More Efficient Real-Valued Gray-Box Optimization through Incremental Distribution Estimation in RV-GOMEA

Renzo J. Scholman
renzo.scholman@cwi.nl
Centrum Wiskunde & Informatica
Amsterdam, The Netherlands
Delft University of Technology
Delft, The Netherlands

Tanja Alderliesten
t.alderliesten@lumc.nl
Leiden University Medical Center
Leiden, The Netherlands

Peter A.N. Bosman
peter.bosman@cwi.nl
Centrum Wiskunde & Informatica
Amsterdam, The Netherlands
Delft University of Technology
Delft, The Netherlands

Abstract

The Gene-pool Optimal Mixing EA (GOMEA) family of EAs offers a specific means to exploit problem-specific knowledge through linkage learning, i.e., inter-variable dependency detection, expressed using subsets of variables, that should undergo joint variation. Such knowledge can be exploited if faster fitness evaluations are possible when only a few variables are changed in a solution, enabling large speed-ups. The recent-most version of Real-Valued GOMEA (RV-GOMEA) can learn a conditional linkage model during optimization using fitness-based linkage learning, enabling fine-grained dependency exploitation in learning and sampling a Gaussian distribution. However, while the most efficient Gaussian-based EAs, like NES and CMA-ES, employ incremental learning of the Gaussian distribution rather than performing full re-estimation every generation, the recent-most RV-GOMEA version does not employ such incremental learning. In this paper, we therefore study whether incremental distribution estimation can lead to efficiency enhancements of RV-GOMEA. We consider various benchmark problems with varying degrees of overlapping dependencies. We find that, compared to RV-GOMEA and VKD-CMA-ES, the required number of evaluations to reach high-quality solutions can be reduced by a factor of up to 1.5 if population sizes are tuned problem-specifically, while a reduction by a factor of 2-3 can be achieved with generic population-sizing guidelines.

CCS Concepts

• **Mathematics of computing → Evolutionary algorithms**.

Keywords

Evolutionary Algorithms, Estimation of Distribution Algorithms, Incremental Learning, Linkage Modeling

1 Introduction

Taking a Gray-Box Optimization (GBO) approach to exploit problem-specific knowledge, can substantially enhance the performance of EAs. Examples include optimization of a billion-parameter problem using problem-specific recombination operators [10], adapted crossover operators for satellite scheduling [18], and tailored mutation operators for wireless communication networks [15].

Sometimes, a problem allows for efficient partial evaluations, meaning that objective values can be efficiently recalculated when only a few variables in a solution are modified, greatly reducing total optimization time, see, e.g., [4, 7, 13].

EAs in the family of Gene-pool Optimal Mixing EAs (GOMEA) are especially suited for exploiting partial evaluations due to the use of optimal mixing [16]. In GOMEA, subsets of interdependent variables are leveraged, called linkage sets, in the GOM variation operator. By only changing the variables in a linkage set, and keeping the change only if the solution did not get worse, partial evaluations can be effectively exploited, which leads to performance benefits if the linkage sets are well-aligned with problem structure.

State-of-the-art performance can be achieved on real-valued problems using the Real-Valued GOMEA (RV-GOMEA), especially in gray-box scenarios, in which Black-Box Optimization (BBO) approaches can be vastly outperformed [7].

Recently, a new version of RV-GOMEA was introduced that learns conditional linkage sets using fitness-based linkage learning techniques during optimization [3]. The use of such linkage models in RV-GOMEA has been essential to also have an performance advantage over black-box methods on optimization problems with overlapping dependencies. The Gaussian distributions pertaining to each linkage set are however still estimated in each generation using only the best-performing solutions in the current population. Doing so discards information from the previous generation and, therefore, typically results in requiring larger population sizes to make reliable estimations of covariances.

The Gaussian sampling model in RV-GOMEA is based on AMaLGaM, a real-valued EDA [7]. It is known that the Gaussian in AMaLGaM can be learned incrementally across generations [5]. This can typically reduce the required population size and often, the required number of function evaluations.

Similar incremental mechanisms are included in the state-of-the-art real-valued EAs CMA-ES [11] and NES [20]. A version of RV-GOMEA, called RV-GOMEA$_C$, exists that uses the CMA-ES sampling model, but for decomposable problems with moderately sized subfunctions, performance differences with RV-GOMEA are small, with RV-GOMEA sometimes outperforming RV-GOMEA$_C$ [7].

Still, the parameter settings of the CMA-ES sampling model were not reconsidered in the integration with RV-GOMEA. Proper calibration of incremental learning may however still lead to achieving even better results, since the incremental AMaLGaM was established entirely through empirical optimization [5]. In this paper, we therefore reconsider the incremental learning model for AMaLGaM, but then within the recent-most version of RV-GOMEA, and revisit optimizing the parameters of this model. We refer to the resulting new EA as the incremental RV-GOMEA (iRV-GOMEA).

In the remainder of this paper, we first explain how incremental learning is incorporated into RV-GOMEA in Section 3. The same section describes how we tune iRV-GOMEA on various problems and determine population-sizing guidelines. Then, in Section 4, we describe the setup of our experiments, testing the resulting iRV-GOMEA in various ways and comparing its performance with that of RV-GOMEA, RV-GOMEA$_C$, and VkD-CMA-ES [1, 2]. The results of the experiments are presented in Section 5. We present a discussion in Section 6 and our final conclusions in Section 7.

2 Background

2.1 Incremental Distribution Estimation

Incremental learning typically takes the following form:

$$\boldsymbol{\theta}(t) = (1 - \eta)\boldsymbol{\theta}(t - 1) + \eta\hat{\boldsymbol{\theta}}(t). \tag{1}$$

In our setting, $\hat{\boldsymbol{\theta}}(t)$ are the probability distribution parameters as estimated from the population in generation t, and $\boldsymbol{\theta}(t)$ are the incrementally learned parameters with a learning rate $\eta \in [0, 1]$.

A general function class has been derived in literature to estimate an effective learning rate for distribution parameters in an EDA in which truncation selection is used and the distribution parameters are estimated from the selected solutions [5]. The function class is defined using the selection size $|\mathcal{S}|$, problem dimensionality ℓ, and three function-class parameters $\boldsymbol{\alpha} = (\alpha_0, \alpha_1, \alpha_2)$ as follows:

$$\eta = f_{\boldsymbol{\alpha}}^{\eta}(|\mathcal{S}|, \ell) = 1 - \exp\left(\frac{\alpha_0|\mathcal{S}|^{\alpha_1}}{\ell^{\alpha_2}}\right). \tag{2}$$

This function class can be instantiated by fitting it to a dataset of N samples $\{(\eta^i, (|\mathcal{S}|^i, \ell^i))\}, i \in \{0, 1, \ldots, N-1\}$, i.e., by performing non-linear regression with a typical sum-of-squares loss function:

$$\min_{\boldsymbol{\alpha}} \sum_{i=0}^{N-1} \left(\eta^i - f_{\boldsymbol{\alpha}}^{\eta}(|\mathcal{S}|^i, \ell^i)\right)^2. \tag{3}$$

Performing this non-linear optimization task can be done using, e.g., an EA. To regress the α_i parameters, a dataset is needed consisting of the (near-)best values for the learning rate associated with various selection sizes and problem dimensionalities. This can be obtained through experimentation, i.e., by trying various values for, or optimizing, the learning rate for different population sizes and different problem dimensionalities. Such experimentation can be done for a *set* of problems, resulting in a more general instantiation of the function class that is well-suited for a class of problems.

