New Epistasis Measures for Detecting Independently Optimizable Partitions of Variables

Dong-Il Seo, Sung-Soon Choi, and Byung-Ro Moon

School of Computer Science & Engineering, Seoul National University
Sillim-dong, Gwanak-gu, Seoul, 151-744 Korea
{diseo, sschoi, moon}@soar.snu.ac.kr
http://soar.snu.ac.kr/~{diseo, sschoi, moon}/

Abstract. An optimization problem is often represented with a set of variables, and the interaction between the variables is referred to as epistasis. In this paper, we propose two new measures of epistasis: *internal epistasis* and *external epistasis*. Then we show that they can quantify the decomposability of a problem, which has a theoretical meaning about how strongly the problem is independently optimizable with a partition of variables. We present examples of the problem decomposition and the results of experiments that support the consistency of the measures.

1 Introduction

An optimization problem is specified by a set of problem instances, each of which is a pair $(\mathcal{U}, f)$, where the universe $\mathcal{U}$ is the set of feasible solutions and the fitness function f is a mapping $f : \mathcal{U} \to \mathbb{R}$ [1]. A solution set is often represented by a set of variables, whose values can have certain ranges. If the ranges are discrete, the problem is said to be combinatorial [2]. The interaction between the variables is referred to as epistasis, which implies that the contribution of a variable to the fitness depends on the values of other variables. The epistasis is one of the main reasons that one cannot solve problems by naive approaches, such as steepest ascent method, which try to optimize each variable independently.

This difficulty was addressed recently with the techniques relevant to the "building blocks" or the "factorization" in the evolutionary algorithms. A building block is a specific assignment to a subset of variables that contributes to a high fitness, which is believed to give a shortcut to the optimal solution. Such building blocks are detected and proliferated in the population by the genetic operators in topological linkage-based genetic algorithms (TLBGAs) [3]. Thus the precise and efficient detection of the building blocks is critical in the black box optimization where no problem-specific knowledge is available. Instead of manipulating the building blocks by the genetic operators, the distribution of the variables is explicitly estimated in estimation-of-distribution algorithms (EDAs) [4,5]. EDAs do not use crossover nor mutation operator. Instead, the new population of individuals is sampled from a probability distribution, which

K. Deb et al. (Eds.): GECCO 2004, LNCS 3103, pp. 150–161, 2004.
© Springer-Verlag Berlin Heidelberg 2004

is estimated from selected individuals from the previous generation. Since the exact estimation of the whole distribution of the variables is computationally prohibitive for large-scale problems, the variable set is factorized into a number of subsets in which variables are believed to have strong dependence with each other. The factorization is often based on probabilistic graphic model [6, 7] in recent EDAs. The building block detection and the factorization of the distribution are strongly correlated with the decomposition of a problem since both of them are, in some sense, to find the subgroups of the variables that are likely to be optimized individually. The approaches are, however, not free from specific algorithms and their running status since the building block detection and the model construction are based on the solutions selected by fitness from the population or the probability distribution at that point of time.

There are a number of algorithm-independent measures of epistasis including the epistasis variance by Davidor [8] and the entropic epistasis by Seo *et al.* [9,10]. The epistasis variance quantifies the nonlinearity lying in the fitness landscape based on experimental design [11], while the entropic epistasis measures, using Shannon's information theory [12,13], the amount of information shared by the variables about the fitness. In this paper, we extend the entropic epistasis to the problem decomposition. To do so, we first show that the epistasis can be split into two factors: *internal epistasis* and *external epistasis*. Then we formally define the theoretical concept of problem decomposition in the sense of independent optimization. Finally, we show the relationship between the two novel measures and the decomposability of a problem.

The reset of this paper is organized as follows. We provide a brief overview of the entropic epistasis and propose new epistasis measures in Section 2. Then we formally define the decomposition of a problem and show the relationship between the decomposability and the epistasis of the problem in Section 3. The examples of problem decomposition and the experimental results are presented in Section 4. The conclusions are given finally in Section 5.

