The generalization of mutual information as the information *between* a set of variables: The information correlation function hierarchy and the information structure of multi-agent systems.

David R. Wolf

NASA Ames Research Center
MS 269-2
Moffet Field, CA 94035

greywolf@ptolemy.arc.nasa.gov

1 May 96

1 Introduction

The topic of this paper is a hierarchy of information-like functions, here named the *information correlation* functions, where each function of the hierarchy may be thought of as the information *between* the variables it depends upon. The information correlation functions are particularly suited to the description of the emergence of complex behaviors due to many-body or many-agent processes. They are particularly well suited to the quantification of the decomposition of the information carried among a set of variables or agents, and its subsets. In more graphical language, they provide the information theoretic basis for understanding the synergistic and non-synergistic components of a system, and as such should serve as a forceful toolkit for the analysis of the complexity structure of complex many agent systems.

The information correlation functions are the natural generalization to an arbitrary number of sets of variables of the sequence starting with the entropy function (one set of variables) and the mutual information function

(two sets). We start by describing the traditional measures of information (entropy) and mutual information.

Then the *between operator* is introduced and the algebra of the between operator is discussed. Then the information correlation function hierarchy is defined using the between operator, and we show that it has the properties desired of the information due to a set of random variables, but not due to any proper subset of those variables, facilitating the interpretation of it being the information between the named variables.

We end with a graphical exploration of the ideas presented here using the Ising model. The full Ising system mathematics is not disussed in detail in this brief introduction. though it may be found in the author's Ph.D. dissertation [6]. For the Ising model we take subsets of sites in the Ising system, and construct the information correlation functions for those sites. The information correlation functions are found to describe well the regions of the phase space of the Ising system where correlations occur that must be described by high order distribution functions ("*complex regions*"), and the regions where low order reductions of the distribution functions are sufficient for description ("*simple regions*").

In [6] the development of the information correlation functions proceeded thru an analysis of the cumulant hierarchy. The current discussion is far simpler, while still capturing many of the subtleties.

The information correlation function hierarchy has been known for some time in different guises, having been noted by the fluid theorists, see for example [1], but it appears to not have been interpreted as an information quantity measuring the information between a set of random variables before, nor has it apparently been applied to infer regions of complexity in systems before. These are the contributions of current work presented here.

2 The information in one variable and the information between two variables

Given a random variable A, the Shannon entropy of A measures the uncertainty of the outcome of A *before* A is seen, and it measures the amount of information in A *after* A is seen. The Shannon entropy of A is given by

$$S(A) := -\sum_a p(a)\, log(p(a)) \tag{1}$$

where $p(a)$ is the probability that symbol $a \in A$ appears. (We confine the discussion to the discrete case in this paper - the continuous case requires

only a brief extension. Also, all of the results of this paper hold when the logarithm is taken with any base.)

The mutual information [4] of, or the information between, two random variables A and B, is given by

$$M(A,B) := \sum_{a,b} p(a,b)\, log\left(\frac{p(a,b)}{p(a)\,p(b)}\right) \tag{2}$$

where the sum is over the mutually exclusive outcomes of the pair of random variables (A,B), and where $p(a,b)$ is the probability that symbol $(a,b) \in A \times B$ appears. Examine equation 2 for a moment. When the joint probability distribution $p(a,b)$ describes two *independent* variables it necessarily factors as $p(a,b) = p(a)p(b)$, leading to the logarithm, and therefore the mutual information, being zero $M(A,B) = 0$. On the other hand, if both of the processes A, B are identical, then $p(a,b) = p(a) = p(b)$ when $a = b$, and $p(a,b) = 0$ otherwise. Then for this case $M(A,B) = -\sum_a p(a)\, log(p(a)) = -\sum_b p(b)\, log(p(b))$, which for this case is seen to be the Shannon entropy of either random variable. (Zero joint probability cases make a contribution of zero to the mutual information since $\lim_{x\to 0} x log(x) = 0$). The mutual information may also be seen as the amount that the uncertainty in A is reduced when B is seen:

$$M(A,B) = S(A) - S(A \mid B) \tag{3}$$

where the second quantity is read as the "uncertainty of A given B", and is given by

$$S(A \mid B) := -\sum_{a,b} p(a,b)\, log(p(a \mid b)) \tag{4}$$

where $p(a \mid b)$ is the probability that $a \in A$ appears given that $b \in B$ is seen, $p(a \mid b) := p(a,b)/p(b)$ (and $p(b) \neq 0$). It is straightforward to show from this that the maximum mutual information possible when A is fixed is $M(A,B) = S(A)$: the minimal uncertainty that A can have after B is seen (the minimal $S(A \mid B)$) is zero (which occurs when B is taken to be A, for example). Note that the mutual information may be written also as

$$M(A,B) = S(B) - S(B \mid A) \tag{5}$$

and symmetrically as

$$M(A,B) = S(A) + S(B) - S(A,B). \tag{6}$$

In the next sections we will develop the generalized mutual information.

