An Information-Theoretic Analysis on the Interactions of Variables in Combinatorial Optimization Problems

Dong-Il Seo diseo@soar.snu.ac.kr
Byung-Ro Moon moon@soar.snu.ac.kr
School of Computer Science & Engineering, Seoul National University, Sillim-dong, Gwanak-gu, Seoul, 151-744 Korea

Abstract

In optimization problems, the contribution of a variable to fitness often depends on the states of other variables. This phenomenon is referred to as epistasis or linkage. In this paper, we show that a new theory of epistasis can be established on the basis of Shannon's information theory. From this, we derive a new epistasis measure called entropic epistasis and some theoretical results. We also provide experimental results verifying the measure and showing how it can be used for designing efficient evolutionary algorithms.

Keywords

Combinatorial optimization, entropy, mutual information, variable interaction, linkage, entropic epistasis.

1 Introduction

An optimization problem is usually encoded into a function called *fitness function* from a set called *universe* to the real numbers $\mathbb{R}$. The universe is the set of feasible solutions which is often represented by a number of variables each of which has its own domain. For decades, the hardness of optimization problems has been studied from various viewpoints focusing on diverse characteristics in the evolutionary computation community. Those characteristics include deception (Goldberg, 1987), multimodality (Horn and Goldberg, 1995), noise (Kargupta, 1995), and epistasis (Holland, 1992; Davidor, 1990). *Epistasis*, also referred to as *linkage*, is one representative characteristic observed in most non-trivial optimization problems. The term epistasis came from biology where it means the suppression of gene expression by one or more other genes. However, in the evolutionary computation community, the term has a somewhat different meaning; it represents the phenomenon whereby the contribution of a variable to the fitness depends on the states of other variables. There is no explicit suppression in this case. Nevertheless we can understand the change of the fitness resulting from the assignment of a variable as an expression and thus the influence of other variables on the change as suppression.

The idea of linkage-based problem solving has been realized in various branches of evolutionary algorithms, such as estimation-of-distribution algorithms (EDAs) (Larrañaga and Lozano, 2002; Pelikan et al., 2002) and topological linkage-based genetic algorithms (TLBGAs) (Seo and Moon, 2003). Many of the methods adopted in

the algorithms for the estimation of epistasis are problem-dependent or algorithm-dependent as the required epistasis information is composed using problem-specific knowledge or extracted from a set of solutions temporarily stored in each generation.

Linkage group detection (Munetomo and Goldberg, 1999; Heckendorn and Wright, 2004; Streeter, 2004) is an example of the problem-independent and algorithm-independent method for the estimation of the epistasis. In those approaches, pseudo-Boolean functions[1] are solved by decomposing the variable set into distinct subsets called *linkage groups*. A linkage group is a minimal set of variables such that no variable in it is linked to (nor has epistasis with) any variable outside it. Thus the variables in a linkage group can be optimized independently of other variables.

The Walsh transform (Bethke, 1981; Goldberg, 1989a; Goldberg, 1989b) is one of the classic problem-independent and algorithm-independent tools for analyzing the epistasis of pseudo-Boolean functions. Given a pseudo-Boolean function of n variables, the Walsh transform generates 2^n coefficients called *partition coefficients* from 2^n fitness values. Each coefficient corresponds to the epistasis of one of 2^n schemata, which mean specific patterns of variable assignments. Recently, it was shown that the partition coefficients of MAXSAT can be computed in linear time to the number of clauses (Rana et al., 1998), and a general algorithm that detects the linkage groups and computes the partition coefficients concurrently was proposed (Heckendorn and Wright, 2004). It is notable that the complete knowledge of all partition coefficients does not directly mean an ability to solve the problem.

Epistasis can be explained in the light of experimental design theory (Reeves and Wright, 1995a; Reeves and Wright, 1995b). According to the theory, the contribution of the variables to the fitness is decomposed into a number of coefficients called *effects*. The effects are classified into two categories, *linear effects* and *interaction effects*, where the latter corresponds to epistasis. It was shown that there is a one-to-one correspondence between these effects and the partition coefficients of the Walsh analysis.

The epistasis variance (Davidor, 1990; Reeves and Wright, 1995b) is a general measure of the epistasis. This measure gauges the nonlinearity lying in the fitness landscape by quantifying the portion of the fitness variation due to the interaction effects of the experimental design analysis. The underlying assumption here is that the term "contribution" in defining the epistasis can be interpreted as the algebraic, or scalar influence of variables on the fitness. This interpretation seems to be natural and is widely accepted nowadays. However, we claim that the interpretation is not the only one. Actually, this study aims to show that a theory of epistasis can be established with a different interpretation of the term.

In this paper, we attempt to interpret the term as the probabilistic, or entropic influence of the variables on the fitness based on Shannon's information theory (Shannon, 1948; Cover and Thomas, 1991). This interpretation leads us to a measure of the contribution called *significance*. From this measure, we can derive a new epistasis measure called *entropic epistasis*. This is a microscopic measure in that with it, the interactions in a variable group can be focused on. To deal with the overall aspect of a fitness function, we derive two additional measures, *mean significance* and *mean entropic epistasis*. The proposed measures are problem-independent and can be used in either an algorithm-independent or -dependent way in evolutionary algorithms. We present some theoretical results on the measures including comparisons with traditional epistasis measures. Also, we provide experimental results verifying the measures and showing how they can be used for designing efficient evolutionary algorithms. This study is an exten-

[1] A fitness function is said to be pseudo-Boolean if it is defined on $\{0,1\}^n$ for some $n \geq 1$.

sion to high-order epistasis of our previous work (Seo et al., 2003a), where the measure definition was limited to pairwise epistasis.

The rest of this paper is organized as follows. The preliminary concepts of Shannon's information theory are introduced in Section 2. The new epistasis-related measures are defined and analyzed in Section 3. The results of experiments on three representative optimization problems are presented in Section 4. Finally, concluding remarks are given in Section 5.

2 Entropy and Mutual Information

Shannon's information theory (Shannon, 1948; Cover and Thomas, 1991) concerns the characteristics of the information carried by a sequence of events which is often formulated as one or more random variables. According to the theory, the amount of information contained in a message notifying the occurrence of an event is defined to be the minimal number of digits being required to describe the event. That is, if the event occurs with probability p, the amount of information is defined to be $\log \frac{1}{p}$. The $\log$ function is to the base 2 and the value is measured in bits. The lower the probability of the event is, the larger amount of information the message carries.

The average amount of information contained in an event notification is the amount of uncertainty of the corresponding random variable. Thus the uncertainty of a random variable X is defined to be

$$H(X) = -\sum_{x \in \mathcal{X}} p(x) \log p(x) \tag{1}$$

where $\mathcal{X}$ is the alphabet (or the domain) and $p(\cdot)$ is the probability mass function (pmf). This quantity is called entropy of X. The convention $0 \log 0 = 0$ is used in the equation, which is easily justified by continuity since $x \log x \rightarrow 0$ as $x \rightarrow 0$. The entropy is always nonnegative. Analogously, the joint entropy of two random variables X and Y is defined to be

$$H(X,Y) = -\sum_{x \in \mathcal{X}} \sum_{y \in \mathcal{Y}} p(x,y) \log p(x,y) \tag{2}$$

where $\mathcal{X}$ and $\mathcal{Y}$ are the alphabets and $p(\cdot,\cdot)$ is the joint pmf. The conditional entropy of X given Y is defined to be

$$\begin{aligned} H(X|Y) &= -\sum_{y \in \mathcal{Y}} p(y) H(X|Y=y) && (3) \\ &= -\sum_{y \in \mathcal{Y}} \sum_{x \in \mathcal{X}} p(x,y) \log p(x|Y=y). && (4) \end{aligned}$$

This quantity represents the average uncertainty of X when the value of Y is known. We can show that

$$H(X|Y) = H(X,Y) - H(Y). \tag{5}$$

We can also show that $H(X|Y) \leq H(X)$, which means that conditioning reduces entropy.

The uncertainty of X reduced by knowing the value of Y is the amount of information about X carried by Y. This quantity is called mutual information between X

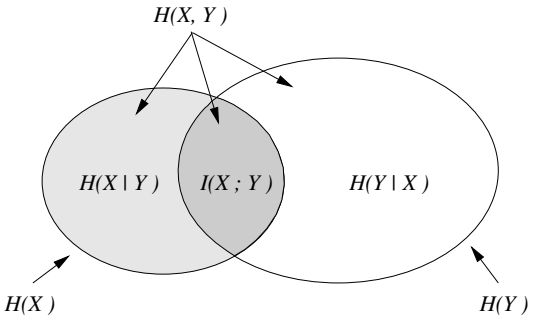


Figure 1: The relationship between the entropy and the mutual information.

and Y, and is formally written as

$$I(X;Y) = H(X) - H(X|Y) \quad (6)$$

$$= \sum_{x\in\mathcal{X}}\sum_{y\in\mathcal{Y}} p(x,y)\log\frac{p(x,y)}{p(x)p(y)}. \quad (7)$$

This is a symmetric and nonnegative measure which has various meanings. It indicates the degree of predictability of one random variable using the other random variable. Also, it is a general measure of dependency between two random variables. That is, the larger the mutual information is, the stronger the dependence between the random variables is. A family of discrete random variables is said to be independent if the joint pmf of any variable subset can be written as the product of the corresponding marginal pmf's, and it is said to be conditionally independent given random variables if the conditional joint pmf of any variable subset can be written as the product of the corresponding marginal conditional pmf's. The mutual information between two random variables has value zero if and only if they are independent. From (5) and (6), we obtain that

$$I(X;Y) = H(X) + H(Y) - H(X,Y). \quad (8)$$

This formula states that two random variables are independent if and only if the joint entropy of them is equal to the sum of the two marginal entropies. Generally, a family of random variables is independent if and only if the joint entropy of the variables is equal to the sum of the marginal entropies (see the Independence Bound on entropy in Appendix A). Figure 1 shows an illustration of the relationship between entropy and mutual information. We append a number of frequently used theorems as Appendix A.

3 New Epistasis Measure

3.1 Dependency between Random Variables

A family of discrete random variables is said to be fully-dependent if the information about any of them is sufficient for predicting all random variables in the family. This is formulated as follows:

Definition 1 *Random variables $X_1, \ldots, X_n$ are said to be fully-dependent if there exists a function $g_{ij} : \mathcal{X}_i \to \mathcal{X}_j$ such that $X_j = g_{ij}(X_i)$ for each pair (i, j).*

This is equivalent to that $H(X_1, \ldots, X_n|X_i) = 0$ for all i (see the proof of Theorem 2). We define conditional full-dependence analogously:

Definition 2 *Let $X_1, \ldots, X_n$ and Y be random variables. $X_1, \ldots, X_n$ are said to be conditionally fully-dependent given Y if there exists a function $g_{ij} : \mathcal{Y} \times \mathcal{X}_i \to \mathcal{X}_j$ such that $X_j = g_{ij}(Y, X_i)$ for each pair (i, j).*

This is equivalent to that $H(X_1, \ldots, X_n|Y, X_i) = 0$ for all i (see the proof of Theorem 4).

