Supply of Order-1 Building Blocks for Functions Linearly Combined of Sinusoidal Bases with Integral Frequencies

Hongqiang Mo[1], Zhong Li[2] and Qiliang Du[1]

[1]*School of Automation Science and Engineering, South China University of Technology, Guangzhou, 510641, P.R. China*

[2]*Faculty of Mathematics and Computer Science, FernUniversitaet in Hagen, Hagen, 58084, Germany*

Keywords: Building Block, Encoding, Genetic Algorithm, Schema Processing, Series Expansion.

Abstract: In line with the theory of schema sampling, a hypothesis could be made that sufficient supply of low-order building blocks (BBs) was one of the necessary conditions for a genetic algorithm(GA) to work. A consequential question of this hypothesis regards, when a certain fitness function is optimized with a commonly used GA, whether it is rare or common that there are plenty of low-order BBs. It is remarked that, when a base-*m* encoded GA is applied to a fitness function that is linearly combined of sinusoidal basis functions with integral frequencies, it is unlikely to obtain order-1 BBs with fixed positions at multiple loci, i.e., it is rare that there are plenty of order-1 BBs. However, if a considerable part of the sinusoidal basis functions are with frequencies exponential to a positive integer *m*, a base-*m* encoding can provide relatively more order-1 BBs compared with the encodings with cardinalities other than *m*.

1 INTRODUCTION

Genetic algorithms (GAs) have been one of the most basic forms of evolutionary algorithms since they were proposed decades ago. Although they have found wide applications in search, optimization, design and machine learning, the features of problems that determine the likelihood of successful GA performance are not fully understood yet.

One of the typical explanations for the mechanisms of genetic search is based on the theory of schema processing, in which the feasible solutions of a problem are represented with strings, and the searching for the optimal or sub-optimal strings are believed to be implicitly implemented by recombining highly fit, low-order schemata (Goldberg, 1989)(Rothlauf, 2006). A schema is a template that identifies a subset of strings with similarities at certain string positions, and a single string belongs to all the schemata in which any of its fixed positions appear. For example, the strings 1011 and 1001 are members of schemata 10** (where the *s stand for unspecified positions), 1**1, *0**, and so forth. The order of a schema refers to the number of its fixed positions. For example, 10** and 1**1 are both order-2 schemata, and the order of *0** is 1. In line with the theory of schema sampling, a schema can be regarded as a particular region in the solution space, and the schemata containing many unspecified positions — the low-order schemata — will typically be sampled by a large fraction of all the strings in a population of a GA. And by manipulating a limited population of strings, a GA actually samples a vastly larger number of regions (Goldberg, 1989)(Rothlauf, 2006)(Holland, 1975). As stated by the schema theorem, successive generations of reproduction produce increasing numbers of trials that lie in the regions represented by highly fit, low-order schemata (Goldberg, 1989)(Holland, 1975). And it is assumed that, when these highly fit, low-order schemata recombine to form even more highly fit, higher-order schemata, a GA rapidly focuses its attention on the most promising parts of the solution space; in this sense, the highly fit, low-order schemata are also called building blocks (BBs) (Goldberg, 1989)(Goldberg, 2002).

If a GA indeed functions in this way, a hypothesis can be made that sufficient supply of low-order BBs is one of the necessary conditions for it to work (Goldberg, 1989)(Rothlauf, 2006)(Holland, 1975). And a consequential question of this hypothesis regards, when a certain fitness function is optimized with a commonly used GA, whether it is rare or common that there are plenty of low-order BBs.

In this paper, we will make an attempt to study this issue by analyzing the cases of fitness functions

Mo, H., Li, Z. and Du, Q..
Supply of Order-1 Building Blocks for Functions Linearly Combined of Sinusoidal Bases with Integral Frequencies.
In *Proceedings of the 7th International Joint Conference on Computational Intelligence (IJCCI 2015) - Volume 1: ECTA*, pages 133-137
ISBN: 978-989-758-157-1

linearly combined of sinusoidal basis functions with integral frequencies, and by finding out whether there are always plenty of order-1 BBs when such a fitness function is optimized with a commonly used GA. The choice of such fitness functions is mainly inspired by Fourier transformation. Linear expansion of a function into sine functions conforms to the common practices in functional analysis. By doing so, we hope that the future analysis of encoding design and GA hardness can get more supports from functional theory.

