

## Research Article

# Epistasis-Based Basis Estimation Method for Simplifying the Problem Space of an Evolutionary Search in Binary Representation

Junghwan Lee  and Yong-Hyuk Kim 

*Department of Computer Science, Kwangwoon University, 20 Kwangwoon-ro, Nowon-gu, Seoul 01897, Republic of Korea*

Correspondence should be addressed to Yong-Hyuk Kim; yhdflly@kw.ac.kr

Received 13 December 2018; Accepted 8 May 2019; Published 28 May 2019

Academic Editor: Thach Ngoc Dinh

Copyright © 2019 Junghwan Lee and Yong-Hyuk Kim. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

An evolutionary search space can be smoothly transformed via a suitable change of basis; however, it can be difficult to determine an appropriate basis. In this paper, a method is proposed to select an optimum basis can be used to simplify an evolutionary search space in a binary encoding scheme. The basis search method is based on a genetic algorithm and the fitness evaluation is based on the epistasis, which is an indicator of the complexity of a genetic algorithm. Two tests were conducted to validate the proposed method when applied to two different evolutionary search problems. The first searched for an appropriate basis to apply, while the second searched for a solution to the test problem. The results obtained after the identified basis had been applied were compared to those with the original basis, and it was found that the proposed method provided superior results.

## 1. Introduction

Binary encoding typically uses a standard basis, and when a nonstandard basis is used, the structure of the problem space may become quite different from that of the original problem. In an evolutionary search, various methods can be used to change a problem space by adjusting the basis, including gene rearrangement, different encoding methods, and the use of an eigen-structure [1–12].

An investigation was conducted to elucidate the possibility of changing the basis in binary encoding and the corresponding effects on the genetic algorithm (GA) [13]; however, it was not possible to determine which basis should be applied to smooth the problem search space. In genetics, epistasis means that the phenotypic effect of one gene is masked by another gene; however, in GA, it refers to any type of gene interaction. In a problem with a large epistasis, as the genes are extremely inter-dependent, the fitness landscape of the problem space is very complex and the problem is difficult [14]. Several studies have been conducted to assess the difficulty of problems from the perspective of epistasis [15–21]. Epistasis has the advantage that it is possible to measure

the extent of nonlinearity only with fitness function. In this paper, we define the difficulty of the problem or problem search space as the nonlinearity level of gene expression. Also, we use epistasis as a measure for the difficulty of the problem.

There are three main contributions of this paper. First, an epistasis approximation is used to identify a basis that will reduce the complexity of an evolutionary search problem. Second, the basis is expressed by a variable-length encoding scheme using an elementary matrix. Finally, a GA is defined that can be used to change the basis of an evolutionary search space. This means that when a basis is given, one can tell how it affects the GA. Our intention in this study is that a nonseparable problem can be transformed into a separable problem by performing an appropriate basis transformation. Such an altered environment enables GA to search space effectively.

This paper is organized as follows: Section 2 describes the principle of reducing the complexity of a problem space in an evolutionary search by changing the basis and presents the motivation for evaluating the basis using the epistasis. In Section 3, a method is introduced for changing a standard basis to another basis for a binary encoding problem. Then,

a GA is introduced that can be used to apply a change of basis. Once an appropriate basis has been selected, this algorithm is more efficient at searching for a solution than the conventional GA. In Section 4, a method is proposed for estimating a basis that reduces the complexity of an evolutionary search problem. Section 5 describes a GA that can be used to search for a basis by applying the proposed estimation method. Here, a variable-length encoding scheme that consists of an elementary matrix is employed so to increase the efficiency of the search for an appropriate basis in the problem space. Section 6 presents a description of the tests used to validate the method and then discusses the results. In the tests, an appropriate basis for the target problem is found via the GA, and then the identified basis is applied to the target problem. The conclusions that can be drawn from this study are presented in Section 7.

## 2. Motivation

In this section, the concept of the epistasis is introduced as a means of estimating a basis that will reduce the complexity of the problem. First, a principal component analysis (PCA) is used to extract important information by changing the basis in real number encoding. Next, an example of changing the basis in binary encoding is presented to illustrate that a complex problem can be converted to a simple problem by changing the basis. Lastly, the epistases between the original and the modified problems are compared. If the epistasis of the problem decreases when the basis is changed, it implies the complexity of the original problem has decreased. Thus, a suitable basis can be identified using the changes in the epistasis before and after the prospective basis has been applied to the problem of interest.

*2.1. An Example of Changing a Basis in  $\mathbb{R}^n$ .* A PCA is used to obtain the principal components of the data by transforming the data into a new coordinate system via an orthogonal transformation. When the data is projected in the coordinate system, the position where the variance is the largest becomes the first principal component. The second principal component is in a position that is orthogonal to the previous component at the position with the second largest variance. Consequently, if the eigenvectors and eigenvalues of the covariance matrix are obtained and sorted in descending order, the principal components can be found. This is identical to changing the basis from the original coordinate system to a coordinate system based on the variance of the data. In general, by using only the important principal components, lost data are used.

*2.2. Change of Basis in Binary Representation.* Binary encoding typically employs a standard basis; however, it is sometimes easier to manipulate a problem in a nonstandard basis. The following example illustrates that the relationship between the basis vectors is dependent on the basis. Here,  $\mathbb{Z}_2$  is a field that has elements of zero and one, the addition operator corresponds to the exclusive-or (XOR) operator, and the multiplication operator corresponds to the AND

operator. The standard basis  $B_s$  for vector space  $\mathbb{Z}_2^n$  is  $\{e_1, e_2, \dots, e_n\}$ , where  $e_i$  consists of column vectors in which the  $i$ -th entry is one and the remaining  $n - 1$  entries are zero.

In the vector space  $\mathbb{Z}_2^n$ , if the vector  $v$  and the evaluation function  $F$  are as follows, then the basis vector  $e_i$  of  $B_s$  has a dependency relationship with the other basis vectors  $e_j$  in  $F$ :

$$v = \sum_{i=1}^n \alpha_i e_i = (\alpha_1, \alpha_2, \dots, \alpha_n), \quad (1)$$

$$F(v) = \sum_{i=1}^n (\alpha_1 \oplus \alpha_2 \oplus \dots \oplus \alpha_n) \oplus \alpha_i,$$

where  $\alpha_i \in \mathbb{Z}_2$  and  $\oplus$  is the XOR operator.

Let us assume a function  $F'$  performs the same operation as  $F$  but in a new basis and suppose  $n$  is even. If a set  $B$  is composed as follows:

$$B = \left\{ e'_i \mid \sum_{e_j \in B_s} e_j - e_i, \forall i = 1, 2, \dots, n \right\}. \quad (2)$$

then  $B$  becomes the basis. One property of a basis is that every vector can be represented as a linear combination of basis vectors. That is,

$$v = \sum_{i=1}^n \alpha_i e_i = \sum_{i=1}^n \alpha'_i e'_i, \quad (3)$$

where  $\alpha'_i = \sum_{j=1}^n \alpha_j + \alpha_i$  and  $[v]_B = (\alpha'_1, \alpha'_2, \dots, \alpha'_n)$ , which is the representation of  $v$  with respect to the basis  $B$ .

Here,  $F'$  is a function that evaluates  $[v]_B$ , has the same operation as  $F(v)$ , and satisfies the following relationship:

$$F(v) = \sum_{i=1}^n (\alpha_1 \oplus \alpha_2 \oplus \dots \oplus \alpha_n) \oplus \alpha_i = \sum_{i=1}^n \alpha'_i \quad (4)$$

$$= F'([v]_B).$$

It can be seen that the basis vector  $e'_i$  of  $B$  is independent of the other basis vectors in  $F'$ . In fact,  $F'$  is identical to the onemax problem that counts the number of ones in a bitstring. Therefore, for a vector in which all  $\alpha'_i$  are set to one, the evaluation value becomes the largest value, and if this vector is transformed with the standard basis, an optimum solution can be obtained. Figure 1 shows the relationships of the basis vectors according to the basis with  $n = 6$  in the graphs.

