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

Ali Ahrari   aliahrari1983@gmail.com
Department of Mechanical Engineering, Michigan State University, East Lansing,
MI 48824, USA

Kalyanmoy Deb   kdeb@egr.msu.edu
Department of Electrical and Computer Engineering, Michigan State University,
East Lansing, MI 48824, 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 empirically verified on some available test problems. They often rely on some particular assumptions associated with the distribution, shape, and size of the basins, which can seldom be made in practical optimization problems. This study utilizes several existing concepts and techniques, such as taboo points, normalized Mahalanobis distance, and the Ursem's hill-valley function in order to develop a new tool for multimodal optimization, which does not make any of these assumptions. In the proposed method, several subpopulations explore the search space in parallel. Offspring of a subpopulation are forced to maintain a sufficient distance to the center of fitter subpopulations and the previously identified basins, which are marked as taboo points. The taboo points repel the subpopulation to prevent convergence to the same basin. A strategy to update the repelling power of the taboo points is proposed to address the challenge of basins of dissimilar size. The local shape of a basin is also approximated by the distribution of the subpopulation members converging to that basin. The proposed niching strategy is incorporated into the covariance matrix self-adaptation evolution strategy (CMSA-ES), a potent global optimization method. 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 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 using the available and the introduced performance measures, RS-CMSA emerges as the most successful method when robustness and efficiency are 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, optimization of practical problems by using meta-heuristic methods has gained much interest (Rechenberg, 2000; Coello and Lamont, 2004), where the problem landscape may exhibit challenging features such as multimodality, discontinuity, ill-condition, and correlation (Hansen et al., 2009). On the other hand, practical considerations have evolved the flexibility of optimization methods in order to solve problems in which more than one objective is pursued (multi-objective optimization) (Deb, 2001), uncertainties exist (robust optimization) (Beyer and Sendhoff, 2007), or distinct good solutions are desired (multimodal optimization) (Das et al., 2011). Finding a set of good solutions, instead of a single one, may provide a variety of distinct and reasonable alternatives for the decision maker. This allows him/her to consider some other factors that were possibly overlooked during the mathematical modeling of the problem, to select the final solution. Moreover, finding all the high-fitness optima might be inherently critical. A typical example is finding all the resonance frequencies that lead to high vibration amplitude (Das et al., 2011). 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, without which diversity of the nondominated set significantly declines and a good representation of the Pareto front cannot be reached (Deb, 2001).

Multimodal optimization is usually achieved by using a diversity preservation strategy, called *niching*, incorporated to a global optimization method, which we call the *core algorithm*, to enable parallel convergence to different minima. Early niching methods were originally proposed for genetic algorithms (GAs), including crowding (De Jong, 1975) and fitness sharing (Goldberg and Richardson, 1987). Several subsequent studies analyzed and developed the niching methods in realm of GAs (Mahfoud, 1995; Sareni and Krahenbuhl, 1998; Mengshoel and Goldberg, 2008). Similar or new niching strategies were also incorporated to other metaheuristics such as particle swarm optimization (PSO) (Qu et al., 2013; Schoeman and Engelbrecht, 2010) and differential evolution (DE) (Basak et al., 2013; Biswas et al., 2014).

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, which reduces the fitness of similar individuals that share a niche. A similar idea is followed in some other techniques including clearing (Pétrowski, 1996), in which inferior individuals within the niche distance are eliminated, and also clustering. A comprehensive review of classical niching methods for continuous parameter optimization can be found in the work of Singh and Deb (2006), and a review paper by Das et al. (2011).

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 the subsequent restarts avoid such regions. The technique relies on definition of a threshold distance, within which the fitness function is degraded. Species conserving genetic algorithm (SCGA) (Li et al., 2002) iteratively identifies sufficiently distant fittest individuals as seeds and allocates close individuals, with respect to a threshold distance, to those seeds. A similar procedure is followed in niching with evolution strategies (Shir and

Bäck, 2009), in which classification into niches is performed iteratively over the whole population, with respect to a preset niche radius parameter.

Multimodal optimization by artificial immune networks (AIN) (de Castro and Timmis, 2002) resembles a nested evolution strategy, where evolution is performed within the subpopulation and the elite of each subpopulation survives to the next generation. The niching strategy employs the Euclidean distance between the elites as affinity of the subpopulations and may suppress inferior ones if their affinity is less than the suppression threshold. New subpopulations are randomly generated to replace the suppressed subpopulations. The user should set the value of the suppression threshold (de Castro and Timmis, 2002), or the rate at which the new subpopulations are introduced (Freschi and Repetto, 2006). The latter case can be interpreted as a radius-free niching method; nevertheless, empirical results from AIN and comparison with other multimodal optimization methods are scarce.

In the 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 the subpopulations is limited to the sampling step, which forces 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 the high amount of tuning effort.

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, in which the descendant competes with the closest among the k randomly selected individuals (Das et al., 2011). A radius-free variant of fitness sharing with an asymmetric sharing function was proposed in another study (van der Goes et al., 2008). This strategy assigns a niche radius to each individual which undergoes self-adaptation.

Stoean et al. (2007) improved SCGA (Li et al., 2002) 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*, the idea of which was proposed by Ursem (1999). This function generates and evaluates a number of samples along the line segment connecting the individuals. If a sample has a lower fitness than both individuals, then it is concluded that those individuals belong to different niches; otherwise, they are assumed to share the same niche. Although the dependency on the user-tuned niche radius parameter was eliminated, the iterative calls of the mentioned function appeared to be costly. Additionally, it seems that the performance of the proposed algorithm depends on well-tuning of some parameters including mutation and crossover rates, since these parameters were tuned for each problem independently. The same authors revised the method later (Stoean et al., 2010), but the challenge of performance sensitivity to the control parameters remained unsolved.

Li (2010) proposed different variants of particle swarm optimization (PSO) with ring topology, which restricts the interaction 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 the Euclidean distance). This method requires only the population size to be tuned and was demonstrated to surpass nine other niching methods, including lbest-PSO, proposed by Li (2010). In 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 a considerable improvement.

Some recent studies perform 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 the 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 multipopulation 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 the 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 seventh among the 15 methods participating in the CEC2013 competition on multimodal optimization (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 in addition to the global minimum. The niche radius was coupled to the mutation strength and adjusted such that every niche consists of ten individuals. The Mahalanobis distance metric, a scaled version of the Euclidean distance metric, was employed to identify the limit of a niche, which allows for an elliptic instead of spherical estimation of the niche shape; however, using a fixed and rather small number of individuals per desired niche can be a drawback when there are many nondesirable local minima, since CMA-ES needs a large population size to find the global minimum in such landscapes (Ahrari and Shariat-Panahi, 2015). The absence of recombination, one of the main operators in ESs, and having only one parent per niche, in comparison with the recommended setting

in CMA-ES (Hansen and Ostermeier, 2001; Beyer and Sendhoff, 2008), can drastically deteriorate the global search capability of the core algorithm.

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 first among the niching methods participating in CEC2013 competition on multimodal optimization (Li et al., 2013a). Particularly, for composite functions in higher dimensions, it outperformed the closest competitors by a clear margin.

Fieldsend (2014) proposed the niching migratory multiswarm optimizer (NMMSO), in which multiple swarms search the space. The goal of NMMSO is that each swarm searches a distinct basin. NMMSO allows a particle in vicinity of another basin to migrate from the parent swarm and form a new swarm. NMMSO can merge two swarms if they are searching the same basin as well. NMMSO is dynamic in the number of swarms and relatively outperformed NEA2 in the CEC2015 competition on multimodal optimization (Li et al., 2015).

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; hence, radius-free niching methods show a significant advantage over those that depend on a user-tuned niche radius parameter. In most radius-based techniques with fixed niche radius, the default value of the niche radius is set based on the recommendation of Deb and Goldberg (1989), which assumes the basins are almost uniformly distributed in the entire search space. 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) do not make such explicit assumptions of the fitness landscape.

Regardless of the underlying idea, niching methods exploit some distance measures which determine whether two solutions share the same basin. Quite often, the Euclidean distance is measured to group individuals to different niches, according to which closer individuals are more 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 considerably 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 a great extent, even though in general, basins might have any arbitrary shape.

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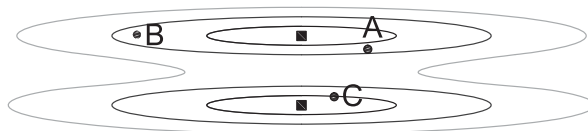
Figure 1: Contour lines around two minima (bold squares). Point C is closer to point A than point B, but A and B share the same basin, while C belongs to a different basin.

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 or the Vincent function; however, the number of global minima exponentially grows with respect to the problem dimension, which practically hinders employing dimensions higher than three. 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 as 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 nonuniform distribution of minima.

In an attempt to introduce more challenging benchmarks for multimodal optimization, Qu and Suganthan (2010) introduced a procedure to pose functions with several global minima using the weighted sum of some basic functions, in which the 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 the high computation cost of a function evaluation if many global minima are desired, since each basic function should be called independently. This method was also employed to compose some scalable test problems for the CEC2013 special session on multimodal optimization (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.

1.4 Contribution of This Study

This study develops a novel niching method for multimodal optimization, called co-variance matrix self-adaptation with repelling subpopulations (RS-CMSA). RS-CMSA adapts and reformulates multiple existing concepts from different methods, as summarized in Table 1. In comparison with the existing methods for multimodal optimization, RS-CMSA can learn the relative size and possibly the shape of the basins, and properly handle the challenge of noncircular basins, as depicted in Figure 1. Limitations of RS-CMSA are only those imposed by the core algorithm, such as continuity of the design parameters. No assumption on the distribution of global minima, the shape, or the size of the basins is required. RS-CMSA has no niche radius parameters and for all other

Table 1: Summary of the contributions of this study with respect to existing concepts.

Concept	Available form	This study
Taboo regions (Glover, 1986) (Siarry and Berthiau, 1997) (Ahrari et al., 2009)	The size of the taboo regions is fixed or decreases dynamically.	Taboo regions are adaptively resized and reshaped.
	Taboo regions are similar for all solutions.	Taboo regions differ for different subpopulations.
	The size of the taboo regions is independent of the basin size.	The size of the taboo region is adapted to the basin size.
	Taboo regions are the recently visited regions.	Taboo regions are the previously identified basins and the center of fitter subpopulations.
	All taboo regions are checked.	Only critical taboo regions are checked.
Core algorithm (CMSA-ES) (Beyer and Sendhoff, 2008)	Comma selection is employed.	Elite selection is employed.
	The global step size is adapted by simple averaging.	A bias compensation is applied in adaptation of the global step size.
Restarts (Auger and Hansen, 2005)	Independent restarts are performed and population size increases.	Outcome of a restart affects the subsequent restarts. (Similarly employed)
Hill-valley function (Ursem, 1999)	Checks whether two solutions share the same basin.	(Similarly employed)
Semi-random initialization (Ahrari et al., 2009)	Initialization maximizes the initial distance to other subpopulations.	Initialization maximizes the initial distance to other subpopulations and the archived points.

parameters, the default values are shown to be robust provided that a rough estimate for the expected/desired number of minima is given.

