Citation for published version


DOI

https://doi.org/10.1145/2908812.2908923

Link to record in KAR

http://kar.kent.ac.uk/55156/

Document Version

Author's Accepted Manuscript

Copyright & reuse
Content in the Kent Academic Repository is made available for research purposes. Unless otherwise stated all content is protected by copyright and in the absence of an open licence (eg Creative Commons), permissions for further reuse of content should be sought from the publisher, author or other copyright holder.

Versions of research
The version in the Kent Academic Repository may differ from the final published version. Users are advised to check http://kar.kent.ac.uk for the status of the paper. Users should always cite the published version of record.

Enquiries
For any further enquiries regarding the licence status of this document, please contact: researchsupport@kent.ac.uk
If you believe this document infringes copyright then please contact the KAR admin team with the take-down information provided at http://kar.kent.ac.uk/contact.html
A Dispersion Operator for Geometric Semantic Genetic Programming

Luiz Otavio V. B. Oliveira  
Dep. Computer Science  
UFMG  
Belo Horizonte, Brazil  
luizvbo@dcc.ufmg.br

Fernando E. B. Otero  
School of Computing  
University of Kent  
F.E.B.Otero@kent.ac.uk

Gisele L. Pappa  
Dep. Computer Science  
UFMG  
Belo Horizonte, Brazil  
glpappa@dcc.ufmg.br

ABSTRACT

Recent advances in geometric semantic genetic programming (GSGP) have shown that the results obtained by these methods can outperform those obtained by classical genetic programming algorithms, in particular in the context of symbolic regression. However, there are still many open issues on how to improve their search mechanism. One of these issues is how to get around the fact that the GSGP crossover operator cannot generate solutions that are placed outside the convex hull formed by the individuals of the current population. Although the mutation operator alleviates this problem, we cannot guarantee it will find promising regions of the search space within feasible computational time. In this direction, this paper proposes a new geometric dispersion operator that uses multiplicative factors to move individuals to less dense areas of the search space around the target solution before applying semantic genetic operators. Experiments in sixteen datasets show that the results obtained by the proposed operator are statistically significantly better than those produced by GSGP and that the operator does indeed spread the solutions around the target solution.

CSC Concepts

• Computing methodologies → Heuristic function construction; Genetic programming;

Keywords

geometric semantic genetic programming; geometric semantic crossover; convex hull

1. INTRODUCTION

In recent years, several studies have attempted to include semantic knowledge into the search mechanism of genetic programming algorithms, producing positive impacts on their search performance [19]. In supervised learning problems, including symbolic regression, the semantics of an individual is defined as the vector of output values produced when applying the function the GP individual represents to the set of training cases [19].

One of the seminal works in the field of semantic genetic programming proposed geometric versions of the crossover and mutation operators, namely geometric semantic crossover (GSX) and mutation (GSM) operators [14]. These operators guarantee the semantic fitness landscape explored by GP is conic, a property with positive effects on the search process. There is formal evidence that indicates evolutionary algorithms with geometric operators can optimise cone landscapes with good results for virtually any metric [13]. However, geometric semantic genetic programming (GSGP) still presents drawbacks.

One of them comes from the fact that, when using only the crossover operator—which produces offspring by performing a convex combination of its parents—the set of possible individuals generated during evolution is delimited by the convex hull of the semantics of the current population [16]. The convex hull of a set of points is given by the set of all convex combinations of these points [17]. Hence, if the target output is not within the convex hull, the algorithm will never be able to find it.

The mutation operator appears as the solution to this problem, as it can expand the semantic search space and, consequently, the population convex hull. However, depending on how far from the target solution the individuals in the initial population are, GSGP may take a prohibitive amount of time to reach the relevant region of the search space where the target output is.

In this context, this paper proposes a new geometric semantic operator, named geometric dispersion (GD). GD takes as input a single individual and moves it in the semantic space aiming to better distribute the population around the target output. By doing that GD increases the chances of generating a convex hull from the population semantics that contains the target output vector (out) (see Figure 1). In contrast with GSM, which randomly moves the individual within a hypersphere around it, the geometric dispersion operator is deterministic, and moves the individual along a line crossing the origin of the semantic space to sparse regions around the target output.

We performed experiments in 16 symbolic regression datasets, and compared the results with the GSGP without using the operator and a canonical GP. Results of root mean square error showed that the proposed operator leads GSGP to obtain statistically significantly better results than the other two methods. An analysis of the distribution of the solutions along the dimensions of the semantic space...
also showed the operator does spread the solutions in the semantic space.

The remainder of this paper is organized as follows. Section 2 presents background on GSGP, and Section 3 discusses different approaches proposed to increase the convex hull of the population. Section 4 introduces the geometric dispersion operator, while Section 5 shows results and empirical comparisons with GSGP and GP. Finally, Section 6 draws conclusions and presents directions for future work.

