust optimization) or distinct good solutions are desired (multimodal optimization). Finding a set of good solutions, instead of a single one, may provide a verity of distinct yet reasonable alternatives

for the decision maker, who can consider some other factors that were probably overlooked in mathematical modeling of the problem, to select the final solution. Moreover, finding all high-fitness optima might be inherently critical. A typical example is finding all resonance frequency that lead to high vibration amplitude (Das et al., 2011).

Multimodal optimization is usually rendered using a diversity preservation strategy, called *niching*, appended to a global optimization method, which we call the *core algorithm*, to enable parallel convergence to different minima. Even when a single minimum is sought, niching can still be utilized to prevent premature convergence. Niching is also an inevitable part of multi-objective optimization methods. Regardless of the core algorithm, niching strategies can be classified into two groups: radius-based and non-radius-based (Stoean et al., 2010).

### 1.1 Radius-based Methods

Radius-based niching methods are those that rely on a distance threshold, generally referred to as *niche radius*, to check whether two individuals share the same niche. The oldest technique in this group is fitness sharing (Singh and Deb, 2006), which reduces the fitness of similar individuals that share a niche. A similar idea is followed in some other techniques including clearing (Das et al., 2011; Singh and Deb, 2006), in which inferior individuals within the niche distance are eliminated and also clustering.

In most radius-based techniques with fixed niche radius, default value of the niche radius is set based on recommendation of Deb and Goldberg (1989), which assumed the niches are placed almost uniformly distributed in the entire search space. Beasley et al. (1993) used a sequential niching technique according to which the fitness function in the region of already found solutions is deteriorated so that subsequent restarts avoid such regions. Similarly, the technique relies on definition of threshold distance within which the fitness function is degraded.

In grenade explosion method (GEM) (Ahrari et al., 2009; Ahrari and Atai, 2010), several subpopulations explore the search space at the same time. Selection and recombination are rendered locally (within the subpopulations). Interaction among subpopulations is limited to the sampling step, which forces all individuals of a subpopulation to lie sufficiently far from the center of superior subpopulations. The required distance is gradually decreased to let the subpopulations converge to the basins that are close to each other. The main problem of the method is high amount of tuning effort.

Species conserving genetic algorithm (Li et al., 2002) iteratively identifies distant fittest individuals as seeds and allocates close individuals, with respect to a threshold distance, to that seeds. Stoean et al. (2007) improved this strategy by proposing topological species conservation (TSC), which checks whether two individuals share the same basin. This task is rendered by calling a function named *DetectMultimodal*, which was introduced by Ursem (1999). Although the dependency on user-tuned niche radius parameter was eliminated, the iterative calls of the mentioned function appeared costly. Additionally, the performance of the proposed algorithm highly depends on well-tuning of some parameters including mutation and crossover rates. The same authors revised the method later (Stoean et al., 2010), but the challenge of performance sensitivity to the control parameters remained unsolved.

### 1.2 Radius-free methods

Radius-free niching methods, in contrast, do not depend upon definition of the threshold distance and hence, can be considered more robust and desirable than radius-based methods. The oldest method in this group is crowding, in which a descendant replaces

the most similar parent. Mengshoel and Goldberg (2008) performed a comprehensive theoretical study on deterministic and stochastic crowding in genetic algorithms. A similar idea is followed in restricted tournament selection, where offspring compete with the closest among the $k$ randomly selected individuals (Das et al., 2011).

Li (2010) proposed different variants of particle swarm optimization (PSO) with ring topology by restricting interaction only to the neighbor particles. In distance-based locally informed particle swarm (LIPS) (Qu et al., 2013), particles are relocated based on information from their nearest neighbors (measured in terms of Euclidean distance). The method requires only population size to be tuned and was demonstrated to surpass 9 other niching methods, including lbest-PSO, proposed by Li (2010). On higher dimensions, only LIPS and the niching covariance matrix adaptation (NCMA) (Shir et al. 2010) could detect a reasonable fraction of the desired optima. A quite similar niching technique was pursued by Qu et al. (2012) to perform local mutation in neighborhood-based species-based differential evolution (NSDE), which demonstrated promising results when compared to earlier niching methods. Biswas et al. (2014) proposed parent-centric normalized mutation with proximity-based crowding differential evolution (PNPCDE), which probabilistically selects parents based on their similarities, measured in terms of their Euclidean distance. Their comparison with different variants of PSO and DE demonstrated considerable improvement.

Some recent studies render multimodal optimization by defining another objective. In shifting balance genetic algorithm (Wineberg and Chen, 2004), several small subpopulations (colonies), search the space around a larger subpopulation (core). The second objective ensures that colonies remain sufficiently far from the core, and thus searching new regions. The method was proposed merely to boost diversity, and multimodal optimization was not pursued. Bi-objective multi-population genetic algorithm (BMPGA) (Yao et al., 2010) defines minimization of the norm of the gradient vector as the second objective. The algorithm requires derivatives of the objective function, calculation of which might be more costly than the objective function itself, even if it is mathematically defined. Performance of the method depends on a user-tuned problem-dependent convergence threshold parameter. Deb and Saha (2012) developed a bi-objective GA for multimodal optimization in which the second objective was minimization of the derivatives of the original function. A heuristic to avoid analytical calculation of the derivatives of the objective function was employed and the method was validated on problems having up to 10 variables, however, in the employed test problems, distribution of the global minima were quite uniform and their attraction regions were roughly circular, which might make them relatively easy problems. The second objective in multimodal optimization with bi-objective DE (MOBiDE) (Basak et al., 2013) is maximization of the average distance to other solutions in the search space. The method requires only population size to be tuned by the user and the recommended values for other parameters were demonstrated to work well on a comparatively large test suite which included some hard composite test problems. Bandaru and Deb (2013) applied a parameter-free bi-objective GA for multimodal optimization, in which the second objective was increasing diversity of the individuals. Despite its robustness, it ranked 7 among 15 methods participated in the CEC2013 special session on niching methods (Li et al., 2013a).

Covariance matrix adaptation with adaptive niching (NCMA) (Shir et al., 2010) was shown to be successful in finding high-fitness local minima as well as the global optimum. The Mahalanobis distance metric, a scaled version of the Euclidean distance metric, was preferred to identify limits of a niche while the niche radius was cou-

pled to the mutation strength. However, it is not clear why they always used (1,10) or (1+10)-ES, even for highly multimodal functions, especially knowing that CMA-ES needs a rather large population size to cope with multimodal functions efficiently. Moreover, their employed variance effective selection mass ($\mu_{\text{eff}}$) does not coincide the recommended value of $0.25\lambda$ (Hansen and Ostermeier, 2001; Beyer and Sendhoff, 2008). Preuss (2010) incorporated the concept of nearest-better clustering (NBC) to CMA-ES. The resulting algorithm (NBC-CMA), identifies basins by finding the nearest better solution and cutting the longest edges to separate them into different clusters. NBC-CMA and its later variant, called NEA2 (Preuss, 2012), were demonstrated to be successful when tested on problems of higher dimensions, however, in the employed experimental setup, the goal was to find a single global minimum. NEA2 also ranked one among the niching methods participated in CEC2013 special session for niching methods (Li et al., 2013a). Particularly, for composite functions in higher dimensions, it outperformed the closest competitors by a clear margin.

### 1.3 Critiques on multimodal optimization

The main advantage of EAs over deterministic methods is their robustness with respect to the assumptions on the fitness landscape of the objective function. Similarly, in multimodal optimization, algorithms with minimal assumptions on properties of the fitness landscape are highly desirable and hence, radius-free niching methods show a significant advantage over those that depend on a user-tuned niche radius parameter. In general, even fine-tuning of the niche radius might be of little use, especially when distribution of minima is not uniform or basins are of dissimilar shape and size. Other control parameters such as window size in restricted tournament or crowding factor in crowding (Das et al., 2011) does not make such explicit assumptions on the fitness landscape.

Regardless of the underlying idea, niching methods exploit some distance measures which determine whether two solutions share the same niche. Quite often, the Euclidean distance is measured to group individuals to different niches, according to which closer individuals are likely to share the same basins, or equivalently, farther individuals belong to different niches. This strategy implicitly presumes that niches are roughly spherical; otherwise it can be misleading. Figure 1 illustrates this challenge for a typical case, where basins have a moderate condition number of 100. Considering ill-condition problems are one of the main challenges in real-parameter optimization (Hansen et al., 2009), exploitation of the Euclidean distance metric may drastically deteriorate reliability of the results. As an alternative, some studies utilized the Mahalanobis distance metric (Shir and Bäck, 2009; Shir et al., 2010), which may handle this problem to great extent, even though in general basins might have any arbitrary shape.

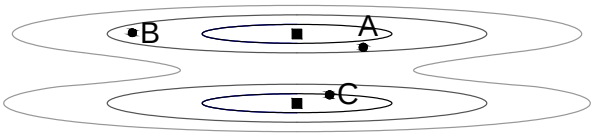



Figure 1: Contour lines around two minima (bold squares). Point C is closer to point A than point B, but A and B share the same basin, while C belongs to a different basin.

Another shortcoming is lack of well-developed comprehensive test suites for analyzing pros and cons of different niching methods. Most commonly used niching

benchmark functions are quite primitive. The most critical point is the low dimension of the search space, probably because these functions were originally defined for 1D or 2D space. Some popular examples are Branin, Himmelblau and Six Hump functions, which have frequently been used to validate niching methods (Ahrari and Atai, 2010; Wang et al., 2012; Schoeman and Engelbrecht, 2010; Liu et al., 2012; Liang and Leung, 2011; Roy et al., 2013; Li et al., 2012; Epitropakis et al., 2011). A few of these functions are scalable, such as the Shubert the or Vincent function, however the number of global minima exponentially grows with respect to the problem dimension which practically hinders employing dimensions more than 3. In such low dimensions, the search space can be exhaustively searched, and hence, results from such evaluations can hardly be generalized to higher dimensions, which are of significant practical interest. Considering multimodal optimization is a generalization of global optimization, niching test problems should reflect a variety of difficulties that are encountered in global optimization, including correlation and ill-condition (Hansen et al., 2009), in addition to those peculiar to multimodal optimization such as non-uniform distribution of minima.

