INFORMATION AND CORRELATION IN STATISTICAL MECHANICAL SYSTEMS.

by

DAVID RICHARD WOLF, B.S. Math / Physics

DISSERTATION

Presented to the Faculty of the Graduate School of

The University of Texas at Austin

in Partial Fulfillment

of the Requirements

for the Degree of

DOCTOR OF PHILOSOPHY

THE UNIVERSITY OF TEXAS AT AUSTIN

May 1996

INFORMATION AND CORRELATION IN STATISTICAL MECHANICAL SYSTEMS.

APPROVED BY
DISSERTATION COMMITTEE:

Supervisor: _____

_____

_____

_____

_____

This dissertation is dedicated to
my parents.

Acknowledgements

This work owes much to Toshiki Tajima my dissertation supervisor and committee chairman, the motivating force behind the investigation of the statistical mechanical systems studied here. Much credit also goes to all of those who have helped along the way by providing direction and insight including, but not limited to: David Wolpert of the Santa Fe Institute and TXN, one of my primary mentors and a strong collaborator in the study of the Bayesian estimation problem. (I have limited the content of chapter 9 to my contribution.) Charlie Strauss of the Los Alamos laser studies group. Kenneth Hanson, Greg Cunningham, Rollin Whitman and Alan Matthews of the Los Alamos DX-13 image analysis group. Harry Watanabe, Will Hemsing, Ed Pogue, Jay Boettner and the rest of the gang at DX-13. Alan Lapedes, John Gibson, Stephen Eubank, Doyne Farmer, Martin Casdagli, Mark Milonnas, Steen Rasmussen, Howard Gutowitz, Stephen Pope, Seth Lloyd, Chris Langton and John Siddorowich of the T-13 complex systems group at Los Alamos, in what, in retrospect, was a hey-day of activity centered around Doyne's prediction efforts. Finally, I want to thank all of the friends who have supported me along the way with ideas and friendships, especially the 99 other students at the 21st Street College House, where the Trusteeship during the "year of the Big Fire" has taught and brought me so much. Everything was all the more fine because of all of you.

Part of this work was funded by DOE contract number W-7405-ENG-36. Part of this work was funded by NSF contract number ATM-9401123. Part of this work was funded by the Japanese Energy Research Institute (JAERI). The author thanks the hospitality and support of Theoretical Division and the

Center for Nonlinear Studies of Los Alamos National Laboratory, the Santa Fe Institute, and the Institute for Fusion Studies of the University of Texas, Austin.

INFORMATION AND CORRELATION IN STATISTICAL MECHANICAL SYSTEMS.

Publication No. _____

David Richard Wolf, Ph.D.
The University of Texas at Austin, 1996

Supervisor: Toshiki T. Tajima

Information comes to the researcher or other system in untold forms. Information is carried in physical objects which interact with the observer or system that they influence. Living systems make use of information in subtle ways, to find and make use of sources of materials and energy.

In this work the focus is on the reduced distribution functions and when they yield information of relevance. Information correlation functions, correlation functions, and entropy are of primary interest. The estimation of functions of the underlying distribution is examined.

Several key theorems of statistical mechanics are shown to be consequences of a single theorem on counting labeled partitions, bringing together the cumulant expansion, linked cluster theorem, and Ursell development as consequences of this theorem. The information correlation functions provide a basis for the notion of the information between set of random variables.

The flow of information is closely examined, in both the classical and quantum frameworks. In the Hamiltonian context, when the unexamined part of the distribution is taken to be the maxent distribution, the information flow into the subsystem is shown to be zero.

The Ising model forms the basis of a non-trivial exactly solvable system for examining the correlations and information correlation functions. The expressions for the entropies of any subset of Ising spins are given, and it is shown that the information correlation functions give the mutual information between the first and last spins considered.

The quantum Heisenberg model forms the basis for a non-trivial system exhibiting dynamics. The measurement entropy and the intrinsic entropy are defined and are shown to be related by an inequality. A time-ordered mutual information that is of great interest when examining the setting and measurement of quantum states is introduced.

Estimating the values of functions of the underlying distribution (i.e. entropy, mutual information, etc.) and their uncertainties forms a large portion of the key results. Closed form expressions for the moments of the entropy and the mutual information are given.

Table of Contents

List of Figures

Chapter 1

Introduction

Information is ubiquitous. Information is carried in physical systems by physical objects, and these objects are consumed by the users of that information. Each and every living system makes use of information to ensure its survival and to continue its existence, from finding food to reproducing to finding destructive infiltrating agents (the immune system), information is presented and acted upon. More speculatively, making use of information is perhaps fundamentally what separates life from non-life. Intrinsic to the use of information is the need to be able to recognize the message being sent by reducing the information in the physical object presented to the relevant message. This is a very general statement. To be more specific, in what ways can a reduced description of a system suffice to bring the relevant information to the consumer of that information?

In this dissertation the question of how information is carried in a physical system is examined. The systems studied here are simple, as are all systems which have a presentable analysis. The point of view that the states of the physical system of interest may be treated probabilistically, that there is an underlying distribution which describes the probability that a particular state occurs, is taken thoroughly. Certainly the thermodynamic systems studied here are treated on this basis, but more generally whenever such a distribution exists and is known, or is learnable, the methods of this work apply. Thus throughout everything starts with these distributions.

The relationships between the full distribution and reductions of the full distribution are of prime interest here, for it is the reductions of the full distribution that represent the process of ignoring, perhaps relevant, information. In making the reduction it is important to know just what is being lost, and it is similarly important to know when a reduction is satisfactory and when it is not.

To begin with a simple but important example (see a similar example in [84]), restrict attention to a typewriter symbol set, and suppose that three magnetic tapes are presented. Unknown to the consumer of the tapes, one contains a poem, one contains random noise, but the symbols occur with the same frequencies as those of the poem, and one, the blank tape, contains nothing but one symbol, zero, repeated forever. What is to be done to distinguish them? This is analogous perhaps to a researcher examining a biological system in which there are many long strands of DNA. In some of the cases the strands carry useful information, and in others there is nothing but random bases joined together. In others still there may be long regions of repeated bases which are there only because some enzyme went out of control at some point in the evolution of this genome. What is to be done to recognize the difference between these strands? What is to be done to distinguish useful DNA from non-useful DNA? In the case of the first two tapes, the poem and the noise tapes, both have the same first order entropies. But the second order entropy of the poem is probably lower than that of the noise. And so on. At very high orders though, the poem probably has very low entropy, with the next unseen symbol being nearly predictable. Thus at high orders and seen in terms of the entropies, it appears similar to the third tape, the blank tape. The blank tape has zero entropy at all orders. Clearly, understanding the entropy of a system at just one order is not enough to make a clear assessment of the content. This example makes it clear that in order to understand a system, it is important to carefully consider the various reductions of that system and their relation-

ships. But no claim that once this is done a useful distinction is possible can be made. This analysis has left all of coding theory and image compression out of the picture; indeed this is not the focus of the effort here. The important issues to be understood here are those surrounding the various reductions. In any system of any large degree of complexity, summary statistics like entropy cannot be expected to provide a clear picture of the system, certainly they do not provide the information present in the full distribution, but for the purposes of engineering or for mapping out relevant biological mechanisms this is necessarily what the researcher must depend upon.

In order to quantify the information in a distribution the entropy is available, see chapter 2. The development here is uniquely centered on counting, with every effort made to show explicitly how counting states and entropy are related. The development of the entropy and conditional entropy is made in this fashion. From this immediately follows the mutual information. The notions of redundancy and relevancy are introduced and several competing definitions from the literature are noted. Several theorems about the entropy are concisely given, see section 2.2, and made use of throughout the work. In particular, the reduced entropy relationship and the reduced entropy per degree of freedom relationship are of immense value.

In order to quantify the goodness of a reduction of the distribution there is an interesting set of techniques arising from many body theory in statistical mechanics which are centered on the notion of partitioning. These includes methods for approximating the moment generating function, the cumulant expansion, see sections 3.1 and 3.3, and approximation of the distribution function itself, the Ursell expansion, see section 3.5, and the cluster expansion, see section 3.8. The usual moment functions, correlation functions, and the cumulants are made explicit. It is important to know when a set of random variables is telling you something new that some subset of the set could not.

For this the information correlation functions, see section 3.12, are introduced. In many ways these functions are the proper generalization of the entropy and mutual information, and generally the notion of the amount of information available between a set of random variables. These functions have been noted for possible use in corners of the statistical mechanics community, especially with the fluid theorists [26] and plasma physicists [39, 84]. In this dissertation they are employed as tools for making the behavior of systems explicit. It is hoped that by presenting the tools of many body physics clearly and concisely, and by providing interpretations for the information correlation functions, the various communities that could make use of these tools will do so - especially the neural networks and machine learning communities - for these techniques indeed present powerful methods for function approximation, function learning, representation theory, etc. That all of these methods are linked together by the central concept of partitioning is the key result of chapter 3. There the linked cluster theorem, cluster expansions, the Ursell development and the cumulant expansion are tied together by one theorem about partitioned structures, see section 3.8. Clearly this approach provides for a rigorous description of many variable interaction and representation, awaiting numerous applications as the techniques become better known outside the theoretical physics community.

Chapter 4 continues with an analysis of an extremely simple statistical mechanical system, the bit string system, where correlations arise not because of some interesting structure in some Hamiltonian, but because there is a simple constraint involved. This is analogous to the situation in quantum mechanics, where a type of symmetry, the symmetries that must be obeyed by fermions and bosons, induces correlations in systems of these particles. The bit string example has an interesting twist - for non-physical values of bits, fractional bits, interesting discontinuities in the entropies occur, see section 4.9. The bit string system is shown equivalent to a lattice hard sphere gas, where in the non-lattice case there is a hypothesized phase transition.

Information is carried in physical objects described by probability distributions. These distributions change in time. Chapter 5 on the classical equations of information flow discusses the time evolution of these distributions, and thus the information that they carry. For Hamiltonian systems, this is the Liouville equation, see section 5.1. The reduction of the full probability distribution also is generally changing in time. The BBGKY hierarchy, see section 5.2 describes the time evolution of the reduced distribution function in two parts, the evolution of the reduced system due to itself, and the evolution of the reduced system due to the rest of the system. The presentation here focuses on the structure of the separation of the reduced system and the rest of the system, applicable to general flows, with the actual Hamiltonian flow description included later. This makes the presentation much simpler to follow than the usual textbook study. It is perhaps somewhat surprising that taking a maximally uninformative rest-of-system makes the entropy change of the reduced system *zero*, see section 5.5.

Models play an important role in the presentation here. Previously mention was made of the study of the simple bit string system with constraint induced correlations. In chapter 6 the Ising coupled spin system is discussed. Expressions for the entropies of collections of neighboring spins are given in closed form in section 6.8 and shown graphically. The presentation shows clearly how the various reduced entropies, moments, and information correlation functions indicate areas of the phase space of interest, in particular the antiferromagnetic coupling case shows more structure than the ferromagnetic coupling case. Entropies for collections of non-neighboring spins are also given in closed form, see section 6.17. It is interesting to note that the reduced entropies per spin are *ordered*, see sections 6.15 and 6.17. Also shown (section 6.18) is that the information correlation functions are given by the mutual information between the first and last spins in the set considered. This indicates that all of the information available between a set of bounded spins is actually available from

the endpoint spins alone, although this result is tied to the interpretation of the information correlation functions, see sections 3.13 and 3.14. Several other topological reductions of the information correlation functions are mentioned, leading to the observation that the information correlation functions may be used to explore many aspects of the redundancy structure of a set of random variables.

Quantum systems are of interest because the world is fundamentally quantum in nature, and because this is the technological frontier for the foreseeable future in computing. Quantum systems are described by density matrices. In chapter 7 the density matrix is described in terms of a state density function. The evolution of this density function is discussed, see section 7.2, giving rise to the quantum Liouville equation (coordinate dependent description). Analogous to the classical case, there is a quantum BBGKY hierarchy, see section 7.3, which breaks the evolution into two parts, the reduced system and the rest of the system. The Heisenberg and pure-state time development equations are discussed here too.

The quantum Heisenberg model of coupled spin-1/2 particles is studied in chapter 8. Here also is presented a brief outline of the classification of spin systems. Entropy and all of the other system functions mentioned so far are presented for the quantum Heisenberg model. The interesting behavior occurs in the antiferromagnetic coupling case, as it did in the Ising system, and is indicated by the structure of the high order information correlations, entropies and other system functions. Again, we find that it is important to consider the high order nature of these systems in order to understand them. Phase transitions between the dominant states are observed at many values of the external field applied to the system, depending on the number of spins in the system. The relationship between the intrinsic entropy of the system (the entropy of the density matrix) and the measurement entropy of the system is made explicit.

One key result is that the measurement entropy is greater than or equal to the intrinsic entropy, see section 8.8. In potential applications, an important quantity to consider is the mutual information between a measurement, and an unmeasured but perhaps more desirable operator, see section 8.9. Because the operators for the energy eigenstates and the measurement eigenstates need not commute, the mutual information here is a time ordered (filter ordered) mutual information. Similarly, in more complex scenarios, the information correlation functions are generally time ordered information correlation functions. The mutual information spoken of here is the mutual information between the measured and unmeasured observables.

Information comes via measurements, and by nature only a finite number of these are available. Chapter 9 is devoted to the extraction of information from finite data, and it is here that we come full circle and complete the picture. Information is carried by a physical system, the system is measured by the observer, and the relevant aspects of the system are then inferred. This is the basis of all pattern recognition, learning, and prediction. It is important to not only infer the value of a quantity, but to also infer the uncertainty that is left after some finite number of data are seen, to know when to stop trying to obtain more data, and because a single value representing a guess at the underlying system is almost completely irrelevant unless it is also known how good that guess is. Here the tools for making these guesses and presenting their uncertainties are developed completely. The chapter starts off with a discussion about learning the values of parameters determining a system. It is shown how, sometimes, more data can lead to greater uncertainty about the parameters, but that the uncertainty about the parameters decreases on the average. The exact amount of decrease is quantified, see section 9.2. Inferring the values of the moments of the entropy is the next problem solved. The expressions give closed forms for the inference of the value of the entropy, its uncertainty, and all of the other moments of the entropy. The presentation is quick, via

the presentation of several theorems which demonstrate how to do the integrations needing to be done, see section 9.3.8. The results for the entropy appear in section 9.3.9. The results for the estimators for moments, correlations and cumulants appear in section 9.3.11. After the entropy and moment inference tools are developed, the tools needed to express the inferences of the moments of the mutual information, chi-squared, and covariance are presented. Again this is done quickly by way of many theorems, see section 9.3.13, with results in section 9.3.14. The appendices for this chapter present the mathematics needed to make the development of this section both rigorous and simple, and should serve as a useful reference to many in similar work to follow.

Derivatives of the entropy in statistical mechanical systems at thermodynamic equilibrium are the subject of chapter 10. It is well known that these may be related to dynamical power spectra and the equations of classical thermodynamics, see section 10.2. There is a similar relationship between the derivatives of the reduced entropy and correlations of the energy and the reduced distribution function, see section 10.4. That the reduced entropies are ordered was shown in chapter 2. That the derivatives of the reduced entropies are not is shown in section 10.5.

In summary, throughout this work a great deal of attention is paid to the nature of the high order entropies, information functions, and cumulant functions, and how they indicate interesting behavior in physical systems, the content of chapters 4, 6 and 8. Also, a great deal of attention is paid to how to infer the underlying physics of a system from actual measurements made on the system, the content of chapter 9. Many things of interest remain to be done yet, and many interesting ideas need to be explored. Especially intriguing is the possibility of using clusters of magnetic spins in the design of quantum computers, manipulating external fields to set the information of the internal state, how information is placed within these systems being analyzed

using tools like the time ordered mutual information of section 8.9. Application of the techniques presented here, especially those of chapters 3 and 9 to the fields of biological system modelling, machine learning, image processing, and pattern recognition are also to come. Not to mention the usefulness of these ideas in such nontraditional fields for physical scientists such as economic analysis. There is doubtless a great deal of pattern recognition work yet to be done in recognizing and understanding the mechanisms of life, and the analysis presented in this work forms an ideal basis for such investigations. This work will serve as a reference and creative motivator for those involved in such work, and apart from the other results presented inside, this is perhaps its most important role.

A brief note on the notation used within. Distribution functions may be denoted by P, p or ρ, with the tendency to be that P or p are used for discrete distributions while ρ is used for continuous distributions. Logarithms in the sections where graphical results are presented are always base e. The coupling strengths in the chapters on the spin systems are referred to as *ferromagnetic* when they favor alignment of the spins, or as *antiferromagnetic* when they favor anti-alignment, in keeping with the usage in [24, 25, 44]. The parameter $\beta = (kT)^{-1}$ always appears as b in the labels of the β axes of graphs.

Chapter 2

Entropy

2.1 Entropy

The entropy of a thing is the asymptotic average of the logarithm of the number of ways that the thing occurs. Thus, if there are two independent event generators A and B each generating events a and b with uniform probabilities for each generator, and there are n_A events possible for the first generator, n_B for the second, then there are $n_A n_B$ equally probable possibilities for both generators taken together. Consider N events from $A \times B$, i.e. N pairs (a, b). There are n_A^N ways that events from A can occur, and n_B^N ways events from B can occur. There are $(n_A n_B)^N$ ways that events from $A \times B$ can occur. Thus, the entropy of the joint process is

$$S(A,B) = \frac{1}{N} log((n_A n_B)^N) = log(n_A n_B) = S(A) + S(B) \tag{2.1}$$

which is an example of a very important and useful property of entropy - the entropy of independent processes is an additive quantity, whereas the number of ways is a multiplicative property. As a further example, consider the probabilities $p_1, \ldots, p_k$ where $\Sigma_{i=1}^k p_i = 1$ and $p_i \geq 0$ for each i. Sample this distribution N times and find that there are n_i events of type i. The number of ways to see distinct vectors $\boldsymbol{n} = (n_i)$ is

$$NumWays[\boldsymbol{n}] = \binom{N}{\boldsymbol{n}} \tag{2.2}$$

The logarithm of this is easily computed, and the asymptotics are simply found from Stirling's approximation (equation 6.1.37 of [3]) so that

$$log(NumWays[\boldsymbol{n}]) \sim -N\Sigma_{i=1}^{k} p_i log(p_i) \tag{2.3}$$

Clearly there were N events in this, so the average of it is the entropy that was defined above

$$S(\boldsymbol{p}) := -\Sigma_{i=1}^{k} p_i log(p_i) \tag{2.4}$$

Now consider the joint process of two random variables, denoted A and B before, except now the two processes will not necessarily be independent. Generate N events from the joint distribution $p_{ij} = P(a_i, b_j)$, with $i \in \{1, \ldots, n_A\}$ and similarly $j \in \{1, \ldots, n_B\}$. Clearly, the joint entropy of this process is given by $S(A, B) = -\Sigma_{ij} p_{ij} log(p_{ij})$. How does this relate to $S(A)$ and $S(B)$? Clearly, the number of ways that two things may occur is no more than the product of the numbers of ways each occurs individually, and this is certainly reflected in this case by the fact that

$$S(A, B) \leq S(A) + S(B) \tag{2.5}$$

with equality holding iff A is independent of B. The proof is trivial, simply note that $log(x) \leq x - 1$, consider $x = (p_{i\cdot} p_{\cdot j})/p_{ij}$ and show that it's average logarithm over p_{ij}, $\Sigma_{ij} p_{ij} log((p_{i\cdot} p_{\cdot j})/p_{ij})$, is non-positive (where $p_{i\cdot} = \Sigma_j p_{ij}$, etc.). See also the generalization of this, the reduced entropy relationship theorem below.

Now, let's consider what happens when we are given one of the two outcomes of each event (a, b) from $A \times B$ consistently, and we want to deduce the other. To be specific, let the events from $A \times B$ be generated N times, and let the value of B be seen each time. What is the asymptotic average log number of ways to see A given that B is seen each time? Well, let $n_b(b_j)$ be the number of times that b_j occurs, and for these occurrences of b_j, let

$n_{a|b}(a_i \mid b_j)$ count the occurrences of a_i. Define the vectors $\boldsymbol{n}_b = (n(b_j))$ and $\boldsymbol{n}_{a|b}(b_j) = (n_{a|b}(a_i \mid b_j))$; then for a fixed value b_j of B there are

$$NumWays[\boldsymbol{n}_{a|b}(b_j)] = \binom{n_b(b_j)}{\boldsymbol{n}_{a|b}(b_j)} \tag{2.6}$$

ways that the A values could be distributed for this b_j. Taking the product of these numbers of ways gives us the number of ways that the A values could be distributed given the B values. This is

$$NumWays[\boldsymbol{n}_{a|b}, \boldsymbol{n}_b] = \Pi_{j=1}^{n_B} \binom{n_b(b_j)}{\boldsymbol{n}_{a|b}(b_j)} \tag{2.7}$$

Taking the logarithm and doing the asymptotics gives gives us

$$\begin{aligned} log(NumWays[\boldsymbol{n}_{a|b}, \boldsymbol{n}_b]) &= \Sigma_{j=1}^{n_B} \, [\, Np_{\cdot j} log(Np_{\cdot j}) - \\ &\quad \Sigma_{i=1}^{n_A} n_b(b_j) p_{i|j} log(n_b(b_j) p_{i|j}) \,] \\ &= -N \Sigma_{ij} p_{ij} log(p_{i|j}) \end{aligned} \tag{2.8}$$

where we have $p_{i|j} := p_{ij}/p_{\cdot j}$. Averaging by dividing by N gives us the entropy of A given B, or $S(A \mid B) := -\Sigma_{ij} p_{ij} log(p_{i|j})$. Note that $S(A, B) = S(A \mid B) + S(B)$. Similarly, $S(A, B) = S(B \mid A) + S(A)$. Note that we could have found the log number of ways that A could occur given b_j as $-n_b(b_j) \Sigma_i p_{i|j} log(p_{i|j})$, and then noted that asymptotically $n_b(b_j) = Np_{\cdot j}$ to average this and find the result above.

After working through these examples, the interpretation of entropy as an uncertainty - an additive quantity representing the state of ignorance of the outcome - is straightforward. For example, if A is determined by B, then there is no uncertainty in A given B, immediately $S(A \mid B) = 0$; further there is no more uncertainty in the joint distribution then there is in the distribution of B, i.e. $S(A, B) = S(B)$. Finally, note that the quantity $S(A) - S(A \mid B)$ gives the uncertainty change between not knowing B and knowing B, and is called

the mutual information. It is symmetric in its arguments, and can be written as

$$M(A,B) = S(A) + S(B) - S(A,B) \tag{2.9}$$
$$= S(A,B) - S(A \mid B) - S(B \mid A) \tag{2.10}$$

The mutual information is clearly a quantity that for two random variables can be labeled the information about one variable that is in the other, and vice-versa. It is the information that each random variable shares about the other. In section 3.12 higher order information functions of this nature are defined, the information correlation functions, and these can be interpreted as the information *between* a set of random variables.

There are several other information functions that are of interest. We may define the *redundancy* of one random variable in another as the mutual information of the two. We might also define the normalized redundancy of two random variables as the mutual information divided by the joint information (entropy), $M(A,B)/S(A,B)$. This is a quantity that has value zero only for independent processes, and has value one when one process completely determines the other. For two or more random variables the redundancy has been defined as the sum of the single entropies minus the joint entropy, $S(A) + S(B) + \ldots - S(A,B,\ldots)$ [66, 93]. This redundancy is distinctly different from that of the information correlation functions to be defined in section 3.12 When there are only two processes this is the mutual information. A measure of correlation has been defined as $1 - S(B|A)/S(A)$ [16]. Note that this is asymmetric in the processes. It is 0 when the entropy of B given A is equal to the entropy of A, which for identically distributed variables occurs only when they are independent. A symmetric function with similar properties is $2(1 - S(A,B)/(S(A) + S(B))) = 2M(A,B)/(S(A) + S(B))$.

2.2 Convexity properties of entropy

Before continuing it is necessary to quantify the convexity properties of the entropy. The following theorem [16] is used several times in this work, and demonstrates the convexity properties of the entropy succinctly

Theorem. Log sum inequality. *Given non-negative* a_i, b_i, $i = 1, \ldots, m$

$$\sum_{i=1}^{m} a_i log\left(\frac{a_i}{b_i}\right) \geq \left(\sum_{i=1}^{m} a_i\right) log\left(\frac{\sum_{i=1}^{m} a_i}{\sum_{i=1}^{m} b_i}\right) \tag{2.11}$$

with equality iff $a_i = b_i$, $i = 1, \ldots, m$.

Proof: By Jensen's inequality and the convexity of $f(x) = xlog(x)$

$$\sum_{i=1}^{m} \alpha_i f(x_i) \geq f(\sum_{i=1}^{m} \alpha_i x_i) \tag{2.12}$$

for $\alpha_i \geq 0$ and $\sum_{i=1}^{m} \alpha_i = 1$. Let

$$\alpha_i = \frac{b_i}{\sum_{i=1}^{m} b_j}, \qquad x_i = \frac{a_i}{b_i} \tag{2.13}$$

in equation 2.12, along with continuity (if any a_i, b_i are zero) to find the result. QED.

Reduced probability distribution functions are defined as integrations of the complete distribution over some subset of the variables of the complete distribution. For example in equation 2.8 use has already been made of the reduced distributions $p_{i\cdot}$ and $p_{\cdot j}$. The next theorem makes explicit the connection between the entropies of different orders based on these reduced distributions. Let $\rho_k(X_{i_1}, \ldots, X_{i_k})$ be the distribution function reduced over those variables not in the set indicated. As usual, define the reduced entropy as

$$S_k(i_1, \ldots, i_k) = -\int dx_{i_1} \ldots dx_{i_k} \rho_k(x_{i_1}, \ldots, x_{i_k}) log(\rho_k(x_{i_1}, \ldots, x_{i_k})) \tag{2.14}$$

The following theorem demonstrates the relationship between entropies based on different reductions of the distribution function.

Theorem. Reduced entropy relationship. *Given the full distribution* $\rho(A \cup B)$ *and the distributions reduced from it,* $\rho(A)$, $\rho(A \backslash B)$ *and* $\rho(B \backslash A)$, *where* A *and* B *are sets of random variables with elements in* $\{X_1, X_2, \ldots\}$ *then*

$$S(A) + S(B \backslash A) \geq S(A \cup B) \tag{2.15}$$

with equality when $\rho(A \cup B)/\rho(B \backslash A)$ *is independent of* $B \backslash A$.

Proof: From the continuous version of the log sum theorem we have

$$\int dx_{A \cup B} \left(\rho(A \cup B) log \left[\frac{\rho(A \cup B)}{\rho(B \backslash A)} \right] \right) \geq \int dx_A \left(\left(\int dx_{B \backslash A} \ \rho(A \cup B) \right) log \left[\frac{\int dx_{B \backslash A} \rho(A \cup B)}{\int dx_{B \backslash A} \rho(x_{B \backslash A})} \right] \right) \tag{2.16}$$

From this the theorem follows almost immediately. QED.