2.2 Gray-box Optimization

Consider a problem where the objective function $f(\mathbf{x}) : \mathbb{R}^{\ell} \rightarrow \mathbb{R}$ is to be minimized and a solution is denoted by $\mathbf{x} = (x_0, \ldots, x_{\ell-1})$.

An objective function $f(\mathbf{x})$ permitting partial evaluations, is composed of q non-decomposable sub-functions $f_0, \ldots, f_{q-1}$. Now, let $\mathcal{I} = \{\mathcal{I}_0, \ldots, \mathcal{I}_{q-1}\}, \mathcal{I}_i \subseteq \{0, 1, \ldots, \ell-1\}$ denote the subsets of indices that are involved with the q sub-functions. The variables pertaining to the i–th subset and subfunction can be denoted as $\mathbf{x}_{\mathcal{I}_i}$. After a modification is made (i.e., during variation) to one or more variables x_j, all sub-functions $f_i(\mathbf{x}_{\mathcal{I}_i})$ are evaluated for which $j \in \mathcal{I}_i$ and subsequently, the fitness of the solution can be recomputed from the values of the subfunctions f_i. For more details on this formulation and its use to perform partial evaluations, see, e.g., [7].

In this work, we consider the cost of the partial evaluation of f_i as $|\mathcal{I}_i| / |\mathcal{I}|$. The cost of changing one or more variables x_j is then computed as the accumulated cost of each affected sub-function f_i.

Two variables x_u and x_v are considered dependent if there exists a sub-function f_i for which $\{u, v\} \subseteq \mathcal{I}_i$ holds. A Variable Interaction Graph (VIG) is an undirected graph $G = (V, E)$ where each vertex $v \in V$ corresponds to a variable x_v and for each pair of dependent variables there exists an edge $(u, v) \in E$, see, e.g., [17, 19].

2.3 RV-GOMEA

RV-GOMEA [7] is a variant of GOMEA for solutions with real-valued variables. In RV-GOMEA, a sampling model is used to generate new solutions that comes from the EDA called AMaLGaM [6]. From the population $\mathcal{P}$, AMaLGaM selects the set $\mathcal{S}$ of $\lfloor \tau|\mathcal{P}| \rfloor$, $\tau \in [1/|\mathcal{P}|, 1]$ best-performing individuals and estimates a Gaussian (normal) distribution $P^{\mathcal{N}}(\cdot)$ from $\mathcal{S}$. Subsequently, variation is performed by sampling new solutions from $P^{\mathcal{N}}(\cdot)$.

RV-GOMEA uses the Family Of Subsets (FOS) linkage model $\mathcal{F} = \{\mathbf{F}^0, \mathbf{F}^1, \ldots, \mathbf{F}^{|\mathcal{F}|-1}\}$, which comprises a set of FOS elements $\mathbf{F}^i$, also called linkage sets. Each linkage set is a set of variable indices $\mathbf{F}^i \subseteq \{0, 1, \ldots, \ell-1\}$ that identify variables that are considered to be interdependent and must be processed *jointly* during variation. For each linkage set $\mathbf{F}^i$, RV-GOMEA estimates a multivariate Gaussian distribution $P^{\mathcal{N},i}(\cdot)$ using the same procedure as in AMaLGaM.

During variation with the GOM operator, all linkage sets $\mathbf{F}^i$ are considered in a random order and for each solution in the population, new values are jointly sampled from $P^{\mathcal{N},i}(\cdot)$ for each of the variables in the linkage set. Changes are in principle only kept if the solution has not become worse. However, in the original, non-conditional, formulation of RV-GOMEA, with a probability of 5%, changes were always accepted [7].

As in AMaLGaM [6], the techniques known as Anticipated Mean Shift (AMS) and Adaptive Variance Scaling (AVS) are used in RV-GOMEA. Through AMS, a part of the population is moved in the direction in which the estimated distribution mean shifted in the recent-most subsequent generations, enhancing optimization on slope-like regions in the search space. AVS is used to counteract the variance-diminishing effect of selection whenever improvements are found far away from the mean. This is done through the use of distribution multipliers $c_{\mathbf{F}^i}^{mult}$ for each linkage set $\mathbf{F}^i$. If the average improvement along any of the principal axes of $P^{\mathcal{N},i}(\cdot)$ is larger than 1 standard deviation, $c_{\mathbf{F}^i}^{mult}$ is increased by $1/0.9$. If no improvements are found, the multiplier is decreased by 0.9. The multipliers are prevented to become smaller than 1 for at least a predefined number of generations, the so-called maximum No Improvement Stretch NIS^{MAX}. For more details, see, e.g., [7].

2.4 Conditional Linkage Modeling

An approach to detecting pairwise dependencies between variables, including a notion of dependency strength, that has proven to be successful, is fitness-based dependency tests [14]. The Dependency Strength Matrix (DSM) [12] is a square matrix $\mathbf{D}_{ij} = d(i, j)$, where $i, j \in \{0, 1, \ldots, \ell - 1\}$ and $d(i, j) \in [0, 1]$ indicates dependency strength with 0 indicating no dependence. Construction of a complete DSM requires $\frac{1}{2}(\ell^2 - \ell)$ fitness dependency tests, imposing a quadratic number of function evaluations as overhead. However, when partial evaluations can be performed, the overhead of these tests may be reduced to a linear order, depending on the size of the sub-functions. The DSM can be built incrementally during optimization, which often leads to performing less dependency tests [3].

A VIG can be constructed based on the DSM by adding an edge (u, v) whenever $d(u, v) > d_{min}$ holds. In the recent-most work on RV-GOMEA, $d_{min} = 10^{-6}$ was used and found to lead to good results on a variety of problems [3].

In each generation, after an update of the DSM, the VIG must be updated. The VIG can then be used to construct the linkage model that is used in RV-GOMEA. In the recent-most version of RV-GOMEA, a technique called clique seeding is used [3]. This technique performs breadth-first searches of cliques in the VIG, starting from each node in the VIG. For each search, the found maximal clique is stored in a set of candidate cliques, any duplicates are discarded. Each clique is conditioned on any variables that are outside the clique but have an edge in the VIG to a variable inside the clique. For each of the candidate cliques a conditional linkage set $\mathbf{F}^i$ is created, which now describes a *conditional* multivariate distribution. I.e., a conditional linkage set consists of *two* sets of variable indices $\mathcal{I}_i^0$ and $\mathcal{I}_i^1$, describing the distribution $P^{\mathcal{N}}(X_{\mathcal{I}_i^0}|X_{\mathcal{I}_i^1})$.

While conditional linkage sets enable explicitly modeling overlapping dependencies, linkages may still be broken when performing GOM, because acceptance selection is performed immediately after partial variation. To repair potentially broken linkages, an additional special GOM step is performed each generation in which all variables are sampled without intermediate GOM acceptance, i.e., following the conditionally factorized distribution defined by all conditional linkage sets together. For details, see [3].

We store a covariance matrix for each linkage set. If a linkage set is a conditional one, a covariance matrix is stored for all involved variables (i.e., both those in $\mathcal{I}_i^0$ and in $\mathcal{I}_i^1$) as the conditional distribution is defined using all covariance values.