In attempt to introduce more challenging benchmarks for multimodal optimization, Qu and Suganthan (2010) introduced a procedure to pose functions with several global minima using weighted sum of some basic functions, where weights are determined based on the distance to the global minimum of the basic functions. This method enables formation of basins with different shape and size while the number of global minima is equal to the number of the basic functions. The drawback is rapid growth of the computation cost of a evaluation when more minima is demanded, since each basic function should be called independently. This method was also employed to compose some niching problems of higher dimensions in CEC2013 workshop on niching methods (Li et al., 2013b).

Although finding high-fitness local minima might be practically interesting, the criteria specifying whether a local minimum is desirable should be provided for the method. For example, asking the method to find the best $\widetilde{N}_{\mathrm{gopt}}$ minima, or local optima within a specific tolerance of the global optimum value. Nevertheless, most available niching methods do not show such flexibility and numerical results commonly aim at finding only global minima.

In this study, a novel niching method based on repelling subpopulations is introduced which aims at alleviating the enumerated shortcoming of most available niching methods. The method, called covariance matrix self-adaptation with repelling subpopulation (RS-CMSA), is elaborated in section 2. In section 3, a descriptive experiment is performed to demonstrate how the method can tackle the challenge of differently sized basin. In section 4, RS-CMSA is compared with some of the state-of-the-art niching methods on CEC2013 test problems for multimodal optimization. Finally, conclusions are drawn in section 5.

## 2 The proposed niching method

Different parts of the proposed method are explained in this section.

### 2.1 The core algorithm

The core algorithm of the proposed method is a variant of CMA-ES, which falls in the category of evolution strategies that adapt the full covariance matrix of multivariate normal distribution (Hansen and Ostermeier, 2001). The adaptation process consists of adaptation of the global mutation strength, the so called step size, $\sigma_{\mathrm{mean}}$, based on the concept of cumulation, and a few other heuristics to update the covariance matrix ($\mathbf{C}$).

In a variant of CMA-ES, called CMSA-ES (Beyer and Sendhoff 2008), the mutative step size control replaced the cumulative step size adaptation and a simplified procedure to adapt the covariance matrix was proposed. The new variant was demonstrated to relatively outperform the original CMA-ES in some multimodal problems, however, in ill-conditioned problems, where efficient adaptation of the covariance matrix plays the critical role, it falls behind CMA-ES. Nevertheless, CMSA has a significant advantage from the practical point of view: It relies on less assumptions on the problem, which makes it more flexible. For example, it can employ intermediate selection schemes, in which a fraction of parents may survive to the next generation. Employing such strategies in CMA-ES violates the assumptions on distribution of samples, upon which the cumulative step size adaptation relies and thus, a performance decline can be assumed. CMSA-ES was preferred in this study and employed as the core algorithm because of the mentioned advantages,

### 2.2 Main niching ideas

In the proposed niching method, the population size is divided into $N_\mathrm{s}$ subpopulations of size $\lambda$ ($\mathbf{P}_i$, $i=1,2,\ldots,N_\mathrm{s}$), which explore the search space in parallel. Every subpopulation has its own mutation parameters ($\sigma_{\mathrm{mean}_i}$, $\mathbf{C}_i$), center ($\boldsymbol{x}_{\mathrm{mean}_i}$) and elite members. Solutions of each subpopulation must maintain sufficient distance from some specific points in the search space, called the taboo points. Taboo points are either previously identified minima (stored in an array called **Archive**), or the center of other subpopulations, if they are not worse than $\mathbf{P}_i$. $\mathbf{P}_i$ is better than $\mathbf{P}_p$ or an archived point ($\boldsymbol{y}_m$), if the best solution of $\mathbf{P}_i$ is better than the best solution of $\mathbf{P}_p$ or $\boldsymbol{y}_m$ respectively. A sampled solution is taboo acceptable if it satisfies the distance criterion with respect to all taboo points. The rest of the samples are discarded without evaluation. This process goes on until $\lambda$ taboo acceptable solutions are generated. The overall effect of this rejection is reshaping distribution of solutions so that subpopulations do not search previously explored regions or regions which are being explored by other subpopulations at the same time. The niching strategy affects only the sampling step, which means selection, recombination and adaptation are performed locally. Therefore, subpopulations may converge at different times and may have totally different strategy parameters, which allows for identification of basins with different shape and size.

The Mahalanobis distance metric, a scaled version of the Euclidean distance metric, is employed to check whether a recently generated solution is sufficiently far from the taboo point. The Mahalanobis distance ($d_{ij-k}$) of the *j*-th solution ($\boldsymbol{x}_{ij}$) of $\mathbf{P}_i$, sampled from normal distribution with covariance matrix $\mathbf{C}_i$ and standard deviation $\sigma_{\mathrm{mean}_i}$, to the *k*-th taboo point $\boldsymbol{y}_k$, is defined as follows:

$$d_{ij-k} = \frac{(\boldsymbol{x}_{ij} - \boldsymbol{y}_k)^T \mathbf{C}_i^{-1} (\boldsymbol{x}_{ij} - \boldsymbol{y}_k)}{\sigma_{\mathrm{mean}_i}} \tag{1}$$

Individuals of each subpopulation should be sufficiently far from the taboo points, but not too far, such that regions close to taboo points can be explored in future. $d_{ij-k}$ in Equation 1 increases inversely proportional to $\sigma_{\mathrm{mean}_i}$, which means for similar values of absolute distance ($|\boldsymbol{x}_{ij} - \boldsymbol{y}_k)|$), the normalized distance $d_{ij-k}$ increases as the subpopulation converges. This can be beneficial for landscapes with multiple global minima close to each other. The probability of rejection of solution $\boldsymbol{x}_{ij}$ due to proximity to taboo point $\boldsymbol{y}_k$ is defined as follows:

$$P_{\text{trej}}(\boldsymbol{x}_{ij}, \boldsymbol{y}_k) = \begin{cases} \exp\left(1 - \left(\frac{d_{ij-k}}{\hat{d}_k}\right)^2\right) & \text{if} \quad d_{ij-k} > \hat{d}_k \\ 1 & \text{if} \quad \text{else} \end{cases} \tag{2}$$

where $\hat{d}_k$ is the normalized taboo distance of the $k$-th taboo point ($\boldsymbol{y}_k$), which is a property of the taboo point. According to Equation 2, if the normalized distance is smaller than the normalized taboo distance, the offspring is always rejected, otherwise rejected with a probability which rapidly decreases as $d_{ij-k}$ increases. Note that any single taboo point can reject $\boldsymbol{x}_{ij}$. The overall taboo rejection probability is thus computed as follows:

$$P_{\text{Trej}}(\boldsymbol{x}_{ij}) = 1 - \prod_{k=1}^{K} (1 - P_{\text{trej}}(\boldsymbol{x}_{ij}, \boldsymbol{y}_k)) \tag{3}$$

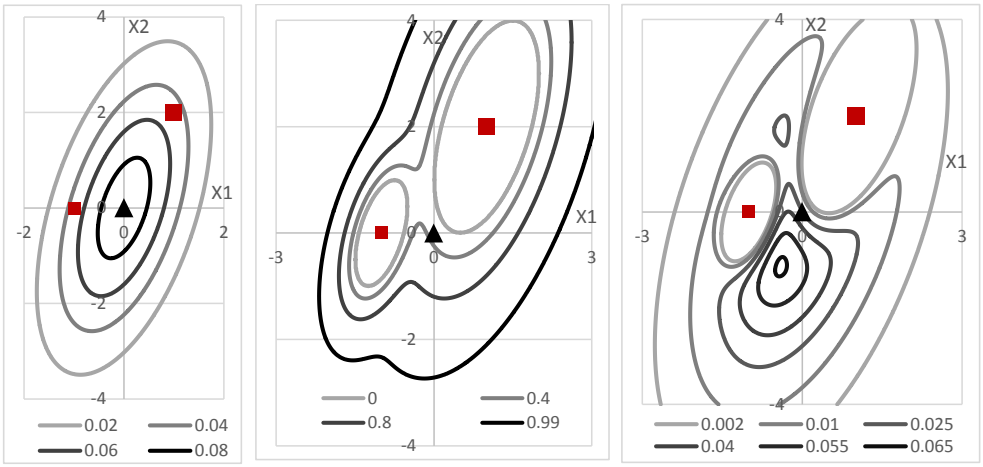
where $K$ is the number of the taboo points. $\boldsymbol{x}_{ij}$ is accepted and evaluated if it satisfies distance conditions with respect to all taboo points, otherwise it is rejected and a new candidate solution is generated. This process continues until $\lambda$ solutions are accepted.

According to Equation 2, the taboo distance is coupled with $\sigma_{\text{mean}_i}$ and $\mathbf{C}_i$, and thus, the challenge of tuning any fixed problem-dependent distance metric is eliminated. The shape of the taboo region is defined by the covariance matrix, which is adapted to conform to the shape of the basin, to which the subpopulation is converging. This means the taboo region gradually conforms to the basin shape to overcome the challenge of non-spherical basins, which was highlighted in section 1.2.