At a secondary level, a general framework to quantify performance sensitivity of multimodal optimization methods to the expected/desired number of minima is provided; and a performance measure, called robust mean peak ratio (*RMPR*), is developed to encompass such sensitivity in comparison.

The rest of this article is organized as follows. Section 2 elaborates different components of RS-CMSA in detail. In Section 3, a few descriptive experiments are performed to demonstrate how the method can tackle the challenge of differently sized basins. In Section 4, RS-CMSA is compared with some of the state-of-the-art niching methods on the 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, σ_{mean}, based on the concept of cumulation, and a few other heuristics to update the covariance matrix ($\mathbf{C}$). By adapting the covariance matrix, distribution of sampled solutions in CMA-ES may gradually conform to the global/local structure of the fitness landscape in order to handle ill-conditioned problems efficiently (Hansen and Kern, 2004). One may speculate that this property can be particularly utilized to learn the shape of the basin to which the algorithm is converging, in order to cope with the challenge of non-spherical basins (see Figure 1), as investigated in a former study (Shir et al., 2010).

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 fewer 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 concept of 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_s subpopulations of size λ ($\mathbf{P}_i$, $i = 1, 2, \ldots, N_s$), which explore the search space in parallel. Every subpopulation has its own mutation parameters (σ_{mean_i}, $\mathbf{C}_i$), center ($\boldsymbol{x}_{mean_i}$) and elite members. Members of each subpopulation must maintain a sufficient distance from some specific points in the search space, called *taboo points*. The set of taboo points for $\mathbf{P}_i$ is the union of:

- Previously identified minima ($\boldsymbol{y}_m$, $m = 1, 2, \ldots, M$), which are stored in an array called **Archive**, unless $\mathbf{P}_i$ is better than $\boldsymbol{y}_m$.

- The centers of fitter subpopulations.

The fitness of a subpopulation is measured by the fitness of its best individual. It is remarkable that the set of taboo points is subpopulation-dependent, and a better subpopulation has a smaller taboo set.

A sampled solution is called *taboo acceptable* if it satisfies the distance criterion with respect to all the taboo points; otherwise, it is discarded without evaluation. This process goes on until λ taboo acceptable solutions are generated. The overall effect of such rejection is reshaping distribution of solutions so that subpopulations do not search previously explored regions or those which are being explored by other subpopulations. This distance-based rejection defines taboo regions around the taboo points, where a subpopulation may not produce any offspring. The taboo regions are ellipsoids whose

centers lie on the taboo points. The shape of the taboo regions for $\mathbf{P}_i$ is exclusively determined by the strategy parameters of $\mathbf{P}_i$, while the size is affected by a property of the taboo point, the *normalized taboo distance* ($\hat{d}_k$) as well. $\hat{d}_k$'s are adapted so that a larger basin has a larger $\hat{d}_k$, which enables the RS-CMSA to handle the challenge of unequally sized basins efficiently. 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 idea of taboo regions may remind of the tabu search (TS) (Glover, 1986), in which a short-term memory helps the algorithm avoid the recently visited regions (Siarry and Berthiau, 1997); however, the size and the shape of the regions are preset in TS while RS-CMSA can learn the relative size and possibly the shape of the basins. The archived solutions in RS-CMSA, in contrast to the memory in TS, contains minima identified during the optimization process, disregarding how recent the identification has occurred. More importantly, taboo regions are coupled to the mutation strength of the subpopulation, which means they shrink as the subpopulation converges. Even for a fixed iteration, the taboo regions perceived by different subpopulations differ in number, shape and size.

2.3 Evolution of Subpopulations

In RS-CMSA, evolution of a subpopulation consists of sampling, selection, and recombination, in which the diversity preservation strategy is applied to the sampling stage. The augmentations applied to RS-CMSA, in comparison with the original CMSA-ES, are discussed in this section.

2.3.1 Distance Metric

A distance metric should be defined to check whether the j-th solution ($\mathbf{x}_{ij}$) of $\mathbf{P}_i$ satisfies the distance condition with respect to the k-th taboo point ($\mathbf{y}_k$). A simple way is to calculate the normalized Euclidean distance (d^E_{ij-k}) as follows:

$$d^E_{ij-k} = \frac{\sqrt{(\mathbf{x}_{ij} - \mathbf{y}_k)^T (\mathbf{x}_{ij} - \mathbf{y}_k)}}{\sigma_{\mathrm{mean}_i} \bar{u}_i}, \quad \bar{u}_i = \sqrt[D]{\prod_{r=1}^{D} u_{ir}}, \tag{1}$$

where u_{ir}'s are the square roots of the eigenvalues of $\mathbf{C}_i$. $\bar{u}_i$ encompasses the effect of $\mathbf{C}_i$ on the mutation strength, although the mutation strength is mainly controlled by σ_{mean_i}.

The shape of a taboo region can be coupled to the covariance matrix by using the Mahalanobis distance metric. The Mahalanobis distance (d_M) of the point x from the center $\mathbf{x}_{\mathrm{mean}}$ with respect to the covariance matrix $\mathbf{\Sigma}$ is defined as follows:

$$d_M = \sqrt{(\mathbf{x} - \mathbf{x}_{\mathrm{mean}})^T \mathbf{\Sigma}^{-1} (\mathbf{x} - \mathbf{x}_{\mathrm{mean}})}. \tag{2}$$

The Mahalanobis distance can be interpreted as the scaled version of the Euclidean distance, where the scaling factors and directions are determined by $\mathbf{\Sigma}$. This definition can be employed to define the normalized Mahalanobis distance (d^M_{ij-k}) between $\mathbf{x}_{ij}$ and $\mathbf{y}_k$ with respect to the parameters of the sampling distribution ($\sigma^2_{\mathrm{mean}_i} \mathbf{C}_i$) as follows:

$$d^M_{ij-k} = \sqrt{(\mathbf{x}_{ij} - \mathbf{y}_k)^T (\sigma^2_{\mathrm{mean}_i} \mathbf{C}_i)^{-1} (\mathbf{x}_{ij} - \mathbf{y}_k)}. \tag{3}$$

Contours of points with identical normalized Euclidean distance from $\mathbf{y}_k$ form concentric spheres whose centers lie on $\mathbf{y}_k$. For the case of the normalized Mahalanobis

distance, they form concentric ellipsoids, the shape of which is defined by $\mathbf{C}_i$. In both cases, the normalized distance increases inversely proportional to σ_{mean_i}, which means for similar values of the absolute distance ($|\boldsymbol{x}_{ij} - \boldsymbol{y}_k|$), the normalized distance (d_{ij-k}^{EM}) increases as the subpopulation converges, which can be beneficial for landscapes with multiple global minima close to each other. The superscript EM denotes that the normalized distance can be either Euclidean or Mahalanobis.

2.3.2 Rejection Rule

The rejection probability of the sampled solution $\boldsymbol{x}_{ij}$ due to proximity to the taboo point $\boldsymbol{y}_k$ is defined as follows:

$$P_{\text{trej}}(\boldsymbol{x}_{ij}, \boldsymbol{y}_k, \sigma_{\text{mean}_i}^2 \mathbf{C}_i, \hat{d}_k) = \begin{cases} 0 & \text{if } d_{ij-k}^{EM} > \hat{d}_k \\ 1 & \text{otherwise} \end{cases}, \tag{4}$$

which means the rejection probability depends on the parameters of the subpopulation ($\sigma_{\text{mean}_i}^2 \mathbf{C}_i$) and $\hat{d}_k$ as well. It is notable that one may define a stochastic rejection rule, such that the rejection probability gradually reduces to zero when d_{ij-k}^{EM} increases; however, for the case of simplicity, the deterministic rejection rule is followed in this study. Note that any single taboo point can reject $\boldsymbol{x}_{ij}$ independently. The overall taboo rejection probability is thus computed as follows:

$$P_{\text{Trej}}(\boldsymbol{x}_{ij}, \sigma_{\text{mean}_i}^2 \mathbf{C}_i) = 1 - \prod_{k=1}^{K_i} \left(1 - P_{\text{trej}}(\boldsymbol{x}_{ij}, \boldsymbol{y}_k, \sigma_{\text{mean}_i}^2 \mathbf{C}_i, \hat{d}_k)\right), \tag{5}$$

where K_i is the number of the taboo points for $\mathbf{P}_i$. $\boldsymbol{x}_{ij}$ is accepted and evaluated if it satisfies the distance condition with respect to all the taboo points; otherwise, it is rejected and a new candidate solution is generated. This process continues until λ solutions are accepted.

If the normalized Euclidean distance is employed, the taboo region around $\boldsymbol{y}_k$ perceived by $\mathbf{P}_i$ is a ball of radius $\sigma_{\text{mean}_i} u_i \hat{d}_k$, while for the Mahalanobis case, it is an ellipsoid with main axes of $2\sigma_{\text{mean}_i} u_{ir} \hat{d}_k, r = 1, 2, \ldots, D$. This means for the latter case, the shape of the taboo region is defined by the covariance matrix, which is assumed to be able to gradually conform to the shape of the basin to which the subpopulation is converging. Consequently, the taboo region may gradually conform to the basin shape to overcome the challenge of nonspherical basins, which was highlighted in Section 1.2. For the rest of this study, the distance metric is assumed to be the Mahalanobis, unless mentioned otherwise.

Taboo regions are illustrated for a typical case in Figure 2, where two subpopulations face two archived points ($\boldsymbol{y}_1, \boldsymbol{y}_2$) with $\hat{d}_1 = 0.8$ and $\hat{d}_2 = 1.4$. $\mathbf{P}_1$ is better than $\mathbf{P}_2$ and both are worse than the archived points. The subpopulations have a default normalized taboo distance of $\hat{d}_0 = 1$. The boundary of a taboo region, depicted by a thick continuous line, forms an ellipsoid, on the center of which lies the taboo point. The taboo regions perceived by $\mathbf{P}_1$ (see Figure 2a) and $\mathbf{P}_2$ (see Figure 2b) differ in three aspects:

- The number of taboo regions is different for $\mathbf{P}_1$ and $\mathbf{P}_2$, since $\boldsymbol{x}_{\text{mean}_1}$ is a taboo point for $\mathbf{P}_2$ but $\boldsymbol{x}_{\text{mean}_2}$ is *not* a taboo point for $\mathbf{P}_1$.

- For each subpopulation, the shape of the taboo regions is similar to the shape of the isodensity contours of the sampling distribution of that subpopulation.