2 Epistasis Measures

2.1 Probability Model

Let the variable indices of a given problem be $\mathcal{V} = \{1, 2, \ldots, n\}$, and the alphabet for each variable be $\mathcal{A}_i, i \in \mathcal{V}$. And let the universe and the fitness function be $\mathcal{U} \subseteq \mathcal{A}_1 \times \mathcal{A}_2 \times \cdots \times \mathcal{A}_n$ and $f : \mathcal{U} \to \mathcal{F}$, respectively. We assume that the alphabet of each variable is finite. Then, the set of all fitness values $\mathcal{F} \subset \mathbb{R}$ is finite as the universe is finite.

Based on the uniform probability model on the universe, we define a random variable X_i for each variable $i \in \mathcal{V}$ and a random variable Y for the fitness value. Then, the joint probabilistic mass function (jpmf) $p : \mathcal{A}_1 \times \cdots \times \mathcal{A}_n \times \mathcal{F} \to \mathbb{R}$ of the random variables are defined as follows:

$$p(x_1, x_2, \ldots, x_n, y) = \begin{cases} \frac{1}{|\mathcal{U}|} & \text{if } (x_1, x_2, \ldots, x_n) \in \mathcal{U} \\ & \quad \text{and } y = f(x_1, x_2, \ldots, x_n) \\ 0 & \text{otherwise.} \end{cases} \tag{1}$$

It means that the probability of a solution $(x_1, x_2, \ldots, x_n)$ and a fitness value y is $\frac{1}{|\mathcal{U}|}$ if the fitness value of the solution $f(x_1, x_2, \ldots, x_n)$ is y, and the probability is zero otherwise. In this paper, we use a conventional notation X_V to denote $(X_{v_1}, X_{v_2}, \ldots, X_{v_k})$ for a variable set $V = \{v_1, v_2, \ldots, v_k\} \subseteq \mathcal{V}$.

It is practical to use a set of sampled solutions instead of the universe $\mathcal{U}$ in Equation (1) for large-scale problems because of the spatial or computational limitations. In the case, the size of the set must be not too small to get results of low levels of distortion (see [10] for details).

In general, the fitness function of a problem is defined on a continuous domain, while each variable has a discrete alphabet in combinatorial optimization problems. Although the set $\mathcal{F}$ of all fitness values of a problem instance is finite, it is less practical to consider each distinct value in $\mathcal{F}$ as a discrete symbol for the random variable Y since we can hardly talk about the statistics including fitness as one of the variables if the solutions rarely share the same fitness. Hence the fitness needs to be discretized into a number of intervals (see [10] for details).

2.2 Significance and Epistasis

The significance of a variable set is defined to be the mutual information between the corresponding random variables and the random variable Y. It is intuitive and natural because if we can get more information about the fitness from the values of the variables then we can say that the variables contribute more to the fitness. Formally, the *significance* $\xi(V)$ of a variable set $V = \{v_1, v_2, \ldots, v_k\} \subseteq \mathcal{V}$, $k \geq 1$, is defined as follows [9,10]:

$$\xi(V) = \frac{I(X_V; Y)}{H(Y)} \tag{2}$$

where I denotes the mutual information [13]. We do not consider the case when the fitness is constant, i.e., the case of $|\mathcal{F}| = 1$, because no optimization is required in the case. Hence we regard the entropy of Y as a nonzero value. The equation in the definition is a normalized formula, as is easily verified in the following. The significance $\xi(V)$ of a variable set $V \subseteq \mathcal{V}$ satisfies the following inequality:

$$0 \leq \xi(V) \leq 1. \tag{3}$$

If the significance is zero, then we cannot get any information about the fitness from the corresponding variables. On the contrary, if the significance is one, then we can fully identify the fitness from the variables. It is clear that $\xi(\mathcal{V}) = 1$ since $I(X_\mathcal{V}; Y) = H(Y)$, and $\xi(V) \leq \xi(W)$ for $V \subseteq W \subseteq \mathcal{V}$ since $I(X_V; Y) \leq I(X_W; Y)$ by the chain rule (see [13] p. 22).