Rejection of some sampled solutions because of proximity to a taboo point reshapes the subpopulation distribution, which is the underlying idea of the proposed niching method. A typical case is illustrated in Figure 2, where the subpopulation $\mathbf{P}$ ($\mathbf{C} = \left[\begin{smallmatrix} 1 & 1 \\ 1 & 4 \end{smallmatrix}\right]$, $\sigma_{\text{mean}}$=1, $\boldsymbol{x}_{\text{mean}}$=[0 0]) approaches two taboo points ($\boldsymbol{y}_1$ and $\boldsymbol{y}_2$). The normalized taboo distance of taboo points are $\hat{d}_1 = 1$, $\hat{d}_2 = 0.5$, where the larger square represents $\boldsymbol{y}_1$.

Figure 2a depicts iso-density contour lines of sampling distribution, which is not affected by presence of the taboo points, however, some of the sampled solutions will be rejected without being evaluated. Lines with equal value of taboo acceptance probability ($P_{\text{Tacc}} = 1 - P_{\text{Trej}}$) are illustrated in Figure 2b. The ellipses which circumscribe the areas where no solution is accepted are detectable. Taboo regions conform to $\sigma_{\text{mean}}$ and $\mathbf{C}$; however, their size is also affected by $\hat{d}_k$, which is a property of the taboo point, and therefore, the taboo region associated with $\boldsymbol{y}_1$ is twice the size of the other one. $P_{\text{Tacc}}$ grows quite fast as the distance to the taboo points increases. Iso-density contours of generating taboo acceptable solutions are illustrated in Figure 2c. High density areas are those that have a high density of sampling (close to the center of the subpopulation) *and* a high taboo acceptance probability (far from the taboo points). Regions with maximum density of generating taboo acceptable solutions have shifted from $\boldsymbol{x}_{\text{mean}}$ towards the left and the bottom, where distance to the taboo points increases. This means the subpopulation is pushed away from the taboo points, however, whether it actually moves that way depends on the fitness landscape of the objective function as well. Regions with minimum density of generating taboo acceptable solutions are those that are either far from $\boldsymbol{x}_{\text{mean}}$ or close to the taboo points.

The evaluated solutions are sorted based on their fitness, and the $\mu$-best sampled solutions are selected to update the strategy and object parameters of $\mathbf{P}_i$ ($\boldsymbol{x}_{\text{mean}_i}$, $\sigma_{\text{mean}_i}$ and $\mathbf{C}_i$), quite similar to the original CMSA-ES (Beyer and Sendhoff, 2008):




(a) (b) (c)

Figure 2: Subpopulation **P** in proximity of two taboo points (square). The triangle represents $\boldsymbol{x}_{\text{mean}}$ a) Iso-density contour lines of sampling distribution. b) contour lines with equal value of $P_{\text{Tacc}}$. c) Iso-density contour lines of generating taboo acceptable solutions.

$$\mathbf{C}_i \leftarrow \left(1-\frac{1}{\tau_c}\right)\mathbf{C}_i + \frac{1}{\tau_c}\sum_{j=1}^{\mu} w_j \left(z_{ij}^T z_{ij}\right) \tag{4a}$$

$$\sigma_{\text{mean}_i} \leftarrow \left(\prod_{j=1}^{\mu} \sigma_{ij}^{w_j}\right) / \left(\prod_{j=1}^{\lambda} \sigma_{ij}^{\frac{1}{\lambda}}\right) \tag{4b}$$

$$\boldsymbol{x}_{\text{mean}_i} \leftarrow \sum_{i=1}^{\mu} w_j \boldsymbol{x}_{ij} \tag{4c}$$

In Equation 4, $\mu = \lfloor \lambda/2 \rfloor$, $w_j$'s are logarithmically decreasing weights (Hansen and Ostermeier, 2001) and $\tau_c = 1+2D(D+1)/\lambda$ is the adaptation interval constant for $\mathbf{C}_i$ (Beyer and Sendhoff, 2008). If there are bounds on the search space, $\boldsymbol{x}_{\text{mean}_i}$ is relocated to the nearest point inside the search space.

### 2.3 Restarts with increasing population size

Optimum population size for CMA-ES and CMSA-ES varies from a small value ($\lfloor 4+3\ln(D) \rfloor$) for unimodal functions to several hundreds for highly multimodal ones. Auger and Hansen (2005) employed the restart strategy with increasing population size (IPOP-CMA-ES) to overcome the challenge of parameter tuning. Starting with the minimum recommended value, the population size is doubled after each restart. The goal is to reach a restart with sufficiently large population which can identify the global minimum of highly multimodal functions. The drawback emerges when a large population size is required, and a considerable number of function evaluations should be wasted to reach this goal. More sophisticated schemes were later proposed to gain more

feedback from the previous restarts (Wessing et al., 2011) and to boost local search capabilities of the method (Hansen, 2009). The method proposed in another study (Ahrari and Shariat-Panahi, 2013) adapts the population size based on variation of the fitness of the recombinant point, which was demonstrated to outperform the original CMA-ES in both efficiency and robustness; however, it still needs a parameter to be tuned.

The restart strategy with increasing population size is preferred in this study, mainly because of its simplicity. For multimodal optimization, there could be situations in which increasing population size could be even more advantageous than a fix, albeit well-tuned, one. For example, the basins that are easy to identify can be detected in the early restarts, when population size is small, while subsequent restarts with a large population size aim at finding harder minima. When all subpopulations in a restart are terminated, $\lambda$ is doubled unless the remaining evaluation budget ($FE_{\mathrm{remained}}$) is not sufficient for the subsequent restart with increased population size to conclude properly. If so, the population size is increased by a smaller factor, or the number of subpopulations might be decreased to make sure most subpopulations will have sufficient budget to converge:

$$\begin{cases} \lambda \leftarrow 2\lambda, N_{\mathrm{s}} \leftarrow N_{\mathrm{s}}^0 & \text{if} \quad FE_{\mathrm{remaind}} > 2FE_{\mathrm{used}} \left(\dfrac{N_{\mathrm{s}}^0}{N_{\mathrm{s}}}\right) \\ \lambda \leftarrow \left\lfloor \lambda \left(\dfrac{FE_{\mathrm{remaind}}}{FE_{\mathrm{used}}}\right) \right\rfloor, N_{\mathrm{s}} \leftarrow N_{\mathrm{s}}^0 & \text{if} \quad \text{else} \\ \lambda \leftarrow \lambda, N_{\mathrm{s}} \leftarrow \max\left\{1, \left\lfloor N_{\mathrm{s}} \left(\dfrac{FE_{\mathrm{remaind}}}{FE_{\mathrm{used}}}\right) \right\rfloor\right\} & \text{if} \quad FE_{\mathrm{remaind}} < FE_{\mathrm{used}} \left(\dfrac{N_{\mathrm{s}}^0}{N_{\mathrm{s}}}\right) \end{cases} \tag{5}$$

where $FE_{\mathrm{used}}$ is the used evaluation budget in the previous restart, which also estimates the required budget for the next restart. $N_{\mathrm{s}}^0$ is the default number of subpopulations. This process continues until the available budget of function evaluations is depleted. Based on available statistical stopping criteria for CMA-ES (Hansen 2009), a subpopulation is terminated if one of the following conditions occurs:

- *ConditionC*: the condition number of $\mathbf{C}$ exceeds $10^{14}$.
- *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 recent function values and the median of recent function values of the last $\lfloor 0.2t + 120 + 30D/\lambda \rfloor$ iterations, excluding elites.
- *TolHistFun*: the range of the best recent function values (excluding elites) during the last $\lfloor 10 + 30D/\lambda \rfloor$ iterations is smaller than *TolHistFun*.

Converged subpopulations either specify a new niche, which is stored in **Archive**, or specify an already discovered niche. In the latter case, still useful information is extracted from such convergence to adapt the normalized taboo distance for the subsequent restart.

### 2.4 Adaptation of the normalized taboo distance

A universal value for the normalized taboo distance may work well if all basins are roughly similar in size, otherwise it would be too small for large basins and too large for the others. Magnitude of $\hat{d}_k$ can be considered as the repelling power of the taboo

point $\boldsymbol{y}_k$. For an arbitrary basin, this parameter is adapted based on the number of subpopulations that converge to that basin. When a restart concludes, the best solution of those subpopulations that have been terminated because of the *TolHistFun* condition is considered to be a local minimum, and compared to the solutions in **Archive** to check whether it is desirable. For the case when only global minima are desired, a converged solution is desirable only if its fitness is not worse than the fitness of best solution in **Archive** minus $\epsilon_f$ (target tolerance on the objective function). All undesirable solutions are discarded. For all the remaining (desirable) solutions ($\widetilde{y}_i, i = 1, 2, \ldots, I$), the strategy proposed by Ursem (1999) is utilized in order to determine whether a solution refers to a new basin; yet, the line search is rendered by Golden Section Search with a maximum of 10 function evaluations. It is added to **Archive** if it appears to be a new basin, otherwise the corresponding archived point is identified. At the end, the normalized taboo distance of the solutions in **Archive** is updated based on the number of subpopulations ($N_{\text{rep}}$) that have converged to the corresponding basin. If this number is large, $\hat{d}_m$ is increased to reduce the probability of convergence to the same basin in the future restarts, otherwise it is decreased since it might be unnecessarily large. Algorithm 1 explains the procedure.

$\hat{d}_0$ is the default value of the normalized taboo distance, which is assigned to the new members of **Archive** as well as subpopulations during the optimization process. $0 \leq \alpha_{\text{new}} \leq 1$ is a parameter which specifies the fraction of solutions that can converge to an already identified basin, without further increase in the corresponding taboo distance. if $N_{\text{rep}_m} = 0$, $\hat{d}_m$ is decreased since it might be unnecessary large. Larger $N_{\text{rep}_m}$ refers to convergence of more subpopulations to the $m$-th archived solution, which demonstrates that current value of $\hat{d}_m$ is too small. $\tau_{\hat{d}}$ specifies the learning rate of the normalized taboo distance. Default values are $\tau_{\hat{d}} = \sqrt{1/D}$ and $\alpha_{\text{new}} = 0.5$. $\hat{d}_0$ is equal to the 25th percentile of normalized taboo distances of the solutions stored in **Archive**. The justification is that basins that are detected in the current restart are probably harder to find than those identified in the previous restarts. Among the factors that make a basin hard to find, the size of the basin is a critical one, and thus the recently detected minima are likely to have a narrower basin than those found previously. if **Archive** is empty, then $\hat{d}_0 = 1$.