Note that in the discrete case a looser upper bound occurs when $\rho(B \backslash A)$ is replaced by $1/d$, where d is the product of the number of objects in each summation over x_i with $X_i \in B \backslash A$. This also gives rise to the interpretation of the entropy as a dimension - here d or $S(B \backslash A)$ is the number of degrees of freedom in the stochastic variable $B \backslash A$, which leads to interesting results regarding the dimension of chaotic time series, see [63].

Intuitively, the reduced distributions contain less information than the full distribution. This is correct, since $S(A \cup B) = S(A|B) + S(B)$, indicating that $S(B) \leq S(A \cup B)$, the reduced entropy being less than the full entropy. But there is another sense in which the reduced distribution contains less information: the entropy per degree of freedom is less for the reduced distributions, as is shown in the next theorem.

Theorem. Reduced entropy per degree of freedom relationship. *Given a full distribution over* n *objects* $X_1, \ldots, X_n$, $\rho_n(x_1, \ldots, x_n)$, *with the probabilities* $\rho(x_k \mid x_{k+1}, \ldots, x_{k+r})$ *independent of* k *(the conditional density is dependent*

only on the values of its arguments, not the position in the sequence of variables) for $1 \leq r \leq n$ *and* $1 \leq k \leq n - r$*, then*

$$\frac{S(1,\ldots,n-1)}{n-1} \geq \frac{S(1,\ldots,n)}{n} \tag{2.17}$$

Proof: Because of the shift invariance, we may indicate the dependence on r neighbor subscripts $X_{k+1},\ldots,X_{k+r}$ by the subscript r. Similarly, indicate conditioning of X_k on $X_{k+1},\ldots,X_{k+r}$ by the subscript $1 \mid r$. Expand $nS_{n-1} - (n-1)S_n$

$$\begin{aligned} nS_{n-1} - (n-1)S_n &= n \sum_{x_1,\ldots,x_n} \rho(x_1,\ldots,x_n) log\left(\frac{\rho(x_1,\ldots,x_n)}{\rho(x_1,\ldots,x_{n-1})}\right) + S_n \\ &= -nS_{1|n-1} + S_n \\ &= -nS_{1|n-1} + (S_{1|n-1} + S_{1|n-2} + \ldots + S_{1|1} + S_1) \end{aligned} \tag{2.18}$$

Note that $S_{1|n-1}$ is conditioned on a superset of the variables that the other conditioned entropies above are conditioned on, i.e. that $S_{1|k} \geq S_{1|n-1}$ for $k = 0,\ldots,n-1$ (note $S_{1|0} = S_1$). Thus rearrange equation 2.18 as

$$-nS_{1|n-1} + (S_{1|n-1} + S_{1|n-2} + \ldots + S_{1|1} + S_1) = \sum_{i=0}^{n-1} S_{1|i} - S_{1|n-1} \geq 0 \tag{2.19}$$

where the difference is clearly positive because each term in the sum is positive. QED.

The statement used in the proof of the theorem above that conditioning on more things decreases the entropy is a direct consequence of the log sum inequality. In chapter 9 on estimating unknown parameters from data, we prove that the uncertainty about the unknown parameters does not generally decrease when the inference is based upon (the distribution of the unknown parameters is

conditioned upon) more data. This result may seem to contradict the theorem above, however, the data is a specific instance of possible data and is thus not averaged over, while the conditional entropy averages over the conditioning variables, too.

An interesting application of the reduced entropy per degree of freedom is [98] where an overall complexity measure for a time series is developed that is the summation of these entropies.

Chapter 3

Information correlation, cumulants, clusters, and partitions

3.1 Cumulants

In this chapter some of the mathematical methods and background of many variable theory are introduced. The cumulant expansion is defined and the relationship between cumulants and moments is described. Then the information correlation hierarchy is defined and its relationship to the cumulant expansion is made explicit. Finally, the linked cluster theorem, the Ursell development, and the cumulant expansion are shown to be direct consequences of a simple theorem about partitions. These techniques form the basis for investigations of the interdependency structure of a set of many variables.

3.2 Cumulants in one dimension

The derivatives of the moment generating function $M(z)$ of X give the moments of X, i.e. $M(z) = \langle e^{zX} \rangle$, where $\frac{d^k M}{dz^k}|_{z=0} = \langle X^k \rangle$.

The cumulant expansion of the function X is given by solving for the coefficients u_i in $M(z) = Exp(\Sigma_{i=1}^{\infty} \frac{z^i}{i!} u_i)$. The solution is given by equating coefficients of z and the first few cumulants are

$$u_1 = \langle X \rangle \tag{3.1}$$

$$u_2 = \langle X^2 \rangle - \langle X \rangle^2 \tag{3.2}$$

$$u_3 = \langle X^3\rangle - 3\langle X\rangle\langle X^2\rangle + 2\langle X\rangle^3 \tag{3.3}$$

3.3 Cumulants in many dimensions

In the case of many random variables we have a similar generating function giving the various moments. $M(z_1,\ldots,z_n) = \langle e^{z_1 X_1 + \ldots + z_n X_n}\rangle$, Here we have

$$\frac{d^{k_1}}{dz_1{}^{k_1}}|_{z_1=0}\ldots\frac{d^{k_n}}{dz_n{}^{k_n}}|_{z_n=0}\, M = \langle X_1{}^{k_1}\ldots X_n{}^{k_n}\rangle \tag{3.4}$$

The cumulant expansion in this case is then given by equating coefficients of the z_k in

$$M(z_1,\ldots,z_n) = Exp\left(\sum_{i=1}^{\infty}\sum_{\{\boldsymbol{m}\}} u_i(\boldsymbol{x})\frac{z_1^{m_1}\ldots z_n^{m_n}}{m_1!\ldots m_n!}\right) \tag{3.5}$$

where $\{\boldsymbol{m}\}$ indicates $\sum_{k=1}^{n} m_k = i$, and $\boldsymbol{x}$ in $u_i(\boldsymbol{x})$ consists of m_1 X_1 indices, etc. The first four cumulants are [27]

$$\begin{aligned}
u_1(i) &= \langle X_i\rangle & (3.6)\\
u_2(i,j) &= \langle X_iX_j\rangle - \langle X_i\rangle\langle X_j\rangle & (3.7)\\
u_3(i,j,k) &= \langle X_iX_jX_k\rangle \\
&\quad -\langle X_i\rangle\langle X_jX_k\rangle - \langle X_j\rangle\langle X_kX_i\rangle - \langle X_k\rangle\langle X_iX_j\rangle \\
&\quad +2\langle X_i\rangle\langle X_j\rangle\langle X_k\rangle & (3.8)\\
u_4(i,j,k,l) &= \langle X_iX_jX_kX_l\rangle \\
&\quad -\langle X_i\rangle\langle X_jX_kX_l\rangle - \langle X_j\rangle\langle X_iX_kX_l\rangle \\
&\quad -\langle X_k\rangle\langle X_iX_jX_l\rangle - \langle X_l\rangle\langle X_iX_jX_k\rangle \\
&\quad -\langle X_iX_j\rangle\langle X_kX_l\rangle - \langle X_iX_k\rangle\langle X_jX_l\rangle - \langle X_iX_l\rangle\langle X_jX_k\rangle \\
&\quad +2\langle X_i\rangle\langle X_j\rangle\langle X_kX_l\rangle + 2\langle X_i\rangle\langle X_k\rangle\langle X_jX_l\rangle \\
&\quad +2\langle X_i\rangle\langle X_l\rangle\langle X_jX_k\rangle + 2\langle X_j\rangle\langle X_k\rangle\langle X_iX_l\rangle \\
&\quad +2\langle X_j\rangle\langle X_l\rangle\langle X_iX_k\rangle + 2\langle X_k\rangle\langle X_l\rangle\langle X_iX_j\rangle \\
&\quad -6\langle X_i\rangle\langle X_j\rangle\langle X_k\rangle\langle X_l\rangle & (3.9)
\end{aligned}$$

In general, all distinct terms where each variable appears exactly once are included, and the coefficient of any term is $(-1)^{n-1}(n-1)!$ where n is the number of averages in the term.

3.4 Independence property

The important property to note is the following: $u_k(\ldots)$ is zero iff there are some arguments X_i, X_j that it depends on which are independent. This property follows directly from the observation that the logarithm of a product is the sum of the logarithms and that no $z_i z_j$ term will be present when there are some X_i, X_j which are independent.

3.5 Ursell functions and the density hierarchy

This section should be taken as an introduction to the Ursell development. A more rigorous treatment appears in section 3.8. Substituting for the $X_i = \delta(x_i - \overline{x_i})$ in equations 3.6–3.9 with $i \neq j,k,\ j \neq k$, we have a hierarchy of functions U_i written in terms of the reduced probability distribution functions. For example, $U_3(x_1,x_2,x_3) = \rho_3(x_1,x_2,x_3) - \rho_1(x_1)\rho_2(x_2,x_3) - \rho_1(x_2)\rho_2(x_3,x_1) - \rho_1(x_3)\rho_2(x_1,x_2) + 2\rho_1(x_1)\rho_1(x_2)\rho_1(x_3)$. The functions U_i are known as the Ursell functions. The relationship between the U_i and the ρ_i is known as the Ursell development. Note that the subscript on the ρ_i indicates the order of the reduced density function, while the argument(s) indicate the variable(s) not averaged over. Note also, that the kernel ρ may in actuality be any function W having been normalized so that summing over its arguments yields one, so in what follows W has been substituted for ρ. Writing the reduced W functions in terms of the Ursell functions turns out to have a simple form, (the sum over all partitions).

$$W_1(i) \;=\; U_1(i) \tag{3.10}$$

$$W_2(i,j) = U_2(i,j) + U_1(i)U_1(j) \tag{3.11}$$

$$\begin{aligned} W_3(i,j,k) &= U_3(i,j,k) + U_1(i)U_2(j,k) + U_1(j)U_2(k,i) \\ &\quad + U_1(k)U_2(i,j) + U_1(i)U_1(j)U_1(k) \end{aligned} \tag{3.12}$$

Here each argument of the W functions is indicated by the argument's subscript.

3.6 Summation property

Note that for the U's we have

$$\Sigma_i U_1(i) = \Sigma_i W_1(i) = 1 \tag{3.13}$$

$$\Sigma_i U_2(i,j) = \Sigma_i \left[W_2(i,j) - U_1(i)U_1(j) \right] = 0 \tag{3.14}$$

$$\Sigma_i U_3(i,j,k) = 0 \tag{3.15}$$

$$\ldots$$

with, in general, the sum over any index of any higher order U function being zero, and as indicated, the proof is by induction. The proof depends on the unit normalization and marginalization properties of the W functions.

3.7 The Ursell functions, partitions, and graphs

Let V_1^k be the set of all containers of k variables from a set of variables X. Let V^N be the set of all possible partitions of N variables from X. Let any element of V^N be called a graph. Each graph in V^N is then a set of containers $v_m \in V_1^m$ with the sum of the m's being N. (In the graphical picture, the v_m are called clusters.) To each v_m associate a function of the variables it contains, the Ursell function $U_m(v_m)$. Now, consider the graphs $g_N \in V^N$. For each graph g_N form the product of the functions that it represents, $\Pi_{v_m \in g_N} U_m(v_m)$, and call this the weight of the graph g_N. Now, sum over the graphs, and define this sum as

$W(V^N) = \Sigma_{g_N \in V^N} \Pi_{v_m \in g_N} U_m(v_m)$. Due to the fact of the partitioning used to create W, the Ursell development of section 3.5 immediately holds.

Examples of the decomposition above occur in the Mayer f-function expansion where each edge connecting two points of the graph represents a factor in a product representing the function [69], and in quantum field theory calculations using Feynman diagrams [55]. The key thing to note is that the Mayer f-function expansion is fundamentally a theorem about labeled partitions, rather than graphs. Similarly, the cumulant expansion is fundamentally a theorem about labeled partitions. In the next section we make explicit this fundamental idea, and rigorously show the connections between the counting of labeled partitions, the partition theorem, and the linked cluster theorem.

3.8 Partition theorem, cluster theorems

The fundamental idea behind the cumulant expansion, the Ursell development, Mayer's cluster theorem, and the linked cluster theorem from quantum field theory can be represented most succinctly by considering the idea of *structuring a set of objects*. To define, each structuring of an n-set of objects is a way that a set of objects may be named or organized. For instance, we have the uniform structure, which is the set itself. We may also have a structure like permutation, which is any ordering of the objects. There is also the structure of partitioning a set of objects. Consider the operation of partitioning the n-set into subsets. For example, $\{\{1,3\},\{2\}\}$ is one partition of $\{1,2,3\}$. How many ways can a structure of a certain type be imposed? Let $M(x) = \Sigma_{i=0}^{\infty} \frac{x^i}{i!} m_i$ be the exponential generating function for a structure of type M. Then there are m_n ways that an n-set can be structured according to M. For example, there is one way that the uniform structure can be imposed on any non-empty n-set. Therefore, the generating function for the uniform structure is $e^x - 1$. Now consider the notion of structuring a set of structures. In how many ways

can this be done? In general, if there are several structures of type M_2, and the structures are themselves organized into a structure of type M_1 then the generating function for M_1 structured M_2 structures is $M_1(M_2(x))$.

Theorem: Structure Composition or Structure Substitution. *Let* $M_1(x) = \sum_{i=1}^{\infty} m_i^{(1)} \frac{x^i}{i!}$ *be the generating function for a structure of type* M_1*, and similarly for* $M_2(x) = \sum_{i=1}^{\infty} m_i^{(2)} \frac{x^i}{i!}$*. The number of ways that an* n*-set can be first partitioned into subsets, then each subset structured by* M_2*, and then these structured subsets taken as objects and structured by* M_1 *is given by the coefficient of* $x^n/n!$ *in* $M_1(M_2(x))$*.*

Proof: Take any partition of the n set. There are b_1 subsets of size 1, b_2 subsets of size 2, ..., b_k subsets of size k. Thus $1b_1 + 2b_2 + \ldots + kb_k = n$. We may permute subsets of the same cardinality, and within any subset we may permute the objects. Thus there are

$$\frac{n!}{1!^{b_1} \ldots k!^{b_k} b_1! \ldots b_k!} \tag{3.16}$$

such partitions. For each of these partitions, each subset of the partition is structured by M_2, and the number of ways that the partitioning and M_2 structuring may be done is then

$$(\frac{m_1^{(2)}}{1!})^{b_1} \ldots (\frac{m_k^{(2)}}{k!})^{b_k} \frac{n!}{b_1! \ldots b_k!} \tag{3.17}$$

Now, the number of ways these structured subsets may be structured by M_1 is given by $m_q^{(1)}$ times this, where $q = b_1 + \ldots + b_k$, and the contribution to the coefficient of x^q in the generating function for the compound structure is found by multiplying by $x^q/q!$. This contribution is found in $M_1(M_2(x))$ from the term $m_q^{(1)}(M_2(x))^q/q!$ as

$$\frac{m_q^{(1)}}{q!} \binom{q}{b_1, \ldots, b_k} (\frac{m_1^{(2)} x}{1!})^{b_1} \ldots (\frac{m_k^{(2)} x^k}{k!})^{b_k} \tag{3.18}$$

Conversely, every such term represents some partition of the n-set into q subsets with structuring of the set of subset objects according to M_1 (the factor $m_q^{(1)}$) and with structuring of each subset according to M_2. QED.

Now we are ready to prove the various forms of the partition theorem which have been found useful in many-body physics.

First, there are two observations that need to be made that make the preceding theorem more general: 1. It is possible to generalize the theorem above to the notion of *weighted* objects. The weight of a single object is assigned then by some function (M_2 above) of the single object, and need not be an integer as might have been assumed in the proof of the structure composition theorem. As before let the weight of a compound object be the product of the weights of the objects making up the compound object. 2. It is possible to label any set in any manner desired, and to each labeled set associate a distinct *cluster* (some abstract mapping for now, later a cluster is defined as a connected graph) dependent only upon the labels that the set has associated with it (independent of any ordering that might have been utilized). The labeling set may be of any nonempty size, even though only as many labels appear (perhaps not distinct) as there are elements in the set. The object that the set of a given size then represents is then the set of all clusters of that size, and its weight is then the sum of the weights of the clusters. These observations will be used in deriving the cumulant and Ursell expansions.

Define a cluster as any labeled subset. Define a graph as any set of clusters. Let the weight of a set of clusters be the sum of the weights of the clusters in the set. Let the weight of any graph be the product of the weights of the clusters in the graph. Let the weight of a set of graphs be the sum of the weights of the graphs. Let the number of vertices of a cluster be the number of objects in the cluster. Let the number of vertices of the graph be the sum of the numbers of vertices of the clusters in the graph. The structure composition

theorem leads directly to the *Linked Cluster Theorem.*

Theorem. Linked Cluster Theorem. *Let V^k represent the set of all graphs on k vertices. Let V_1^k be the set of all clusters on k vertices. Let the weight of any set of graphs be the sum of the weights of the graphs in the set, and let the weight of any graph be the product of the weights of the clusters in the graph. Represent the weight of a graph g as $\langle g \rangle$. Let the weight of the empty graph be 0. Then*

$$\langle e^{Vx} \rangle = e^{\langle e^{V_1 x} \rangle} - 1 \tag{3.19}$$

Proof: The coefficient of $x^N/N!$ on the left hand side may be taken as the sum of the weights of all graphs on N vertices, so the left hand side is the generating function for the sum of the weights of all graphs on N vertices. The generating function of the uniform structure is $e^x - 1$. By the structure composition theorem the generating function for the compound structure consisting of the uniform structure of M_2 structures is $e^{M_2(x)} - 1$. Let m_k^2 be the weight associated with a subset of size k. Make the identification indicated above the theorem of labeled subset and cluster. Then m_k^2 is the weight of the set of all clusters of size k. The function $\langle e^{V_1 x} \rangle$ is then the generating function $M_2(x)$ just defined. Make the identification of uniform structure of clusters and graph to find that the right side is immediately another expression for the generating function given by the left side. QED.

We may rewrite the linked cluster theorem in symbols more easily read directly. Let $G(N)$ be any graph on N vertices, $G_1(N)$ be any connected graph on N vertices and $W(G)$ be the weight of any graph G. We then have that the linked cluster theorem is

$$1 + \Sigma_{N=1}^{\infty} \frac{x^N}{N!} \Sigma_{G(N)} W(G(N)) = Exp[\Sigma_{m=1}^{\infty} \frac{x^m}{m!} \Sigma_{G_1(m)} W(G_1(m))] \tag{3.20}$$

The weights of clusters may be any desired functions, and the rule stating that the weight of a graph made up of clusters is the product of the weights of

the clusters immediately yields the Ursell development of 3.5 with sum of the weights of the graphs on vertices, the $W(G(N))$ here, being the W_k there. A proper choice for the weights of the clusters immediately leads to the cumulant expansion theorem we have already discussed, as is proven next.

Theorem: Cumulant expansion theorem. *See section 3.3*

Proof: Let the labels of the vertices in equation 3.20 range over a set $\{1,\ldots,k\}$, and set $x = 1$ in that equation. Let $G_1(m)$ be any labeling of an m-vertex cluster consisting of m_1 labels 1, ..., m_k k labels k. Define $u_m(\boldsymbol{x})$ by $u_m(\boldsymbol{x}) = (z_{x_1}^{m_1}\ldots z_{x_k}^{m_k})^{-1}W(G_1(m))$, where $\boldsymbol{x}$ is an argument consisting of m_1 X_1's, etc. Since an m vertex cluster may be labeled with m_1 1's, ..., m_k k's in $\binom{m}{m_1\ldots m_k}$ ways we find that the expression $\sum_{G_1(m)} W(G_1(m))$ of equation 3.20, is $\sum_{\{\boldsymbol{m}\}}\binom{m}{m_1\ldots m_k}u_i(\boldsymbol{x})z_{x_1}^{m_1}\ldots z_{x_k}^{m_k}$. Now, with the choice of the u given after equation 3.5, the right side of equation 3.20 becomes the right side of equation 3.5, and similarly for the left sides of these equations, thus showing that the cumulant expansion theorem follows as a special case of the linked cluster theorem, with the weights of the clusters given by special sums of products of averages of the variables that the labels of the cluster vertices represent. QED.

To summarize, in this section we have shown that the cumulant expansion, and the various cluster theorems are simply re-namings of a theorem about labeled partitions.

3.9 Factorization property

Referring back to the property that any cumulant involving two independent variables is zero, it follows that whenever any W is representable as a product of W's then the corresponding U function is zero. In other words, any U function depending on any two independent functions (variables in the case that the functions are simply random variables) is zero.

3.10 Clustering

The cumulant expansion and the linked cluster theorem make it clear that we may cluster in various ways. In the linked cluster theorem we have clustered by the number of variables involved (vertices of a graph). We might also consider clustering by the types of variables involved (support), or by the types of functions involved (functions of arguments within certain sets of variables, say). If we cluster by support then we have an expansion which is useful in defining hierarchical decompositions of functions, as seen in the next section.

3.11 Function hierarchies

Now, consider the case where we have clustered by support. The $U(v)$'s which appear all appear as a sum in an exponential when the representation $\langle Exp(\Sigma_i z_i f_i)\rangle = Exp(\Sigma_v C(v,z)U(v))$ is made. This yields the product representation $\Pi_v Exp(\phi(v,z))$ for $\langle Exp(\Sigma_i z_i f_i)\rangle$. By extension to power series other than that for Exp, we may write any average of any function $F(X_1,...,X_n)$ as $\langle F\rangle = Exp(\Sigma[\phi_n(1,\ldots,n)+\phi_{n-1}(1,\ldots,n-1)+\ldots+\phi_1(1)])$, where $\Sigma[]$ means take the sum of the functions arguments over arguments that are unique ordered subsets of $\{1,\ldots,n\}$. Thus, $F_1(1) = Exp(\phi(1)), F_2(1,2) = Exp(\phi(1,2)+\phi(1)+\phi(2))$, etc. When we have a hierarchy of functions F_i, or when we generate such a hierarchy from a single function's series expansion, we can solve this system for the ϕ_i.

3.12 Information correlation functions

In particular, consider the reduced density functions $\rho_1(i), \rho_2(i,j)$, etc. Make the hierarchical expansion

$$log(\rho_1(i)) = \phi_1(i) \tag{3.21}$$

$$log(\rho_2(i,j)) = \phi_2(i,j)+\phi_1(i)+\phi_1(j) \tag{3.22}$$

$$\begin{aligned} & \ldots \\ log(\rho_n(1,\ldots,n)) &= \phi_n(1,\ldots,n) + \ldots + \phi_2(1,2) \\ &\quad + \phi_2(1,3) + \ldots + \phi_1(1) + \ldots + \phi_1(n) \end{aligned} \tag{3.23}$$

and find

$$\phi_1(i) = log(\rho_1(i)) \tag{3.24}$$

$$\phi_2(i,j) = log(\frac{\rho_2(i,j)}{\rho_1(i)\rho_1(j)}) \tag{3.25}$$

$$\ldots$$

$$\phi_n(1,\ldots,n) = log(\frac{\rho_n \Pi[\rho_{n-2}]\ldots}{\Pi[\rho_{n-1}]\ldots}) \tag{3.26}$$

where $\Pi[]$ means to take the product over unique ordered subsets of $\{1,\ldots,n\}$. Multiplying both sides of this hierarchy by ρ_n and integrating over the variables gives a very suggestive information correlation hierarchy. The order-1 expressions in the hierarchy are the negative entropies of the individual random variables. The order-2 terms are the mutual informations. These are one measure of the degree of mutual dependence of two variables.The order-3 information correlation functions are less amenable to interpretation (see sections 3.13 and 3.14 for interpretations of the information correlation functions), but give an information correlation of three random variables, which is zero iff any two subsets are independent. These are one measure of the degree of mutual dependence of three variables.

$$C_1(i) := \int \rho_n log(\rho_1(i))\, d\mathbf{x}_n = -S_1(i) \tag{3.27}$$

$$C_2(i,j) := \int \rho_n log(\frac{\rho_2(i,j)}{\rho_1(i)\rho_1(j)})\, d\mathbf{x}_n = M(i,j) \tag{3.28}$$

$$C_3(i,j,k) := \int \rho_n log(\frac{\rho_3(i,j,k)\rho_1(i)\rho_2(j)\rho_3(k)}{\rho_2(i,j)\rho_2(j,k)\rho_2(k,i)})\, d\mathbf{x}_n \tag{3.29}$$

Note that the full negentropy (information) of the system of variables is given by the permutation sum of the information correlations as in

$$I_n = -S_n(1,\ldots,n) = \int \rho_n log(\rho_n)\, d\mathbf{x}_n = \Sigma_{i=1}^n \Sigma[C_i] \tag{3.30}$$

It is useful to point out that we now have two quantities which are zero iff the density function factors, C_k and u_k. In later chapters we will analyze the information correlation structure of several physical systems and draw conclusions about how the correlation structure indicates underlying physical phenomena. Information correlation functions have attracted the attention of liquid theorists [26] and plasma theorists [39, 84].