As mentioned in Section 2, there is a consensus that the mutual information is a general measure of the dependence between two random variables. Recall that the mutual information between random variables X_1 and X_2 is written as

$$I(X_1; X_2) = H(X_1) + H(X_2) - H(X_1, X_2).$$

This formula states that mutual information is defined to be the difference between the sum of marginal entropies and the joint entropy. By the Independence Bound on entropy (see Appendix A), $H(X_1) + H(X_2)$ has a minimum value $H(X_1, X_2)$ for fixed $H(X_1, X_2)$. This happens when X_1 and X_2 are independent. Also, since $H(X_1) \leq H(X_1, X_2)$ and $H(X_2) \leq H(X_1, X_2)$ by the Chain Rule for entropy (see Appendix A), $H(X_1) + H(X_2)$ has a maximum value $2H(X_1, X_2)$ for fixed $H(X_1, X_2)$. We can show that this happens when X_1 and X_2 are fully-dependent (see the proof of Theorem 2). Based on these observations, we can extend the mutual information to a general measure of the dependence among more than two random variables:

Definition 3 (Dependency) *Given random variables $X_1, \ldots, X_n$, the dependency $D(X_1, \ldots, X_n)$ is defined to be*

$$D(X_1, \ldots, X_n) = \sum_{i=1}^{n} H(X_i) - H(X_1, \ldots, X_n). \tag{9}$$

We can obtain the following theorems regarding the range of the measure:

Theorem 1 *Given random variables $X_1, \ldots, X_n$, $D(X_1, \ldots, X_n)$ has a minimum value zero for fixed $H(X_1, \ldots, X_n)$ if and only if X_i are independent.*

Proof. This is immediate from the Independence Bound on entropy (see Appendix A). □

Theorem 2 *Let $X_1, \ldots, X_n$ be random variables with $p(x_i) > 0$ for all $x_i \in \mathcal{X}_i$. $D(X_1, \ldots, X_n)$ has a maximum value $(n-1)H(X_1, \ldots, X_n)$ for fixed $H(X_1, \ldots, X_n)$ if and only if X_i are fully-dependent.*

Proof. By the Chain Rule for entropy (see Appendix A), $H(X_i) \leq H(X_1, \ldots, X_n)$ for all i. Thus $D(X_1, \ldots, X_n) \leq (n-1)H(X_1, \ldots, X_n)$. Now, suppose that $D(X_1, \ldots, X_n) = (n-1)H(X_1, \ldots, X_n)$. Then

$$\sum_{i=1}^{n} H(X_i) = H(X_1, \ldots, X_n) + D(X_1, \ldots, X_n) = nH(X_1, \ldots, X_n).$$

Since $H(X_i) \leq H(X_1, \ldots, X_n)$ for all i,

$$H(X_i) = H(X_1, \ldots, X_n)$$

for $i = 1, \ldots, n$. Thus

$$H(X_2, \ldots, X_n|X_1) = H(X_1, \ldots, X_n) - H(X_1) = 0.$$

This can be rewritten as

$$H(X_2, \ldots, X_n|X_1) = \sum_{x_1 \in \mathcal{X}_1} p(x_1) H(X_2, \ldots, X_n|X_1 = x_1) = 0.$$

Since $p(x_1) > 0$ and $H(X_2, \ldots, X_n|X_1 = x_1) \geq 0$ for all $x_1 \in \mathcal{X}_1$,

$$H(X_2, \ldots, X_n|X_1 = x_1) = 0$$

for all $x_1 \in \mathcal{X}_1$. By the fact that $H(Z) = 0$ is equivalent to that Z is constant in general, $X_2, \ldots, X_n$ are constant under conditioning $X_1 = x_1$ for all $x_1 \in \mathcal{X}_1$. Thus, with the trivial function $g_{ii}(x) = x$, there exists a function $g_{1j} : \mathcal{X}_1 \to \mathcal{X}_j$ such that $X_j = g_{1j}(X_1)$ for $j = 1, 2, \ldots, n$. We can prove the existence of g_{ij} for $i = 2, 3, \ldots, n$ analogously.

The converse is proved by reversing the order of the arguments. □

We define conditional dependency analogously:

Definition 4 (Conditional Dependency) *Given random variables $X_1, \ldots, X_n$ and Y, the conditional dependency $D(X_1, \ldots, X_n|Y)$ is defined to be*

$$D(X_1, \ldots, X_n|Y) = \sum_{i=1}^{n} H(X_i|Y) - H(X_1, \ldots, X_n|Y). \tag{10}$$

We can obtain the following theorems similar to those of dependency:

Theorem 3 *Given random variables $X_1, \ldots, X_n$ and Y, $D(X_1, \ldots, X_n|Y)$ has a minimum value zero for fixed $H(X_1, \ldots, X_n|Y)$ if and only if X_i are conditionally independent given Y.*

Proof. This is immediate from the Independence Bound on conditional entropy (see Appendix A). □

Theorem 4 *Let $X_1, \ldots, X_n$ and Y be random variables with $p(x_i, y) > 0$ for all $x_i \in \mathcal{X}_i$ and $y \in \mathcal{Y}$. $D(X_1, \ldots, X_n|Y)$ has a maximum value $(n-1)H(X_1, \ldots, X_n|Y)$ for fixed $H(X_1, \ldots, X_n|Y)$ if and only if X_i are conditionally fully-dependent given Y.*

Proof. By arguments similar to the proof of Theorem 2, we obtain that

$$H(X_2, \ldots, X_n|Y = y, X_1 = x_1) = 0$$

for all $y \in \mathcal{Y}$ and $x_1 \in \mathcal{X}_1$. By the fact that $H(Z) = 0$ is equivalent to that Z is constant in general, $X_2, \ldots, X_n$ are constant under conditioning $Y = y$ and $X_1 = x_1$ for all $y \in \mathcal{Y}$ and $x_1 \in \mathcal{X}_1$. Thus, with the trivial function $g_{ii}(y, x) = x$, there exists a function $g_{1j} : \mathcal{Y} \times \mathcal{X}_1 \to \mathcal{X}_j$ such that $X_j = g_{1j}(Y, X_1)$ for $j = 1, 2, \ldots, n$. We can prove the existence of g_{ij} for $i = 2, 3, \ldots, n$ analogously.

The converse is proved by reversing the order of the arguments. □

Dependency and conditional dependency will be used for analyzing a new epistasis measure in Section 3.3.

3.2 Probability Space

In this section, we construct probability spaces where new epistasis-related measures will be defined.

Consider an optimization problem instance specified by $(\mathcal{U}, f)$, where $\mathcal{U}$ is the universe and f is the fitness function. Let $\mathcal{V} = \{1, 2, \ldots, n\}$ be the variable indices[2] and $\mathcal{A}_v, v \in \mathcal{V}$ be the alphabet of the v^{th} variable x_v. Then $\mathcal{U} \subseteq \mathcal{A}_1 \times \cdots \times \mathcal{A}_n$. In this paper, we use a variable set and its corresponding index set interchangeably. Also, we use a conventional notation x_V to denote $(x_{v_1}, x_{v_2}, \ldots, x_{v_k})$ for a variable set $V = \{v_1, v_2, \ldots, v_k\} \subseteq \mathcal{V}$. We assume that the alphabet of each variable is finite. Then the set of all fitness values $\mathcal{F} \subset \mathbb{R}$ is also finite as the universe is finite.

Assume that we are given a *sampling model* where each solution is sampled at some rate from the universe. That is, assume that a *sampling rate function* $r : \mathcal{U} \to \mathbb{R}$ satisfying

$$r(x_\mathcal{V}) \geq 0, \quad x_\mathcal{V} \in \mathcal{U} \quad \text{and} \tag{11}$$

$$\sum_{x_\mathcal{V} \in \mathcal{U}} r(x_\mathcal{V}) = 1 \tag{12}$$

is given. By this sampling model, a probability space is automatically constructed; a random variable X_v for each variable $v \in \mathcal{V}$ and a random variable Y for the fitness are defined with the joint pmf $p_r : \mathcal{A}_1 \times \cdots \times \mathcal{A}_n \times \mathcal{F} \to \mathbb{R}$ defined to be

$$p_r(x_\mathcal{V}, y) = \begin{cases} r(x_\mathcal{V}) & \text{if } x_\mathcal{V} \in \mathcal{U} \text{ and } y = f(x_\mathcal{V}) \\ 0 & \text{otherwise.} \end{cases} \tag{13}$$

Note that the joint pmf is defined not on $\mathcal{U}$ but on $\mathcal{A}_1 \times \cdots \times \mathcal{A}_n \times \mathcal{F}$. Using an indicator function $1(\cdot)$ defined as $1(\mathbf{true}) = 1$ and $1(\mathbf{false}) = 0$, we can rewrite (13) as

$$p_r(x_\mathcal{V}, y) = r(x_\mathcal{V}) 1(x_\mathcal{V} \in \mathcal{U}) 1(y = f(x_\mathcal{V})). \tag{14}$$

In this paper, the notations with a subscript r such as p_r, H_r, I_r, and D_r, indicate that they are defined on the probability space constructed with a sampling rate function denoted by r. This is, in some sense, a two-stage sampling since it is assumed that the sampler already has the full information on the universe including the fitness of solutions.

Consider the sampling models where the sampling rate of a solution is determined by its fitness, i.e., there is a function $w : \mathcal{F} \to \mathbb{R}$ such that $r(x_\mathcal{V}) = w(f(x_\mathcal{V})), x_\mathcal{V} \in \mathcal{U}$ for the sampling rate function r. Such sampling rate functions are said to be *based on fitness*. A special sampling rate function that has a constant value is said to be *uniform*. By (12), there is only one such function $u(x_\mathcal{V}) = \frac{1}{|\mathcal{U}|}, x_\mathcal{V} \in \mathcal{U}$. The uniform sampling model is a special case of the fitness-based sampling model. We give an individual notation u to the uniform sampling rate function to differentiate it from others. The fitness-based sampling models are natural and convincing since the selection operators and the replacement policies used in the evolutionary algorithms are basically based on the fitness of solutions. As it is not easy to say something about general sampling models, we concentrate on the fitness-based sampling models, particularly on the uniform sampling model in this paper.

We can use a sample set instead of the universe in the probability space construction for reducing the time/space complexity. In this case, the selection of the sample set corresponds to the first stage of the two-stage sampling and the selection of solutions from the sample set in accordance with the sampling rate function corresponds to the second-stage. The size of the sample set cannot be too small to obtain results with low distortion in this case. One general method for the first-stage sampling is to use

[2] The variables are labeled with indices 1 through n starting from the left-most position.

random solutions. For the algorithm-dependent estimation of the epistasis in an evolutionary algorithm, the population can be used as the sample set. Further discussion on this will be given in Section 3.5.

The following theorem shows that the random variables $X_v, v \in \mathcal{V}$ are independent when the universe is the Cartesian product of the alphabets and the sampling model is uniform.

Theorem 5 *Let u be the uniform sampling rate function. Given a problem with universe $\mathcal{U} = \mathcal{A}_1 \times \cdots \times \mathcal{A}_n$,*

$$D_u(X_V) = 0 \tag{15}$$

for any nonempty variable set $V \subseteq \mathcal{V}$.

Proof. From (14),

$$\begin{aligned} p_u(x_V) &= \sum_{z_\mathcal{V} \in \mathcal{A}_1 \times \cdots \times \mathcal{A}_n} \sum_{y \in \mathcal{F}} p_u(z_\mathcal{V}, y) 1(z_V = x_V) \\ &= \sum_{z_\mathcal{V} \in \mathcal{A}_1 \times \cdots \times \mathcal{A}_n} \sum_{y \in \mathcal{F}} u(z_\mathcal{V}) 1(z_\mathcal{V} \in \mathcal{U}) 1(y = f(z_\mathcal{V})) 1(z_V = x_V) \\ &= \frac{1}{|\mathcal{U}|} \sum_{z_\mathcal{V} \in \mathcal{A}_1 \times \cdots \times \mathcal{A}_n} 1(z_V = x_V) \\ &= \frac{\prod_{v \in \mathcal{V} \setminus V} |\mathcal{A}_v|}{\prod_{v \in \mathcal{V}} |\mathcal{A}_v|} = \prod_{v \in V} \frac{1}{|\mathcal{A}_v|} \end{aligned}$$

for any nonempty variable set $V \subseteq \mathcal{V}$. Similarly,

$$p_u(x_v) = \frac{1}{|\mathcal{A}_v|}$$

for any variable $v \in \mathcal{V}$. Thus

$$p_u(x_V) = \prod_{v \in V} p_u(x_v)$$

for any nonempty variable set $V \subseteq \mathcal{V}$. By Theorem 1, the theorem follows. □

This result is useful as many optimization problems have fitness functions defined on the Cartesian product of the alphabets.