The discussions will be limited to the commonly used encodings. In the applications of GAs to continuous optimization problems, the feasible solutions are usually expressed with binary encodings. In this paper, we will extend them to base-m encodings — a more general form of representations — where m is an integer larger than 1. When a solution x, also called string or individual, is represented with the base-m encoding of string length l, $x = \sum_{h=1}^{l} x_h m^{h-l-1}$, where $x_h \in \{0,\cdots,m-1\}$. The commonly used binary representations are base-2 encodings. And for the base-3 encoding of string length 12, $x=\sum_{h=1}^{12} x_h 3^{h-13}$, and $x_h \in \{0,1,2\}$.

There have been models to estimate the population size required to guarantee the presence of all raw BBs in a GA (Goldberg et al., 2001) population or a genetic-programming (Sastry et al., 2003) population, but the models only established necessary population size for building-block supply, and did not tell about whether or not there were indeed low-order BBs. The previous work of this paper has investigated the supply of order-1 BBs for fitness functions that were linearly combined of sinusoidal basis functions with frequencies exponential to a positive integer (Mo et al., 2009)(Mo et al., 2015). In this paper, we will extend the discussions to sinusoidal basis functions with integral frequencies, and will focus on the existence of order-1 BBs.

The rest of this paper is organized as follows: Section 2 introduces an index to the existence of an order-1 BB at a certain locus. Section 3 explains why it is unlikely to generate order-1 BBs simultaneously at multiple loci when a base-m encoding is used to express the fitness functions linearly combined of sinusoidal basis functions with arbitrary integral frequencies, and then proposes an encoding suggestion for the cases that the frequencies of a considerable part of the sinusoidal basis functions are exponential to a positive integer m. Finally, Section 4 summarizes the paper.

2 INDEX TO THE EXISTENCE OF ORDER-1 BUILDING BLOCK

Specially, the fitness functions discussed herein are $G(x) = \sum_{i=1}^{n_B} a_i \sin(2\pi p_i x + \varphi_i) + c$, where n_B and p_i are positive integers, a_i, φ_i, and c are real numbers, $a_i \geq 0$, $\varphi_i \in [0,\ 2\pi)$, and c is large enough to ensure $G(x) \geq 0$. When x is unlimited, $G(x)$ is periodical, and there is at least one complete cycle of $G(x)$ within $[0,\ 1)$. Therefore, without loss of generality, the discussions are restricted within $x \in [0,\ 1)$. And for the sake of convenience, we do not distinguish between the fitness of string $x_l \cdots x_1$, which is denoted as $G(x_l \cdots x_1)$, and that of its decoded value, x, i.e., $G(x_l \cdots x_1) = G(x)$.

A schema is said to match an individual if they are identical at the fixed positions of the former. The fitness of a schema can be defined as the average fitness of all the individuals matched by the schema in a certain population or in the whole search space. Let's take the base-3 encoding of length 2 as an example: Under the former definition, given a population consisting of $01, 01, 22, 10$, the fitness of *1 is equal to $(G(01)+G(01))/2$; Under the latter definition, the fitness of schema *1 is equal to $(G(01)+G(11)+G(21))/3$ regardless of population members. With the former definition, schema fitness is dynamic during evolution, and its value depends not only on encoding, but also on the formation of initial population, genetic operators, and selection strategy. The latter, usually used to determine the static fitness distributions of a schema (Goldberg, 1989)(Whitley et al., 2003), is especially suitable to study the sole effect of encoding on schema fitness. Therefore, here and throughout, the latter definition is adopted.

In this paper, the fitness are compared among the order-1 schemata with the same fixed positions. For convenience, if the fixed position of an order-1 schema is at the h-th position of the string counted from the rightmost, we call it an order-1 schema at locus h, where h is a positive integer no larger than string length l. For example, 1***, **0* are order-1 schemata at the 4th and 2nd loci, respectively.