3 The information between several variables

In the previous section the mutual information was discussed in the context of measuring the information between two sets of variables. What properties are desired of the "information between variables" for cases when more than two variables are involved, which we identify as the "generalized mutual information"?

Consider three variables, A, B, and C. There are immediately the three quantities $S(A)$, $S(B)$, and $S(C)$ which quantify the uncertainties or informations in single variables, but say little about information between any of them. There are also the three mutual informations $M(A,B)$, $M(B,C)$, and $M(C,A)$ which together quantify the information between all pairs of two variables. What, though, is the information *between all three variables*? Could it be the quantity $S(A,B)-S(A,B \mid C)$? No, because this is quantity is not permutation symmetric, and we expect the information between a set of random variables to be unchanged by their ordering. Thus *symmetry* is an important property of the desired quantity.

Consider the notion of the information between a *set* of variables, that information due *only* to the whole set and not to any proper subset of the set. We might expect that this information, in the case where one variable of the set is independent of all others, is zero because there is a natural decomposition of the information that the variables provide into the informations from the independent subsets. We see that the quantity above also fails to have the property that it generally is zero when any of the variables is independent of the others. Generalizing to cases where there is a natural decomposition of the set into independent subsets, we have another important desired property of the information between a set of variables: The information between any set of variables should be zero whenever any proper subset of variables from the set is independent of its complement in the set. We call this property *subset decomposition.*

Taking a heuristic approach, we might pretend as if a *between process* existed and take the information between three variables to be the information in the *between process* of the *between process* of the first two variables, and the third variable. For example, imagine a channel between A and B through which passes only the necessary information in A which could be used to reduce the uncertainty of B. This channel is then a representation of the information *between* A and B. Call the process in this channel $A \cap B$. Then the information *between* all three variables is given by $(A \cap B) \cap C$. If the between operation is to be meaningful in this sense, one might ex-

pect that this reduction be associative, that is the reduction above should be equivalnt to the reduction $A \cap (B \cap C)$. This leads to *associativity* as another desired property.

Similarly, *commutativity* is clearly a desired property, that is $A \cap B = B \cap A$ is desired since the location of the channels is irrelevant in the present discussion.

In the next section we define an algebra for the *between operator*, and demonstrate that the desired properties are satisfied. The desired properties for the information between variables are *symmetry, associativity, commutativity, and subset decomposition.* We will assume known properties of the entropy, the entropy of joint processes, and one other assumption, the distributive property, and find that symmetry, associativity, commutativity are reuired for consistency, while subset decomposition follows as an implicit property.

In later sections we define the information correlation function hierarchy and lend to each member of the hierarchy the interpretation of being the information between the named variables, the generalized mutual information.

We end the paper with a concrete example using the Ising model, computing the entropies of subsets of sites in the Ising spin system, and constructing the information correlation functions for those sites. The mutual information of a sequence of spin sites has a particularly simple form which aids in understanding the structure of the information correlation function hierarchy. We note that the information correlation functions serve as quantitative measures in describing the hierarchy of subsystem complexities of a system.

4 The algebra of the *between* operator ($\cap$)

The information between two random variables X_1 and X_2 is the mutual information of those variables. We motivate the definition of the between operator by writing the mutual information as the entropy of the between process

$$M(X_1, X_2) = S(X_1 \cap X_2) \tag{7}$$

We do not define the between process, nor do we ever need to define this process, since at this and any point equation 7 may be viewed as a purely motivational statement, and since the definitional statements to follow may be taken without reference to any between process. Here and as previously

stated, the notion of the between process is a convenient heuristic in discussion. Define the joint process operator of two random variables as

$$X_1 \cup X_2 := \{X_1, X_2\} \tag{8}$$

where we note that the joint process operator $\cup$ is intrinsically associative and commutative. Equation 8 leads, using equations 2 and 7, to the equation we take as the definition of the entropy of the between operator of two processes,

$$S(X_1 \cap X_2) := S(X_1) + S(X_2) - S(X_1 \cup X_2) \tag{9}$$

Because the joint process operator $\cup$ is commutative, and because addition is commutative, consistency requires that the between operator $\cap$ is commutative,

$$X_1 \cap X_2 \;:=\; X_2 \cap X_1. \tag{10}$$

We assume one more property, distributivity,

$$\begin{aligned} (X_1 \cap X_2) \cup X_3 &:= (X_1 \cup X_3) \cap (X_2 \cup X_3) \\ (X_1 \cup X_2) \cap X_3 &:= (X_1 \cap X_3) \cup (X_2 \cap X_3). \end{aligned} \tag{11}$$