The following theorem explains the relationship between the joint pmf in the uniform sampling model and that in an arbitrary fitness-based sampling model.

Theorem 6 *Let r be a fitness-based sampling rate function with $r(x_\mathcal{V}) = w(f(x_\mathcal{V})), x_\mathcal{V} \in \mathcal{U}$ and u be the uniform sampling rate function. Given a nonempty variable set $V = \{v_1, \ldots, v_k\} \subseteq \mathcal{V}$,*

$$p_r(y) = |\mathcal{U}| w(y) p_u(y) \tag{16}$$

and

$$p_r(x_V | Y = y) = p_u(x_V | Y = y) \tag{17}$$

for all $y \in \mathcal{F}$ and $x_V \in \mathcal{A}_{v_1} \times \cdots \times \mathcal{A}_{v_k}$.

Proof. From (14),

$$\begin{aligned} p_r(y) &= \sum_{z_\mathcal{V} \in \mathcal{A}_1 \times \cdots \times \mathcal{A}_n} p_r(z_\mathcal{V}, y) \\ &= \sum_{z_\mathcal{V} \in \mathcal{A}_1 \times \cdots \times \mathcal{A}_n} w(f(z_\mathcal{V})) 1(z_\mathcal{V} \in \mathcal{U}) 1(y = f(z_\mathcal{V})) \\ &= \sum_{z_\mathcal{V} \in \mathcal{A}_1 \times \cdots \times \mathcal{A}_n} w(y) 1(z_\mathcal{V} \in \mathcal{U}) 1(y = f(z_\mathcal{V})) \\ &= |\mathcal{U}| w(y) \sum_{z_\mathcal{V} \in \mathcal{A}_1 \times \cdots \times \mathcal{A}_n} u(z_\mathcal{V}) 1(z_\mathcal{V} \in \mathcal{U}) 1(y = f(z_\mathcal{V})) \\ &= |\mathcal{U}| w(y) p_u(y) \end{aligned}$$

for all $y \in \mathcal{F}$. From (14) and (16),

$$\begin{aligned} p_r(x_V | Y = y) &= \frac{\sum_{z_\mathcal{V} \in \mathcal{A}_1 \times \cdots \times \mathcal{A}_n} p_r(z_\mathcal{V}, y) 1(z_V = x_V)}{p_r(y)} \\ &= \frac{\sum_{z_\mathcal{V} \in \mathcal{A}_1 \times \cdots \times \mathcal{A}_n} w(f(z_\mathcal{V})) 1(z_\mathcal{V} \in \mathcal{U}) 1(y = f(z_\mathcal{V})) 1(z_V = x_V)}{|\mathcal{U}| w(y) p_u(y)} \\ &= \frac{\sum_{z_\mathcal{V} \in \mathcal{A}_1 \times \cdots \times \mathcal{A}_n} u(z_\mathcal{V}) 1(z_\mathcal{V} \in \mathcal{U}) 1(y = f(z_\mathcal{V})) 1(z_V = x_V)}{p_u(y)} \\ &= \frac{\sum_{z_\mathcal{V} \in \mathcal{A}_1 \times \cdots \times \mathcal{A}_n} p_u(z_\mathcal{V}, y) 1(z_V = x_V)}{p_u(y)} \\ &= p_u(x_V | Y = y) \end{aligned}$$

for all $y \in \mathcal{F}$ and $x_V \in \mathcal{A}_{v_1} \times \cdots \times \mathcal{A}_{v_k}$. □

Notice that $p_u(x_\mathcal{V} | Y = y) = u(x_\mathcal{V}) 1(x_\mathcal{V} \in \mathcal{U}) 1(y = f(x_\mathcal{V})) / p_u(y)$ is constant for all solutions $x_\mathcal{V} \in \mathcal{U}$ with a fixed fitness value $y \in \mathcal{F}$. Thus (17) implies that if the sampling is based on the fitness, all solutions with the same fitness are equally probable.

3.3 Significance and Entropic Epistasis

In this section, we define a new epistasis measure using the information-theoretic measures introduced in Section 2. Note that all probabilistic or entropic quantities are defined on the probability spaces described in Section 3.2.

We define the significance of a variable set as follows.

Definition 5 (Significance) *Let r be a sampling rate function. The significance $\xi_r(V)$ of a nonempty variable set $V \subseteq \mathcal{V}$ is defined to be*

$$\xi_r(V) = \frac{I_r(X_V; Y)}{H_r(Y)}. \tag{18}$$

In general, the proportion $\frac{I(X;Y)}{H(Y)}$ for random variables X and Y is known as the uncertainty coefficient (Theil, 1972) and is used to gauge how strongly Y is predictable with the information about X. The above definition is intuitive and natural since if we can get much information on the fitness from the variables, we can say that the variables

play a significant role in the determination of the fitness. We do not consider the case of constant fitness $|\mathcal{F}| = 1$ where no optimization is required. Thus we can regard the entropy $H_r(Y)$ as a positive value. Note that $I_r(X_V; Y) = 0$ is equivalent to $\xi_r(V) = 0$.

The significance is bounded as follows:

Theorem 7 *Let r be a sampling rate function. For a nonempty variable set $V \subseteq \mathcal{V}$,*

$$0 \leq \xi_r(V) \leq 1. \tag{19}$$

Proof. From that $I_r(X_V; Y) \geq 0$, $\xi_r(V) \geq 0$. Also,

$$\begin{aligned} \xi_r(V) H_r(Y) &= I_r(X_V; Y) \\ &= H_r(Y) - H_r(Y|X_V) \\ &\leq H_r(Y). \end{aligned}$$

□

If the significance is zero, we cannot get any information on the fitness from the corresponding variables. On the other hand, if the significance is 1, we can fully identify the fitness from the variables as they are fully-dependent by Theorem 2. It is clear that $\xi_r(\mathcal{V}) = 1$ as $I_r(X_\mathcal{V}; Y) = H_r(Y)$, and that $\xi_r(V) \leq \xi_r(W)$ for $V \subseteq W \subseteq \mathcal{V}$ as $I_r(X_V; Y) \leq I_r(X_W; Y)$ by the Chain Rule for information (see Appendix A).

The entropic epistasis is derived from the significance. By Definition 5, the significance of a nonempty variable set $V \subseteq \mathcal{V}$ is denoted by $\xi_r(V)$ and the significance of a variable $v \in V$ is denoted by $\xi_r(v)$. We define the entropic epistasis of V to be the difference between $\xi_r(V)$ and the sum of $\xi_r(v)$ for all $v \in V$:

Definition 6 (Entropic Epistasis) *Let r be a sampling rate function. The entropic epistasis $\varepsilon_r(V)$ of a nonempty variable set $V \subseteq \mathcal{V}$ is defined to be*

$$\varepsilon_r(V) = \begin{cases} \dfrac{\xi_r(V) - \sum_{v \in V} \xi_r(v)}{\xi_r(V)} & \text{if } \xi_r(V) \neq 0 \\ 0 & \text{otherwise,} \end{cases} \tag{20}$$

which is rewritten as

$$\varepsilon_r(V) = \frac{I_r(X_V; Y) - \sum_{v \in V} I_r(X_v; Y)}{I_r(X_V; Y)} \tag{21}$$

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

Figure 2 shows an illustration of the pairwise (or order-2) entropic epistasis. Entropic epistasis has a positive value when $\xi_r(V) > \sum_{v \in V} \xi_r(v)$, and a negative value when $\xi_r(V) < \sum_{v \in V} \xi_r(v)$. The former case means that the corresponding variables interact constructively with each other with respect to the fitness and the latter case means that they interact destructively.

Negative entropic epistasis is connected to the redundancy in the problem encoding. Consider a 3-bit pseudo-Boolean function defined to be $f(000) = 0, f(011) = 1, f(100) = 2, f(111) = 3$ on $\mathcal{U} = \{000, 011, 100, 111\}$. Using $I_u(X_2; Y) = I_u(X_3; Y) = I_u(X_2, X_3; Y) = 1$, we get $\varepsilon_u(2, 3) = -1$, which indicates that x_2 and x_3 interact destructively. Notice that there is some redundancy in this encoding as x_2 and x_3 always have the same values. In fact, this function can be reduced to a 2-bit pseudo-Boolean function defined to be $f'(00) = 0, f'(01) = 1, f'(10) = 2, f'(11) = 3$ on $\mathcal{U}' = \{00, 01, 10, 11\}$. Thus a solution for only two of the three bits is sufficient for the optimization purpose.

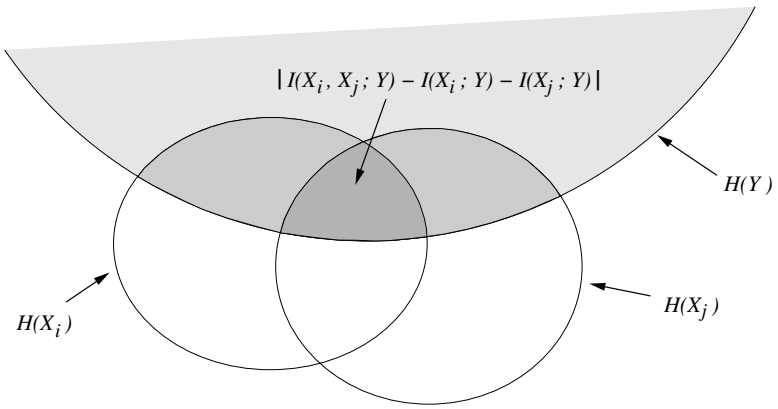


Figure 2: An illustration of the pairwise entropic epistasis.

In summary, negative entropic epistasis indicates a reduction of the search space, which makes the problem easier than it looks.

The following theorem explains the case when the entropic epistasis has a minimum value:

Theorem 8 *Let r be a sampling rate function. A nonempty variable set $V \subseteq \mathcal{V}$ with nonzero significance has minimum entropic epistasis $\varepsilon_r(V) = 1 - |V|$ if and only if $\xi_r(v) = \xi_r(V)$ for all $v \in V$.*

Proof. From that $\xi_r(V) \neq 0$, $I_r(X_V;Y) \neq 0$. Since $0 \leq I_r(X_v;Y) \leq I_r(X_V;Y)$ for all $v \in V$ by the Chain Rule for information (see Appendix A),

$$\begin{aligned} \varepsilon_r(V) I_r(X_V;Y) &= I_r(X_V;Y) - \sum_{v \in V} I_r(X_v;Y) \\ &\geq I_r(X_V;Y) - \sum_{v \in V} I_r(X_V;Y) \\ &= (1 - |V|) I_r(X_V;Y) \end{aligned}$$

with equality if and only if $I_r(X_v;Y) = I_r(X_V;Y)$ for all $v \in V$. □

Consider a pseudo-Boolean function defined to be $f(00) = 0, f(11) = 1$ on $\mathcal{U} = \{00, 11\}$. This is an example of the case $\varepsilon_u(V) = 1 - |V|$ as it has identical mutual information $I_u(X_1;Y) = I_u(X_2;Y) = I_u(X_1, X_2;Y) = 1$. Using that $H_u(Y) = 1$, we obtain that $\xi_u(1) = \xi_u(2) = \xi_u(1,2) = 1$ and $\varepsilon_u(1,2) = -1$.