The epistasis can be defined rigorously from the significance. From (2), the significance of a variable set $V \subseteq \mathcal{V}$ is denoted by $\xi(V)$ and the significance of each variable v in $\mathcal{V}$ is denoted by $\xi(v)$. We define the epistasis between the variables in V to be the difference between $\xi(V)$ and the summation of all $\xi(v)$'s,

$v \in V$. Formally, the *epistasis* $\varepsilon(V)$ of a variable set $V = \{v_1, v_2, \ldots, v_k\} \subseteq \mathcal{V}$, $k \geq 1$, is defined as follows [9,10]:

$$\varepsilon(V) = \begin{cases} \dfrac{\xi(V) - \sum_{i=1}^{k} \xi(v_i)}{\xi(V)} & \text{if } I(X_V; Y) \neq 0 \\ 0 & \text{otherwise,} \end{cases} \tag{4}$$

which is rewritten as

$$\varepsilon(V) = \frac{I(X_V; Y) - \sum_{i=1}^{k} I(X_{v_i}; Y)}{I(X_V; Y)} \tag{5}$$

when $I(X_V; Y) \neq 0$. The epistasis $\varepsilon(V)$ of a variable set $V \subseteq \mathcal{V}$ satisfies the following inequality:

$$1 - |V| \leq \varepsilon(V) \leq 1. \tag{6}$$

The epistasis has a positive value when $\xi(V)$ is greater than $\sum_{i=1}^{k} \xi(v_i)$ and it has a negative value when $\xi(V)$ is smaller than $\sum_{i=1}^{k} \xi(v_i)$. The former case means that the corresponding variables interact constructively with each other, and the latter case means that they interact destructively with each other.

The epistasis has a nonnegative value if the universe contains the whole combinations of the alphabets. Formally, if the universe $\mathcal{U}$ of a problem is defined to be $\mathcal{A}_1 \times \mathcal{A}_2 \times \cdots \times \mathcal{A}_n$, then the epistasis $\varepsilon(V)$ of a variable set $V \subseteq \mathcal{V}$ is nonnegative, i.e.,

$$0 \leq \varepsilon(V) \leq 1. \tag{7}$$

The equation means that a set of variables always interact constructively in such a fitness function. If we get a negative epistasis from a sampled solution set, then it implies that the solution set was mis-sampled in the case. The epistasis has a zero value if and only if the corresponding variables are conditionally independent given Y, i.e., $p(x_{v_1}, x_{v_2}, \ldots, x_{v_k}|y) = \prod_{i=1}^{k} p(x_{v_i}|y)$ for all $y \in \mathcal{F}$.

2.3 Internal/External Epistasis

A set of disjoint nonempty subsets of a set V is said to be a partition of V if the union of the subsets is V. A partition composed of k subsets is said to be a k-way partition. Based on the epistasis measures shown in Section 2.2, we devise two novel epistasis measures for a partition, *internal epistasis* and *external epistasis*.

The internal epistasis of a partition is defined to be the weighted sum of the epistases of the subsets, where each weight corresponds to the relative significance of a subset.

Definition 1 (Internal Epistasis). *Let $\pi = \{V_1, V_2, \ldots, V_q\}$ be a partition of a variable set $V \subseteq \mathcal{V}, |V| \geq 1$. The internal epistasis $\zeta(\pi)$ of π is defined as follows:*

$$\zeta(\pi) = \sum_{i=1}^{q} \frac{\xi(V_i)\varepsilon(V_i)}{\xi(V)}, \tag{8}$$

which is rewritten as

$$\zeta(\pi) = \sum_{i=1}^{q} \frac{I(X_{V_i}; Y) - \sum_{v \in V_i} I(X_v; Y)}{I(X_V; Y)} \tag{9}$$

when $I(X_V; Y) \neq 0$.

The value of internal epistasis is bounded, which is verified in the following proposition.

Proposition 1. *Let $\pi = \{V_1, V_2, \ldots, V_q\}$ be a partition of a variable set $V = \{v_1, v_2, \ldots, v_k\} \subseteq \mathcal{V}, k \geq 1$. The internal epistasis $\zeta(\pi)$ of π satisfies the following inequality:*

$$q - k \leq \zeta(\pi) \leq q. \tag{10}$$

Proof. Omitted by space limitation [14]. □

The internal epistasis has a nonnegative value if the universe contains the whole combinations of the alphabets.