Assuming consistency of equations 9, 10 and 11, the between process is required to be associative. Thus

$$(X_1 \cap X_2) \cap X_3 \;:=\; X_1 \cap (X_2 \cap X_3) \tag{12}$$

It is clear at this point that the $\cap$ and $\cup$ operators form a *boolean algebra* equivalent to set intersection and union, respectively. As things stand now, because of this algebra and equation 9 we are always able to reduce the entropy of an expression involving the between operator to a sum of entropies of expressions involving only joint processes. Since we understand the entropy of joint processes without further clarification, the mathematics of the entropy of expressions involving between processes is consistent. With the understanding that values of entropy expressions involving the between operator are to be determined by their equivalent expressions involving only the joint operator, all values of entropies of expressions involving between operators are well-defined.

The between operator has been successfully defined. The degree to which the between process may be defined therefore has significant impact on the degree to which the interpretation of the entropies of an expression involving

the between operator may be interpreted as an entropy of a real process. Since we do not construct between processes in this paper, this question of interpretation remains open, and we only point to the fact that the information correlation functions have all of the desired properties desired of the information between, or generalized mutual information of, a set of variables, except for one so-far unmentioned property - positivity. In fact, it is not difficult to show that there exist joint processes of several variable for which the entropy of the metaphoric between discrete process is *negative.* Thus it is ill advised to attempt to construct any between process, since the entropy of any discrete process is positive. A related interpretation issue which remains is this sign issue - from the results to follow the *magnitude* of the information correlation functions is highly indicative of the complexity due to the full process and not to any subprocess, however the dominant sign of the information correlation functions changes from one order to the next. Perhaps there is a rational interpretation in terms of the confusion a new member of the group of variables brings to the group: perhaps the mathematics is simply telling us that adding a new member to a group to bring the group to an odd number of members is always prone to causing confusion!

In the next section we present a different approach to the information correlation functions, and demonstrate the subset independence property, the only remaining property desired of the information between a set of variables, the generalized mutual information.

5 The information correlation function hierarchy

Using the algebraic properties of the information between random variables given by equations 8–11, we find that the information between any set of random variables may be expanded in terms of a sum of informations of single and joint processes. We have then the following hierarchy of functions, which we call the *information correlation functions.*

$$\begin{aligned}
S(X_1) &= S(X_1) \\
S(X_1 \cap X_2) &= S(X_1) + S(X_2) - S(X_1 \cup X_2) \\
S(X_1 \cap X_2 \cap X_3) &= S(X_1) + S(X_2) + S(X_3) \\
&\quad - S(X_1 \cup X_2) - S(X_2 \cup X_3) - S(X_3 \cup X_1) \\
&\quad + S(X_1 \cup X_2 \cup X_3) \qquad (13) \\
&\ldots
\end{aligned}$$

In general, the pattern is that the signs are $(-1)^{(k+1)}$, where k is the number of random variables mentioned as arguments, and all subsets appear as arguments exactly once.

There have been several previous characterizations of the *redundancy* of a set of random variables (see for example [3], all involving a comparison of the full distribution of the variables to the product of the marginal distributions of the single variables. However, none of these characterizations is able to properly sort out the information due to a set of more than two variables and not to any proper subset of that set. For more details on this see [6].

6 Another look at the information correlation hierarchy

Consider the n variable joint process $(X_1, \ldots, X_n)$ and let $p_k(i_1, \ldots, i_k)$, $k < n$, denote the marginal probability distribution of the k variable joint process $(X_{i_1}, \ldots, X_{i_k})$, with its arguments being given lowest index to highest index ($i_m < i_n$ for $m < n$). Now consider the hierarchy

$$\begin{aligned}
log(p_1(i)) &= \phi_1(i) \\
log(p_2(i,j)) &= \phi_2(i,j) + \phi_1(i) + \phi_1(j) \\
&\ldots \\
log(p_n(1,\ldots,n)) &= \phi_n(1,\ldots,n) + \ldots + \phi_2(1,2) \\
&\quad +\phi_2(1,3) + \ldots + \phi_1(1) + \ldots + \phi_1(n)
\end{aligned} \tag{14}$$

