Reducing Antagonism between Behavioral Diversity and Fitness in Semantic Genetic Programming

Marcin Szubert
Dept. of Computer Science
University of Vermont
mszubert@uvm.edu

Anuradha Kodali
UC Santa Cruz / NASA Ames
Research Center

Sangram Ganguly
BAERI / NASA Ames
Research Center

Kamalika Das
UC Santa Cruz / NASA Ames
Research Center

Josh C. Bongard
Dept. of Computer Science
University of Vermont

ABSTRACT

Maintaining population diversity has long been considered fundamental to the effectiveness of evolutionary algorithms. Recently, with the advent of novelty search, there has been an increasing interest in sustaining behavioral diversity by using both fitness and behavioral novelty as separate search objectives. However, since the novelty objective explicitly rewards diverging from other individuals, it can antagonize the original fitness objective that rewards convergence toward the solution(s). As a result, fostering behavioral diversity may prevent proper exploitation of the most interesting regions of the behavioral space, and thus adversely affect the overall search performance. In this paper, we argue that an antagonism between behavioral diversity and fitness can indeed exist in semantic genetic programming applied to symbolic regression. Minimizing error draws individuals toward the target semantics but promoting novelty, defined as a distance in the semantic space, scatters them away from it. We introduce a less conflicting novelty metric, defined as an angular distance between two program semantics with respect to the target semantics. The experimental results show that this metric, in contrast to the other considered diversity promoting objectives, allows to consistently improve the performance of genetic programming regardless of whether it employs a syntactic or a semantic search operator.

Keywords

genetic programming; program semantics; novelty search; diversity; geometric crossover; symbolic regression

1. INTRODUCTION

In analogy to the importance of genetic diversity in natural evolution, preserving population diversity has long been perceived as being crucial to the performance of evolutionary algorithms. Intuitively, maintaining a diverse pool of candidate solutions provides better exploration of the search space and thus gives more opportunities to discover novel, potentially fitter individuals. On the other hand, losing diversity can lead to the well-known problem of *premature convergence*, where a population stagnates at local optima and is unlikely to make any further progress.

A number of diversity maintenance techniques have been proposed to mitigate the problem of premature convergence [10, 26]. Most of these methods modify the selection process by promoting the individuals that are most different from the rest of the population. One particular approach relies on multiobjective evaluation of individuals with two objectives: the original fitness of the solution and some measure of its *novelty* designed to promote diversity. Although earlier studies measured novelty by comparing genotypes [6], recent work has successfully employed novelty metrics based on the distance between behaviors [16, 17, 23].

However, since behavioral novelty promotes increasing distance between behaviors while the fitness function typically rewards minimizing distance to the target behavior, we hypothesize that in some cases these two objectives can be overly antagonistic with each other. Consequently, promoting diversity can result in spreading individuals over the behavioral space and slowing down the convergence of the search process. In other words, under certain conditions, employing such conflicting objectives may result in excessive exploration of the entire behavioral space and insufficient exploitation of its most promising regions.

In this paper, we investigate the relationship between behavioral diversity and fitness of evolved individuals in the context of genetic programming (GP), where behavior of an individual can be identified with program semantics. In particular, we attempt to determine whether and under what conditions promoting behavioral diversity can adversely affect the search effectiveness. To this end, we consider four diversity promoting objectives and examine how each of them, used along with the fitness objective, affects the performance of tree-based GP. Moreover, we compare the fitness of programs evolved with two types of search operators: traditional subtree-swapping crossover and locally geometric semantic crossover. Since fitness landscapes induced by the latter are supposedly smoother and easier to search with the fitness objective alone, we expect to observe different effects of promoting diversity.

GECCO '16, July 20-24, 2016, Denver, CO, USA

DOI: http://dx.doi.org/10.1145/2908812.2908939

The results obtained on a set of symbolic regression problems demonstrate that some diversity objectives can be indeed detrimental to the search performance, supposedly because of being overly antagonistic with the fitness objective. In particular, we show that using straightforward Euclidean semantic novelty metric can lead to reduced performance with respect to the conventional genetic programming. By contrast, the introduced angular semantic novelty metric, designed to be less antagonistic with the fitness objective, allows to consistently improve both fitness and generalization performance, regardless of the employed search operator.

The remainder of this paper is structured as follows. The next section describes the paradigm of semantic genetic programming and presents geometric semantic operators. Section 3 gives a brief overview of diversity maintenance methods applied in GP and introduces the two aforementioned semantic novelty metrics. Sections 4 and 5 describe experimental setup and present the results. Finally, sections 6 and 7 provide discussion and concluding remarks.

2. SEMANTIC GENETIC PROGRAMMING

Standard tree-based GP searches the space of programs using traditional operators of subtree-swapping crossover and subtree-replacing mutation. These operators are designed to be generic and produce syntactically correct offspring regardless of the problem domain. However, their actual effects on the behavior of the program, and thus its fitness, are generally hard to predict. Because of the complex genotype-phenotype mapping characterized by low locality, even a minimal change at the syntax level may diametrically alter program semantics. Such large phenotypic changes are often considered problematic because, according to Fisher's geometric model [7], the probability of the mutation being beneficial is inversely proportional to its magnitude.