Proposition 2. *Let $\pi = \{V_1, V_2, \ldots, V_q\}$ be a partition of a variable set $V \subseteq \mathcal{V}, |V| \geq 1$. If the universe $\mathcal{U}$ of a problem is defined to be $\mathcal{A}_1 \times \mathcal{A}_2 \times \cdots \times \mathcal{A}_n$, then the internal epistasis $\zeta(\pi)$ of π is nonnegative, i.e.,*

$$0 \leq \zeta(\pi) \leq 1. \tag{11}$$

Proof. Omitted by space limitation [14]. □

The external epistasis of a partition is defined similarly to the epistasis of a variable set. From (2), the significance of a variable set $V \subseteq \mathcal{V}$ is denoted by $\xi(V)$, and the significance of each subset V_i in a partition π of V is denoted by $\xi(V_i)$. We define the external epistasis of π to be the difference between $\xi(V)$ and the summation of all $\xi(V_i)$'s, $V_i \in \pi$.

Definition 2 (External Epistasis). *Let $\pi = \{V_1, V_2, \ldots, V_q\}$ be a partition of a variable set $V \subseteq \mathcal{V}, |V| \geq 1$. The* external epistasis *$\theta(\pi)$ of π is defined as follows:*

$$\theta(\pi) = \begin{cases} \dfrac{\xi(V) - \sum_{i=1}^{q} \xi(V_i)}{\xi(V)} & \text{if } I(X_V; Y) \neq 0 \\ 0 & \text{otherwise,} \end{cases} \tag{12}$$

which is rewritten as

$$\theta(\pi) = \frac{I(X_V;Y) - \sum_{i=1}^{q} I(X_{V_i};Y)}{I(X_V;Y)} \tag{13}$$

when $I(X_V;Y) \neq 0$.

The value of external epistasis is also bounded as in the following proposition.

Proposition 3. *Let* $\pi = \{V_1, V_2, \ldots, V_q\}$ *be a partition of a variable set* $V \subseteq \mathcal{V}, |V| \geq 1$. *The* external epistasis $\theta(\pi)$ *of* π *satisfies the following inequality:*

$$1 - q \leq \theta(\pi) \leq 1. \tag{14}$$

Proof. Omitted by space limitation [14]. □

The external epistasis has a nonnegative value like the internal epistasis if the universe contains the whole combinations of the alphabets.

Proposition 4. *Let* $\pi = \{V_1, V_2, \ldots, V_q\}$ *be a partition of a variable set* $V \subseteq \mathcal{V}, |V| \geq 1$. *If the universe* $\mathcal{U}$ *of a problem is defined to be* $\mathcal{A}_1 \times \mathcal{A}_2 \times \cdots \times \mathcal{A}_n$, *then the external epistasis* $\theta(\pi)$ *of* π *is nonnegative, i.e.,*

$$0 \leq \theta(\pi) \leq 1. \tag{15}$$

Proof. Omitted by space limitation [14]. □

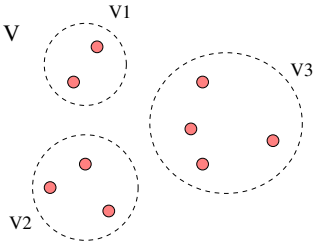
The summation of the internal epistasis and external epistasis of a partition is exactly the same as the epistasis of the variable set.

Theorem 1 (Internal/External Epistasis). *The following holds for a partition* π *of a variable set* $V \subseteq \mathcal{V}, |V| \geq 1$.