The following theorem explains the case when the entropic epistasis has a maximum value:

Theorem 9 *Let r be a sampling rate function. A nonempty variable set $V \subseteq \mathcal{V}$ with nonzero significance has maximum entropic epistasis $\varepsilon_r(V) = 1$ if and only if $\xi_r(v) = 0$ for all $v \in V$.*

Proof. From that $\xi_r(V) \neq 0$, $I_r(X_V;Y) \neq 0$. Since $I(X_v;Y) \geq 0$ for all $v \in V$,

$$\begin{aligned} \varepsilon_r(V) I_r(X_V;Y) &= I_r(X_V;Y) - \sum_{v \in V} I_r(X_v;Y) \\ &\leq I_r(X_V;Y) \end{aligned}$$

with equality if and only if $I_r(X_v;Y) = 0$ for all $v \in V$. □

An example of this extreme case is a Boolean function called XOR defined to be $f(00) = f(11) = 0$ and $f(01) = f(10) = 1$ on $\mathcal{U} = \{00, 01, 10, 11\}$. In this case, each variable alone has no information on the fitness as $I_u(X_1; Y) = I_u(X_2; Y) = 0$, while the two variables together have the full information on the fitness as $I_u(X_1, X_2; Y) = 1$. Using that $H_u(Y) = 1$, we obtain that $\xi_u(1) = \xi_u(2) = 0$, $\xi_u(1, 2) = 1$, and $\varepsilon_u(1, 2) = 1$.

Using (6), we can derive that

$$\begin{aligned} & I_r(X_V; Y) - \sum_{v \in V} I_r(X_v; Y) \\ &= H_r(X_V) - H_r(X_V|Y) - \sum_{v \in V} (H_r(X_v) - H_r(X_v|Y)) \\ &= D_r(X_V|Y) - D_r(X_V), \end{aligned}$$

from which we obtain the following theorem:

Theorem 10 *Let r be a sampling rate function. Then*

$$\varepsilon_r(V) I_r(X_V; Y) = D_r(X_V|Y) - D_r(X_V). \tag{22}$$

This formula gives rise to an interesting interpretation of entropic epistasis. Entropic epistasis is the difference between the dependency and the conditional dependency given the fitness; the average increment of the dependence between variables by the information on the fitness.

If the universe is the Cartesian product of the alphabets and the sampling model is uniform, we can obtain a simplified formula for entropic epistasis:

Theorem 11 *Let u be the uniform sampling rate function. Given a problem with universe $\mathcal{U} = \mathcal{A}_1 \times \cdots \times \mathcal{A}_n$,*

$$\varepsilon_u(V) I_u(X_V; Y) = D_u(X_V|Y) \tag{23}$$

for any nonempty variable set $V \subseteq \mathcal{V}$.

Proof. This is immediate from Theorem 5 and Theorem 10. □

This theorem states that the entropic epistasis under the constraints is equivalent to the conditional dependency of the corresponding variables given the fitness; zero entropic epistasis indicates the conditional independence of the variables given the fitness and maximum entropic epistasis indicates the conditional full-dependence of the variables given the fitness. Further discussion of this will be given in Section 3.4.

In this special case, entropic epistasis has tighter bounds:

Theorem 12 *Let u be the uniform sampling rate function. Given a problem with universe $\mathcal{U} = \mathcal{A}_1 \times \cdots \times \mathcal{A}_n$,*

$$0 \leq \varepsilon_u(V) \leq 1 \tag{24}$$

for any nonempty variable set $V \subseteq \mathcal{V}$.

Proof. The upper bound is by Theorem 9. If $I_u(X_V; Y) = 0$, then, by definition, $\varepsilon_u(V) = 0$. Now, assume that $I_u(X_V; Y) \neq 0$. Since the mutual information and the conditional dependency are always nonnegative,

$$\varepsilon_u(V) = \frac{D_u(X_V|Y)}{I_u(X_V; Y)} \geq 0$$

by Theorem 11. □

This means that a set of variables always interact non-destructively when the universe is the Cartesian product of the alphabets and the sampling model is uniform. An example of zero entropic epistasis is a pseudo-Boolean function defined to be $f(00) = 2, f(01) = 1, f(10) = 0, f(11) = 3$ on $\mathcal{U} = \{00, 01, 10, 11\}$. This function has entropic quantities $I_u(X_1; Y) = I_u(X_2; Y) = 1$, $I_u(X_1, X_2; Y) = 2$, and $H_u(Y) = 2$. Thus $\xi_u(1) = \xi_u(2) = \frac{1}{2}$, $\xi_u(1, 2) = 1$, and $\varepsilon_u(1, 2) = 0$.

By Theorem 6, we can derive the following formulas for the relationship between the dependency in the uniform sampling model and that in an arbitrary fitness-based sampling model:

Theorem 13 *Let r be a fitness-based sampling rate function and u be the uniform sampling rate function. For a nonempty variable set $V \subseteq \mathcal{V}$,*

$$D_r(X_V|Y = y) = D_u(X_V|Y = y) \tag{25}$$

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

Proof. This is immediate from the definition of the conditional dependency and Theorem 6. □

Theorem 14 *Let r be a fitness-based sampling rate function with $r(x_\mathcal{V}) = w(f(x_\mathcal{V})), x_\mathcal{V} \in \mathcal{U}$ and u be the uniform sampling rate function. For a nonempty variable set $V \subseteq \mathcal{V}$,*

$$D_r(X_V|Y) = |\mathcal{U}| \sum_{y \in \mathcal{F}} w(y) p_u(y) D_u(X_V|Y = y). \tag{26}$$

Proof. By the definition of conditional dependency, Theorem 6, and Theorem 13,

$$\begin{aligned} D_r(X_V|Y) &= \sum_{y \in \mathcal{F}} p_r(y) D_r(X_V|Y = y) \\ &= |\mathcal{U}| \sum_{y \in \mathcal{F}} w(y) p_u(y) D_u(X_V|Y = y). \end{aligned}$$

□

Recall that

$$D_u(X_V|Y) = \sum_{y \in \mathcal{F}} p_u(y) D_u(X_V|Y = y). \tag{27}$$

Theorem 14 and (27) state that $D_u(X_V|Y)$ is a weighted sum of $D_u(X_V|Y = y)$ with weights $p_u(y)$, while $D_r(X_V|Y)$ in a fitness-based sampling model is a weighted sum of the same $D_u(X_V|Y = y)$ with different weights $|\mathcal{U}|w(y)p_u(y)$.

Theorem 15 *Let r be a fitness-based sampling rate function and u be the uniform sampling rate function. For a nonempty variable set $V \subseteq \mathcal{V}$, if $D_u(X_V|Y) = 0$, then $D_r(X_V|Y) = 0$. Conversely, if $r(x_\mathcal{V}) > 0$ for all $x_\mathcal{V} \in \mathcal{U}$ and $D_r(X_V|Y) = 0$, then $D_u(X_V|Y) = 0$.*

Proof. By Theorem 14 and the fact that $p_u(y) > 0$ and $D_u(X_V|Y = y) \geq 0$, the theorem follows. □

This means that given a fitness-based sampling function $r(x_\mathcal{V}) > 0$ for all $x_\mathcal{V} \in \mathcal{U}$, $\varepsilon_r(V) I_r(X_V; Y) = -D_r(X_V)$ is equivalent to that $\varepsilon_u(V) I_u(X_V; Y) = -D_u(X_V)$. Thus, by Theorem 5, we obtain that if $\mathcal{U} = \mathcal{A}_1 \times \cdots \times \mathcal{A}_n$ and $r(x_\mathcal{V}) > 0$ for all $x_\mathcal{V} \in \mathcal{U}$, then $\varepsilon_r(V) I_r(X_V; Y) = -D_r(X_V)$ is equivalent to that $\varepsilon_u(V) = 0$.

We define the mean significance and the mean entropic epistasis of all combinations of k variables as new measures for the fitness functions:

Definition 7 (Mean Significance) *Let r be a sampling rate function. The order-k mean significance $\xi_r^{(k)}$ is defined to be*

$$\xi_r^{(k)} = \frac{1}{\binom{n}{k}} \sum_{V \subseteq \mathcal{V}, |V|=k} \xi_r(V) \tag{28}$$

for $1 \leq k \leq n$.

Definition 8 (Mean Entropic Epistasis) *Let r be a sampling rate function. The order-k mean entropic epistasis $\varepsilon_r^{(k)}$ is defined to be*

$$\varepsilon_r^{(k)} = \frac{1}{\binom{n}{k}} \sum_{V \subseteq \mathcal{V}, |V|=k} \varepsilon_r(V) \tag{29}$$

for $1 \leq k \leq n$.

Clearly, the mean significance ranges from 0 to 1 and the order-k mean entropic epistasis ranges from $1 - k$ to 1. In the case that the universe is the Cartesian product of the alphabets and the sampling model is uniform, $0 \leq \varepsilon_r^{(k)} \leq 1$.

3.4 Epistasis Variance versus Entropic Epistasis

In this section, we attempt to provide an in-depth insight into entropic epistasis by comparing it with the epistasis variance.

Epistasis variance is an epistasis measure that quantifies the nonlinearity lying in the fitness landscape based on experimental design theory (Davidor, 1990; Reeves and Wright, 1995a; Reeves and Wright, 1995b). Experimental design is a branch of statistics that attempts to conduct the way in which experiments should be carried out so that the data gathered will have statistical value. Interestingly, the epistatic behavior of variables is well explained by this theory. According to the theory, the fitness $f(x_\mathcal{V})$ of a solution $x_\mathcal{V} = (x_1, x_2, \ldots, x_n) \in \mathcal{U}$ can be expressed as

$$\begin{aligned} f(x_\mathcal{V}) \quad = \quad & (\text{constant}) + \sum_{v \in \mathcal{V}} (\text{effect of } x_v) \\ & + \sum_{\substack{v_1 < v_2 \\ v_1, v_2 \in \mathcal{V}}} (\text{joint effect of } x_{v_1} \text{ and } x_{v_2}) \\ & + \cdots \\ & + (\text{joint effect of } x_1,\ x_2, \ldots, \text{ and } x_n). \end{aligned} \tag{30}$$

For instance, a fitness function of three variables can be written as

$$\begin{aligned} f(a_1, a_2, a_3) \quad = \quad & (\text{constant}) + f_1(a_1) + f_2(a_2) + f_3(a_3) \\ & + f_{1,2}(a_1, a_2) + f_{1,3}(a_1, a_3) + f_{2,3}(a_2, a_3) \\ & + f_{1,2,3}(a_1, a_2, a_3) \end{aligned} \tag{31}$$

where $f_1(a_1)$ is the effect of $x_1 = a_1$, $f_2(a_2)$ is the effect of $x_2 = a_2$, $f_{1,2}(a_1, a_2)$ is the joint effect of $x_1 = a_1$ and $x_2 = a_2$, and so on. These effects proved to be equivalent to the partition coefficients of the Walsh transform (Reeves and Wright, 1995a). In (30), the terms "$\sum_{v \in \mathcal{V}}$(effect of x_v)" are known as linear effects and other terms (except the constant) as interaction effects. It is easy to show that the total sum of squares (SS), or the total variation of the fitness is equal to the sum of the linear effects SS and interaction effects SS, i.e.,

$$\text{Total SS} \ = \ \text{Linear effects SS} \ + \ \text{Interaction effects SS}.$$

Davidor's epistasis variance (Davidor, 1990) corresponds to the interaction effects SS and the normalized epistasis variance (Reeves and Wright, 1995a; Reeves and Wright, 1995b) corresponds to the interaction effects SS over the total SS. The normalized epistasis variance η is written as

$$\eta = \frac{\sum_{x_{\mathcal{V}} \in \mathcal{U}} (\text{interaction effect of } x_{\mathcal{V}})^2}{\sum_{x_{\mathcal{V}} \in \mathcal{U}} (f(x_{\mathcal{V}}) - (\text{constant}))^2}. \tag{32}$$

We see that this measure has value zero if and only if the interaction effects SS is zero.