Recently, many alternative search operators have been proposed that take into account the effect of syntactic modifications on program semantics [1, 3, 15, 22, 30]. In order to control the scope of behavioral change, most of these methods adopt common definition of program semantics, known as *sampling semantics* [30], which is identified with the vector of outputs produced by a program for a sample of possible inputs. For instance, in supervised learning, where n input-output pairs are given as a training set $T = \{(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_n, y_n)\}$, semantics of a program p is equal to vector $\mathbf{s}(p) = [p(\mathbf{x}_1), \ldots, p(\mathbf{x}_n)]$, where $p(\mathbf{x})$ is a result obtained by running program p on input $\mathbf{x}$. Consequently, each program p corresponds to a point in n-dimensional semantic space and a metric d can be adopted to measure semantic distance between two programs. Furthermore, fitness of a program p can be calculated as a distance between its semantics $\mathbf{s}(p)$ and the target semantics $\mathbf{t} = [y_1, \ldots, y_n]$ defined by the training set, i.e., $f(p) = d(\mathbf{s}(p), \mathbf{t})$.

Importantly, the information about program semantics can be exploited not only at the level of search operators but also for other purposes, e.g, to maintain semantic diversity [11], to initialize the population [2] or to drive the selection process [18]. All such semantic-aware methods are collectively captured by the umbrella term of semantic genetic programming [31]. Recently, a paradigm of behavioral program synthesis [13] has been proposed, which extends semantic GP by using information not only about final program results but also about behavioral characteristics of program execution.

2.1 Geometric Semantic Operators

One particularly interesting class of semantic-aware search operators are geometric semantic operators introduced by Moraglio *et al.* [22]. These operators not only incorporate knowledge about program semantics but also exploit geometric structure of the semantic space endowed by a metric-based fitness function. As a result, fitness landscapes seen by these operators are smooth conic landscapes, which are in principle easy to search. In particular, a geometric semantic crossover under the metric d guarantees that semantics of each offspring p' is located in the d-metric segment connecting semantics of its parents p_1 and p_2, i.e., $d(\mathbf{s}(p_1), \mathbf{s}(p_2)) = d(\mathbf{s}(p_1), \mathbf{s}(p')) + d(\mathbf{s}(p'), \mathbf{s}(p_2))$

Although *exact* geometric crossover has been proposed [22], its practical applicability is limited because it leads to exponential growth of the program size. For this reason, alternative operators exist that employ heuristic methods to produce an *approximately* geometric offspring [14, 15]. Previous studies demonstrate that such approximately geometric operators can be still effective while producing much shorter offspring programs than exactly geometric ones.

2.2 Locally Geometric Crossover

In this paper we use Locally Geometric Crossover (LGX) proposed by Krawiec and Pawlak [15]. This operator is arguably the easiest to implement among existing approximately geometric crossover operators. Before applying a crossover, a library of short programs (procedures) must be created. Typically, a static library is generated by enumerating all possible trees lower than a predefined height. Alternatively, a dynamic library could be created at each generation from all subtrees existing in the population.

Given two parents p_1 and p_2, the operator starts by identifying their structurally common region, i.e., the largest region where the parent trees have the same topology. Two crossover points are selected by drawing a pair of corresponding nodes from the common region. Then, for the subtrees p_1' and p_2' rooted at the crossover points, semantics of the midpoint between them (i.e., semantically intermediate subprogram) is calculated as $\mathbf{s}_m = (\mathbf{s}(p_1') + \mathbf{s}(p_2'))\,/\,2$. The library is searched for programs that are semantically closest to $\mathbf{s}_m$ according to adopted metric d. From a set of k closest programs found in a library, a random one is selected and used to replace subtrees p_1' and p_2' in both parents, producing two offspring. In a rare situation when both subtrees p_1' and p_2' are semantically equivalent, a random procedure is drawn from a library.

3. PROMOTING DIVERSITY IN GP

Diversity maintenance has been a long-standing issue in GP and a number of methods have been proposed to preserve diversity in a population [4]. Most of the early studies in this area focus on genotypic diversity, which refers to structural differences between programs in a population [6, 21]. In recent years, with the advent of semantic GP, more attention has been paid to semantic or behavioral diversity [2, 9, 11, 19]. The notion of semantic diversity is particularly important in GP, because the mapping between programs and their semantics is usually a complex, non-injective function. In particular, since many syntactically different programs may exhibit the same behavior, genotypic diversity does not necessarily imply semantic diversity while the converse is often true.

Despite the assumed importance of semantic diversity [31], there have been few empirical investigations into effects of promoting behavioral diversity on the effectiveness of genetic programming. Moreover, almost all of the studied methods are limited to ensuring that the genetic operators do not produce offspring that is semantically equivalent to their parents [1, 30, 11, 9]. To the best of our knowledge, the only exception is the work of Nguyen et al. [24]. The authors apply both syntactic and semantic distance metrics in the fitness sharing mechanism and demonstrate that only using the latter improves GP performance on selected symbolic regression problems.

Here, rather than fitness sharing we adopt multiobjective approach treating diversity as a separate objective. In the following we describe four considered variants of multiobjective GP, which differ only with respect to the objective used to encourage diversity. In particular, two of the objectives (age and structural density) have already proved successful in improving GP performance. Additionally, we propose two other objectives which are essentially behavioral novelty metrics designed to promote semantic diversity.

3.1 Age-Fitness Pareto Optimization

Age-Fitness Pareto Optimization (AFPO, [27]) is a multiobjective method that relies on the concept of *genotypic age* of an individual, defined as the number of generations its genetic material has been in the population [10]. The age attribute is intended to protect young individuals before being dominated by older already optimized solutions. Each randomly initialized individual starts with age of one which is then incremented by one every generation. An offspring inherits age of the older parent.