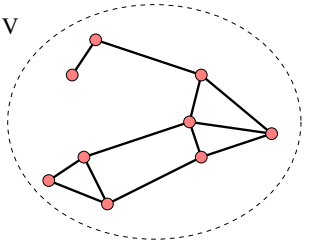
$$\varepsilon(V) = \theta(\pi) + \zeta(\pi) \tag{16}$$

Proof. Omitted by space limitation [14]. □

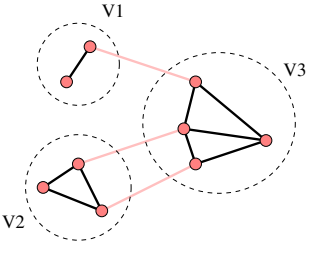
The theorem shows that the internal epistasis and the external epistasis of a partition can be interpreted as the intra-partition epistasis and the inter-partition epistasis, respectively. Figure 1 shows an illustration of the relationship between the epistasis, the internal epistasis, and the external epistasis. Since the epistasis is constant with a given variable set, the internal epistasis and the external epistasis are competitive with each other, i.e., the maximality of the internal epistasis implies the minimality of the external epistasis. This property is utilized in Section 3.2.



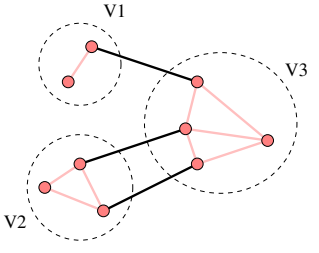
(a) A partition $\pi = \{V_1, V_2, V_3\}$.



(b) The epistasis of V.



(c) The internal epistasis of π.



(d) The external epistasis of π.

Fig. 1. An illustration of the internal epistasis and the external epistasis of a partition. A partition $\pi = \{V_1, V_2, V_3\}$ of a variable set V is shown denoting the epistatic variable pairs by edge connections.

3 Problem Decomposition

3.1 Decomposable Problem

The decomposability of a problem is formally defined in this section. At first, we define a schema as follows:

Definition 3 (Schema). *Let $V = \{v_1, v_2, \ldots, v_k\} \subseteq \mathcal{V}$ be an ordered set. A pair (V, b) is said to be a schema if $b \in \mathcal{A}_{v_1} \times \mathcal{A}_{v_2} \cdots \mathcal{A}_{v_k}$.*

The schema "$*10**$" expressed in Holland's notation [15], for example, is represented as $(\{2, 3\}, (1, 0))$.

The conjunction of a number of schemata, whose variable sets are disjoint with each other, is defined to be a schema composed of the union of the variable sets and the union of the assignments.

Definition 4 (Schema Conjunction). *Let $\pi = \{V_1, V_2, \ldots, V_q\}$ be a partition of $V \subseteq \mathcal{V}$, and (V_i, b_i) be a schema for each $i = 1, 2, \ldots, q$. A schema (V, c) is said to be a conjunction of (V_i, b_i)'s if $c_{V_i} = b_i$ for all $i = 1, 2, \ldots, q$.*

For example, given two schemata, $(\{1,2\},(1,1))$ and $(\{4,5\},(0,0))$, which are expressed as "11∗∗∗" and "∗∗∗00" in Holland's notation, respectively, the conjunction of the two schemata is defined to be $(\{1,2,4,5\},(1,1,0,0))$, which is expressed as "11∗00".

A solution $e \in \mathcal{U}$ is said to be an instance of a schema (V,b) if the solution contains the schema, i.e., $e_V = b$. Given $y \in \mathcal{F}$, a schema is said to be instantiable for y if there exists a solution which is an instance of the schema and whose fitness is y.

Definition 5 (Instantiable Schema). *A schema (V,b) is said to be instantiable for y if there exists a solution $e \in \mathcal{U}$ such that $e_V = b$ and $f(e) = y$.*

An optimization is in a sense the process of finding desirable schemata. Given a partition of the variables, we can individually optimize each subset of the variables if we can guarantee that the conjunction of the optimal schemata corresponding to the subsets is also optimal. In this context, we can define a decomposable problem as follows:

Definition 6 (Schema Independence Condition). *Let $\pi = \{V_1, V_2, \ldots, V_q\}$ be a partition of $\mathcal{V}$. The following statement is defined to be a schema independence condition for π and $y \in \mathcal{F}$.*

"If there exists an instantiable schema for y for each $V_i, i = 1, 2, \ldots, q$, then the conjunction of the schemata is also instantiable for y."