- The size of a taboo region for a subpopulation is proportional to the normalized taboo distance of the corresponding taboo point; therefore, for either of the
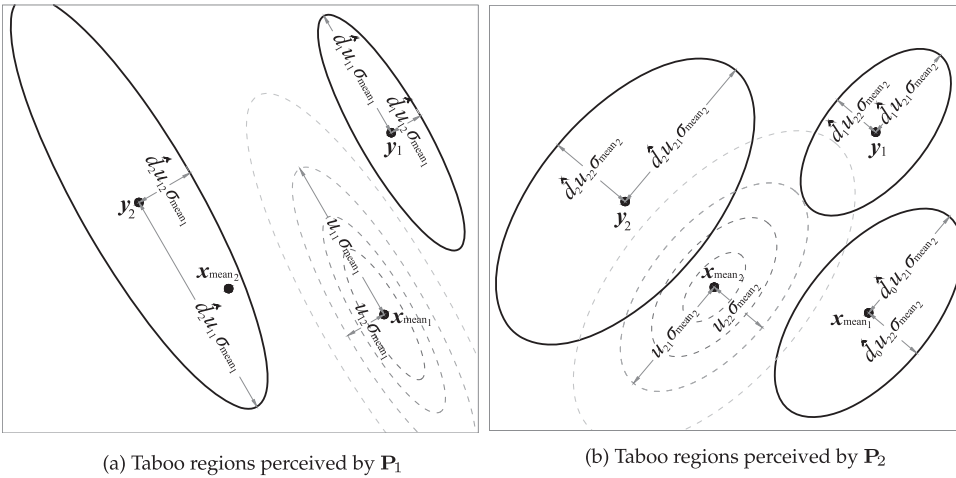
(a) Taboo regions perceived by $\mathbf{P}_1$ (b) Taboo regions perceived by $\mathbf{P}_2$

Figure 2: Taboo regions perceived by $\mathbf{P}_1$ and $\mathbf{P}_2$ ($\hat{d}_0 = 1$), in vicinity of two taboo points ($\mathbf{y}_1$ and $\mathbf{y}_2$) with $\hat{d}_1 = 0.8$ and $\hat{d}_2 = 1.4$. u_{ir} is the square root of the r-th eigenvalue of $\mathbf{C}_i$.

subpopulations, the taboo region of $\mathbf{y}_2$ is $(\hat{d}_2/\hat{d}_1)^2 = 1.75^2$ times larger than the taboo region of $\mathbf{y}_1$.

Sampled solutions that fall inside the taboo regions are rejected without evaluation. Rejection because of proximity to a taboo point results in an anisotropy in distribution of offspring such that $\mathbf{x}_{\mathrm{mean}}$ is biased to move farther from the taboo regions; however, whether it actually moves that way depends on the fitness landscape of the objective function as well.

2.3.3 Selection and Recombination

While contemporary evolution strategies, including CMSA, employ the comma selection scheme (Bäck et al., 2013), preserving the elite member has been preferred for niching with ESs (Debski et al., 2008). Some previous studies demonstrated superiority of the plus scheme in multimodal optimization for the special cases of $(1+\lambda)$-ES and $(1,\lambda)$-ES (Shir and Bäck, 2007, 2009), in which the global step-size is updated iteratively while the covariance matrix is updated only if the selected offspring is better than the parent.

RS-CMSA-ES employs a more general elite preservation scheme, in which $0 \leq N_{\mathrm{elt}} \leq \mu$ parents may survive to the next generation. The surviving parents may participate in subsequent update of the subpopulation parameters ($\mathbf{x}_{\mathrm{mean}}$, $\mathbf{C}_i$, σ_{mean}). Similar to the concept of finite life span (Schwefel and Rudolph, 1995), in which the life span of offspring is limited to κ generations, or gradual fitness decay (Ahrari and Kramer, 2015), in which the fitness of individuals declines as they grow older, this general case of elitism enables arbitrary trade-off between advantages and disadvantages of both extreme selection schemes, and it can even surpass both extremes in some landscapes (Ahrari and Kramer, 2015).

In RS-CMSA, the number of parents is more than one and recombination is performed. Some parents may be selected from the surviving elites and the rest from the recently generated offspring. The employed approach for including the elites in the

update of the strategy variables, in addition to the object variables, follows the main goals of elitism. When the traditional concept of mutative strategy parameter control is employed, the underlying goal of elitism is to preserve the successful mutations, those that resulted in better offspring, so that such success may be repeated in the subsequent generations. This motivation can be extended to *successful directions* in a similar way: preserving successful directions $(z_{ij} = (x_{ij} - x_{\text{mean}})/\sigma_{ij})$ of the surviving parents, which is employed in this study. At the end of each generation, the solutions (the union of λ recently generated and N_{elt} surviving solutions from the previous iteration) are sorted in increasing order of the function value. Parameters of the subpopulations are then updated as follows:

$$\mathbf{C}_i \leftarrow \left(1 - \frac{1}{\tau_c}\right) \mathbf{C}_i + \frac{1}{\tau_c} \sum_{j=1}^{\mu} w_j \left(z_{ij}^T z_{ij}\right), \tag{6a}$$

$$\sigma_{\text{mean}_i} \leftarrow \frac{\sigma_{\text{mean}_i} \left(\prod_{j=1}^{\mu} \sigma_{ij}^{w_j}\right)}{\left(\prod_{j=1}^{\lambda+N_{\text{elt}}} \sigma_{ij}^{\frac{1}{\lambda+N_{\text{elt}}}}\right)}, \tag{6b}$$

$$x_{\text{mean}_i} \leftarrow \sum_{i=1}^{\mu} w_j x_{ij}, \tag{6c}$$

$$\omega_j = \frac{\ln(\mu + 1) - \ln(j)}{\sum_{j=1}^{\mu} (\ln(\mu + 1) - \ln(j))}, j = 1, 2, \ldots, \mu, \tag{6d}$$

where ω_j's are logarithmically decreasing weights (Hansen and Ostermeier, 2001), $\mu = \lfloor \lambda/2 \rfloor$ is the number of parents in the subpopulation, and $\tau_c = 1 + D(D+1)/\mu$ is the adaptation interval constant for $\mathbf{C}_i$ (Beyer and Sendhoff, 2008). Equivalently, $\frac{1}{\tau_c}$ can be interpreted as the learning rate for the covariance matrix. If there are bounds on the search space, x_{mean_i} is relocated to the nearest point inside the search space. Note that parameters of the surviving elites $(x_{ij}, z_{ij}, \sigma_{ij})$ from the previous iteration are included in Equations 6a, 6b, and 6c provided that they are among the best μ solutions of the current iteration. The N_{elt} best solutions of this iteration will survive to the next iteration accordingly. By default, $N_{\text{elt}} = \max \{1, \lfloor 0.15\lambda \rfloor\}$.

A few deviations from the original CMSA-ES (Beyer and Sendhoff, 2008) are observed in the update scheme. First, the parental subpopulation size has been doubled, however, since weighted recombination is used, the effective number of parents (μ_{eff}) (Hansen et al., 2013) is similar for both cases. The adaption of the global mutation strength has also been revised to compensate for anisotropic distribution of σ_{ij}'s in the subpopulation, caused by rejection of certain samples, such that under random selection of the parents, the expected change of σ_{mean_i} is zero. Ignoring such a compensation results in a bias towards increasing σ_{mean_i}. The reason is that a greater σ_{ij} is more likely to produce a sample farther from x_{mean_i}, which is more likely to lie outside the taboo regions. A larger σ_{mean_i} means larger taboo regions, and preference of greater σ_{ij}'s again. This ever increasing σ_{mean_i} risks divergence of $\mathbf{P}_i$. Although some statistical criteria will be provided to terminate nonconverging subpopulations, such subpopulations waste many evaluations.

One may speculate such bias compensation for the covariance matrix update is also required. For example, the following equation should be used instead of Equation 6a:

$$\mathbf{C}_i \leftarrow \mathbf{C}_i + \frac{1}{\tau_c} \left(\sum_{j=1}^{\mu} w_j \left(z_{ij}^T z_{ij} \right) - \frac{1}{\lambda + N_{\text{elt}}} \sum_{j=1}^{\lambda + N_{\text{elt}}} \left(z_{ij}^T z_{ij} \right) \right). \tag{7}$$

The subpopulation should be terminated if the updated $\mathbf{C}_i$ has a negative eigenvalue. We did not notice any substantial effect using this modification in our preliminary results, probably because the size of a taboo region is mainly controlled by σ_{mean_i}. Therefore, such bias compensation for the covariance matrix is not performed in this study.

2.4 Restarts with Increasing Population Size

The optimal population size for CMA-ES and presumably CMSA-ES varies from a small value ($\lfloor 4 + 3 \ln(D) \rfloor$) for unimodal functions to several hundreds for highly multimodal ones (Ahrari and Shariat-Panahi, 2015). 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 recommended minimum value, the population size is doubled after each restart. The goal is to reach a restart with a sufficiently large population size which can identify the global minimum of highly multimodal functions. The drawback emerges when a large population size is required. In such a situation, several restarts should be performed so that the population size becomes large enough to be able to find the global minimum. Since the restarts are independent, the evaluation effort in the earlier restarts is actually wasted, resulting in considerable increase in the number of evaluations (Ahrari and Shariat-Panahi, 2015). More sophisticated schemes were later proposed to gain more feedback from the previous restarts (Wessing et al., 2011) or to enhance local search capabilities of the method (Hansen, 2009). The method proposed in another study (Ahrari and Shariat-Panahi, 2015) 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 requires 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 can be situations in which this strategy even surpasses a well-tuned but fixed population size. For example, basins that are easy to identify can be detected in the early restarts, when the population size is small, while the subsequent restarts with a larger population size aim at finding harder minima. When all subpopulations in a restart are terminated, λ is doubled unless the remaining evaluation budget (FE_{remained}) is not sufficient for the subsequent restart with the 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 a sufficient budget to converge:

$$\begin{cases} \lambda \leftarrow 2\lambda, \ N_s \leftarrow N_s^0 & \text{if } FE_{\text{remained}} > 2\lambda N_s^0 \bar{I}_{\text{used}} \\ \lambda \leftarrow \lambda, \ N_s^0 \leftarrow \max \left\{ 1, \left\lfloor \frac{FE_{\text{remained}}}{\lambda \bar{I}_{\text{used}}} \right\rfloor \right\} & \text{if } FE_{\text{remained}} < \lambda N_s^0 \bar{I}_{\text{used}} \\ \lambda \leftarrow \left\lfloor \lambda \left(\frac{FE_{\text{remained}}}{N_s^0 \bar{I}_{\text{used}}} \right) \right\rfloor, \ N_s \leftarrow N_s^0 & \text{otherwise,} \end{cases} \tag{8}$$

where $\bar{I}_{\text{used}}$ is the average number of iterations used by the subpopulations in the previous restart and N_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:

- *Condition C*: the condition number of **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*.

The terminated subpopulations are analyzed at the end of the restart to check whether they specify new basins. The newly identified basins are added to **Archive** and the normalized taboo distance of each archived solution is updated based on the attraction power of the corresponding basin.