Let $\bar{f}$ denote the average fitness of the solutions in the universe. Given a variable subset $V = \{v_1, v_2, \ldots, v_k\} \subseteq \mathcal{V}$, let $\bar{f}_V(a_{v_1}, a_{v_2}, \ldots, a_{v_k})$ denote the average fitness of the solutions $x_{\mathcal{V}} = (x_1, x_2, \ldots, x_n) \in \mathcal{U}$ such that $x_{v_1} = a_{v_1}, x_{v_2} = a_{v_2}, \ldots, x_{v_k} = a_{v_k}$, i.e.,

$$\bar{f}_V(a_{v_1}, a_{v_2}, \ldots, a_{v_k}) = \frac{\sum_{x_{\mathcal{V}} \in \mathcal{U}} f(x_{\mathcal{V}}) \prod_{i=1}^{k} 1(x_{v_i} = a_{v_i})}{\sum_{x_{\mathcal{V}} \in \mathcal{U}} \prod_{i=1}^{k} 1(x_{v_i} = a_{v_i})}. \tag{33}$$

Let $f_v(x_v)$ denote the average excess fitness of x_v, i.e., $f_v(x_v) = \bar{f}_v(x_v) - \bar{f}$. Then, in (30), $(\text{constant}) = \bar{f}$ and $(\text{effect of } x_v) = f_v(x_v)$. Thus we obtain that $(\text{interaction effect of } x_{\mathcal{V}}) = f(x_{\mathcal{V}}) - \left(\bar{f} + \sum_{v \in \mathcal{V}} f_v(x_v)\right)$. Accordingly, we can rewrite (32) as

$$\eta = \frac{\sum_{x_{\mathcal{V}} \in \mathcal{U}} \left(f(x_{\mathcal{V}}) - \bar{f} - \sum_{v \in \mathcal{V}} f_v(x_v) \right)^2}{\sum_{x_{\mathcal{V}} \in \mathcal{U}} \left(f(x_{\mathcal{V}}) - \bar{f} \right)^2}. \tag{34}$$

From this, we obtain that $\eta = 0$ if and only if

$$f(x_{\mathcal{V}}) = \bar{f} + \sum_{v \in \mathcal{V}} f_v(x_v) \tag{35}$$

for all $x_{\mathcal{V}} \in \mathcal{U}$. In this case, f is said to be *additively separable* into f_v for $v \in \mathcal{V}$. For instance, $f(x_1, x_2) = 2 + x_1 + 3e^{x_2}$ is additively separable into x_1 and $3e^{x_2}$, but $f(x_1, x_2) = x_1 x_2$ is a function which is not additively separable.

Based on (34), we can define the epistasis variance of a variable set as follows:

Definition 9 (Epistasis Variance) *Given a nonempty variable set $V \subseteq \mathcal{V}$, the epistasis variance $\eta(V)$ is defined to be*

$$\eta(V) = \frac{\sum_{x_{\mathcal{V}} \in \mathcal{U}} \left(\bar{f}_V(x_V) - \bar{f} - \sum_{v \in V} f_v(x_v) \right)^2}{\sum_{x_{\mathcal{V}} \in \mathcal{U}} \left(\bar{f}_V(x_V) - \bar{f} \right)^2}. \tag{36}$$

From this definition, we obtain the following theorem:

Theorem 16 *Given a nonempty variable set $V \subseteq \mathcal{V}$, $\eta(V) = 0$ if and only if*

$$\bar{f}_V(x_V) = \bar{f} + \sum_{v \in V} f_v(x_v) \tag{37}$$

for all $x_\mathcal{V} \in \mathcal{U}$.

Theorem 16 states that the definition of the epistasis variance is based on the algebraic, or scalar influence of variables on the fitness. In the case of entropic epistasis, however, different aspects are focused on.

Consider the case of zero entropic epistasis. By Theorem 10, $\varepsilon_r(V) = 0$ if and only if $D_r(X_V|Y) = D_r(X_V)$ for a nonempty variable set $V \subseteq \mathcal{V}$. This means that the conditioning of the fitness does not affect the dependence between the variables. Particularly, when the universe is the Cartesian product of the alphabets and the sampling model is uniform, zero entropic epistasis means that the variables are conditionally independent given the fitness by Theorem 11. Now, consider the conditional probability of the fitness given variables:

Theorem 17 *Let u be the uniform sampling rate function. Given a problem with universe $\mathcal{U} = \mathcal{A}_1 \times \cdots \times \mathcal{A}_n$ and a nonempty variable set $V = \{v_1, \ldots, v_k\} \subseteq \mathcal{V}$ with nonzero significance, $\varepsilon_u(V) = 0$ if and only if*

$$p_u(y|X_V = x_V) = p_u(y)^{1-|V|} \prod_{v \in V} p_u(y|X_v = x_v) \tag{38}$$

for all $x_V \in \mathcal{A}_{v_1} \times \cdots \times \mathcal{A}_{v_k}$ and $y \in \mathcal{F}$.

Proof. By Theorem 5,

$$p_u(x_V) = \prod_{v \in V} p_u(x_v) \tag{39}$$

for all $x_V \in \mathcal{A}_{v_1} \times \cdots \times \mathcal{A}_{v_k}$. Also, by Theorem 11, $\varepsilon_u(V) = 0$ if and only if $D_u(X_V|Y) = 0$. By Theorem 3, this is equivalent to

$$p_u(x_V|Y = y) = \prod_{v \in V} p_u(x_v|Y = y) \tag{40}$$

for all $y \in \mathcal{F}$ and $x_V \in \mathcal{A}_{v_1} \times \cdots \times \mathcal{A}_{v_k}$. From (39) and (40),

$$\begin{aligned} p_u(y|X_V = x_V) &= \frac{p_u(y)p_u(x_V|Y = y)}{p_u(x_V)} \\ &= \frac{p_u(y) \prod_{v \in V} p_u(x_v|Y = y)}{\prod_{v \in V} p_u(x_v)} \\ &= p_u(y) \prod_{v \in V} \frac{p_u(y|X_v = x_v)}{p_u(y)} \\ &= p_u(y)^{1-|V|} \prod_{v \in V} p_u(y|X_v = x_v) \end{aligned}$$

for all $x_V \in \mathcal{A}_{v_1} \times \cdots \times \mathcal{A}_{v_k}$ and $y \in \mathcal{F}$. □

This theorem states that if a nonempty variable set $V \subseteq \mathcal{V}$ has zero entropic epistasis, we can obtain the conditional probability $p_u(y|X_V = x_V)$ by multiplying the marginal conditional probabilities $p_u(y|X_v = x_v)$ to $p_u(y)^{1-|V|}$. This means that if $\varepsilon_u(V) = 0$, the conditional probability of Y given X_v contributes independently to the conditional probability of Y given X_V, by which we can predict[3] the fitness Y. This is different from the fact that zero epistasis variance implies that the function is additively separable. In this sense, we say that the definition of entropic epistasis is based on the probabilistic, or entropic influence of variables on the fitness. A function $f(x_1, x_2) = x_1 + x_2$ is an example that has zero epistasis variance, but nonzero entropic epistasis. A pseudo-Boolean function defined to be $f(00) = 2, f(01) = 1, f(10) = 0, f(11) = 3$ is an example that has zero entropic epistasis, but nonzero epistasis variance. A pseudo-Boolean function defined to be $f(00) = 0, f(01) = 1, f(10) = 2, f(11) = 3$ is an example whose epistasis variance and entropic epistasis are both zero.

We can find a secondary difference between the two epistasis measures in terms of how they deal with fitness. That is, the epistasis variance regards the fitness as a scalar quantity, whereas the entropic epistasis treats it as a categorical index. In information theory, entropic quantities are not directly affected by the magnitude of elements in the alphabets. Entropic epistasis behaves in a similar way. (Thus it is more statistically stable to discretize the fitness, as will be explained in Section 3.7, before computing the entropic quantities particularly in the optimization problems with large ranges of fitness.)

We can summarize the above observations as follows: Both of the two measures concern the interactions of variables. But, the term contribution means the algebraic influence of variables on the fitness in the epistasis variance, while it means the probabilistic influence in the entropic epistasis. Also, the fitness is treated differently in the two measures.

3.5 Fitness-Based Sampling and Algorithm-Dependent Entropic Epistasis

The sampling rate of a solution in a sampling model represents the degree of consideration of the solution in the computation of entropic epistasis. In fitness-based sampling models, the sampling rate is determined by the fitness. Given any nonnegative function $h : \mathbb{R} \to [0, \infty)$, we can construct a fitness-based sampling rate function $r : \mathcal{U} \to \mathbb{R}$ by normalizing h as

$$r(x_\mathcal{V}) = \frac{h(f(x_\mathcal{V}))}{\sum_{x_\mathcal{V} \in \mathcal{U}} h(f(x_\mathcal{V}))}. \tag{41}$$

We see that $r(\cdot)$ satisfies (11) and (12). Figure 3 shows four example fitness-based sampling rate functions based on a constant function, a truncated function, a linear function, and a half-Gaussian function, respectively. In a sampling model using a truncated function, we can ignore all solutions that have fitness values lower than a given bound. In a sampling model using a linear function or a half-Gaussian function, high-fitness solutions have larger sampling rates than low-fitness solutions.

If we normalize the function h with respect to a sample set $S \subsetneq \mathcal{U}$, we obtain a

[3] Although we are using the term predict, we can obtain only a probability distribution over the fitness space rather than a crisp value unless $V = \mathcal{V}$.

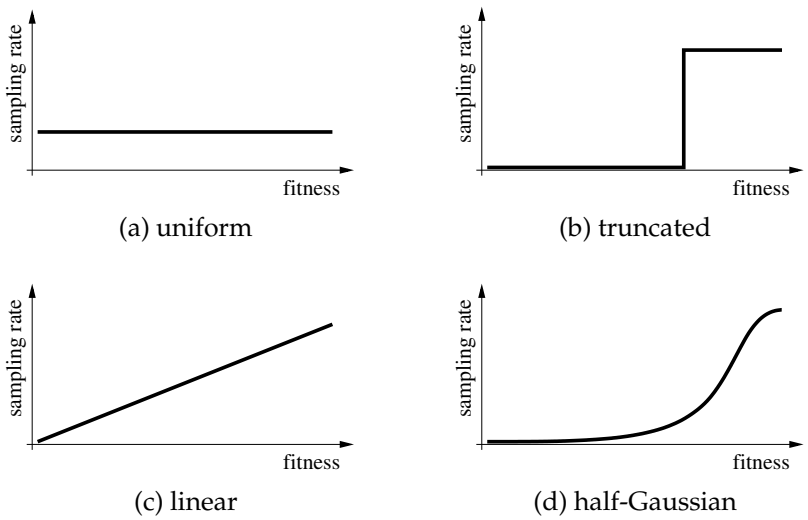


Figure 3: Four examples of the fitness-based sampling rate function.

sampling rate function $r_S : \mathcal{U} \to \mathbb{R}$ as follows:

$$r_S(x_\mathcal{V}) = \begin{cases} \dfrac{h(f(x_\mathcal{V}))}{\sum_{x_\mathcal{V} \in S} h(f(x_\mathcal{V}))} & \text{if } x_\mathcal{V} \in S \\ 0 & \text{otherwise.} \end{cases} \tag{42}$$

Note that this is not a fitness-based sampling function since the sampling rate of a solution depends not only on its fitness but also on whether it is in S or not. Consider a dynamic computation of the entropic epistasis in an evolutionary algorithm. As mentioned in Section 3.2, we can use the population of the evolutionary algorithm as a sample set. By replacing S in (42) with population P, we obtain a sampling rate function $r_P(\cdot)$ with respect to P. This algorithm-dependent computation of the entropic epistasis can be seen as a sequential application of the algorithm-independent computation with a sampling rate function that varies with the population.

3.6 Time Complexity

Let $V \subseteq \mathcal{V}$ be a variable set and l denote the size of the sample set used in the probability space construction. For any sampling rate function r, the entropy $H_r(Y)$ and the mutual information $I_r(X_V; Y)$ can be computed in $\Theta(l+\beta)$ time and $\Theta(kl+\beta\prod_{v\in V}\alpha_v)$ time, respectively, where $\alpha_v = |\mathcal{A}_v|$, $\beta = |\mathcal{F}|$, and $k = |V|$. The first terms l and kl are for constructing the probability tables by accumulating the occurrences of events in the samples and the other terms β and $\beta\prod_{v\in V}\alpha_v$ for calculating the $\log$ values of the probability table entries. Thus both the significance $\xi_r(V)$ and the entropic epistasis $\varepsilon_r(V)$ are computed in $\Theta(kl + \beta\prod_{v\in V}\alpha_v)$ time. In the case of constant alphabet size $|\mathcal{A}_v| = \alpha$ for $v \in V$, the time complexity is $\Theta(kl + \alpha^k\beta)$. Thus, for fixed k, this is polynomial in α, β, and l.