3.13 Interpretation of the information correlation functions.

Suppose that we are given information about the distribution of n variables in the form of distributions of some subsets of those variables. Let the subsets be denoted s_i, subsets of $S = \{1, ..., n\}$. For example let $n = 3$ and let s_1, s_2 be given with $s_1 = \{1,2\}$, and $s_2 = \{2,3\}$, indicating that we are given $\rho_2(1,2) = \int \rho_3(1,2,3)\, dx_3$ and similarly $\rho_2(2,3)$. Call this information K, for *known*. Now infer the distribution of the n variables given this information, and call this inference $\rho_{n|K} = \rho_n(1, \ldots, n \mid K)$. The inferred uncertainty of the full distribution is then given by $S_{n|K} = -\int \rho_{n|K} log(\rho_{n|K})\, d\mathbf{x}_n$. Because $\rho_{n|K}$ is consistent with K we have for each i the marginal of $\rho_{n|K}$ over the variables not in s_i is one of the given distributions, $\rho_{N(s_i)}(s_i)$. Doing some algebra, for two sets A, B,

$$\begin{aligned} S(A \cup B \mid K) &= S((A \setminus B) \cup (A \cap B) \cup (B \setminus A) \mid K) && (3.31) \\ &= S(A \setminus B \mid (A \cap B) \cup (B \setminus A), K) \\ &\quad + S(B \setminus A \mid (A \cap B), K) + S(A \cap B \mid K) && (3.32) \end{aligned}$$

Now, note that $S(A \setminus B \mid (A \cap B) \cup (B \setminus A), K) \leq S(A \setminus B \mid (A \cap B), K)$. Then it follows that $S(A \cup B) \leq S(A) + S(B) - S(A \cap B)$ (all conditioned on K). Note that equality holds iff $A \setminus B$ is independent of $B \setminus A$ when conditioned on $A \cap B$. This can be extended to more than two subsets by noting that the

algebra with the equality holding is the same as that for counting set elements upon intersection or union, i.e. additive. The approximate result is then

$$S(\cup_{i=1}^{n} A_i) \approx \Sigma_{s \subseteq \{1,\ldots,n\}} (-1)^{N(s)+1} S(\cap_{i \in s} A_i) \tag{3.33}$$

The connection to the information correlation functions C_i can be made by taking the $A_i = \{1, \ldots, n\} \setminus \{i\}$. Then the right side of equation 3.33 is just $C_n + S_n$, so that C_n equals how much the expression on the right side misestimates the entropy of all of the variables. (In general, the sign of C_n may be positive or negative. Take $S(\emptyset) = 0$ to make everything consistent.) This observation also allows us to generalize the notion of the information correlation function to include those functions defined similarly when any set of indicator sets is given, that is

$$C_A := -S(\cup_{i=1}^{n} A_i) + \Sigma_{s \subseteq \{1,\ldots,n\}} (-1)^{N(s)+1} S(\cap_{i \in s} A_i) \tag{3.34}$$

3.14 Another interpretation of the information correlations

Note, if we let go of the notions of intersections and unions of sets of random variables, if the notations of union and intersection are considered as *joint* and, somewhat metaphorically, as *between*, then we may *define* the information *between* two random variables A and B as $S(A \cap B) := S(A) + S(B) - S(A \cup B)$ and so on for more than two random variables. We then need to be able to imagine the "*between*" process of two random variables, and be able to take its union and intersection with other random variables in a fashion that satisfies the set union and intersection algebra. Then the inequalities above become equalities, and the information correlation functions on n random variables become $(-1)^n$ *times the information between the n random variables.* The algebra is straightforward. The previous section shows how this fairly seductive labeling may lead the worker astray. For the generalized information correlation

functions mentioned at the end of the previous section, the interpretation here is the information between the named reduced densities.

3.15 Information correlation functions - example

As a clearly contrived example of what the information correlation functions indicate, consider the three binary variables $x_i \in \{0,1\}$, $i = 1,2,3$. Let $P_{ij} = \frac{1}{2}\delta(x_i, x_j)$. Then $P_{123} = \frac{1}{2}\delta(x_1, x_2)\delta(x_2, x_3)$, i.e. it is zero unless all three have the same value.

Consider the full information correlation function, C_3 of equation 3.29. It is straightforward to show that

$$C_3 = \int P_{123} log(\frac{P_{123}P_1P_2P_3}{P_{12}P_{23}P_{31}}) dx_1 dx_2 dx_3 = -log(2) \tag{3.35}$$

Thus the information between the reduced distributions is $log(2)$ (the n of section 3.14 is odd), as might be expected from the fact that there is one bit determined by the full distribution not determined by the any of the reduced distributions. Again, we must warn the reader that care must be used in interpreting the information correlation functions.

3.16 Correlation functions

The standard approach to understanding correlations involves averaging products of zero-mean random variables. We have both a moment hierarchy and a correlation hierarchy. The moment hierarchy is simply

$$m_1(i) = \langle X_i \rangle \tag{3.36}$$

$$m_2(i,j) = \langle X_i X_j \rangle \tag{3.37}$$

$$m_3(i,j,k) = \langle X_i X_j X_k \rangle \tag{3.38}$$

$$\ldots$$

The correlation hierarchy is

$$c_2(i,j) = \langle (X_i - \langle X_i \rangle)(X_j - \langle X_j \rangle) \rangle \tag{3.39}$$

$$c_3(i,j,k) = \langle (X_i - \langle X_i \rangle)(X_j - \langle X_j \rangle)(X_k - \langle X_k \rangle) \rangle \tag{3.40}$$

$$\ldots$$

It is useful to note that both $c_2(i,j) = u_2(i,j)$ and $c_3(i,j,k) = u_3(i,j,k)$, and that this is not true for any higher order. It is also useful to note that for zero-moment random variables, $c_k = m_k$.

3.17 Information flow

With regard to information flow within a physical system, we may consider the time evolution of the cumulant hierarchy, u_k and the information correlation hierarchy, C_k, even the moment m_k and correlation c_k hierarchies, once we have evolution equations for the density functions. See chapter 5.

3.18 Estimating the information functions

In any real data acquisition scenario the amount of data available for analysis is necessarily limited. In the case of small samples, it is important to understand the significance of the information correlation functions, etc., and to know whether they say anything meaningful about information and correlations in the data. See chapter 9.

Chapter 4

Constraint induced correlations

4.1 What is done here

A simple bit-string system of weakly interacting bits is defined. The system consists of a microcanonical ensemble of bit strings of fixed length, with a fixed number of 1 and 0 bits for each member of the ensemble.

A fixed portion of the string is observed. The fixed overall length of the whole string and the fixed number of ones and zeros in the whole string induces weak correlations between the observed bit locations.

The relationships between the hypergeometric probability distribution, which describes the probability of having some observation of the bit string system with a fixed number of bits, the binomial probability distribution, and the Poisson probability distribution is discussed.

The first and second order entropies for this simple bit-string system are then computed. Then asymptotic limits are found for both large strings and small bit density. The exact first and second order entropies are compared. How each changes as the bit density of the system is changed is found.

A phase transition in the bit string system is found. This is not the usual temperature-dependent transition, as the system is temperature independent and we are considering a microcanonical ensemble. It is a phase transition that occurs as the bit density is changed. Specifically, the derivative of the second order entropy diverges logarithmically at two transition bit densities

in the range $(0, 1)$. This phase transition may be useful in understanding the hypothesized Kirkwood phase transition in a hard-sphere gas. The bit string system is identical to a lattice hard sphere gas on the simplex lattice. Higher order entropies exhibit phase transitions at increased numbers of densities. The third order entropy has a phase transition at four densities, etc.

The computation is generalized to entropies of arbitrary order, and the information correlation functions for the system are computed, their asymptotics are noted, etc. These functions thus characterize a system having one of the weakest forms of correlations possible - correlations induced by a single constraint on a single extensive variable.

4.2 The bit string system and observed states

Consider the bit string system microstates as bit strings of length N, with $0 <= B <= N$ 1-bits and $N - B$ 0-bits. Define the bit density of the system $q := B/N$. Observe $0 < m \leq N$ bits of these strings. These are the observed states. There are b 1-bits observed, $0 \leq b \leq min(m, B)$. In pictures:

B ones, $N - B$ zeros.	
observed	unobserved
m bits	$N - m$ bits
b ones, $m - b$ zeros	$B - b$ ones, $N - B - m + b$ zeros

4.3 Distribution of the observed states

We may easily compute the distribution of the observed states from the assumption that all microstates of the system with the same total number of bits B are equally probable.

The total number of states (where there are B one bits among the N bits) is

$$C(N, B) = \frac{N!}{B!(N - B)!} \tag{4.1}$$

Now suppose an m bit state having b one bits is observed. Fix the locations of the one bits in the observed state. How many microstates correspond to this observed state? To find this, just distribute the remaining $N-b$ one bits among the remaining $N-m$ locations. Then each distinct $N-m$ bit pattern is a microstate corresponding to the observed state. The number of observed states with b 1-bits in fixed locations of the m observed bits is thus

$$C(N-m, B-b) \tag{4.2}$$

and therefore the probability of *any* observed state x_b having b 1-bits and $m-b$ 0-bits is given by

$$P_m(x_b) = C(N-m, B-b)/C(N, B) \tag{4.3}$$

Note that the probability of seeing *some* observed state of length m having b one bits is $C(m,b)P_m(x_b)$, so that

$$\sum_{b=0}^{m} C(m,b)P_m(x_b) = 1 \tag{4.4}$$

Note that $C(m,b)P_m(x_b) =: P_m(b)$ is the hypergeometric distribution, the distribution giving the probability that b balls of m balls are labeled one, where the m balls are chosen randomly without replacement from an urn holding N balls of which B are labeled one [43]. This allows us another approach to deriving the distribution $P_m(x_b)$. First note that drawing m balls from an urn randomly without replacement is equivalent to randomly lining up all N balls, and taking the first m of them. Then note that for each arrangement of the undrawn balls there are $C(m,b)$ distinct orderings of the m drawn balls of which b are labeled one. Each of these orderings is equally probable by the assumption of equal probability microstates, so the probability of any one must be $P_m(x_b) = P_m(b)/C(m,b)$. Thus $P_m(x_b) = C(N-m, B-b)/C(N,B)$, as already derived in equation 4.3.

It is useful to express the hypergeometric distribution in the large N limit when $q := B/N$ is fixed. Doing this shows that $P_m(x_b)$ becomes the binomial probability distribution for a fixed sequence involving b ones and $m-b$ zeros, and that the hypergeometric distribution becomes the binomial distribution for bit sequences of length m having b one bits,

$$P_m(x_b) \rightarrow q^b(1-q)^{m-b} \tag{4.5}$$

$$P_m(b) \rightarrow \binom{m}{b} q^b(1-q)^{m-b} \tag{4.6}$$

Now, if we consider $q \ll 1$ and $b \ll m$ then the factor $(1-q)^{m-b}$ may be written $(1-q)^{m-b} \approx 1 - \binom{m-b}{1} q \approx 1 - mq$, and in turn this may be approximated as $exp(-mq)$. Similarly, $m!/(m-b)! \approx m^b$. Finally, in this limit we have with $\nu := mq$ being the expected number of ones in m locations at rate q that the hypergeometric distribution becomes the Poisson distribution

$$P_m(b) \rightarrow \frac{e^{-\nu}\nu^b}{b!} \tag{4.7}$$

Thus, the hypergeometric, binomial, and Poisson distributions are simply related - the binomial distribution is the hypergeometric distribution for an urn with a large number of balls and a fixed fraction of one balls, and the Poisson distribution occurs when additionally the fraction of one balls in the urn is small. In the bit string system these cases correspond to having a large unobserved portion of the bit string (binomial case) and to having a small number of ones relative to the number of locations of the bit string (Poisson case).

4.4 High order entropies of the observed states

The entropy at order m is given by

$$S_m(N,B) = -\sum_{b=0}^{m} \sum_{x_b} P_m(x_b) log(P_m(x_b)) \tag{4.8}$$

$$= -\sum_{b=0}^{m} \binom{m}{b} P_m(x_b) log(P_m(x_b)) \tag{4.9}$$

where in the second equation x_b may be understood as the probability of any observed m bit state having b one bits - all such states have the same probabilty. If the distributions of equations 4.6 or 4.7 are taken for the distribution of the bit strings it is immediately obvious that because effectively the bits are treated by these distributions as independent (a large unobserved portion is assumed) that the entropy will simply be the entropy for a single bit times the number of observed bits. Thus, letting q be the probability that a single bit is one we have for the large string case

$$S_m(q) = mS_1(q) = m(-qlog(q) - (1-q)log(1-q)) \tag{4.10}$$

which is trivially verified for the binomial form $P_m(x_b) = q^b(1-q)^{m-b}$. For the Poisson form, note that effectively we have assumed that $m \to \infty$ while mq is constant. The argument is slightly trickier and the limit of m must be taken carefully, but the result is again that the entropy per bit is $-qlog(q) - (1-q)log(1-q)$.

Clearly, if the entropy is to show the effects of the finite unobserved portion then we must compute with the non-limiting case form of the distribution, equation 4.3.

4.5 First and second order entropies

The first order entropy has already been found, and it is dependent on the ratio $B/N = q$ only.

$$S_1(N,B) = -qlog(q) - (1-q)log(1-q) \tag{4.11}$$

The second order entropy may be computed directly from the combinatorial coefficients. The result is that

$$\begin{aligned} S_2(N,B) &= 2S_1(N,B) + log(1-\frac{1}{N}) - \frac{C(N-2,B)}{C(N,B)}log(1-\frac{1}{(N-B)}) \\ &\quad -\frac{C(N-2,B-2)}{C(N,B)}log(1-\frac{1}{B}) \end{aligned} \tag{4.12}$$

The combinatorial identities useful in deriving this result are given in appendix 4.13.

4.6 Asymptotic S_2 is twice S_1

Let $N \to \infty$ with B/N fixed in equation 4.12. The last three terms there have the arguments of their logarithms going to 1, with $log(1) = 0$, and therefore these terms disappear, leaving

$$S_2(N,B) = 2S_1(N,B). \tag{4.13}$$

Clearly, the result is more general, from section 4.4 we saw in fact that asymptotically $S_m = mS_1$. Thus all correction terms $S_m - mS_1$ disappear in the thermodynamic limit.

4.7 $S_2 \leq 2S_1$

Since both $P_1(a)P_1(b)$ and $P_2(a,b)$ are both probability distributions, then by the information inequality (see chapter 2) we find

$$-\sum_{a,b\in\{0,1\}} P_2(a,b) log(\frac{P_2(a,b)}{P_1(a)P_1(b)}) \leq 0 \tag{4.14}$$

but the left hand side is also $S_2 - 2S_1$. Thus the result $S_2(N,B) \leq 2S_1(N,B)$. Also, then, the last three terms on the right side of equation 4.12 sum to a non-positive number, i.e.

$$log(1-\frac{1}{N}) - \frac{C(N-2,B)}{C(N,B)} log(1-\frac{1}{(N-B)}) - \frac{C(N-2,B-2)}{C(N,B)} log(1-\frac{1}{B}) \leq 0 \tag{4.15}$$

This result is also general, in fact by considering S_{m-1} and S_1 resursively we can easily show $S_m \leq mS_1$.

4.8 Correction terms in S_2

The last three terms of S_2 in equation 4.12 for $N \to \infty$ with B/N fixed are $O(1/N^2)$. Specifically, expanding these correction terms in a Taylor series for N with $B = Nq$ yields

$$S_2 - 2S_1 \simeq -\frac{1}{2}\frac{1}{N^2} - \frac{1 + 2q(1-q)}{6q(1-q)}\frac{1}{N^3} + h.o.t. \tag{4.16}$$

Note 1) this is independent of q only in the first term; 2) this is negative, which is required to satisfy the theorem of section 4.7; 3) in deriving this result it was assumed that $N \geq 2$, so that the second order entropy exists; however 4) in deriving this result a range for B was not assumed other than $0 \leq B \leq N$, contrary to the perhaps misleading form of the correction terms in equation 4.12; 5) the correction terms of equation 4.12 are not necesarily real for noninteger N and B or, therefore, away from the discrete values of q allowed. However it is clear from the form of equation 4.12 that the correction terms become imaginary for noninteger B only for $0 < B < 1$ $(0 < q < 1/N)$ and $0 < N - B < 1$ $(1 - 1/N < q < 1)$; 6) the third and higher order correction terms are not independent of q; 7) the third order correction term becomes singular at $q = 0, 1$ $(B = 0, N)$, indicating that for fixed N, as $q \to 0, 1$ $(B \to 0, N)$, the correction terms can become quite large. However, entropies are always positive, and as $q \to 0, 1$ all entropies converge to zero, so either the series for the correction terms must also converge to zero in sum in these limits, or the radius of convergence of the series does not include these limits, which is actually the case; 8) each coefficient of the series of equation 4.16 is a real function of q, yet the overall correction of equation 4.12 is not real for values of q in the ranges mentioned above. This is an effect of the fact that the radius of convergence of the Taylor series includes only those values of N for which the arguments of the logarithms in equation 4.12 are all positive.

Not unexpectedly after the discussion just given, the maximum absolute value of the correction of equation 4.12 occurs at the points $q = 1/N, 1 - 1/N$,

and has a slope which diverges logarithmically as these points are approached. *In this unphysical sense, the derivative of the second order entropy with respect to bit density shows a phase transition as the number of one bits per string of length N drops to one, or increases to $N-1$.* Again, what this indicates physically is not defined, as it is not possible to have fractional bits.

4.9 The Kirkwood phase transition

The discontinuous behavior demonstrated here may prove useful in investigating the hypothesized phase transition of Kirkwood [87] for a hard sphere gas. The Kirkwood transition hypothetically occurs for a density smaller than the close-packing density, while here the phase transition occurs for two densities of bits smaller than the maximum bit density of one. Because of the hard-sphere potential, the system analyzed there is independent of temperature, while the system described here is de-facto independent of temperature. For the hard-sphere gas it is known that, if a phase transition does occur at a density less than the close-packed density, the dimension must be greater than one. Here the dimension is effectively infinite because only the number of bit locations is considered, not the spatial organization of these locations. Finally, to complete the analogy to the hard sphere gas, let the bits interact with a hard sphere potential: the energy of any configuration of bits where no two bits are at the same location may be considered zero independent of the configuration of the occupied locations, while the energy of any configuration of bits some two of which occupy the same location may be taken as infinite, thus disallowing such configurations at any temperature. That is, the bits interact with a pairwise potential that is zero for bits at different locations, and infinite for bits at the same location. With this interaction potential the bit string system becomes temperature independent when considered as a thermodynamic system. Thus the infinite bit string system is a lattice hard-sphere gas on the infinite simplex.

Before taking this too seriously though, note that the hard sphere gas exists in a continuous space, and that all these discontinuities indicate for now is that fractional bits do not occur. At the very least, the analogy to and the exactly solvable nature of the bit system may afford an avenue toward insightful work on the hypothesized Kirkwood phase transition in the hard-sphere gas. For this reason, the plots are given for continuous ranges of q and N values, showing where interesting behavior may be found/utilized in some other investigation. Like the operation $sqrt$ on the integers and the subsequent introduction of the algebraic numbers (like $sqrt(2)$), perhaps there is a sensible interpretation of these fractional bit objects.

A plot of the correction at the maximal absolute correction ($q = 1/N$) as a function of string length is given in figure 4.1. A plot of the correction as a function of bit density for a string length of $N = 10$ is given in figures 4.2, and again magnified for the second order correction in the region of the lower phase transition in figure 4.3. A plot of the derivative of the correction (one sided) at the lower phase transition point and $N = 10$ is given in figure 4.4.

4.10 General S_m

The third and higher order entropies also exhibit unphysical phase transitions. The derivative of the third order entropy has transitions at $q = 1/n, 2/N, 1 - 1/N$, and $1 - 2/N$. Note again that it is difficult to interpret such transitions physically because the systems described away from the discrete values of $q = k/N, k = 0, 1, \ldots, N$ must have fractional numbers of bits to be realized. In general, the derivative of the mth order entropy exhibits a phase transition at $2(m-1)$ locations, up to $m = \lfloor N/2 \rfloor + 1$. Graphs of the corrections $S_m/m - S_1$, $m = 2, \ldots, 4$ as a function of N at $q = 1/N$ appear in figure 4.1, and as a function of bit density q for $N = 10$ in figure 4.2. Note that complex values occur at unphysical values of q only when too few or too many one bits are

sought in the observed region, or for N less than the number of bits needed to make the entropies exist.

The variation of the entropy differences at physical values of q are of interest. Graphs of these variations as a function of the order of the entropy and $N = 10$ appear next. The differences are normalized by the order of the entropy. In figure 4.5 the entropy differences between the points $q = k/N$ and $q = (k-1)/N$ are plotted for $k = 1, \ldots, 4$. These figures demonstrate that even for this simple system it is important to consider the higher order functions of the probability distributions of observed states in certain regions of the parameter space.

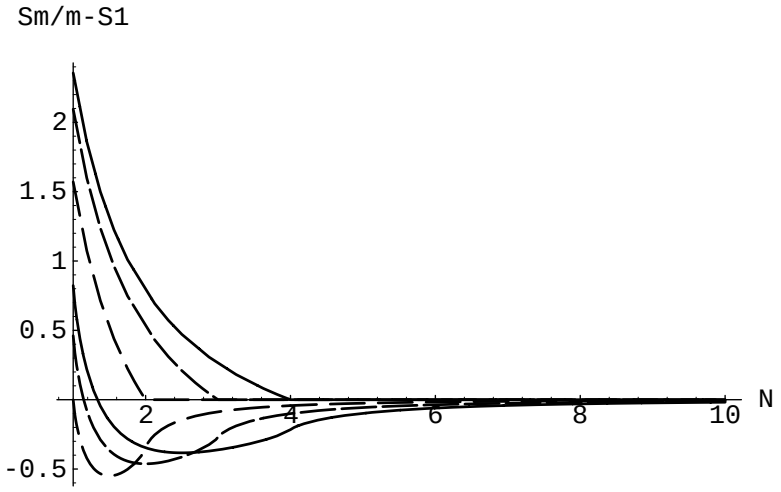

Figure 4.1: Correction terms $S_m/m - S_1$, $m = 2, 3, 4$ at $q = 1/N$ as a function of the length of the strings, N. Wider spaces in the dashing indicates lower m. Plot indicates that thermodynamic limit of corrections is zero. Low N behavior is complex due to having strings shorter than needed to define the entropy involved.

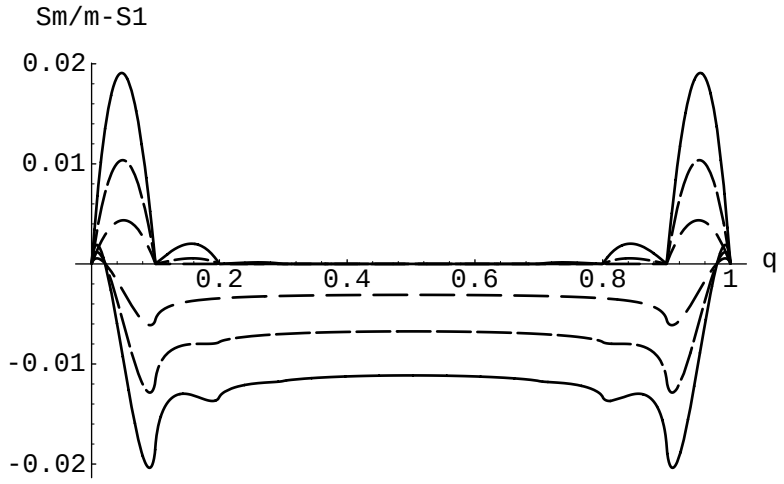

Figure 4.2: Correction terms $S_m/m - S_1$, $m = 2, 3, 4$ at $N = 10$ as a function of q. Wider spaces in the dashing indicates lower m. Off $q = k/N$ for integer k behavior is complex due to having nonphysical numbers of bits, fractional numbers of bits.

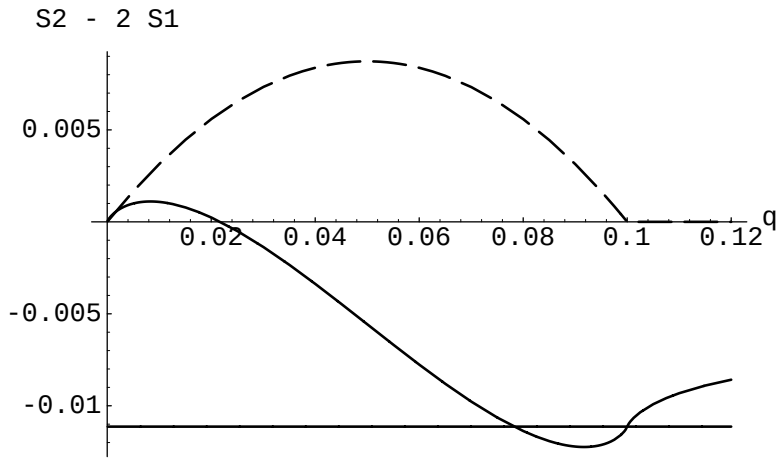

Figure 4.3: Second order correction term $S_2 - 2S_1$ at $N = 10$ as a function of q. Real part is solid. Imaginary part is dashed. Line crosses curve at divergent derivative.

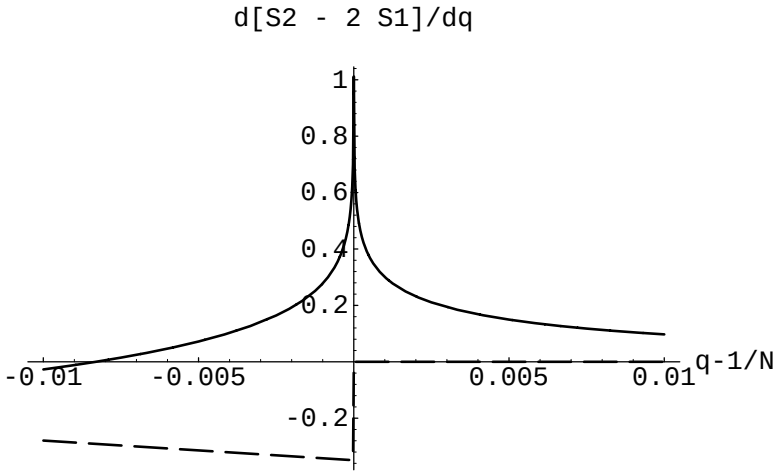

Figure 4.4: Derivative of second order correction term $S_2 - 2S_1$ at $N = 10$ as a function of q near $1/N$. Real part is solid. Imaginary part is dashed.

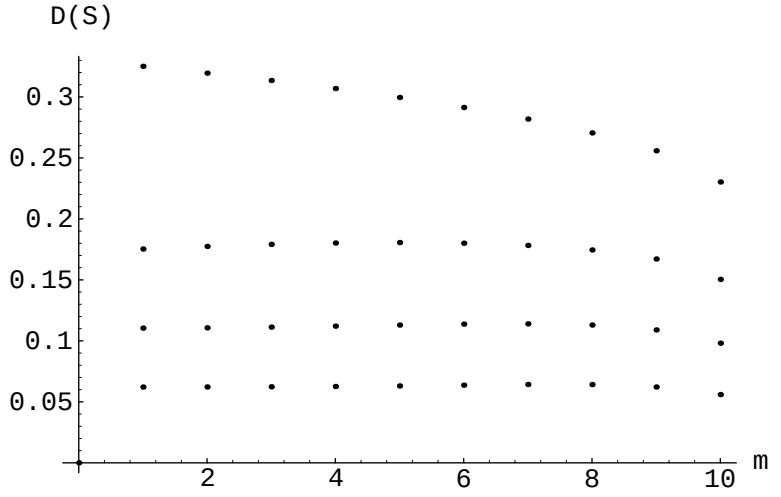

Figure 4.5: Entropy differences. Entropy order on x axis. The quantity plotted is $D(S) = (S_m(N, B) - S_m(N, B-1))/m$ for $B = 1, \ldots, 4$ and strings of length $N = 10$. Higher graph indicates lower B. These are effectively the derivative functions at small one bit counts per string.

4.11 Information correlation functions

Because of the complete symmetry between observed bit locations, the information correlation functions (see section 3.12) are given by

$$C_k(i_1, \ldots, i_k) = \sum_{j=1}^{k} (-1)^{k-j+1} \binom{k}{j} S_j \tag{4.17}$$

Plots of the information correlation functions to order four are shown as a function of the bit density for $N = 10$ in figures 4.6- 4.9.