*2.3. Epistasis According to the Basis.* In a GA, the epistasis indicates the correlation between the genes. If the epistasis for a particular problem is large, then the genes are very inter-dependent, the fitness landscape of the problem space is extremely complex and the problem is difficult. In Section 2.2, it was shown that the complexity of a problem varies depending on the basis. The epistasis numerically expresses the complexity of such a problem. In general, when the genes in a problem are very dependent, the epistasis has a large

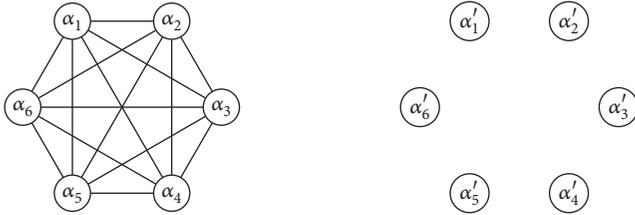


FIGURE 1: Dependency relationship of the different basis vectors of  $n = 6$  (left side: standard basis, right side: basis  $B$ ).

value. In contrast, when the genes are independent, the value is zero.

The results of calculating the epistasis according to the problem size  $n$  of evaluation functions  $F$  and  $F'$  in Section 2.2 are shown in Table 1. In this paper, the method proposed by Davidor [14] is used to compute the epistasis. In  $F$ , because the dependency relationship with other basis vectors increases as  $n$  increases, the epistasis also increases. However, for  $F'$ , since the basis vectors are independent, the epistasis is zero. Thus, it is expected that the search space can be simplified via an appropriate change of basis.

The epistasis can be used to check if the search space can be simplified by using a particular basis. If the epistasis of the problem after changing the basis is lower than the epistasis of the original problem, then this indicates that the problem has become easier. However, using the epistasis in this way requires all solutions to be searched. An alternative is to estimate the actual epistasis by calculating the epistasis of a sample set of solutions. Note that nonlinearity may be misleading due to approximation error by solution sampling. It hinders to find the proper basis for the target problem. The target problem may be transformed into a more complex problem through a basis transformation. That is, the basis transformation can rather prevent a GA from efficiently finding the solution.

### 3. Change of Basis

This section presents a GA that performs an effective search through a change of basis. Before presenting the GA, we introduce the related terminologies and theories of change of basis in binary representation. Next, we apply the change of basis in the onemax problem to show how the problem actually transformed. In addition, a methodology for evaluating solutions in the transformed problem will be described. Finally, we propose a GA that effectively searches solutions through applying the change of basis. On the other hand, searching for an appropriate basis will be covered in Sections 4 and 5.

**3.1. Change of Basis in  $\mathbb{Z}_2^n$ .** A basis for an  $n$ -dimensional vector space is a subset that consists of  $n$  vectors and every element of the space can be uniquely represented as a linear combination of basis vectors. Since it is possible to use one or more bases in a vector space, the coordinate representation of a vector with respect to the basis can be transformed via

TABLE 1: Epistasis of evaluation functions  $F$  and  $F'$ .

$n$	Epistasis in $F$	Epistasis in $F'$
2	0.0	0
4	1.0	0
6	1.5	0
8	2.0	0
10	2.5	0
12	3.0	0
14	3.5	0
16	4.0	0

an equivalent representation to other bases via the invertible linear transformation. Such a transformation is called a *change of basis*. The following theorem was derived from the basic theory of linear algebra [22].

**Theorem 1.** *Let  $B_1$  and  $B_2$  be two bases for  $\mathbb{Z}_2^n$ . Then, there exists a nonsingular matrix  $T \in M_{n \times n}(\mathbb{Z}_2)$  such that for every  $v \in \mathbb{Z}_2^n$ ,  $T[v]_{B_1} = [v]_{B_2}$ , where  $[v]_B$  is the representation of  $v$  with respect to the basis  $B$ .*

A matrix  $A$  is defined as binary if  $A \in M_{n \times n}(\mathbb{Z}_2)$ . In general, if  $B$  is the standard basis,  $[v]_B$  is the representation of  $v$  with respect to the basis  $B$ . In Theorem 1, nonsingular binary matrix  $T = [T]_{B_1}^{B_2}$  is a coordinate-change matrix from basis  $B_1$  to  $B_2$ . When a  $T$  is given,  $T$  can be viewed as a coordinate-change matrix from the standard basis to  $B_T$ , which is related to the  $T$ . For every vector  $v \in \mathbb{Z}_2^n$ ,  $Tv = [v]_{B_T}$  holds and  $B_T$  is  $\{Te_1, Te_2, \dots, Te_n\}$ . This study considers a change of basis from a standard basis to another basis. Thus, estimating the basis is equivalent to estimating an appropriate  $T$ .

**3.2. Analysis of Changing a Basis in the Onemax Problem.** The onemax problem maximizes the number of ones in a bitstring and has zero epistasis. Here, a onemax problem in which the basis was changed using a selected nonsingular binary matrix  $T$  is compared to the original onemax problem. The specific onemax problem of interest has a size of three. The  $T$  is defined as follows:

$$T = \begin{pmatrix} 1 & 0 & 0 \\ 1 & 0 & 1 \\ 0 & 1 & 0 \end{pmatrix}. \quad (5)$$

Then, it can be shown that  $B_T = \{Te_1, Te_2, Te_3\} = \left\{ \begin{pmatrix} 1 \\ 1 \\ 0 \end{pmatrix}, \begin{pmatrix} 0 \\ 0 \\ 1 \end{pmatrix}, \begin{pmatrix} 0 \\ 1 \\ 0 \end{pmatrix} \right\}$ . Table 2 shows the original vector and that obtained using  $Tv = [v]_{B_T}$ . From this, it can be seen that, after the basis change, the problem became more complex.

The evaluation function  $F$  of the onemax problem is as follows:

$$F(v) = \sum_{i=1}^n \alpha_i, \quad \text{where } v = \sum_{i=1}^n \alpha_i e_i. \quad (6)$$

On the other hand, from Table 2, it is difficult to identify a rule for the fitness of  $[v]_{B_T}$  for the onemax problem. The

*Step 1.* The population  $P$  of the GA is initialized and the fitness is evaluated.  
*Step 2.*  $P$  is replaced by the population  $P'$  whereby the standard basis  $B_s$  is changed to the basis  $B$ .  
*Step 3.* By using the genetic operator on the GA, the offspring population  $O'$  is produced from  $P'$ .  
*Step 4.* The fitness of  $O'$  is evaluated using the population  $O$  that was used to change the basis from  $B$  to  $B_s$ .  
*Step 5.*  $P'$  and  $O'$  are used to create a new generation and update  $P'$  to the new generation.  
*Step 6.* The process from Step 3 onward is repeated as many times as there are generations. When the number of generations has been exceeded, then we return  $P'$  whereby the basis  $B$  is changed to the standard basis  $B_s$ .

ALGORITHM 1: A GA with a change of basis.

TABLE 2: The vectors with a modified basis  $B_T$  and the original vectors in the onemax problem of size 3.

$v$	$[v]_{B_T}$	Fitness
$(1, 1, 1)^T$	$(1, 0, 1)^T$	3
$(1, 1, 0)^T$	$(1, 1, 1)^T$	2
$(1, 0, 1)^T$	$(1, 0, 0)^T$	
$(0, 1, 1)^T$	$(1, 1, 1)^T$	1
$(1, 0, 0)^T$	$(1, 1, 0)^T$	
$(0, 1, 0)^T$	$(0, 1, 0)^T$	
$(0, 0, 1)^T$	$(0, 0, 1)^T$	0
$(0, 0, 0)^T$	$(0, 0, 0)^T$	

evaluation function  $F'$  of  $[v]_{B_T}$  can be obtained by computing  $v$  by changing the basis from  $B_T$  to  $B_s$  and evaluating  $v$  with  $F$ . That is,

$$F'([v]_{B_T}) = F(T^{-1}[v]_{B_T}) = F(v), \quad (7)$$

where  $T^{-1}$  is the inverse matrix of  $T$ . The above equation is obtained by multiplying the left side by  $T^{-1}$  in  $Tv = [v]_{B_T}$  and then applying  $F$  to both sides. In this way, the basis on both sides can be easily changed using  $T$  and  $T^{-1}$ .