3 iRV-GOMEA
3.1 Incremental Distribution Estimation

We consider two scenarios to incorporate incremental learning into RV-GOMEA: a static scenario where the linkage model is obtained using clique seeding based on the true VIG, and an online scenario where the VIG is obtained through incremental DSM updates using fitness-based dependency tests. Incremental learning is considered for each linkage set in the current generation, in turn, i.e., $\mathbf{F}_g^i$.

3.1.1 Covariance Matrices. In the static scenario, the covariance matrices $\Sigma^{\mathbf{F}^i}$ of distributions $P^{\mathcal{N},i}(\cdot)$ are initialized with maximum-likelihood estimates of variances from the selection in the first generation on the diagonal and covariance values of 0 off-diagonal. In subsequent generations, Eq. 1 is applied to each $\Sigma^{\mathbf{F}^i}$.
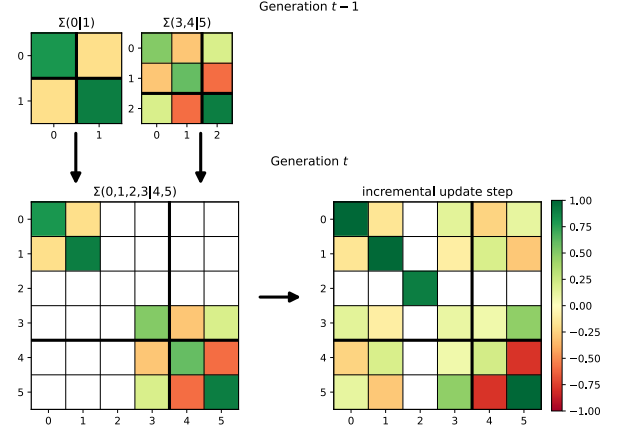

Figure 1: Illustration of copying covariance information between generations and performing a subsequent incremental update step. Colors visualize correlation coefficients.

In the online scenario, the linkage model changes in each generation. Consequently, the size and involved variables of the covariance matrices $\Sigma^{\mathbf{F}^i}$ associated with the linkage sets $\mathbf{F}^i$ also change in each generation. To incrementally update covariances, a method is needed to transfer as much covariance information as possible.

When there exists a linkage set in the previous generation that considers the same variables as used in $\mathbf{F}_g^i$, i.e., $\bigcup \mathbf{F}_{g-1}^j = \bigcup \mathbf{F}_g^i$, then the update step is the same as in the static scenario. I.e., Eq. 1 can be used where $\Sigma^{\mathbf{F}_g^i}$ is $\boldsymbol{\theta}(t)$, $\Sigma^{\mathbf{F}_{g-1}^j}$ is used for $\boldsymbol{\theta}(t-1)$, and $\hat{\boldsymbol{\theta}}(t)$ are the maximum-likelihood estimates from the selection set $\mathcal{S}_g$ in generation g for all (co)variances pertaining to the variables in $\mathbf{F}_g^i$.

Otherwise, we proceed as follows:

(1) Initialize a matrix Σ of the same dimensions as $\Sigma^{\mathbf{F}_g^i}$ to a matrix of zeros and let $\mathbf{V} = \bigcup \mathbf{F}_g^i$.
(2) Find all possible linkage sets in the previous generation from which information can be reused for the distribution parameters pertaining to $\mathbf{F}_g^i$, i.e., $\mathbb{F} = \left\{ \mathbf{F}_{g-1}^j \in \mathbf{F}_{g-1} \left| \bigcup \mathbf{F}_{g-1}^j \subset \mathbf{V} \right. \right\}$.
(3) Find the largest set in $\mathbb{F}$ (with ties being broken randomly), i.e.,: $\mathbf{F}^{max} = \arg \max_{\mathbf{F}_{g-1}^j \in \mathbb{F}} \left\{ \left| \bigcup \mathbf{F}_{g-1}^j \right| \right\}$.
(4) Copy variances and covariances from $\Sigma^{\mathbf{F}^{max}}$ to Σ.
(5) Update $\mathbf{V}$ by setting it to $\mathbf{V} - \bigcup \mathbf{F}^{max}$ and $\mathbb{F}$ to $\mathbb{F} - \mathbf{F}^{max}$.
(6) If either either $|\mathbb{F}|$ or $|\mathbf{V}|$ is zero, then exit, otherwise, repeat the process, starting at finding $\mathbb{F}$ (step 2).

After this, $\Sigma^{\mathbf{F}_g^i}$ can be computed with Eq. 1 by using Σ for $\boldsymbol{\theta}(t-1)$ and the variances and covariances estimated using maximum-likelihood estimates from the selection set $\mathcal{S}_g$ in generation g for $\hat{\boldsymbol{\theta}}(t)$. However, for any indices v that remain in $\mathbf{V}$, we estimate only the maximum-likelihood variances. Any covariance associated with remaining variables in $\mathbf{V}$ are kept at zero. This is similar to the initialization of the covariance matrices in static linkage models. See Figure 1 for a visual representation of the described approach.

3.1.2 Anticipated Mean Shift. In RV-GOMEA, the anticipated mean shift (AMS), denoted $\boldsymbol{\mu}_{AMS}$ is estimated once for the full set of variables. For each linkage set $\mathbf{F}^i$, the $\boldsymbol{\mu}_{AMS}^{\mathbf{F}^i}$ that is used, is equal to the subset of $\boldsymbol{\mu}_{AMS}$ pertaining to the variables in $\mathbf{F}^i$. Of note, in case

of conditional linkage sets, this pertains only to the conditioned variables, not the ones they are conditioned on. As was the case for the incremental AMaLGaM [5], we aim to apply incremental learning also to the AMS. Ultimately, the applied AMS depends on the size of the linkage set. However, Equation 2 is dependent on the number of variables. By incrementally estimating the AMS only for the full set of variables at once, the learning rate may not properly reflect the fact that the AMS is in fact applied in GOM steps that pertain to subsets of variables. We therefore change this such that in iRV-GOMEA the AMS is also learned independently for each linkage set $\mathbf{F}^i$ in the form of $\boldsymbol{\mu}_{AMS}^{\mathbf{F}^i}(t)$.

3.2 Distribution Multipliers

Preliminary experiments with iRV-GOMEA indicated that the direct implementation of incremental learning in RV-GOMEA led to unexpected poor results for small population sizes. This is counter-intuitive because the use of incremental learning has the potential to reduce the minimally required population size. The number of required generations will then typically go up. While this balance ultimately is typically in favor of smaller population sizes, resulting in fewer function evaluations required, the number of times the distribution parameters are updated, also goes up. In RV-GOMEA, different from CMA-ES or AMaLGaM, updates are performed for each linkage set separately and there is a distribution multiplier per linkage set. Without linkage sets, or for very large linkage sets, multiple variables are changed at the same time and while some changes separately would lead to worse fitness, the joint change could still lead to an improvement. The smaller the linkage sets, the stronger individual variable changes must be correlated with making improvements. The use of linkage sets in RV-GOMEA may therefore lead to less improvements per generation to be found and the distribution multipliers converting back to the value of 1 quickly after they were increased due to an improvement. Moreover, with incremental learning, smaller population sizes are made possible, necessitating however more generations before the covariances become aligned with the (local) landscape of the problem being solved and exacerbating the problem of finding improvements per variation of single FOS elements. The increased number of generations also entails more updates to the distribution multipliers being done in iRV-GOMEA (with a smaller population size) than in RV-GOMEA. Altogether, the consequence is that the distribution multipliers are reverted back to 1 too quickly, leading to limited numbers of improvements and premature convergence taking place if the population size becomes too small, which contradicts what we are trying to achieve. For this reason, we changed the distribution multiplier decrease factor to be asymmetrical to the distribution multiplier increase factor, decreasing the value more slowly than it is increased. Specifically, we set the decrease rate to half the increase rate, i.e.,: 0.95 (and $1/0.9$ for the increase rate).