4.12 Moments, Correlations, Cumulants

Because of the fact that the only term in the sum for the moments that contributes is the term where all bits are one, the moments, correlation functions, and cumulants are trivial to compute. The moments of the bit string observed state distribution are given by

$$m_1(i) := \langle b_i \rangle = \sum_{b=0}^{m} \sum_{x_b} b_i P_m(x_b) = P_1(1) = q \tag{4.18}$$

$$m_2(i,j) := \langle b_i b_j \rangle = P_2(2) = \frac{q(q - 1/N)}{1 - 1/N} \tag{4.19}$$

$$\ldots \tag{4.20}$$

$$m_k(i_1, \ldots, i_k) = \langle b_{i_1} \ldots b_{i_k} \rangle = P_k(k) \tag{4.21}$$

As before, the correlations and cumulants are identical to order three. Plots of the moments (figure 4.10), correlations (figures 4.11- 4.13) and cumulants to order four (figure 4.14) (the first three cumulants are the same as the correlations of the same order) as a function of bit density for $N = 10$ are shown.

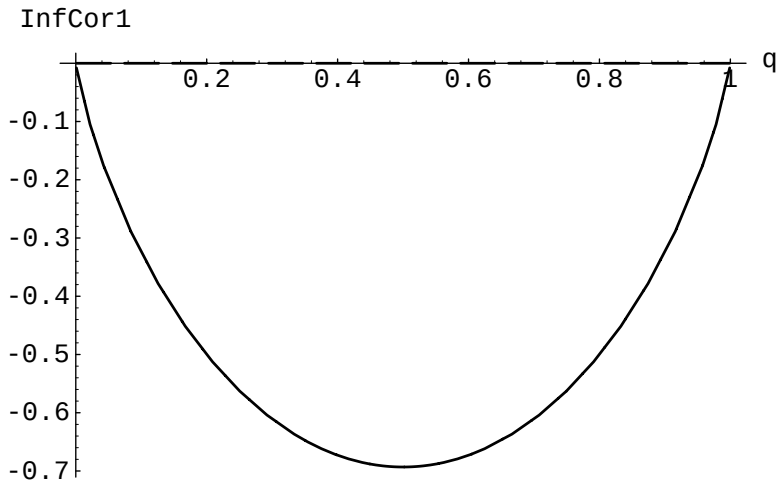

Figure 4.6: First order information correlation at $N = 10$ as a function of q. Real part is solid. Imaginary part is dashed. This is also known as negative of the first order entropy.

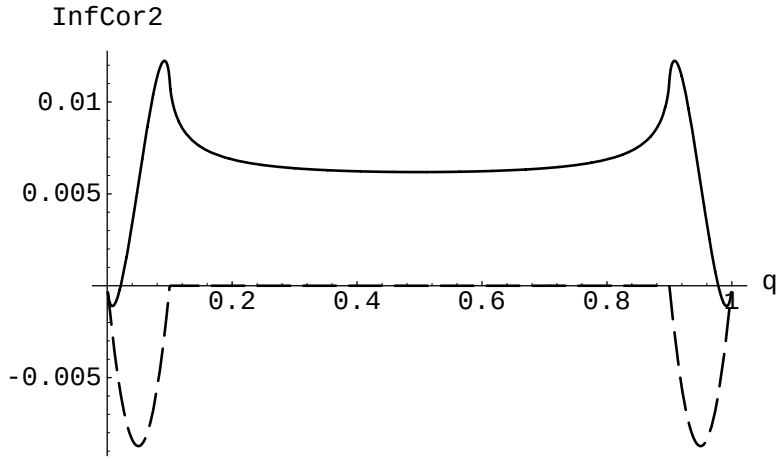

Figure 4.7: Second order information correlation at $N = 10$ as a function of q. Real part is solid. Imaginary part is dashed. This is also known as the mutual information.

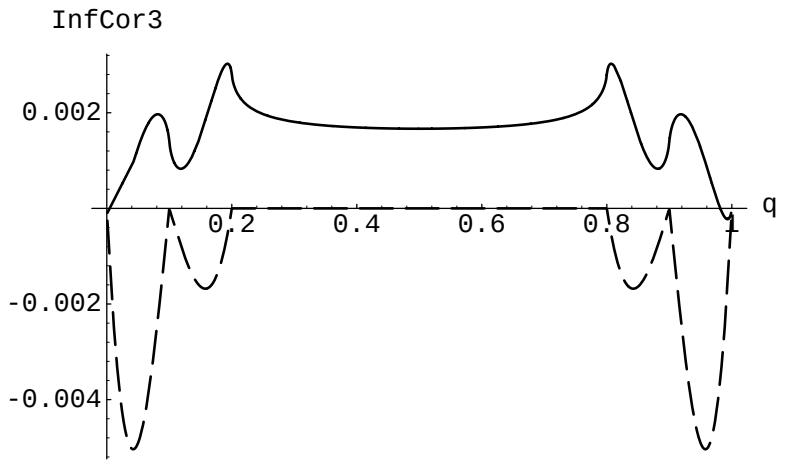

Figure 4.8: Third order information correlation at $N = 10$ as a function of q. Real part is solid. Imaginary part is dashed.

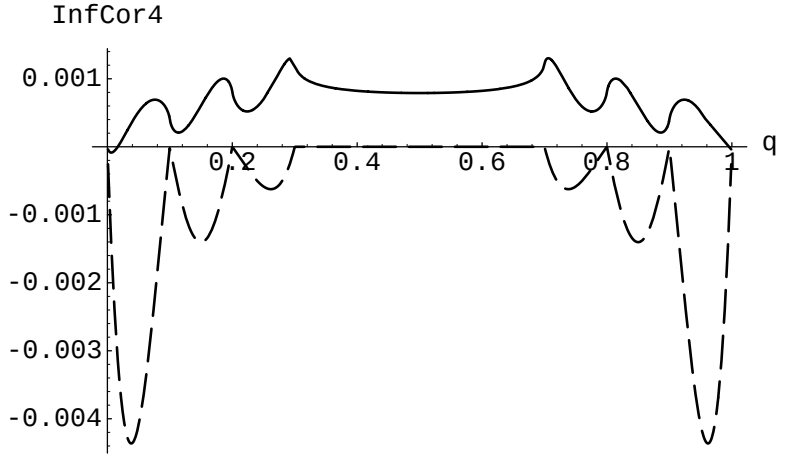

Figure 4.9: Fourth order information correlation at $N = 10$ as a function of q. Real part is solid. Imaginary part is dashed.

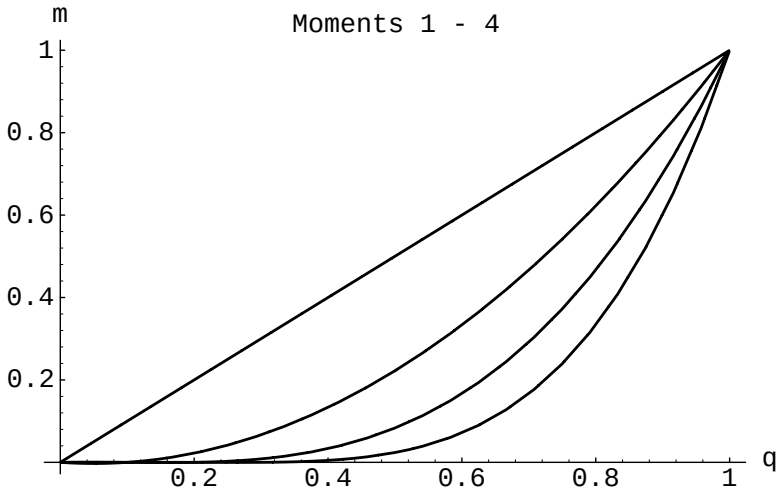

Figure 4.10: Moments 1 thru 4 at $N = 10$ as a function of q. Increasing order moments are more curved.

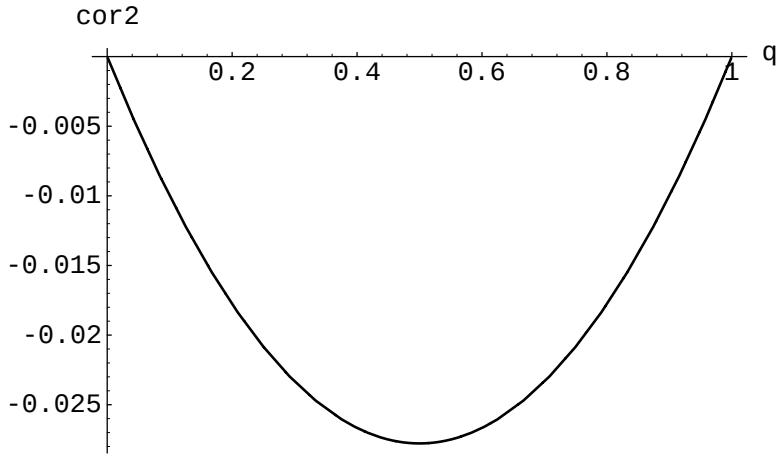

Figure 4.11: Second order correlation at $N = 10$ as a function of q. This is also the second order cumulant.

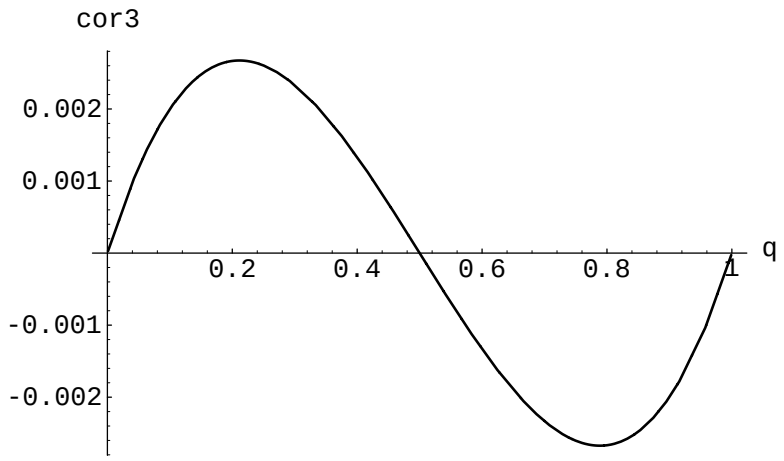

Figure 4.12: Third order correlation at $N = 10$ as a function of q. This is also the third order cumulant.

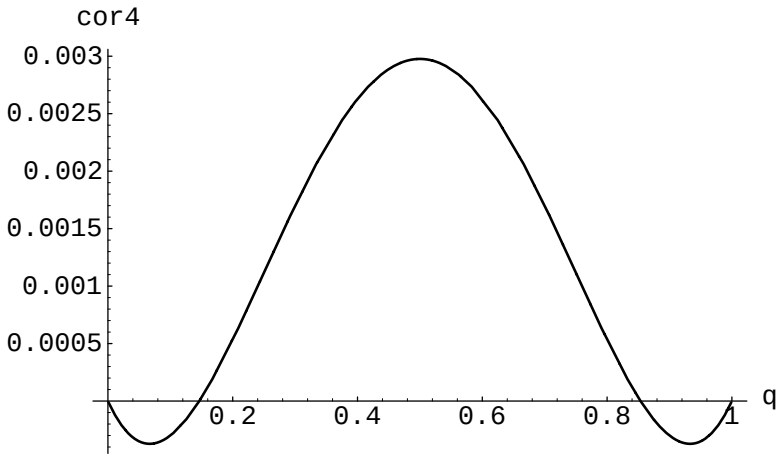

Figure 4.13: Fourth order correlation at $N = 10$ as a function of q.

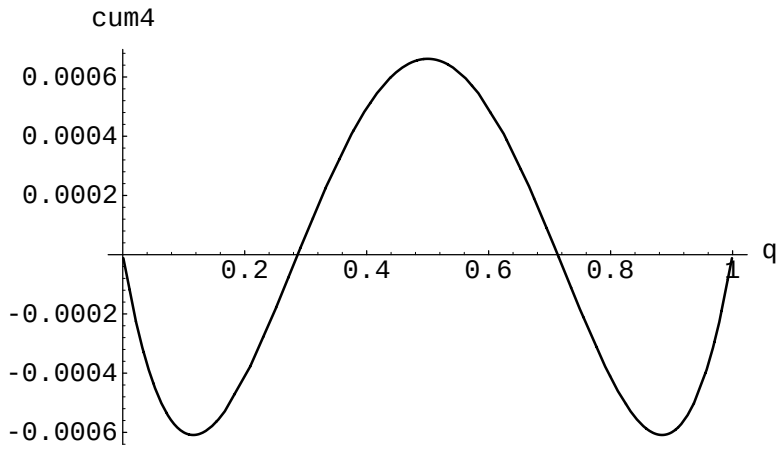

Figure 4.14: Fourth order cumulant at $N = 10$ as a function of q.

4.13 Appendix A - Combinatorial identities

The identities

$$\begin{aligned} C(N-2,B) + C(N-2,B-1) &= C(N-1,B) \\ C(N-2,B-1) + C(N-2,B-2) &= C(N-2,B-1) \\ C(N-2,B) + 2C(N-2,B-1) + C(N-2,B-2) &= C(N,B) \end{aligned}$$

were found useful in deriving the result for S_2 in section 4.5. The proof follows.

Consider $N-1$ slots containing B ones. When the first slot is 0 then B one bits are distributed into into $N-2$ slots, and when the first slot is 1, then $B-1$ one bits are distributed into $N-2$ slots. The sum of the number of ways that each case can happen counts the number of ways that B one bits may be distributed into $N-1$ slots, or

$$C(N-2,B) + C(N-2,B-1) = C(N-1,B-1). \tag{4.22}$$

$N-1$ slots with B ones X $\binom{N-1}{B}$ ways		$\binom{N-1}{B}$ ways
0	$N-2$ slots with B ones	$\binom{N-2}{B}$ ways
1	$N-2$ slots with $B-1$ ones	$\binom{N-2}{B-1}$ ways

Similarly $C(N-2,B-1)+C(N-2,B-2) = C(N-2,B-1)$. Finally, consider N slots, and look at the possible ways to have the first two slots $00, 01, 10, 11$ respectively. Their sum counts B one bits into N slots.

Chapter 5

Classical equations of information flow

5.1 Liouville equation

The time evolution of the density function is needed to discuss the time evolution of the correlations and information correlation functions. The key characteristic of the density function is that it integrates to one, which implies the probability conservation equation in general, then more specifically a flow equation when a velocity field is definable (deterministic continuous dynamics), and the Liouville equation when in addition the dynamics is Hamiltonian. In this section we discuss the development of the Liouville equation for the full density function, and the corresponding equations for the reduced density functions.

In general, let $\rho(\boldsymbol{x}, t)$ be the density function describing the probabilities of seeing state $\boldsymbol{x}$ at time t. Since the probability at any time of seeing some state is one, we have $\int \rho(\boldsymbol{x}, t)\, d\boldsymbol{x} = 1$. Microscopically, consider a volume element V and the probability in this element at time t, $p(V, t) = \int_V \rho(\boldsymbol{x}, t)\, d\boldsymbol{x}$. In general the time evolution of a set of many variables is not going to be deterministic (in the sense that the future state of the system is not specified by the present state of the system alone), such as in any relaxation process driven by radiative cooling or by other interactions of the system with a high complexity and uncertainty outside world. However, if the systems described by the density function evolve deterministically and continuously then it is possible to define a velocity field on the space of states $\boldsymbol{x}$. The velocity field

is simply given by the rate of change of system points. Then, the change in the probability within V with respect to time is due to the flow through the surface S of the volume element, so we write

$$\frac{d}{dt}p(V,t) = \int_V \frac{\partial}{\partial t}\rho(\boldsymbol{x},t)\,d\boldsymbol{x} = -\int_S \rho(\boldsymbol{x},t)\,\dot{\boldsymbol{x}}\cdot d\boldsymbol{S} \tag{5.1}$$

With the use of Gauss' theorem, the surface integral may be rewritten and we find

$$\int_V \frac{\partial}{\partial t}\rho(\boldsymbol{x},t)\,d\boldsymbol{x} = -\int_V \nabla\cdot(\rho(\boldsymbol{x},t)\,\dot{\boldsymbol{x}})\,d\boldsymbol{x} \tag{5.2}$$

Taking the small volume limit in equation 5.2, the probability conservation equation, gives us the equation of continuity

$$\frac{\partial}{\partial t}\rho(\boldsymbol{x},t) = -\nabla\cdot(\rho(\boldsymbol{x},t)\,\dot{\boldsymbol{x}}) \tag{5.3}$$

So far this equation holds whenever a velocity field $\dot{\boldsymbol{x}}$ is definable. In the case of Hamiltonian evolution, it turns out that the probability density behaves like an incompressible fluid. We show this by considering Hamilton's equations $\partial H/\partial p = \dot{q}$ and $\partial H/\partial q = -\dot{p}$ for each component of the coordinate and conjugate momentum. Operate with the divergence so that

$$\nabla\cdot(\rho(\boldsymbol{x},t)\,\dot{\boldsymbol{x}}) = (\,\nabla\rho\cdot\dot{\boldsymbol{x}} + \rho\nabla\cdot\dot{\boldsymbol{x}}\,) \tag{5.4}$$

Then, noting that $\boldsymbol{x} = (\boldsymbol{p},\boldsymbol{q})$ and making the substitutions from Hamilton's equations we find for Hamiltonian evolution that $\rho\nabla\cdot\dot{\boldsymbol{x}} = 0$ and

$$\nabla\cdot(\rho(\boldsymbol{x},t)\,\dot{\boldsymbol{x}}) = \{\rho,H\} \tag{5.5}$$

where $\{A,B\} := \frac{\partial A}{\partial q}\frac{\partial B}{\partial p} - \frac{\partial B}{\partial q}\frac{\partial A}{\partial p}$, the sum over components being implicit. Finally, noting that we may take the volume small, and noting that $d\rho/dt = \partial\rho/\partial t + \nabla\rho\cdot\dot{\boldsymbol{x}}$ we find for Hamiltonian evolution

$$\frac{d\rho}{dt} = \frac{\partial\rho}{\partial t} + \{\rho,H\} = 0 \tag{5.6}$$

which is known as Liouville's equation. The total time derivative of the density being zero indicates that the density at any time evolving system point remains constant, and that in turn the volume integral of any point function of the density, such as the entropy or information correlation functions, remains constant.

In summary we have gone from a very general notion of probability conservation, through the addition of the information that the system evolves deterministically, to the additional constraint that the system evolves according to a Hamiltonian.

5.2 Reduced density functions and the BBGKY hierarchy

Integrating the density function over a subset of the variables of dependence yields a reduced density function. In this section the subscript of the density function indicates its order. For example, for $m < n$, $\rho_m(\boldsymbol{x}_m) := \int d\boldsymbol{x}_{\overline{m}}\, \rho_n(\boldsymbol{x}_n)$, where $d\boldsymbol{x}_{\overline{m}} := dx_{m+1} \ldots dx_n$. The reduced density function also has a dynamics, which may be found by integrating the solution found from the Liouville equation. Integrating both sides of equation 5.6 over $\boldsymbol{x}_{\overline{m}}$ yields what is known as the BBGKY hierarchy (Born, Bogoliubov, Green, Kirkwood, Yvon) [12]. First, take $H = H_m + H_{\overline{m}}$, where H_m is the part of H that is the sum of functions of p_i's and q_i's with all indices $i \leq m$, to find

$$\int \frac{d\rho}{dt} d\boldsymbol{x}_{\overline{m}} = 0 \tag{5.7}$$

$$= \int [\, \frac{\partial \rho}{\partial t} + \{\rho, H\}\,]\, d\boldsymbol{x}_{\overline{m}} \tag{5.8}$$

$$= \frac{\partial \rho_m}{\partial t} + \sum_{i=1}^{m} \{\rho_m, H_m\}_i + \sum_{i=1}^{m} \int \{\rho, H_{\overline{m}}\}_i\, d\boldsymbol{x}_{\overline{m}} \tag{5.9}$$

Except for the last term, this equation is the Liouville equation for the reduced system with the reduced Hamiltonian. Because of this form, the last term of

equation 5.9 is the only place where the reduced system couples to the rest of the system. As an aside, in deriving equation 5.9 the following expression of the poisson bracket is useful.

$$\{\rho, H\} = \sum_{i=1}^{m} \{\rho, H\}_i + \sum_{i=m+1}^{n} [\, \frac{\partial}{\partial q_i}(\rho \frac{\partial H}{\partial p_i}) - \frac{\partial}{\partial p_i}(\rho \frac{\partial H}{\partial q_i}) \,] \tag{5.10}$$

The last summation always integrates to zero over $\boldsymbol{x}_{\overline{m}}$. In standard presentations (see, for example [69, 38, 57]), the BBGKY hierarchy is a set of equations which takes into account the information that the Hamiltonian has the form $H = H_n^{(1)} + H_n^{(2)}$ with $H_n^{(1)} = \sum_{i=1}^{n} [\, f(p_i) + g(q_i) \,]$ and $H_n^{(2)} = \sum_{1 \le i < j \le n}^{n} V(q_i, q_j)$. If we use this information in equation 5.9 then we have $\{\rho, H_{\overline{m}}\}_i = \{\rho, H_{\overline{m}}^{(2)}\}_i$ for $i < m$, and the coupling of the reduced system to the rest of the system then explicitly occurs only through the potential function. Also for this case, when the distribution function is symmetric under permutation of the indices of its arguments, the coupling term reduces to an integral over $\boldsymbol{x}_{m+1}$ with integrand involving ρ_{m+1}.

5.3 Geometry of the BBGKY equations - general flows

There is a simple geometric derivation of the BBGKY hierarchy that is worthwhile discussing. Note that we cannot directly apply the notion of flow to the reduced density function. There can be multiple solutions to the dynamics at any given time which have the same first m coordinates. These solutions will evolve in separate directions, depending on the initial values of the $n - m$ other coordinates. However, we may define a flow by averaging all of these solutions. The probability density flow at $\boldsymbol{x}_n$ is given by $\rho_n \dot{\boldsymbol{x}}_n$, and the truncation of this to the m space of interest is $\rho_n \dot{\boldsymbol{x}}_m$. We average this over $\boldsymbol{x}_{\overline{m}}$ and define the reduced flow velocity through $\rho_m \dot{\boldsymbol{x}}_m^R := \int \rho_n \dot{\boldsymbol{x}}_m d\boldsymbol{x}_{\overline{m}}$. With this flow defined,

then we have from equation 5.3 the conservation equation

$$\frac{\partial}{\partial t}\rho(\boldsymbol{x}_m, t) + \nabla \cdot (\rho_m \dot{\boldsymbol{x}}_m^R) = 0 \tag{5.11}$$

and making the substitution for the definition of $\dot{\boldsymbol{x}}_m^R$ puts the conservation equation in the form

$$\frac{\partial}{\partial t}\rho(\boldsymbol{x}_m, t) + \nabla \cdot \int \rho_n \dot{\boldsymbol{x}}_m d\boldsymbol{x}_{\overline{m}} = 0 \tag{5.12}$$

Interchanging the ∇ and the integral is possible, and the resulting equation is easily seen to be the same as equation 5.9 (the quantity $\partial\dot{q}_i/\partial q_i + \partial\dot{p}_i/\partial p_i = \nabla \cdot \dot{\boldsymbol{x}} = 0$). Thus, the BBGKY equation for the reduced density function is simply the flow equation for the reduced density; the reduced dynamics is no longer Hamiltonian, but that of the superposition of Hamiltonian systems indicated by ρ_n. This makes it quite clear why no equation involving only ρ_m can succeed: the information from the variables $x_{m+1} \ldots x_n$ affects the evolution at all times.

5.4 Information flow, correlation flow

The observation from the Liouville equation that an isolated Hamiltonian system's entropy is constant (see section 5.1), and in fact that any volume integral of a point function of the density is constant, makes it simple to define the information flow into a subsystem. For if the subsystem were isolated, not coupled to the rest of the system, then the entropy of the subsystem would be a constant S_u. When the subsystem is coupled, let the entropy be $S_c(t)$, a function of time. The negative rate of change of this entropy is the magnitude of information flow into the system, (negative flow indicates that the subsystem is losing information or "structure") With $S_c(t) := -\int \rho_m log(\rho_m)\, d\boldsymbol{x}_m$ we have immediately that

$$F_I = \int (1 + log(\rho_m)) \frac{\partial \rho_m}{\partial t} d\boldsymbol{x}_m \tag{5.13}$$

Note that the above expression is equivalent to $\int log(\rho_m)\frac{\partial \rho_m}{\partial t} d\boldsymbol{x}_m$ since probability is conserved. If the evolution of the subsystem was completely specified by the Hamiltonian H_m, substituting $\partial\rho_m/\partial t = -\{H_m, \rho_m\}$ and using Gauss' theorem to integrate would give zero exactly. Since the flow is not completely determined by H_m we substitute from the BBGKY equation 5.9 to find that the information change in the subsystem is due solely to the coupling of the subsystem to the rest of the system and is

$$F_I = -\int (1 + log(\rho_m))\,[\sum_{i=1}^{m} \int \{H_{\overline{m}}, \rho\}_i \, d\boldsymbol{x}_{\overline{m}}]\, d\boldsymbol{x}_m \tag{5.14}$$

With a slight adjustment to the derivation above we note that in fact the rate of change of any functional of a point function f of the density ρ_m is given by the above with f' substituted for $1 + log(\rho_m)$.

5.5 Maxent outside implies no flow into subsystem

One surprising point. If the maximal entropy distribution ρ_{ME} consistent with ρ_m is substituted for ρ above, i.e. substitute $\rho_{ME} = C^{-1} \times \rho_m$, where $C = \int 1 d\boldsymbol{x}_{\overline{m}}$, for ρ, then the subsystem density effectively decouples from the rest of the system. The change in entropy of the subsystem is then zero. The result is more general.