2. GEOMETRIC SEMANTIC GENETIC PROGRAMMING

This paper focuses on geometric semantic genetic programming for symbolic regression problems. Given a finite set of input-output pairs representing the training cases, defined as \( T = \{(x_i, y_i)\}_{i=1}^n \) where \((x_i, y_i) \in \mathbb{R}^d \times \mathbb{R} \) \((i = 1, 2, \ldots, n)\)—symbolic regression consists in inducing a model \( p : \mathbb{R}^d \rightarrow \mathbb{R} \) that maps inputs to outputs, such that \( \forall (x_i, y_i) \in T : p(x_i) = y_i \).

In this scenario, let \( \text{in} = [x_1, x_2, \ldots, x_n] \) and \( \text{out} = [y_1, y_2, \ldots, y_n] \) be the input and the output vectors, respectively, associated to the training cases. The semantics of a program \( p \) represented by an individual evolved by GSGP, denoted by \( s(p) \), is the vector of outputs it produces when applied to the set of inputs \( \text{in} \), i.e., \( s(p) = p(\text{in}) = [p(x_1), p(x_2), \ldots, p(x_n)] \). This notation is extended to the semantics of a population of programs \( P = \{p_1, p_2, \ldots, p_n\} \), i.e., \( s(P) = \{s(p_1), s(p_2), \ldots, s(p_n)\} \). The semantics of any program can be represented as a point in a \( n \)-dimensional topological space \( S \) (called semantic space), where \( n \) is the size of the training set. Notice that the target output \( \text{out} \) can also be represented in the semantic space.

GSGP introduces geometric semantic operators for GP that acts on the syntax of the programs, inducing a geometric behaviour on the semantic level [14]. Let \( P^0 \) be the solution set comprising all the possible candidate solutions to a problem in the arithmetic domain, the geometric semantic crossover and mutation operators are defined as follows:

**Definition 1.** Given two parent functions \( p_1, p_2 \in P^0 \), the geometric semantic arithmetic crossover \( \text{GSX} : P^0 \times P^0 \rightarrow P^0 \) returns the offspring arithmetic function

\[
\alpha = r \cdot p_1 + (1 - r) \cdot p_2 ,
\]

where \( r \) is a random real constant in \([0, 1]\) (for fitness function based on Euclidean distance) or a random real function based on Manhattan distance.

**Definition 2.** Given a parent function \( p \in P^0 \), the geometric semantic arithmetic mutation \( \text{GSM} : P^0 \times \mathbb{R}^+ \rightarrow P^0 \) with mutation step \( \varepsilon \) returns the real function

\[
\alpha = p + \varepsilon \cdot (r_1 - r_2) ,
\]

where \( r_1 \) and \( r_2 \) are random real functions.

Despite the unimodal fitness landscape, the randomness present in these operators, available in the form of random real functions or constants, has been shown to be a better way to explore the space, in terms of generalisation, when compared to modifications of these operators where the randomness is replaced by decisions based on the training error [1, 6, 9].

**Definition 3.** The convex hull of a set \( S \) of points in \( \mathbb{R}^n \), denoted as \( \mathcal{C}(S) \), is the set of all convex combinations of points in \( S \) [17].

GSGP is, by definition, a geometric crossover operator [14]. This property leads us to the following theorem regarding the convex hull of the population:

**Theorem 1.** Let \( P \) be the population at generation \( g \). For a GSGP, where the GSX operator is the only search operator available, we have \( \mathcal{C}(s(P_{g+1})) \subseteq \mathcal{C}(s(P_g)) \subseteq \ldots \subseteq \mathcal{C}(s(P_0)) \).

Theorem 1 is a particular case of the Theorem 3 defined and proved in [13], and has an important implication regarding the GSX operator. Given a population \( P \) and a semantic vector \( q \in S \), the offspring resulting from the application of GSGP to any pair of individuals in \( P \) can reach \( q \) if and only if \( q \in \mathcal{C}(s(P)) \). Consequently, if GSGP has no other search operators (only the GSX), a semantic \( q \) is reachable only if \( q \in \mathcal{C}(s(P_0)) \), i.e., if \( q \) is located inside the convex hull of the initial population.

Figure 1 illustrates this situation for a bi-dimensional semantic space. Let \( \text{out} = [0, 0] \) be the desired output vector defined by the training cases. Now consider two different populations \( P_1 \) and \( P_2 \), where the individuals from \( P_2 \) are concentrated in the first quadrant and, consequently, \( \mathcal{C}(s(P_1)) \) cannot reach the origin \( (\text{out}) \). On the other hand, the set \( s(P_2) \) is distributed along the four quadrants and \( \mathcal{C}(s(P_2)) \) embraces the desired vector. In the first scenario, GSGP needs the mutation operator to expand the convex hull to reach the solution. In the second scenario, as it is already inside the convex hull, the desired vector can be found using the crossover operator alone or it can be calculated analytically with no need to use GSGP.