The AFPO algorithm is based on the ParetoGP method which was originally proposed to address the issue of bloat in GP [28]. The algorithm starts with a population of n randomly initialized individuals. In each generation, it proceeds by selecting random parents from the population and applying crossover and mutation operators (with certain probability) to produce $n-1$ offspring. The offspring, together with a single randomly initialized individual, are added to the population extending its size to $2n$. Then, Pareto tournament selection is iteratively applied by randomly selecting a subset of individuals and removing the dominated ones until the size of the population is reduced back to n. To determine which individuals are dominated, the algorithm identifies the Pareto front using two objectives (both minimized): age and fitness (distance to the target semantics).

3.2 Density-Fitness Pareto Optimization

Recently, Burks and Punch [5] proposed an alternative variant of the AFPO algorithm called Density-Fitness Pareto Optimization (DFPO). This method relies on the idea of a *genetic marker*, which refers to concatenated fragments of a program tree. The authors used markers based on the topmost part of a tree and calculated *structural density* of each individual as a fraction of individuals in the population that share the same marker. Employing such a density measure as a minimized objective is intended to maintain a specific form of structural diversity focused on the rooted portions of the trees. According to the reported results obtained on three different problems (including symbolic regression), using density instead of age allows DFPO to further improve the performance achieved by AFPO.

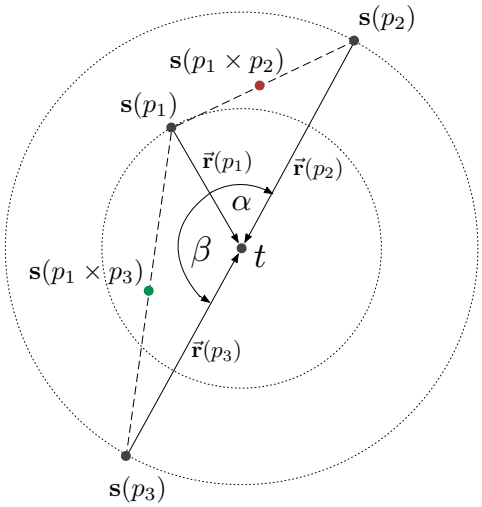

Figure 1: Residual vectors in two-dimensional semantic space where fitness is expressed using Euclidean distance.

3.3 Novelty-Fitness Pareto Optimization

Inspired by novelty search [16], we propose two behavioral novelty metrics that can be used as search objectives. Since both objectives refer to the distribution of programs in the semantic space, maximizing them is intended to promote some form of behavioral diversity. The bi-objective algorithm employing fitness and a behavioral novelty objective is termed Novelty-Fitness Pareto Optimization.

Euclidean Semantic Novelty. Since in this work we focus on real-valued symbolic regression problems, the semantic space is n-dimensional real space. Consequently, we can calculate behavioral distance between programs as a Euclidean distance between their semantics. We define Euclidean semantic novelty of a program p as a mean Euclidean distance between its semantics $\mathbf{s}(p)$ and semantics of its k nearest neighbors in the semantic space:

$$\rho(p) = \frac{1}{k}\sum_{i=1}^{k} d(\mathbf{s}(p), \mathbf{s}(\mu_i)),$$

where k is user-defined parameter and μ_i is i-th nearest program with respect to the semantic distance.

Angular Semantic Novelty. The second proposed novelty metric focuses on angles in the semantic space (see Fig. 1). Measuring angular distance between program semantics has been recently applied in GP [25] but not for the purpose of maintaining diversity. For each program p, we define *residual vector* $\mathbf{r}(p)$ as a difference between target semantics and the program semantics, i.e., $\mathbf{r}(p) = \mathbf{t} - \mathbf{s}(p)$. We define angular semantic novelty of a program p as a mean angle between its residual vector $\mathbf{r}(p)$ and residual vectors of its k nearest neighbors with respect to the angular distance:

$$\rho(p) = \frac{1}{k}\sum_{i=1}^{k} \arccos \frac{\mathbf{r}(p)\cdot\mathbf{r}(\mu_i)}{\|\mathbf{r}(p)\|\|\mathbf{r}(\mu_i)\|}.$$

Table 1: Symbolic regression benchmarks.

Problem	Objective function
QUARTIC	$4x^4 + 3x^3 + 2x^2 + x$
NONIC	$\sum_1^9 x^i$
R1	$(x+1)^3/(x^2 - x + 1)$
R2	$(x^5 - 3x^3 + 1)/(x^2 + 1)$
KEIJZER-4	$x^3 e^{-x} \cos(x) \sin(x)(\sin^2(x)\cos(x) - 1)$

We expect that using this novelty metric as an additional search objective can be beneficial for two reasons. First, this objective is less conflicting with fitness than Euclidean semantic novelty — a population of very fit individuals can at the same time exhibit high angular semantic diversity. Second, promoting large angles between residual vectors makes it more likely that the parents occupy the opposite slopes of the fitness landscape, which is advantageous for geometric semantic crossover. For instance, consider three programs illustrated in Fig. 1. Let us assume that p_1 is the first parent and we need to pick the second parent among programs p_2 and p_3, which are equally fit (have the same distance to the target semantics). By considering possible offspring $p_1 \times p_3$ and $p_1 \times p_2$, it can be observed that fitness of the geometric offspring is inversely proportional to the angle between residual vectors of its parents.

4. EXPERIMENTAL SETUP

The main goal of the experiments is to investigate whether and how promoting particular forms of diversity affect the fitness of programs evolved with tree-based GP. For this purpose, we analyze the performance of multiobjective diversity promoting methods described in Section 3 and compare them to the standard GP driven by the fitness objective alone. All the considered algorithms were implemented as an extension[1] of the Distributed Evolutionary Algorithms in Python (DEAP) framework [8].