### 2.5 Boosting time efficiency

The main time complexity of the proposed niching strategy originates from the sampling part, according to which distance of each sampled solution should be checked against all taboo points. It is also likely that most sampled solutions are rejected because of the distance criteria (Equation 2), especially in the early iterations of each restart when the mutation strength is great. Three different alternatives are proposed to alleviate this problem.

First, taboo points for a subpopulation are classified into two groups, critical and non-critical taboo points. Critical taboo points are those that significantly affect distribution of taboo acceptable solutions, when compared to distribution of the sampled solutions. Non-critical taboo points are those far from the center of the subpopulation, or with small normalized taboo distance, and therefore, the probability that a solution is sampled and rejected because of the proximity to them is minimal. When checking acceptability of a sampled solution, the distance criterion (Equation 2) is checked only for the critical taboo points. This means it is assumed that the sampled solution always satisfy distance criterion with respect to non-critical taboo points. The set of crit-

ical taboo points for each subpopulation is determined iteratively. Second, the critical taboo points are checked for satisfaction of the distance criterion one after another, such that those with more level of criticality are checked earlier. This helps identification of taboo unacceptable samples faster, and saves some computation time. The measure to quantify level of criticality of a taboo point is explained in this section. Third, the distance condition is loosened whenever a sampled solution ($\boldsymbol{x}_{ij}$) is rejected by temporary reduction of the normalized taboo distance of all taboo points: $\hat{d}_k \leftarrow c_{\mathrm{red}}\hat{d}_k, 0<c_{\mathrm{red}}<1$. This modification is valid only for this subpopulation and only for the current iteration. When $\lambda$ solutions are evaluated, the original values of $\hat{d}_k$ are restored. As default setting, $c_{\mathrm{red}}=0.99^{(1/D)}$, which means the size of the taboo region shrinks by 1% whenever a solution is rejected.

**Algorithm 1:** Updating **Archive** and the normalized taboo distance of its members.

**Data**: **Archive** with $M$ members, desirable solutions, $(\widetilde{y}_i, i=1,2,\ldots,I)$, default value of the normalized taboo distance ($\hat{d}_0$), target rate of basin identification ($\alpha_{\mathrm{new}}$)
**Result**: Updated **Archive**
initialization
$N_{\mathrm{rep}_m} \leftarrow 0, m=1,2,\ldots,M$
**for** *i* ← *1 to I* **do**
  Sort the points in **Archive** $(\boldsymbol{y}_m)$ based on their Euclidean distance to $\widetilde{y}_i$.
  $isnew \leftarrow 1$
  **for** *m* ←*1 to M* **do**
    **if** $\boldsymbol{y}_m$ *and* $\widetilde{\boldsymbol{y}}_i$ *share the same basin* **then**
      Replace $\boldsymbol{y}_m$ with $\widetilde{\boldsymbol{y}}_i$ in **Archive** if the latter is fitter
      $N_{\mathrm{rep}_m} \leftarrow N_{\mathrm{rep}_m}+1$
      $isnew \leftarrow 0$
      Exit the loop
    **end**
  **end**
  **if** $isnew=1$ **then**
    Add $\widetilde{y}_i$ to **Archive**.
    $N_{\mathrm{rep}_{m+1}} \leftarrow 0$
    $\hat{d}_m \leftarrow \hat{d}_0$ %
    $M \leftarrow M+1$
  **end**
  **for** $m=1$ *to* $M$ **do**
    **if** $N_{\mathrm{rep}_m} > \alpha_{\mathrm{new}}(I/M)$ **then**
      $\hat{d}_m \leftarrow \hat{d}_m \left(1+N_{\mathrm{rep}_m}-\alpha_{\mathrm{new}}(I/M)\right)^{(+\tau_{\hat{d}})}$
    **else**
      $\hat{d}_m \leftarrow \hat{d}_m \left(1-N_{\mathrm{rep}_m}+\alpha_{\mathrm{new}}(I/M)\right)^{(-\tau_{\hat{d}})}$
    **end**
  **end**
**end**

To find a metric to quantify criticality of a taboo point, it is initially assumed that mutation is isotropic with a single parameter $\sigma_{\mathrm{mean}}$ (Figure 3). Let the origin of the auxiliary coordinate system $[\hat{e}_1, \hat{e}_2, \ldots, \hat{e}_D]$ coincide $\boldsymbol{x}_{\mathrm{mean}}$, while $\hat{e}_1$ is along the line

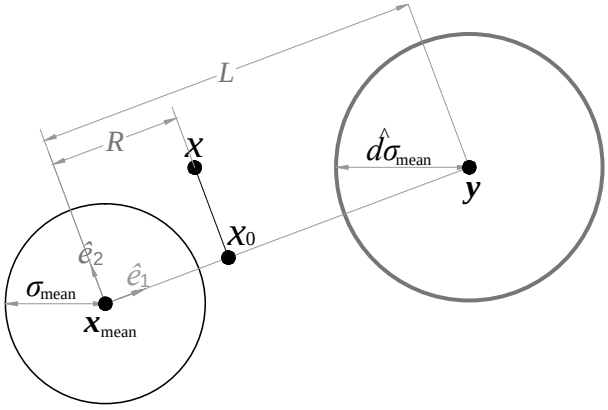



Figure 3: Subpopulation **P** samples $\boldsymbol{x}$ at the proximity of taboo point $\boldsymbol{y}$.

segment connecting $\sigma_{\text{mean}}$ to $\boldsymbol{y}$. $x_1, x_2, \ldots, x_D$ are coordinates of the sampled solution $\boldsymbol{x}$ in the auxiliary coordinate system, which are realizations of $X_1, X_2, \ldots, X_D \sim \mathcal{N}(\mu = 0, \sigma^2)$. Let $p_1(R)$ be the probability density that $\boldsymbol{x}$ lies on plane $x_1 = R$ (line $x_1 = R$ in 2D space):

$$p_1(R) = \int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} \cdots \int_{-\infty}^{+\infty} f(R, x_2, x_3, \ldots, x_D)\, dx_2\, dx_3 \cdots dx_D \tag{6}$$

where $f(x_1, x_2, \ldots, x_D)$ is the joint probability density of $X_1, X_2, \ldots, X_D$. Since the covariance matrix is the identity matrix, $X_1, X_2, \ldots, X_D$ are independent and identically distributed (iid) random variables, therefore, the joint density function is the product of the marginal density functions:

$$p_1(R) = \int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} \cdots \int_{-\infty}^{+\infty} p_{X_1}(R)\, p_{X_2}(x_2)\, p_{X_3}(x_3)\, \cdots p_{X_D}(x_D)\, dx_2\, dx_3 \cdots dx_D \tag{7}$$

where $p_{X_1} = p_{X_2} = \ldots = p_{X_D}$ are the marginal density functions. Terms inside the integral in Equation 7 are independent and $\int_{+\infty}^{-\infty} p_{X_i}(x_i)\, dx_i = 1$. This simplifies Equation 7 as follows:

$$p_1(R) = p_{X_1}(R) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(\frac{-R^2}{2\sigma^2}\right) \tag{8}$$

where $\sigma$ is the step size of solution $\boldsymbol{x}$, derived from slight mutation of the subpopulation step size ($\sigma \approx \sigma_{\text{mean}}$). For simplicity, we assume $\sigma = \sigma_{\text{mean}}$. Let $\boldsymbol{x}_0$ be projection of $\boldsymbol{x}$ on the first axis of the auxiliary coordinate system ($\boldsymbol{x}_0 = [x_1, 0, 0, \ldots, 0]$). According to Equation 2, the rejection probability of $\boldsymbol{x}_0$ is:

$$P_{\text{trej}}(\boldsymbol{x}_0, \boldsymbol{y}) = \begin{cases} \exp\left(1 - \left(\dfrac{L - R}{\hat{d}\,\sigma_{\text{mean}}}\right)^2\right) & \text{if } \left|\dfrac{L - R}{\hat{d}\,\sigma_{\text{mean}}}\right| > 1 \\ 1 & \text{if else} \end{cases} \tag{9}$$

For all points with $x_1 = R$, $\boldsymbol{x}_0$ has the highest rejection probability. The mean (expected)

rejection probability of $\boldsymbol{x}_0$ when $R$ varies from $-\infty$ to $+\infty$ is computed as follows:

$$\bar{P}_{\text{trej}}(\boldsymbol{y}) = \int_{-\infty}^{+\infty} p_1(R)\, P_{\text{trej}}(\boldsymbol{x}_0, \boldsymbol{y})\, \mathrm{d}R \tag{10}$$

where $P_{\text{trej}}(\boldsymbol{x}_0, \boldsymbol{y})$ is a function of $R$. For simplicity, all points with $R > L + \hat{d}\,\sigma_{\text{mean}}$ are assumed to be rejected, which leads to a larger mean rejection probability than the actual one:

$$\bar{P}_{\text{trej}}(\boldsymbol{y}) < \int_{-\infty}^{L-\hat{d}\,\sigma_{\text{mean}}} p_1(R)\, P_{\text{trej}}(\boldsymbol{x}_0, \boldsymbol{y})\, \mathrm{d}R + \int_{L-\hat{d}\,\sigma_{\text{mean}}}^{+\infty} p_1(R)\, \mathrm{d}R \tag{11}$$