2.5 Adaptation of the Normalized Taboo Distance

A universal value for the normalized taboo distance may work well if all the basins are roughly similar in size; otherwise, it would be too small for large basins and too large for the others. The magnitude of $\hat{d}_k$ can be considered as the repelling power of the taboo point 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 every terminated subpopulation is 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 the best solution in **Archive** minus ϵ_f (target tolerance on the objective function). All undesirable solutions are discarded. For all the remaining (desirable) solutions ($\tilde{y}_i, i = 1, 2, \ldots, I$), the hill-valley function of Ursem (1999) is utilized in order to determine whether a solution refers to a new basin. The line search is rendered using Golden Section Search with a maximum of ten function evaluations. The solution 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 each solution in **Archive** is updated based on the number of subpopulations (N_{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 the subpopulations during the restart. $0 \leq \alpha_{new} \leq 1$ 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 $N_{rep_m} = 0$, $\hat{d}_m$ is decreased since it might be unnecessary large. A greater N_{rep_m} indicates convergence of more subpopulations to the m-th archived solution, which demonstrates that the current value of $\hat{d}_m$ is too small. $\tau_{\hat{d}}$ specifies the learning rate of the normalized taboo distance. Default values are $\tau_{\hat{d}} = 1/\sqrt{D}$ and $\alpha_{new} = 0.5$. $\hat{d}_0$ is equal to the 25th percentile of normalized taboo distances of the solutions stored in **Archive**. The justification is that basins detected in the latest restart are probably harder

---

Algorithm 1: Updating **Archive** and the normalized taboo distances

---

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 (α_{new})

Result: Updated **Archive**

initialization

$N_{\text{rep}_m} \leftarrow 0, m = 1, 2, \ldots, M$

for $i \leftarrow 1$ to I **do**

  Sort the points in **Archive** (y_m) based on their Euclidean distance to $\widetilde{y}_i$.

  $isnew \leftarrow 1$

  for all the archived solutions **do**

    if y_m and $\widetilde{y}_i$ share the same basin **then**

      if $\widetilde{y}_i$ is fitter than y_m **then**

        | Replace y_m with $\widetilde{y}_i$

      end

      $N_{\text{rep}_m} \leftarrow N_{\text{rep}_m} + 1$

      $isnew \leftarrow 0$

      Break

    end

  end

  if $isnew = 1$ **then**

    Let $\widetilde{y}_i$ be the $(M+1)$-th solution in **Archive**

    $N_{\text{rep}_{M+1}} \leftarrow 0$

    $\hat{d}_{M+1} \leftarrow \hat{d}_0$

    $M \leftarrow M + 1$

  end

end

for $m = 1$ to M **do**

  if $N_{\text{rep}_m} > \alpha_{\text{new}}(I/M)$ **then**

    $\hat{d}_m \leftarrow \hat{d}_m \left(1 + N_{\text{rep}_m} - \alpha_{\text{new}}(I/M)\right)^{(+\tau_{\hat{d}})}$

  else

    $\hat{d}_m \leftarrow \hat{d}_m \left(1 - N_{\text{rep}_m} + \alpha_{\text{new}}(I/M)\right)^{(-\tau_{\hat{d}})}$

  end

end

---

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 $\hat{d}_0 = 1$.

2.6 Boosting Time Efficiency

The main time complexity of the proposed niching strategy originates from the sampling part, according to which the distance of each sampled solution should be checked against all the taboo points. It is also likely that most sampled solutions are rejected because of the distance criteria (Equation 4), especially in the early iterations of each restart when the mutation strength is great. Each sampled solution should on average be checked against $M + (N_s - 1)/2$ taboo points, where M and N_s refer to the number of

archived points and subpopulations, respectively. For the best-case scenario, when no sampled solutions are rejected, the time complexity is proportional to N_s, assuming that $M \propto N_s$. This intensifies the time complexity of generating a taboo acceptable solution when the number of desired/estimated minima is great. Two different strategies are proposed and employed in RS-CMSA to alleviate this problem: temporary shrinkage of the taboo regions and checking for more critical taboo points earlier. These two strategies are explained in this section.

2.6.1 Temporary Shrinkage of the Taboo Regions

The fraction of the rejected sampled points is directly related to the overall size of the taboo regions. There might be situations in which the taboo regions are so huge that almost all the sampled solutions of a subpopulation are rejected. To avoid stagnation, the distance condition is loosened by temporary reduction of the normalized taboo distance of every taboo point whenever a sampled solution is rejected: $\hat{d}_k \leftarrow c_{\text{red}}\hat{d}_k, 0 < c_{\text{red}} < 1$. This modification is valid only for the current subpopulation and only for the current iteration. When λ taboo acceptable solutions are generated and evaluated, the original values of $\hat{d}_k$'s are restored. A greater value of c_{red} speeds up reduction of the normalized taboo distances and thus reduces the fraction of taboo-rejected samples; however, it may detectably suppress the effect of the proposed niching strategy. The default setting is $c_{\text{red}} = 0.99^{(1/D)}$, which means the size of each taboo region shrinks by 1% whenever a sample is rejected.

2.6.2 Checking for More Critical Taboo Points Earlier

According to Equation 4, each taboo point (y_k) may reject the sampled solution (x_{ij}) with the probability of $P_{\text{trej}}(x_{ij}, y_k, \sigma^2_{\text{mean}_i} C_i, \hat{d}_k)$. If x_{ij} violates the distance condition to any taboo point, it is rejected, and checking the distance condition with respect to the other taboo points is not required. This can save a lot of computation if the most critical taboo points, those that are most likely to reject a sampled solution, are known so that the distance condition is checked for them earlier. This means that if the sampled solution is going to be rejected, the algorithm discovers it sooner.

We propose mean rejection probability ($\bar{P}_{\text{trej}}(y_k, \sigma^2_{\text{mean}_i} C_i, \hat{d}_k)$) as a measure to quantify *criticality* of y_k for $\mathbf{P}_i$. Before $\mathbf{P}_i$ starts sampling, the criticality of the taboo points is calculated. A sampled solution is checked against the taboo points in order of their criticality. Note that the criticality measure is not a function of the sampled solution; therefore, for each subpopulation, it is calculated once per iteration and used to generate all λ the taboo acceptable solutions. Moreover, it can be safely assumed that a taboo point never rejects a sampled solution if its criticality is sufficiently small. Such a taboo point is called *non-critical*, and ignored while checking the distance condition. It will be demonstrated in Section 3.2 that most taboo points become non-critical as the subpopulation converges. By default, if $\bar{P}_{\text{trej}}(y_k, \sigma^2_{\text{mean}_i} C_i, \hat{d}_k) < 0.01$, then y_k is considered as a non-critical taboo point. A conservative estimation for $\bar{P}_{\text{trej}}(y_k, \sigma^2_{\text{mean}_i} C_i, \hat{d}_k)$ is provided in this section.

2.6.3 Calculation of the Mean Rejection Probability

Figure 3 depicts the subpopulation $\mathbf{P}$ and the taboo point y'. The origin of the auxiliary Cartesian coordinate system lies on x_{mean}, while x_1 is along the eigenvector corresponding to the greatest eigenvalue of $\mathbf{C}$. For simplicity of calculations, y' is relocated to y such that $y - x_{\text{mean}}$ is aligned with x_1 and $L = \|x_{\text{mean}} - y\| = \|x_{\text{mean}} - y'\|$. The iso-density contour line of sampling distribution that passes through an arbitrary point in the relocated taboo region has a higher value than the one passing through the corresponding
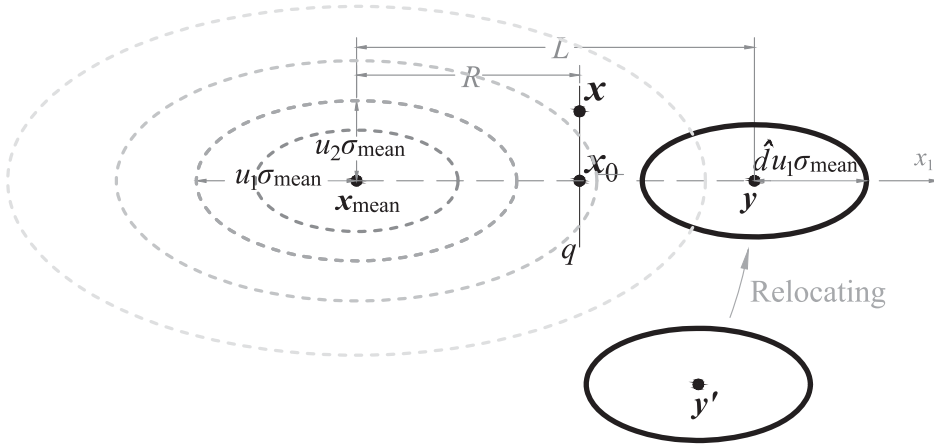
Figure 3: Subpopulation $\mathbf{P}$ samples $\boldsymbol{x}$ at the proximity of taboo point $\boldsymbol{y}'$. The taboo point is relocated to $\boldsymbol{y}$.

point in the original taboo region. This means for an arbitrary sample, the probability of falling inside the relocated taboo region is more than the original taboo region; therefore, this relocation results in a greater $\bar{P}_{\text{trej}}(\boldsymbol{y}, \sigma^2_{\text{mean}}\mathbf{C}, \hat{d})$ than the actual one.

Let $x_1, x_2, \cdots, x_D$ be coordinates of the sampled solution $\boldsymbol{x}$, which are realizations of $X_1 \sim \mathcal{N}(0, \sigma^2 u_1^2)$, $X_2 \sim \mathcal{N}(0, \sigma^2 u_2^2)$, $\cdots$, $X_D \sim \mathcal{N}(0, \sigma^2 u_D^2)$. Let $p_1(R)$ be the probability density that $\boldsymbol{x}$ lies on plane $x_1 = R$ (line $x_1 = R$ in 2D space):

$$p_1(R) = p_{X_1}(R) = \frac{1}{\sigma u_1 \sqrt{2\pi}} \exp\left(\frac{-R^2}{2\sigma^2 u_1^2}\right), \tag{9}$$

where σ is the step size of $\boldsymbol{x}$, 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 x_1 axis of the auxiliary coordinate system ($\boldsymbol{x}_0 = [x_1, 0, 0, \ldots, 0]$). Among all the points on $x_1 = R$, $\boldsymbol{x}_0$ has the highest rejection probability. The mean rejection probability of $\boldsymbol{x}_0$ when R varies from $-\infty$ to $+\infty$ is computed as follows:

$$\bar{P}_{\text{trej}}(\boldsymbol{y}, \sigma^2_{\text{mean}}\mathbf{C}, \hat{d}) = \int_{-\infty}^{+\infty} p_1(R)\, P_{\text{trej}}(\boldsymbol{x}_0, \boldsymbol{y}, \sigma^2_{\text{mean}}\mathbf{C}, \hat{d})dR, \tag{10}$$

$P_{\text{trej}}(\boldsymbol{x}_0, \boldsymbol{y}, \sigma^2_{\text{mean}}\mathbf{C}, \hat{d})$ is one inside the taboo region and zero outside it. Therefore:

$$\bar{P}_{\text{trej}}(\boldsymbol{y}, \sigma^2_{\text{mean}}\mathbf{C}, \hat{d}) = \int_{L-\hat{d}u_1\sigma_{\text{mean}}}^{L+\hat{d}u_1\sigma_{\text{mean}}} p_1(R)\, dR = \Phi\left(\frac{L + \hat{d}u_1\sigma_{\text{mean}}}{u_1\sigma_{\text{mean}}}\right) - \Phi\left(\frac{L - \hat{d}u_1\sigma_{\text{mean}}}{u_1\sigma_{\text{mean}}}\right), \tag{11}$$

where Φ computes the *cdf* of standard normal distribution. $\bar{P}_{\text{trej}}(\boldsymbol{y}, \sigma^2_{\text{mean}}\mathbf{C}, \hat{d})$ computed using Equation 11 gives an upper limit for the actual mean rejection probability of $\boldsymbol{x}$ because of the violation of the distance metric to the taboo point $\boldsymbol{y}$. It is the employed measure to calculate and compare criticality of taboo points.