3.3 Regressing the Learning Rates

To obtain learning rates for each linkage set $\mathbf{F}^i$, and for each of the categories of distribution parameters (covariance matrices and AMS), function $f_{\boldsymbol{\alpha}}^{\eta}$ (Eq. 2) can be used where ℓ is replaced by the number of variables in $\mathbf{F}^i$, which we will denote by κ in the remainder of this paper. To determine proper values for the $\boldsymbol{\alpha}$ parameters

in $f_{\boldsymbol{\alpha}}^{\eta}$ for both the covariance matrix learning rate and the AMS learning rate, a set of samples $\{((\eta^{\Sigma}, \eta^{AMS}), (|\mathcal{S}|, \kappa))\}$, is required to perform regression using Equation 3. These samples must span a large variety of different selection sizes, linkage set sizes, and benchmark problems to obtain a generically usable $f_{\boldsymbol{\alpha}}^{\eta}$.

To create the samples, RV-GOMEA is used to optimize the learning rates η^{Σ} and η^{AMS} for each problem in Table 1. This is done by considering them as variables in a two-dimensional problem that involves running iRV-GOMEA and minimizing the required number of function evaluations to achieve the Value-To-Reach (VTR).

Learning rates are optimized for non-conditional linkage set sizes $\kappa \in \{5, 10, 15, ..., 100\}$ and population sizes $|\mathcal{P}| \in \{20, 40, 60, ..., 400\}$, with the selection percentile the same as in AMaLGaM and RV-GOMEA, $\tau = 0.35$ ($|\mathcal{S}| = \lfloor \tau|\mathcal{P}| \rfloor$). Solutions in iRV-GOMEA are initialized in $[-10, 5]$, a range located asymmetrically around the optimum (at 0). For each optimization problem and for each value of κ and $|\mathcal{P}|$, iRV-GOMEA is given a budget of twice the number of evaluations that are required to reach the VTR using the non-incremental RV-GOMEA with a full covariance matrix, for which the guideline population size of $3\kappa^{1.5} + 17$ is used [5]. If iRV-GOMEA exceeds that budget, the sample $((\eta^{\Sigma}, \eta^{AMS}), (|\mathcal{S}|, \kappa))$ for that problem is discarded.

(i)RV-GOMEA reverts to univariate sampling of the distribution when the (incrementally) estimated covariance matrix is singular. This can happen when the selection size $|\mathcal{S}|$ is simply too small for a linkage set size κ. Since the goal is to determine the learning rates for the correct sampling of full covariance matrices, one of the included optimization problems is the ill-conditioned rotated ellipsoid problem. Reverting to univariate sampling will cause iRV-GOMEA to no longer be able to solve this problem. Moreover, when iRV-GOMEA fails to find the optimum for the rotated ellipsoid problem of a certain linkage set size κ and with a population size $|\mathcal{P}|$, all samples for other problems with linkage set size κ and population size $|\mathcal{P}|$ are removed from the set of samples.

The resulting set of samples for both η^{Σ} and η^{AMS} are used as input for Equation 3. The problem defined by this equation is then optimized using RV-GOMEA to instantiate $f_{\boldsymbol{\alpha}}^{\eta}$ for both the AMS and covariance matrix separately.

Table 1: Optimization problems used in the experiments.

Problem	Function Definition	VTR
Sphere	$\sum_{i=0}^{\ell-1} x_i^2$	10^{-10}
Rotated Ellipsoid	$\sum_{i=0}^{\ell-1} 10^{6\frac{i}{\ell-1}} (x_i^r)^2$, where $\boldsymbol{x}^r = \boldsymbol{R}\boldsymbol{x}$ for rotation matrix $\boldsymbol{R}$	10^{-10}
Cigar	$x_0^2 + \sum_{i=1}^{\ell-1} 10^6 x_i^2$	10^{-10}
Tablet	$10^6 x_0^2 + \sum_{i=1}^{\ell-1} x_i^2$	10^{-10}
Cigar Tablet	$x_0^2 + \sum_{i=1}^{\ell-2} 10^4 x_i^2 + 10^8 x_{\ell-1}^2$	10^{-10}
Two Axes	$\sum_{i=0}^{\lfloor \ell/2 \rfloor-1} 10^6 x_i^2 + \sum_{i=\lfloor \ell/2 \rfloor-1}^{\ell-1} x_i^2$	10^{-10}
Different Powers	$\sum_{i=0}^{\ell-1} |x_i|^{2+10\frac{i}{\ell-1}}$	10^{-10}
Rosenbrock	$\sum_{i=0}^{\ell-2} \left(100(x_i^2 - x_{i+1})^2 + (x_i - 1)^2\right)$	10^{-10}
Parabolic Ridge	$-x_1 + 100 \sum_{i=0}^{\ell-1} x_i^2$	-10^{10}
Sharp Ridge	$-x_1 + 100 \sqrt{\sum_{i=0}^{\ell-1} x_i^2}$	-10^{10}

The regressed $\boldsymbol{\alpha}$ parameters, after performing the above for all problems in Table 1 and merging all samples into one dataset, can be found in Table 2. A visualization of the learning rates η^{Σ} and η^{AMS} can be found in Figure 2. Using the regressed learning rate formulas

in the incremental distribution estimation procedure as outlined above within RV-GOMEA, the final iRV-GOMEA is obtained. Next, we turn to experimenting with the now fully defined iRV-GOMEA.
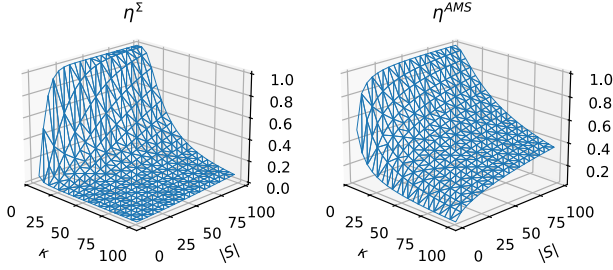

Figure 2: Regressed f_α^η functions over all problems.

Table 2: Regressed α values over all problems.

	α_0	α_1	α_2
η^Σ	-1.01	1.32	1.94
η^{AMS}	-2.95	0.47	0.87

3.4 Determining Population Sizing Guidelines for a Single Linkage Set in iRV-GOMEA

One of the main motivations for incremental distribution estimation is the potential for requiring smaller population sizes to solve problems. We determine the required population size for each problem in Table 1 for linkage set sizes up to $\kappa = 200$. Note that in iRV-GOMEA now a single linkage set is used, and as such, a full covariance matrix is estimated over all involved problem variables, regardless of whether the problem is actually decomposable. Beginning at the default population size of RV-GOMEA, the population size is halved as long as the required number of evaluations (averaged over 100 runs) to achieve the VTR, reduces. Once the required number of evaluations stops decreasing, binary search is performed between the last two tried population sizes, until the best performing population size is found. The required number of evaluations is divided by the success probability to account for failed runs with very small population sizes.