Definition 7 (Decomposable Problem). *Let $\pi = \{V_1, V_2, \ldots, V_q\}$ be a partition of $\mathcal{V}$. A problem is said to be decomposable with π if the schema independence condition is satisfied for $y = \max \mathcal{F}$.*

Unfortunately, we can not always assure of the optimality of the partial solutions obtained from the individual optimizations of the decomposed subproblems. Thus the schema independence condition only for the maximum $y \in \mathcal{F}$ do not guarantee that the problem is possibly solved independently. This is the reason why it is necessary to define the decomposability of a problem with more strong conditions. Hence we define a strongly decomposable problem as follows:

Definition 8 (Strongly Decomposable Problem). *Let $\pi = \{V_1, V_2, \ldots, V_q\}$ be a partition of $\mathcal{V}$. A problem is said to be strongly decomposable with π if the schema independence condition is satisfied for all $y \in \mathcal{F}$.*

The decomposition of a problem in this paper has different meaning from the concept of decomposition in additively decomposed functions (ADFs) [16]. The strong decomposability of a problem implies that the distributions of the variable subsets in the partition are conditionally independent given the fitness value, while the additive decomposability in ADFs means that the fitness function of a problem can be represented as a summation of subfunctions defined on the variable subsets.

3.2 Decomposition Theorem

The Conditional Independence Theorem and the Problem Decomposition Theorem, the main results of this paper, are shown in this section.

Firstly, we show that the strong decomposability of a problem is equivalent to the conditional independence of the variables given Y.

Theorem 2 (Conditional Independence). *Let $\pi = \{V_1, V_2, \ldots, V_q\}$ be a partition of $\mathcal{V}$. Given a problem of which universe $\mathcal{U}$ is defined to be $\mathcal{A}_1 \times \mathcal{A}_2 \cdots \mathcal{A}_n$, the problem is strongly decomposable with π if and only if X_{V_i}'s are conditionally independent given Y, i.e.,*

$$p(x_{\mathcal{V}}|Y = y) = \prod_{i=1}^{q} p(x_{V_i}|Y = y) \tag{17}$$

for all $y \in \mathcal{F}$.

Proof. Omitted by space limitation [14]. □

Secondly, we show that the strong decomposability of a problem with a partition is equivalent to the zero external epistasis of the partition.

Theorem 3 (Problem Decomposition). *Let $\pi = \{V_1, V_2, \ldots, V_q\}$ be a partition of $\mathcal{V}$. Given a problem of which universe $\mathcal{U}$ is defined to be $\mathcal{A}_1 \times \mathcal{A}_2 \cdots \mathcal{A}_n$, the problem is strongly decomposable with π if and only if the external epistasis $\theta(\pi)$ of π is zero.*

Proof. Omitted by space limitation [14]. □

We have the following corollary to Theorem 3.

Corollary 1. *Let $\pi = \{V_1, V_2, \ldots, V_q\}$ be a partition of $\mathcal{V}$. Given a problem of which universe $\mathcal{U}$ is defined to be $\mathcal{A}_1 \times \mathcal{A}_2 \cdots \mathcal{A}_n$, the problem is decomposable with π if the external epistasis $\theta(\pi)$ of π is zero.*

Proof. Omitted by space limitation [14]. □

Since the minimal external epistasis means the maximal internal epistasis by Theorem 1, the strong decomposability of a problem is equivalent to the maximality of the internal epistasis. Consequently, the previous theorems show that the internal epistasis and the external epistasis are capable of quantifying the decomposability of a problem with a given partition of variables.

4 An Example

We tested the proposed measures on a well-known problem, Royal Road function. This example is just for more concrete explanation of the measures and the concepts previously mentioned, not for providing a new optimization method.

Table 1. A decomposition example for the Royal Road function. Given a Royal Road function R, we can obtain three subproblem pairs (R_{11}, R_{12}), (R_{21}, R_{22}), and (R_{31}, R_{32}) from decomposing R with partitions $\pi_1 = \{\{1,2,3,4\},\{5,6,7,8\}\}$, $\pi_2 = \{\{1,2,3,8\},\{4,5,6,7\}\}$, and $\pi_3 = \{\{1,3,5,7\},\{2,4,6,8\}\}$, respectively.