**3.3. Genetic Algorithm with a Change of Basis.** In general, a GA is expected to be more efficient when searching for a solution to a simple problem than a complex problem. As shown in Section 2.2, a complex problem can be changed to a simple problem by changing the basis. With this in mind, if an appropriate change of basis is applied to a problem space to be searched by a GA, this will greatly improve the efficiency of the search process. A flowchart of the proposed algorithm is shown in Figure 2 and the corresponding steps are detailed in Algorithm 1.

If Steps 2 and 4 are excluded, then Algorithm 1 produces a typical GA. However, if the problem is transformed with an appropriate basis in Step 2, the original problem space is transformed into an easier problem space, which is expected to make it easier for the GA to find an optimum solution. On the other hand, Step 4 shows that the generated offspring vector is evaluated by changing the basis to the standard basis. This is identical to the method in Section 3.2 that evaluates a solution in another basis.

## 4. Evaluation of a Basis

The objective is to identify a basis that can be used to change a complex problem into a simple problem. While such a basis was examined in Section 2.2, in that case, the change in basis converted the onemax problem from a simple to a complex problem.

When a basis and a target problem are given, a method is proposed that uses the epistasis to evaluate whether the basis is appropriate for the problem space. A meta-genetic algorithm (Meta-GA) is generally used as a method for estimating a hyperparameter of a GA. The two methods are compared to analyze the advantages and disadvantages of the proposed method.

**4.1. Evaluation with Epistasis.** Assume a target problem  $P$  and basis  $B$  are given. To determine the smoothing effect of  $B$  on  $P$ , a sampling population  $S$  can be obtained from  $P$ . Then,  $S'$  can be obtained by changing the basis for  $S$  from the standard basis to  $B$ . The epistasis of  $S'$  that numerically shows the difficulty of the problem can then be calculated. The lower the epistasis is, the more appropriate  $B$  is as a basis for  $P$ . The epistasis calculation method proposed by Davidor [14] is shown in Algorithm 2. Suppose the chromosome length is  $l$  and the number of samples in  $S$  is  $s$ . Then, the time complexity of evaluating a single basis becomes  $O(l^2s)$ . This is because the cost of executing the change of basis is  $l^2s$ . The change of basis is performed for a total of  $s$  vectors, and the cost of the change of basis is  $l^2$  since each vector  $v$  becomes  $[v]_B$  through  $[T]_{B_s}^B v$ .

**4.2. Evaluation with a Meta-Genetic Algorithm.** The use of a meta-GA to optimize the parameters and tune GAs was first proposed by Grefenstette [23]. Here, a meta-GA to determine whether the basis is appropriate for the problem space of the GA. A method of evaluating a basis with a meta-GA is shown in Algorithm 3. By applying Algorithm 1 with a given  $B$  and an instance of GA,  $k$  populations are searched. Then, using the best fitness in each population, the basis is evaluated. That is, when  $k$  units of fitness are found to be acceptable, it is estimated that  $B$  is an appropriate basis of the instance. The reason for searching  $k$  populations is because even with a basis that is not appropriate, a good solution may be obtained by using the GA to search once. To calculate the time complexity, with respect to the target GA, let the number of generations be  $g$ , population size  $p$ , and chromosome length  $l$ . The time cost of line (10) in Algorithm 3 is the largest.

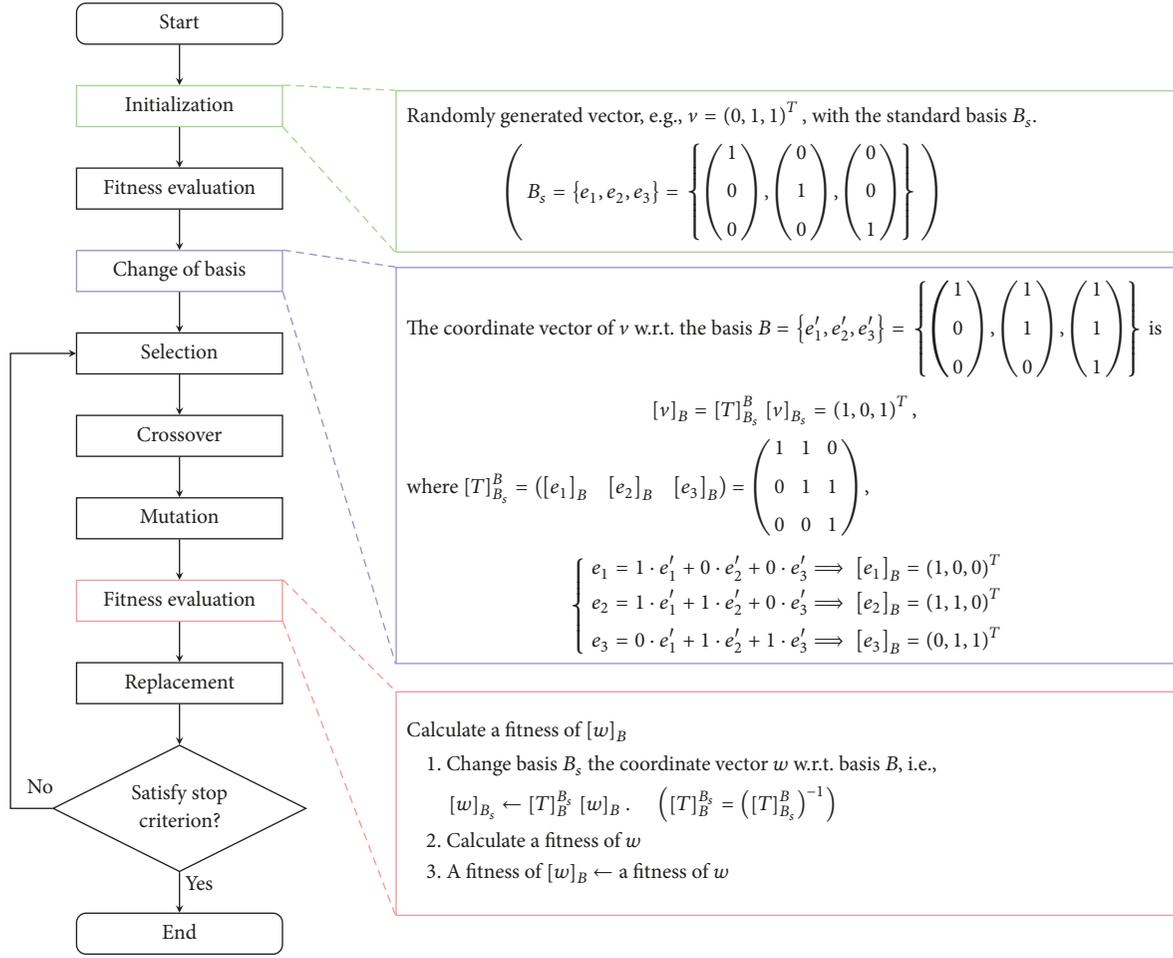


FIGURE 2: Flowchart of a GA with a change of basis.

When  $p$  offspring are generated, the time consumed is  $pl^2$ . Since this is repeated  $kg$  times, the worst case time complexity becomes  $O(kgpl^2)$ . Note that, in the experiment evaluated in this paper,  $k$  is set to 5 and  $g$  is set to the chromosome length.

## 5. Finding a Basis Using a Genetic Algorithm

This section describes the components of the GA used to search for a basis for the problem space with the evaluation method outlined in Section 4. The method of applying a basis and the genetic operator for the encoding are discussed, and the fitness of the basis is evaluated using the method of either Algorithm 2 or 3.

**5.1. Encoding with an Elementary Matrix.** A nonsingular binary matrix can be regarded as a change from a standard basis to another basis. That is, a basis corresponds to an appropriate the matrix. If a typical 2D type of encoding is used to encode the matrix, a repair mechanism may be required after recombination. In this case, one option is to conduct the repair using the Gauss-Jordan method; however, this will require a length of time equal to  $O(n^3)$  time.

Every nonsingular matrix can be expressed as a product of elementary matrices [22]. Therefore, in  $GL_n(\mathbb{Z}_2)$ , if a solution is expressed as a product of elementary matrices, it is possible to maintain their invertibility. Each element in an elementary matrix can be expressed by a variable-length linear string [24], which allows a new encoding to be applied. Note that any recombination method for a variable-length string can be used. In the following, an elementary row operation is defined and then the elementary matrix in  $M_{n \times n}(\mathbb{Z}_2)$  is introduced.