Theorem. (Factorization of complete density into subsystem and outside densities implies no flow into subsystem). *Let the complete distribution function of n particles factor into ρ_m over the subsystem and $\rho_{\overline{m}}$ over the rest of the system. Then the time derivative of any point function of the subsystem reduced distribution function $f(\rho_m)$ integrated over the subsystem variables has zero time derivative. i.e.*

$$\frac{d}{dt} \int f(\rho_m) d\boldsymbol{x}_m = 0 \tag{5.15}$$

Proof: The time derivative may be written

$$\frac{d}{dt}\int f(\rho_m)d\boldsymbol{x}_m = \int f'(\rho_m)\frac{\partial \rho_m}{\partial t}d\boldsymbol{x}_m \tag{5.16}$$

Substitute for $\frac{\partial \rho_m}{\partial t}$ from the BBGKY equation 5.9,

$$\frac{\partial \rho_m}{\partial t} = -\sum_{i=1}^{m}\{\rho_m, H_m\}_i - \sum_{i=1}^{m}\int \{\rho, H_{\overline{m}}\}_i \, d\boldsymbol{x}_{\overline{m}} \tag{5.17}$$

The $\sum_{i=1}^{m}\{\rho_m, H_m\}_i$ term of the resulting integral may be written as

$$\begin{aligned} &-\int f'(\rho_m)\sum_{i=1}^{m}\{\rho_m, H_m\}_i \, d\boldsymbol{x}_m \\ &\quad = -\sum_{i=1}^{m}\int\left(\frac{\partial f(\rho_m)}{\partial q_i}\frac{\partial H_m}{\partial p_i} - \frac{\partial f(\rho_m)}{\partial p_i}\frac{\partial H_m}{\partial q_i}\right) d\boldsymbol{x}_m \\ &\quad = 0 \end{aligned} \tag{5.18}$$

where the last line, equation 5.18, is found by integrating the line above it using Gauss' theorem. Note that in applying Gauss' theorem, $\int_V \nabla\cdot(f\boldsymbol{v})dV = \int_S f\boldsymbol{v}\cdot d\boldsymbol{a}$, the velocity is not defined as $(\dot{\boldsymbol{q}}, \dot{\boldsymbol{p}})$ because the Hamiltonian appearing is H_m and not the full Hamiltonian $H = H_m + H_{\overline{m}}$. Instead, the velocity is defined as $(\boldsymbol{v}_q, \boldsymbol{v}_p) = (\frac{\partial H_m}{\partial \boldsymbol{p}}, -\frac{\partial H_m}{\partial \boldsymbol{q}})$. The structure that makes $\nabla\cdot(f\boldsymbol{v}) = \{f, H\}$ is built into this definition because $\frac{\partial^2 H_m}{\partial p \partial q} = \frac{\partial^2 H_m}{\partial q \partial p}$. (This proves the statement leading to equation 5.14 that the flow into the subsystem from the H_m term is always zero.) The $\sum_{i=1}^{m}\int\{\rho, H_{\overline{m}}\}_i \, d\boldsymbol{x}_{\overline{m}}$ term may be rewritten as

$$\begin{aligned} &-\int f'(\rho_m)\sum_{i=1}^{m}\left(\int \rho_{\overline{m}}\{\rho_m, H_{\overline{m}}\}_i \, d\boldsymbol{x}_{\overline{m}}\right) d\boldsymbol{x}_m \\ &\quad = -\int \rho_{\overline{m}}\left(\sum_{i=1}^{m}\int f'(\rho_m)\{\rho_m, H_{\overline{m}}\}_i \, d\boldsymbol{x}_m\right) d\boldsymbol{x}_{\overline{m}} \\ &\quad = 0 \end{aligned} \tag{5.19}$$

where the last line, equation 5.19, is found by integrating the inner integral using Gauss' theorem in a manner similar to that leading to 5.18. The sum of these terms is zero; therefore no information flows into the subsystem when the distribution function factors into a subsystem part and an outside part. QED.

Note that the theorem above does not imply that there will be no information flow into the subsystem for all time. It simply states that whenever the ensemble density factors that there is no flow at those times.

The theorem above leads to a somewhat surprising counter-intuitive result: coupling *any* subsystem density to a maximum entropy outside system density does not immediately lead to information flow to or from the subsystem. The corollary below immediately follows from the theorem above.

Corollary. (Maxent outside implies no information flow into subsystem). *If the complete distribution function factors as* $\rho = C^{-1}\rho_m$, *where* $C = \int 1 d\boldsymbol{x}_{\overline{m}}$, *then there is no information flow into the subsystem.*

Some comments on irreversibility and the equations of motion of the reduced density function are in order. These equations may form the basis for an approach to the understanding of the nature of irreversibility in thermodynamical systems. See, for example, [37]. Also [5] is an interesting discussion of the evolution of the correlation structure in Hamiltonian systems with two-body potentials. Other approaches exist, however. A proposal to understand the nature of irreversibility by modifying the equations of motion of physics is reviewed in [15]. Macroscopic coarse-graining plays a large role in some presentations of irreversibility, see for example [53].

It should be pointed out that studying irreversibility from the point of view of the ensemble density is somewhat flawed. In the case of the single universe that we live within there is only one system to consider. The key is to understand why irreversibility is a phenomenon in *this* universe.

Chapter 6

Information and correlation in the Ising system

6.1 Entropy in the Ising System

The following details the calculation of the high order entropies of k neighboring or non-neighboring sites of the linear (1D) Ising system. The system has no intrinsic dynamics, and stochastic dynamics are not considered. Limits as the number of spins in the complete system is increased and as the number of sites being considered is increased are considered. Information correlation functions of the sites are investigated and found to give quantities which are the mutual information between the first and last sites of those considered in the chain up to a sign. Graphs show that the information correlation function indicate where in the phase space of the system the high order distribution functions provide new information.

6.2 Ising Model - Closed chain

For simplicity consider the closed chain Ising model - a string of n neighbor coupled spin sites $\sigma = (s_1, ..., s_n)$ joined at the ends. Thus spin s_1 is coupled to spin s_n. The spin values s_i take on the values -1 or 1. The energy of the system is given by the sum of the internal interaction energies of neighboring spin states and the energy of interaction of the set of states with an external field, which for the Ising model is taken to be a sum of spin-field energies. Thus,

with the spin-spin coupling given by $-Q$ and the spin-field constant $-R$, the energy in the Ising model is given by

$$E(\sigma) = -\sum_{i=1}^{n}[Qs_is_{i+1} + Rs_i] \tag{6.1}$$

6.3 Probability of states

Using full state maxent or the thermodynamic assumption we find that the probability $P_n(\sigma)$ of any state σ with energy $E(\sigma)$ is proportional to $e^{-\beta E(\sigma)}/Z_n(\beta)$ with $\beta = 1/kT$ unspecified for now. The normalization constant (partition function) is given by

$$Z_n(\beta) = \sum_{\sigma} e^{-\beta E(\sigma)} \tag{6.2}$$

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

$$P_n(\sigma) = \frac{e^{-\beta E(\sigma)}}{Z_n(\beta)} \tag{6.3}$$

6.4 Reduced probabilities

Summing $P_n(\sigma)$ over the spins $\sigma_{\overline{k}} := (s_{k+1}, ..., s_n)$ gives the probability of the state $\sigma_k = (s_1, ..., s_k)$. For simplicity, define the k-partition function $Z_{n,k}(\sigma_k, \beta)$ by

$$Z_{n,k}(\sigma_k, \beta) = \sum_{\sigma_{\overline{k}}} e^{-\beta E(\sigma_k, \sigma_{\overline{k}})}. \tag{6.4}$$

The probability of the k-spin state σ_k is given in terms of the k-partition function by

$$P_{n,k}(\sigma_k) = \frac{Z_{n,k}(\sigma_k, \beta)}{Z_n(\beta)} \tag{6.5}$$

where $Z_n(\beta) = Z_{n,0}(\beta)$, and where $P_n = P_{n,n}$ follows.

6.5 Computing the partition functions

We find the exact results for the reduced entropies using the transfer matrix method, previously applied to find correlations (see [87, 30]). Write the energy $E(\sigma)$ symmetrically as

$$\begin{aligned} E(\sigma) &= \sum_{i=1}^{n} E_i(\sigma) \\ E_i(\sigma) &= -[Q s_i s_{i+1} + R/2(s_i + s_{i+1})]. \end{aligned} \tag{6.6}$$

Define the 2×2 transfer matrix L by

$$\begin{aligned} L(s_i, s_{i+1}) &= e^{\beta[Q s_i s_{i+1} + R/2(s_i + s_{i+1})]} \\ &= e^{-\beta E_i(\sigma)} \end{aligned} \tag{6.7}$$

With

$$q := \beta Q \tag{6.8}$$

$$r := \beta R \tag{6.9}$$

we have

$$L(s_i, s_{i+1}) = e^{q s_i s_{i+1} + \frac{r}{2}(s_i + s_{i+1})} \tag{6.10}$$

or

$$L = \begin{bmatrix} e^{q+r} & e^{-q} \\ e^{-q} & e^{q-r} \end{bmatrix}. \tag{6.11}$$

In terms of the transfer matrix L the partition function is given by

$$Z_{n,k}(\sigma_k, \beta) = \sum_{\sigma_{\overline{k}}} \Pi_{i=1}^{n} L(s_i, s_{i+1}), \tag{6.12}$$

These forms may be simplified directly, noting that the sums of the products above amount to matrix products, so that we have

$$Z_{n,k}(\sigma_k, \beta) = L(s_1, s_2) ... L(s_{k-1}, s_k) L^{n-k+1}(s_k, s_1) \tag{6.13}$$

and

$$Z_n(\beta) = Z_{n,0}(\beta) = Tr(L^n) \tag{6.14}$$

6.6 Simplifying the transfer matrix

The mth power of the transfer matrix L may be written using the unit eigenvectors of L, v_1 and v_2, and their corresponding eigenvalues λ_1 and λ_2 as

$$\begin{aligned} L^m &= (v_1^T \lambda_1 v_1 + v_2^T \lambda_2 v_2)^m \\ &= v_1^T \lambda_1^m v_1 + v_2^T \lambda_2^m v_2 \end{aligned} \tag{6.15}$$

$$\begin{aligned} L^m[a,b] &= v_1[a]v_1[b]\lambda_1^m + v_2[a]v_2[b]\lambda_2^m \\ &= \sum_{i=1}^{2} v_i[a]v_i[b]\lambda_i^m. \end{aligned} \tag{6.16}$$

6.7 Expression for the partition function

Using the result for the transfer matrix L^m we find

$$Z_n(\beta) = Tr(L^n) = \lambda_1^n + \lambda_2^n. \tag{6.17}$$

6.8 High order entropies in the Ising system

Define the kth-order entropy of the n-spin Ising system by

$$\begin{aligned} S_{n,k}(\beta) &:= -\sum_{\sigma_k} P_{n,k}(\sigma_k) log(P_{n,k}(\sigma_k)) \\ &= -\sum_{\sigma_k} \frac{Z_{n,k}}{Z_n} log(\frac{Z_{n,k}}{Z_n}) \\ &= -\sum_{\sigma_k} \frac{Z_{n,k}}{Z_n} [log(Z_{n,k}) - log(Z_n)] \\ &= -\frac{1}{Z_n} \sum_{\sigma_k} Z_{n,k} log(Z_{n,k}) + log(Z_n) \end{aligned} \tag{6.18}$$

We may substitute the result for $Z_{n,k}$ into the expression for the entropy, and using $xlog(x) = \frac{d}{d\eta} x^\eta|_{\eta=1}$ (the derivative with respect to η of x^η, setting η to one) then find

$$S_{n,k}(\beta) = - \frac{1}{Z_n} \frac{d}{d\eta}|_{\eta=1} [\sum_{s_1,s_2} ((k-1)L(s_1,s_2)^\eta L^{n-1}(s_2,s_1)$$

$$+L(s_1,s_2)^{k-1}L^{n-k+1}(s_2,s_1)^{\eta})\,]$$
$$+\quad log(Z_n) \tag{6.19}$$

6.9 The full Ising entropy

Setting $k=n$ in 6.19 the expression for the full entropy per spin of the n spin state Ising system is given by

$$\begin{aligned}
\frac{S_{n,n}(\beta)}{n} &= \frac{-1}{nZ_n}\frac{d}{d\eta}|_{\eta=1}\,[\sum_{s_1,s_2}((n-1)L(s_1,s_2)^{\eta}L^{n-1}(s_2,s_1) \\
&\quad +L(s_1,s_2)^{n-1}L(s_2,s_1)^{\eta})\,] + \frac{log(Z_n)}{n} \\
&= \frac{-1}{nZ_n}\frac{d}{d\eta}|_{\eta=1}\left[\sum_{s_1,s_2}(nL(s_1,s_2)^{\eta}L^{n-1}(s_2,s_1))\right] \\
&\quad +\frac{log(Z_n)}{n} \\
&= \frac{-\sum_{s_1,s_2}\left(t(1)log(\lambda_1 t(1))t^T(n-1)\right)}{1+(\frac{\lambda_2}{\lambda_1})^n} \\
&\quad +log(\lambda_1)+log(1+(\frac{\lambda_2}{\lambda_1})^n)
\end{aligned} \tag{6.20}$$

where

$$\begin{aligned}
t(m) &:= L^m(s_1,s_2)/\lambda_1^m \\
&= \sum_{i=1}^{2} v_i[s_1]v_i[s_2](\lambda_i/\lambda_1)^m
\end{aligned} \tag{6.21}$$

6.10 Thermodynamic limit of transfer matrix

With $\lambda_1 > \lambda_2$ we find that

$$\lim_{m\to\infty} t(m) = v_1[s_1]v_1[s_2] \tag{6.22}$$

6.11 Thermodynamic limit of full Ising entropy

Using the result of equation 6.22 and the orthonormality of the eigenvectors v_1 and v_2 of L we find

$$\begin{aligned} S_I &:= \lim_{n\to\infty} S_{n,n}/n \\ &= -\sum_{s_1,s_2} \{t(1)log(t(1))v_1[s_1]v_1[s_2] \\ &\quad +t(1)v_1[s_1]v_1[s_2]log(\lambda_1)\} + log(\lambda_1) \\ &= -\sum_{s_1,s_2} (t(1)log(t(1))v_1[s_1]v_1[s_2]) \end{aligned} \tag{6.23}$$

6.12 Thermodynamic limit of high order entropies

We may also find the thermodynamic limit of the k-spin system. The algebra is omitted here because it's similar to the algebra in the S_I case.

$$\begin{aligned} S_k &:= \lim_{n\to\infty} S_{n,k}/k \\ &= (-1/k) \sum_{s_1,s_2} [\,(k-1)t(1)log(t(1)) \\ &\quad +t(k-1)log(v_1(s_1)v_1(s_2))\,]\, v_1(s_1)v_1(s_2) \end{aligned} \tag{6.24}$$

Note that taking $k \to \infty$ here gives S_I. Graphs of the entropy per spin for orders $1-4$ including both the ferromagnetic $Q > 0$ and $\beta > 0$ and the antiferromagnetic $Q < 0$ and $\beta > 0$ ($Q > 0$ and $\beta < 0$) appear in figures 6.1-6.4. 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.

6.13 Eigenvalues and Eigenvectors

We may write the eigenvalues λ_i and the eigenvectors $v_i, i = 1, 2$ as

$$\lambda_1 = e^q cosh(r) + sqrt(e^{2q} sinh^2(r) + e^{-2q})$$

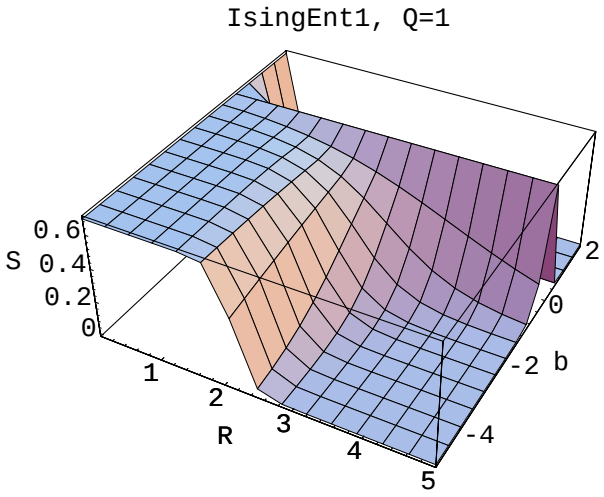

Figure 6.1: First order entropy per spin of the Ising system. Entropy of one spin.

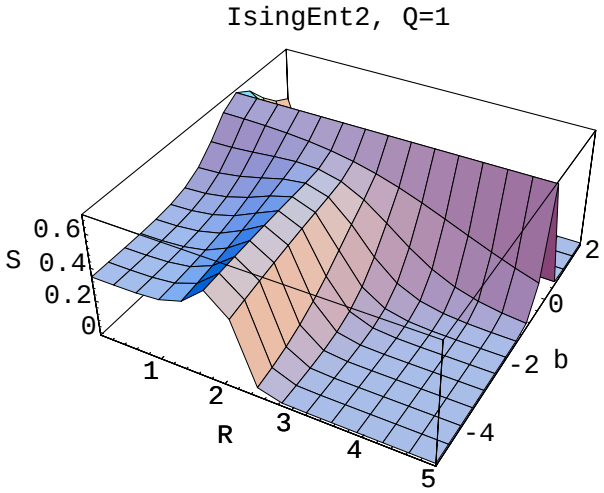

Figure 6.2: Second order entropy per spin of the Ising system. Note the difference between the first and second order entropies, and the similarities of the second-fourth order entropies. The bump on the antiferromagnetic phase side occurs as the external field is increased, and indicates a transition between the $\uparrow\downarrow \ldots$ states, and the $\uparrow\uparrow \ldots$ state.

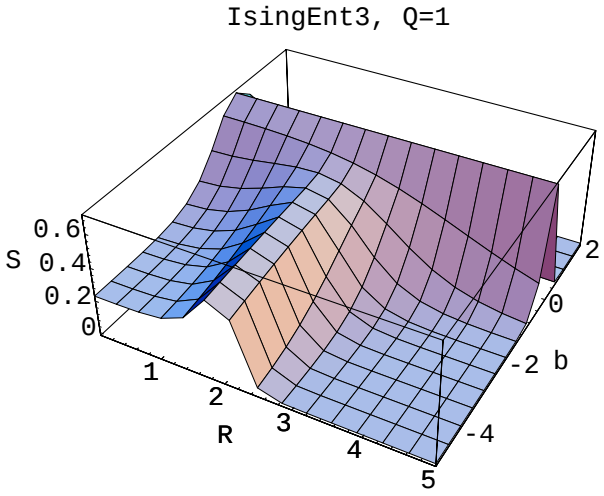

Figure 6.3: Third order entropy per spin of the Ising system.

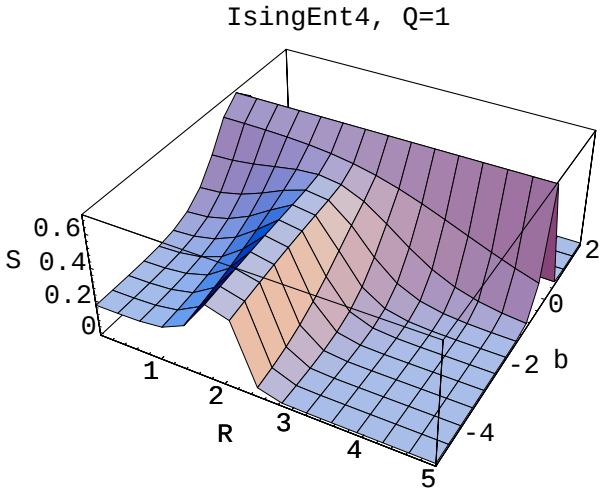

Figure 6.4: Fourth order entropy per spin of the Ising system.

$$\lambda_2 = e^q cosh(r) - sqrt(e^{2q} sinh^2(r) + e^{-2q}) \tag{6.25}$$

$$\begin{aligned} v_1 &= (e^q(e^q sinh(r) + sqrt(e^{2q} sinh^2(r) + e^{-2q}]), 1)/n_1 \\ v_2 &= (e^q(e^q sinh(r) - sqrt[e^{2q} sinh^2(r) + e^{-2q}]), 1)/n_2 \end{aligned} \tag{6.26}$$

where n_1 and n_2 are the normalization constants needed to make v_1 and v_2 unit vectors.

6.14 Eigenvalues and Eigenvectors in zero external field

When the external field $R = 0$ things become simple. The following quantities are needed:

$$\begin{aligned} \lambda_1 &= 2cosh(q), \lambda_2 = 2sinh(q), \\ v_1 &= (1,1)/Sqrt(2), v_2 = (-1,1)/Sqrt(2), \\ t(m) &= [1 + (2\delta(s_1, s_2) - 1)(\lambda_2/\lambda_1)^m]/2 \end{aligned} \tag{6.27}$$

where $\delta(s_1, s_2) := 1$ if $s_1 = s_2, 0$ else.

6.15 High order entropies in zero external field

Substituting for the various quantities with $R = 0$ in the expression for S_k and simplifying we find

$$S_k = (1 - 1/k)(-p^+ log(p^+) - p^- log(p^-)) + log(2)/k. \tag{6.28}$$

where p^+ and p^- are defined as

$$\begin{aligned} p^+ &= [1 + sinh(q)/cosh(q)]/2 \\ p^- &= [1 - sinh(q)/cosh(q)]/2 \end{aligned} \tag{6.29}$$

In terms of the $R = 0$ value of $S_I = -p^+ log(p^+) - p^- log(p^-)$ (see equation 6.23) we have

$$S_k = (1 - 1/k)S_I + log(2)/k = S_I + [log(2) - S_I]/k \tag{6.30}$$

From equation 6.30 we see that the entropy per spin decreases as the order increases, consistent with the reduced entropy per degree of freedom theorem of chapter 2. From the form of equation 6.30 it is implied that the change in entropy with changes in the average energy (or any parameter) decreases in magnitude as the order increases. Thus it is important to observe the high-order structure indicated by the high order entropy in order to understand the creation of that structure as the average energy changes. In fact, since $\frac{\partial S_I}{\partial \langle E \rangle} = \beta$ (see chapter 10, equation 10.22), note that

$$\frac{\partial S_k}{\partial \langle E \rangle} = \beta(1 - \frac{1}{k}) \tag{6.31}$$

Note the extremely simple k-dependence of the kth order entropy S_k. Note that the quantities p^+ and p^- obey

$$p^+ + p^- = 1, \quad 0 \leq p^+, p^- \leq 1, \tag{6.32}$$

so that they may be interpreted as probabilities, and referring to equation 6.5 we see that p^+ and p^- are the probabilities that any two neighboring spins are the *same* and *different* respectively. Note that $0 \leq S_I \leq log(2)$.

Note that all derivatives of S_k with respect to system parameters or variables have the same functional dependence on system parameters up to the factor $1 - 1/k$.

6.16 High order reduced distributions of non-contiguous sites

In this section the general form for the distribution of any subset of the system sites is developed. Consider any set of k distinct spins of the n spin system.

Let $\iota = \{i_1, \ldots, i_k\}$ be the indices of these spins, and without loss of generality let $i_1 < \ldots < i_k$. Label the spin values $\sigma_\iota = (s_{i_1}, \ldots, s_{i_k})$, and the rest of the spin values $\sigma_{\overline{\iota}}$. With this notation, define the partition functions $Z_{n,k}(\sigma_\iota)$, where the number of spins is made explicit in the subscript, but their locations are only made explicit as arguments, as is typical in descriptions of many body distribution functions.

$$Z_{n,k}(\sigma_\iota) = \sum_{\sigma_{\overline{\iota}}} \Pi_{i=1}^{n} L(s_i, s_{i+1}) \tag{6.33}$$

where we identify s_{n+1} with s_1. Simplifying $Z_{n,k}$ we find immediately that

$$Z_{n,k}(\sigma_\iota) = \Pi_{m=1}^{k} L^{i_{m+1}-i_m}(s_{i_m}, s_{i_{m+1}}) \tag{6.34}$$

where $i_{k+1} - i_k$ is taken to be $(n - i_1 + i_k)$. Write the probability for the spin state σ_ι directly as

$$P_{n,k}(\sigma_\iota) = \frac{Z_{n,k}(\sigma_\iota)}{Z_n} \tag{6.35}$$

where $Z_n := Z_{n,0}$.

6.17 High order entropies of non-contiguous sites

Using the probability for the state of k non-contiguous sites of equation 6.35 we find that the entropy of these sites is given by

$$\begin{aligned} S_{n,k}(\iota) &= -\sum_{\sigma_\iota} P_{n,k}(\sigma_\iota) log(P_{n,k}(\sigma_\iota)) \\ &= -\frac{1}{Z_n} \sum_{m=1}^{k} \sum_{s_1, s_2} \{L^{i_{m+1}-i_m}(s_1, s_2)\, log(L^{i_{m+1}-i_m}(s_1, s_2)) \\ &\quad \times L^{n-(i_{m+1}-i_m)}(s_2, s_1)\} + log(Z_n) \end{aligned} \tag{6.36}$$

In the thermodynamic limit ($n \to \infty$) and using the substitution of equations 6.21 and 6.22 we find that the entropy per spin is given by

$$S_k(\iota) = \lim_{n\to\infty} \frac{S_{n,k}(\iota)}{k}$$

$$= -\frac{1}{k}\sum_{s_1,s_2}\{\sum_{m=1}^{k-1} t(i_{m+1}-i_m)\, log(t(i_{m+1}-i_m)) + t(i_k-i_1)\, log(v_1[s_1]v_1[s_2])\ \}\ v_1[s_1]v_1[s_2] \quad (6.37)$$

Compare equation 6.37 to equation 6.24. If the spins are equally spaced, and we let the spacing be $d = i_{m+1} - i_m$, $m = 1, \ldots, k-1$, and define $S_I(d)$ to be the entropy per spin of an infinite ising system, taking every dth spin, while $S_k(d)$ is the similar quantity, but for k of the spins with the spacing d, then

$$S_k(d) = (1-\frac{1}{k})S_I(d) - \frac{1}{k}\sum_{s_1,s_2} t(i_k-i_1)\, log(v_1[s_1]v_1[s_2])\, v_1[s_1]v_1[s_2] \quad (6.38)$$

In zero external field, where from equation 6.27 the eigenvector components have values $\pm 1/\sqrt{2}$, this becomes

$$S_k(d) = S_I(d) - [log(2) - S_I(d)]/k \quad (6.39)$$

which should be compared with equation 6.30. Probabilities of same and different spins separated by d are also similarly given (see equation 6.29) by

$$p^{\pm}(d) = \frac{1 \pm tanh(q)^d}{2} \quad (6.40)$$

6.18 Information correlation functions in the Ising system

Referring to section 3.12 and using the general distribution for the Ising system we find that the ϕ_k of equation 3.26 (here denoted by $\phi_{n,k}$) of the information correlation function C_k equations 3.29 (here denoted by $C_{n,k}$) is given by

$$e^{(-1)^k \phi_{n,k}(\sigma_\iota)} = L^{i_k - i_1}(s_{i_1}, s_{i_k}) \frac{Z_n \Pi_{m=1}^{k-1} L^{n-(i_{m+1}-i_m)}(s_{i_m}, s_{i_{m+1}})}{\Pi_{m=1}^{k} L^n(s_{i_m}, s_{i_m})} \quad (6.41)$$

Taking the logarithm of this and averaging yields $C_{n,k}$. In the large n limit, where there are a large number of spins in the system, the terms in the fraction largely cancel, leaving the simpler result that

$$e^{(-1)^k \phi_{n,k}(\sigma_\iota)} = 1 + \frac{v_2[s_{i_1}]v_2[s_{i_k}]}{v_1[s_{i_1}]v_1[s_{i_k}]}(\frac{\lambda_2}{\lambda_1})^{i_k - i_1} \quad (6.42)$$

A comparison with the distribution of two spins and its marginals shows that C_k is $(-1)^k$ times the mutual information between the spins (see section 6.16). In general this mutual information and C_k are then

$$\begin{aligned} M(i_1, i_k) &= \sum_{s_1, s_2} v_1(s_1)^2 v_1(s_2)^2 \times \\ &\quad (1 + \frac{v_2[s_1]v_2[s_2]}{v_1[s_1]v_1[s_2]}(\frac{\lambda_2}{\lambda_1})^{i_k - i_1}) log(1 + \frac{v_2[s_1]v_2[s_2]}{v_1[s_1]v_1[s_2]}(\frac{\lambda_2}{\lambda_1})^{i_k - i_1}) \end{aligned} \tag{6.43}$$

$$C_k(\sigma_\iota) = (-1)^k M(i_1, i_k) \tag{6.44}$$

Graphs of the information correlation functions of $2, 3$ and 4 neighboring spins are given in figures 6.5-6.7. When the external field is zero, we find that the ratio of v's is either $+1$ or -1 depending on whether s_{i_1} and s_{i_k} differ. (Note that the logarithms are base e, and that β appears as b in the axis label.) Thus, referring to section 6.14 for the values of the eigenvectors and eigenvalues the nth information correlation function is given simply in this case by

$$C_k(\sigma_\iota) = \frac{(-1)^k}{2} [\, (1 + r^d) log(1 + r^d) + (1 - r^d) log(1 - r^d) \,] \tag{6.45}$$

where $r := \lambda_2/\lambda_1 = tanh(q)$ and $d := i_k - i_1$. Thus, we have a straightforward interpretation of $C_k(\sigma_\iota)$ in the large number of spins limit (regardless of whether the external field is zero) - it is defined on a set of k possibly widely-separated spins, but gives us $(-1)^k$ times the mutual information between the first spin and the last spin. The asymptotics of $C_k(\sigma_\iota)$ in zero field are trivial, since $(-1)^k C_k(\sigma_\iota) \sim r^{2d}/2 + r^{4d}/12 + O(r^{6d})$. Finally, defining $p^+_{i_k - i_1}$ and $p^-_{i_k - i_1}$ to be the probability that spins i_1 and i_k are the same and different, respectively, we have $p^\pm_{i_k - i_1} = (1 \pm r^d)/2$, and

$$C_k(\sigma_\iota) = (-1)^k [\, log(2) - S(p^+_{i_k - i_1}, p^-_{i_k - i_1}) \,] \tag{6.46}$$

where $S(p, 1-p) := -p log(p) - (1-p) log(1-p)$.