The best performing population size for each problem and linkage set size κ can be seen in Figure 3. We want to use these results to derive a population size guideline. Sharp Ridge results were however ignored when determining the guideline, as deviating behavior was observed. Specifically, the criterion to only accept improvements in GOM seems to prevent finding improvements over multiple generations. The fraction of the population that improves if the acceptance criterion in GOM is ignored with increasing probability (to the point where all samples are always accepted, which corresponds to classical EDA configurations as used in AMaLGaM and CMA-ES), can be seen in Figure 3 on Sharp Ridge with $\kappa = 100$. Clearly, when the acceptance selection in GOM is ignored more frequently, the optimum is found in fewer generations. While this can be considered a shortcoming, it is important to note that this holds only when using the full covariance matrix on a (sub)function of κ variables with κ fairly large. Arguably, oftentimes, problems have subfunctions pertaining to smaller sets of fully dependent variables that should be considered jointly. Moreover, the observed limitation holds for a set of variables in the Sharp Ridge function, the variables of which are actually not interdependent. As such,

these variables in practice should be modeled as independent and be in separate linkage sets of size $\kappa = 1$. For similar reasons, we also disregarded the results on Two Axes, for which similar behavior was observed. Finally, for various problems, e.g., Cigar, a drop in population size can clearly be identified in Figure 3 as κ goes up. This is because there is a large basin of approximately equally-well performing population sizes (for an example, see supplementary material A.2), but we have automated finding the best performing population size. Ultimately, we therefore opted to fit a function to the upper envelope of the problem that leads to the largest population size without deviating behavior, which was Rotated Ellipsoid. Based on that function, we determined guideline for the required population size for linkage set sizes κ in iRV-GOMEA to be:

$$10 + 3\kappa. \tag{4}$$

4 Experiments

In this section, we describe the various experiments that we performed with the finally obtained iRV-GOMEA. First, we consider the use of a single, static, non-conditional linkage set in iRV-GOMEA and compare performance with the non-incremental RV-GOMEA and RV-GOMEA$_C$, the CMA-ES variant of RV-GOMEA. This can be considered a classical black-box evaluation setting. We subsequently evaluate the performance of iRV-GOMEA when the number of static non-conditional linkage sets is scaled up and partial evaluations can be performed, i.e., a gray-box setting. We compare again with RV-GOMEA and RV-GOMEA$_C$, but also with VkD-CMA-ES [1, 2], a variant of CMA-ES that learns to decompose the Gaussian model, however in a manner that cannot leverage partial evaluations. Lastly, we consider again the gray-box setting, and evaluate iRV-GOMEA with *conditional* linkage sets, defined both statically and via online fitness-based linkage learning, on a variety of problems and we compare its performance to RV-GOMEA and VkD-CMA-ES. Overview of algorithm hyperparameters can be found in supplementary material A.1. Source code of iRV-GOMEA is provided on Github[1].

4.1 Single Non-Conditional Linkage Sets

We conducted experiments for a single linkage set of size $\kappa = \ell$ up to 200 variables for each problem in Table 1. Such sizes likely cover most real-world scenarios where the maximum number of fully dependent variables is expected to be smaller [4, 8, 9, 13].

We consider iRV-GOMEA as defined in this paper, the non-incremental RV-GOMEA and RV-GOMEA$_C$, and VkD-CMA-ES. The population sizes are set to their guideline values:

- For iRV-GOMEA we followed Equation 4 with $\ell = \kappa$.
- For RV-GOMEA we used $3\ell^{1.5} + 17$ [5].
- For RV-GOMEA$_C$ and VkD-CMA-ES we used $4 + \lfloor 3 * \ln(\ell) \rfloor$, following guidelines for the CMA-ES sampling model [1, 11].

Solutions are uniformly randomly initialized in $[-115, -100]$ and the required number of evaluations to reach the VTR is averaged over 100 runs.

4.2 Scaling Multiple Linkage Sets

All algorithms are tested using static non-conditional linkage models on the Sphere and Rosenbrock problem as well as on the SoREB

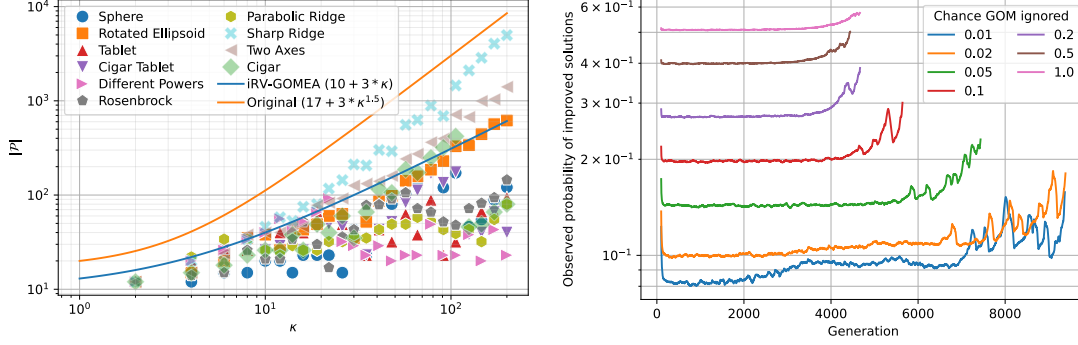---

[1]https://github.com/renzoscholman/irv-gomea

Figure 3: Population size guideline results. Left: Best performing population size $|\mathcal{P}|$ per problem and linkage set size κ (symbols) and guidelines (solid lines). Right: Observed probability of improvement when ignoring GOM on Sharp Ridge with different probabilities.

problem, which is derived from the Rotated Ellipsoid Blocks (REB), that is defined as follows:

$$f^{REB}(\mathbf{x}, c, \theta, \kappa, s) = \sum_{i=0}^{\left\lceil \frac{|\mathbf{x}|-\kappa}{s} \right\rceil - 1} f^{ellipsoid}(R^{\theta}(x_{\lceil is:is+\kappa-1 \rceil}), c)$$

$$f^{ellipsoid}(\mathbf{x}, c) = \sum_{i=0}^{|\mathbf{x}|-1} 10^{c\frac{i}{|\mathbf{x}|-1}} x_i^2$$

Here, c is the condition number of a subfunction, θ the rotation angle, κ the block size, and $s \geq 1$ the block stride. The SoREB problem is now defined as $f^{SoREB}(\mathbf{x}, \kappa) = f^{REB}(\mathbf{x}, 6, 45, \kappa, \kappa)$, i.e., it is a sum of κ-dimensional consecutively encoded rotated ellipsoids.