**Definition 2.** Let  $A \in M_{n \times n}(\mathbb{Z}_2)$ . Any one of the following two operations on the rows of  $A$  is called an elementary two operation:

- (i) Interchanging any two rows of  $A$ , and
- (ii) Adding a row of  $A$  to another row.

Elementary row operations are Type 1 or Type 2 depending on whether they were obtained using (i) or (ii) of Definition 2.

**Definition 3.** An  $n \times n$  elementary matrix in  $M_{n \times n}(\mathbb{Z}_2)$  is a matrix obtained by performing an elementary operation on

```

Require: Sampling population S
(1) procedure EVALUATION( $B, S$ )
(2)    $S' \leftarrow$  Change of basis from  $B_s$  to  $B$  on  $S$ 
(3)
(4)   for each ind in  $S'$ 
(5)      $\mu \leftarrow \mu + v(\text{ind})/\text{SIZE}(S')$ 
(6)     for  $i \leftarrow 1$  to  $\text{SIZE}(\text{ind})$  do
(7)        $a \leftarrow \text{ind}[i]$ 
(8)        $A[i][a] \leftarrow A[i][a] + v(\text{ind})$ 
(9)        $C[i][a] \leftarrow C[i][a] + 1$ 
(10)     end for
(11)   end for
(12)
(13)   for  $i \leftarrow 1$  to  $\text{SIZE}(\text{ind})$  do
(14)     for each  $a$  in allele values
(15)        $A[i][a] \leftarrow A[i][a]/C[i][a]$ 
(16)        $E[i][a] \leftarrow A[i][a] - \mu$ 
(17)     end for
(18)   end for
(19)
(20)   for each ind in  $S'$ 
(21)      $G \leftarrow 0$ 
(22)     for  $i \leftarrow 1$  to  $\text{SIZE}(\text{ind})$  do
(23)        $G \leftarrow G + E[i][\text{ind}[i]]$ 
(24)     end for
(25)      $G \leftarrow G + \mu$ 
(26)      $\sigma_{S'} \leftarrow \sigma_{S'} + (v(\text{ind}) - G)^2$ 
(27)   end for
(28)   return  $\sigma_{S'}$ 
(29) end procedure

```

ALGORITHM 2: Basis evaluation based on epistasis.

$I_n$ . The elementary matrix is said to be of Type 1 or Type 2 depending on whether the elementary operation performed on  $I_n$  is a Type 1 or Type 2 operation, respectively.

Let us define  $S_n^{ij}$  as an elementary matrix of Type 1 that interchanges the  $i$ -th row and the  $j$ -th one for  $i$  and  $j$ . Also define  $A_n^{ij}$  as an elementary matrix of Type 2 that adds the  $i$ -th row to the  $j$ -th row for  $i$  and  $j$ .

When the representation of a nonsingular binary matrix is considered in the order of an elementary matrix, this representation is not unique. Also, it is difficult to determine how many equivalent representations exist for a nonsingular binary matrix. Several equivalences were proposed by Yoon and Kim [24] as Propositions 4 and 5 by way of a simple idea. The newly discovered equivalences proposed in this paper are denoted in Proposition 6. Their proof is provided in the Appendix.

**Proposition 4** (exchange rule). *For each  $i, j, k$  such that  $i \neq j$ ,  $j \neq k$ , and  $k \neq i$ , the following five exchange rules hold:*

(i)  $A_n^{ik}A_n^{jk} = A_n^{jk}A_n^{ik}$ , (ii)  $A_n^{ij}A_n^{jk} = A_n^{ik}A_n^{ij}$ , (iii)  $S_n^{ij}A_n^{ik} = A_n^{jk}S_n^{ij}$ , (iv)  $S_n^{ij}A_n^{ki} = A_n^{kj}S_n^{ij}$ , and (v)  $S_n^{ij}S_n^{jk} = S_n^{jk}S_n^{ij} = S_n^{ik}S_n^{ij}$ .

**Proposition 5** (compaction rules). *For each  $i, j, k$  such that  $i \neq j$ ,  $j \neq k$ , and  $k \neq i$ , the following two exchange rules hold:*

(i)  $A_n^{ik}A_n^{jk}A_n^{ij} = A_n^{ij}A_n^{jk}$  and (ii)  $A_n^{kj}A_n^{ki}A_n^{ij} = A_n^{ji}A_n^{kj}$ .

**Proposition 6.** *For each  $i$  and  $j$  such that  $i \neq j$ , the following three rules hold:*

(i)  $A_n^{ij}S_n^{ij} = A_n^{ji}A_n^{ij}$ , (ii)  $S_n^{ij} = A_n^{ij}A_n^{ji}A_n^{ij}$ , and (iii)  $(A_n^{ij}A_n^{ji})^2 = A_n^{ji}A_n^{ij}$ .

For example, the encodings of matrices  $P_1$  and  $P_2$  are as follows: let  $P_1 = S_4^{12}A_4^{21}A_4^{12}$  and  $P_2 = A_4^{21}S_4^{12}$ . Then, calculate  $d_e(P_1, P_2)$  based on a sequence alignment between  $P_1$  and  $P_2$ , where  $d_e$  is the edit distance and the insertion, deletion, and replacement functions have weights of one, one, and two, respectively. First, consider the original form:

$$\begin{aligned}
 P_1 &= S_4^{12}A_4^{21}A_4^{12}, \\
 P_2 &= \text{---}A_4^{12}S_4^{12}.
 \end{aligned} \tag{8}$$

Then,  $d_e(P_1, P_2) = 3$ . This allows the parents to be changed into other forms. Note that

$$\begin{aligned}
 P_1 &= S_4^{12}A_4^{21}A_4^{12} = (S_4^{12}A_4^{21}A_4^{12})(A_4^{21}A_4^{21}) \\
 &= S_4^{12}(A_4^{21}A_4^{12}A_4^{21})A_4^{21} = S_4^{12}S_4^{21}A_4^{21} = A_4^{21}.
 \end{aligned} \tag{9}$$

From these rules,  $d_e(P_1, P_2) = 2$ . Thus, the propositions can produce offspring that are more similar to the parents.

<b>Require:</b> Target GA, Search GA $k$ times, Generations of GA $g$	
(1) <b>procedure</b> EVALUATION( $B, GA, k, g$ )	▷ Evaluation of a basis $B$
(2)   BestFits[ $k$ ]	▷ Return array
(3) <b>for</b> $i \leftarrow 1$ <b>to</b> $k$ <b>do</b>	
(4) $P \leftarrow$ GA.InitPopulation	▷ Initialization of population
(5)     GA.EvalPopulation( $P$ )	▷ Evaluation of the population
(6) $P \leftarrow$ Change of basis from $B_s$ to $B$ on $P$	
(7) <b>for</b> $j \leftarrow 1$ <b>to</b> $g$ <b>do</b>	
(8) $P'' \leftarrow$ GA.Selection( $P'$ )	
(9) $O' \leftarrow$ GA.Recombination( $P''$ )	▷ Perform crossover and mutation operations
(10) $O \leftarrow$ Change of basis from $B$ to $B_s$ on $O'$	
(11)      GA.EvalPopulation( $O$ )	
(12) $P' \leftarrow$ GA.Replace( $O', P'$ )	
(13) <b>end for</b>	
(14)     BestFits[ $i$ ] $\leftarrow$ the best fitness of $P'$	
(15) <b>end for</b>	
(16) <b>return</b> BestFits	
(17) <b>end procedure</b>	

ALGORITHM 3: Basis evaluation in a meta-GA.

5.2. *Crossover.* Any recombination for a variable-length string can be used as a recombination operator for the encoding and the edit distance is typically used as the distance for the variable-length string. This changes one string into another by using a minimum number of insertions, deletions, and replacements of the elementary matrix. A geometric crossover that is associated with this distance is called a homologous geometric crossover [25].