We may always solve equations 14 for the ϕ_i. The result of doing this is

$$\begin{aligned}
\phi_1(i) &= log(p_1(i)) \\
\phi_2(i,j) &= log(p_2(i,j)) - log(p_1(i)) - log(p_1(j)) \\
&\ldots \\
\phi_n(1,\ldots,n) &= log(p_n(1,\ldots,n)) - \sum_{[n-1]} log(p_{n-1}) + \\
&\quad \ldots + (-1)^{n+1} \sum_{[1]} log(p_1)
\end{aligned} \tag{15}$$

where $\sum_{[k]}$ indicates that the summation is over all subsets of k arguments. Now, multiply the right sides of equations 15 by $(-1)^n$ and then average over p_n to find, as we will see immediately after the next equation, the hierarchy

of *information correlation functions* $C_k := \langle(-1)^k \phi_k\rangle$:

$$
\begin{aligned}
C_1(i) &:= -\sum_i p_1(i)\, log(p_1(i)) \\
C_2(i,j) &:= \sum_{ij} p_2(i,j)\; log\left(\frac{p_2(i,j)}{p_1(i)p_1(j)}\right) \\
&\cdots \\
C_n(1,\ldots,n) &:= (-1)^n \sum_{1,\ldots,n} p(1,\ldots,n)\; log\left(\frac{p_n(1,\ldots,n)\prod_{[n-2]} p_{n-2}\cdots}{\prod_{[n-1]} p_{n-1}\cdots}\right)
\end{aligned}
\tag{16}
$$

where $\prod_{[m]}$ indicates that the product is over all subsets of m arguments. Comparing equations 13 and 16 shows that

$$S(X_1 \cap \ldots \cap X_n) = C_n(1,\ldots,n) \tag{17}$$

so that at this point in the paper we have developed the information correlation hierarchy in two distinct ways, the first development being with the boolean process algebra of section 4. Further, we note that the ϕ_k are the terms in the exponents of a product expansion of p_n in terms of subsets of it arguments. From equations 15 we easily find

$$p_n = e^{\phi_n} \times \prod_{[n-1]} e^{\phi_{n-1}} \times \ldots \times \prod_{[1]} e^{\phi_1} \tag{18}$$

From equation 18 it is clear that if the probability distribution p_n factors into the product of two probability distributions, p_r and p_s with $n = r + s$, (which occurs iff the sets of variables corresponding to the arguments of p_r and p_s are independent of each other) that there will be no ϕ_k present for $k > max(r,s)$. Thus, such ϕ_k, and their corresponding information correlation functions, taken in this paper to represent the generalized mutual information of, or information between, the arguments taken as a set, are zero. This establishes the last desired property, that whenever two such sets of independent variables occur, there will be no information between the full set of variables.

In the next sections we develop the entropies of the Ising system, develop the information correlation function hierarchy from these entropies, and then demonstrate that the information correlation functions are useful in determining where in the parameter space of the Ising system "complex" behavior (here behavior requiring more of the Ising spins to describe) is located, and conversely, where the Ising system behaves "simply".

7 Information correlation functions in the Ising System

The following graphs present the various quantities that might be considered in an analysis of a complex system, for the linear (1D) Ising system. These quantities are entropies, moments, correlations, cumulants, and the information correlation functions.

Briefly, the Ising system considered here is a one-dimensional chain of nearest neighbor coupled spins, taking the values ± 1, with energy (Hamiltonian) of the state σ (representing the values of the spins) given by

$$E(\sigma) = -\sum_{i=1}^{n}[Qs_i s_{i+1} + Rs_i]. \tag{19}$$

Here, the parameters Q and R are spin-spin coupling strength, and spin-external field strength respectively, and the probability of states is proportional to $e^{-\beta E(\sigma)}/Z_n(\beta)$ with $\beta = 1/kT$, and T the temperature. The normalization constant (partition function) is given by

$$Z_n(\beta) = \sum_{\sigma} e^{-\beta E(\sigma)} \tag{20}$$

so that $P(\sigma)$ is given by

$$P_n(\sigma) = \frac{e^{-\beta E(\sigma)}}{Z_n(\beta)} \tag{21}$$

The parameter b seen on the graphs is the β defined above, proportional to the inverse temperature. The parameter Q is taken to be one, so that positive values of b correspond to the ferromagnetic region, while negative values of b correspond to the antiferromagnetic region.

The mathematics presented in [6] concerns the calculation of the various quantities mentioned above in closed form for this system, for any subset of the spins. In particular, the graphs shown here range over sets of spins of sizes one to four.

The important thing to note, above all, is that the information correlation functions are the most diagnostic for complexity structure occurring because of the synergistic interaction of the spins. They indicate when set of spins contains information not attributable to any subset of the set, the synergistic content. They are zero whenever the set of spins may be decomposed into subsets which behave largely independently of one another, and nonzero otherwise.

Graphs of the entropy per spin for orders $1-4$ including both the ferromagnetic $\beta > 0$ and the antiferromagnetic $\beta < 0$ appear in figures 1- 4. Note the sharp difference in the antiferromagnetic case between the first and second order entropies, and the marked similarities between the second and higher order entropies. Note that the logarithms are base e, and that β appears as b in the axis label.