3. RELATED WORK

Previous work on GSGP have proposed different approaches to take advantage of the properties of the geometric semantic space to improve search. However, to the best of our knowledge, so far only one has investigated ways to increase the area covered by the population convex hull, specially focusing on the coverage of the initial population, as discussed in this section.

Regarding operators that take advantage of the conic shape of the geometric semantic space, in [18] the authors explore the geometry of the semantic space through the concept of error vector. An error vector is represented by a point in the \( n \)-dimensional space, called error space, given by the translation \( t_e(p) = s(p) - \text{out} \). This notion is used to introduce the concept of optimally aligned individuals in the error space, i.e., given a number of dimensions \( \mu = 1, 2, \ldots, n \), where \( n \) is the size of the training set, \( \mu \) individuals are optimally aligned in the error space if they belong to the same \( \mu \)-dimensional hyperplane intersecting the origin of the error space. The authors show that if \( \mu \) individuals are optimally aligned, we can analytically obtain an equation to express the target output vector \( \text{out} \). In this context, they present GP-based methods to find optimally aligned individuals in two and three dimensions, called ESAGP-1 (Error Space
Alignment GP) and ESAGP-2, respectively. Experimental results suggest that searching for optimally aligned individuals (in two and three dimensions) is easier than directly searching for a globally optimal solution.

The authors in [8], in contrast, extend ESAGP-1 to what they call Pair Optimization GP (POGP). Unlike the original method, which represents individuals as simple expressions and computes the fitness by the angle between the error vector of an individual and a particular point called attractor, POGP represents individuals as pairs of expressions, and calculates the fitness as the angle between the error vectors of these two expressions. POGP experimental results indicate that the method deserves attention in future studies.

Concerning methods taking the area covered by the convex hull, in [16] the authors propose the Competent Initialization (CI) method, which aims to increase the convex hull of the initial population. The algorithm adopts a generalized version of the Semantically Driven Initialization (SDI) method [3], initially proposed for non-geometric spaces, to generate individuals semantically distinct. SDI picks a node randomly from the function set to combine individuals already in the population. If the resulting program has semantics different from the other members of the population, it is accepted. Otherwise, the method makes a new attempt of generating an individual. The process continues until a semantically distinct individual is created, in a trial-and-error strategy. CI, on the other hand, accepts the semantically distinct individual only if it is not in the current convex hull. The main drawback of this method is the possible waste of resources, since individuals are randomly created, evaluated and discarded when they are semantically similar to an existing member of the population or when it is already in the population’s convex hull.

Although not taking advantage of the geometric properties of the search space, in [7] the authors propose a semantic-based algorithm that keeps a distribution of different semantics during the evolution to drive GP to search in areas of the semantic space where previous good solutions were found. The method outperformed standard GP and bacterial GP [4] in the test bed adopted. However, the individuals generated presented statistically bigger sizes than the individuals generated by the other two GP variants.

In contrast with the aforementioned works, the geometric dispersion operator proposed here has a different rationale: to improve the distribution of the points around the target point, extending the convex hull of the current population.

4. THE GEOMETRIC DISPERSION OPERATOR

This section introduces the geometric dispersion (GD) operator. It takes as input a single individual $p$ and moves it in the semantic space in order to better distribute the population around the desired output in each dimension of $S$. By distributing the individuals around the desired output we increase the chances of generating a convex hull from the population semantics space that contains the desired output vector $out$ (see Figure 1b).

The aim of GD is to move $p$ to the region of the semantic space around $out$ with the lowest concentration of individuals. In order to find this region, the population distribution in the semantic space is measured. The operator calculates the proportion of individuals in each side of $out$ for each dimension of $S$, i.e., the proportion of individuals greater and smaller than $out$ in each dimension of the semantic space.

Knowing the region of the semantic space around $out$ where we want to have individuals shifted to, different methods can be used to move individual $p$. GD does that by adding a multiplicative factor ($m$) to $p$, in the form $m \cdot p$. Note that, by definition, the movement performed by GD is limited to the line crossing both the origin and individual $p$ in the semantic space (see Figure 2a). The problem is then to find the value of $m$ that maximizes the number of dimensions benefited by the displacement of $p$.

Let $count_{gt}$ and $count_{lt}$ be $n$-dimensional arrays, where the value in position $i$ corresponds to the number of individuals from $P$ greater and smaller than $out$ in dimension $i$, respectively. The GD operator moves $p$ in a way that if $count_{lt}[i] < count_{gt}[i]$—there is a concentration of individuals on the right side of $out$ in dimension $i$—GD moves $p$ in dimension $i$ to the left side of $out$. When the concentration of individuals in dimension $i$ is on the left side of $out$, GD has the opposite effect—it moves $p$ in dimension $i$ to the right side of $out$. However, given that the operator can only move $p$ in a straight line, there is no guarantee of achieving this behaviour for all dimensions.