4.1 Symbolic Regression Problems

We consider five univariate symbolic regression problems that are adopted from previous studies [5, 20]. Selected benchmarks (see Table 1) include polynomial, rational and trigonometric functions. For each problem, fitness was calculated as Euclidean distance to the target semantics on 20 training cases distributed equidistantly in the $[-1, 1]$ interval. The only exception is KEIJZER-4, for which the training cases were sampled from the range $[0, 10]$.

4.2 Genetic Programming Variants

We compare the performance of the following five variants of tree-based GP. Four of them rely on multiobjective fitness evaluation where one of the objectives actively promotes some form of diversity. These setups differ only with respect this objective. All the other settings remain unchanged and they are summarized in Table 2.

GP. To observe the relative impact of promoting diversity, as a baseline method we use standard generational tree-based GP with single-objective tournament selection.

[1]The source code necessary for reproducing our results is available at https://github.com/mszubert/gecco_2016.

Table 2: Genetic programming settings

Parameter	Value
population size	256
generations	1000
initialization	ramped half-and-half height range $2 - 6$
instruction set	$\{+, -, \times, /, \exp, \log, \sin, \cos\}$
tournament size	7
crossover probability	0.9
reproduction probability	0.1
mutation probability	0.0
node selection	90% internal nodes 10% leaves
maximum tree height	17
maximum tree size	300
number of runs	100

AFPO. Age-Fitness Pareto Optimization algorithm described in Section 3.1.

DFPO. Density-Fitness Pareto Optimization algorithm (see Section 3.2). To calculate the density objective, genetic markers were constructed using first three levels of each tree.

ESNFPO. Novelty-Fitness Pareto Optimization (see Section 3.3) with Euclidean semantic novelty objective using $k = 15$ nearest neighbors to calculate novelty score.

ASNFPO. Novelty-Fitness Pareto Optimization (see Section 3.3) with angular semantic novelty objective using $k = 15$ nearest neighbors to calculate novelty score.

4.3 Search Operators

To gain deeper understanding about usefulness of diversity under different conditions, we combine each of the considered GP variants with the following search operators.

Standard syntactic crossover. Traditional subtree-swapping crossover operator with Koza-style node selection: 0.9 probability of choosing an internal node [12].

Geometric semantic crossover. Locally geometric semantic crossover (LGX, see Section 2.2) based on a static precomputed library of procedures. The library is generated by enumerating all possible trees of height at most 3, built from the given instruction set. When queried with a desired semantics, library returns a random program among $k = 8$ with closest semantics.

4.4 Diversity Measures

To analyze the relationship between behavioral diversity of a population and fitness of evolved programs, the following diversity measures were calculated for each generation.

Median Euclidean Semantic Distance. To assess Euclidean semantic diversity we calculate median of semantic distances between each pair of programs in the population.

Mean Angular Semantic Distance. Angular semantic diversity is measured as a mean angle between residual vectors of each pair of programs in the population.

5. RESULTS

In order to conduct an accurate analysis of the relationship between diversity and performance, we conducted 100 independent runs (with different random seeds) of each of 10 considered configurations (5 GP variants $\times$ 2 crossover operators) on each of 5 symbolic regression problems.

5.1 Search Performance

Figure 2 shows the average best-of-generation fitness (calculated as a Euclidean distance to the target) achieved by particular methods on different benchmark problems, with 95% confidence intervals marked as semi-transparent bands. The left part of the figure illustrates the results obtained with traditional subtree-swapping crossover. Clearly, each of the considered diversity promoting methods significantly improves the performance of the standard GP algorithm. The best performance is achieved either by DFPO or ASNFPO, depending on the problem. The impact of promoting diversity on the fitness of evolved solutions is much less clear in the case of LGX crossover (right part of Figure 2). The only method that consistently improves the results of the baseline GP algorithm on all considered problems is ASNFPO. All the other diversity preserving approaches are detrimental to the search performance at least on some benchmarks.

Further observations can be made by comparing the results achieved by the same algorithm but equipped with different crossover operators. The largest performance improvement is observed for the standard GP algorithm, which when equipped with the LGX operator, achieves significantly higher convergence speed and final performance. As a matter of fact, it converges so quickly, that if the runs were stopped after 100 generations, it would be the best of the considered setups. Besides GP, the only other method that regularly benefits from replacing traditional syntactic crossover with geometric semantic crossover is ASNFPO. Importantly, the synergistic interplay of LGX crossover and the angular semantic novelty objective leads to the best overall results in terms of the ultimate achieved fitness.

5.2 Diversity Analysis

To analyze the relationship between behavioral diversity and fitness, we assessed diversity of populations evolved by particular methods using measures listed in Section 4.4. Table 3 shows Spearman correlation coefficients calculated between behavioral diversity measured at selected generations and best fitness in the last generation of each run.

In the context of the Euclidean semantic distance measure (left part of Table 3), correlation is stronger for semantic crossover than for standard syntactic crossover. More importantly, at the end of runs correlation is positive — large Euclidean semantic diversity is seen with high (bad) fitness values. This observation is consistent with relatively weak performance achieved by ESNFPO method which uses Euclidean semantic novelty objective to promote behavioral diversity. Taken together, these results suggest that high levels of Euclidean semantic diversity can not be considered as being generally beneficial to the search performance.

Moreover, it can be noticed that at the beginning of runs correlation coefficients are much lower (sometimes even negative) and only later start to increase. Therefore, behavioral diversity may play different role at different evolutionary times. Indeed, further analysis revealed that in the most successful runs, Euclidean semantic diversity stays relatively high in the early, exploratory phase of evolution but then gradually decreases which corresponds to exploitation of the most promising parts of the behavioral space. Thus, high diversity at the beginning of the evolution may be not only less harmful but even advantageous. On the other hand, keeping diversity high throughout entire runs typically leads to inferior performance.