Several general string genetic operators can be used. In the case of a string encoding of the elementary matrix, a mathematically designed genetic operator was proposed [24]. Specifically, the geometric crossover by sequence alignment is expected to be effective. Here, alignment refers to allowing the strings to stretch in order to provide a better match. A stretched string involves interleaving the symbol ‘—’ anywhere in the string to create two stretched strings of the same length with a minimum Hamming distance. The offspring is generated by applying a uniform crossover to the aligned parents after removing the ‘—’ symbols. Here, two offspring solutions are generated as solutions of the two parents.

The optimal alignment of the two strings is as per the Wagner-Fischer algorithm [26], which is a dynamic programming (DP) algorithm that computes the edit distance between two strings of characters. This algorithm has a time complexity of  $O(mn)$  and a space complexity of  $O(mn)$  when the full dynamic programming table is constructed, where  $m$  and  $n$  are the lengths of the two strings.

5.3. *Initial Population, Selection, Mutation, and Replacement.* An initial population is generated with a random number of random elementary matrices. The random number is generated from a normal distribution where the mean is  $3n$  and the standard deviation is  $n$  when the problem size is  $n$ . If the random number is smaller than one, it is fixed at one. The selection operator applies a tournament selection method by

choosing three parents. The mutation operator applies one of three operations, namely insertion, deletion, or replacement, to each string with a 5% probability. Furthermore, the probability that each individual will be mutated is set at 0.2. Lastly, replacement refers to replacing the parent generation with an offspring generation. The details of this process are as follows: the selection operator is used for candidates of the offspring generation. When the population of the parent generation is  $p$ , then  $p$  parents are extracted by applying the selection operator  $p$  times. The probability of two parents pairing up and applying the crossover is 0.5. When the crossover is not applied, the two parents become candidates for members of the next generation, while in the opposite case, the two offspring become candidates for members of the next generation. Each candidate proceeds with a mutation probability of 0.2 and replaces the parent generation with the next generation.

## 6. Experiments

6.1. *Target Problem in Binary Representation.* In this section, two problems are described for which better solutions can be obtained with an appropriate basis.

- (1) Variant-onemax: for the evaluation function  $F$  of the onemax problem, vector  $v$  has an evaluation value of one. Variant-onemax is defined as counting the number of ones by changing vector  $v$  from the standard basis to a certain basis  $B$ . That is, in variant-onemax,  $F([v]_B)$  becomes the evaluation value for vector  $v$ .

If the basis is changed for  $v$  with the nonsingular binary matrix  $[T]_B^B$ , then we have  $[T]_B^B v = [v]_B$ . Then,  $F([v]_B)$  becomes a function that counts the number of 1s in  $[v]_B$ . This is therefore identical to the onemax problem as a result of an appropriate change of basis in variant-onemax. Meanwhile, from  $F([v]_B) =$

$F([T]_{B_i}^B, \nu)$ , an evaluation function of variant-onemax can be generated even when a nonsingular binary matrix is given. As for the optimum solution of variant-onemax, when the problem size is  $n$ , the number of ones becomes  $n$  through the change of basis, and  $n$  becomes the optimal solution.

- (2) *NK*-landscape: the *NK*-landscape model consists of a string of length  $N$  and a fitness contribution is attributed to each character depending on the other  $K$  characters. These fitness contributions are often randomly chosen from a particular probability distribution. In addition, the number of hills and valleys can be adjusted by varying  $N$  and  $K$ . One of the reasons why the *NK*-landscape model is used in optimization is that it is a simple instance of an NP-hard problem.

In the experiments, the GA is used to search for solutions to the above the two problems. The GA consists of tournament selection, one-point crossover, and flip mutation, and the replacement replaces all the parent generations with offspring generations. The tournament selection process chooses the best solution among three randomly selected parents, the one-point crossover combines a solution involving two offspring with the solution of two parents, while in flip mutation, each gene is flipped from zero to one or from one to zero with a probability of 0.05. The replacement method is the same as that described in Section 5. In other words, in the composition of the next generation, the number of parents extracted is equal to the number in the population. Two parents are paired up with a 50% probability that the crossover will be applied. When the crossover is not applied, the two parents become member candidates of the next generation, while in the opposite case, the two offspring become member candidates of the next generation. Each member candidate undergoes mutation with a 20% probability that it will replace an existing parent. When the chromosome length of variant-onemax or *NK*-landscape is  $n$ , the size of the population is set to  $4n$ . Because the fitness of the optimum solution of the variant-onemax problem is  $n$ , solutions of 10,000 generations have to be searched until an optimum solution has been identified. In the *NK*-landscape, the fitness of the optimum solution is different for each  $N$ ,  $K$ , and all solutions must be searched to obtain an optimum solution. Thus, 300,000 generations must be searched to find an optimum solution for the *NK*-landscape problem.

**6.2. Results.** The evaluation function of variant-onemax requires a nonsingular binary matrix that corresponds to a basis. For the basis of variant-onemax that has a chromosome length of  $n$ , a random number of elementary matrices are generated and then are multiplied sequentially. The number of elementary matrices is generated from a normal distribution that has a mean of  $3n$  and a standard deviation of  $n/2$ .

In the experiment, instances of variant-onemax where  $n$  was 20, 30, and 50 were generated. With the GA described in Section 5, the following bases were searched for each instance: meta-GA-based basis  $B_1$ , epistasis-based basis  $B_2$  where the

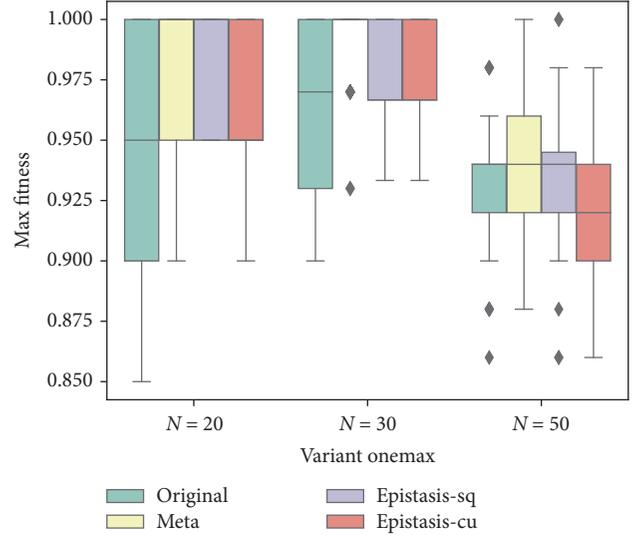


FIGURE 3: A box plot of each of the best solutions obtained by conducting the GA experiment 100 times on an instance of the variant-onemax problem.

sampling number was  $n^2$ , and epistasis-based basis  $B_3$  where the sampling number was  $n^3$ .

A total of 100 independent searches were conducted for each instance, and the number of times that an optimum solution was identified was counted along with the execution time. The results for the variant-onemax experiment are shown in Table 3. In the table, a type of ‘Original’ indicates that a solution instance was evaluated without a change of basis. Similarly, ‘Meta’, ‘Epistasis-sq’, and ‘Epistasis-cu’ refer to evaluating solution instances by applying  $B_1$ ,  $B_2$ , and  $B_3$ , respectively, to change the basis. In addition, the box plot in Figure 3 depicts the fitness distribution of the 100 best solutions obtained by performing 100 independent searches for each instance. A fitness is a value between zero and one that can be obtained by dividing the fitness of the optimum solution. That is, a value of one on the  $y$ -axis indicates the fitness of an optimum solution, while values approaching zero indicate a lower fitness. In most cases, it can be seen that the search performance of the GA is efficient with the change of basis. When  $N$  is 50, ‘Epistasis-cu’ does not seem to improve the search performance of the GA. This was likely because the population of the GA was not evenly distributed throughout the sample population.