Graphs of the information correlation functions of $2, 3$ and 4 neighboring spins are given in figures 5-7. Note that these functions are always largest in the regions of antiferromagnetic bahevior where is is impossible to decompose the system into simpler subsystems.

Graphs of the moments of orders 1 through 4 of neighboring spins appear in figures 8- 11. As seen in the graphs, the moments are particulary disgnostic of transition regions, but not complexity.

Graphs of the correlation functions of orders 2 through 4 of neighboring spins appear in figures 12- 14.

The graph of the cumulant function of order 4 of neighboring spins appears in figure 15 (the first three correlation and cumulant functions are equal by order).

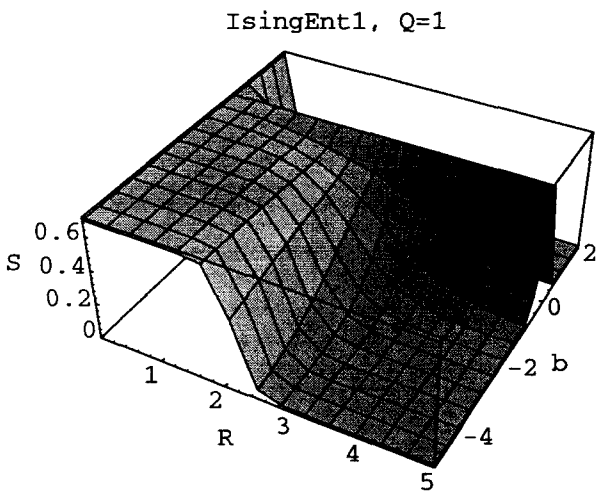

Figure 1: First order entropy per spin of the Ising system. Entropy of one spin.

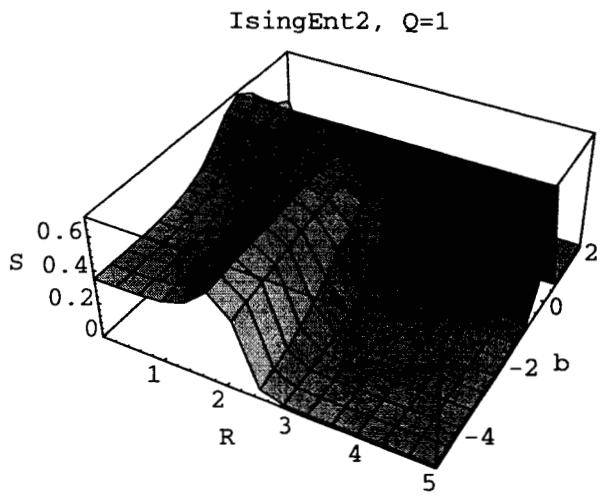

Figure 2: Second order entropy per spin of the Ising system. Note the difference between the first and second order entropies, and the similarities of the second-fourth order entropies. The bump on the antiferromagnetic phase side occurs as the external field is increased, and indicates a transition between the $\uparrow\downarrow\ldots$ states, and the $\uparrow\uparrow\ldots$ state.

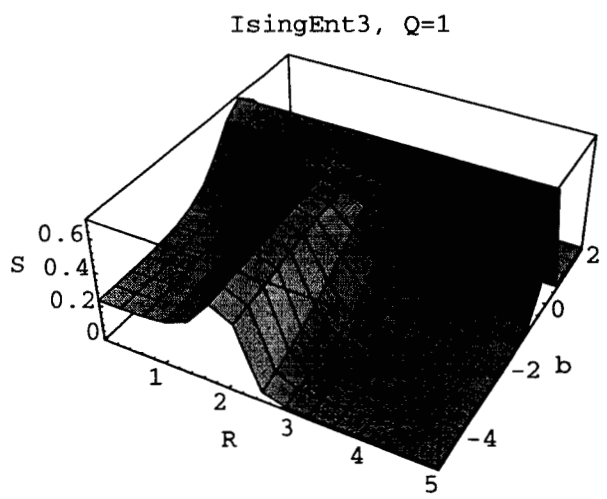

Figure 3: Third order entropy per spin of the Ising system.