When the normalized Euclidean distance metric is used (Equation 1), the taboo region is a ball of radius $\sigma_{\text{mean}}\bar{u}$. Equation 11 can still calculate the mean rejection probability in this case if the lower and upper bounds of the integral change to $L - \hat{d}\bar{u}\sigma_{\text{mean}}$ and $L + \hat{d}\bar{u}\sigma_{\text{mean}}$, respectively; therefore:

$$\bar{P}_{\text{trej}}(\boldsymbol{y}, \sigma^2_{\text{mean}}\mathbf{C}, \hat{d}) = \Phi\left(\frac{L + \hat{d}\bar{u}\sigma_{\text{mean}}}{u_1\sigma_{\text{mean}}}\right) - \Phi\left(\frac{L - \hat{d}\bar{u}\sigma_{\text{mean}}}{u_1\sigma_{\text{mean}}}\right). \tag{12}$$

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

Data: N_s, D, c_{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 $(\boldsymbol{x}_{\mathrm{mean}_i})$ and their global step size $(\sigma_{\mathrm{mean}_i} = \sigma_{\mathrm{mean}}, i = 1, 2, \ldots, N_s)$

initialization;

$i \leftarrow 1$; $N_{\mathrm{rej}} \leftarrow 0$

$\sigma_{\mathrm{mean}} \leftarrow 0.25$

$\mathbf{C} \leftarrow diag(\boldsymbol{U_x} - \boldsymbol{D_x})^2$

while $i \leq N_s$ **do**

    Sample $\boldsymbol{D_x} \leq \boldsymbol{x}_{\mathrm{mean}_i} \leq \boldsymbol{U_x}$ randomly.

    if $\forall (m \leq M) : (2\hat{d}_0 + \hat{d}_m) \leq d^M_{i-m}$ **then**

        if $\forall (q < i) : 3\hat{d}_0 \leq d^M_{i-q}$ **then**

            $i \leftarrow i+1$

            $N_{\mathrm{rej}} \leftarrow 0$

        end

    end

    $N_{\mathrm{rej}} \leftarrow N_{\mathrm{rej}}+1$

    if $N_{\mathrm{rej}} > 100$ **then**

        $\sigma_{\mathrm{mean}} \leftarrow c_{red}\sigma_{\mathrm{mean}}$

        $N_{\mathrm{rej}} \leftarrow 0$

    end

end

Let $\boldsymbol{x}'_{\mathrm{mean}_i}s$ be center of the subpopulations, with strategy parameters of $\mathbf{C}_i = \mathbf{C}, \sigma_{\mathrm{mean}_i} = \sigma_{\mathrm{mean}} \ (i = 1, 2, \ldots, N_s)$

Since the normalized distance is inversely proportional to σ_{mean}, the number of critical taboo points gradually decreases as the algorithm converges. This significantly reduces time complexity of sampling especially for large values of N_s. The employed model for computation of the mean rejection probability reminds one of the corridor model described by Schwefel (1993). In our model, the infeasible region is cut out of the feasible one whereas it is the other way around in the corridor model.

2.7 Initialization of Subpopulations

At the beginning of each restart, N_s subpopulation are initialized such that their centers lie far from each other and the archived solutions. All new subpopulations have similar initial strategy parameters: $\sigma_{\mathrm{mean}_i} = \sigma_{\mathrm{mean}}, \mathbf{C}_i = diag(\boldsymbol{U_x} - \boldsymbol{D_x})^2, i = 1, 2, \ldots, N_s$, in which function $diag(\boldsymbol{z})$ returns a diagonal matrix whose diagonal elements are the elements of $\boldsymbol{z}$. Starting from a conservatively great σ_{mean}, a candidate point for the center $(\boldsymbol{x}_{\mathrm{mean}_i})$ is randomly sampled in the search space and its normalized Mahalanobis distance to the archived solutions (d^M_{i-m}) and the centers of other subpopulations $(d^M_{i-q}, q < i)$ is calculated. If the distance condition is satisfied, the candidate is accepted as the center of $\mathbf{P}_i$, otherwise discarded, and a new candidate is sampled. If multiple consecutive tries (say, 100) fail, σ_{mean} is slightly reduced $(\sigma_{\mathrm{mean}} \leftarrow c_{red}\,\sigma_{\mathrm{mean}})$, which reduces the overall size of the taboo regions. This process continues until N_s subpopulations are generated (Algorithm 2).

Such initialization, although not critically important, helps RS-CMSA avoid taboo regions more efficiently in early iterations of a restart, especially in lower dimensions.

Algorithm 3: Pseudocode of RS-CMSA

Data: N_s^0, D, *MaxFE*
Result: One run of RS-CMSA
initialization;
Archive $\leftarrow \{\}$; $\lambda \leftarrow \lfloor 4.5 + 3 \times \ln D \rfloor$; $N_s \leftarrow N_s^0$; $FE_{\text{remained}} \leftarrow MaxFE$;
Set the value of the control parameters;
while $FE_{\text{remained}} > 0$ **do**
    Initialize N_s subpopulations using Algorithm 2
    while *there is at least one nonterminated subpopulation* **do**
        for $i \leftarrow 1 to N_s$ **do**
            if $\mathbf{P}_i$ *was not terminated* **then**
                Find and sort critical taboo points for $\mathbf{P}_i$ based on their criticality;
                Generate λ taboo acceptable solutions;
                Append the surviving elites from the previous generation;
                Update x_{mean_i}, $\mathbf{C}_i$ and σ_{mean_i} using Equation 6;
            end
            Terminate $\mathbf{P}_i$ if a stopping criterion is satisfied;
        end
    end
    Update **Archive** using Algorithm 1;
    Update λ and N_s using Equation 8;
end

In higher dimensions, contribution of Algorithm 2 subsides due to the distance concentration effect (Beyer et al., 1999); however, for dimensions commonly used for empirical evaluation of multimodal optimization methods ($D \leq 30$), the effect of Algorithm 2 should be detectable, since there is still a considerable variance among distances of randomly generated points (Beyer et al., 1999).

2.8 Pseudocode

Algorithm 3 presents the pseudocode of RS-CMSA.

2.9 Parameter Tuning

Except for the default number of subpopulations (N_s^0), all the parameters are set to their default values, as follows: $\lambda = \lfloor 4.5 + 3 \times \ln D \rfloor$, $\mu = \lfloor \lambda/2 \rfloor$, $\tau = \sqrt{0.5/D}$, $\tau_c = 1 + D(D+1)/\mu$, $\hat{d}_0 = 1$, $\tau_{\hat{d}} = \sqrt{1/D}$, $\alpha_{\text{new}} = 0.5$, $c_{\text{red}} = 0.99^{(1/D)}$, $N_{\text{elt}} = \max\{1, \lfloor 0.15\lambda \rfloor\}$ and criticality threshold $= 0.01$.

    N_s^0 should be proportional to the desired number of minima. Smaller values result 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

A few 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 these four factors:

(a) Top view of the landscape

(b) Restart #0: $\bar{u}\sigma_{\mathrm{mean}} = 0.33$

(c) Restart #1: $\bar{u}\sigma_{\mathrm{mean}} = 0.39$

(d) Restart #2: $\bar{u}\sigma_{\mathrm{mean}} = 0.35$

(e) Restart #3: $\bar{u}\sigma_{\mathrm{mean}} = 0.32$
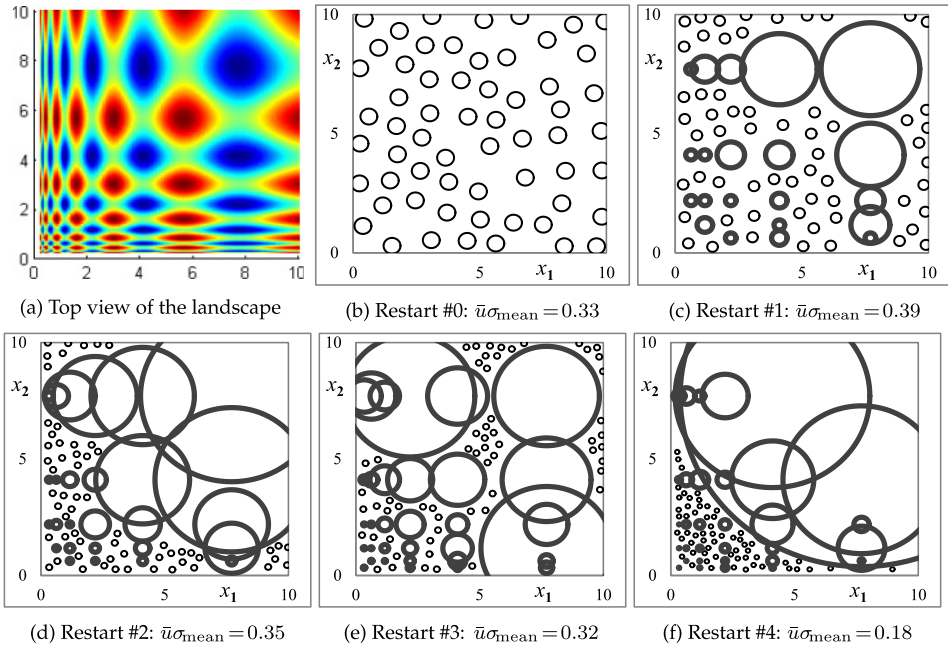
(f) Restart #4: $\bar{u}\sigma_{\mathrm{mean}} = 0.18$

Figure 4: Taboo regions around the archived points (gray circles) and the centers of the subpopulations (black circles) at different restarts, immediately after initialization.

- Contribution of proper initialization of subpopulations (Algorithm 2).

- Adaptation of the normalized taboo distance and its importance when basins significantly vary in size, shape, and relative distance (Algorithm 1).

- Effect of the criticality measure (Equation 11) on reduction of the number of the critical taboo points.

- Possible advantages of the Mahalanobis distance metric, in comparison with the Euclidean distance metric.