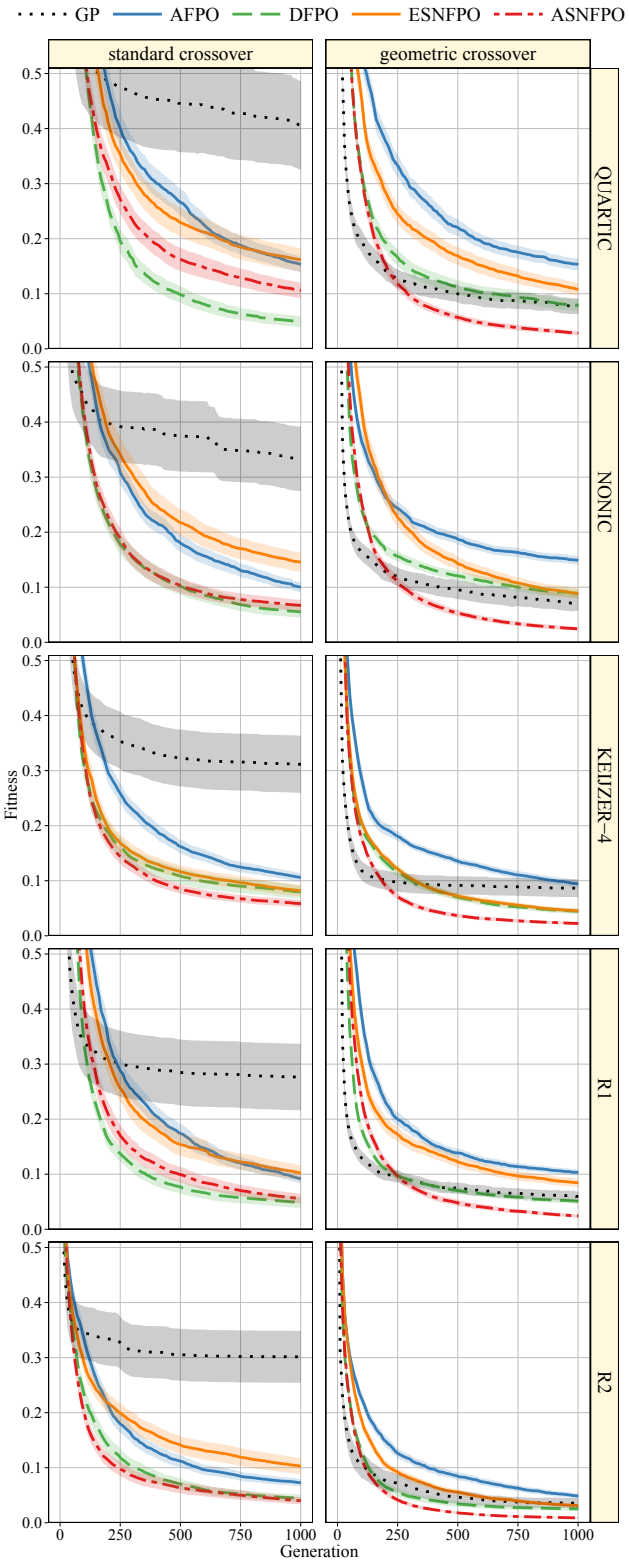

Figure 2: Average best fitness achieved by different variants of multiobjective GP equipped with either standard syntactic crossover (left column) or locally geometric semantic crossover (right column).

Table 3: Correlation between best fitness (lowest error) in the last generation and behavioral diversity measured at selected generations as: 1) median euclidean semantic distance 2) mean angular semantic distance.

	Median Euclidean Semantic Distance										Mean Angular Semantic Distance									
	Standard syntactic crossover					Geometric semantic crossover					Standard syntactic crossover					Geometric semantic crossover				
generation	QUA	NON	KEI	R1	R2	QUA	NON	KEI	R1	R2	QUA	NON	KEI	R1	R2	QUA	NON	KEI	R1	R2
0	−.017	−.034	−.003	−.003	+.031	+.002	+.044	−.012	−.066	+.003	+.010	−.002	+.042	−.010	+.041	−.031	+.061	−.015	−.101	+.001
10	+.010	−.003	−.283	−.227	−.353	−.119	−.203	+.004	−.127	−.131	+.153	+.173	−.427	−.009	−.220	−.464	−.533	−.115	−.499	−.421
25	−.072	−.082	−.250	−.222	−.201	−.109	−.272	+.036	+.026	−.196	−.263	−.187	−.375	−.300	−.547	−.585	−.642	−.324	−.610	−.570
50	+.021	+.071	−.261	−.037	−.078	−.065	−.222	−.227	+.232	−.300	−.392	−.358	−.477	−.448	−.609	−.651	−.675	−.607	−.680	−.624
100	+.018	+.109	−.242	+.015	−.002	+.119	+.027	−.233	+.334	−.080	−.457	−.450	−.600	−.515	−.649	−.654	−.647	−.610	−.677	−.599
250	+.087	+.204	−.121	+.134	+.134	+.393	+.381	+.000	+.502	+.214	−.519	−.441	−.639	−.533	−.650	−.630	−.560	−.599	−.607	−.578
500	+.123	+.265	−.031	+.204	+.180	+.489	+.567	+.177	+.558	+.257	−.587	−.463	−.663	−.478	−.635	−.533	−.456	−.513	−.506	−.530
1000	+.175	+.296	+.020	+.243	+.229	+.630	+.694	+.251	+.567	+.380	−.549	−.411	−.658	−.376	−.607	−.284	−.251	−.266	−.301	−.324

The second form of behavioral diversity we investigate is angular semantic diversity. The right part of Table 3 illustrates relatively strong negative correlation between this diversity measure and final fitness of evolved programs, regardless of the type of employed crossover operator. Since high levels of angular semantic diversity are frequently seen with low (good) fitness, we can hypothesize that this form of diversity facilitates genetic programming. Together with high performance of the ASNFPO method, these results provide empirical evidence that angular semantic diversity tends to be more useful than Euclidean semantic diversity.