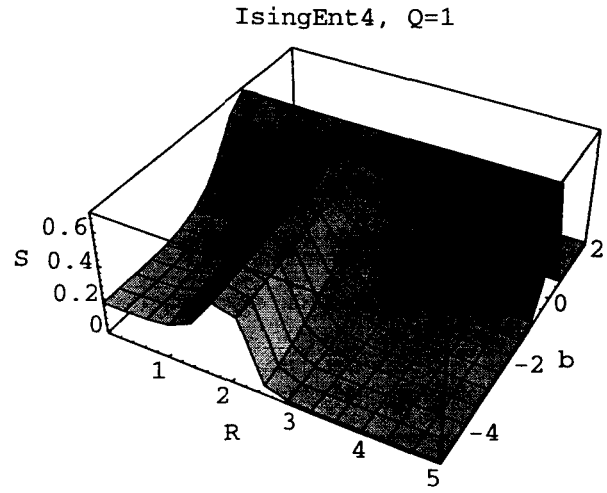

Figure 4: Fourth order entropy per spin of the Ising system.

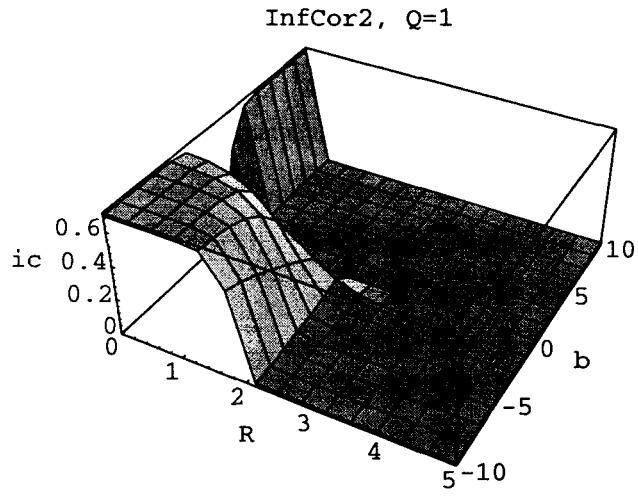

Figure 5: Second order information correlation function of the Ising system. Information correlation of two neighboring spins. Note that the information correlation functions are similar up to a sign at each order for this system. The information correlation functions are the mutual information of the first and last spins along the chain in the k spins considered at order k times $(-1)^k$

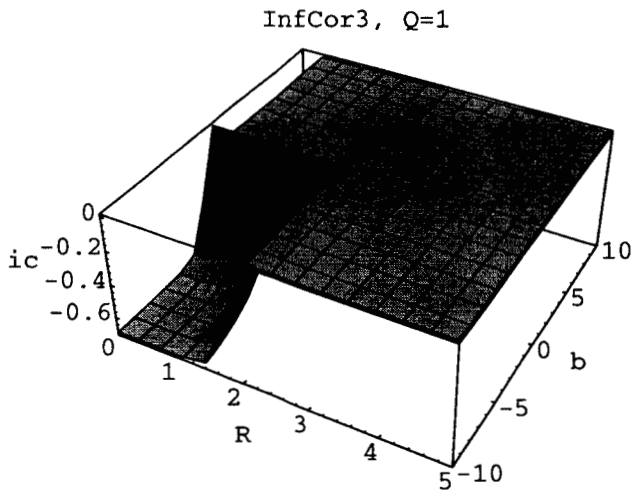

Figure 6: Third order information correlation function of the Ising system. Information correlation of three neighboring spins.

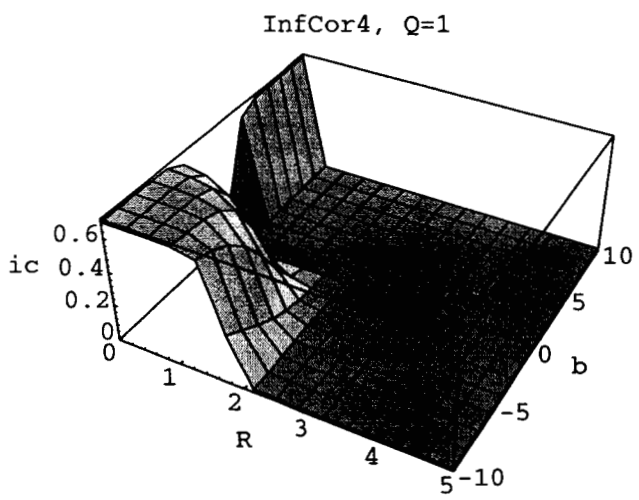

Figure 7: Fourth order information correlation function of the Ising system. Information correlation of four neighboring spins.