In Table 3, ‘Meta’ found optimal solutions more frequently than the other methods. In particular, when  $n$  was 30, the 82nd most optimal solution was obtained out of 100. This indicates that the corresponding basis was appropriate. However, because the computation time for this approach was very long, it cannot be applied in practice. Note that when  $n$  is 50, it was over 2 hours. Furthermore, no difference was observed when compared to the case in which the basis was not changed. The method of evaluating the basis using the epistasis provides a good indication of when changing the basis will provide a better result. In particular, when  $n$  is 20, the number of optima found in ‘Original’ is 30, and the

TABLE 3: Results of each of the best solutions obtained by conducting the GA experiments 100 times on an instance of the variant-onemax problem. ('# of optima' is the number of optima found during 100 experiments, 'Average' is the average of 100 best solutions, and 'SD' is the standard deviation of 100 best solutions.  $Q_1$ ,  $Q_2$ , and  $Q_3$  are the first, second, and third quartiles, respectively. 'Time' is the sum of the time to search for the basis and that for the GA experiments.)

$n$	Type	# of optima	Average	SD	$Q_1$	$Q_2$	$Q_3$	Time (mm:ss)*
20	Original	30	0.945	0.0452	0.900	0.950	1.000	0:44
	Meta	66	0.980	0.0302	0.950	1.000	1.000	3:07
	Epistasis-sq	64	0.982	0.0241	0.950	1.000	1.000	1:01
	Epistasis-cu	33	0.964	0.2760	0.950	0.950	1.000	3:11
30	Original	31	0.963	0.0329	0.930	0.970	1.000	1:09
	Meta	82	0.993	0.0155	1.000	1.000	1.000	12:15
	Epistasis-sq	47	0.979	0.0216	0.967	0.967	1.000	3:49
	Epistasis-cu	40	0.979	0.0187	0.967	0.967	1.000	7:03
50	Original	0	0.931	0.0257	0.920	0.940	0.940	2:58
	Meta	0	0.939	0.0240	0.920	0.940	0.960	136:46
	Epistasis-sq	2	0.934	0.0272	0.920	0.940	0.945	7:48
	Epistasis-cu	0	0.927	0.0272	0.900	0.920	0.940	67:59

\*On Intel (R) Core TM i7-6850K CPU @ 3.60GHz

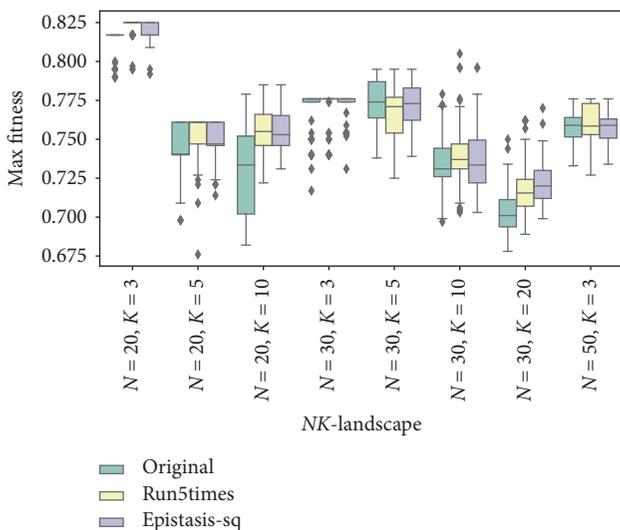


FIGURE 4: A box plot of each of the best solutions obtained by conducting the GA experiment 100 times on an instance of the  $NK$ -landscape problem.

numbers of optima found in 'Epistasis-sq' and 'Epistasis-cu' are 64 and 33, respectively. In summary, these tests confirmed that a sample size of  $n^2$  provided good results while requiring less time than a sample size of  $n^3$ . Therefore, in terms of time and performance, a sample size of  $n^2$  was deemed reasonable for estimating an epistasis.

The value of  $N$  in the  $NK$ -landscape experiment represents the size of the problem. In this experiment, there were  $N$  characters of zero and one and the total number of populations was  $2^N$ . The evaluation functions were randomly generated according to  $K$ . In terms of the instance generation, each gene was dependent on  $K$  other genes and a value between  $[0, 1]$  was assigned. The fitness of the  $NK$ -landscape is based on the fitness of each gene. Therefore, the maximum

and minimum fitness values, which are between zero and one, may be different for each instance. In the experiment, 100 independent searches for a solution are conducted for each instance. Table 4 shows the results of  $NK$ -landscape experiment in which the best solution and the computation time for each of the 100 searches were compared. In the table, when the type is 'Epistasis,' this indicates that a basis was obtained based on the epistasis using a sample set of size  $n^2$ , and which 100 independent searches were conducted for that instance. A box plot showing the distribution of the 100 best solutions is shown in Figure 4.

Upon analysis, the method of searching for the solution after changing the basis exhibited better performance than the original problem. In particular, in the box plot, it can be seen that the distribution of solutions obtained by changing the basis was more concentrated and had a higher mean. In the  $NK$ -landscape, when 'Meta' and 'Epistasis' were compared, neither side exhibited better performance. However, it can be seen that the computation time of 'Meta' was about 4–30 times longer than that of 'Epistasis.' Furthermore, although the 'Epistasis' consumed slightly more time than the 'Original,' it tended to have a more efficient evolutionary search. For these reasons, the method used to obtain the 'Epistasis' results was found to be the best among the three methods evaluated.

6.3. *Experimental Analysis.* The results of the above experiments confirmed that a basis obtained by estimating the epistasis improved the efficiency of searching for a solution using a GA. In this section, an analysis is performed to examine how much the basis found in the experiment reduced the epistasis. The basis was estimated in such a way that the epistasis of the sample population  $S$  was reduced. Whether the GA proposed in Section 5 was effective can be confirmed by comparing the epistasis of  $S$  and that of  $S'$  in which the basis was changed to the one identified by the search  $S$ . It is expected that the latter epistasis will be smaller.

TABLE 4: Results of each of the best solutions obtained by conducting the GA experiments 100 times on an instance of the  $NK$ -landscape problem. ('Best' is the best fitness among solutions found in 100 experiments, 'Average' is the average of 100 best solutions, and 'SD' is the standard deviation of 100 best solutions.  $Q_1$ ,  $Q_2$ , and  $Q_3$  are the first, second, and third quartiles, respectively. 'Time' is the sum of the time to search for the basis and that for the GA experiments.)

$N, K$	Type	Best	Average	SD	$Q_1$	$Q_2$	$Q_3$	Time (mm:ss)*
20, 3	Original	0.817	0.8135	0.0085	0.8170	0.8170	0.8170	1:02
	Meta	0.825	0.8226	0.0057	0.8250	0.8250	0.8250	5:52
	Epistasis	0.825	0.8200	0.0056	0.8170	0.8170	0.8250	1:32
20, 5	Original	0.761	0.7449	0.0157	0.7400	0.7405	0.7610	1:03
	Meta	0.761	0.7533	0.0131	0.7470	0.7610	0.7610	5:39
	Epistasis	0.761	0.7505	0.0109	0.7460	0.7470	0.7610	1:40
20, 10	Original	0.779	0.7306	0.0253	0.7020	0.7335	0.7520	1:10
	Meta	0.785	0.7572	0.0155	0.7660	0.7550	0.7660	7:13
	Epistasis	0.785	0.7558	0.0136	0.7460	0.7530	0.7653	2:16
30, 3	Original	0.776	0.7687	0.1373	0.7740	0.7760	0.7760	2:06
	Meta	0.776	0.7719	0.0109	0.7760	0.7760	0.7760	5:39
	Epistasis	0.776	0.7718	0.0090	0.7740	0.7760	0.7760	1:40
30, 5	Original	0.795	0.7725	0.0125	0.7638	0.7740	0.7870	2:06
	Meta	0.795	0.7661	0.0170	0.7540	0.7710	0.7770	32:28
	Epistasis	0.795	0.7706	0.0136	0.7623	0.7730	0.7830	2:50
30, 10	Original	0.779	0.7349	0.0181	0.7260	0.7310	0.7443	2:06
	Meta	0.805	0.7391	0.0179	0.7310	0.7370	0.7470	49:47
	Epistasis	0.796	0.7366	0.0198	0.7220	0.7335	0.7960	3:48
30, 20	Original	0.750	0.7039	0.0152	0.6938	0.7010	0.7113	2:51
	Meta	0.762	0.7181	0.0163	0.7070	0.7155	0.7243	49:47
	Epistasis	0.770	0.7220	0.0133	0.7120	0.7200	0.7300	3:48
50, 3	Original	0.776	0.7576	0.0102	0.7515	0.7590	0.7640	5:31
	Meta	0.776	0.7599	0.0119	0.7530	0.7585	0.7730	220:14
	Epistasis	0.776	0.7578	0.0096	0.7508	0.7590	0.7630	6:34

\*On Intel (R) Core TM i7-6850K CPU @ 3.60GHz

TABLE 5: Epistasis of the original and modified basis sampling in the variant-onemax problem.