When a $G(x)$ is expressed with a base-m encoding of string length l, where l is much larger than n_B, the fitness of order-1 schema $* \cdots * x_h * \cdots *$ at locus h, denoted as $f_G(x_h)$, is

$$f_G(x_h) = \frac{\sum_{k=0}^{m^{l-h}-1} \sum_{o=0}^{m^{h-1}-1} G((km+x_h)m^{h-l-1} + om^{-l})}{m^{l-1}}, \tag{1}$$

where the symbols f and x_h stand for average fitness and position value (allele value), respectively, and the subscript G indicates the name of the fitness function. The maximal fitness of the order-1 schemata at locus h is denoted as $\max_h(f_G(x_h))$.

The average fitness of all the order-1 schemata at locus h is

$$f_G(*) = \frac{\sum_{x_h=0}^{m-1} f_G(x_h)}{m}. \tag{2}$$

In line with the theory of schema processing, a GA implicitly attempts to allocate trials to different regions of the search space based on schema fitness. If $f_G(x_h) > f_G(*)$ at locus h, which means that $*\cdots * x_h * \cdots *$ is fitter than the average fitness of all the order-1 schemata at locus h, the schema will have a chance higher than average to reproduce more samples in the subsequent generation. Consulting the definition given in (Goldberg, 1989), we have **Definition 1** for order-1 BB.

Definition 1. Order-1 BB at locus *h*: An order-1 schema at locus h that fitter than the average fitness of all the order-1 schemata at this locus.

Therefore, $f_G(x_h) > f_G(*)$ indicates that $*\cdots * x_h * \cdots *$ is an order-1 BB at locus h.

3 SUPPLY OF ORDER-1 BBS FOR $G(X)$ EXPRESSED WITH A BASE-*M* ENCODING

To simplify expressions, let's denote $\sin(2\pi p_i x + \varphi_i)$ as $B_i(x)$, and m^{h-l-1} as Δ, respectively. By definition (1) and (2), the values of $f_G(x_h)$ and $f_G(*)$ are equal to the weighted sums of $f_{B_i}(x_h)$ and $f_{B_i}(*)$, respectively, as shown in (3) and (4), respectively.

$$f_G(x_h) = \sum_{i=1}^{n_B} a_i f_{B_i}(x_h) + c, \tag{3}$$

and

$$f_G(*) = \sum_{i=1}^{n_B} a_i f_{B_i}(*) + c. \tag{4}$$

As illustrated in Fig.1, when $m^{-l} \to 0$,

$$\sum_{o=0}^{m^{h-1}-1} \frac{B_i((km+x_h)\Delta + om^{-l})}{m^l} \approx \int_{(km+x_h)\Delta}^{(km+x_h+1)\Delta} B_i(x)dx, \tag{5}$$

where $h \in \{1, \cdots, l\}$. Applying sum-to-product trigonometric formulas, one has

$$\int_{(mk+x_h)\Delta}^{(mk+x_h+1)\Delta} B_i(x)dx = \frac{\sin(\pi p_i \Delta)\sin(2\pi p_i \Delta(mk+x_h) + \pi p_i \Delta + \varphi_i)}{\pi p_i}. \tag{6}$$

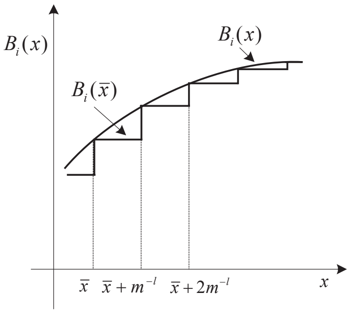

Figure 1: The illustration of an approximation of $\int_{\bar{x}}^{\bar{x}+m^{-l}} B_i(x)dx$ with $B_i(\bar{x})$, where $\bar{x} = (km+x_h)\Delta + om^{-l}$. When $m^{-l} \to 0$, $\int_{\bar{x}}^{\bar{x}+m^{-l}} B_i(x)dx \to m^{-l}B_i(\bar{x})$.