Table 1: The two 8-bit Royal Road functions.

(a) R_1

Schema s_i (1 2 3 4 5 6 7 8)	Coeff c_i
1 1 * * * * * *	2
* * 1 1 * * * *	2
* * * * 1 1 * *	2
* * * * * * 1 1	2
$s_{opt} = 1\,1\,1\,1\,1\,1\,1\,1$	

(b) R_2

Schema s_i (1 2 3 4 5 6 7 8)	Coeff c_i
1 1 * * * * * *	2
* * 1 1 * * * *	2
* * * * 1 1 * *	2
* * * * * * 1 1	2
1 1 1 1 * * * *	4
* * * * 1 1 1 1	4
$s_{opt} = 1\,1\,1\,1\,1\,1\,1\,1$	

3.7 Fitness Discretization

In the combinatorial optimization problems, the range of the fitness is the real numbers $\mathbb{R}$, while each variable has a discrete alphabet. Although the set $\mathcal{F}$ of all fitness values is finite, it is not so desirable to consider each distinct value in $\mathcal{F}$ as a discrete symbol of the random variable Y since we can hardly talk about the statistics including the fitness as a variable if the solutions rarely share the same fitness. In this case, it is helpful to discretize the fitness before applying the equations in Section 3.3. The most simple and intuitive discretization methods are the equal-width discretization and the equal-frequency discretization (Chiu et al., 1991). The range of the fitness is divided into K_Y intervals with the same width in the equal-width discretization, while it is divided into K_Y intervals with the same number of samples in the equal-frequency discretization. As a result, $\mathcal{F}' = \{0, 1, \ldots, K_Y - 1\}$ can be used instead of the original $\mathcal{F}$ in the computation of the entropic epistasis. We will use the equal-frequency discretization with $K_Y = 10$ in the experiments in Sections 4.2 and 4.3.

4 Experiments

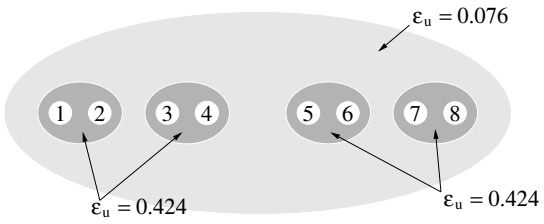
4.1 Royal Road Function

We tested the proposed measures on the Royal Road function (Forrest and Mitchell, 1993), which is a family of pseudo-Boolean functions designed to investigate how schema processing actually takes place inside the evolutionary algorithm. These functions are suitable for this kind of test as they have explicit building blocks. Formally, a Royal Road function $R : \{0, 1\}^n \to \mathbb{R}$ is defined as

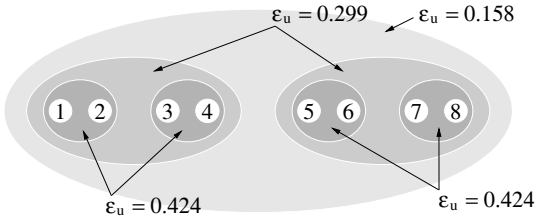
$$R(x_1, x_2, \ldots, x_n) = \sum_i c_i \delta_i(x_1, x_2, \ldots, x_n) \tag{43}$$

where c_i is a predefined coefficient corresponding to a schema s_i and $\delta_i : \{0, 1\}^n \to \{0, 1\}$ is an indicator function that returns 1 if s_i contains[4] a given solution and returns 0 otherwise. The order of schema s_i is usually used as the coefficient c_i. Table 1 shows the two 8-bit Royal Road functions we used in the experiments. The function R_1 has four order-2 building blocks, while R_2 has two more building blocks of order-4.

[4] A schema is sometimes regarded as a subset of solutions that have a specific pattern of variable assignments. Thus a solution that has the pattern is said to be contained in the schema.



(a) R_1



(b) R_2

Figure 4: The pairwise entropic epistasis $\varepsilon_u(u, v)$ in the 8-bit Royal Road functions.

In the first experiment, we computed the pairwise entropic epistasis in the Royal Road functions. The results were illustrated in Figure 4. We can see that variable pairs that belong to some building block(s) have relatively stronger entropic epistasis than those that belong to no building block in both R_1 and R_2, and variable pairs that belong to an order-2 building block and, at the same time, an order-4 building block have stronger entropic epistasis than those that belong only to an order-4 building block in R_2. These results are consistent with the intuition that a building block means the existence of the interactions between the corresponding variables.

In the second experiment, we computed some high-order entropic epistasis values in the Royal Road functions. The results are shown in Table 2. This table lists the number of building blocks (# 2-BBs, # 4-BBs) and the entropic epistasis ($\varepsilon_u(V)$) of eleven representative variable sets (V). We can see that a variable set has stronger entropic epistasis as it contains more building blocks. The results above show that the proposed measure quantifies the variable interactions well. Also, they show the possibility of using the measures for detecting building blocks in black box optimizations where no problem-specific knowledge is available.

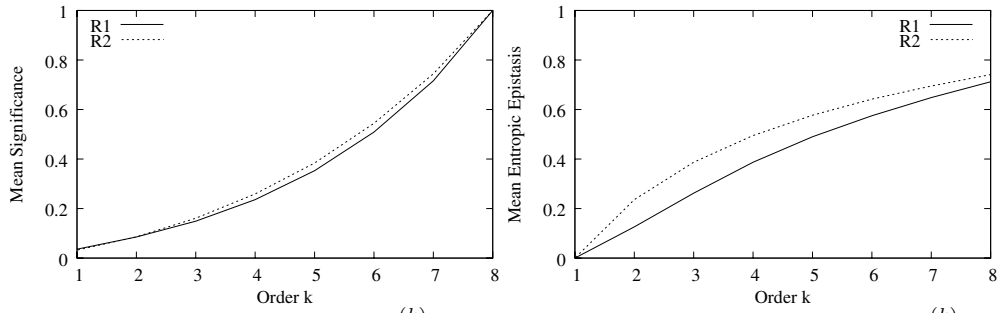
In the final experiment, we computed the mean significance and the mean entropic epistasis of the Royal Road functions. Figure 5 shows the results. We can see that R_2 has stronger mean significance and mean entropic epistasis than R_1. Interestingly, this result is consistent with the experimental results in (Forrest and Mitchell, 1993) where R_1 was more easily solved than R_2 with genetic algorithms.

4.2 NK-Landscape Function

The NK-landscape model (Kauffman, 1989) is a scheme to define a family of pseudo-Boolean functions that have controllable degrees of epistasis. In this model, a fitness function is parameterized by N and K; N is the number of variables and K indicates the degree of epistasis. Formally, an NK-landscape function $f : \{0, 1\}^N \rightarrow \mathbb{R}$ is defined

Table 2: The high-order entropic epistasis $\varepsilon_u(V)$ in the 8-bit Royal Road functions.

	R_1		R_2		
V	# 2-BBs	$\varepsilon_u(V)$	# 2-BBs	# 4-BBs	$\varepsilon_u(V)$
$\{1,3,5\}$	0	0.173	0	0	0.336
$\{1,2,5\}$	1	0.381	1	0	0.425
$\{1,2,3\}$	1	0.381	1	0	0.519
$\{1,3,5,7\}$	0	0.315	0	0	0.444
$\{1,2,5,7\}$	1	0.395	1	0	0.482
$\{1,2,5,6\}$	2	0.507	2	0	0.553
$\{1,2,3,4\}$	2	0.507	2	1	0.642
$\{1,2,3,5,6,7\}$	2	0.571	2	0	0.632
$\{1,2,3,4,5,7\}$	2	0.571	2	1	0.648
$\{1,2,3,4,5,6\}$	3	0.600	3	1	0.668
$\{1,2,3,4,5,6,7,8\}$	4	0.712	4	2	0.741



(a) The mean significance $\xi_u^{(k)}$ (b) The mean entropic epistasis $\varepsilon_u^{(k)}$

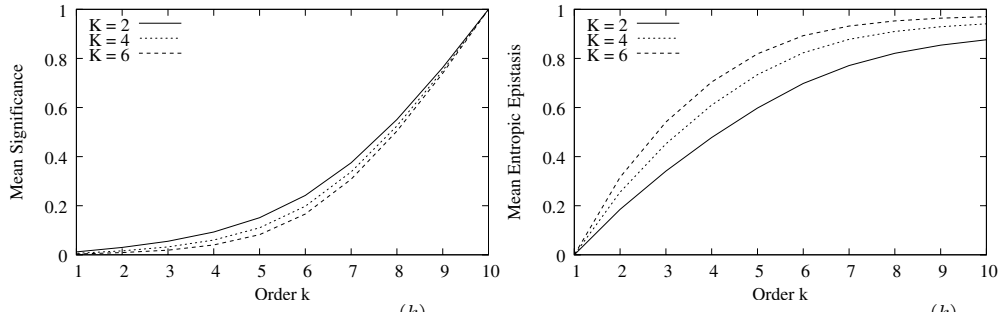
Figure 5: The mean significance $\xi_u^{(k)}$ and the mean entropic epistasis $\varepsilon_u^{(k)}$ of the 8-bit Royal Road functions.

as follows:

$$f(x_1, x_2, \ldots, x_N) = \frac{1}{N} \sum_{i=1}^{N} f_i(x_i, x_{j_{i,1}}, x_{j_{i,2}}, \ldots, x_{j_{i,K}}) \tag{44}$$

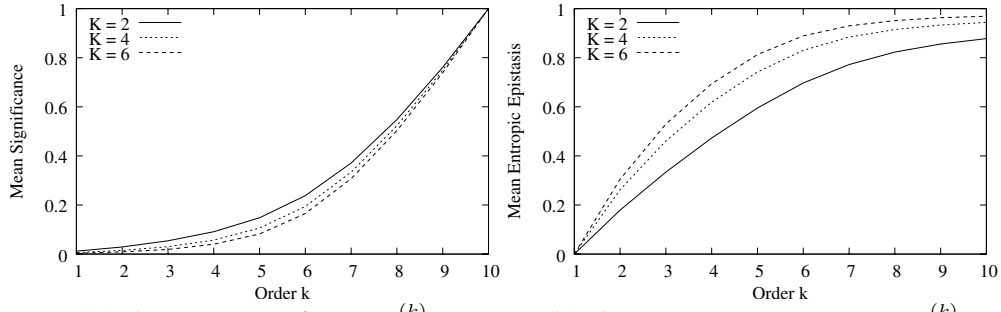
where f_i is a function of $(K+1)$ variables x_i, $x_{j_{i,1}}$, $x_{j_{i,2}}$, ..., $x_{j_{i,K}}$. Each function $f_i : \{0,1\}^{K+1} \rightarrow [0,1]$ assigns a real number chosen uniformly at random from the interval $[0,1]$ to each of 2^{K+1} inputs. Usually, the K indices $j_{i,1}, j_{i,2}, \ldots, j_{i,K}$ are determined by two types of neighborhood models: the random neighborhood model and the adjacent neighborhood model. The indices for each i are chosen uniformly at random from $\{1,2,\ldots,N\} \setminus \{i\}$ without duplication in the random neighborhood model, while each index $j_{i,k}$ is set to $1+(i+k-1)\ mod\ N$ in the adjacent neighborhood model. The optimization of the NK-landscape functions in the random neighborhood model proved to be NP-hard for any $K \geq 2$ (Weinberger, 1996). We apply the equal-frequency discretization with ten intervals to the fitness in computing the entropic epistasis in the following experiments. The analysis in Section 3.6 shows that the entropic epistasis of k variables is computed in $\Theta(kl + 2^k)$ time as $\alpha = 2$ and $\beta = 10$.