Given this limitation, the GD operator builds a system of linear inequalities that reflects the desired behaviour of the operator and finds the value of $m$ that maximizes the number of inequalities satisfied, i.e., the number of dimensions satisfied.

\begin{align*}
\text{GD operator: } m & = \frac{\text{count}_{lt}[i] - \text{count}_{gt}[i]}{\text{count}_{lt}[i] + \text{count}_{gt}[i]} \\
\text{Conditions: } & \begin{cases}
\text{if } \text{count}_{lt}[i] < \text{count}_{gt}[i] \quad \text{move } p \text{ to the left side of } out \\
\text{if } \text{count}_{lt}[i] > \text{count}_{gt}[i] \quad \text{move } p \text{ to the right side of } out
\end{cases}
\end{align*}

By solving these linear inequalities, we can find the optimal value of $m$ for each dimension $i$. The resulting program $\hat{p}$ is the individual after applying the GD operator.
After the operator is applied to individual $p$ (showed by the number in the coloured rectangles in Figure 2b), the individuals are distributed unequally in both dimensions respectively. Figure 2 presents the GD operator in action. Initially, the method isolates $B$ upper bounds, respectively) in the set $B$ (lines 5-13). If the best interval corresponds to an upper bound, we decrement the $cSatisfied$ counter, since the interval in the right side of the bound does not satisfy it. On the other hand, if the examined value corresponds to a lower bound, the right side interval satisfies the bound and $cSatisfied$ is incremented.

After finding the best interval for $m$, the procedure assigns an actual value for $m$ (lines 14-24). If the best interval corresponds to $(-\infty, B[1])$, $m$ takes the value of $B[1]$ subtracted by one.\footnote{Although we shift the value of $m$ by one, any other constant could be adopted.} If the best interval corresponds to $(B[|B|], \infty)$, $m$ takes one added to the value of $B[|B|]$. Otherwise, the method selects a random value in the interval $(B[index], B[index+1])$. The value of $m$ returned by Algorithm 2 is then used to move individual $p$ in the semantic space. GD is applied during the evolution at every generation, right before other genetic semantic operators (crossover and mutation).

Notice that the process of finding the set of inequalities satisfied by $m$ is deterministic. The only stochastic part of the method is the selection of a value between two bounds (line 24). This fact justifies the use of other operators (crossover and mutation) in order to further explore the search space. Although the location of the optimal solution is known, making decisions based only on the distance to the solution (equivalent to the training error) may lead to overfitting, as presented in other works [1, 6, 9].

The time complexity of the operator is $O(n \log n)$, where $n$ is the number of the training cases, bounded by the sort method applied in line 4 of Algorithm 2 (we adopted merge sort). The construction of the arrays $count_{gt}$ and $count_{lt}$ occurs once per generation with time $O(n|P|)$, where $|P|$ is the population size.

5. EXPERIMENTAL ANALYSIS

In this section we present an experimental analysis of the geometric dispersion operator within the GSGP in a diver-
Algorithm 1: Build System of Inequalities

Input: Individual program \( (p) \), desired output \( (out) \), population distribution \( (count_{gt}, count_{lt}) \)

Output: Set of bounds for \( m \) \((B)\)

1. \( B \leftarrow \{\} \);
2. \( \text{for } i \leftarrow 1 \text{ to } |s(p)| \text{ do} \) // Calculate the bounds
3. \( \text{if } s(p)[i] \neq 0 \text{ then} \)
4. \( \text{value} \leftarrow \frac{val[i]}{s(p)[i]} \);
5. \( \text{if } count_{gt}[i] > count_{lt}[i] \text{ then} \)
6. \( \text{if } s(p)[i] > 0 \text{ then} \)
7. \( B \leftarrow B \cup \{\text{value, 'ub'}\}; \)
8. \( \text{else} \)
9. \( B \leftarrow B \cup \{\text{value, 'lb'}\}; \)
10. \( \text{if } count_{lt}[i] < count_{gt}[i] \text{ then} \)
11. \( \text{if } s(p)[i] > 0 \text{ then} \)
12. \( B \leftarrow B \cup \{\text{value, 'ub'}\}; \)
13. \( \text{else} \)
14. \( B \leftarrow B \cup \{\text{value, 'lb'}\}; \)
15. return \( B \);

Algorithm 2: Finds \( m \)

Input: Set of bounds for \( m \) given by Alg. 1 \((B)\)
Output: Individual multiplier \( (m) \)