$L$ and $R$ are absolute values, which are normalized with respect to $\sigma_{\text{mean}}$ and $P_{\text{trej}}(\boldsymbol{x}_0, \boldsymbol{y})$ is computed using Equation 9:

$$\begin{aligned}
\bar{P}_{\text{trej}}(\boldsymbol{y}) &< \int_{-\infty}^{l-\hat{d}} \frac{1}{\sqrt{2\pi}} \exp\!\left(-\frac{r^2}{2}\right) \exp\!\left(1 - \left(\frac{d-r}{\hat{d}}\right)^2\right) \mathrm{d}r + \left(1 - \Phi\left(d - \hat{d}\right)\right) \\
&= \int_{-\infty}^{l-\hat{d}} \frac{1}{\sqrt{2\pi}} \exp\!\left(-\frac{(r-\eta)^2}{2\tau^2} + \delta\right) \mathrm{d}r + \left(1 - \Phi\left(d - \hat{d}\right)\right) \\
&= \tau \exp(\delta) \Phi\left(\frac{d - \hat{d} - \eta}{\tau}\right) + \left(1 - \Phi\left(d - \hat{d}\right)\right) \\
&= \bar{P}^u_{\text{trej}}(\boldsymbol{y})
\end{aligned} \tag{12}$$

where

$$d = \frac{L}{\sigma_{\text{mean}}}, r = \frac{R}{\sigma_{\text{mean}}}$$

are the normalized distances, $\Phi$ computes *cdf* of standard normal distribution and

$$\tau = \frac{\hat{d}}{\sqrt{\hat{d}^2 + 2}}\,,\ \eta = \frac{2d}{2 + \hat{d}^2}\,,\ \delta = 1 - \left(\frac{d}{\hat{d}}\right)^2 + \frac{\eta^2}{2\tau^2}$$

$\bar{P}^u_{\text{trej}}(\boldsymbol{y})$ gives an upper limit for the mean rejection probability of $\boldsymbol{x}$ because of the violation of the distance metric to the taboo point $\boldsymbol{y}$. This value is the employed measure to compare criticality of the taboo points.

It is remarkable that all terms in Equation 12 are normalized with respect to the mutation strength. Using the Mahanolobis distance metric to compute $d$, Equation 12 is generalized to the case when the full covariance matrix is used for sampling. $\bar{P}_{\text{trej}}(\boldsymbol{y}_k), k = 1, 2, \ldots, K$ for a subpopulation in an arbitrary iteration does not change during the sampling stage. This means that for an arbitrary subpopulation, $\bar{P}_{\text{trej}}(\boldsymbol{y}_k)$ is calculated once per iteration. Since $d$ in Equation 12 is inversely proportional to $\sigma_{\text{mean}}$, the number of the critical taboo points gradually decreases when the algorithm converges. This significantly boosts time complexity of sampling especially for large values of $N^0_{\text{s}}$. By default, if $\bar{P}_{\text{trej}}(\boldsymbol{y}_k)$ is smaller than $0.01/\sqrt{K}$, then $\boldsymbol{y}_k$ is considered non-critical.

### 2.6 Initialization of subpopulations

At the beginning of each restart, subpopulations are formed such that they distribute uniformly over the search space excluding the taboo regions. All new subpopulations

have similar initial strategy parameter ($\sigma_{\text{mean}_i} = \sigma_{\text{mean}}, \mathbf{C}_i = \mathbf{I}_{D\times D}, i = 1, 2, \ldots, N_{\text{s}}$). The absolute taboo distance of the $m$-th archived point is proportional to $\sigma_{\text{mean}}$. The proper value of $\sigma_{\text{mean}}$ which leads to the maximum diversity in distribution of subpopulations, is not known beforehand, therefore, a conservatively large value for $\sigma_{\text{mean}}$ is initially used. This initial estimate is utilized to generate the centers of subpopulations such that they lie far from the taboo regions and other subpopulation. If this condition is violated, a new point is randomly selected as the center for the subpopulation. If multiple consecutive tries (say, 100) fail, $\sigma_{\text{mean}}$ is slightly reduced ($\sigma_{\text{mean}} \leftarrow c_{\text{red}}\, \sigma_{\text{mean}}$), which reduces the overall size of the taboo regions. This process continues until $N_{\text{s}}$ subpopulations are generated (Algorithm 2).

**Algorithm 2:** Initialization of subpopulations at the beginning of each restart

**Data**: $N_{\text{s}}, D, c_{\text{red}}$, default value of the normalized taboo distance ($\hat{d}_0$), **Archive** with $M$ members, range of the design variables ($\boldsymbol{D_x}, \boldsymbol{U_x}$)
**Result**: Center of subpopulations ($\boldsymbol{x}_{\text{mean}_i}$) and their global step size ($\sigma_{\text{mean}_i} = \sigma_{\text{mean}}$, $i = 1, 2, \ldots, N_{\text{s}}$)
initialization;
$i \leftarrow 1$ ; $N_{\text{rej}} \leftarrow 0$
$\sigma_{\text{mean}} \leftarrow 0.33 \times \|\boldsymbol{D_x} - \boldsymbol{U_x}\|$
**while** $i \leq N_s$ **do**
  Sample $\boldsymbol{D_x} \leq \boldsymbol{x}_{\text{mean}_i} \leq \boldsymbol{U_x}$ randomly.
  **if** $\forall (m \leq M) : (2\hat{d}_0 + \hat{d}_m)\sigma_{\text{mean}} \leq \|\boldsymbol{x}_{\text{mean}_i} - \boldsymbol{y}_m\|$ **then**
    **if** $\forall (q < i) : 3\hat{d}_0 \sigma_{\text{mean}} \leq \|\boldsymbol{x}_{\text{mean}_i} - \boldsymbol{x}_{\text{mean}_q}\|$ **then**
      $i \leftarrow i+1$ % *accept* $\boldsymbol{x}_{\text{mean}_i}$
      $N_{\text{rej}} \leftarrow 0$
    **end**
  **end**
  $N_{\text{rej}} \leftarrow N_{\text{rej}} + 1$
  **if** $N_{\text{rej}} > 100$ **then**
    $\sigma_{\text{mean}} \leftarrow c_{\text{red}} \sigma_{\text{mean}}$
    $N_{\text{rej}} \leftarrow 0$
  **end**
**end**
Let $\boldsymbol{x}_{\text{mean}_i}$ be center of the subpopulations, with strategy parameters of $\mathbf{C}_i = \mathbf{I}_{D\times D}, \sigma_{\text{mean}_i} = \sigma_{\text{mean}}$ $(i = 1, 2, \ldots, N_{\text{s}})$

### 2.7 Elitism

Both CMA-ES and CMSA-ES are non-elitist methods, however, our preliminary experiments revealed that an intermediate selection scheme, where a fraction of parents may survive to the next generation, can improve performance in multimodal optimization, and thus was preferred in this study. By default, $N_{\text{elt}} = \max\{1, \lfloor 0.15\lambda \rfloor\}$.

### 2.8 Pseudo code

The pseudo code of RS-CMSA is as follows:
Let **Archive**={}
While the evaluation budget is not depleted do:
  Initialize $N_{\text{s}}$ subpopulations according to Algorithm 2.
  While all subpopulations are not terminated, do:
    For $i$=1 to the No. of remaining (not terminated) subpopulations do:

Find the taboo points for $\mathbf{P}_i$, which are the points in **Archive**, unless the elite of $\mathbf{P}_i$ is fitter than the archived point, plus the center of $\mathbf{P}_p, p = 1, 2, \ldots, i-1$. Use Equation 12 to determine which taboo points are critical.
Perform mutation to sample new solutions, while all of them satisfy distance condition to the critical taboo points, according to Equation 2. Perform selection locally. Update elite individuals and strategy parameters of $\mathbf{P}_i$ according to Equation 4.
End *%iteration completed*
Sort the subpopulations according to the fitness of their best individual. Stop a subpopulation for which a termination criterion is satisfied from further sampling. Save the best individual of terminated subpopulations only if convergence has occurred.
End *%restart completed*
Analyze the saved solutions to update **Archive** according to Algorithm 1.
Update the number of subpopulations and population size according to Equation 5.
End *%run completed*

### 2.9 Parameter tuning

Except the default number of subpopulations ($N_{\mathrm{s}}^0$), all parameters are set equal to their default values. This parameter should be proportional to the desired number of minima. Smaller values results in a higher short-term success. One subpopulation per desired minimum is the recommended minimal value. Stopping criteria *TolHistFun* is set equal to the desired tolerance of the objective function. The crafting effort is thus zero provided that an estimate for the number of the desired/available minima is given.

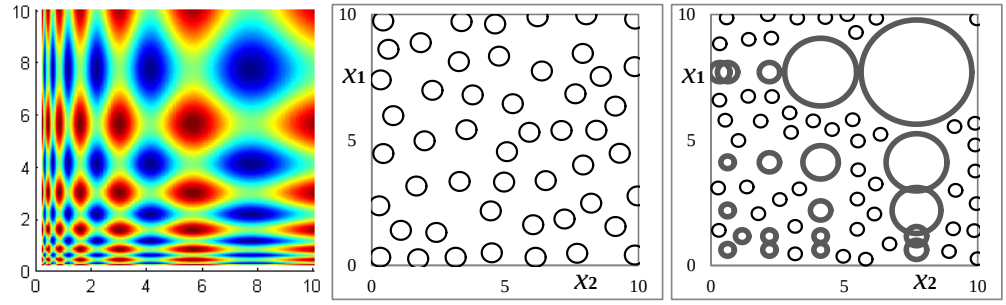
## 3 Descriptive experiments

Two descriptive experiments are rendered in this section to highlight some aspects of RS-CMSA. At this point, the objective is not to benchmark and compare RS-CMSA with other methods but to show:

- Adaptation of the normalized taboo distance and its importance when basins significantly vary in size, shape and relative distance.
- Effect of using Equation 12 on reduction of the number of the critical taboo points.