Problem	Schema s_i	Coefficient c_i
	1 2 3 4 5 6 7 8	
R	1 1 * * * * * *	2
	* * 1 1 * * * *	2
	* * * * 1 1 * *	2
	* * * * * * 1 1	2
R_{11}	1 1 * * * * * *	2
	* * 1 1 * * * *	2
R_{12}	* * * * 1 1 * *	2
	* * * * * * 1 1	2
R_{21}	1 1 * * * * * *	2
R_{22}	* * * * 1 1 * *	2
R_{31}	–	–
R_{32}	–	–

The Royal Road function is suitable for this kind of test since it has explicit building blocks.

The Royal Road functions are special functions proposed by Forrest and Mitchell [17] to investigate how schema processing actually takes place inside evolutionary algorithms. To do so, the function was designed to have obvious building blocks and an optimal solution. A Royal Road function is defined as follows:

$$f(x_1, x_2, \ldots, x_n) = \sum_i c_i \delta_i(x_i, x_2, \ldots, x_n) \tag{18}$$

where c_i is a predefined coefficient corresponding to a schema s_i, and $\delta_i : \{0,1\}^n \to \{0,1\}$ is a function that returns 1 if the solution contains the schema s_i, and returns 0 otherwise. Generally, the coefficient c_i is defined to be equal to the order of schema s_i.

Table 1 shows decomposition examples for the Royal Road function. In the table, a Royal Road function R, which contains four order-2 building blocks, is decomposed into (R_{11}, R_{12}), (R_{21}, R_{22}), and (R_{31}, R_{32}), respectively, with three different 2-way partitions $\pi_1 = \{\{1,2,3,4\}, \{5,6,7,8\}\}$, $\pi_2 = \{\{1,2,3,8\}, \{4,5,6,7\}\}$, and $\pi_3 = \{\{1,3,5,7\}, \{2,4,6,8\}\}$. It is notable that the subproblems (R_{11}, R_{12}) have more building blocks than (R_{21}, R_{22}), which have more building blocks than (R_{31}, R_{32}). We can solve the decomposed subproblems individually by an exhaustive search algorithm as shown in Table 2.

Table 3 shows the external epistasis $\theta(\cdot)$ and the internal epistasis $\zeta(\cdot)$ of the partitions, and the average fitness of the solutions obtained by joining the mutually exclusive partial solutions in Table 2. We can see that $\theta(\pi_1)$ is less than $\theta(\pi_2)$ and $\theta(\pi_2)$ is less than $\theta(\pi_3)$. By Theorem 3, it follows that the function

Table 2. The partial solutions obtained from solving the Royal Road sub-functions individually.

Subproblem	Partial solution	Subproblem	Partial solution
	1 2 3 4 5 6 7 8		1 2 3 4 5 6 7 8
R_{11}	1 1 1 1	R_{31}	0 0 0 0
R_{12}	1 1 1 1		1 0 0 0
R_{21}	1 1 0 0		0 1 0 0
	1 1 1 0		...
	1 1 0 1		1 1 1 1
	1 1 1 1	R_{32}	0 0 0 0
R_{22}	0 1 1 0		1 0 0 0
	1 1 1 0		0 1 0 0
	0 1 1 1		...
	1 1 1 1		1 1 1 1

Table 3. The epistasis and the average quality of the conjuncted partial solutions of the Royal Road sub-functions.

Partition π	Subproblems	$\epsilon(\mathcal{V})$	$\theta(\pi)$	$\zeta(\pi)$	Quality
$\pi_1 = \{\{1,2,3,4\},\{5,6,7,8\}\}$	$(R_{11},\ R_{12})$	0.712	0.416	0.296	8.000
$\pi_2 = \{\{1,2,3,8\},\{4,5,6,7\}\}$	$(R_{21},\ R_{22})$	0.712	0.524	0.188	4.500
$\pi_3 = \{\{1,3,5,7\},\{2,4,6,8\}\}$	$(R_{31},\ R_{32})$	0.712	0.580	0.132	2.000

R is more decomposable with π_1 than with π_2, and more decomposable with π_2 than with π_3. This is consistent with the fact that π_1 preserves more building blocks than π_2 which preserves more building blocks than π_3. These results also agree with the fact that the average quality corresponding to π_1 is greater than that of π_2, and the average quality corresponding to π_2 is greater than that of π_3. The function R is not strongly decomposable with any of the example partitions since neither of them has external epistasis zero, although it is decomposable with all of them by the schema independence condition for $y = \max \mathcal{F} = 8$.