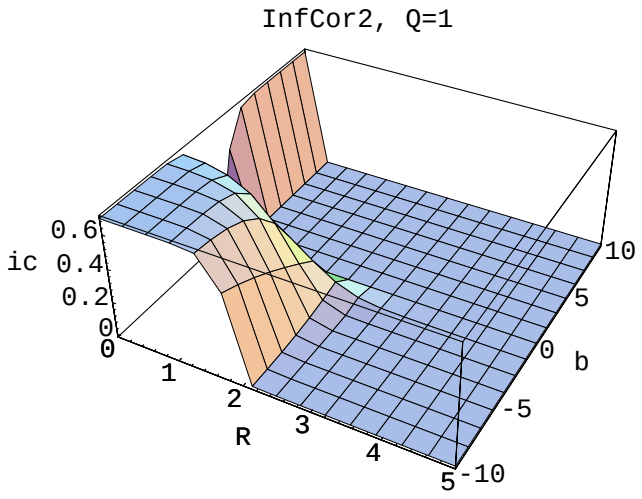

Figure 6.5: Second order information correlation function of the Ising system. Information correlation of two neighboring spins. Note that the information correlation functions are similar up to a sign at each order for this system. The information correlation functions are the mutual information of the first and last spins along the chain in the k spins considered at order k times $(-1)^k$

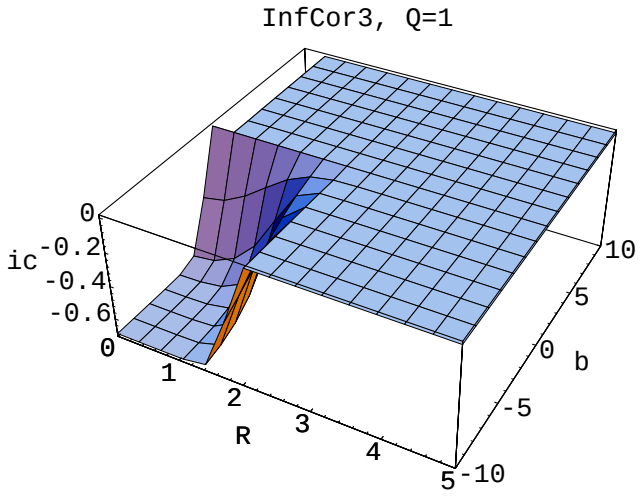

Figure 6.6: Third order information correlation function of the Ising system. Information correlation of three neighboring spins.

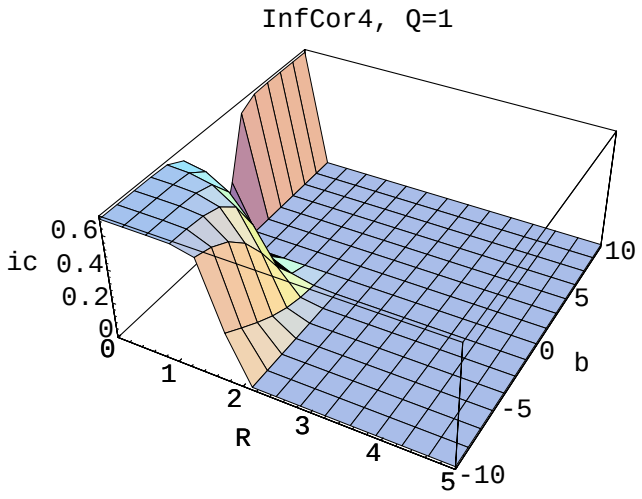

Figure 6.7: Fourth order information correlation function of the Ising system. Information correlation of four neighboring spins.

6.19 Forms of the information correlation function

The result of equation 6.44 is a specific case of a more general result. It is easily seen that any $C_k(\sigma_\iota)$ on a set of random variables where there are *two* of the random variables v_1 and v_2 which are conditionally independent given *any other* random variable(s) is the mutual information between the two random variables. I.e. if $P(v_1, v_2 \mid v_i) = P(v_1 \mid v_i)\, P(v_2 \mid v_i)$ for $i \neq 1, 2$ then $C_k(v_1, \ldots, v_k) = (-1)^k C_2(v_1, v_2)$. This occurs in any chain of random variables, where fixing any random variable in the middle of the chain effectively removes any dependence that the random variable at one end of the chain has on the random variable at the other end of the chain. Other interesting results of this nature are obtainable. For instance, C_4 for four random variables where three of them decouple given the fourth (imagine a three pointed star tree for the dependency of coupling, with the fourth variable at the center of the tree) reduces to $-C_3$ on the three that decouple when the fourth is specified, and so on. Another result of this nature is based on similar topological considerations: if a random variable is a cut point of the dependency tree (removing it splits the tree into two nonempty parts), then the information correlation function

on all of the random variables reduces to a lower order information correlation function on the subsets of random variables that are generated by removing the cut point random variable.

6.20 Moments, correlations, and cumulants

From equation 6.35, the moments of k spins in the Ising system are given by

$$\langle s_{i_1} \dots s_{i_k} \rangle = \sum_{\sigma_\iota} P_{n,k}(\sigma_\iota) \ \Pi_{m=1}^{k} s_{i_m} \tag{6.47}$$

$$= \frac{1}{Z_n} \sum_{\sigma_\iota} \Pi_{m=1}^{k} s_{i_m} L^{i_{m+1}-i_m}(s_{i_m}, s_{i_{m+1}}) \tag{6.48}$$

In the thermodynamic limit this simplifies as

$$\langle s_{i_1} \rangle = f(1,1) \tag{6.49}$$

$$\langle s_{i_1}, s_{i_2} \rangle = \sum_{a} f(1,a)^2 \left(\frac{\lambda_a}{\lambda_1} \right)^{i_2 - i_1} \tag{6.50}$$

$$\langle s_{i_1} \dots s_{i_k} \rangle = \sum_{\boldsymbol{a}} f(1,a_1) \ \Pi_{m=1}^{k-2} \left[f(a_m, a_{m+1}) \left(\frac{\lambda_{a_m}}{\lambda_1} \right)^{i_{m+1}-i_m} \right] \times f(a_{k-1}, 1) \left(\frac{\lambda_{a_{k-1}}}{\lambda_1} \right)^{i_k - i_{k-1}} \tag{6.51}$$

where

$$f(a,b) = \sum_{s} s v_a[s] v_b[s] \tag{6.52}$$

From the moments, the correlations and cumulants may be calculated. Figures 6.8- 6.11 show moments of orders 1 through 4 of neighboring spins. Figures 6.12- 6.14 show the correlation functions of orders 2 through 4 of neighboring spins. Figure 6.15 shows the cumulant function of orders 4 of neighboring spins. Recall that the first three correlation and cumulant functions are equal by order.

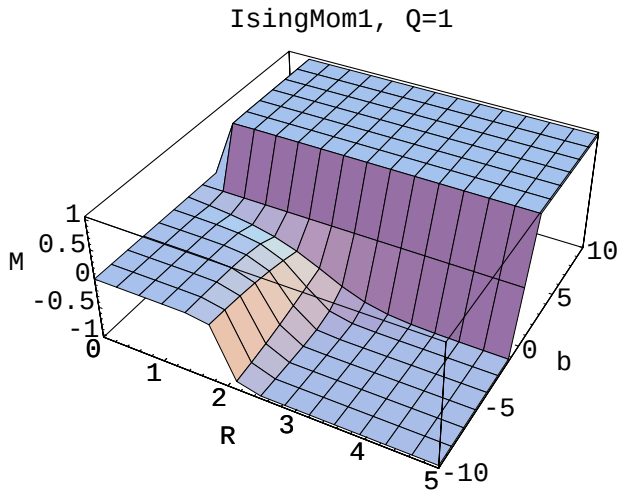

Figure 6.8: First order moment - moment of one spin. Note the transitional behavior in the region of the increase in the entropy, where the states $\uparrow\downarrow \ldots$ becomes dominated by $\uparrow\uparrow \ldots$.

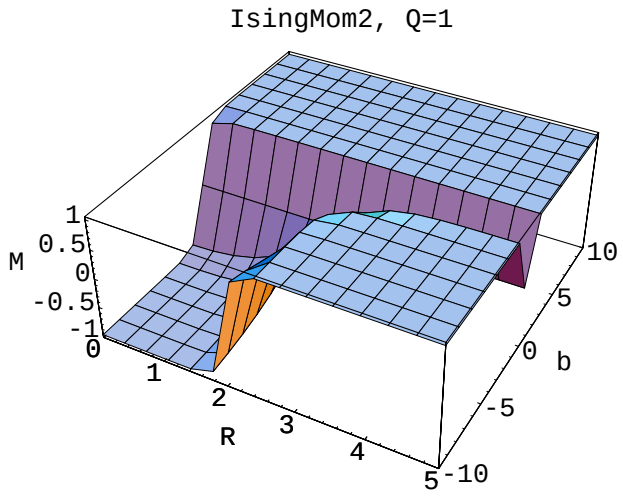

Figure 6.9: Second order moment of two neighboring spins.

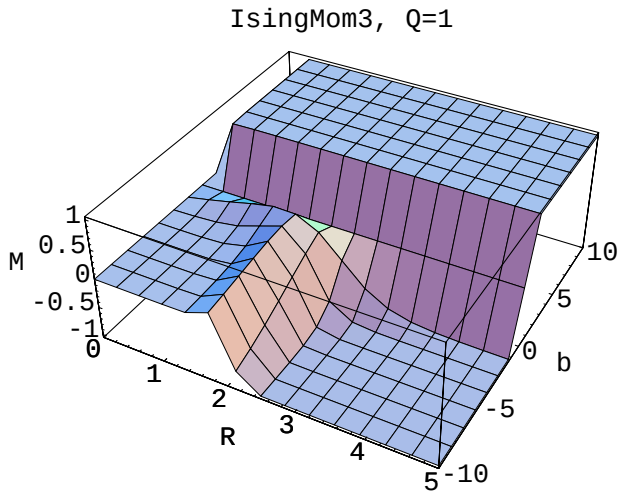

Figure 6.10: Third order moment of three neighboring spins.

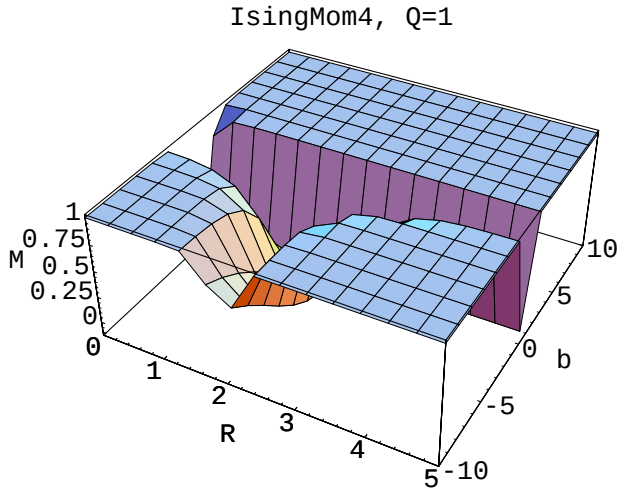

Figure 6.11: Fourth order moment of four neighboring spins.

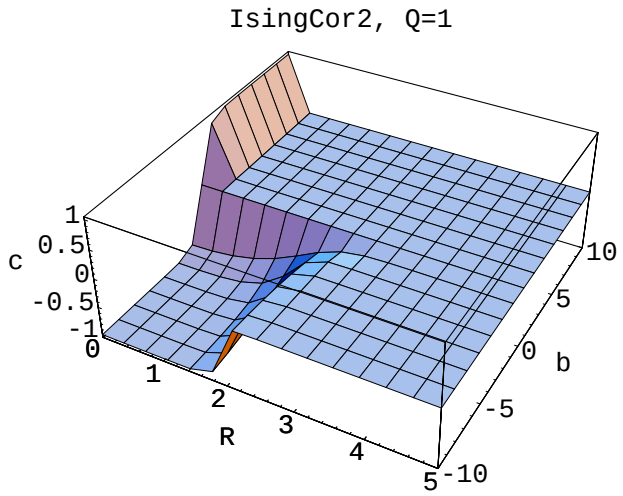

Figure 6.12: Second order correlation of two neighboring spins.

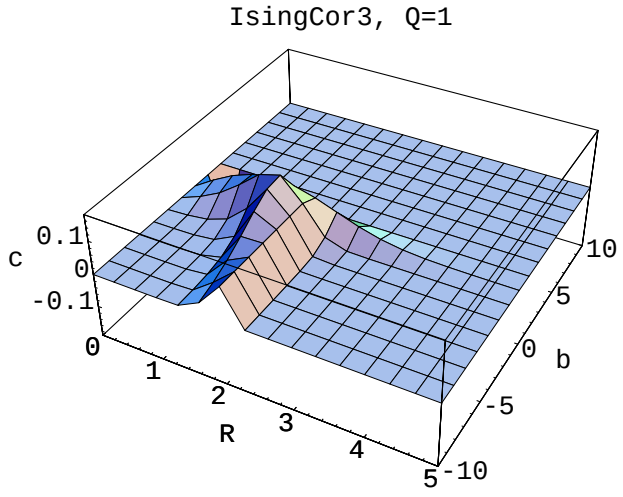

Figure 6.13: Third order correlation of three neighboring spins.

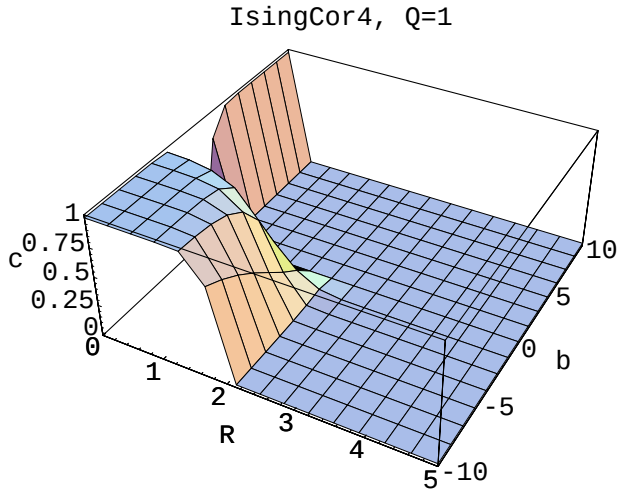

Figure 6.14: Fourth order correlation of four neighboring spins.

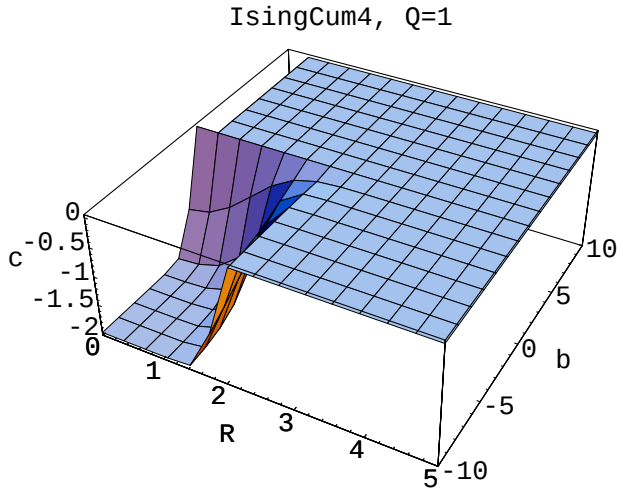

Figure 6.15: Fourth order cumulant of four neighboring spins.

Chapter 7

Quantum equations of information flow

7.1 Equation of motion of pure states

In the quantum case the dynamics of pure states $\boldsymbol{\alpha}$ are governed by Schrodinger's equation,

$$i\hbar\dot{\boldsymbol{\alpha}} = H\boldsymbol{\alpha} \tag{7.1}$$

Note that $\boldsymbol{\alpha}$ is a vector representing the state. In a particular normalized basis $\{\boldsymbol{\alpha}_k\}$ of states $({\boldsymbol{\alpha}_k}^\dagger\boldsymbol{\alpha}_k = 1)$ we have $\boldsymbol{\alpha}(t) = \sum_k \alpha_k(t)\boldsymbol{\alpha}_k$, where the coefficients $\alpha_k(t)$ are complex functions of time representing the probability amplitudes of the basis states $\boldsymbol{\alpha}_k$. The solution of Schrodinger's equation for $\boldsymbol{\alpha}(t)$ is given by

$$\boldsymbol{\alpha}(t) = e^{-iHt/\hbar}\boldsymbol{\alpha}(0) \tag{7.2}$$

7.2 Density of states, quantum Liouville equation

To capture the notion of a distribution of pure states which have no phase relationship between them consider a collection of states $\boldsymbol{\alpha}$ each occurring at time t with probability density $\rho(\boldsymbol{\alpha}, t)$. In order to simplify the discussion in the following calculations, let $\alpha_k \equiv x_k + iy_k$, ${\alpha_k}^* \equiv x_k - iy_k$, and write the distribution function in conjugate coordinates as $\rho(\boldsymbol{\alpha}, \boldsymbol{\alpha}^*, t)$. Note that ρ is positive semidefinite, and nonzero only on the surface $\boldsymbol{\alpha}^\dagger\boldsymbol{\alpha} = 1$. In anticipation of a quantum Liouville equation, define a flow on the space $(\boldsymbol{\alpha}, \boldsymbol{\alpha}^*)$ by $\boldsymbol{v} = (\dot{\boldsymbol{\alpha}}, \dot{\boldsymbol{\alpha}}^*)$. In a manner analogous to the derivation of the classical Liouville

equation of section 5 we may derive the quantum Liouville equation in terms of the conjugate coordinates. To be technically precise, we must go through the algebra using the coordinates $\{x_k, y_k\}$, and the result of doing this is

$$\frac{d\rho}{dt} = 0 = i\hbar\frac{\partial\rho}{\partial t} + \sum_k [\frac{\partial\rho}{\partial\alpha_k}\frac{\partial(\boldsymbol{\alpha}^\dagger H\boldsymbol{\alpha})}{\partial{\alpha_k}^*} - \frac{\partial(\boldsymbol{\alpha}^\dagger H\boldsymbol{\alpha})}{\partial\alpha_k}\frac{\partial\rho}{\partial{\alpha_k}^*}] \tag{7.3}$$

$$= i\hbar\frac{\partial\rho}{\partial t} + \sum_k \{\rho, \boldsymbol{\alpha}^\dagger H\boldsymbol{\alpha}\}_k \tag{7.4}$$

This is the precise analogy of the classical Liouville equation and, except for the factor $i\hbar$ and the complex coordinates, it has the same form as that equation.

Given ρ, the expected value of an operator Q is given by

$$\langle Q\rangle = \int \rho(\boldsymbol{\alpha}, \boldsymbol{\alpha}^*, t)\ \boldsymbol{\alpha}^\dagger Q\boldsymbol{\alpha}\ d\boldsymbol{\alpha}d\boldsymbol{\alpha}^* \tag{7.5}$$

where we note that we depend on the fact that the probability density ρ is nonzero only on the normalized surface of $\boldsymbol{\alpha}$'s to connect $\boldsymbol{\alpha}$ and $\boldsymbol{\alpha}^*$. Equation 7.5 leads immediately to the coordinate-independent or invariant representation of the expectation value by (note that the operator $\sum_k \boldsymbol{\alpha}_k{\boldsymbol{\alpha}_k}^\dagger = [1]$, where $[1]_{ij} \equiv \delta_{ij}$)

$$\langle Q\rangle = \sum_k \int \rho(\boldsymbol{\alpha}, \boldsymbol{\alpha}^*, t)\ \boldsymbol{\alpha}^\dagger Q\boldsymbol{\alpha}_k{\boldsymbol{\alpha}_k}^\dagger\boldsymbol{\alpha}\ d\boldsymbol{\alpha}d\boldsymbol{\alpha}^* \tag{7.6}$$

$$= \sum_k {\boldsymbol{\alpha}_k}^\dagger [\int \boldsymbol{\alpha}\rho(\boldsymbol{\alpha}, \boldsymbol{\alpha}^*, t)\boldsymbol{\alpha}^\dagger\ d\boldsymbol{\alpha}d\boldsymbol{\alpha}^*] Q\boldsymbol{\alpha}_k \tag{7.7}$$

$$= Tr[[\rho]Q] \tag{7.8}$$

where

$$[\rho] \equiv \int \boldsymbol{\alpha}\rho(\boldsymbol{\alpha}, \boldsymbol{\alpha}^*, t)\boldsymbol{\alpha}^\dagger\, d\boldsymbol{\alpha}d\boldsymbol{\alpha}^* \tag{7.9}$$

is the density operator.

We move from the equation of motion of the probability density of states to the equation of motion of the density operator. Take the equation of motion of the probability density of states, equation 7.4, multiply it on the left by $\boldsymbol{\alpha}$,

multiply it on the right by $\boldsymbol{\alpha}^\dagger$, and then integrate over $\boldsymbol{\alpha}$ and $\boldsymbol{\alpha}^*$. The term involving the partial time derivative of the density of states is trivial to find, but the other terms involve more effort to work out. The partial time derivative term is

$$\int \boldsymbol{\alpha} \frac{\partial \rho}{\partial t} \boldsymbol{\alpha}^\dagger \, d\boldsymbol{\alpha} d\boldsymbol{\alpha}^* = \frac{\partial [\rho]}{\partial t} \tag{7.10}$$

while the other terms are found using the identity

$$\{\rho, \boldsymbol{\alpha}^\dagger H \boldsymbol{\alpha}\}_k = \frac{\partial}{\partial \alpha_k}[\rho \frac{\partial}{\partial \alpha_k{}^*}(\boldsymbol{\alpha}^\dagger H \boldsymbol{\alpha})] - \frac{\partial}{\partial \alpha_k{}^*}[\rho \frac{\partial}{\partial \alpha_k}(\boldsymbol{\alpha}^\dagger H \boldsymbol{\alpha})] \tag{7.11}$$

and integrating by parts, noting that ρ is nonzero only on the normalized surface of $\boldsymbol{\alpha}$'s. Doing the algebra and substituting yields the coordinate-independent equation of motion of the density operator

$$i\hbar \frac{\partial [\rho]}{\partial t} + [[\rho], H] = 0 \tag{7.12}$$

where $[A, B] \equiv AB - BA$ is the operator commutator. (It is perhaps easiest to demonstrate this result by showing that the ijth elements of the operators are equal for all i, j). This relationship can also be derived by doing a direct differentiation with respect to time of $[\rho]$ as defined in equation 7.9 on the one hand, and then on the other, noting that equation 7.4 shows that the probability density at a point moving with the Schrodinger flow is constant, transforming to time zero coordinates, and doing the time derivative on that form. The derivation done here shows how the coordinate-dependent and coordinate-independent descriptions of the motion interrelate, demonstrates how to begin with the coordinate independent quantum Liouville equation and then derive the coordinate independent quantum Liouville equation, and prepares the reader for the quantum reduced density equations of motion, directly analogous to the classical BBGKY equations of section 5.2.

7.3 Quantum BBGKY equations

As in the classical case, section 5.2, the quantum BBGKY equations are the expression of the reduction of the full system to a susbsystem of the full system. Consider an orthonormal basis denoted by the set of vectors $\{\mid \boldsymbol{\alpha}_k\rangle\}$, and define the subsystem by the condition $k \in \{1, \ldots, m\}$. Analogous to the classical case, the Hamiltonian may be written as a part specific to the subsystem and the remaining part, each term of which includes a coupling to the rest of the system,

$$H = H_m + H_{\overline{m}} \tag{7.13}$$

For example, the Hamiltonian

$$H = \sum_{k=1}^{m} \mid \boldsymbol{\alpha}_k\rangle\langle\boldsymbol{\alpha}_k \mid \alpha_k + \sum_{rs=1}^{m} \mid \boldsymbol{\alpha}_r\rangle\langle\boldsymbol{\alpha}_s \mid \alpha_{rs} + H_{\overline{m}} \tag{7.14}$$

with $\alpha_{rs} = \alpha_{sr}^*$ and $\alpha_k \in R$, and where each term in $H_{\overline{m}}$ contains a basis vector with index in $\{m+1, \ldots, n\}$, is a possible expression for a Hamiltonian with pairwise state interactions. Take the quantum coordinate independent form of the density evolution equation, equation 7.12, insert the Hamiltonian in the form of equation 7.13, and trace over all of the states outside the subsystem to find the quantum BBGKY equations for the reduced density matrices

$$i\hbar \frac{\partial [\rho]_m}{\partial t} + [\, [\rho]_m, H_m \,] + Tr_{\overline{m}}[\, [[\rho], H_{\overline{m}}] \,] = 0 \tag{7.15}$$

where $[\rho]_m = Tr_{\overline{m}}[\, [\rho] \,]$. Compare this equation to equation 5.9. As in the classical case, the entropy of the reduced system is a constant unless there is information flow from the rest of the system into the subsystem.