1. \( n_{ub} \leftarrow \text{number of 'ub's in } B; \)
2. \( \text{maxSatisfied} \leftarrow cSatisfied \leftarrow n_{ub}; \)
3. \( \text{index} \leftarrow 0; \)
4. \( \text{Sort } B \text{ by value in ascending order}; \)
5. \( \text{for } i \leftarrow 1 \text{ to } |B| \text{ do} \) // Find best interval for \( m \)
6. \( \text{bound} \leftarrow B[i]; \)
7. \( \text{if } \text{bound.type} = \text{ 'lb'} \text{ then} \)
8. \( cSatisfied \leftarrow cSatisfied + 1; \)
9. \( \text{if } cSatisfied > \text{maxSatisfied} \text{ then} \)
10. \( \text{maxSatisfied} \leftarrow cSatisfied; \)
11. \( \text{index} \leftarrow i; \)
12. \( \text{else} \)
13. \( cSatisfied \leftarrow cSatisfied - 1; \)
14. \( \text{if } \text{index} = 0 \text{ then} \) // Calculate \( m \)
15. \( \text{if } B \text{ is empty then} \) // No need to move \( p \)
16. \( m \leftarrow 1; \)
17. \( \text{else} \)
18. \( m \leftarrow B[1].\text{value} - 1; \) // Shift by one
19. \( \text{else} \)
20. \( \text{if } \text{index} = |B| \text{ then} \)
21. \( m \leftarrow B[|B|].\text{value} + 1; \) // Shift by one
22. \( \text{else} \)
23. \( \delta \leftarrow B[\text{index} + 1].\text{value} - B[\text{index}].\text{value}; \)
24. \( \text{// rand() returns a random value in (0, 1)} \)
25. \( m \leftarrow B[\text{index}].\text{value} + \delta \cdot \text{rand();} \)
26. return \( m; \)

sified collection of real-world and synthetic datasets. The results obtained by GSGP with the GD operator, from now on referred as GSGP+ [4], compared with a canonical GP [2] and the GSGP without the use of the operator [5].

The experimental test bed, presented in Table 1, is composed of datasets selected from the UCI machine learning repository [12] and GP benchmarks [21]. For each real-world dataset, we performed a 5-fold cross-validation with 10 replications, resulting in 50 executions. For the synthetic datasets (except keijzer-6 and keijzer-7), we generated 5 samples and, for each sample, applied the algorithms 10 times, resulting again in 50 executions. For keijzer-6 and keijzer-7, the test set is fixed, so we performed 50 executions. The categorical attributes, namely ‘vendor name’ and ‘model name’ from the cpu dataset and ‘month’ and ‘day’ from the forestFires dataset, were removed for compatibility purposes.

All executions used a population of 1000 individuals evolved for 2000 generations with tournament selection of size 10. The grow method [11] was adopted to generate the random functions inside the geometric semantic crossover and mutation operators, and the ramped half-and-half method [11] used to generate the initial population, both with maximum individual depth equal to 6. The function set included three binary arithmetic operators \((+, -, \times)\) and the analytic quotient (AQ) [15] as an alternative to the arithmetic division. The terminal set included the variables of the problem and constant values randomly picked from the interval \([-1, 1]\). The GP method employed the canonical crossover and mutation operators [11] with probabilities 0.9 and 0.1, respectively. GSGP employed the geometric semantic crossover for fitness function based on Manhattan distance and mutation operators, as presented in [5], both with probability 0.5.

The differences in the probabilities of applying the mutation and crossover operators in GP and GSGP/GSGP+ are justified based on previous work. While GP typically applies mutation with probability much smaller than the crossover [2], GSGP adopts relatively high mutation rate in order to be able to effectively explore the search space [20].

5.1 Parameter tuning

As we are interested in understanding the impact of the GD operator, we fix the other GSGP parameters and focus on looking at the results as we varied the probability values of GD. Note that the proposed operator cannot be used in isolation or without the mutation operator (i.e. only with crossover). This is because GD works with the restriction that individuals have to be moved along the line that crosses the desired output and the individual to be moved. As individuals movements are restricted to this line, it can stagnate the search. Hence, mutation still plays an important role when GD is applied, while GD main advantage is its ability to move individuals in the semantic space to enhance the convex hull generated by the population. For this reason, all experiments reported in this paper use GD, crossover and mutation operators with their respective probabilities.

One thing we realized during preliminary experiments was that moving individuals around the semantic spaces in the initial generations has a high positive impact in search. However, this behaviour is not maintained during the whole evolutionary process, leading to negative or no effects in later generations. Results of these experiments are omitted due to space constraints. For this reason, we proposed to dynamically adjust the probability \( pgd \) of applying GD as follows:
Table 1: Median training RMSE of the GSGP+ with different values of $\alpha$ and $pgd_0$ for the adopted test bed. The smallest RMSE for each dataset is presented in bold.