$n$	Sampling size	Epistasis		
		Before	After	Decrease rate (%)*
20	square	4.46	3.23	27.6
	cubic	4.35	3.83	12.0
30	square	4.57	3.20	30.0
	cubic	5.00	3.72	25.6
50	square	9.27	7.53	18.8
	cubic	9.69	8.93	7.8

\*Decrease rate =  $100 \times (\text{Before} - \text{After}) / \text{Before}$

A comparison of the epistasises between  $S$  and  $S'$  in the variant-onemax and  $NK$ -landscape experiments can be seen in Tables 5 and 6, respectively. First, in Table 5,  $n$  is the chromosome length of the variant-onemax experiment. The sizes of the sample sets were  $n^2$  and  $n^3$ , respectively; 'Before' and 'After' show the epistasises of  $S$  and  $S'$ , respectively. For every  $n$ , it was confirmed that a lower epistasis value was obtained when the basis was changed. Moreover, when the sampling size was 'square', the epistasis was reduced more compared to the 'cubic'. Thus, there was a higher possibility

that the GA would conduct a more efficient search and find a better solution. When  $n$  was 20, since there were  $2^{20}$  solutions, the epistasis for all the solutions, not the sample sets, can be obtained. Here, it was confirmed that the epistasis was 4.50, and since the epistasis was 4.46 and 4.35 when the sampling sizes were square and cubic, respectively, this indicates that the original epistasis was accurately estimated.

In Section 5.2, the size of the sample set  $S$  in the  $NK$ -landscape experiment was  $N^2$ . Table 6 shows the epistasises of  $S$  and  $S'$  after the basis was changed, respectively, for

TABLE 6: Epistasis of the original and modified basis sampling in the NK-landscape problem.

$N, K$	Epistasis		
	Before	After	Decrease rate (%)*
20, 3	$3.17e^{-3}$	$2.25e^{-3}$	29.0
20, 5	$3.16e^{-3}$	$2.90e^{-3}$	8.2
20, 10	$4.28e^{-3}$	$3.82e^{-3}$	10.7
30, 3	$1.85e^{-3}$	$1.60e^{-3}$	13.5
30, 5	$2.61e^{-3}$	$2.37e^{-3}$	9.2
30, 10	$2.68e^{-3}$	$2.39e^{-3}$	10.8
30, 20	$2.78e^{-3}$	$2.50e^{-3}$	10.1
50, 3	$1.13e^{-3}$	$9.32e^{-4}$	17.5

\*Decrease rate =  $100 \times (\text{Before} - \text{After}) / \text{Before}$

the values of  $N, K$  used in the experiment. The ‘Before’ and ‘After’ results indicate the epistases of  $S$  and  $S'$ , respectively. As in the case of the variant-onemax experiment, it was confirmed that for every  $n$ , a lower value of epistasis was obtained when a change of basis was applied. When  $N$  was 20, the epistasis for all the solutions, but not the samplings, was obtained. When  $K$  was 3, 5, and 10, the epistasis was  $3.24e^{-3}$ ,  $3.38e^{-3}$ , and  $4.13e^{-3}$ , respectively. These values are close to the respective epistases of  $S$ ,  $3.17e^{-3}$ ,  $3.16e^{-3}$ , and  $4.28e^{-3}$ .

## 7. Conclusions

In this paper, an epistasis-based evolutionary search method was proposed for estimating a basis that would simplify a particular problem. Two test problems were constructed, a basis was identified by estimating the epistasis, and after the basis was changed, the results before and after the basis change were compared. The epistasis-based basis estimation method was found to be extremely efficient compared to a meta-GA in terms of time. This was also found for the NK-landscape in which the epistasis-based basis estimation method provided similar results. Thus, it is reasonable to estimate the basis by using the epistasis rather the meta-GA algorithm.

To estimate an epistasis, sample sets of size  $n^2$  or  $n^3$  sampling data were used. It was therefore necessary to conduct a study to find an appropriate sampling number. However, the method of finding the basis was carried out using a simple GA. In the future, a study should be conducted to identify a better basis. Also, by applying various factors in the GA or other genetic operators or by applying the method shown in the Appendix, a higher quality search can be performed.

Furthermore, the experiment evaluated specific problems that could be simplified with a change of basis. In further research, it will be necessary to identify the characteristics of problems that could benefit from a change of basis. Note that the basis evaluation method is applicable to not only binary encoding, but also to  $k$ -ary encoding. In addition, it can be used to evaluate any vector space in which the epistasis can be calculated.

## Appendix

We present the following lemma to prove Proposition 6:

**Lemma A.1.** *Letting  $M = (m_{ij})$  be an  $n \times n$  binary matrix. For each  $i$  and  $j$  such that  $i \neq j$ , the following four rules hold:*

- (i)  $\text{Row}_i(A_n^{ij}M) = \text{Row}_i(M)$ ,
- (ii)  $\text{Row}_j(A_n^{ij}M) = \text{Row}_j(M) + \text{Row}_i(M)$ ,
- (iii)  $\text{Row}_i(S_n^{ij}M) = \text{Row}_j(M)$ , and
- (iv)  $\text{Row}_j(S_n^{ij}M) = \text{Row}_i(M)$ ,

where  $\text{Row}_i(M)$  is the  $i$ -th row vector of matrix  $M$ .

*Proof.* Let  $m_i$  be the  $i$ -th row vector of  $M$ ; that is,  $M = (m_1, m_2, \dots, m_n)^T$ . Without loss of generality, we assume that  $i < j$ . Note that

$$A_n^{ij}M = \begin{pmatrix} \vdots \\ \text{Row}_i(A_n^{ij}M) \\ \vdots \\ \text{Row}_j(A_n^{ij}M) \\ \vdots \end{pmatrix} = \begin{pmatrix} \vdots \\ m_i \\ \vdots \\ m_j + m_i \\ \vdots \end{pmatrix}, \text{ and}$$

$$S_n^{ij}M = \begin{pmatrix} \vdots \\ \text{Row}_i(S_n^{ij}M) \\ \vdots \\ \text{Row}_j(S_n^{ij}M) \\ \vdots \end{pmatrix} = \begin{pmatrix} \vdots \\ m_j \\ \vdots \\ m_i \\ \vdots \end{pmatrix}. \quad (\text{A.1})$$

So, we have the following:

- (i)  $\text{Row}_i(A_n^{ij}M) = m_i = \text{Row}_i(M)$ ,
- (ii)  $\text{Row}_j(A_n^{ij}M) = m_j + m_i = \text{Row}_j(M) + \text{Row}_i(M)$ ,
- (iii)  $\text{Row}_i(S_n^{ij}M) = m_j = \text{Row}_j(M)$ , and
- (iv)  $\text{Row}_j(S_n^{ij}M) = m_i = \text{Row}_i(M)$ .

□

*Proof of Proposition 6.* Let  $M = (m_{ij})$  be an  $n \times n$  binary matrix which  $m_i$  is the  $i$ -th row vector; that is,  $M = (m_1, m_2, \dots, m_n)^T$ .