In the first experiment, we computed the mean significance and the mean entropic



(a) The mean significance $\xi_u^{(k)}$ (b) The mean entropic epistasis $\varepsilon_u^{(k)}$

Figure 6: The mean significance $\xi_u^{(k)}$ and the mean entropic epistasis $\varepsilon_u^{(k)}$ of the NK-landscape functions with $N = 10$ in the random neighborhood model.

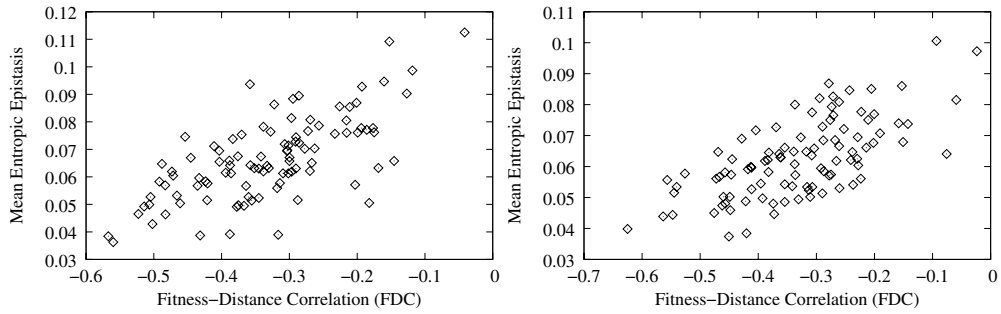


(a) The mean significance $\xi_u^{(k)}$ (b) The mean entropic epistasis $\varepsilon_u^{(k)}$

Figure 7: The mean significance $\xi_u^{(k)}$ and the mean entropic epistasis $\varepsilon_u^{(k)}$ of the NK-landscape functions with $N = 10$ in the adjacent neighborhood model.

epistasis of the NK-landscape functions with $N = 10$ and varying K using the whole solutions in the universe. Figures 6–7 show the results. We can see that both the mean significance and the mean entropic epistasis increase monotonously as the order k increases. Also, for a fixed order k, the mean significance decreases, but the mean entropic epistasis increases with the increase of K. The following interpretation is possible: As K increases, the influence of a variable on the fitness tends to be mixed up with the influences of other variables causing the decrease of the significance and the increase of entropic epistasis.

In the second experiment, we compared the order-2 mean entropic epistasis $\varepsilon_u^{(2)}$ with the fitness distance correlation (FDC) for the NK-landscape functions. The FDC (Jones and Forrest, 1995) is a well-known measure of the GA-hardness of fitness functions, which is defined as the correlation coefficient between the fitness and the distance to the global optimum. So, it ranges from -1 to 1. The measure value close to -1 indicates that the solutions similar to the global optimum have high fitness, while the measure value close to 1 indicates that they have low fitness. At the same time, the measure value close to 0 indicates that the similarity to the global optimum gives little information about the fitness. With the FDC value ρ, Jones and Forrest roughly classified the fitness landscapes into three classes, straightforward ($\rho \leq -0.15$), misleading



(a) Random neigh. model (b) Adjacent neigh. model

Figure 8: A comparison between the order-2 mean entropic epistasis $\varepsilon_u^{(2)}$ and the FDC for the NK-landscape functions with $N = 20$ and $K = 2$.

Table 3: The correlation coefficients between the order-2 mean entropic epistasis $\varepsilon_u^{(2)}$ and the FDC for the NK-landscape functions with varying N and K.

N	K	Neighbor Model	
		Random	Adjacent
20	2	0.688	0.674
	4	0.393	0.470
	6	0.250	0.281
40	2	0.509	0.486
	4	0.460	0.197
	6	0.228	0.215

($\rho \geq 0.15$), and difficult ($-0.15 < \rho < 0.15$). We used the Hamming distance[5] as the distance measure between solutions and used one million randomly generated solutions instead of the universe in the probability space construction due to the time/space limitation. Figure 8 shows graphs comparing the two measures for one hundred independently generated NK-landscape functions with $N = 20$ and $K = 2$ in each of the two neighborhood models. We can observe considerable correlations between the two measures. The correlation coefficients are 0.688 and 0.674, respectively. This is interesting enough as the two measures focusing on different aspects of the fitness functions coincide with each other. We computed the correlation coefficients for varying N and K as shown in Table 3. As we can see, the coefficients have a tendency to decrease with the increase of K.

The final experiment is an application of the entropic epistasis to the performance improvement of genetic algorithms. In Figure 8, we can see that all the function instances have negative FDC values and most of them fall in the straightforward class. These results can be interpreted as follows. In spite of the local ruggedness, the solutions that have more features of the global optimum tend to have higher fitness values. We can understand such features as building blocks. Moreover, the negative FDC values indicate that the fitness landscapes are not deceptive in global sense. In optimizing fitness functions with these characteristics, the power of the traditional k-point cross-

[5]The number of variables whose values are not identical.

Table 4: The performance improvements of the genetic algorithms by the EGR for the NK-landscape functions.

N	K	Reordering	Random Neigh. Model			Adjacent Neigh. Model		
			Avg	$\sigma/\sqrt{m}$	Time (s)	Avg	$\sigma/\sqrt{m}$	Time (s)
40	2	None	29.954	0.006	0.92	29.858	0.008	1.15
		EGR	**29.959**	0.006	0.91	**29.880**	0.008	1.16
	4	None	31.239	0.007	0.96	31.213	0.009	1.18
		EGR	**31.253**	0.007	0.96	**31.264**	0.009	1.19
	6	None	31.167	0.005	0.99	31.426	0.005	1.25
		EGR	**31.179**	0.005	0.99	**31.513**	0.004	1.26
80	2	None	60.092	0.014	1.53	60.203	0.009	1.18
		EGR	**60.154**	0.014	1.55	**60.303**	0.009	1.18
	4	None	61.515	0.010	1.63	61.258	0.013	1.24
		EGR	**61.599**	0.010	1.98	**61.504**	0.013	1.25
	6	None	61.600	0.009	1.70	61.397	0.009	1.32
		EGR	**61.646**	0.009	1.77	**61.715**	0.009	1.33

over is known to depend on the arrangement of variables in the chromosome. That is, a variable arrangement where mutually epistatic variables are placed near to one another is advantageous. Based on this observation, we can consider preprocessing the loci of variables before running the genetic algorithms; we can design the following simple reordering method: At first, we choose a variable at random as the first variable v_1 and place it at the left most locus in the chromosome. In i^{th} stage, a variable that has the strongest pairwise entropic epistasis with v_{i-1} is chosen as the i^{th} variable v_i among those not chosen yet and is placed at the i^{th} locus in the chromosome. This stage is repeated until all variables are placed in the chromosome. We call this reordering method *epistasis-based greedy reordering (EGR)*.

We solved the NK-landscape functions with varying N and K using genetic algorithms with and without the EGR, respectively, and gauged the performance improvements by the reordering. For the EGR, the pairwise entropic epistasis $\varepsilon_u(u, v)$ for each $u, v \in \mathcal{V}$ was computed using one hundred thousand sample solutions generated at random. The EGR is performed once before running each genetic algorithm. It is assumed that the variables are given in a random order. We used a steady-state genetic algorithm where each offspring generated is immediately inserted into the population. The main genetic operators we used are a roulette-wheel selection, a traditional 2-point crossover, a fixed-bit mutation, and a preselection-like replacement (Bui and Moon, 1994). The genetic algorithms stop after one hundred thousand generations in each run. For each N and K, we used ten independently-generated function instances in this experiment and, for each instance, we ran each genetic algorithm one thousand times. Table 4 shows the results. In this table, listed are N, K, the reordering method (Reordering), the average solution quality (Avg), the standard deviation over the square root of the number of runs ($\sigma/\sqrt{m}$), and the running time (Time). We can observe considerable performance improvements by the EGR, which indicates that the EGR is indeed helpful to the crossover efficiency. It is notable that the improvements were more obvious in the adjacent neighborhood model than in the random neighborhood model. This is a good example showing how entropic epistasis can be used to improve the performance

of evolutionary algorithms.

4.3 Traveling Salesman Problem

The traveling salesman problem (TSP) is a problem of finding the shortest Hamiltonian cycle visiting given cities. This problem is a well-known NP-hard problem (Garey and Johnson, 1979) that has served for decades as an initial proving ground for new optimization techniques. To bind each variable to a phenotype feature (namely, a city), we used a locus-based encoding as in (Bui and Moon, 1994); one variable is assigned to each city to indicate the index of the next city in the Hamiltonian cycle. For example, a tour $2\rightarrow1\rightarrow3\rightarrow4\rightarrow5$ is encoded into $(3, 1, 4, 5, 2)$. The domain of x_i is thus $\mathcal{A}_i = I \setminus \{i\}$. We confine the feasible solutions to only those corresponds to complete tours among the solutions in $\mathcal{A}_1 \times \cdots \times \mathcal{A}_n$. We define the fitness of a solution $x_{\mathcal{V}} \in \mathcal{U}$ as

$$f(x_{\mathcal{V}}) = C_{max} - \sum_{i=1}^{n} d_{ix_i} \tag{45}$$

where d_{uv} is the distance from city u to city v and C_{max} is a constant larger than the worst-case cycle length. The subtraction in the equation forces the problem to become a maximization problem. We applied the equal-frequency discretization with ten intervals to the fitness as in Section 4.2. Note that C_{max} does not affect entropic epistasis with this discretization. The analysis in Section 3.6 shows that the time complexity of computing the order-k entropic epistasis is $\Theta(kl + n^k)$ as $\alpha = n - 1$ and $\beta = 10$.

An empirical study (Boese et al., 1994) showed that the fitness landscapes of the TSP have a globally convex structure called Big Valley. This result implies that the fitness landscapes are not deceptive in a global sense and there exist features of solutions that can be understood as building blocks. In this experiment, we applied the EGR to genetic algorithms for the TSP similarly to the case of the NK-landscape functions in Section 4.2. For the EGR, the pairwise entropic epistasis $\varepsilon_u(u, v)$ for each $u, v \in \mathcal{V}$ was computed using one hundred thousand sample solutions generated at random. The EGR is performed once before running each genetic algorithm. It is assumed that the variables are given in a random order. We used the same genetic algorithms as that used in our previous study (Seo and Moon, 2002) except the crossover and the local search heuristic; we used steady-state genetic algorithms adopting a roulette-wheel selection, a preselection-like replacement (Bui and Moon, 1994), and no local search heuristic. Crossover is a traditional k-point crossover with $k = 2 + \lfloor \ln n + \frac{1}{2} \rfloor$ where n is the problem size, and the mutation is one double-bridge kick move (Lin and Kernighan, 1973) to each offspring with probability $\frac{1}{10}$. The genetic algorithms stop when 70 percent of the population get to have the same fitness as the best of the population. All solutions in the population are maintained feasible by repairing each offspring obtained from the crossover. This is done by a heuristic that aims to maximize the number of edges inherited from the parent cycles to the offspring. As this repair mechanism is not directly affected by the order of variables in the chromosome, we believe that its disturbance in the comparison of the algorithms is negligible. We independently ran each genetic algorithm one hundred times for each of eight problem instances obtained from the TSPLIB[6].

Table 5 shows the results. This table lists the instance name (Graph) with the optimal cycle length (Opt), the reordering method (Reordering), the average cycle length (Cycle Length) with the percentage of the excess over the optimum (%), the standard

[6] Available at `http://www.iwr.uni-heidelberg.de/groups/comopt/software/TSPLIB95/`.

Table 5: The performance improvements of the genetic algorithms by the EGR for the TSP.