7.4 Schrodinger representation and Heisenberg representation

Until now we have discussed the evolution of a quantum system from the point of view that the wavefunctions are evolving and the operators are fixed

(Schrodinger basis). In this picture, the time evolution of an expectation value is

$$\frac{d\langle Q\rangle}{dt} = \frac{d}{dt}Tr[[\rho]Q] = \frac{1}{i\hbar}[Q,H] \tag{7.16}$$

When the operator Q depends on time, this is modified to

$$\frac{d\langle Q\rangle}{dt} = \frac{1}{i\hbar}[Q,H] + \langle\frac{\partial Q}{\partial t}\rangle \tag{7.17}$$

Now, when the operator Q is time-independent, and the state basis vectors are transformed so that all of the time dependence is in the operators, putting us in the Heisenberg picture, $\mid \boldsymbol{\alpha}(t)\rangle_H \rightarrow \mid \boldsymbol{\alpha}(0)\rangle = U^\dagger \mid \boldsymbol{\alpha}(t)\rangle_S$, and $Q_H = U^\dagger Q U$, then the equation of motion is easily found to be (H_H commutes with H)

$$\frac{d\langle Q_H\rangle}{dt} = \frac{1}{i\hbar}[Q_H,H] \tag{7.18}$$

7.5 Time development of pure state measurements.

When a thermodynamic quantum system is at equilibrium there is no definite phase relationship between the energy eigenstates of the system. The system is in a mixed state, and the probability of any measured value of any quantity is independent of time because the phases of the energy components of the system cannot interfere. In the case of an isolated quantum system, however, the state may be represented by a pure state (taken at time zero, say), $\mid \alpha\rangle = \sum_k c_k \mid k\rangle$, with the sum of the component amplitudes being one, $\sum_k \mid c_k \mid^2 = 1$, and the $\mid k\rangle$ are taken as energy eigenstates. The time evolution of such a state is simple, with each energy component being multiplied by $Exp(-i/\hbar E(k)t)$ following immediately from the Schrodinger equation, equation 7.1. Thus, the time evolution of the state $\mid \alpha\rangle$ above is given by

$$\mid \alpha(t)\rangle = \sum_k c_k \; e^{-\frac{i}{\hbar}E(k)t} \;\; \mid k\rangle \tag{7.19}$$

and the probability of finding the system in the state $\mid \gamma \rangle$ at time t is given by

$$\begin{aligned} p_\gamma(t) &= \mid \sum_k c_k \; e^{-\frac{i}{\hbar}E(k)t} \; \langle \gamma \mid k \rangle \mid^2 \\ &= \sum_{jk} c_j c_k^* e^{-\frac{i}{\hbar}(E(j)-E(k))t} \langle \gamma \mid j \rangle \langle k \mid \gamma \rangle \\ &= \sum_k \mid c_k \mid^2 \mid \langle \gamma \mid k \rangle \mid^2 \\ &\quad +2Re[\sum_{j>k} c_j c_k^* e^{-\frac{i}{\hbar}(E(j)-E(k))t} \langle \gamma \mid j \rangle \langle k \mid \gamma \rangle] \end{aligned} \tag{7.20}$$

Clearly this is not independent of time because the second term is time dependent, unless $\mid \gamma \rangle$ is itself an energy eigenstate, or unless all of the energy eigenstates with nonzero $\mid \gamma \rangle$ amplitude are degenerate - having the same energy.

Chapter 8

Information and correlation in the quantum Heisenberg system

8.1 Description of system, Hamiltonian

This chapter describes the quantum spin system. Various model Hamiltonians are described, the history of previous solution attempts for special forms of these Hamiltonians (quantum and classical) is briefly outlined. The Heisenberg model (spin dimension of three) with a fully cross-coupled set of spins (hypersimplex spatial structure, or infinite space dimension), a uniform coupling constant, and zero external field is taken as being a good non-trivial candidate for exact solution. Failing to find an exact solution, a numerical solution is presented. The equilibrium solution is explored. A classical system which reproduces the equations of motion of the quantum system is explored. The the entropy, moment, correlation, and information correlation functions for the quantum Heisenberg system are discussed.

The quantum spin system is described by a Hamiltonian of the form

$$H = -\frac{1}{2}\sum_{ij} \boldsymbol{S}_i^T \cdot \boldsymbol{Q}_{ij} \cdot \boldsymbol{S}_j - \sum_i \boldsymbol{S}_i \cdot \boldsymbol{R}_i \tag{8.1}$$

where the individual spin vectors in the system are labeled with a subscript. In general the spin-spin coupling $\boldsymbol{Q}$ is a tensor and the external field $\boldsymbol{R}$ is a vector, indicating that the coupling between the spins and between the field and the spins in different spatial directions may not be constant. Further, the subscripts

on $\boldsymbol{Q}$ and $\boldsymbol{R}$ indicate that the coupling between the spins and between the field and the spins may vary depending on their spatial organization. In the next two sections we discuss the various special forms that this Hamiltonian may take, the existing known solutions of these forms, and the dynamical equations that accompany this Hamiltonian.

8.2 Hamiltonian forms and known solutions

Compare the Hamiltonian of equation 8.1 to the Hamiltonian of section 6.2 where the classical Ising model was described. There the dimensionality of the spins was taken to be one, and the dimensionality of the space that the spins are ordered within was taken to be one. The coupling parameter $\boldsymbol{Q}$ was taken constant for nearest neighbors and zero for non-nearest neighbors, and the external field $\boldsymbol{R}$ was taken constant for all spins. The common usage is to name the models after the dimensionality of the spins regardless of the spatial dimension. Thus, one dimensional spins organized in any way spatially with whatever coupling strengths are denoted by the term Ising model. When the dimensionality of the spins is three, the coupled spin system is called the Heisenberg model.

In general, the spin-spin coupling may be anisotropic or isotropic, both in the space and the spin dimensions. The spin-spin coupling can be nearest neighbor, farther than nearest neighbor, etc. We can also consider either a zero or a nonzero external field, $\boldsymbol{R}$, which may be coupled either isotropically or anisotropically in the space and spin dimensions.

Another detail that arises is whether or not the spins have a fixed size. There are so-called soft spin models, where the spins are allowed to have any size, but the Hamiltonian for these soft-spin models is generally taken to be different from that of equation 8.1. See [24] for further details.

There is also the relevant detail of whether an equilibrium solution to

any particular model system has been found, and what the properties of the solution are - does it demonstrate a phase transition, etc. The form of the mean field theory solution is also of interest.

There is no compilation of the various systems that have been considered, and new solutions are regularly being found. In summary, we may categorize by spin and space dimensions, known solutions and properties of the solutions (classical or quantum, in mean field, nonzero or zero external field, isotropic or anisotropic spin-spin coupling in space and spin dimensions, isotropic or anisotropic field-spin coupling in space and spin dimensions, and whether the solution demonstrates interesting properties, such as phase transitions, etc.). For classical spin models the interested reader may consult [6, 7, 83, 30] Classical spin glasses are discussed in [10, 24, 60]. For quantum spin models the reader may consult [86, 44, 89, 64, 1, 70, 82, 56] among others. In all cases no reduced distributions are considered, and no reduced entropies or information correlation functions are considered. [89] is a good example of a numerical paper treating the two-dimensional quantum Heisenberg spin-1/2 model with exchange and dipole interactions. At the time of preparing the final draft of this document the authors of [13] (Los Alamos) indicated that they had heard rumors of a collaboration which was approaching the solution of the long-range quantum Heisenberg model using the results of [13]. A result like this would be useful in finding the complete time-evolution solution of the model.

8.3 Dynamics, comparison of classical and quantum systems.

Dynamics of spin systems can be discussed from both the exact and the stochastic points of view. Here the stochastic point of view is bypassed in favor of exact dynamical solutions. For work relevant to the stochastic point of view with Ising

spins see [24] Strictly, there is no classical dynamics for the Ising model, the spins are fixed values. In fact, classically, the Hamiltonian of equation 8.1 has no dynamics because there is no explicit dependence on the conjugate moentum: $\partial H/\partial p_i = 0 = \dot{q}_i$. However, it is possible to construct a classical model that has non-trivial dynamics which depends on underlying variables not yet stated in equation 8.1. It turns out that the equations of motion of this classical model have the same form as the quantum dynamical equations in the Heisenberg representation. This is demonstrated next for a system of rotating "corner-charged squares". The example given appears to be the simplest classical system involving discrete charged particles that reproduces in form the equations of motion of the angular momentum of the quantum coupled spin system. Another example, perhaps simpler, but not involving discrete charges, would be a system of charged rotating circles or current loops. Note that the exchange interaction approximation that the magnetic field at any square is given in direction by the sum of the spins of the other rotating squares is unphysical. For the magnetic interaction we need to consider a dipole field, but this yields a different equation of motion, as is shown. In effect, the true spin-spin coupling arises not from a fictitious magnetic field as is assumed here, but from the fermion exchange asymmetry. See for example [48].

Consider a classical system of charges consisting of quadruples of charges, with the four charges of the ith quadruple each having the same charge value $e_i = 1$ and mass value $m_i = 1$, and with the charges in each quadruple constrained to lie at the corners of a rigid rotating massless square. Let the positions of the corners of the ith square relative to its center be given by $\boldsymbol{n}_{i,1}, \boldsymbol{n}_{i,2}, \boldsymbol{n}_{i,3}, \boldsymbol{n}_{i,4}$ in counterclockwise order (the direction of the rotation), with $\mid \boldsymbol{n}_{i,\alpha} \mid = 1$. Each square rotates about the axis thru its center and perpendicular to itself, with the corners of the square moving with unit magnitude velocity $\mid \dot{\boldsymbol{n}}_{i,\alpha} \mid = 1$. Thus the mass, length, and charge values in this example can be ignored for now; we may reconstitute them at the end of the calculation. Let

there be a uniform magnetic field $\boldsymbol{B}_i$ in the vicinity of the square. We easily compute that the net force on any square is zero

$$\boldsymbol{F}_i = e_i \sum_{\alpha=1}^{4} \dot{\boldsymbol{n}}_{i,\alpha} \times \boldsymbol{B}_i = e_i \frac{d}{dt}(\sum_{\alpha=1}^{4} \boldsymbol{n}_{i,\alpha}) \times \boldsymbol{B}_i = 0 \times \boldsymbol{B}_i = 0 \tag{8.2}$$

(Note that this is true for any charge configuration which is any number of charges equally spaced on a circle.) The torque on the square is nonzero. We find the torque for two interaction Hamiltonians: the exchange energy H_e and the dipole energy H_d. These Hamiltonians are given by the expressions

$$H_e(i,j) = -\boldsymbol{S}_i \cdot \boldsymbol{S}_j \tag{8.3}$$

$$H_d(i,j) = \sum_{\alpha} \frac{3}{d^2}((\boldsymbol{n}_{i,\alpha} + \boldsymbol{d}) \cdot \boldsymbol{S}_i)((\boldsymbol{n}_{i,\alpha} + \boldsymbol{d}) \cdot \boldsymbol{S}_j) - \boldsymbol{S}_i \cdot \boldsymbol{S}_j \tag{8.4}$$

where $\boldsymbol{d}$ is the vector separation of the squares. Using Hamilton's equation $\partial H/\partial q_k = -\dot{p}_k$ with $\boldsymbol{q} = \boldsymbol{n}_{i,\alpha}$, $\boldsymbol{p} = \dot{\boldsymbol{n}}_{i,\alpha}$, noting that the spin $\boldsymbol{S}_i = \frac{1}{2}\sum_\alpha \boldsymbol{n}_{i,\alpha} \times \dot{\boldsymbol{n}}_{i,\alpha}$ we have

$$Exchange : \ddot{\boldsymbol{n}}_{i,\alpha} = \frac{1}{2}\dot{\boldsymbol{n}}_{i,\alpha} \times \boldsymbol{S}_j \tag{8.5}$$

$$Dipole : \ddot{\boldsymbol{n}}_{i,\alpha} = \frac{1}{2}\dot{\boldsymbol{n}}_{i,\alpha} \times \boldsymbol{S}_j - \frac{3}{2d^2}[(\dot{\boldsymbol{n}}_{i,\alpha} \times \boldsymbol{d})((\boldsymbol{n}_{i,\alpha} + \boldsymbol{d}) \cdot \boldsymbol{S}_j) - (\boldsymbol{n}_{i,\alpha} \times \dot{\boldsymbol{n}}_{i,\alpha} \cdot \boldsymbol{d})\boldsymbol{S}_j] \tag{8.6}$$

The torque is $\dot{\boldsymbol{S}}_i = \frac{1}{2}\sum_\alpha \boldsymbol{n}_{i,\alpha} \times \ddot{\boldsymbol{n}}_{i,\alpha}$. In summing over the charges on the square note also that only terms with an even number of $\boldsymbol{n}$'s appearing may be nonzero (this is true for any charge configuration which is an even number of charges equally spaced on a circle). The identity $\boldsymbol{a} \times (\boldsymbol{b} \times \boldsymbol{c}) = \boldsymbol{b}(\boldsymbol{a} \cdot \boldsymbol{c}) - \boldsymbol{c}(\boldsymbol{a} \cdot \boldsymbol{b})$ is useful when computing the torque, and note that $\boldsymbol{n}_{i,\alpha} \cdot \dot{\boldsymbol{n}}_{i,\alpha} = 0$ always (this is true for any charge configuration which is an even number of charges equally spaced on a circle). Note also that for the square $\boldsymbol{n}_{i,\alpha} = -\dot{\boldsymbol{n}}_{i,\alpha+1}$, where the corner subscript is taken $mod(4)$. Finally, the torque for square i due to $\boldsymbol{S}_j$ is

$$Exchange : \quad \dot{\boldsymbol{S}}_i = \boldsymbol{S}_i \times \boldsymbol{S}_j \tag{8.7}$$

$$Dipole : \quad \dot{\boldsymbol{S}}_i = \boldsymbol{S}_i \times \boldsymbol{S}_j - \frac{3}{d^2}(\boldsymbol{S}_i \times \boldsymbol{d})(\boldsymbol{S}_j \cdot \boldsymbol{d}) \tag{8.8}$$

In three places so far the generality of the various even charge symmetry simplifications has been noted, so we might ask if the result above holds for all even number charge configurations. The answer is no! The single symmetry used to arrive at the equations of motion above that is not general is $\boldsymbol{n}_{i,\alpha} = -\dot{\boldsymbol{n}}_{i,\alpha+1}$, which holds for the square, octagon (in the form $\boldsymbol{n}_{i,\alpha} = -\dot{\boldsymbol{n}}_{i,\alpha+2}$), $4k$-gons, etc. For the circular charged loop there is square symmetry too, and the above result also holds. However, for the rod (two charges), the hexagon, $4k+2$-gons, etc., the result does not hold. For simple derivations of the results above for current loops, see for example [40].

The quantum equations of motion of the operators $\dot{\boldsymbol{S}}_i$, for the Hamiltonian of equation 8.1, in the Heisenberg representation are, from equation 7.18 and assuming that $\boldsymbol{Q}_{ij} = \boldsymbol{Q}_{ji}$, and that $[\boldsymbol{Q}_{ij}]_{\mu\nu} = \delta_{\mu\nu} Q_{ij}$ (diagonal and spin direction isotropic)

$$\dot{\boldsymbol{S}}_i = \frac{i}{\hbar}[H, \boldsymbol{S}_i] = \sum_j \boldsymbol{Q}_{ij} \boldsymbol{S}_i \times \boldsymbol{S}_j + \boldsymbol{S}_i \times \boldsymbol{R}_i \tag{8.9}$$

The identity $[S_{i\mu}, S_{j\nu}] = i\hbar\delta_{ij}\epsilon_{\mu\nu\rho}S_{i\rho}$ is useful in showing this result. Compare equation 8.9 with equation 8.7 to see that the spin-spin coupling terms have the same form in the classical description and the quantum description, though the objects in the two equations are different. The spin-field terms also have the same form.

8.4 Symmetric spin Hamiltonian

After some consideration the Hamiltonian of equation 8.1 was taken for consideration with $\boldsymbol{Q}_{ij}$ a constant, i.e. isotropic in both spin and space dimensions, infinite space dimension, i.e. every spin interacting equally with every other spin, a spin dimension of three, i.e. Heisenberg spins, and zero external field. This is analogous to the Sherrington-Kirkpatrick (SK) model (infinite space dimension) [45], but the interaction strength in the SK case is taken as a site-

random quantity and the spins are Ising, which implies that the SK analysis is necessarily classical. Thus the Hamiltonian of interest is

$$H = -Q/2 \sum_{i,j=1, i\neq j}^{N} \boldsymbol{S}_i \cdot \boldsymbol{S}_j \tag{8.10}$$

When this Hamiltonian is written, as below, in terms of the total spin $\boldsymbol{S} = \sum_i \boldsymbol{S}_i$ and the individual spins then the eigenstates are immediately seen to be those where the system total and individual angular momentum j values are listed.

$$H = -Q/2(\boldsymbol{S}^2 - \sum_i \boldsymbol{S}_i{}^2) \tag{8.11}$$

8.5 Energy eigenvectors of the symmetric spin Hamiltonian

It is necessary to diagonalize this $2N$ dimensional Hamiltonian in order to find the distribution of states at thermodynamic equilibrium. The $2N$ operators $\{\boldsymbol{S}_i^2\}, \{\boldsymbol{S}_{1k}^2 = (\boldsymbol{S}_1 + \ldots + \boldsymbol{S}_k)^2\}$, $k = 2, \ldots, N$, and $S_z = S_{1z} + \ldots + S_{kz}$ form a complete set of operators commuting with H and each other. The N total spin eigenvalues $s_i, i = 1, \ldots, N$, along with the $N-1$ successively larger susbsystem s values $s_{1-2} := s_{12}, s_{1-3} := s_{123}, s_{1-4}, \ldots, s := s_{1-N}$ may be taken. Taking the total S_z value m rounds out the set of $2N$ eigenvalues needed. The energy eigenvalue of any eigenstate depends only on its s value and is $E = s(s+1) - N(1/2)(1/2+1) = s(s+1) - 3N/4$ for spin 1/2 particles. Since spin correlations are of interest, the basis where the S_{iz} are easily found must be used to express the energy eigenstates. Thus step one is to generate a complete list of energy eigenvectors $\mid s_1, s_2, s_{1-2}, s_3, s_{1-3}, \ldots, s, m\rangle$ with the energy eigenvalues described above. The second step is to make use of Clebsch-Gordon coefficients to transform the energy eigenbasis just generated to the $\mid (s_i m_i)\rangle$ basis. The description of how this is done is given in appendix B.

8.6 Spin correlations in the symmetric spin Hamiltonian equilibrium distribution

Once the eigenstates of the Hamiltonian are found and written in a form where S_{iz} is easily found then the distributions for seeing values of $(S_{1z}, \ldots, S_{Nz}) = (m_1, \ldots, m_N)$ are readily generated. First, the energy of each of the energy eigenvectors is found. From these energies the probabilities of the eigenvectors at equilibrium are given by the Boltzmann distribution $P_E(\boldsymbol{v}) \propto exp(-\beta E(\boldsymbol{v}))$. The distribution of the vectors of S_{iz} values is then generated. From this distibution the correlation functions and information correlation functions desired may be computed as a function of temperature and coupling strength.

The distribution of measured spin values is simply found from the representation of the energy eigenvectors in the $\mid (s_i m_i)\rangle$ basis. Given an eigenvector in the energy eigenbasis, the probability of measuring the S_{iz} eigenvalues $(m_1, \ldots, m_k)$ is the square of the amplitude of the component vector $\mid (s_i m_i)_k\rangle$, where the m_i values for $i = k+1, \ldots, N$ are ignored. Then sum this probability times the probability that the energy eigenvector occurs (Boltzmann distribution) over the energy eigenvectors to find the overall probability of the measured values. This is the same thing as tracing over the reduced density matrix for spins $1, \ldots, k$ times the operator $\mid (s_i m_i)_k\rangle\langle(s_i m_i)_k \mid$, i.e.

$$P_M((m_1, \ldots, m_k)) = Tr_k[\rho_k \mid (s_i m_i)_k\rangle\langle(s_i m_i)_k \mid] \tag{8.12}$$

Evaluating the information correlation functions (see section 3.12) is trivial because the information correlation functions are simply a sum involving functions of the probabilities of the observed states, which is what is generated in equation 8.12 above.

Evaluating the moments, and therefore the correlations and cumulants (see section 8.12) is straightforward. In evaluating the average of a product of spins $\langle m_1 m_2, \ldots, m_k\rangle$ each term involving an even number of $m_i = -1/2$

values contributes a positive quantity. Every other term contributes a negative quantity. The magnitude of each term is $(1/2)^k$ times the probability of the state. The moments are given by the trace

$$\langle m_1 \dots m_k \rangle = Tr_k[\rho_k S_{1z} \dots S_{kz}] \tag{8.13}$$

8.7 Symmetric spin Hamiltonian with external field

When the external field $\boldsymbol{R}$ is applied to the system then the zero field spin-flip symmetry is broken. The operator $S_z = S_{1z} + \ldots + S_{Nz}$ was chosen as one of the complete commuting set in section 8.5 to label the energy eigenvectors. Thus those eigenvectors are also eigenvectors of

$$H = -Q/2(\boldsymbol{S}^2 - \sum_i \boldsymbol{S}_i{}^2) - R_z S_z \tag{8.14}$$

which is the Hamiltonian of equation 8.11 with an external field applied only in the z direction. The eigenvalues in the energy eigenvector basis are now given by $s(s+1) - N(1/2)(3/2) + mR_z$. The probabilities of each eigenstate are calculated from the Boltzmann distribution using these energies, and as before the probability distribution of the $S_{1z} \dots S_{kz}$ eigenstates, the information correlation functions, moments, correlations, and cumulants may be calculated.

8.8 Entropy of the spin system

The spin system is in thermodynamic equilibrium with the environment and therefore the density function is diagonal in the energy eigenstates, each energy eigenstate $\mid \boldsymbol{v}\rangle$ having probability $P_E(\boldsymbol{v}) \propto exp(-\beta E(\boldsymbol{v}))$ as already mentioned. The density matrix is therefore given by

$$\rho = \sum_{\boldsymbol{v}} P_E(\boldsymbol{v}) \mid \boldsymbol{v}\rangle\langle \boldsymbol{v} \mid \tag{8.15}$$

The entropy of this system is then given by

$$S_E(V) \quad = \quad -Tr[\rho log(\rho)] \tag{8.16}$$

$$= \quad -\sum_{\boldsymbol{v}} P_E(\boldsymbol{v}) log(P_E(\boldsymbol{v})) \tag{8.17}$$

As already mentioned, the probabilities of the various measured values of the z-components of the spins are given by

$$P_M(\boldsymbol{m}_k) = Tr_k[\rho_k \mid \boldsymbol{m}_k\rangle\langle\boldsymbol{m}_k \mid] \tag{8.18}$$

where $\boldsymbol{m}_k = (m_1, \ldots, m_k)$ and the s_i values are dropped here in the representation of the S_{iz} eigenvectors. Based on this probability distribution we have the measurement entropy,

$$S_M(M_k) = -\sum_{\boldsymbol{m}_k} P_M(\boldsymbol{m}_k) log(P_M(\boldsymbol{m}_k)) \tag{8.19}$$

which, because of the full symmetry of the Hamiltonian is independent of which particular k spins are chosen. It may come as some surprise that $S_M(M_k) \neq S_E(V)$ (almost always) even when $k = N$, the number of spins in the system. However, the two entropies are related, and their difference is a physically meaningful quantity. The fact that they are not equal is shown next, and the generality of the argument to follow indicates that care must be taken when defining the entropy of a system from measurements of the states of that system.

Consider the distribution of measured $\boldsymbol{m}_k = (m_1, \ldots, m_k)$ states given in equation 8.18. We can make the following expansion of this as

$$\begin{aligned}
P_M(\boldsymbol{m}_k) &= Tr_k[\rho_k \mid \boldsymbol{m}_k\rangle\langle\boldsymbol{m}_k \mid] && (8.20)\\
&= Tr_k[\sum_{\boldsymbol{m}_{\overline{k}}} \langle\boldsymbol{m}_{\overline{k}} \mid \left(\sum_{\boldsymbol{v}} P_E(\boldsymbol{v}) \mid \boldsymbol{v}\rangle\langle\boldsymbol{v} \mid\right) \mid \boldsymbol{m}_{\overline{k}}\rangle \mid \boldsymbol{m}_k\rangle\langle\boldsymbol{m}_k \mid] && (8.21)\\
&= \sum_{\boldsymbol{v}} P_E(\boldsymbol{v}) \left(\sum_{\boldsymbol{m}_{\overline{k}}} C_{\boldsymbol{v}}(\boldsymbol{m}_k, \boldsymbol{m}_{\overline{k}}) C^*_{\boldsymbol{v}}(\boldsymbol{m}_k, \boldsymbol{m}_{\overline{k}})\right) && (8.22)\\
&= \sum_{\boldsymbol{v}} P_M(\boldsymbol{m}_k \mid \boldsymbol{v}) P_E(\boldsymbol{v}) && (8.23)
\end{aligned}$$

where

$$\mid \boldsymbol{v}\rangle \;=\; \sum_{\boldsymbol{q}_k,\boldsymbol{q}_{\overline{k}}} C_{\boldsymbol{v}}(\boldsymbol{q}_k,\boldsymbol{q}_{\overline{k}}) \mid \boldsymbol{q}_k\rangle \mid \boldsymbol{q}_{\overline{k}}\rangle \tag{8.24}$$

$$\rho_k \;=\; \sum_{\boldsymbol{m}_{\overline{k}}} \langle \boldsymbol{m}_{\overline{k}} \mid \left(\sum_{\boldsymbol{v}} P_E(\boldsymbol{v}) \mid \boldsymbol{v}\rangle\langle \boldsymbol{v} \mid \right) \mid \boldsymbol{m}_{\overline{k}}\rangle \tag{8.25}$$

$$\;=\; \sum_{\boldsymbol{v}} P_E(\boldsymbol{v}) \sum_{\boldsymbol{q}_k,\boldsymbol{r}_k} \left(\sum_{\boldsymbol{m}_{\overline{k}}} C_{\boldsymbol{v}}(\boldsymbol{q}_k,\boldsymbol{m}_{\overline{k}}) C^*_{\boldsymbol{v}}(\boldsymbol{r}_k,\boldsymbol{m}_{\overline{k}}) \right) \mid \boldsymbol{q}_k\rangle\langle \boldsymbol{r}_k \mid \tag{8.26}$$

$$P_M(\boldsymbol{m}_k \mid \boldsymbol{v}) \;=\; \sum_{\boldsymbol{m}_{\overline{k}}} C_{\boldsymbol{v}}(\boldsymbol{m}_k,\boldsymbol{m}_{\overline{k}}) C^*_{\boldsymbol{v}}(\boldsymbol{m}_k,\boldsymbol{m}_{\overline{k}}) \tag{8.27}$$

and where $\boldsymbol{m}_{\overline{k}} = (m_{k+1},\ldots,m_N)$, and the $\boldsymbol{q}_k$ and $\boldsymbol{r}_k$ are also (S_{iz}) eigenstates, etc.