- (1) It is enough to show that the  $i$ -th and  $j$ -th row vectors of  $A_n^{ij}S_n^{ij}M$  are the same as those of  $A_n^{ji}A_n^{ij}M$ . Consider the left side: using Lemma A.1, we have

$$\begin{aligned} \text{Row}_i(A_n^{ij}S_n^{ij}M) &= \text{Row}_i(S_n^{ij}M) = \text{Row}_j(M) \\ &= m_j, \text{ and} \\ \text{Row}_j(A_n^{ij}S_n^{ij}M) &= \text{Row}_j(S_n^{ij}M) + \text{Row}_i(S_n^{ij}M) \\ &= \text{Row}_j(M) + \text{Row}_i(M) \\ &= m_j + m_i. \end{aligned} \quad (\text{A.2})$$

Now consider the right side:

$$\begin{aligned} \text{Row}_i(A_n^{ji}A_n^{ij}M) &= \text{Row}_i(A_n^{ij}M) + \text{Row}_j(A_n^{ij}M) \\ &= \text{Row}_i(M) \\ &\quad + (\text{Row}_j(M) + \text{Row}_i(M)) \\ &= m_j, \text{ and} \end{aligned} \quad (\text{A.3})$$

$$\begin{aligned} \text{Row}_j(A_n^{ji}A_n^{ij}M) &= \text{Row}_j(A_n^{ij}M) \\ &= \text{Row}_i(M) + \text{Row}_i(M) \\ &= m_i + m_j. \end{aligned}$$

- (2) We know  $A_n^{ij}S_n^{ij} = A_n^{ji}A_n^{ij}$ . We multiply  $A_n^{ij}$  in both sides. Then, the left side is  $A_n^{ij}A_n^{ij}S_n^{ij} = I_n S_n^{ij} = S_n^{ij}$ , and so  $S_n^{ij} = A_n^{ij}A_n^{ji}A_n^{ij}$ .

- (3)  $S_n^{ij} = S_n^{ji}$  by the definition of  $S_n^{ij}$ . Now, consider  $S_n^{ij}A_n^{ji} = S_n^{ji}A_n^{ij}$ . Note that the left side

$$S_n^{ij}A_n^{ji} = (A_n^{ij}A_n^{ji}A_n^{ij})A_n^{ji} = (A_n^{ij}A_n^{ji})^2, \quad (\text{A.4})$$

and note that the right side

$$S_n^{ji}A_n^{ij} = (A_n^{ji}A_n^{ij}A_n^{ji})A_n^{ij} = (A_n^{ji}A_n^{ij})(A_n^{ji})^2 = A_n^{ji}A_n^{ij}. \quad (\text{A.5})$$

□

## Data Availability

The data used to support the findings of this study are included within the article.

## Conflicts of Interest

The authors declare that they have no conflicts of interest.

## Acknowledgments

The present research has been conducted by the Research Grant of Kwangwoon University in 2019. This research was supported by a grant (KCG-01-2017-05) through the Disaster and Safety Management Institute funded by Korea Coast Guard of Korean government and by Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Education (No. 2015R1D1A1A01060105).

## References

- [1] I. Hwang, Y.-H. Kim, and B.-R. Moon, "Multi-attractor gene reordering for graph bisection," in *Proceedings of the 8th Annual Genetic and Evolutionary Computation Conference*, pp. 1209–1215, July 2006.
- [2] D. X. Chang, X. D. Zhang, and C. W. Zheng, "A genetic algorithm with gene rearrangement for K-means clustering," *Pattern Recognition*, vol. 42, no. 7, pp. 1210–1222, 2009.
- [3] D. Sankoff and M. Blanchette, "Multiple genome rearrangement and breakpoint phylogeny," *Journal of Computational Biology*, vol. 5, no. 3, pp. 555–570, 1998.
- [4] G. R. Raidl and B. A. Julstrom, "A weighted coding in a genetic algorithm for the degree-constrained minimum spanning tree problem," in *Proceedings of the ACM Symposium on Applied Computing (SAC '00)*, vol. 1, pp. 440–445, ACM, March 2000.
- [5] E. Falkenauer, "A new representation and operators for genetic algorithms applied to grouping problems," *Evolutionary Computation*, vol. 2, no. 2, pp. 123–144, 1994.
- [6] M. Gen, F. Altiparmak, and L. Lin, "A genetic algorithm for two-stage transportation problem using priority-based encoding," *OR Spectrum*, vol. 28, no. 3, pp. 337–354, 2006.
- [7] Y. M. Wang, H. L. Yin, and J. Wang, "Genetic algorithm with new encoding scheme for job shop scheduling," *The International Journal of Advanced Manufacturing Technology*, vol. 44, no. 9–10, pp. 977–984, 2009.

- [8] M. M. Lotfi and R. Tavakkoli-Moghaddam, "A genetic algorithm using priority-based encoding with new operators for fixed charge transportation problems," *Applied Soft Computing*, vol. 13, no. 5, pp. 2711–2726, 2013.
- [9] F. Pernkopf and P. O'Leary, "Feature selection for classification using genetic algorithms with a novel encoding," in *Proceedings of the International Conference on Computer Analysis of Images and Patterns*, vol. 2001, pp. 161–168.
- [10] Y. Wang, L. Han, Y. Li, and S. Zhao, "A new encoding based genetic algorithm for the traveling salesman problem," *Engineering Optimization*, vol. 38, no. 1, pp. 1–13, 2006.
- [11] J.-Z. Wu, X.-C. Hao, C.-F. Chien, and M. Gen, "A novel bi-vector encoding genetic algorithm for the simultaneous multiple resources scheduling problem," *Journal of Intelligent Manufacturing*, vol. 23, no. 6, pp. 2255–2270, 2012.
- [12] D. Wyatt and H. Lipson, "Finding building blocks through eigenstructure adaptation," in *Proceedings of the Genetic and Evolutionary Computation Conference*, pp. 1518–1529, 2003.
- [13] Y.-H. Kim and Y. Yoon, "Effect of changing the basis in genetic algorithms using binary encoding," *KSII Transactions on Internet and Information Systems*, vol. 2, no. 4, pp. 184–193, 2008.
- [14] Y. Davidor, "Epistasis variance: suitability of a representation to genetic algorithms," *Complex Systems*, vol. 4, no. 4, pp. 369–383, 1990.
- [15] D. Seo, Y. Kim, and B. R. Moon, "New entropy-based measures of gene significance and epistasis," in *Proceedings of the Genetic and Evolutionary Computation Conference*, vol. 2724, pp. 1345–1356, 2003.
- [16] D. Seo, S. Choi, and B. Moon, "New epistasis measures for detecting independently optimizable partitions of variables," in *Proceedings of the Genetic and Evolutionary Computation Conference*, pp. 150–161, 2004.
- [17] M. Ventresca and B. Ombuki-Berman, "Epistasis in multi-objective evolutionary recurrent neuro-controllers," in *Proceedings of the 1st IEEE Symposium on Artificial Life, IEEE-ALife'07*, pp. 77–84, USA, April 2007.
- [18] D.-I. Seo and B.-R. Moon, "Computing the variance of large-scale traveling salesman problems," in *Proceedings of the GECCO 2005 - Genetic and Evolutionary Computation Conference*, pp. 1169–1176, USA, June 2005.
- [19] C. R. Reeves and C. C. Wright, "Epistasis in genetic algorithms: an experimental design perspective," in *Proceedings of the International Conference on Genetic Algorithms*, pp. 217–224, 1995.
- [20] B. Naudts and L. Kallel, "A comparison of predictive measures of problem difficulty in evolutionary algorithms," *IEEE Transactions on Evolutionary Computation*, vol. 4, no. 1, pp. 1–15, 2000.
- [21] D. Beasley, R. David, and R. Ralph, "Reducing epistasis in combinatorial problems by expansive coding," in *Proceedings of the International Conference on Genetic Algorithms*, pp. 400–407, 1993.
- [22] S. H. Friedberg, A. J. Insel, and L. E. Spence, *Linear Algebra*, Prentice Hall, Upper Saddle River, NJ, USA, 3rd edition, 1997.
- [23] J. J. Grefenstette, "Optimization of control parameters for genetic algorithms," *IEEE Transactions on Systems, Man, and Cybernetics*, vol. 16, no. 1, pp. 122–128, 1986.
- [24] Y. Yoon and Y.-H. Kim, "A mathematical design of genetic operators on  $GL_n(\mathbb{Z}_2)$ ," *Mathematical Problems in Engineering*, vol. 2014, Article ID 540936, 8 pages, 2014.
- [25] A. Moraglio, P. Riccardo, and R. Seehuus, "Geometric crossover for biological sequences," in *In Proceedings of the European Conference on Genetic Programming*, pp. 121–132, 2006.
- [26] R. A. Wagner and M. J. Fischer, "The string-to-string correction problem," *Journal of the ACM*, vol. 21, pp. 168–173, 1974.