In the first descriptive experiment, RS-CMSA is tested on the 2D Vincent function. It has 36 minima, all global, which significantly vary in size and relative distance (Figure 4a). Restarts are performed, but for this experiment, the population size is kept unchanged ($\lambda = 10, N_{\mathrm{s}}^0 = 50$). Initialized subpopulations are depicted with black circles, where the center and radius of each circle represent $\boldsymbol{x}_{\mathrm{mean}_i}$ and $\hat{d}_0 \sigma_{\mathrm{mean}_i}$ respectively. The center and radius of the gray circles correspond to the archived points and their taboo region ($\hat{d}_m \sigma_{\mathrm{mean}}$) respectively. At this point of each restart, $\mathbf{C}_i$ is the identity matrix and $\sigma_{\mathrm{mean}_i} = \sigma_{\mathrm{mean}}$ is identical for all subpopulations, therefore, all taboo regions are spherical and the difference in size of the taboo regions originates merely from the difference in the normalized taboo distance ($\hat{d}_m$). It is remarkable that illustration of these regions is much harder for an arbitrary point during the restart, since the shape, the size, and the number of taboo regions would be different for each subpopulation.

Figure 4b depicts the generated subpopulations immediately after the initialization part in the zeroth restart. There is no archived point at this moment, and centers of the subpopulations are at least $3\hat{d}_0 \sigma_{\mathrm{mean}}$ far from each other.




(a) Top view of the landscape (b) Restart #0: $\sigma_{\text{mean}} = 0.38$ (c) Restart #1: $\sigma_{\text{mean}} = 0.45$

(d) Restart #2: $\sigma_{\text{mean}} = 0.40$ (e) Restart #3:: $\sigma_{\text{mean}} = 0.38$ (f) Restart #4: $\sigma_{\text{mean}} = 0.29$

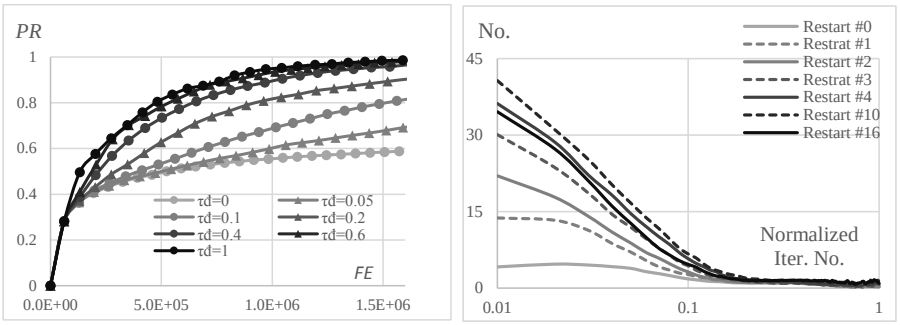
Figure 4: Taboo regions (gray circles) and current subpopulations and their step sizes (black circles) at different restarts immediately after the initialization step.

Figure 4c illustrates the solutions in **Archive** and their absolute taboo distance immediately after initialization of the subpopulations in the first restart. There are 20 archived solutions, which implies 20 global minima were identified during the zeroth restart. Comparing taboo regions of the archived points with the basin size of the global minima (Figure 4a) demonstrates that RS-CMSA has correctly increased the normalized taboo distance of minima with wider basins.

At the beginning of the second restart (Figure 4d), there are 27 solutions in the **Archive**, which means 7 new minima were located during the first restart. The three largest taboo regions pertaining to the three largest basins, have defined huge taboo regions in the top-right part of the search space. This not only pushes subpopulations away from these minima in the current restart, but also disrupts distribution of the initialized subpopulations such that almost all of them are generated on the bottom-left side of the search space, where most of the unidentified minima lie. These factors increase probability of convergence to new minima in the current restart.

There are 34 archived solutions at the beginning of the third restart (Figure 4e), which implies 7 new minima were identified in the second restart. It seems few solutions of the second restart have converged to the largest minima, and thus their normalized taboo distances have dwindled for the third restart. Not many subpopulations have initiated at the bottom-left side, and hence the algorithm can hardly detect any new minima.

At the beginning of the fourth restart, the size of **Archive** is still 34, which means no new minima were identified in the third restart. Nevertheless, the third restart has significantly resized the normalized taboo distance of the archived pints. It seems many




(a) (b)

Figure 5: Results on 3D Vincent function: a) Peak ratio ($PR$) versus the adaptation rate of the normalized taboo distance ($\tau_{\hat{d}}$) and b) Average number of the critical taboo points versus the normalized iteration number, plotted for some selected restarts when $\tau_{\hat{d}} = 0.6$.

subpopulations of the third restart have converged to the largest basin, such that its taboo region covers most part of the search space in the fourth restart. All new subpopulations are initialized at the bottom-left side, and at the end of this restart, all 36 minima were identified.

As it can be observed, this procedure may reasonably adapt the normalized taboo distance of the archived points. If it is larger than what it should be, more subpopulations will converge to it in the subsequent restarts, which increases the corresponding normalized taboo distance, and vice versa.

In the second experiment, importance of the adaptation of the normalized taboo distance and the effect of ignoring non-critical taboo points on time complexity of the method are studied on the 3D Vincent function, which has 216 minima, all global. Different learning rates ($\tau_{\hat{d}}$) for the normalized taboo distance are tried with $N_{\mathrm{s}}^0 = 108$ and $\lambda = 10$ (fixed). Figure 5a illustrates peak ratio ($PR$), which is the fraction of identified minima, versus the number of function evaluations, averaged over 20 independent runs. Markers represent the average time in which a restart was performed. For $\tau_{\hat{d}}$=0, adaptation of the normalized taboo distances is suppressed and thus $\hat{d}_m = \hat{d}_0 = 1$ for $m = 1, 2, \ldots, M$. Results after the zeroth restart are predictably similar, since the adaptation of the normalized taboo distance is performed when a restart is concluded. A considerable decline in the rate of identification of new basins is observed when $\tau_{\hat{d}}$ is small. It seems for this problem, $0.6 \leq \tau_{\hat{d}} \leq 1$ is a logical choice.

Figure 5b illustrates the average number of the critical taboo points for some selected restarts, when $1\%, 2\%, \ldots, 100\%$ of iterations of the restart have completed. The average number of the critical taboo points increases with the restart number for the early restarts, mainly because of the growth in the size of **Archive**, however, there remain only a few critical taboo points per subpopulation after 10% of the iterations of a restart. The result is a significant reduction of the algorithmic computation time, since samples of each subpopulation are checked only against the critical taboo points.

## 4 Performance evaluation

A number of the best available methods in the literature for multimodal optimization are selected. These algorithms are tested and compared to RS-CMSA on the CEC2013

test suite on multimodal optimization.

### 4.1 Selection of the methods

Several factors are considered to select the niching methods for comparing RS-CMSA with, including the year of publication, strength of the reported numerical results, minimal tuning effort and niche-radius independence. Following these criteria, the following methods are selected: NEA2 (Preuss, 2012), LIPS (Qu et al., 2013) , NSDE (Qu et al., 2012) and PNPCDE (Biswas et al., 2014).

In CEC2013 workshop, NEA2 was benchmarked without any parameter tuning, however, for PNPCDE, LIPS and NSDE, population size was tuned for each problem in the referenced studies. To follow a unified parameter setting for each method, we assume that the optimum population size for these methods (including NEA2) is a function of the number of global minima $(N_{\mathrm{gopt}})$ and probably problem dimension, however, our preliminary results suggested dependence on $N_{\mathrm{gopt}}$ only, which is followed through the rest of parameter setting.

In practice, the exact number of minima is not known a priori, however, one usually a rough idea of the desired number of minima. For benchmarking, it is assumed a rough estimate for $N_{\mathrm{gopt}}$ is available which can be used for parameter setting:

$$\begin{cases} \lambda = 5 \times 2^{\beta} \times \widetilde{N}_{\mathrm{gopt}} & \text{for} \quad \text{NEA2, PNPCDE, LIPS, NSDE} \\ N_{\mathrm{s}}^{0} = 2^{\beta} \times \widetilde{N}_{\mathrm{gopt}} & \text{for} \quad \text{RS-CMSA} \end{cases} \tag{13}$$

where $\widetilde{N}_{\mathrm{gopt}}$ is the estimated number of global minima. $\beta$ specifies the number of individuals (subpopulation for RS-CMSA) for a given $\widetilde{N}_{\mathrm{gopt}}$. Parameter study is performed in the next section to find the best value of $\beta$ for each method. Performance sensitivity to accuracy of the estimated number of global minima is also analyzed by varying $\widetilde{N}_{\mathrm{gopt}}$ for a given problem.

### 4.2 Performance measures

Parameter setting and consequently performance measures would depend on $\beta$ and $\widetilde{N}_{\mathrm{gopt}}$. For each method, mean peak ratio (*MPR*) is computed as a function of $\beta$ in Equation 13, assuming that $\widetilde{N}_{\mathrm{gopt}} = N_{\mathrm{gopt}}$:

$$MPR_0(\beta) = \sum_{i=1}^{20} \sum_{j=2}^{5} PR_{ij} \tag{14}$$

where subscript '0' refers to the assumption that $\widetilde{N}_{\mathrm{gopt}} = N_{\mathrm{gopt}}$. $PR_{ij}$ is the average fraction of global minima found over 50 independent runs of the *i*-th problem with respect to the *j*-th target precision. There are 20 problems (*PID* $= 1, 2, \ldots, 20$) and the target precisions are $\epsilon_f = 10^{-j}$. $\epsilon_f = 10^{-1}$ is considered to be a too loose tolerance and hence it is removed. The precision of the provided global minimum value for the 2D Shubert function (*PID*=6) was not sufficient and thus, for this problem, *PR* for $\epsilon_f = 10^{-3}$ and $\epsilon_f = 10^{-4}$ is extrapolated to compute *PR* for $\epsilon_f = 10^{-5}$.