3.1 First Descriptive Experiment

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 (see 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 x_{mean_i} and $\hat{d}_0\sigma_{\mathrm{mean}_i}$, respectively. The gray circles correspond to the archived points and their taboo regions ($\hat{d}_m\sigma_{\mathrm{mean}}$), respectively. At this point of each restart, $\mathbf{C}_i = 9.75^2\mathbf{I}$ and $\sigma_{\mathrm{mean}_i} = \sigma_{\mathrm{mean}}$; therefore, all of the taboo regions are spherical and the difference in size of the taboo regions originates merely from the difference in the normalized taboo distances ($\hat{d}_m$). 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 the centers of the subpopulations are at least $3\hat{d}_0\bar{u}\sigma_{\text{mean}}$ far from each other, in which $\bar{u}$ is the geometric average of the square root of the initial covariance matrix eigenvalues.

Figure 4c illustrates the solutions in **Archive** and their absolute taboo distances, immediately after initialization of the subpopulations in the first restart. There are 21 archived solutions, which implies 21 global minima were identified during the zeroth restart. Comparing taboo regions of the archived points with the basin sizes of the global minima (see Figure 4a) demonstrates that RS-CMSA has correctly increased the normalized taboo distances of minima with a wider basin.

At the beginning of the second restart (see Figure 4d), there are 27 solutions in **Archive**, which means six new minima were located during the first restart. The four largest taboo regions pertaining to the four 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 the distribution of the initialized subpopulations such that most of them are generated on the bottom-left side of the search space, where most of the unidentified minima lie. These factors result in a more efficient identification of new minima in the current restart.

There are 33 archived solutions at the beginning of the third restart (see Figure 4e), which implies six 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.

At the beginning of the fourth restart, the size of **Archive** is 34, which means only one new minimum was identified in the third restart. Nevertheless, the third restart has significantly resized the normalized taboo distances of the archived points. It seems many subpopulations of the third restart have converged to the largest basin, such that its taboo region covers most 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 each archived point. If it is larger than what it should be, more subpopulations will converge to that basin in the subsequent restarts, which increases the corresponding normalized taboo distance, and vice versa.

3.2 Second Descriptive Experiment

In the second experiment, the 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 $N_s^0 = 108$ 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 50 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, $0.5 \leq \tau_{\hat{d}} \leq 1$ is a logical choice.

Figure 5b illustrates the fraction of the taboo points that turned out to be critical, averaged over all subpopulations in each iteration. The horizontal axis is the normalized
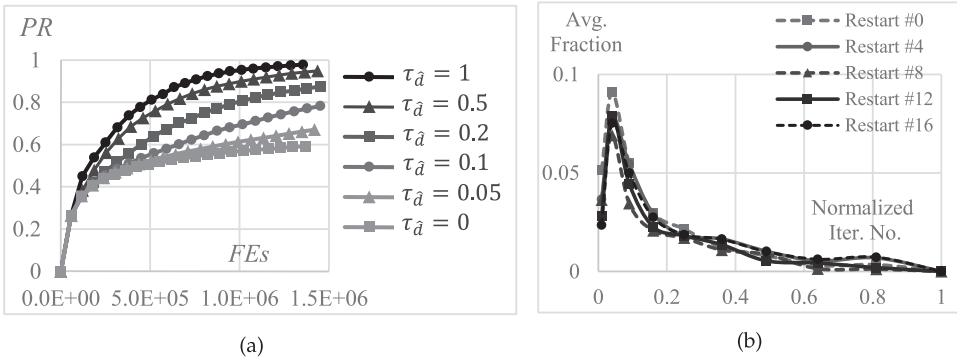
Figure 5: Results on 3D Vincent function: a) Peak ratio (PR) versus the adaptation rate of the normalized taboo distance ($\tau_{\hat{d}}$) and b) Average number of the critical taboo points versus the normalized iteration number, plotted for some selected restarts when $\tau_{\hat{d}} = 1/\sqrt{3}$ (default value).

iteration number, which is obtained by dividing the iteration number by the maximum number of the iterations in the corresponding restart. As it can be observed, the fraction of the critical taboo points is fairly small, which demonstrates that most taboo points are not critical, and therefore, they can be ignored. This fraction, as predicted earlier, reduces quite fast with the iteration number of the restart. For example, when 20% of the iterations of a restart are completed, only about 2% of the taboo points remain critical. The result is a significant reduction of the algorithmic computation time, since samples of each subpopulation are checked against only the critical taboo points.

3.3 Third Descriptive Experiment

In the third experiment, the benefits of the Mahalanobis distance metric is investigated and the capability of CMSA in learning the fitness landscape structure is explored. Five variants of RS-CMSA are tested and compared to reach this goal:

- RS-CMSA-M: The proposed method with the Mahalanobis distance metric.

- RS-CMSA-E: The proposed method with the Euclidean distance metric.

- RS-CMSA-M-I: Similar to RS-CMSA-M, but the covariance matrix is set to $\bar{u}^2\mathbf{I}$ after initialization in each restart, in which $\bar{u}$ is the geometric average of the search space side lengths.

- RS-CMSA-E-I: Similar to RS-CMSA-E, but the covariance matrix is set to $\bar{u}^2\mathbf{I}$ after initialization in each restart.

- RS-CMSA-M-I0: Similar to RS-CMSA-M-I, but τ_C is set to ∞ (no learning for the covariance matrix).

These variants are tested on the 3D conditioned Shubert function:

$$f(\boldsymbol{x}) = \prod_{j=1}^{3} \sum_{i=1}^{5} i \cos\left((i+1)\acute{x}_j + i\right);$$

$$\acute{x} = [10^{-s}x_1, x_2, 10^s x_3]; \quad -10^{1+s(2-j)} \le x_j \le 10^{1+s(2-j)},$$
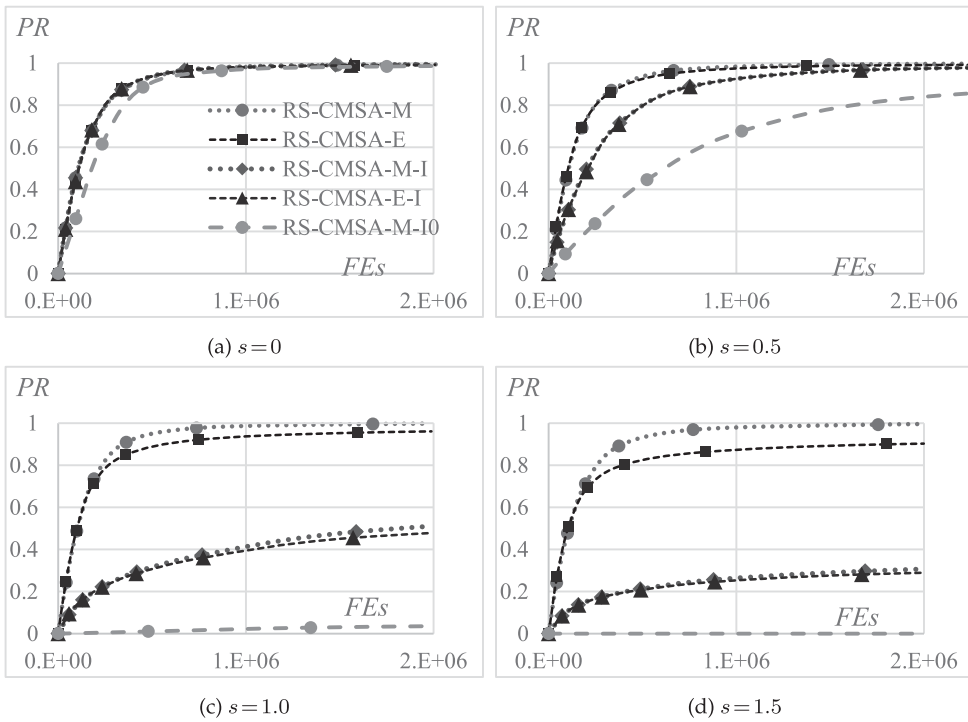
Figure 6: Peak ratio of RS-CMSA-M, RS-CMSA-E, RS-CMSA-M-I, RS-CMSA-M-E and RS-CMSA-M-I0 on the rescaled Shubert function with different scaling factors.

where exponent s determines the scaling factor. Since the condition number of the original Shubert function ($s = 0$) is about one and the search space is linearly scaled in our problem, the condition number of this function is about 10^{4s}. The search space undergoes the same scaling; therefore, the number of the global minima is independent of the scaling factor.

Each variant is tested on this problem for $s = 0$, 0.5, 1.0, and 1.5 with $N_s = 50$ and $MaxEvals = 6 \times 10^6$, for a maximum of seven restarts. Each problem is solved 50 times independently and the fraction of identified global minima is plotted versus the number of evaluations in Figure 6, in which the markers specify the end of restarts. This figure demonstrates that:

- increasing the scaling factor has no effect on the performance of RS-CMSA-M. This was predictable, since the initial covariance matrix is scaled accordingly and the Mahalanobis distance metric neutralizes the scaling effect when calculating the normalized taboo distance.

- RS-CMSA-E falls slightly behind RS-CMSA-M when the scaling factor increases. This inferiority exacerbates for greater scaling factors, but can be observed only for a great RR.

- no considerable difference between performance of RS-CMSA-M-I and RS-CMSA-E-I can be detected. It can be justified by the insignificance of performance difference between RS-CMSA-E and RS-CMSA-M for a low RR and the fact that the adaptation of the covariance matrix is gradual.

- starting with the isotropic distribution instead of the near-optimal one results in a significant performance decline; however, it is much more drastic for RS-CMSA-M-I0 than RS-CMSA-M-I. This demonstrates that RS-CMSA can gradually learn the structure of the fitness landscape, the contribution of which is significant.

It might be concluded that the normalized Euclidean distance metric should be employed; however, since there could be situations in which the Mahalanobis distance metric may appear superior, it is selected as the default choice. Computation of $\mathbf{C}^{-1}$, required by the Mahalanobis metric, is not expensive, since the core algorithm performs eigen decomposition of $\mathbf{C}$, from which $\mathbf{C}^{-1}$ can be easily computed.

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, which has also been employed for the CEC2015 and CEC2016 competitions on multimodal optimization. By selecting a standard and up-to-date test suite, we avoid any ad-hoc selection of test problems and biased evaluation of the methods.

4.1 Selection of the Methods

Several factors are considered to select the niching methods for comparing RS-CMSA, 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: NMMSO (Fieldsend, 2014), NEA2 (Preuss, 2012), LIPS (Qu et al., 2013), NSDE (Qu et al., 2012), PNPCDE (Biswas et al., 2014), and IPOP-CMA-ES (Auger and Hansen, 2005).