<table>
<thead>
<tr>
<th>Dataset</th>
<th>$\alpha = 0$</th>
<th>$\alpha = 2$</th>
<th>$\alpha = 5$</th>
<th>$\alpha = 10$</th>
<th>$\alpha = 0$</th>
<th>$\alpha = 2$</th>
<th>$\alpha = 5$</th>
<th>$\alpha = 10$</th>
<th>$\alpha = 0$</th>
<th>$\alpha = 2$</th>
<th>$\alpha = 5$</th>
<th>$\alpha = 10$</th>
</tr>
</thead>
<tbody>
<tr>
<td>airfoil†</td>
<td>1.5180</td>
<td>1.4754</td>
<td>1.4517</td>
<td>1.4600</td>
<td>1.6117</td>
<td>1.4912</td>
<td>1.4566</td>
<td>1.4566</td>
<td>1.8437</td>
<td>1.5306</td>
<td>1.4715</td>
<td>1.4654</td>
</tr>
<tr>
<td>bioavailability†</td>
<td>1.3029</td>
<td>1.1031</td>
<td>1.0323</td>
<td>1.0464</td>
<td>1.8538</td>
<td>1.1723</td>
<td>1.0593</td>
<td>1.0241</td>
<td>3.1383</td>
<td>1.3819</td>
<td>1.0892</td>
<td>1.0352</td>
</tr>
<tr>
<td>concrete†</td>
<td>2.8193</td>
<td>2.7705</td>
<td>2.7500</td>
<td>2.7469</td>
<td>2.9370</td>
<td>2.7824</td>
<td>2.7503</td>
<td>2.7462</td>
<td>3.0971</td>
<td>2.8285</td>
<td>2.7881</td>
<td>2.7319</td>
</tr>
<tr>
<td>cpu†</td>
<td>1.1506</td>
<td>0.9829</td>
<td>1.0358</td>
<td>1.0358</td>
<td>1.2904</td>
<td>1.1042</td>
<td>1.0083</td>
<td>0.9685</td>
<td>1.4872</td>
<td>1.2693</td>
<td>1.0930</td>
<td>1.0863</td>
</tr>
<tr>
<td>energyCooling†</td>
<td>0.9632</td>
<td>1.0203</td>
<td>0.9966</td>
<td>0.9995</td>
<td>1.1199</td>
<td>1.0520</td>
<td>1.0036</td>
<td>0.9853</td>
<td>1.2435</td>
<td>1.0769</td>
<td>1.0301</td>
<td>1.0086</td>
</tr>
<tr>
<td>energyHeating†</td>
<td>0.7577</td>
<td>0.7216</td>
<td>0.7329</td>
<td>0.7025</td>
<td>0.8191</td>
<td>0.7710</td>
<td>0.7440</td>
<td>0.7299</td>
<td>0.9325</td>
<td>0.8011</td>
<td>0.7448</td>
<td>0.7374</td>
</tr>
<tr>
<td>keijzer-5†</td>
<td>0.0481</td>
<td>0.0461</td>
<td>0.0447</td>
<td>0.0449</td>
<td>0.0430</td>
<td>0.0447</td>
<td>0.0443</td>
<td>0.0431</td>
<td>0.0261</td>
<td>0.0262</td>
<td>0.0311</td>
<td>0.0347</td>
</tr>
<tr>
<td>keijzer-6†</td>
<td>0.0067</td>
<td>0.0072</td>
<td>0.0069</td>
<td>0.0071</td>
<td>0.0092</td>
<td>0.0091</td>
<td>0.0097</td>
<td>0.0090</td>
<td>0.0095</td>
<td>0.0097</td>
<td>0.0086</td>
<td>0.0101</td>
</tr>
<tr>
<td>keijzer-7†</td>
<td>0.0166</td>
<td>0.0173</td>
<td>0.0162</td>
<td>0.0146</td>
<td>0.0195</td>
<td>0.0177</td>
<td>0.0192</td>
<td>0.0194</td>
<td>0.0178</td>
<td>0.0189</td>
<td>0.0173</td>
<td>0.0171</td>
</tr>
<tr>
<td>ppb†</td>
<td>0.0043</td>
<td>0.0033</td>
<td>0.0028</td>
<td>0.0029</td>
<td>0.0070</td>
<td>0.0041</td>
<td>0.0030</td>
<td>0.0028</td>
<td>0.0045</td>
<td>0.0055</td>
<td>0.0036</td>
<td>0.0030</td>
</tr>
<tr>
<td>vladislavleva-1†</td>
<td>0.0122</td>
<td>0.0119</td>
<td>0.0118</td>
<td>0.0116</td>
<td>0.0128</td>
<td>0.0122</td>
<td>0.0120</td>
<td>0.0119</td>
<td>0.0148</td>
<td>0.0121</td>
<td>0.0118</td>
<td>0.0116</td>
</tr>
<tr>
<td>vladislavleva-4†</td>
<td>0.0401</td>
<td>0.0387</td>
<td>0.0383</td>
<td>0.0376</td>
<td>0.0428</td>
<td>0.0396</td>
<td>0.0386</td>
<td>0.0380</td>
<td>0.0476</td>
<td>0.0404</td>
<td>0.0389</td>
<td>0.0386</td>
</tr>
<tr>
<td>wineRed†</td>
<td>0.3397</td>
<td>0.3259</td>
<td>0.3217</td>
<td>0.3192</td>
<td>0.3731</td>
<td>0.3375</td>
<td>0.3265</td>
<td>0.3222</td>
<td>0.4225</td>
<td>0.3498</td>
<td>0.3296</td>
<td>0.3234</td>
</tr>
<tr>
<td>wineWhite†</td>
<td>0.5469</td>
<td>0.5388</td>
<td>0.5346</td>
<td>0.5332</td>
<td>0.5669</td>
<td>0.5447</td>
<td>0.5377</td>
<td>0.5350</td>
<td>0.5938</td>
<td>0.5533</td>
<td>0.5406</td>
<td>0.5363</td>
</tr>
</tbody>
</table>