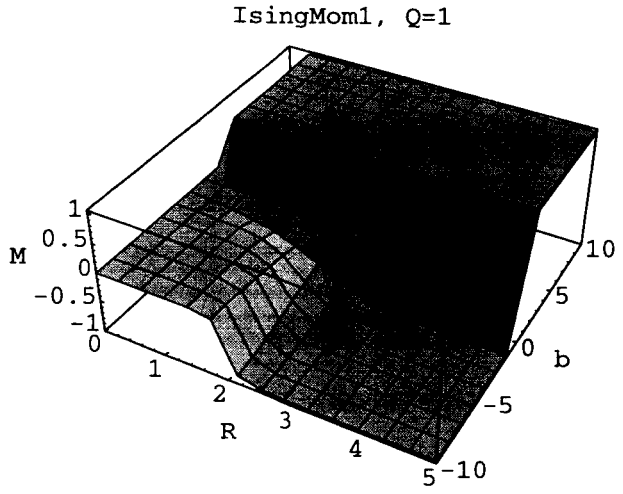

Figure 8: First order moment - moment of one spin. Note the transitional behavior in the region of the increase in the entropy, where the states $\uparrow\downarrow\ldots$ becomes dominated by $\uparrow\uparrow\ldots$.

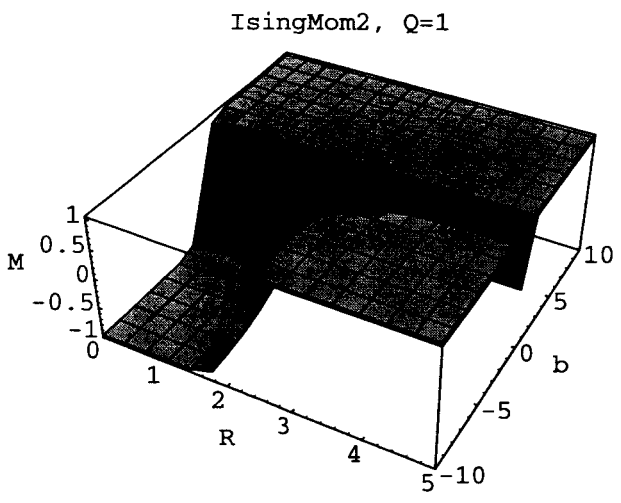

Figure 9: Second order moment of two neighboring spins.

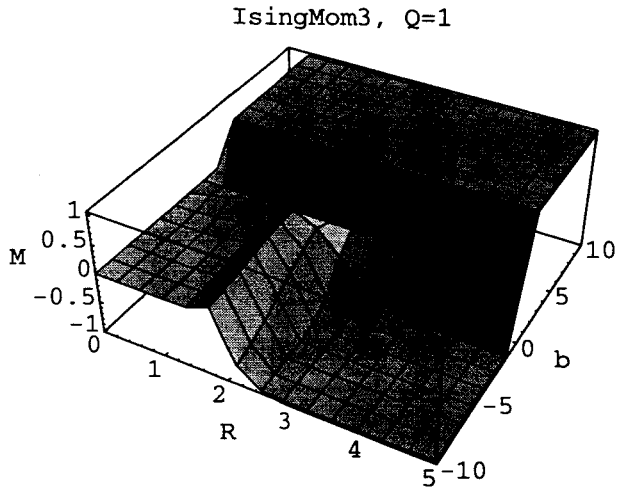

Figure 10: Third order moment of three neighboring spins.

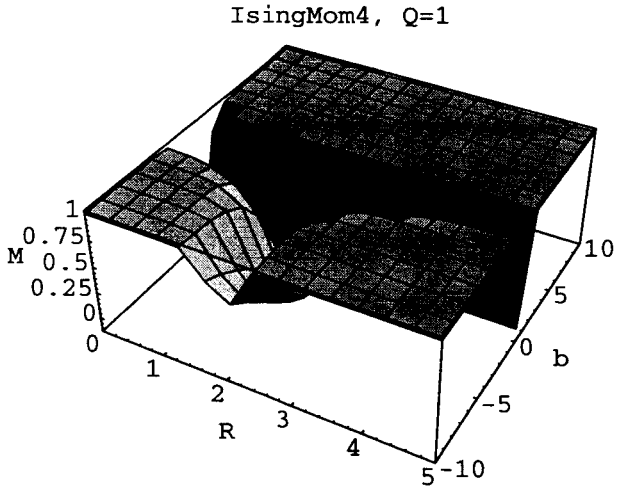

Figure 11: Fourth order moment of four neighboring spins.

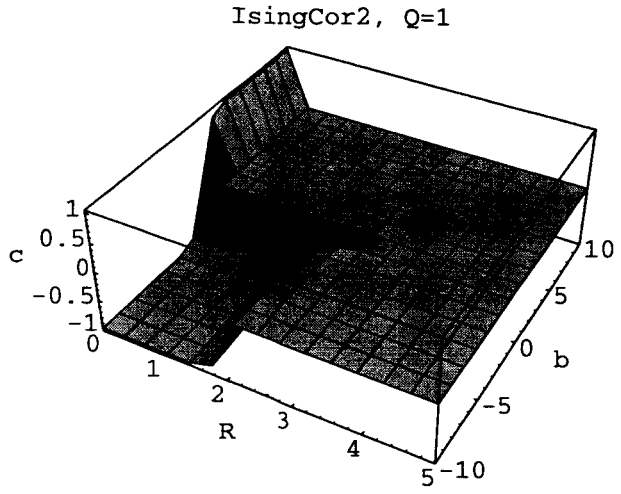

Figure 12: Second order correlation of two neighboring spins.