Graph (Opt)	Reordering	Cycle Length (%)		$\sigma/\sqrt{m}$	Gen	Time (s)
kroA100	None	33628.59	(58.014)	350.63	83522	1.48
(21282)	EGR	**32683.45**	**(53.573)**	256.61	66292	1.20
lin105	None	23098.80	(60.643)	276.87	93587	1.74
(14379)	EGR	**22543.29**	**(56.779)**	204.72	74770	1.40
kroA200	None	60492.45	(105.981)	550.81	158954	4.62
(29368)	EGR	**55933.18**	**(90.456)**	454.40	147145	4.28
lin318	None	110988.66	(164.076)	768.85	3367045	177.12
(42029)	EGR	**103874.41**	**(147.149)**	761.48	2981903	159.44
att532	None	89793.99	(224.330)	643.55	9065603	804.05
(27686)	EGR	**80501.64**	**(190.767)**	359.32	6433279	562.54
rat575	None	21168.03	(212.536)	114.21	8245914	779.75
(6773)	EGR	**19760.76**	**(191.758)**	92.27	8204245	789.51
gr666	None	999007.49	(239.385)	5464.65	666115	67.38
(294358)	EGR	**904613.27**	**(207.317)**	4459.56	485421	46.96
rat783	None	33699.82	(282.692)	169.30	537883	56.32
(8806)	EGR	**32170.60**	**(265.326)**	157.04	545666	57.13

deviation over the square root of the number of runs ($\sigma/\sqrt{m}$), the average generation (Gen), and the average running time (Time). We can see that the genetic algorithms with the EGR significantly outperforms those with no reordering. This indicates again that the EGR is indeed helpful to crossover efficiency. It is also notable that the genetic algorithms with the EGR consume less running time than the others. These results show that entropic epistasis carries useful information for designing efficient evolutionary algorithms.

5 Concluding Remarks

In this paper, we showed that a theory of epistasis can be well established based on Shannon's information theory. As a consequence, we presented four new epistasis-related measures, significance, entropic epistasis, mean significance, and mean entropic epistasis. The definitions of the measures are based on a new interpretation of the term contribution in defining the epistasis. Comparing the entropic epistasis with the epistasis variance, it was shown that each measure focuses on different aspects of the fitness functions. The proposed measures are microscopic and problem-independent. Moreover, they can be used in either an algorithm-independent or -dependent way in evolutionary algorithms.

From the tests on three important optimization problems, the Royal Road function, the NK-landscape function, and the traveling salesman problem, it was empirically shown that the proposed measures quantify their object features well. By the simple application of entropic epistasis to rearranging the loci of variables in the chromosome, we showed how entropic epistasis can be used for improving the performance of evolutionary algorithms. The applications of entropic epistasis, however, are not limited to that; their possible applications include most topics related to epistasis. For example,

we can try to design new operators such as mutation and crossover based on analyses using the measures. Also, the measures can be used for detecting the linkage structure of given fitness functions in either an algorithm-independent or -dependent way in black-box optimizations. Applications of the measures to the estimation of distribution algorithm (EDA) are also challengeable. We hope these measures to be widely used in the future.

Finally, it should be noted that the entropic epistasis is a measure sensitive to the involved parameters such as the number of variables, the size of the sample set, and the problem size; thus much attention should be paid to comparing the measure values with different parameter settings.

Acknowledgments

The authors would like to thank Yong-Hyuk Kim and Sung-Soon Choi for the discussions. They also wish to acknowledge the helpful comments from the anonymous referees, which improved this paper. This work was partly supported by Brain Korea 21 Project. The ICT at Seoul National University provided research facilities for this study.

References

Bethke, A. D. (1981). *Genetic Algorithms as Function Optimizers*. Ph.D. thesis, University of Illinois, Urbana, IL.

Boese, K. D., Kahng, A. B., and Muddu, S. (1994). A new adaptive multi-start technique for combinatorial global optimizations. *Operations Research Letters*, 16(2):101–113.

Bui, T. N. and Moon, B. R. (1994). A new genetic approach for the traveling salesman problem. In Michalewicz, Z. et al., editors, *Proc. IEEE Conference on Evolutionary Computation*, pages 7–12, IEEE Press, Piscataway, NJ.

Chiu, D., Wong, A., and Cheung, B. (1991). Information discovery through hierarchical maximum entropy discretization and synthesis. In Piatetsky-Shapiro, G. and Frawley, W. J., editors, *Knowledge Discovery in Databases*. The MIT Press, Cambridge, MA.

Cover, T. and Thomas, J. (1991). *Elements of Information Theory*. John Wiley & Sons, New York.

Davidor, Y. (1990). Epistasis variance: Suitability of a representation to genetic algorithms. *Complex Systems*, 4(4):369–383.

Forrest, S. and Mitchell, M. (1993). Relative building-block fitness and the building-block hypothesis. In Whitley, L. D., editor, *Foundations of Genetic Algorithms 2*, pages 109–126, Morgan Kaufmann Publishers, San Francisco, CA.

Garey, M. R. and Johnson, D. S. (1979). *Computers and Intractability: A Guide to the Theory of NP-Completeness*. W. H. Freeman and Company, New York.

Goldberg, D. E. (1987). Simple genetic algorithms and the minimal deceptive problem. In Davis, L., editor, *Genetic Algorithms and Simulated Annealing*, pages 74–88, Morgan Kaufmann Publishers, San Francisco, CA.

Goldberg, D. E. (1989a). Genetic algorithms and Walsh functions: Part I, a gentle introduction. *Complex Systems*, 3:129–152.

Goldberg, D. E. (1989b). Genetic algorithms and Walsh functions: Part II, deception and its analysis. *Complex Systems*, 3:153–171.

Heckendorn, R. B. and Wright, A. H. (2004). Efficient linkage discovery by limited probing. *Evolutionary Computation*, 12(4):517–545.

Holland, J. (1992). *Adaptation in Natural and Artificial Systems*. The MIT Press, Cambridge, MA.

Horn, J. and Goldberg, D. E. (1995). Genetic algorithm difficulty and the modality of fitness landscapes. In Whitley, L. D. and Vose, M. D., editors, *Foundations of Genetic Algorithms 3*, pages 243–270, Morgan Kaufmann Publishers, San Francisco, CA.

Jones, T. and Forrest, S. (1995). Fitness distance correlation as a measure of problem difficulty for genetic algorithms. In Eshelman, L. J., editor, *Proceedings International Conference on Genetic Algorithms*, pages 184–192, Morgan Kaufmann Publishers, San Francisco, CA.

Kargupta, H. (1995). Signal-to-noise, crosstalk, and long range problem difficulty in genetic algorithms. In Eshelman, L. J., editor, *Proceedings International Conference on Genetic Algorithms*, pages 193–200, Morgan Kaufmann Publishers, San Francisco, CA.

Kauffman, S. A. (1989). Adaptation on rugged fitness landscapes. In Stein, D. L., editor, *Lectures in the Sciences of Complexity*, pages 527–618. Addison-Wesley, Reading, MA.

Larrañaga, P. and Lozano, J. A. (2002). *Estimation of Distribution Algorithms: A New Tool for Evolutionary Computation*. Kluwer Academic Publishers, Boston, MA.

Lin, S. and Kernighan, B. (1973). An effective heuristic algorithm for the traveling salesman problem. *Operations Research*, 21(2):498–516.

Munetomo, M. and Goldberg, D. E. (1999). Linkage identification by non-monotonicity detection for overlapping functions. *Evolutionary Computation*, 7(4):377–398.

Pelikan, M., Goldberg, D. E., and Lobo, F. (2002). A survey of optimization by building and using probabilistic models. *Computational Optimization and Applications*, 21(1):5–20.

Rana, S., Heckendorn, R. B., and Whitley, D. (1998). A tractable Walsh analysis of SAT and its implications for genetic algorithms. In Rich, C. and Mostow, J., editors, *Proceedings National Conference on Artificial Intelligence*, pages 392–397, AAAI Press, Menlo Park, CA.

Reeves, C. R. and Wright, C. C. (1995a). An experimental design perspective on genetic algorithms. In Whitley, L. D. and Vose, M. D., editors, *Foundations of Genetic Algorithms 3*, pages 7–22, Morgan Kaufmann Publishers, San Francisco, CA.

Reeves, C. R. and Wright, C. C. (1995b). Epistasis in genetic algorithms: An experimental design perspective. In Eshelman, L. J., editor, *Proceedings International Conference on Genetic Algorithms*, pages 217–224, Morgan Kaufmann Publishers, San Francisco, CA.

Seo, D. I., Kim, Y. H., and Moon, B. R. (2003a). New entropy-based measures of gene significance and epistasis. In Cantú-Paz, E. et al., editors, *Genetic and Evolutionary Computation Conference*, pages 1345–1356, Springer-Verlag, Berlin, Germany.

Seo, D. I. and Moon, B. R. (2002). Voronoi quantized crossover for traveling salesman problem In Langdon, W. B. et al., editors, *Genetic and Evolutionary Computation Conference*, pages 544–552, Springer-Verlag, Berlin, Germany.

Seo, D. I. and Moon, B. R. (2003). A survey on chromosomal structures and operators for exploiting topological linkages of genes. In Cantú-Paz, E. et al., editors, *Genetic and Evolutionary Computation Conference*, pages 1357–1368, Springer-Verlag, Berlin, Germany.

Shannon, C. E. (1948). A mathematical theory of communication. *Bell System Technical Journal*, 27:379–423, 623–656.

Streeter, M. J. (2004). Upper bounds on the time and space complexity of optimizing additively separable functions. In Deb, K. et al., editors, *Genetic and Evolutionary Computation Conference*, pages 186–197, Springer-Verlag, Berlin, Germany.

Theil, H. (1972). *Statistical Decomposition Analysis*. North-Holland Publishing Company, Amsterdam, Netherlands.

Weinberger, E. D. (1996). NP completeness of Kauffman's NK model, a tunable rugged fitness landscape. Technical report 96-02-003, Santa Fe Institute, Santa Fe, NM.

Appendix

A Frequently Used Theorems

Theorem 18 (Chain Rule for entropy) *For random variables* $X_1, \ldots, X_n$,

$$H(X_1, \ldots, X_n) = \sum_{i=1}^{n} H(X_i | X_{i-1}, X_{i-2}, \ldots, X_1).$$

Proof. See (Cover and Thomas, 1991, p. 21). □

Theorem 19 (Chain Rule for information) *For random variables* $X_1, \ldots, X_n$ *and* Y,

$$I(X_1, \ldots, X_n; Y) = \sum_{i=1}^{n} I(X_i; Y | X_{i-1}, X_{i-2}, \ldots, X_1).$$

Proof. See (Cover and Thomas, 1991, p. 22). □

Theorem 20 (Independence Bound on entropy) *For random variables* $X_1, \ldots, X_n$,

$$H(X_1, \ldots, X_n) \leq \sum_{i=1}^{n} H(X_i)$$

with equality if and only if X_i *are independent.*

Proof. See (Cover and Thomas, 1991, p. 28). □

Theorem 21 (Independence Bound on conditional entropy) *For random variables X_1, ..., X_n, and Y,*

$$H(X_1, \ldots, X_n|Y) \leq \sum_{i=1}^{n} H(X_i|Y)$$

with equality if and only if X_i are conditionally independent given Y.

Proof. By the Independence Bound on entropy,

$$\begin{aligned} H(X_1, \ldots, X_n|Y) &= \sum_{y \in \mathcal{Y}} p(y) H(X_1, \ldots, X_n|Y = y) \\ &\leq \sum_{y \in \mathcal{Y}} p(y) \sum_{i=1}^{n} H(X_i|Y = y) \\ &= \sum_{i=1}^{n} \sum_{y \in \mathcal{Y}} p(y) H(X_i|Y = y) \\ &= \sum_{i=1}^{n} H(X_i|Y) \end{aligned}$$

with equality if and only if X_i are conditionally independent given Y. □