5.3 Generalization Performance

In order to assess generalization performance of evolved programs, we calculated the root-mean-square error committed by the best-of-run individuals on 1 000 tests drawn uniformly from the same range as for the training set. Table 4 shows median training error, test error and size (number of nodes) of the individuals evolved by particular methods. To confirm statistically significant differences between the results obtained by the five compared GP variants, for each problem and crossover operator we conducted the Kruskal-Wallis test followed by a post-hoc analysis using pairwise Mann-Whitney tests (with sequential Bonferroni correction). We set the level of significance at $p \leq 0.05$. Table 4 shows with an underline the results that were found significantly better than those achieved by every other GP variant.

On most problems, the significantly lowest test error is obtained by either DFPO or ASNFPO. Interestingly, while DFPO achieves the highest generalization performance in the context of standard crossover, ASNFPO is the winner among methods paired with the LGX operator. These results suggest that there is a synergy between particular variation operators and diversity promoting methods. Traditional syntactic crossover is able to exploit structural diversity maintained by DFPO, whereas semantic crossover benefits from angular semantic diversity. Another important observation is that ASNFPO is the only method that achieves higher generalization performance than standard GP on all problems, regardless of the crossover operator.

Finally, by comparing training and test errors achieved on particular benchmarks, we can observe that AFPO and DFPO methods overfit less than the other methods. One reason explaining less severe overfitting is that these two methods tend to produce shorter programs than the other methods (especially when equipped with the LGX operator). In particular, AFPO usually produces the significantly smallest trees among the considered methods.

6. DISCUSSION

One of the most interesting findings from experiments is the discrepancy between results obtained with different crossover operators (see left vs. right part of Fig. 2 and upper vs. lower part of Table 4). With traditional crossover, all the considered diversity promoting methods improve the performance of standard GP. On the other hand, with geometric crossover, ASNFPO is the only algorithm that consistently outperforms standard GP on all five symbolic regression problems. These findings raise the following questions: Why is angular semantic novelty so effective in the context of geometric crossover? Why are other diversity objectives beneficial with one crossover operator while being detrimental with another? We attempt to answer these questions by referring to the notion of fitness-diversity antagonism.

For the purpose of this discussion, let us say that there is an antagonism between fitness and diversity in a given population if improving fitness of any single individual is impossible without reducing population diversity. Under this definition, angular semantic diversity is never antagonistic with fitness. Indeed, by moving program semantics straight in the direction of the target (along residual vectors), angular semantic diversity does not change while fitness of any solution can be arbitrarily improved. In contrast, Euclidean semantic diversity is at least sometimes antagonistic with fitness — there are populations which can not be optimized without reducing their diversity. Indeed, minimizing error pulls individuals toward the target semantics but maximizing Euclidean diversity scatters them away from it.

Intuitively, one could expect that antagonistic diversity objectives would be detrimental to the search performance. However, this may not be the case in deceptive fitness landscapes, where local fitness gradient is misleading. In such a situation, increasing semantic distance to the target (fitness) can in fact reduce the distance to the target measured in the search space seen by specific operators. We hypothesize that semantically-blind standard crossover induces relatively rugged and deceptive landscape. This hypothesis would to some extent explain why any type of diversity objective, regardless of its antagonism, improves the search performance in the context of this crossover operator.

On the other hand, according to Moraglio et al. [22], geometric semantic operators see cone landscapes which are easy to search by fitness objective alone as they are not deceptive at all. Even though we employ *approximately* geometric crossover, we expect that the corresponding fitness landscape is still much smoother than the one induced by traditional crossover. In such landscapes fitness-diversity

Table 4: Median training error, test error and size of best-of-run individuals. For each problem and crossover operator the best results are shown in bold. Underline indicates statistically significant superiority.

			QUARTIC			NONIC			KEIJZER-4			R1			R2		
			TRAIN	TEST	SIZE	TRAIN	TEST	SIZE	TRAIN	TEST	SIZE	TRAIN	TEST	SIZE	TRAIN	TEST	SIZE
crossover	standard	GP	0.071	0.088	124	0.059	0.062	94	0.057	0.284	229	0.040	0.059	92	0.060	0.064	**75**
		AFPO	0.031	0.034	**72**	0.022	0.027	**84**	0.024	0.111	**121**	0.019	0.019	**69**	0.016	0.016	77
		DFPO	**0.008**	**0.009**	123	**0.009**	**0.013**	138	0.015	**0.089**	161	**0.009**	**0.009**	100	0.008	**0.008**	135
		ESNFPO	0.026	0.050	128	0.029	0.052	125	0.015	0.210	143	0.018	0.030	111	0.016	0.022	87
		ASNFPO	0.019	0.059	136	0.011	0.043	137	**0.012**	0.169	166	**0.009**	0.023	127	**0.007**	0.015	125
	geometric	GP	0.011	0.033	284	0.010	0.029	295	0.012	0.415	300	0.010	0.019	257	0.005	0.005	183
		AFPO	0.033	0.032	**78**	0.034	0.032	**63**	0.020	**0.198**	**144**	0.023	0.021	**66**	0.010	0.009	**61**
		DFPO	0.016	0.016	89	0.019	0.018	86	0.010	0.340	248	0.011	0.011	83	0.005	0.005	80
		ESNFPO	0.023	0.028	185	0.018	0.023	188	0.010	0.508	287	0.017	0.018	149	0.006	0.007	153
		ASNFPO	**0.005**	**0.008**	186	**0.005**	**0.011**	226	**0.005**	0.379	293	**0.005**	**0.007**	202	**0.002**	**0.002**	192

antagonism is much more likely to be detrimental. This would explain weak performance achieved by using antagonistic Euclidean semantic novelty objective. Since angular semantic novelty, by contrast, is the only diversity objective known to be non-antagonistic, it proves successful in the context of the geometric crossover operator.