It is notable that the tests conducted on other combinatorial optimization problems, such as the MAXSAT problem [18], showed consistent results with the case of the Royal Road function.

5 Conclusions

We showed that the epistasis can be factorized into the internal epistasis and the external epistasis, and these new measures provide strong evidences for the decomposability of a problem. The theorems and the experimental results on an example function showed that the proposed measures were well defined and consistent with other properties of the function. We believe that the contribution of this paper is not merely providing new measures of the decomposability but also presenting a way to in-depth understanding about the epistatic behaviors of the

variables in combinatorial optimization problems. Future work includes applying these measures to the design of efficient hierarchical optimization methods.

Acknowledgments. This work was supported by Brain Korea 21 Project. The ICT at Seoul National University provided research facilities for this study.

References

1. E. H. Aarts and J. K. Lenstra. Introduction. In *Local Search in Combinatorial Optimization*, pages 1–17. John Wiley & Sons, 1997.
2. C. H. Papadimitriou and K. Steiglitz. *Combinatorial Optimization: Algorithms and Complexity*. Dover Publications, 1998.
3. D. I. Seo and B. R. Moon. A survey on chromosomal structures and operators for exploiting topological linkages of genes. In *Genetic and Evolutionary Computation Conference*, 2003.
4. M. Pelikan, D. E. Goldberg, and F. Lobo. A survey of optimization by building and using probabilistic models. *Computational Optimization and Applications*, 21(1):5–20, 2002.
5. P. Larrañaga and J. A. Lozano. *Estimation of Distribution Algorithms: A New Tool for Evolutionary Computation*. Kluwer Academic Publishers, 2002.
6. S. L. Lauritzen. Graphical models. Oxford: Clarendon Press, 1996.
7. B. J. Fery. *Graphical Models for Machine Learning and Digital Communication*. Cambridge: MIT Press, 1998.
8. Y. Davidor. Epistasis variance: Suitability of a representation to genetic algorithms. *Complex Systems*, 4:369–383, 1990.
9. D. I. Seo, Y. H. Kim, and B. R. Moon. New entropy-based measures of gene significance and epistasis. In *Genetic and Evolutionary Computation Conference*, 2003.
10. D. I. Seo, S. S. Choi, Y. H. Kim, and B. R. Moon. New epistasis measures based on Shannon's entropy for combinatorial optimization problems. in preparation, 2004.
11. C. R. Reeves and C. C. Wright. An experimental design perspective on genetic algorithms. In *Foundations of Genetic Algorithms 3*, pages 7–22. 1995.
12. C. E. Shannon. A mathematical theory of communication. *Bell System Technical Journal*, 27:379–423, 623–656, 1948.
13. T. Cover and J. Thomas. *Elements of Information Theory*. John Wiley & Sons, 1991.
14. D. I. Seo, S. S. Choi, and B. R. Moon. The epistasis and problem decomposition for combinatorial optimization problems. in preparation, 2004.
15. J. Holland. *Adaptation in Natural and Artificial Systems*. The University of Michigan Press, 1975.
16. H. Mühlenbein and T. Mahnig. FDA — A scalable evolutionary algorithm for the optimization of additively decomposed functions. *Evolutionary Computation*, 7(4):353–376, 1999.
17. S. Forrest and M. Mitchell. Relative building-block fitness and the building-block hypothesis. In *Foundations of Genetic Algorithms 2*, pages 109–126. 1993.
18. M. R. Garey and D. S. Johnson. *Computers and Intractability: A Guide to the Theory of NP-Completeness*. Freeman, 1979.