Again, solutions are initialized in $[-115, -100]$. For SoREB, block sizes $\kappa \in \{5, 10, 15, 20, 25, 30\}$ are used for ℓ up to 2000. For RV-GOMEA the population size guideline of $3\kappa^{1.5}+17$ is used, i.e., based on the block size. Similarly for iRV-GOMEA and RV-GOMEA$_C$, the block size κ is used instead of ℓ for the population guideline. For VkD-CMA-ES, the population size is kept at $4 + \lfloor 3 * \ln(\ell) \rfloor$, conforming with literature [1, 2]. For Rosenbrock, problem sizes up to $\ell = 10^4$, and for Sphere up to $\ell = 10^5$, are tested. A univariate linkage model is used for Sphere. For Rosenbrock, a linkage model with non-conditional, but overlapping blocks of 2 is used defined according to the additive composition of the problem definition. We investigate the average number of required evaluations to reach the VTR, which we set to 10^{-10}, over 100 runs.

4.3 Gray-Box Optimization

For online Fitness-Based (FB) linkage learning, we employ the method from [3]. We further consider static conditional linkage models based on the true VIG. For both cases, we use the recently introduced Clique Seeding (CS) method [3] to maximize the size of the conditional linkage sets. These two cases are denoted as: "FB-CS-LM" and "Static-CS-LM".

We consider various GBO problems, both separable and non-separable. Besides Sphere and Rosenbrock, as seen in Table 1, the remaining problems are defined as [3]:

$$f^{REB2Weak}(\mathbf{x}) = f^{REB}(\mathbf{x}, 1, 5, 2, 1)$$
$$f^{REB2Strong}(\mathbf{x}) = f^{REB}(\mathbf{x}, 6, 5, 2, 1)$$
$$f^{REB2Alternating}(\mathbf{x}) = f^{REB}(\mathbf{x}, c_i, \theta_i, 2, 1)$$
$$f^{REB5NoOverlap}(\mathbf{x}) = f^{REB}(\mathbf{x}, 6, 45, 5, 0)$$
$$f^{REB5SmallOverlap}(\mathbf{x}) = f^{REB}(\mathbf{x}, 6, 45, 5, 1)$$
$$f^{REB5LargeOverlap}(\mathbf{x}) = f^{REB}(\mathbf{x}, 6, 45, 5, 4)$$
$$f^{REB5Alternating}(\mathbf{x}) = f^{REB}(\mathbf{x}, c_i, \theta_i, 5, 4)$$

$$f^{REB5DisjointPairs}(\mathbf{x}) = f^{REB}(\mathbf{x}, 6, 45, 5, s_i)$$
$$f^{REB10NoOverlap}(\mathbf{x}) = f^{REB}(\mathbf{x}, 6, 45, 10, 0)$$
$$f^{REB10SmallOverlap}(\mathbf{x}) = f^{REB}(\mathbf{x}, 6, 45, 10, 1)$$
$$f^{REB10LargeOverlap}(\mathbf{x}) = f^{REB}(\mathbf{x}, 6, 45, 10, 4)$$
$$f^{REB10Alternating}(\mathbf{x}) = f^{REB}(\mathbf{x}, c_i, \theta_i, 10, 4)$$

For the alternating problems, $c_i = 1$ and $\theta_i = 5$ if i (block index) is even, else $c_i = 6$ and $\theta_i = 45$. For the disjoint pair problems, $s_i = 4$ if i is even, else $s_i = 5$.

Lastly, we use the OSoREB problem with two degrees of overlapping REB functions and the REBGrid problem. The REBGrid problem arranges the ℓ variables as vertices $v \in V$ in a square grid of dimensions $\sqrt{\ell} * \sqrt{\ell}$, and horizontally and vertically connects neighboring vertices $a \in N(b)$ through an edge. These two functions are defined as:

$$f^{OSoREB}(\mathbf{x}) = f^{REB}(\mathbf{x}, 6, 45, 5, 4) + f^{REB}(x_{[f:|\mathbf{x}|-1]}, 6, 45, 2, 5)$$

$$f^{REBGrid}(\mathbf{x}) = \sum_{v \in V} f^{ellipsoid}(R^{45}(x_{v \cup N(v)}), 6)$$

For each of the problems above, iRV-GOMEA, RV-GOMEA, and VkD-CMA-ES are tested for $\ell \in [10, 20, 40, 80, 160, 320]$. Some of the REB problems are not defined for these dimensionalities. In those cases, a proper value for ℓ was used that was as similar as possible. All algorithms have an evaluation budget of 10^8 to reach the VTR of 10^{-10}.

Population sizes are optimized for each algorithm, problem instance, and linkage model using bisection search up to $\ell = 80$. The bisection search is performed similar to how the general population guideline was determined. Due to large computation time requirements for bisection on larger dimensionalities (160, 320), the population size results are extrapolated for those. This allows testing whether incremental learning can still lead to improvements even when fully tuning the algorithms to a problem.

For iRV-GOMEA and RV-GOMEA both conditional linkage learning models are compared. In the comparison of iRV-GOMEA to VkD-CMA-ES, only the FB linkage learning is used so that no information is provided to the algorithm a priori and both algorithms learn the problem decomposition during optimization.

5 Results

5.1 Single Non-Conditional Linkage Sets

The results for varying the maximum linkage set size κ can be seen in Figure 4. We report the ratio of the required evaluations
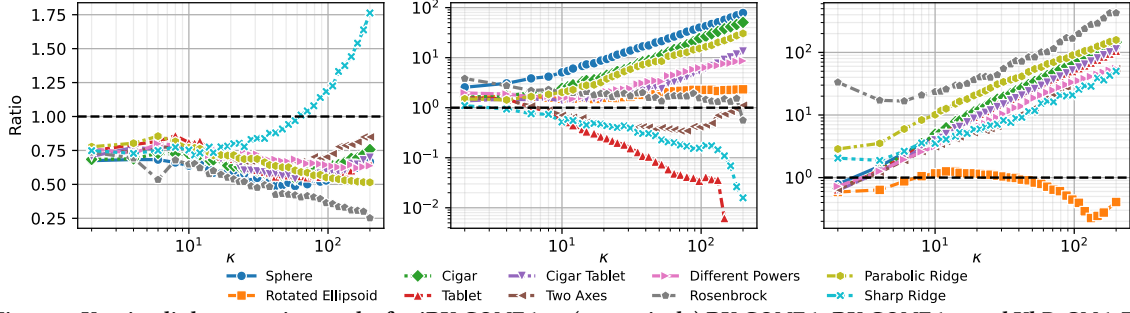
Figure 4: Varying linkage set size results for iRV-GOMEA vs (respectively) RV-GOMEA, RV-GOMEA$_C$, and VkD-CMA-ES.

between iRV-GOMEA and the other two algorithms (< 1 indicates iRV-GOMEA was better). In the supplementary material A.2 we show the required evaluations to reach the VTR.

For most problems, the required number of evaluations is decreased notably when comparing iRV-GOMEA to RV-GOMEA. However, for larger κ the ratio deteriorates, especially for Sharp Ridge. These results are in contrast to those found for iAMaLGaM versus AMaLGaM [5], in which GOM is not used. For larger linkage sets, the larger population size in the guideline of RV-GOMEA works better with the GOM operator, as more opportunities arise in each generation to find improvements. In practice, this will most likely not have many implications since linkage sets larger than several tens of parameters are not expected to typically be present.

The CMA-ES sampling model in RV-GOMEA$_C$ leads to better performance than iRV-GOMEA on various problems. This is expected, as in this setting of a single linkage set we are essentially comparing iAMaLGaM and CMA-ES in a black-box setting and CMA-ES is the state-of-the-art in this case. For Sharp Ridge, GOM was already shown to be a limiting factor that prevents improvements. This is reflected again here in the results of all three RV-GOMEA variants.