Finally, let us discuss two other reasons that could explain aforementioned discrepancy in results. First, by analyzing how fitness of perfectly geometric offspring depends on the angular distance between its parents (cf. Fig. 1), we expect that geometric crossover operator is able to effectively exploit angular semantic diversity. Another synergistic combination involves structural diversity (promoted by the DFPO algorithm) and traditional syntactic crossover operator. Both combinations of diversity objective and search operator result in superior performance when compared to other considered methods. Second, the reason why diversity maintenance plays such an important (and beneficial) role in GP equipped with traditional crossover is that in our experiments we do not employ any mutation operator which could supply new genetic material and explicitly sustain genetic diversity in a population. In absence of mutation, we expect that standard GP with subtree-swapping crossover is particularly vulnerable to the problem of premature convergence. This problem is less severe with locally geometric crossover because it relies on a large library of procedures which provides the population with new subtrees acting as a simple diversity preserving mechanism.

7. CONCLUSIONS

In recent years, the issue of behavioral diversity and its impact on the performance of evolutionary algorithms has been studied in many different contexts [17, 23, 29]. To the best of our knowledge, this is the first study that investigates the role of behavioral diversity in genetic programming equipped with semantic search operators. The main goal of this work was to determine whether and under what conditions promoting behavioral diversity can adversely affect the performance of GP applied to symbolic regression.

The most important finding is that using an additional diversity promoting objective can be indeed detrimental to the search performance. However, such a situation was observed only when both of the following conditions were met. First, a specific search operator was employed, which supposedly induced a smooth, non-deceptive fitness landscape. Second, the behavioral diversity objective was inherently antagonistic with the fitness objective. On the other hand, by introducing a non-antagonistic angular semantic novelty objective, we were able to improve the results regardless of the employed search operator. Importantly, this objective was the only one that proved successful in the context of locally geometric crossover operator.

The major limitation of this study is that our experimental investigations were conducted using a small set of five univariate symbolic regression benchmarks. Although promoting angular semantic diversity proved useful in this context, further work is needed to verify whether these results could be extended to more complex real-world problems. In particular, it would be interesting to analyze how much dimensionality of both feature space and semantic space impacts the performance of particular diversity promoting methods. Another direction of future research would be to investigate the importance of behavioral diversity for other semantic search operators.

In a broader perspective, our investigation indicates that a diversity objective needs to be carefully chosen with respect to the problem at hand and employed search algorithm. As demonstrated by this study, using objectives that are antagonistic with fitness was detrimental to the performance of semantic GP. However, we expect that with increasing deceptiveness in the fitness landscape, the consequences of using antagonistic objectives become more difficult to predict. In particular, one could hypothesize that in highly deceptive fitness landscapes antagonistic diversity objectives are more likely to be beneficial. This would be consistent with previous studies demonstrating that in extremely deceptive cases a successful way to increase search effectiveness is to ignore fitness and use a novelty objective alone [16].

Acknowledgments

This work was supported by the National Aeronautics and Space Administration under grant number NNX15AH48G.

8. REFERENCES

[1] L. Beadle and C. G. Johnson. Semantically Driven Crossover in Genetic Programming. In *Proceedings of the IEEE Congress on Evolutionary Computation, CEC 2008*, pages 111–116. IEEE, 2008.

[2] L. Beadle and C. G. Johnson. Semantic Analysis of Program Initialisation in Genetic Programming. *Genetic Programming and Evolvable Machines*, 10(3):307–337, 2009.

[3] J. C. Bongard. A Probabilistic Functional Crossover Operator for Genetic Programming. In *Proceedings of the Genetic and Evolutionary Computation Conference*, pages 925–932. ACM, 2010.

[4] E. K. Burke, S. M. Gustafson, and G. Kendall. Diversity in Genetic Programming: An Analysis of Measures and Correlation with Fitness. *IEEE Trans. Evolutionary Computation*, 8(1):47–62, 2004.

[5] A. R. Burks and W. F. Punch. An Efficient Structural Diversity Technique for Genetic Programming. In *Proceedings of the Genetic and Evolutionary Computation Conference*, pages 991–998. ACM, 2015.

[6] E. D. de Jong, R. A. Watson, and J. B. Pollack. Reducing Bloat and Promoting Diversity using Multi-Objective Methods. In *Proceedings of the Genetic and Evolutionary Computation Conference*, pages 11–18. Morgan Kaufmann, 2001.

[7] R. A. Fisher. *The Genetical Theory of Natural Selection*. Oxford University Press, Oxford, 1930.

[8] F.-A. Fortin, F.-M. De Rainville, M.-A. G. Gardner, M. Parizeau, and C. Gagné. DEAP: Evolutionary Algorithms Made Easy. *Journal of Machine Learning Research*, 13(1):2171–2175, 2012.