Each of the selected methods was demonstrated, in the corresponding publications, to surpass several other methods, including more traditional ones like crowding. For example, LIPS was compared in the referenced study with nine other niching methods, including crowding differential evolution, and was demonstrated to outperform all of them; therefore, we exclude the other nine algorithms. The other methods have been selected on a similar basis. Methods for which several problem-dependent parameters should be manually set have been excluded and the selected methods were tested either by using a default parameter setting or at most one parameter was tuned for each problem so that by performing an exhaustive search in the control parameter space, a reasonable and robust default value for the control parameter could be derived. Most of these methods were also tested and compared on more recent and challenging test problems including those proposed by Qu and Suganthan (2010) or employed in CEC2013 niching test suite (Li et al., 2013b). Although IPOP-CMA-ES is not tailored for multimodal optimization, it is employed as a benchmark to demonstrate the contribution of a niching strategy. We employed the default parameter setting for this method, except the increase factor, which may strongly affect the number of identified basins. *TolHhistFun* was set to 10^{-5}. NMMSO, NEA2, and IPOP-CMA-ES were benchmarked without any ad-hoc parameter tuning; however, for PNPCDE, LIPS, and NSDE, population size was tuned for each problem in the referenced studies.

Table 2: Test problems for benchmarking the niching methods, directly adapted from the CEC2013 special session on multimodal optimization (Li et al., 2013b). The niche radius parameter is used only for post-processing of the results.

PID	Function	Dimension	Niche radius	N_{gopt}	MaxEvals
1	Five-Uneven-Peak Trap	1	0.01	2	50,000
2	Equal Maxima	1	0.01	5	50,000
3	Uneven Decreasing Maxima	1	0.01	1	50,000
4	Himmelblau	2	0.01	4	50,000
5	Six-Hump Camel Back	2	0.5	2	50,000
6	Shubert	2	0.5	18	200,000
7	Vincent	2	0.2	36	200,000
8	Shubert	3	0.5	81	400,000
9	Vincent	3	0.2	216	400,000
10	Modified Rastrigin	2	0.01	12	200,000
11	Composition Function 1	2	0.01	6	200,000
12	Composition Function 2	2	0.01	8	200,000
13	Composition Function 3	2	0.01	6	200,000
14	Composition Function 3	3	0.01	6	400,000
15	Composition Function 4	3	0.01	8	400,000
16	Composition Function 3	5	0.01	6	400,000
17	Composition Function 4	5	0.01	8	400,000
18	Composition Function 3	10	0.01	6	400,000
19	Composition Function 4	10	0.01	8	400,000
20	Composition Function 4	20	0.01	8	400,000

In practice, the exact number of minima is not known a priori; however, one usually has a rough idea of the desired number of minima. For benchmarking, it is assumed a rough estimate for N_{gopt} is available which can be used for parameter setting. This estimate ($\widetilde{N}_{\text{gopt}}$) is used for parameter setting as follows:

- The optimum population size for LIPS, NSDE, PNPCDE, and NEA2 is a function of the number of global minima $\widetilde{N}_{\text{gopt}}$ and probably problem dimension; however, our preliminary results suggested dependence on $\widetilde{N}_{\text{gopt}}$ only.

- The optimum increase factor for the population size in IPOP-CMA-ES depends on $\widetilde{N}_{\text{gopt}}$.

- The only problem sensitive parameter of NMMSO is the swarm size, which was set to $S_s = 10D$ by Fieldsend (2014). $\widetilde{N}_{\text{gopt}}$ can hardly be used to set this parameter, since the swarm size is intuitively independent of the number of global minima; therefore, the optimal swarm size is assumed to be proportional to the problem dimension only.

- The optimum N_s^0 for RS-CMSA is presumably proportional to $\widetilde{N}_{\text{gopt}}$. We also investigate the case where $\widetilde{N}_{\text{gopt}}$ is not available and a predefined N_s^0 should be used for the problem at hand (RS-CMSA$_0$)

Following these assumptions, the parameter setting is formulated as follows:

$$\begin{cases} \lambda = 5 \times 2^\beta \times \widetilde{N}_{\text{gopt}} & \text{for NEA2, PNPCDE, LIPS, NSDE} \\ N_s^0 = 2^\beta \times \widetilde{N}_{\text{gopt}} & \text{for RS-CMSA} \\ N_s^0 = 10 \times 2^\beta & \text{for RS-CMSA}_0 \\ S_{\text{swarm}} = 5 \times 2^\beta \times D & \text{for NMMSO} \\ c_\lambda = 2^{\frac{0.5}{2^\beta \times \widetilde{N}_{\text{gopt}}}} & \text{for IPOP-CMA-ES.} \end{cases} \tag{13}$$

A parameter study is performed in the next section to find the best value of β for each method. Performance sensitivity to accuracy of the estimated number of global minima is also analyzed by varying $\widetilde{N}_{\text{gopt}}$ for a given problem.

4.2 Performance Measures

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

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

where subscript 0 refers to the assumption that $\widetilde{N}_{\text{gopt}} = N_{\text{gopt}}$. RR_{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. $\epsilon_f = 10^{-1}$ 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, peak ratios for $\epsilon_f = 10^{-3}$ and $\epsilon_f = 10^{-4}$ are 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 β and γ:

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

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

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

where γ 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 γ at the same time.

It is notable that $MPR(\beta, \gamma)$ can be derived from the $MPR_0(\beta)$ curve. Parameter setting for NMMSO and RS-CMSA$_0$ is independent of $\widetilde{N}_{\text{gopt}}$, and thus: $RMPR(\beta) = MPR_0(\beta)$. For other methods, the effect of increasing β and γ on parameter setting is identical. For example, increasing γ by one means $\widetilde{N}_{\text{gopt}}$ is doubled (Equation 15), which in turn doubles λ or N_s^0 (Equation 13). Increasing β by one also doubles λ or N_s^0 (Equation 13). Therefore, $MPR(\beta, \gamma)$ in Equation 16 can be replaced with $MPR_0(\beta + \gamma)$:

$$RMPR(\beta) = \int_{-\infty}^{+\infty} \rho(\gamma) MPR_0(\beta + \gamma) \mathrm{d}\gamma \approx \sum_{\gamma=-\tilde{M}}^{\gamma=+\tilde{M}} \rho(\gamma) MPR_0(\beta + \gamma). \tag{17}$$
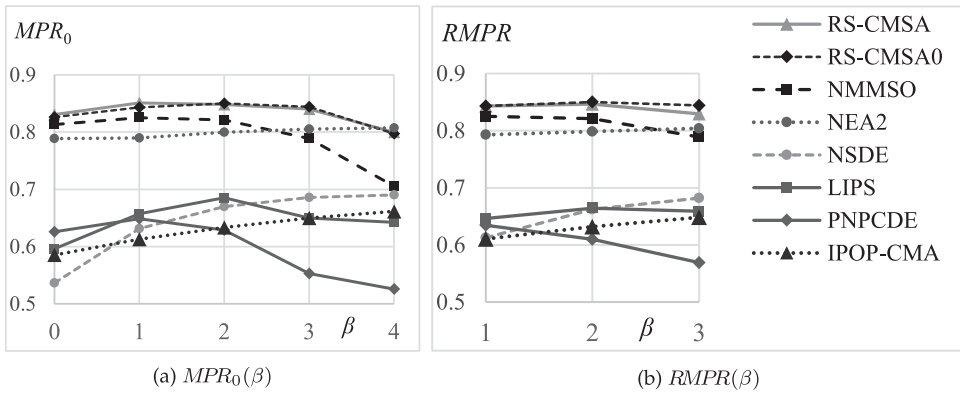
Figure 7: $MPR_0(\beta)$, $RMPR(\beta)$ of different methods.

To calculate $RMPR$ according to Equation 17, one should evaluate $MPR_0(\beta)$ for many values of β. To avoid this problem, Equation 17 is further simplified by assuming $\rho(\gamma) = 0$ if $|\gamma| > \tilde{M}$, which means $\tilde{N}_{\text{gopt}}$ can be at most a few times smaller/larger than N_{gopt}. For this study, we set $\tilde{M} = 1$:

$$RMPR(\beta) \approx \rho(-1)MPR_0(\beta - 1) + \rho(0)MPR_0(\beta) + \rho(1)MPR_0(\beta + 1). \quad (18)$$

Symmetry of $\rho(\gamma)$ necessitates $\rho(-1) = \rho(1)$ and besides, $\rho(-1) + \rho(0) + \rho(1) = 1$, therefore:

$$RMPR(\beta) \approx 0.5\,(1 - \rho(0))\,MPR_0(\beta - 1) + \rho(0)MPR_0(\beta) + 0.5\,(1 - \rho(0))\,MPR_0(\beta + 1), \quad (19)$$

which leaves only $\rho(0)$ as a decision parameter. For this study, $\rho(0) = \frac{1}{3}$ is used. For each method, the best value of β (β^*) which results in the greatest $RMPR(\beta)$ is sought. $RMPR(\beta^*)$ is employed as the primary performance measure to compare different methods.

4.3   Results and Discussion

Several aspects of these methods are compared by post-processing the results. Figure 7 illustrates $MPR_0(\beta)$ and $RMPR(\beta)$ for all the methods. The range of 99% confidence interval for these plots is fairly narrow (on the 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. Figure 8 plots PR (averaged over all the target function tolerances) versus PID for each method when $\beta = \beta^*$ and $\gamma = 0$. Figure 9 demonstrates PR versus PID for different values of β, assuming $\gamma = 0$.

- Figure 7a demonstrates that for a large range of β, RS-CMSA and RS-CMSA$_0$ outperform NMMSO and NEA2, and they all significantly surpass PNPCDE, LIPS, NSDE, and IPOP-CMA when MPR_0 is regarded.

- Figure 7b demonstrates that for $\beta = \beta^*$, RS-CMSA$_0$ ($RMPR = 0.850$) and RS-CMSA ($RMPR = 0.846$) relatively outperform NMMSO ($RMPR = 0.825$) and NEA2 ($RMPR = 0.804$), and they all significantly surpass NSDE, LIPS, PN-PCDE, and IPOP-CMA. It is remarkable that although IPOP-CMA is not specialized for multimodal optimization, it can fairly compete with LIPS and PNPCDE and to some extent, with NSDE.
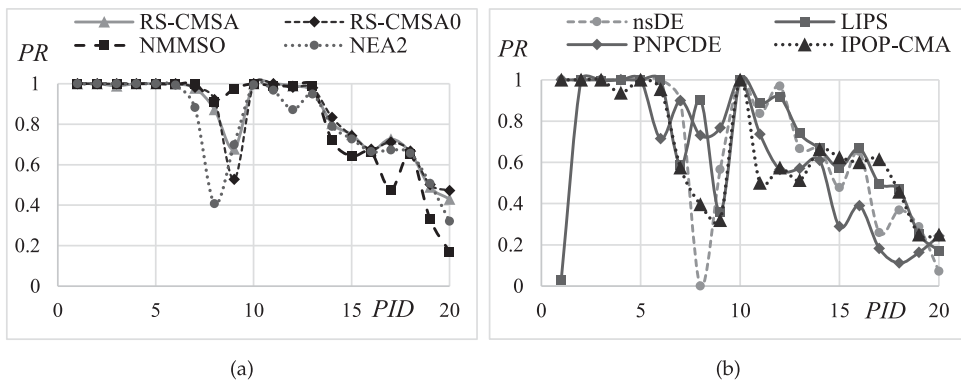
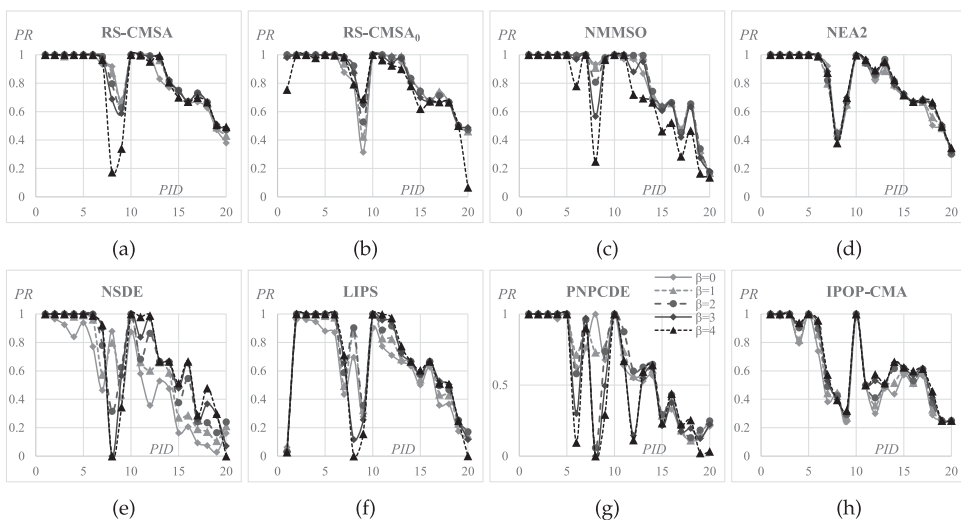Figure 8: *RR* for $\beta = \beta^*$ for each problem and method.