According to (1), (5), and (6),

$$f_{B_i}(x_h) \approx \sum_{k=0}^{m^{l-h}-1} \frac{m\sin(\pi p_i \Delta)\sin(\pi p_i \Delta(2mk + 2x_h + 1) + \varphi_i)}{\pi p_i}. \tag{7}$$

Further denote $m\sin(\pi p_i \Delta)/(\pi p_i)$ as A_i and $\pi p_i \Delta(2x_h+1) + \varphi_i$ as $\bar{\varphi}_i$, respectively, we have

$$f_{B_i}(x_h) \approx \sum_{k=0}^{m^{l-h}-1} A_i \sin(2kp_i\pi m^{h-l} + \bar{\varphi}_i). \tag{8}$$

When $p_i \neq m^{l-h}$ and $h < l$, the amplitudes of the m^{l-h} sinusoidal functions in (8), i.e., $A_i\sin(2kp_i\pi m^{h-l} + \bar{\varphi}_i)$, $k \in \{0, \cdots, m^{l-h}-1\}$, are the same, and the phase difference between each adjacent pair of the functions is $2p_i\pi m^{h-l}$. Therefore, their sum is equal to 0, and $f_{B_i}(x_h) \approx 0$.

By a straightforward derivation of (8), we have

$$f_{B_i}(x_h) \approx \begin{cases} \frac{m\sin\frac{\pi p_i}{m}\sin(\frac{\pi p_i(2x_h+1)}{m}+\varphi_i)}{\pi p_i}, & \text{if } h = l; \\ \frac{m\sin\frac{\pi}{m}\sin(\frac{\pi(2x_h+1)}{m}+\varphi_i)}{\pi}, & \text{if } p_i = m^{l-h}; \\ 0, & \text{if } p_i \neq m^{l-h} \text{ and } h < l. \end{cases} \tag{9}$$

Once again, when $h = l$, the amplitudes of the m sinusoidal functions in (9) are the same, and the phase difference between each adjacent pair of the functions is $2p_i\pi/m$. As a result, their sum is equal to 0. The same thing happens when $p_i = m^{l-h}$. Thus, in line with (2) and (9), we have

$$f_{B_i}(*) \approx 0, \tag{10}$$

regardless of the values of h, p_i, and m. Substituting (10) into (4), we have

$$f_G(*) \approx c. \tag{11}$$

- **Remark 1.** If p_i is randomly assigned with a positive integer, the probability that the value happens to be m^{l-h} is 0. Therefore, according to (9), for most of the sinusoidal basis functions $B_i(x) = \sin(2\pi p_i x + \varphi_i)$, $f_{B_i}(x_h) \approx 0$ at locus $h < l$. Substituting $f_{B_i}(x_h) \approx 0$ into (3), and according to (11), we have $f_G(x_h) \approx c \approx f_G(*)$. Thus, the fitness of all the order-1 schemata at loci $h < l$ are almost the same regardless of the allele value of x_h, and no fitness differences among the order-1 schemata can be achieved at these loci. In a word, it is unlikely to obtain order-1 BBs at loci $h < l$ when a base-*m* encoding is used to express $G(x) = \sum_{i=1}^{n_B} a_i \sin(2\pi p_i x + \varphi_i) + c$.

  It should be noted that the above-mentioned conclusion is independent of the value of the encoding base. It applies to all base-*m* encodings, including the most commonly used binary ones. And for such kind of fitness functions, no choice of encoding cardinality offers intrinsic advantage over another on the supply of order-1 BBs.

- **Remark 2.** According to (9), when a considerable part of the sinusoidal basis functions of a fitness function $G(x)$ are with frequencies exponential to m, expressing the fitness function with a base-*m* encoding can result in order-1 BBs simultaneously at multiple loci. Substituting $p_i = m^{l-h_i}$ into (9), we have $f_{B_i}(x_{h_i}) = m \sin(\pi/m) \sin(\pi(2x_{h_i} + 1)/m + \varphi_i)/\pi$. When x_{h_i} increases from 0 to $m-1$, there will be at least one locus setting, $x_{h_i} = x^*_{h_i}$, that satisfies $f_{B_i}(x^*_{h_i}) > 0$. In line with (3) and (4), $f_G(x^*_{h_i}) > f_G(*)$. In other words, $* \cdots * x^*_{h_i} * \cdots *$ is an order-1 BB at this locus. Therefore, for the fitness functions defined as in (12), if a_i is significantly larger than 0, and if a base-*m* encoding is used, there will be an order-1 BB at locus h_i, where $i \in \{1, \cdots, n_1\}$.