* Real-world dataset † Synthetic dataset

$$pgd_0 = pgd_0 \cdot \exp \left( -\frac{\alpha \cdot g}{g_{\max}} \right),$$

where $pgd_0$ is the base probability, $\alpha$ is the decay rate, $g$ is current generation index and $g_{\max}$ is total number of generations. Equation 5 ensures the probability of applying the operator decays exponentially with the generations.

We tested the application of GD with different initial probabilities, and here we show the best results, which were obtained by setting $pgd_0$ to 0.2, 0.4 and 0.6. For these three values, we vary the parameter $\alpha$ from Equation 5, which defines the decay rate of $pgd_0$ along the generations. Note that when $\alpha$ equals to 0, the decay function is not used, and $pgd_0$ becomes constant along the generations.

Table 1 presents the results of root mean squared error (RMSE) for different combinations of $pgd_0$ and $\alpha$ in the training set. The results show that lower values for $pgd_0$ (0.2, 0.4), allied with higher $\alpha$ values (10) tend to reduce the training RMSE. This fact indicates the application of the GD operator with low probability only in the early generations, which can reduce the training RMSE in relation to other methods.

The second aspect we evaluate is whether GD actually improves the distribution of the population around the desired output $out$ or not. For that, we propose a new measure, called mean dimension distribution (MDD). For each dimension $d$ of the training set, it is defined as

$$mdd(P, d) = \text{abs} \left( \frac{1}{|P|} \sum_{i=1}^{|P|} \text{ge}(s(p_i)[d], out[d]) - 0.5 \right),$$

where $P$ is the population, $|P|$ is the population size, $p_i$ is its $i$-th individual, $\text{ge}(a, b)$ returns 1 if $a \geq b$ and 0 otherwise and $\text{abs}(a)$ returns the absolute value of $a$. MDD is defined in the [0, 0.5] interval. A higher $mdd(P, d)$ value means that $P$ is evenly distributed in dimension $d$, i.e., half the individuals are greater than or equal to and half are smaller than $out$ in dimension $d$, while $mdd(P, d) = 0.5$ means the population is badly distributed, i.e., all individuals are greater than or equal to $out$ and none is smaller than $out$ in the dimension $d$ or vice-versa.

Figure 3a shows an analysis of the MDD throughout the generations for dataset bioavailability using a heatmap. We reported the mean of the values resulting from ten replications of the first fold adopted in the test bed, using the best parameter combination from Section 5.1. The values in
Table 2: Training and test RMSE’s (median and IQR) obtained by the algorithms for each dataset.

<table>
<thead>
<tr>
<th>Dataset</th>
<th>GSGP+ Training Median</th>
<th>GSGP+ Test Median</th>
<th>GSGP Training Median</th>
<th>GSGP Test Median</th>
<th>GP Training Median</th>
<th>GP Test Median</th>
</tr>
</thead>
<tbody>
<tr>
<td>Dataset</td>
<td>Median</td>
<td>IQR</td>
<td>Median</td>
<td>IQR</td>
<td>Median</td>
<td>IQR</td>
</tr>
<tr>
<td>airfoil</td>
<td>1.886</td>
<td>0.041</td>
<td>2.131</td>
<td>0.243</td>
<td>7.885</td>
<td>0.527</td>
</tr>
<tr>
<td>bioavailability</td>
<td>9.695</td>
<td>0.690</td>
<td>30.860</td>
<td>4.426</td>
<td>9.885</td>
<td>0.689</td>
</tr>
<tr>
<td>concrete</td>
<td>3.654</td>
<td>0.118</td>
<td>5.144</td>
<td>0.635</td>
<td>3.647</td>
<td>0.138</td>
</tr>
<tr>
<td>cpu</td>
<td>6.151</td>
<td>0.905</td>
<td>30.837</td>
<td>14.563</td>
<td>6.126</td>
<td>0.665</td>
</tr>
<tr>
<td>energyCooling</td>
<td>1.271</td>
<td>0.066</td>
<td>1.531</td>
<td>0.159</td>
<td>1.257</td>
<td>0.070</td>
</tr>
<tr>
<td>energyHeating</td>
<td>0.798</td>
<td>0.083</td>
<td>0.971</td>
<td>0.128</td>
<td>0.802</td>
<td>0.113</td>
</tr>
<tr>
<td>forestfires</td>
<td>30.967</td>
<td>4.201</td>
<td>50.227</td>
<td>48.373</td>
<td>30.737</td>
<td>4.626</td>
</tr>
<tr>
<td>keijzer-5</td>
<td>0.026</td>
<td>0.008</td>
<td>0.028</td>
<td>0.009</td>
<td>0.045</td>
<td>0.003</td>
</tr>
<tr>
<td>keijzer-6</td>
<td>0.007</td>
<td>0.006</td>
<td>0.281</td>
<td>0.282</td>
<td>0.007</td>
<td>0.005</td>
</tr>
<tr>
<td>keijzer-7</td>
<td>0.016</td>
<td>0.009</td>
<td>0.017</td>
<td>0.009</td>
<td>0.017</td>
<td>0.010</td>
</tr>
<tr>
<td>ppb</td>
<td>0.954</td>
<td>0.305</td>
<td>28.568</td>
<td>6.170</td>
<td>0.917</td>
<td>0.266</td>
</tr>
<tr>
<td>towerData</td>
<td>20.405</td>
<td>0.621</td>
<td>21.769</td>
<td>1.252</td>
<td>20.436</td>
<td>0.610</td>
</tr>
<tr>
<td>vladislavleva-1</td>
<td>0.012</td>
<td>0.002</td>
<td>0.044</td>
<td>0.030</td>
<td>0.012</td>
<td>0.002</td>
</tr>
<tr>
<td>vladislavleva-4</td>
<td>0.038</td>
<td>0.001</td>
<td>0.051</td>
<td>0.004</td>
<td>0.038</td>
<td>0.001</td>
</tr>
<tr>
<td>wineRed</td>
<td>0.494</td>
<td>0.011</td>
<td>0.615</td>
<td>0.046</td>
<td>0.493</td>
<td>0.011</td>
</tr>
<tr>
<td>wineWhite</td>
<td>0.642</td>
<td>0.003</td>
<td>0.696</td>
<td>0.015</td>
<td>0.641</td>
<td>0.003</td>
</tr>
</tbody>
</table>