$MPR_0$ plots can compare different niching methods should the exact number of global minima be known beforehand. This is usually not the case in practice, where one may have only a rough estimate of the number of global or desirable minima. This

means *MPR* is a function of both $\beta$ and $\gamma$:

$$MPR(\beta,\gamma) = \sum_{i=1}^{20}\sum_{j=2}^{5} PR_{ij}, \gamma = \log_2\left(\frac{\widetilde{N}_{\text{gopt}}}{N_{\text{gopt}}}\right) \tag{15}$$

Robust mean peak ratio (*RMPR*) computes the expected performance when $\gamma$ varies from $-\infty$ to $+\infty$

$$RMPR(\beta) = \int_{-\infty}^{+\infty} \rho(\gamma) MPR(\gamma,\beta)\mathrm{d}\gamma \tag{16}$$

where $\gamma$ is a random variable with probability density function of $\rho(\gamma)$, which is assumed to be symmetric around the origin. *RMPR* measures efficiency and robustness with respect to $\gamma$ at the same time.

It is notable that $MPR(\beta,\gamma)$ can be derived from the $MPR_0(\beta)$ curve because the effect of increasing $\beta$ and $\gamma$ on parameter setting is identical. For example, increasing $\gamma$ by one means $\widetilde{N}_{\text{gopt}}$ is doubled (Equation 15), which in turn doubles $\lambda$ or $N_{\text{s}}^0$ (Equation 13). Increasing $\beta$ by one also doubles $\lambda$ or $N_{\text{s}}^0$ (Equation 13). Therefore, $MPR(\beta,\gamma)$ in Equation 16 can be replaced with $MPR_0(\beta+\gamma)$:

$$RMPR(\beta) = \int_{-\infty}^{+\infty} \rho(\gamma) MPR_0(\beta+\gamma)\mathrm{d}\gamma \approx \sum_{\gamma=-\infty}^{\gamma=+\infty} P(\gamma) MPR_0(\beta+\gamma) \tag{17}$$

To calculate *RMPR* according to Equation 17, one should evaluate $MPR_0(\beta)$ for many values of $\beta$. To avoid this problem, Equation 17 is further simplified as follows:

$$RMPR(\beta) \approx (1-2\gamma_0)MPR(\beta) + \gamma_0 MPR(\beta-1) + \gamma_0 MPR(\beta+1) \tag{18}$$

which only includes $\gamma = -1, 0, 1$ terms. For this study, $\gamma_0 = 0.33$ is used.

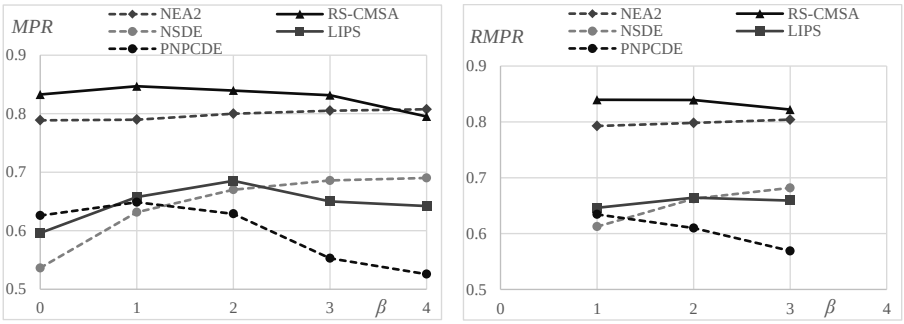
For each method, the best value of $\beta$ ($\beta^\star$) which results in the greatest $RMPR(\beta)$ is sought. $RMPR(\beta^\star)$ is employed as the primary performance measure to compare different methods.

### 4.3 Results and comparison

Several aspects of these methods are compared by post-processing the results. Figure 6 illustrates $MPR_0(\beta)$ and $RMPR(\beta)$ for all methods. The range of 99% confidence interval for these plots is fairly narrow (in order of 0.001) and thus the effect of random nature of the results is ignored, unless the difference between two methods is in the same range. Figures 7a-7e plots *PR* (averaged over all target function tolerances) versus *PID* for different values of $\beta$, assuming $\gamma = 0$. *PR* versus *PID* is plotted for each method when $\beta = \beta^\star$ and $\gamma = 0$ in Figure 7f.

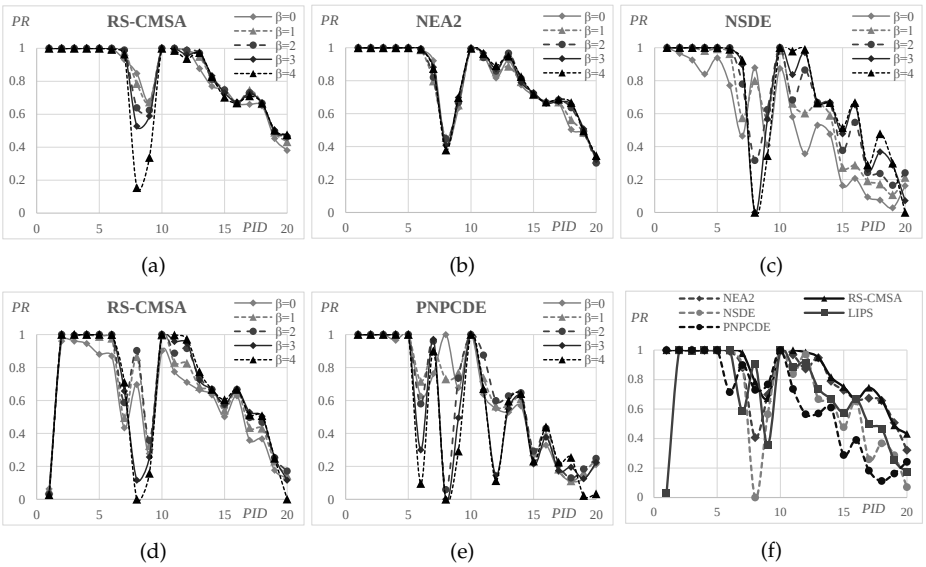
- Figure 6a demonstrates that for a large range of $\beta$, RS-CMSA outperforms NEA2, and both significantly surpass PNPCDE, LIPS and NSDE when $MPR_0$ is regarded. NEA2 shows minimum performance sensitivity to $\beta$, followed by RS-CMSA.

- According to *RMPR* curves, $\beta^\star$= 1, 3, 3, 2 and 1 for RS-CMSA, NEA2, NSDE, LIPS and PNPCDE respectively. For $\beta = \beta^\star$, RS-CMSA relatively outperforms NEA2 (*RMPR*=0.840 versus *RMPR*=0.804), and both significantly surpass NSDE, LIPS and PNPCDE.




(a) $MPR_0(\beta)$ (b) $RMPR(\beta)$

Figure 6: $MPR_0(\beta)$ and $RMPR(\beta)$ for different methods.




(a) (b) (c)

(d) (e) (f)

Figure 7: a-e) Effect of $\beta$ on $PR$ for each method on each problem, when $\gamma = 0$. f) $PR$ when $\beta = \beta^\star, \gamma = 0$ for each method on each problem.

- The first five problems are the easy ones for which most niching methods can reach *PR*=1 within a small evaluation budget (Figure 7). Strangely, LIPS cannot reach a good *PR* for *PID*=2, a 1D problem with two peaks at the bounds. This implies that LIPS cannot handle this situation efficiently, and some revisions on its operators seem necessary.
- It is unlikely that for identical number of iterations, increasing population size ($N_{\mathrm{s}}^0$ for RS-CMSA) results in a lower *PR* and hence, performance decline for larger (sub-) populations is attributed to termination of the optimization process before convergence. This means increasing/decreasing the evaluation budget can modify the optimum value of $\beta$ for each method.
- *PR* of NEA2 and RS-CMSA varies slightly when $\beta$ is modified, while this variation is quite significant for the other methods (Figure 7). The more critical point is that the optimum values of $\beta$ is highly problem dependent for these methods, which challenges finding a good universal parameter setting. For example, $\beta^\star$=3 for NSDE, however, the highest *PR* for *PID*=8 is achieved when $\beta$=0, and increasing $\beta$ significantly reduces *PR* for this method. The reason is depletion of the evaluation budget before convergence. It seems the sequential nature of NEA2 and to some extent RS-CMSA, mitigates this problem, since at least a few restarts are performed even if the evaluation budget might be small.
- RS-CMSA considerably outperforms NEA2 for the 3D Shubert function (*PID*=8, *PR*=0.786 versus *PR*=0.407). For a few other problems, minor superiority (*PID*=7, 12, 15, 20) is observed. For the rest, both methods have quite similar *PR*.
- For the composite functions (*PID*=11-20), a detectable domination of RS-CMSA and NEA2 over the other methods is observed (Figure 7f). NSDE and LIPS show similar performance, while both on average outperform PNPCDE.

## 5 Summary and conclusions

A novel niching method called covariance matrix self-adaptation with repelling subpopulations (RS-CMSA) was developed in this study. RS-CMSA mainly consists of CMSA evolution strategy as the core search engine of several equally sized subpopulations. Distribution of members of each subpopulation is reshaped by presence of taboo points to prevent searching previously explored or already being explored regions.

The challenge of niches with dissimilar basin size, shape or distribution have alluded EA researchers since the pioneering multimodal studies in late eighties. One of the main contributions of this work has been to suggest a self-adaptive update of the niche radius referred here as the normalized taboo distance and its integration with one of the most efficient real-parameter evolutionary optimization algorithms – covariance matrix self-adaptation evolution strategy. Another hallmark of this study is that for a rough estimate for the desired number of minima, there is no parameter tuning required which makes RS-CMSA a robust and practically useful method.