$$G(x) = \sum_{i=1}^{n_1} a_i \sin(2\pi m^{l-h_i} x + \varphi_i) + \sum_{j=1}^{n_2} a_{n_1+j} \sin(2\pi p_j x + \varphi_{n_1+j}) + c, \quad (12)$$

  where $n_1 + n_2 = n_B$.

  According to the schema theory, more than average samples would be allocated to the regions represented by the order-1 BBs during evolution, and the search would soon be guided to the promising regions. Therefore, it can be expected that, for such kind of fitness functions, the GAs with the right encoding cardinality outperform those with other cardinalities. Thus, it is suggested to adopt a base-*m* encoding to express the kind of fitness functions defined in (12).

4 FURTHER DISCUSSIONS AND CONCLUSIONS

This paper discussed the supply of order-1 BBs for the fitness functions that were linearly combined of sinusoidal basis functions with integral frequencies, $\sum_{i=1}^{n_B} a_i \sin(2\pi p_i x + \varphi_i) + c$. It was remarked that, if the positive integers p_i's were randomly chosen, and if a base-*m* encoded GA was used, it was unlikely to obtain order-1 BBs at loci $h < l$, no matter what the value of *m* was. Therefore, for this kind of fitness functions, no cardinality of encoding could exhibit advantage over other choices on the supply of order-1 BBs. The results to some degree supported the known facts that no representation should be superior for all classes of problems (Fogel and Ghozeil, 1997)(Wolpert and Macready, 1997)(Whitley, 1999).

However, things would change if one was restricted to consider a special subclass of these fitness functions, in which the frequencies of a considerable part of the sinusoid basis functions were exponential to a positive integer *m*. It was proved that, for a fitness function in this subclass, a base-*m* encoding could provide relatively more order-1 BBs compared with the encodings with cardinalities other than *m*.

It should be noted that the discussions in this paper have been focused on order-1 BBs. That an encoding can not achieve fitness differences among order-1 schemata does not imply that it can not gain them among higher-order ones. However, it is necessary to employ a population of relatively large size if one tries to gain sufficient supply of high-order BBs. And also, it iss not a trivial matter to dig them out. Actually, plenty of skillful memetics algorithms (Goldberg, 2002)(Goldberg et al., 2003)(Chen and Lim., 2009)(Krasnogor and Smith, 2005) (Chen et al., 2011) have been established to facilitate the formation of deep BBs and discover them.

It should also be noted that the idea that GAs search by schema sampling has received many different criticisms: Increasing the sampling rate of schemata that are above average compared to other competing schemata does not guarantee convergence to a global optimum, since the search may be misled to wrong directions due to deception (Goldberg, 1992)(Deb and Goldberg, 1994)(Deb et al., 1993), hyper-plane inconsistency (Whitley et al., 1995)(Whitley et al., 2003), synchronization (Hoyweghen et al., 2001), sampling errors (Goldberg, 1989)(Forrest and Mitchell, 1993), etc. It is accepted that the notion of using schema information to guide search at best be viewed as a heuristic (Whitley, 2001).

We will expand the analyses to higher-order BBs, and seek for more explanations for genetic behaviors in future study. Meanwhile, we will also explore the possibility of extending the analysis to Gray-coded and real-coded GAs in the future.

ACKNOWLEDGEMENTS

This work was supported by “National Natural Science Foundation of China, 61105062 and 61305038”, “Fundamental Research Funds for the Central Universities, SCUT, 2012ZZ0106 and 2014ZZ0045”.

REFERENCES

Chen, X., Ong, Y.-S., Lim, M.-H., and Tan, K. C. (2011). A multi-facet survey on memetic computation. *IEEE Transactions on Evolutionary Computation*, 15(5):591–607.

Chen, Y. and Lim., M. (2009). *Linkage in Evolutionary Computation (Studies in Computational Intelligence)*. Springer Verlag.