5.2 Scaling Multiple Linkage Sets

RV-GOMEA is known to perform really well in GBO settings, even outperforming RV-GOMEA$_C$ [7]. Adding incremental learning following the approach in this paper, can lead to even better results, as can be seen in Figure 5. On the SoREB problem, iRV-GOMEA outperforms RV-GOMEA substantially by reaching the VTR two to three times as fast (for problems with more than one block). Compared to RV-GOMEA$_C$ and VkD-CMA-ES, the improvements are even larger. As ℓ is increased beyond a hundred variables, VkD-CMA-ES starts to give out compared to the scalability of the RV-GOMEA variants. Even with its decomposition capabilities, it converged prematurely above $\ell = 100$ in our experiments. In the supplementary material A.3 we show the required evaluations to reach the VTR.

5.3 Gray-Box Optimization

Figure 6 shows the resulting reduction in the required number of evaluations for iRV-GOMEA versus RV-GOMEA for both linkage models when solving every benchmark problem in a GBO setting and additionally finding the optimally performing population size for each problem dimensionality. On nearly all problems, the required number of evaluations is reduced. The optimized population sizes and required evaluations can be found in the supplementary material A.4.

The only case where iRV-GOMEA is clearly worse is the fully separable Sphere problem. There, the use of the GOM operator with

univariate linkage sets allows for very small population sizes to be used also in RV-GOMEA, making incremental learning superfluous.

On average, iRV-GOMEA can be observed to achieve equal performance compared to VkD-CMA-ES when problem dimensionalities are small and there are few subfunctions. However, as problem dimensionality increases and the relative utility of partial evaluations grows, iRV-GOMEA demonstrates a substantial reduction in the number of required evaluations.

6 Discussion

Many real-world problems require the exploitation of problem-specific knowledge to reach good results within a reasonable amount of time. Of particular value can be the ability to exploit problem decomposition together with the ability to perform partial evaluations. While RV-GOMEA is already state-of-the-art and capable of outperforming traditional EAs and EDAs that cannot exploit partial evaluations, the incremental learning of Gaussian distributions had not yet been studied within the GOMEA framework, nor applied to a GBO scenario where partial variation and evaluations are possible. Therefore, in this paper we have investigated for the first time how incremental learning can be incorporated in RV-GOMEA by empirically discovering a formula for learning rates to be used when incrementally learning covariances and the AMS in RV-GOMEA.

The use of incremental learning as proposed in this paper, improves the performance of RV-GOMEA in many cases. However, when linkage set sizes increase, the performance advantage of iRV-GOMEA over RV-GOMEA and RV-GOMEA$_C$ decreases. Further research would be required to determine the root cause of this how and that could be tackled. Likely, the main reason is the fact that solutions may not change during GOM (i.e., if changes are rejected) and incremental learning then is poised to learn from the same solutions, deteriorating the covariance learning process.

While the approach taken in this paper to arrive at learning rate formulas for the covariance matrix and the AMS has proven mostly successful, the approach does involve tailoring the learning rate formulas to a set of benchmark problems. For a real-world problem, especially if is a recurring one where instances need to be solved repeatedly, it may work well to use the same setup but then for the specific problem at hand. Additional performance improvements could then well be obtained, which could have important real-world implications. It would be interesting to explore this in future work.

7 Conclusion

In this paper, we have introduced iRV-GOMEA, a new variant of RV-GOMEA that employs generation-wise incremental learning of
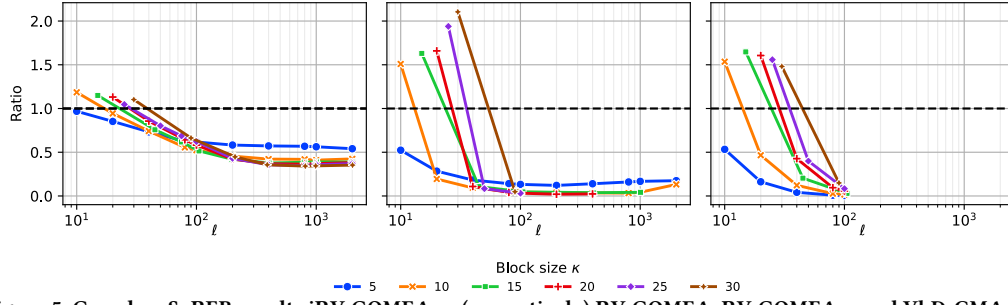
Figure 5: Gray-box SoREB results iRV-GOMEA vs (respectively) RV-GOMEA, RV-GOMEA$_C$, and VkD-CMA-ES.
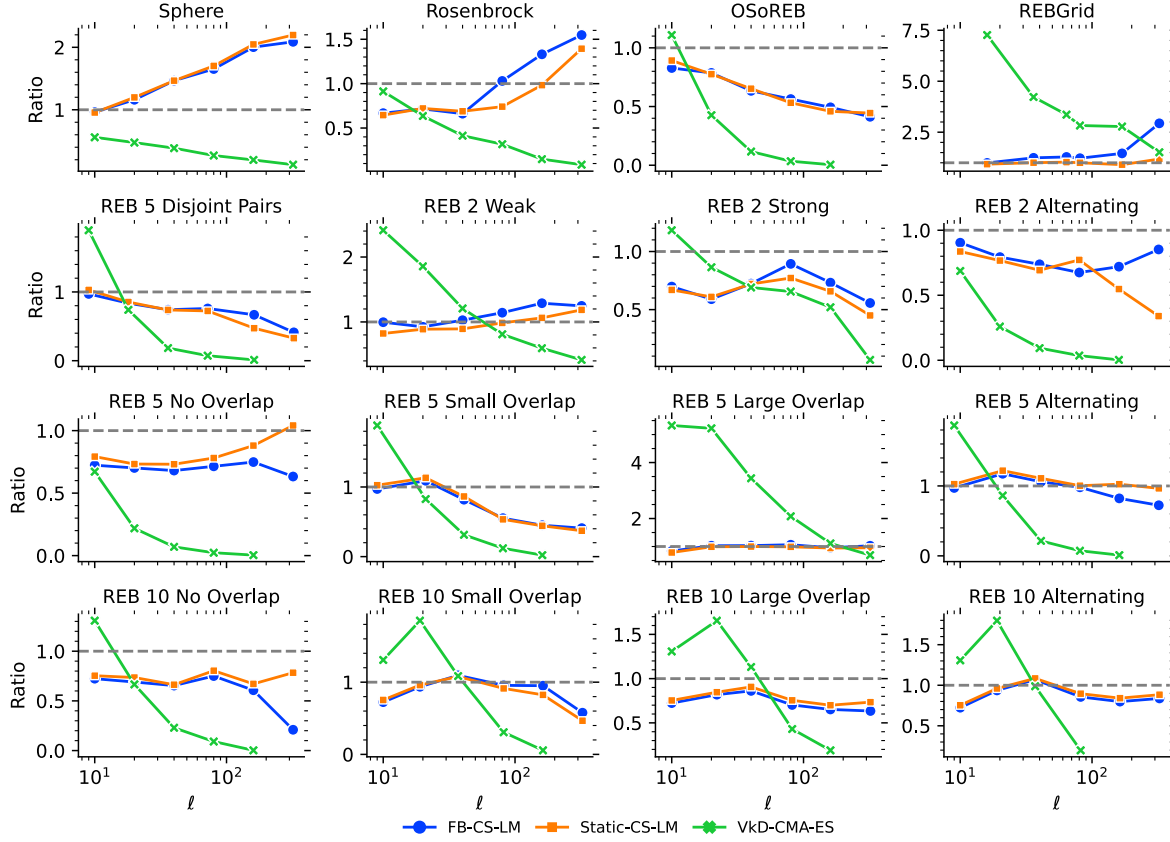