Figure 9: Effect of β on *RR* for each method on each problem.

- Superiority of RS-CMSA$_0$ over RS-CMSA is an unexpected and interesting observation. It demonstrates that the proposed method can still excel even if no estimate for the number of minima is available. The evaluation budget in the CEC2013 competition setup is not proportional to the number of global minima, which may justify suitability of a fixed N_s^0 for all the problems. A parameter setting scheme in which N_s^0 grows at a slower rate, for example, proportional to the square root of the estimated number of minima, can further improve *RMPR* of RS-CMSA, at least on this test suite.

- The first five problems are the easy ones for which most niching methods can reach $PR = 1$ within a small evaluation budget (see Figure 8); nevertheless, 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 of its operators seem necessary. IPOP-CMA cannot reach $PR = 1$

for the 4th problem, a classical and rather easy test problem for multimodal optimization.

- NMMSO outperforms RS-CMSA$_0$, RS-CMSA, and NEA2 on 3D Shubert and Vincent functions (PID = 8, 9). For the composite functions (*PID* = 11–20), a detectable domination of RS-CMSA$_0$, RS-CMSA, and NEA2 over the other methods is observed.

- It is unlikely that for identical number of iterations, increasing β results in a lower *PR*, and hence performance decline for a larger β is attributed to termination of the optimization process before convergence (see Figure 9). This means increasing/decreasing the evaluation budget can modify the optimum value of β for each method.

- Figure 9 demonstrates that NEA2 has the minimum sensitivity to the value of β when *PR* for each problem is regarded. Aside from $\beta = 4$, this sensitivity is small for RS-CMSA$_0$, RS-CMSA, and NMMSO as well, while it is quite significant for the other methods. The more critical point is that the optimum value of β is highly problem dependent for these methods, which challenges finding a good universal parameter setting. For example, $\beta^* = 3$ for NSDE; however, the highest *PR* for *PID* = 8 is achieved when $\beta = 0$, and increasing β significantly reduces *PR* for this problem. The reason is depletion of the evaluation budget before convergence. It seems the sequential nature of NEA2 and IPOP-CMA 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.

5   Summary and Conclusions

A new niching method called covariance matrix self-adaptation with repelling subpopulations (RS-CMSA) was developed in this study, which hybridizes and integrates several concepts and techniques from different existing methods, such as the taboo points in tabu search, normalized Mahalanobis distance, and the hill-valley function. RS-CMSA mainly consists of the covariance matrix self-adaptation evolution strategy (CMSA-ES), reinforced with elitism, 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 concurrently under exploration regions.

RS-CMSA was evaluated and compared with a number of the state-of-the-art niching methods on a recent test suite proposed for the CEC2013 competition on multimodal optimization. Robust mean peak ratio was proposed and utilized to compare the benchmarked methods, not only to quantify their success in finding different minima, but also to evaluate their performance sensitivity to accuracy of the given estimate for the number of global minima. According to the employed performance measures, RS-CMSA emerged as the most successful method, followed by NMMSO and NEA2. A considerable superiority of NEA2 and RS-CMSA over the other methods was observed on the more difficult composite functions.

RS-CMSA can adapt the niche radius parameter, referred to here as the normalized taboo distance, for each taboo point independently. This enables RS-CMSA to address the critical challenge of dissimilar and nonuniformly distributed minima, without de-

pendence on a user-specified niche radius parameter. Another hallmark of this method is that no problem-dependent parameter tuning is required, although providing a rough estimate for the desired number of minima can improve the performance. Simplicity of the main niching idea, its robustness, and superior results amply demonstrate the usefulness and practicality of RS-CMSA as a potent niching evolutionary algorithm.

Acknowledgments

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., and Kramer, O. (2015). Finite life span for improving the selection scheme in evolution strategies. *Soft Computing*, 1–13.

Ahrari, A., and Shariat-Panahi, M. (2015). An improved evolution strategy with adaptive population size. *Optimization*, 64(12): 2567–2586.

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 *The 2005 IEEE Congress on Evolutionary Computation, 2005*, vol. 2, pp. 1769–1776.

Bäck, T., Foussette, C., and Krause, P. (2013). *Contemporary evolution strategies*. New York: Springer Science and Business Media.

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

Basak, A., Das, S., and Tan, K. C. (2013). Multimodal optimization using a biobjective differential evolution algorithm enhanced with mean distance-based selection. *IEEE Transactions on Evolutionary Computation*, 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. (2007). Robust optimization—A comprehensive survey. *Computer Methods in Applied Mechanics and Engineering*, 196(33): 3190–3218.

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

Beyer, K., Goldstein, J., Ramakrishnan, R., and Shaft, U. (1999). When is nearest neighbor meaningful? In *Database Theory ICDT99*, pp. 217–235.

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.

Coello, C. A. C., and Lamont, G. B. (2004). *Applications of multi-objective evolutionary algorithms*, vol. 1. Toh Tuck Link, Singapore: World Scientific.

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.

de Castro, L. N., and Timmis, J. (2002). An artificial immune network for multimodal function optimization. In *CEC'02, Proceedings of the 2002 Congress on Evolutionary Computation, 2002*, vol. 1, pp. 699–704.

De Jong, K. A. (1975). *The analysis of the behaviour of a class of genetic adaptive systems*. Doctoral dissertation, University of Michigan, Ann Arbor.

Deb, K. (2001). *Multi-objective optimization using evolutionary algorithms*, vol. 16. New York: John Wiley & Sons.

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*, pp. 42–50.

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

Debski, R., Dreżewski, R., and Kisiel-Dorohinicki, M. (2008). Maintaining population diversity in evolution strategy for engineering problems. In *New Frontiers in Applied Artificial Intelligence*, pp. 379–387.

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

Fieldsend, J. E. (2014). Running up those hills: Multi-modal search with the niching migratory multi-swarm optimiser. In *2014 IEEE Congress on Evolutionary Computation (CEC)*, pp. 2593–2600.

Freschi, F., and Repetto, M. (2006). Comparison of artificial immune systems and genetic algorithms in electrical engineering optimization. *COMPEL—The International Journal for Computation and Mathematics in Electrical and Electronic Engineering*, 25(4): 792–811.

Glover, F. (1986). Future paths for integer programming and links to artificial intelligence. *Computers and Operations Research*, 13(5): 533–549.

Goldberg, D. E., and Richardson, J. (1987). Genetic algorithms with sharing for multimodal function optimization. In *Genetic Algorithms and Their Applications: Proceedings of the Second International Conference on Genetic Algorithms*, pp. 41–49.

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*, pp. 2389–2396.

Hansen, N., Arnold, D. V., and Auger, A. (2013). Evolution strategies. *Handbook of computational intelligence*. Berlin: Springer.

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 Kern, S. (2004). Evaluating the cma evolution strategy on multimodal test functions. In *Parallel Problem Solving from Nature-PPSN VIII*, pp. 282–291.

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. *IEEE Transactions on Evolutionary Computation*, 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. Retrieved from http://goanna.cs.rmit .edu.au/~xiaodong/cec13-niching/competition/Niching-final-presentation-CEC2013.pdf

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.

Li, X., Engelbrecht, A., and Epitropakis, M. (2015). Results of the 2015 IEEE CEC competition on niching methods for multimodal optimization. Retrieved from http://goanna.cs .rmit.edu.au/~xiaodong/cec15-niching/competition/

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.

Mahfoud, S. W. (1995). Niching methods for genetic algorithms. Doctoral dissertation, University of Illinois at Urbana-Champaign, Champaign, IL, USA. UMI Order No. GAX95-43663.

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

Pétrowski, A. (1996). A clearing procedure as a niching method for genetic algorithms. In *International Conference on Evolutionary Computation*, pp. 798–803.

Preuss, M. (2010). Niching the CMA-ES via nearest-better clustering. In *Proceedings of the 12th Annual Conference Companion on Genetic and Evolutionary Computation*, pp. 1711–1718.

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

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

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

Rechenberg, I. (2000). Case studies in evolutionary experimentation and computation. *Computer Methods in Applied Mechanics and Engineering*, 186(2): 125–140.

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.

Sareni, B., and Krahenbuhl, L. (1998). Fitness sharing and niching methods revisited. *IEEE Transactions on Evolutionary Computation*, 2(3): 97–106.

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.

Schwefel, H.-P., and Rudolph, G. (1995). *Contemporary evolution strategies*. Berlin: Springer.

Schwefel, H.-P. P. (1993). *Evolution and optimum seeking: The sixth generation*. New York: John Wiley & Sons.

Shir, O. M., and Bäck, T. (2007). Performance analysis of niching algorithms based on derandomized-es variants. In *Proceedings of the 9th Annual Conference on Genetic and Evolutionary Computation*, pp. 705–712.

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.

Siarry, P., and Berthiau, G. (1997). Fitting of tabu search to optimize functions of continuous variables. *International Journal for Numerical Methods in Engineering*, 40(13): 2449–2457.

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*, pp. 1305–1312.

Stoean, C., Preuss, M., Stoean, R., and Dumitrescu, D. (2010). Multimodal optimization by means of a topological species conservation algorithm. *IEEE Transactions on Evolutionary Computation*, 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*, pp. 1420–1427.

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

van der Goes, V., Shir, O. M., and Bäck, T. (2008). Niche radius adaptation with asymmetric sharing. In *Parallel Problem Solving from Nature–PPSN X*, pp. 195–204.

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*, pp. 821–828.

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*, pp. 318–329.

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