RS-CMSA has been evaluated and compared with a number of the state-of-the-art niching methods on a recent test suite proposed in the CEC2013 special session on multimodal optimization. Robust mean peak ratio has been utilized to compare the benchmarked methods, not only to quantify their success in finding different niches but also to evaluate their performance sensitivity to accuracy of the provided estimate for the number of the global minima. According to the employed performance measure, RS-CMSA has been emerged as the most successful method, followed by NEA2.

In particular, a considerable superiority of NEA2 and RS-CMSA over the other methods has been observed on the more difficult composite functions. Simplicity of the main niching idea, its robustness, and superior results have amply demonstrated the usefulness and practicality of RS-CMSA as a potential niching evolutionary algorithm.

## 6 Acknowledgment

The source codes for NSDE and LIPS were provided by Ponnuthurai N. Suganthan. The source code of PNPCDE was provided by Subhodip Biswas. The authors would like to thank them for sharing their source codes and providing guidelines to tune the control parameters.

## References

Ahrari, A. and Atai, A. A. (2010). Grenade explosion method - a novel tool for optimization of multimodal functions. *Applied Soft Computing*, 10(4):1132–1140.

Ahrari, A., Atai, A. A., and Deb, K. (2014). Simultaneous topology, shape and size optimization of truss structures by fully stressed design based on evolution strategy. *Engineering Optimization*, (ahead-of-print):1–22.

Ahrari, A. and Shariat-Panahi, M. (2013). An improved evolution strategy with adaptive population size. *Optimization*, (ahead-of-print):1–20.

Ahrari, A., Shariat-Panahi, M., and Atai, A. A. (2009). GEM: a novel evolutionary optimization method with improved neighborhood search. *Applied Mathematics and Computation*, 210(2):376–386.

Auger, A. and Hansen, N. (2005). A restart CMA evolution strategy with increasing population size. In *Evolutionary Computation, 2005. The 2005 IEEE Congress on*, volume 2, pages 1769–1776. IEEE.

Bandaru, S. and Deb, K. (2013). A parameterless-niching-assisted bi-objective approach to multimodal optimization. In *Evolutionary Computation (CEC), 2013 IEEE Congress on*, pages 95–102. IEEE.

Basak, A., Das, S., and Tan, K. C. (2013). Multimodal optimization using a biobjective differential evolution algorithm enhanced with mean distance-based selection. *Evolutionary Computation, IEEE Transactions on*, 17(5):666–685.

Beasley, D., Bull, D. R., and Martin, R. R. (1993). A sequential niche technique for multimodal function optimization. *Evolutionary computation*, 1(2):101–125.

Beyer, H.-G. and Sendhoff, B. (2008). Covariance matrix adaptation revisited–the CMSA evolution strategy–. In *Parallel Problem Solving from Nature–PPSN X*, pages 123–132. Springer.

Biswas, S., Kundu, S., and Das, S. (2014). An improved parent-centric mutation with normalized neighborhoods for inducing niching behavior in differential evolution. *IEEE TRANSACTIONS ON CYBERNETICS*, 44(10):1726–1737.

Das, S., Maity, S., Qu, B.-Y., and Suganthan, P. N. (2011). Real-parameter evolutionary multimodal optimization - a survey of the state-of-the-art. *Swarm and Evolutionary Computation*, 1(2):71–88.

Deb, K. and Goldberg, D. E. (1989). An investigation of niche and species formation in genetic function optimization. In *Proceedings of the 3rd International Conference on Genetic Algorithms*, pages 42–50. Morgan Kaufmann Publishers Inc.

Deb, K. and Saha, A. (2012). Multimodal optimization using a bi-objective evolutionary algorithm. *Evolutionary computation*, 20(1):27–62.

Epitropakis, M. G., Plagianakos, V. P., and Vrahatis, M. N. (2011). Finding multiple global optima exploiting differential evolution's niching capability. In *Differential Evolution (SDE), 2011 IEEE Symposium on*, pages 1–8. IEEE.

Hansen, N. (2009). Benchmarking a bi-population CMA-ES on the BBOB-2009 function testbed. In *Proceedings of the 11th Annual Conference Companion on Genetic and Evolutionary Computation Conference: Late Breaking Papers*, pages 2389–2396. ACM.

Hansen, N., Finck, S., Ros, R., Auger, A., et al. (2009). Real-parameter black-box optimization benchmarking 2009: noiseless functions definitions. Technical Report RR-6829, INRIA.

Hansen, N. and Ostermeier, A. (2001). Completely derandomized self-adaptation in evolution strategies. *Evolutionary computation*, 9(2):159–195.

Li, J.-P., Balazs, M. E., Parks, G. T., and Clarkson, P. J. (2002). A species conserving genetic algorithm for multimodal function optimization. *Evolutionary computation*, 10(3):207–234.

Li, M., Lin, D., and Kou, J. (2012). A hybrid niching PSO enhanced with recombination-replacement crowding strategy for multimodal function optimization. *Applied Soft Computing*, 12(3):975–987.

Li, X. (2010). Niching without niching parameters: particle swarm optimization using a ring topology. *Evolutionary Computation, IEEE Transactions on*, 14(1):150–169.

Li, X., Engelbrecht, A., and Epitropakis, M. (2013a). Results of the 2013 IEEE CEC competition on niching methods for multimodal optimization. In *IEEE congress on evolutionary computation competition on: Niching methods for multimodal, optimization*.

Li, X., Engelbrecht, A., and Epitropakis, M. G. (2013b). Benchmark functions for CEC2013 special session and competition on niching methods for multimodal function optimization. Technical report, Evolutionary Computation and Machine Learning Group, RMIT University.

Liang, Y. and Leung, K.-S. (2011). Genetic algorithm with adaptive elitist-population strategies for multimodal function optimization. *Applied Soft Computing*, 11(2):2017–2034.

Liu, L., Yang, S., and Wang, D. (2012). Force-imitated particle swarm optimization using the near-neighbor effect for locating multiple optima. *Information Sciences*, 182(1):139–155.

Mengshoel, O. J. and Goldberg, D. E. (2008). The crowding approach to niching in genetic algorithms. *Evolutionary computation*, 16(3):315–354.

Preuss, M. (2010). Niching the CMA-ES via nearest-better clustering. In *Proceedings of the 12th annual conference companion on Genetic and evolutionary computation*, pages 1711–1718. ACM.

Preuss, M. (2012). Improved topological niching for real-valued global optimization. In *Applications of Evolutionary Computation*, pages 386–395. Springer.

Qu, B.-Y. and Suganthan, P. N. (2010). Novel multimodal problems and differential evolution with ensemble of restricted tournament selection. In *Evolutionary Computation (CEC), 2010 IEEE Congress on*, pages 1–7. IEEE.

Qu, B.-Y., Suganthan, P. N., and Das, S. (2013). A distance-based locally informed particle swarm model for multimodal optimization. *Evolutionary Computation, IEEE Transactions on*, 17(3):387–402.

Qu, B.-Y., Suganthan, P. N., and Liang, J.-J. (2012). Differential evolution with neighborhood mutation for multimodal optimization. *IEEE transactions on evolutionary computation*, 16(5):601–614.

Roy, S., Islam, S. M., Das, S., and Ghosh, S. (2013). Multimodal optimization by artificial weed colonies enhanced with localized group search optimizers. *Applied Soft Computing*, 13(1):27–46.

Schoeman, I. L. and Engelbrecht, A. P. (2010). A novel particle swarm niching technique based on extensive vector operations. *Natural Computing*, 9(3):683–701.

Shir, O. M. and Bäck, T. (2009). Niching with derandomized evolution strategies in artificial and real-world landscapes. *Natural Computing*, 8(1):171–196.

Shir, O. M., Emmerich, M., and Bäck, T. (2010). Adaptive niche radii and niche shapes approaches for niching with the CMA-ES. *Evolutionary Computation*, 18(1):97–126.

Singh, G. and Deb, K. (2006). Comparison of multi-modal optimization algorithms based on evolutionary algorithms. In *Proceedings of the Genetic and Evolutionary Computation Conference (GECCO-2006), New York*, pages 1305–1312. Proceedings of the Genetic and Evolutionary Computation Conference (GECCO-2006), New York.

Stoean, C., Preuss, M., Stoean, R., and Dumitrescu, D. (2010). Multimodal optimization by means of a topological species conservation algorithm. *Evolutionary Computation, IEEE Transactions on*, 14(6):842–864.

Stoean, C. L., Preuss, M., Stoean, R., and Dumitrescu, D. (2007). Disburdening the species conservation evolutionary algorithm of arguing with radii. In *Proceedings of the 9th annual conference on Genetic and evolutionary computation*, pages 1420–1427. ACM.

Ursem, R. K. (1999). Multinational evolutionary algorithms. In *Evolutionary Computation, 1999. CEC 99. Proceedings of the 1999 Congress on*, volume 3. IEEE.

Wang, H., Moon, I., Yang, S., and Wang, D. (2012). A memetic particle swarm optimization algorithm for multimodal optimization problems. *Information Sciences*, 197:38–52.

Wessing, S., Preuss, M., and Rudolph, G. (2011). When parameter tuning actually is parameter control. In *Proceedings of the 13th annual conference on Genetic and evolutionary computation*, pages 821–828. ACM.

Wineberg, M. and Chen, J. (2004). The shifting balance genetic algorithm as more than just another island model GA. In *Genetic and Evolutionary Computation–GECCO 2004*, pages 318–329. Springer.

Yao, J., Kharma, N., and Grogono, P. (2010). Bi-objective multipopulation genetic algorithm for multimodal function optimization. *Evolutionary Computation, IEEE Transactions on*, 14(1):80–102.