Figure 6: Ratio of required number of function evaluations to reach the VTR for iRV-GOMEA vs RV-GOMEA and for iRV-GOMEA vs VkD-CMA-ES on all tested GBO benchmark problems. A ratio below 1.0 means that iRV-GOMEA has superior performance.

the parameters of the sampling model. On several GBO problems with varying dependency structures, iRV-GOMEA was shown to be capable of outperforming RV-GOMEA, RV-GOMEA$_C$, and VkD-CMA-ES. When establishing and using general population-sizing guidelines, the required number of function evaluations could be reduced by a factor of 2 to 3 times when using iRV-GOMEA instead of RV-GOMEA. Even when tuning population sizes specifically for a problem and dimensionality, the use of incremental learning in RV-GOMEA could still lead to a reduction in the required number of evaluations. While incremental learning was not found to improve the performance of RV-GOMEA on fully decomposable (i.e., univariate) problems, when larger sets of interdependent variables are present in a problem, the efficiency gains of using incremental

learning were clear, both in case static linkage models are provided, and when fitness-based linkage learning is employed.

Acknowledgments

References

[1] AKIMOTO, Y., AND HANSEN, N. Online model selection for restricted covariance matrix adaptation. In *Parallel Problem Solving from Nature – PPSN XIV* (2016),

J. Handl, E. Hart, P. R. Lewis, M. López-Ibáñez, G. Ochoa, and B. Paechter, Eds., Springer International Publishing, pp. 3–13.

[2] Akimoto, Y., and Hansen, N. Projection-based restricted covariance matrix adaptation for high dimension. In *Proceedings of the Genetic and Evolutionary Computation Conference 2016* (2016), GECCO '16, Association for Computing Machinery, pp. 197–204.

[3] Andreadis, G., Alderliesten, T., and Bosman, P. A. N. Fitness-based linkage learning and maximum-clique conditional linkage modelling for gray-box optimization with RV-GOMEA. In *Proceedings of the Genetic and Evolutionary Computation Conference 2024* (2024), GECCO '24, Association for Computing Machinery, pp. 647 – 655.

[4] Andreadis, G., Bosman, P. A., and Alderliesten, T. MOREA: a GPU-accelerated evolutionary algorithm for multi-objective deformable registration of 3d medical images. In *Proceedings of the Genetic and Evolutionary Computation Conference 2023* (2023), GECCO '23, Association for Computing Machinery, pp. 1294–1302.

[5] Bosman, P. A. On empirical memory design, faster selection of bayesian factorizations and parameter-free gaussian EDAs. In *Proceedings of the 11th Annual Conference on Genetic and Evolutionary Computation* (2009), GECCO '09, Association for Computing Machinery, pp. 389–396.

[6] Bosman, P. A. N., Grahl, J., and Thierens, D. Enhancing the performance of maximum–likelihood gaussian EDAs using anticipated mean shift. In *Parallel Problem Solving from Nature – PPSN X* (2008), G. Rudolph, T. Jansen, N. Beume, S. Lucas, and C. Poloni, Eds., Springer, pp. 133–143.

[7] Bouter, A., Alderliesten, T., and Bosman, P. A. Achieving highly scalable evolutionary real-valued optimization by exploiting partial evaluations. *Evolutionary Computation 29*, 1 (2021), 129–155.

[8] Cai, W., Hu, Y., Fang, F., Yao, L., and Liu, J. Wind farm power production and fatigue load optimization based on dynamic partitioning and wake redirection of wind turbines. *Applied Energy 339* (2023), 121000.

[9] Chen, W., Whitley, D., Tinós, R., and Chicano, F. Tunneling between plateaus: improving on a state-of-the-art MAXSAT solver using partition crossover. In *Proceedings of the Genetic and Evolutionary Computation Conference* (2018), GECCO '18, Association for Computing Machinery, pp. 921–928.

[10] Deb, K., and Myburgh, C. Breaking the billion-variable barrier in real-world optimization using a customized evolutionary algorithm. In *Proceedings of the Genetic and Evolutionary Computation Conference 2016* (2016), GECCO '16, Association for Computing Machinery, pp. 653–660.

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

[12] Hsu, S.-H., and Yu, T.-L. Optimization by pairwise linkage detection, incremental linkage set, and restricted / back mixing: DSMGA-II. In *Proceedings of the 2015 Annual Conference on Genetic and Evolutionary Computation* (2015), GECCO '15, Association for Computing Machinery, pp. 519–526.

[13] Luong, N. H., Alderliesten, T., Bel, A., Niatsetski, Y., and Bosman, P. A. Application and benchmarking of multi-objective evolutionary algorithms on high-dose-rate brachytherapy planning for prostate cancer treatment. *Swarm and Evolutionary Computation 40* (2018), 37–52.

[14] Munetomo, M., and Goldberg, D. E. Linkage identification by non-monotonicity detection for overlapping functions. *Evolutionary Computation 7*, 4 (1999), 377–398.

[15] Neumann, A., Gounder, S., Yan, X., Sherman, G., Campbell, B., Guo, M., and Neumann, F. Diversity optimization for the detection and concealment of spatially defined communication networks. In *Proceedings of the Genetic and Evolutionary Computation Conference* (2023), GECCO '23, Association for Computing Machinery, pp. 1436–1444.

[16] Thierens, D., and Bosman, P. A. Optimal mixing evolutionary algorithms. In *Proceedings of the 13th Annual Conference on Genetic and Evolutionary Computation - (2011)*, GECCO '11, Association for Computing Machinery, pp. 617–624.

[17] Tinós, R., Whitley, D., Chicano, F., and Ochoa, G. Partition crossover for continuous optimization: ePX. In *Proceedings of the Genetic and Evolutionary Computation Conference 2021* (2021), GECCO '21, Association for Computing Machinery, pp. 627–635.

[18] Whitley, D., Quevedo De Carvalho, O., Roberts, M., Shetty, V., and Jampathom, P. Scheduling multi-resource satellites using genetic algorithms and permutation based representations. In *Proceedings of the Genetic and Evolutionary Computation Conference 2023* (2023), GECCO '23, Association for Computing Machinery, pp. 1473–1481.

[19] Whitley, L. D., Chicano, F., and Goldman, B. W. Gray box optimization for Mk landscapes (NK landscapes and MAX-kSAT). *Evolutionary Computation 24*, 3 (2016), 491–519.

[20] Wierstra, D., Schaul, T., Peters, J., and Schmidhuber, J. Natural evolution strategies. In *2008 IEEE Congress on Evolutionary Computation (IEEE World Congress on Computational Intelligence)* (2008), pp. 3381–3387. ISSN: 1941-0026.