Avg. rank* 1.250
Finner APV* 2.063

Table 3: Median elapsed time (in seconds) of the GSGP and GSGP*. All differences are statistically significant according to a Wilcoxon test with a confidence level of 95%.

<table>
<thead>
<tr>
<th>Dataset</th>
<th>GSGP+</th>
<th>GSGP</th>
<th>Difference (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>airfoil</td>
<td>1708.3</td>
<td>1523.9</td>
<td>12.10</td>
</tr>
<tr>
<td>bioavailability</td>
<td>421.0</td>
<td>387.6</td>
<td>8.62</td>
</tr>
<tr>
<td>concrete</td>
<td>1229.7</td>
<td>1195.0</td>
<td>2.90</td>
</tr>
<tr>
<td>cpu</td>
<td>327.4</td>
<td>313.4</td>
<td>4.48</td>
</tr>
<tr>
<td>energyCooling</td>
<td>852.9</td>
<td>801.1</td>
<td>6.47</td>
</tr>
<tr>
<td>energyHeating</td>
<td>826.8</td>
<td>782.3</td>
<td>5.70</td>
</tr>
<tr>
<td>forestfires</td>
<td>585.5</td>
<td>555.8</td>
<td>5.34</td>
</tr>
<tr>
<td>keijzer-5</td>
<td>10579.3</td>
<td>9287.9</td>
<td>13.90</td>
</tr>
<tr>
<td>keijzer-6</td>
<td>234.0</td>
<td>224.5</td>
<td>4.23</td>
</tr>
<tr>
<td>keijzer-7</td>
<td>1021.2</td>
<td>995.1</td>
<td>2.62</td>
</tr>
<tr>
<td>ppb</td>
<td>215.4</td>
<td>203.4</td>
<td>5.92</td>
</tr>
<tr>
<td>towerData</td>
<td>5357.9</td>
<td>4997.3</td>
<td>7.22</td>
</tr>
<tr>
<td>vladislavleva-1</td>
<td>1847.4</td>
<td>1870.3</td>
<td>-1.23</td>
</tr>
<tr>
<td>vladislavleva-4</td>
<td>5344.2</td>
<td>5074.2</td>
<td>5.32</td>
</tr>
<tr>
<td>wineRed</td>
<td>1597.2</td>
<td>1447.7</td>
<td>10.33</td>
</tr>
<tr>
<td>wineWhite</td>
<td>4690.9</td>
<td>4339.7</td>
<td>8.09</td>
</tr>
</tbody>
</table>

* Regarding the median test RMSE
(a) Heatmap with the differences between GSGP+ and GSGP according to MDD for each instance in generations numbers multiple of 10.

(b) Median RMSE in the training set.

(c) Median RMSE in the test set.

Figure 3: MME Heatmap and evolution of the RMSE over the generations for GSGP+ and GSGP in the dataset bioavailability.

7. ACKNOWLEDGEMENTS
The authors would like to thank CAPES, CNPq (141985/2015-1) and Fapemig for their financial support.

8. REFERENCES