Deb, K. and Goldberg, D. E. (1994). Sufficient conditions for deceptive and easy binary functions. *Annals of Mathematics and Artificial Intelligence*, 10(4):385–408.

Deb, K., Horn, J., and Goldberg, D. (1993). Multimodal deceptive functions. *Complex Systems*, 7(2):131–153.

Fogel, D. B. and Ghozeil, A. (1997). A note on representations and variation operators. *IEEE Transactions on Evolutionary Computation*, 1(2):159–161.

Forrest, S. and Mitchell, M. (1993). What makes a problem hard for a genetic algorithm? some anomalous results and their explanation. *Machine Learning*, 13(2-3):285–319.

Goldberg, D. E. (1989). *Genetic Algorithms in Search, Optimization and Machine Learning*. Addison-Wesley.

Goldberg, D. E. (1992). Construction of high-order deceptive functions using low-order walsh coefficients. *Annals of Mathematics and Artificial Intelligence*, 5(1):35–48.

Goldberg, D. E. (2002). *Design of Innovation: Lessons from and for Competent Genetic Algorithms*. Kluwer Academic Publishers.

Goldberg, D. E., Sastry, K., and Latoza, T. (2001). On the supply of building blocks. In Spector, L., editor, *Proc. of the 2001 Genetic and Evolutionary Computation Conference*, pages 336–342. San Francisco, Calif., Kaufmann.

Goldberg, D. E., Sastry, K., and Ohsawa, Y. (2003). Discovering deep building blocks for competent genetic algorithms using chance discovery via keygraphs. In Ohsawa, Y. and Mcburney, P., editors, *Chance Discovery*, pages 276–301. Springer-Verlag.

Holland, J. H. (1975). *Adaptation in natural and artificial systems: An introductory analysis with applications to biology, control, and artificial intelligence.* U Michigan Press.

Hoyweghen, C. V., Goldberg, D. E., and Naudts, B. (2001). Building block superiority, multimodality and synchronization problems. In *Proceedings of the Genetic and Evolutionary Computation Conference (GECCO-2001)*, pages 694–701. Morgan Kaufmann.

Krasnogor, N. and Smith, J. (2005). A tutorial for competent memetic algorithms: Model, taxonomy, and design issues. *IEEE Transactions on Evolutionary Computation*, 9(5):474–488.

Mo, H. Q., Li, Z., Park, J. B., and Joo, Y. H. (2009). On the supply of superior order-1 building blocks for a class of periodical fitness functions. *International Journal of Computational Intelligence Systems*, 2(1):91–98.

Mo, H. Q., Li, Z., Tian, L. F., and Tian, X. (2015). Selection of encoding cardinality for a class of fitness functions to obtain order-1 building blocks. *International Journal of Computational Intelligence Systems*, 8(1):62–74.

Rothlauf, F. (2006). *Representations for genetic and evolutionary algorithms*. Springer.

Sastry, K., O’Reilly, U.-M., Goldberg, D. E., and Hill, D. (2003). Building block supply in genetic programming. In Riolo, R. L. and Worzel, B., editors, *Genetic Programming Theory and Practice*, chapter 9, pages 137–154. Kluwer.

Whitley, D. (1999). A free lunch proof for gray versus binary encodings. In *Proceedings of the Genetic and Evolutionary Computation Conference*, volume 1, pages 726–733. Citeseer.

Whitley, D. (2001). An overview of evolutionary algorithms: Practical issues and common pitfalls. *Information and software technology*, 43(14):817–831.

Whitley, D., Heckendorn, R. B., and Stevens, S. (2003). Hyperplane ranking, nonlinearity and the simple genetic algorithm. *Information Sciences*, 156(3):123–145.

Whitley, D., Mathias, K. E., and Pyeatt, L. D. (1995). Hyperplane ranking in simple genetic algorithms. In Eshelman, L., editor, *Proceedings of the 6th International Conference on Genetic Algorithms*, pages 231–238.

Wolpert, D. H. and Macready, W. G. (1997). No free lunch theorems for optimization. *IEEE Transactions on Evolutionary Computation*, 1(1):67–82.