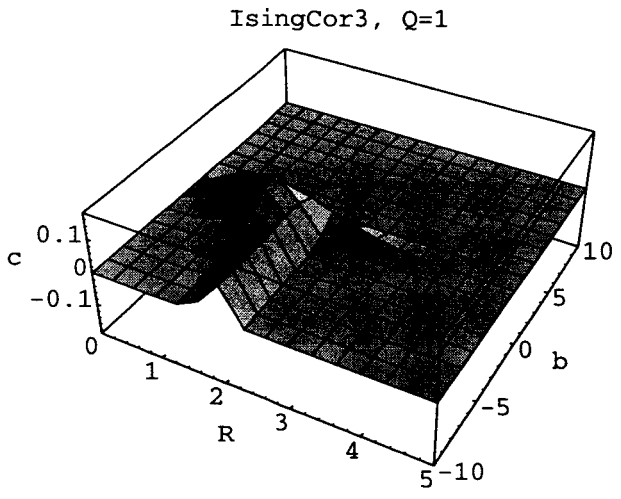

Figure 13: Third order correlation of three neighboring spins.

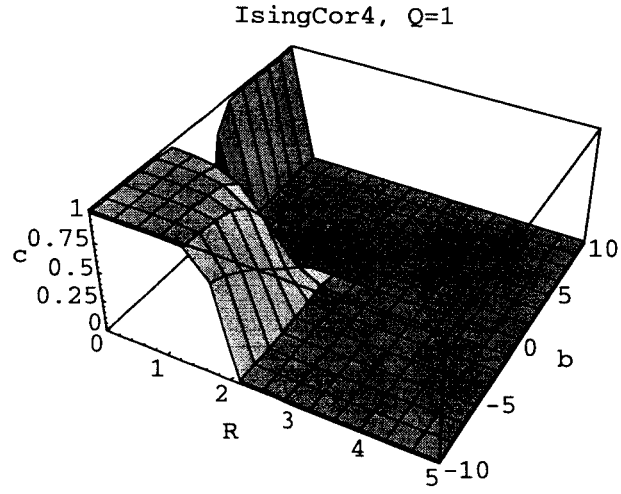

Figure 14: Fourth order correlation of four neighboring spins.

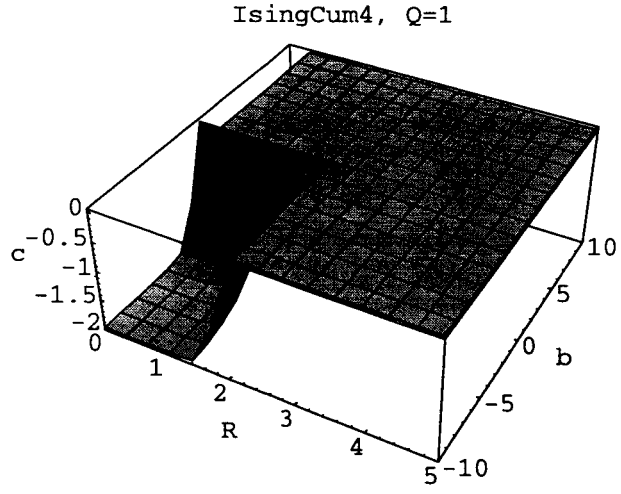

Figure 15: Fourth order cumulant of four neighboring spins.

References

[1] I. Z. Fisher. *Statistical Theory of Liquids.* University of Chicago Press, Chicago, IL, 1964.

[2] Nigel Goldenfeld. *Lectures on Phase Transitions and the Renormalization Group.* Addison-Wesley, Reading, MA, 1992.

[3] Milan Palus. Identifying and quantifying chaos by using information theoretic functionals. In *Time Series Prediction: Forecasting the Future and Understanding the Past*, pages 387–414, 1992.

[4] Claude E. Shannon and Warren Weaver. *The Mathematical Theory of Communication.* University of Illinois Press, Chicago, 1963.

[5] Colin J. Thompson. *Mathematical Statistical Mechanics.* Princeton University Press, Princeton, NJ, 1972.

[6] David R. Wolf. *Information and Correlation in Statistical Mechanical Systems.* University of Texas, http://dino.ph.utexas.edu/ ~ wolf, 1996.