[9] E. Galvan-Lopez, B. Cody-Kenny, L. Trujillo, and A. Kattan. Using Semantics in the Selection Mechanism in Genetic Programming: A Simple Method for Promoting Semantic Diversity. In *2013 IEEE Congress on Evolutionary Computation (CEC)*, pages 2972–2979, 2013.

[10] G. S. Hornby. ALPS: The Age-layered Population Structure for Reducing the Problem of Premature Convergence. In *Proceedings of the Genetic and Evolutionary Computation Conference*, pages 815–822. ACM, 2006.

[11] D. Jackson. Promoting Phenotypic Diversity in Genetic Programming. In *Parallel Problem Solving from Nature, PPSN XI*, volume 6239 of *Lecture Notes in Computer Science*, pages 472–481. Springer, 2010.

[12] J. R. Koza. *Genetic Programming: On the Programming of Computers by Means of Natural Selection*. MIT Press, Cambridge, MA, USA, 1992.

[13] K. Krawiec. *Behavioral Program Synthesis with Genetic Programming*, volume 618 of *Studies in Computational Intelligence*. Springer, 2016.

[14] K. Krawiec and P. Lichocki. Approximating Geometric Crossover in Semantic Space. In *Proceedings of the Genetic and Evolutionary Computation Conference*, pages 987–994. ACM, 2009.

[15] K. Krawiec and T. Pawlak. Locally Geometric Semantic Crossover: A Study on the Roles of Semantics and Homology in Recombination Operators. *Genetic Programming and Evolvable Machines*, 14(1):31–63, 2013.

[16] J. Lehman and K. O. Stanley. Abandoning Objectives: Evolution through the Search for Novelty Alone. *Evolutionary Computation*, 19(2):189–223, 2011.

[17] J. Lehman, K. O. Stanley, and R. Miikkulainen. Effective Diversity Maintenance in Deceptive Domains. In *Proceedings of the Genetic and Evolutionary Computation Conference*, GECCO '13, pages 215–222. ACM, 2013.

[18] P. Liskowski, K. Krawiec, T. Helmuth, and L. Spector. Comparison of Semantic-aware Selection Methods in Genetic Programming. In *Proceedings of th Genetic and Evolutionary Computation Conference*, pages 1301–1307, New York, NY, USA, 2015. ACM.

[19] Y. Martínez, E. Naredo, L. Trujillo, and E. G. López. Searching for Novel Regression Functions. In *Proceedings of the IEEE Congress on Evolutionary Computation*, pages 16–23. IEEE, 2013.

[20] J. McDermott, D. R. White, S. Luke, L. Manzoni, M. Castelli, L. Vanneschi, W. Jaskowski, K. Krawiec, R. Harper, K. De Jong, and U.-M. O'Reilly. Genetic Programming Needs Better Benchmarks. In *Proceedings of the Genetic and Evolutionary Computation Conference*, pages 791–798. ACM, 2012.

[21] N. F. McPhee and N. J. Hopper. Analysis of genetic diversity through population history. In *Proceedings of the Genetic and Evolutionary Computation Conference*, volume 2, pages 1112–1120, 1999.

[22] A. Moraglio, K. Krawiec, and C. G. Johnson. Geometric Semantic Genetic Programming. In *Parallel Problem Solving from Nature - PPSN XII*, volume 7491 of *Lecture Notes in Computer Science*, pages 21–31. Springer Berlin Heidelberg, 2012.

[23] J. B. Mouret and S. Doncieux. Encouraging Behavioral Diversity in Evolutionary Robotics: An Empirical Study. *Evolutionary Computation*, 20(1):91–133, 2012.

[24] Q. U. Nguyen, X. H. Nguyen, M. O'Neill, and A. Agapitos. An Investigation of Fitness Sharing with Semantic and Syntactic Distance Metrics. In *Proceedings of the 15th European Conference on Genetic Programming*, EuroGP'12, pages 109–120, Berlin, Heidelberg, 2012. Springer-Verlag.

[25] S. Ruberto, L. Vanneschi, M. Castelli, and S. Silva. *Genetic Programming: 17th European Conference, EuroGP 2014*, chapter ESAGP – A Semantic GP Framework Based on Alignment in the Error Space, pages 150–161. Springer Berlin Heidelberg, 2014.

[26] B. Sareni and L. Krahenbuhl. Fitness Sharing and Niching Methods Revisited. *IEEE Transactions Evolutionary Computation*, 2(3):97–106, 1998.

[27] M. Schmidt and H. Lipson. Age-Fitness Pareto Optimization. In *Genetic Programming Theory and Practice VIII*, volume 8 of *Genetic and Evolutionary Computation*, pages 129–146. Springer, 2011.

[28] G. Smits and M. Kotanchek. Pareto-Front Exploitation in Symbolic Regression. In *Genetic Programming Theory and Practice II*, chapter 17, pages 283–299. Springer, Ann Arbor, 2004.

[29] M. Szubert, W. Jaśkowski, P. Liskowski, and K. Krawiec. The Role of Behavioral Diversity and Difficulty of Opponents in Coevolving Game-Playing Agents. In *18th European Conference on Applications of Evolutionary Computation*, pages 394–405, 2015.

[30] N. Q. Uy, N. X. Hoai, M. O'Neill, R. I. Mckay, and E. Galván-López. Semantically-based Crossover in Genetic Programming: Application to Real-valued Symbolic Regression. *Genetic Programming and Evolvable Machines*, 12(2):91–119, 2011.

[31] L. Vanneschi, M. Castelli, and S. Silva. A Survey of Semantic Methods in Genetic Programming. *Genetic Programming and Evolvable Machines*, 15(2):195–214, 2014.