Now from equation 8.27 note that for $k = N$ we have $C_{\boldsymbol{mv}} := P_M(\boldsymbol{m} \mid \boldsymbol{v}) = C_{\boldsymbol{v}}(\boldsymbol{m}) C^*_{\boldsymbol{v}}(\boldsymbol{m})$, where the subscript on $\boldsymbol{m}$ has been dropped. Because of the orthonormality of the various vectors involved $\sum_{\boldsymbol{m}} C_{\boldsymbol{mv}} = \sum_{\boldsymbol{v}} C_{\boldsymbol{mv}} = 1$. Note also that $P_M(\boldsymbol{m}) = \sum_{\boldsymbol{v}} C_{\boldsymbol{mv}} P_E(\boldsymbol{v})$. Making the appropriate substitutions in the log sum inequality we find

Theorem. Measurement increases entropy. *Given a density matrix ρ describing a quantum mechanical system of the form $\rho = \sum_{\boldsymbol{v}} P_E(v) \mid \boldsymbol{v}\rangle\langle \boldsymbol{v} \mid$, where the $\mid \boldsymbol{v}\rangle$ are orthonormal, and an orthonormal measurement basis $\mid \boldsymbol{m}\rangle$, define the intrinsic entropy $S_E = -\sum_{\boldsymbol{v}} P_E(\boldsymbol{v}) log(P_E(\boldsymbol{v}))$ and the measurement entropy $S_M = -\sum_{\boldsymbol{m}} P_M(\boldsymbol{m}) log(P_M(\boldsymbol{m}))$, where $P_M(\boldsymbol{m}) = Tr[\rho \mid \boldsymbol{m}\rangle\langle \boldsymbol{m} \mid]$. Then $S_M \geq S_E$, with equality iff $P_E(\boldsymbol{v})$ is independent of $\boldsymbol{v}$.*

Proof: By the log sum inequality of chapter 2 we have

$$\sum_{\boldsymbol{mv}} C_{\boldsymbol{mv}} P_E(\boldsymbol{v}) log\left(\frac{C_{\boldsymbol{mv}} P_E(\boldsymbol{v})}{C_{\boldsymbol{mv}}} \right) \geq \sum_{\boldsymbol{m}} \left(\sum_{\boldsymbol{v}} C_{\boldsymbol{mv}} P_E(\boldsymbol{v}) log\left(\frac{\sum_{\boldsymbol{v}} C_{\boldsymbol{mv}} P_E(\boldsymbol{v})}{\sum_{\boldsymbol{v}} C_{\boldsymbol{mv}}} \right) \right) \tag{8.28}$$

Consider the quantity on the left, cancel the $C_{\boldsymbol{mv}}$'s and sum over $\boldsymbol{m}$, then consider the quantity on the right, sum over $\boldsymbol{v}$ where these sums appear, finally change signs on both sides, and using equation 8.23 find that

$$S_E(V) \leq S_M(M_N) \tag{8.29}$$

Note that the entropy of the reduced measurements ($k < N$) may be less than the intrinsic entropy simply because of the reduction in dimensionality involved.

It is important to note that quantities like $P(\boldsymbol{m}, \boldsymbol{v})$ need to be carefully considered. For instance, the notion of making simultaneous measurements of $\boldsymbol{m}$ and $\boldsymbol{v}$ is impossible, because the operators for these eigenvalues do not generally commute. It is possible to time-order the "measurements" (now called filters), and then quantities like $P(\boldsymbol{m}_1, \boldsymbol{v}_2 \mid \rho)$ appear, where the M filter is applied first, followed by the V filter. This quantity indicates the probability that the quantum object follows the route through channel $\boldsymbol{m}$ of M, then through channel $\boldsymbol{v}$ of V. This allows us to perform strange inferences of retrodiction, like $P(\boldsymbol{m}_1 \mid \boldsymbol{v}_2, \rho)$, the probability that the object passed through channel $\boldsymbol{m}$ of M having been found in channel $\boldsymbol{v}$ of V, addressed briefly in the next section. We can go much further though, and do in the next section: we may ask the question *what is the information in one measurement that would have been available in another measurement that could have been, but wasn't, made?*.

In summary, we have the relationship that $S_E(V) \leq S_M(M_N)$.

In figure 8.1 an example of the relationship $S_E(V) \leq S_M(M_N)$ is demonstrated. (Note that the logarithms are base e, and that β appears as b in the axis label.)

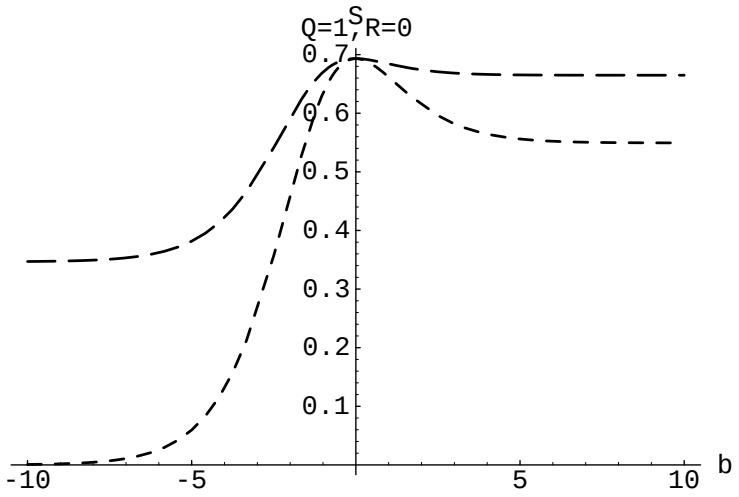

Figure 8.1: Typical plot of the entropy of the density matrix S_E (narrow dashes) and the entropy of the measured values of the z-components of the spins S_M (wide dashes) for the two-spin case and order two entropies. Note $S_M \geq S_E$.

8.9 Mutual information between measured and unmeasured variables

Suppose that the measurement basis M is all that is available to the experimenter, but that the result of a computation is given by the probabilities of the various eigenstates of another operator, V. This is similar to the situation we now have in the Heisenberg spin system, where M is the z-component spin operator, and V is the energy operator, but for now these operators will be taken to be any arbitrary operators. The question of interest is *What is the information about V, which could have been measured but was not measured, gained in a measurement of M?*. Note that it is somewhat bizarre to be asking this question, because quantum objects, being what they are, do not generally allow the measurement of both M and V simultaneously. The act of measuring V would indelibly change forever the object being measured. We are discussing the possibility of gaining information about a measurement result that, in fact, cannot exist.

Let the underlying density matrix of the system being measured be

given by ρ. Then the probabilities for measuring the eigenstates of M and V are given by $P(\boldsymbol{m}_1 \mid \rho)$ and $P(\boldsymbol{v}_1 \mid \rho)$ respectively, where the subscript indicates the time-order of the filter application. If the filter V is applied first, followed by the filter M, then the joint probability of seeing eigenstates of M and V is given by $P(\boldsymbol{m}_2\boldsymbol{v}_1 \mid \rho)$. It is tempting to answer the question posed above with the quantity below, the time-ordered mutual information between the two filters

$$\sum_{\boldsymbol{m}\boldsymbol{v}} P(\boldsymbol{m}_2\boldsymbol{v}_1 \mid \rho) log \left(\frac{P(\boldsymbol{m}_2\boldsymbol{v}_1 \mid \rho)}{P(\boldsymbol{m}_2 \mid \rho)P(\boldsymbol{v}_1 \mid \rho)} \right) \tag{8.30}$$

where the probability distribution $P(\boldsymbol{m}_2 \mid \rho)$ is the reduction of $P(\boldsymbol{m}_2\boldsymbol{v}_1 \mid \rho)$ over $\boldsymbol{v}$. This is indeed the mutual information between the measurements of M and V, assuming that the filter V is applied before the filter M. However, in the question posed above, no actual measurement of V is to be made! In fact, the proposed measurement of V and the measurement of M are conditionally independent once ρ is specified.

What has been ignored, and is needed, is that the measurement of M is indirectly a measurement of ρ, and it is this ρ that *would have been* measured by V. In other words, the probability that ρ was the underlying distribution, given that some eigenvalue of M was measured, and then the probability that V was measured on this underlying ρ, must appear in the information between the unmade measurement of V and the measurement of M. In order to do this, the prior probability of the underlying ρ must be specified, as is seen in the following development. Admittedly, it is strange to be discussing a distribution over potential, but unmade, measurements; yet it is consistent. The probability of a measurement that *could have been* made can be defined!

As before, let the distributions of measured eigenstates of M and V, given ρ, be $P(\boldsymbol{m} \mid \rho)$ and $P(\boldsymbol{v} \mid \rho)$ respectively. Let the prior probability that ρ was the underlying state be given by $P(\rho)$, and let the probability that ρ was the underlying state, given that the eigenvalue $\boldsymbol{m}$ of M was measured, be

given by $P(\rho \mid \boldsymbol{m})$. This probability distribution is given by Bayes' theorem (see section 9.1), once both $P(\boldsymbol{m} \mid \rho)$ and $P(\rho)$ are given:

$$P(\rho \mid \boldsymbol{m}) = \frac{P(\boldsymbol{m} \mid \rho)P(\rho)}{P(\boldsymbol{m})} \tag{8.31}$$

where $P(\boldsymbol{m}) = \int P(\boldsymbol{m} \mid \rho)\, P(\rho)\, d\rho$. The probability $P(\boldsymbol{v} \mid \boldsymbol{m})$ that the eigenvalue $\boldsymbol{v}$ of V *would have been* measured given that the eigenvalue $\boldsymbol{m}$ of M was measured is conceptually clearly given by

$$P(\boldsymbol{v} \mid \boldsymbol{m}) = \int P(\boldsymbol{v} \mid \rho, \boldsymbol{m})\, P(\rho \mid \boldsymbol{m})\, d\rho \tag{8.32}$$

Now, $P(\boldsymbol{v} \mid \rho, \boldsymbol{m}) = P(\boldsymbol{v} \mid \rho)$, and similarly $P(\boldsymbol{v}, \boldsymbol{m} \mid \rho) = P(\boldsymbol{v} \mid \rho)P(\boldsymbol{m} \mid \rho)$. The last of these indicates that the joint distribution conditioned on ρ of the unmeasured V eigenvalue and the measured M eigenvalue may be treated as if both were measured on the same state ρ. Finally, applying these identities and Bayes' theorem, the unconditioned joint distribution is easily seen to be

$$P(\boldsymbol{v}, \boldsymbol{m}) = \int P(\boldsymbol{v}, \boldsymbol{m} \mid \rho)\, P(\rho)\, d\rho \tag{8.33}$$

and as usual

$$P(\boldsymbol{v}, \boldsymbol{m}) = P(\boldsymbol{v} \mid \boldsymbol{m})P(\boldsymbol{m}) \tag{8.34}$$

so that all of the usual identities for events can be defined for the unmeasured V events and the measured M events. The time-ordered mutual information between the measurement and the unmeasurement is given by

$$I_2(M_{meas}, V_{unmeas}) = \sum_{\boldsymbol{m}\boldsymbol{v}} P(\boldsymbol{v}, \boldsymbol{m}) log \left(\frac{P(\boldsymbol{v}, \boldsymbol{m})}{P(\boldsymbol{v})P(\boldsymbol{m})} \right) \tag{8.35}$$

It is interesting to note that this mutual information is exactly zero if ρ is known in advance (the prior probability of ρ is a delta function distribution), and so there must be some uncertainty in the underlying physical state for the mutual information just defined to be nonzero. This is admittedly the usual case. I propose that this is the quantity that should be considered when making

measurements (M) of a system that is to yield information about some aspect of the system that another measurement (V) *would have been* able to provide more directly.

I leave it to another day to come to terms with what it means to have a probability distribution, like that just defined, over events that are assumed to not happen, and which, if happened, would fundamentally change the result of the measurement that is made. It is as if a step out of the usual probability theory has been taken.

8.10 Entropies

The entropies for the various spin systems considered were computed. There are three principal results that may be seen in the figures. (Note that the logarithms are base e, and that β appears as b in the axis label.)

At infinite temperature, $\beta = 0$, every state is equally probable, and the entropy/spin reaches its maximum value of $log(2)$.

For ferromagnetic, or positive, spin-spin coupling constant $Q > 0$ the behavior of the entropy of any order and any number of spins is simple. For both decreasing temperature ($\beta \to \infty$) and increasing external field R the entropy decreases. This is seen in the plots of the first and second order entropy of the two spin system in figures 8.2 and 8.3 which show the typical ferromagnetic behavior. When the external field is zero, the entropy approaches a nonzero value at zero temperature, determined by the degeneracy of the lowest energy level. The entropy S_E in zero field for $2, 3$, and 4 spins is graphed in figure 8.4. The temperature at which the entropy "transitions" from its infinite temperature value to its zero temperature value is given approximately by the mean-field theory value of the temperature, which is different for each system depending upon the number of spins in the system. Note that the "transition" temperature is increasing as the number of spins increases. This is in accord

with the mean field theory result in the thermodynamic limit of an infinite number of spins that $k_B T_c = dQ/2$, where d is the coordination number of each spin and the critical β is $\beta_c = 1/(k_B T_c)$ [30]. Thus we expect for a system of N spins, where because the spins are fully coupled to each other $d = N - 1$, that $\beta_c = 2/[(N-1)Q]$. All of the plots are with $Q = 1$, so that this is $\beta_c = 2/(N-1)$ here, decreasing for increasing number of spins in the system. The decreasing behavior is seen. That the transition is not sharp, nor properly given by the mean field theory value, may be attributed to the fact that a finite number of spins has been considered.

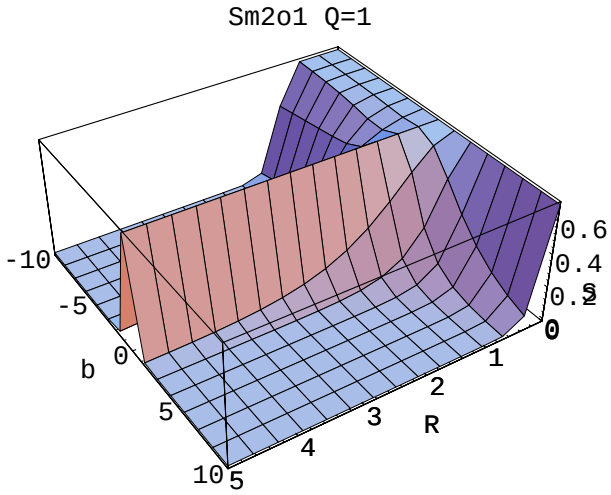

Figure 8.2: The ferromagnetic side of the first order entropy S_M for the 2 spin case showing relatively little structure.

Antiferromagnetic, or negative spin-spin coupling Q, also effectively occurs for positive Q and $\beta < 0$. The graphs of S_M in the $2, 3$, and 4 spin cases in figures 8.5–8.13 show more complex structure in this case than in the ferromagnetic case, with several entropy maxima occurring as the external field varies. These maxima correspond to states having higher energy in zero external field becoming energetically favorable in nonzero fields. Consider three probabilities of energy eigenstates, each having a different coupling to the external field given by $p_1 \propto Exp[-\beta(1-R)], p_2 \propto Exp[-\beta(2-2R)], p_2 \propto Exp[-\beta(4-3R)]$,

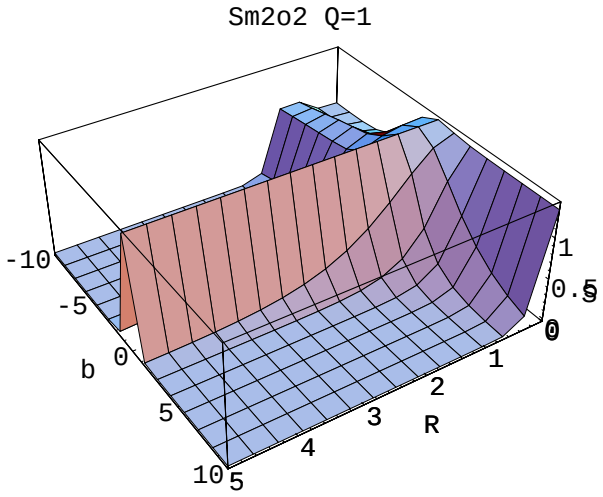

Figure 8.3: The ferromagnetic side of the second order entropy S_M for the 2 spin case showing relatively little structure.

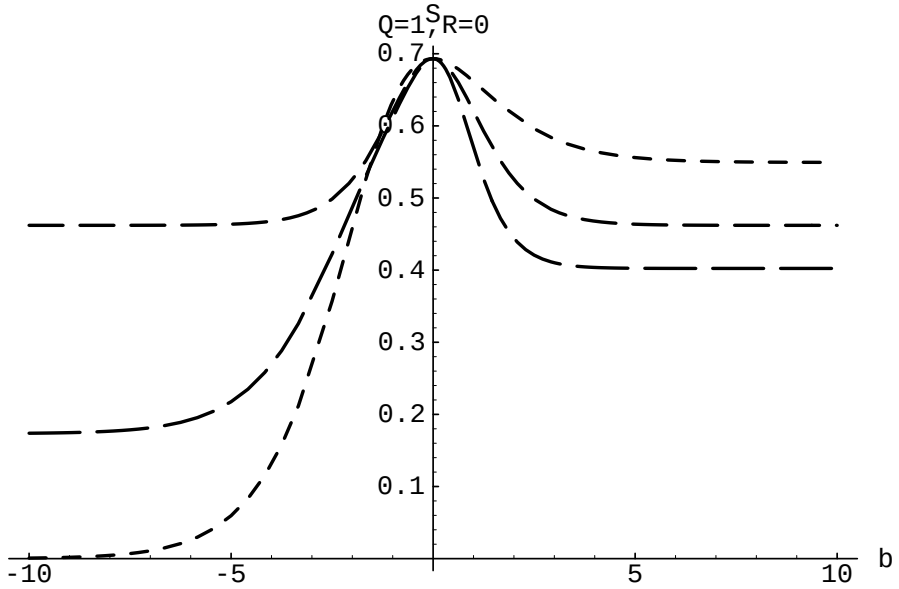

Figure 8.4: Entropy of the density matrix S_E for the 2 (narrowest dashes), 3 and 4 spin (widest dashes) cases. Note the increasing temperature at which the larger spin system transitions on the ferromagnetic side of the plots (positive β) in approximate agreement with the mean field theory result that the transition temperature is proportional to the coordination number.

and let $\beta = 1$ for now. For $0 \leq R < 1/2$ they are ordered as $p_1 > p_2 > p_3$. For $R = 1/2$ there is a crossing, $p_2 = p_3$, and then for $1/2 < R < 2/3$ we have $p_1 > p_3 > p_2$. For $R = 2/3$ there is another crossing, and $p_1 = p_3$. For $2/3 < R < 1$ we have $p_3 > p_1 > p_2$. Finally, for $R = 1$ we find $p_1 = p_2$ and for $R > 1$ we have $p_3 > p_2 > p_1$. In this manner there can be phase transitions where the dominant behavior switches between different eigenstates. At zero temperature, such transitions are sharp. At nonzero temperature, the entropy increases at such transitions because there are more states then contributing on the average. This is seen clearly in a cross section of the four-spin system, where there are two entropy local maxima as the external field changes, figure 8.14. Figures 8.15 and 8.16 show the probabilities of states and the transition behavior just described for the three spin system. Figure 8.15 shows the energy eigenstate probabilities, while figure 8.16 shows the probabilites of measuring (S_{iz}) values.

Figure 8.14 also shows the emergence of the structure in the entropy. Plotted here are the entropies of orders one through four. Note the increasing amount of structure as the order increases. Note also that the complete structure does not make itself apparent until the fourth order, though the third order entropy does show a hint that there are two transition regions in the phase space. The entropies have been normalized by the number of spins involved in the summation for the entropy, and they are ordered consistent with the entropy per degree of freedom relationship theorem of chapter 2.

Consider the sequence of graphs of figures 8.10–8.13 showing the reduced entropies and the full entropy for the four spin system. The reduced entropy graphs show much less structure. It is thus important to consider the full entropy for such systems in order to understand their behavior. For instance, the local entropy maxima seen in the order four plot indicate that there are two distinct states that might be set with an external field. The information

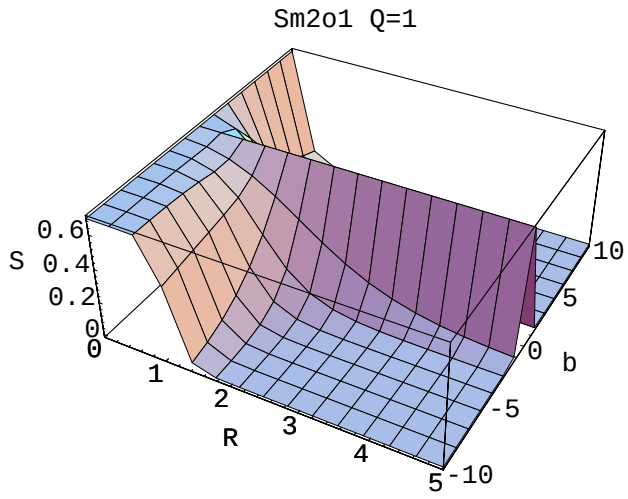

Figure 8.5: First order entropy S_M for the 2 spin case. Detail in this and the next eight plots is of the antiferromagnetic phase.

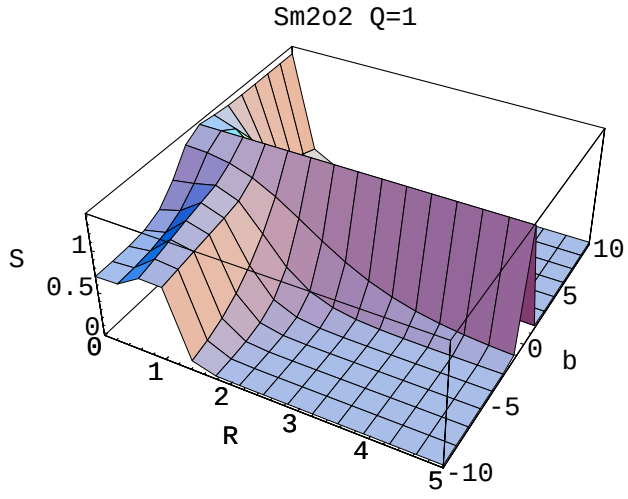

Figure 8.6: Second order entropy S_M for the 2 spin case.

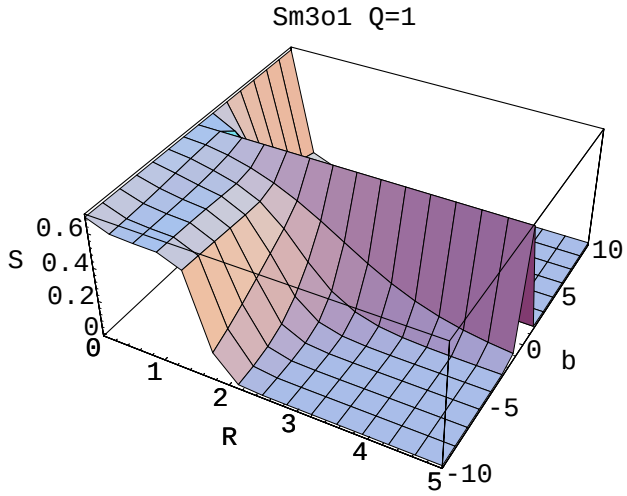

Figure 8.7: First order entropy S_M for the 3 spin case.

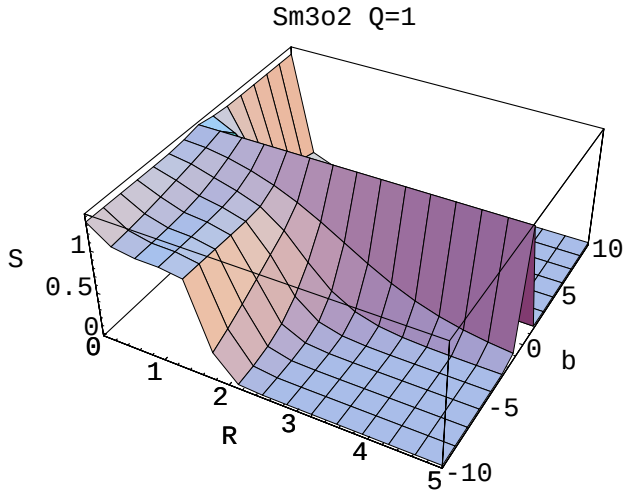

Figure 8.8: Second order entropy S_M for the 3 spin case.

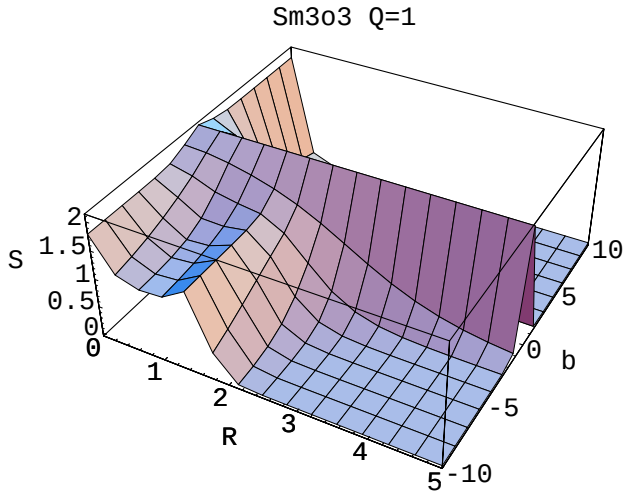

Figure 8.9: Third order entropy S_M for the 3 spin case.

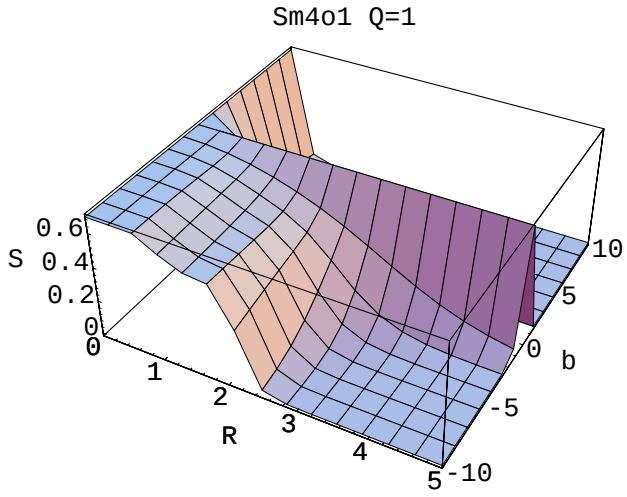

Figure 8.10: First order entropy S_M for the 4 spin case.

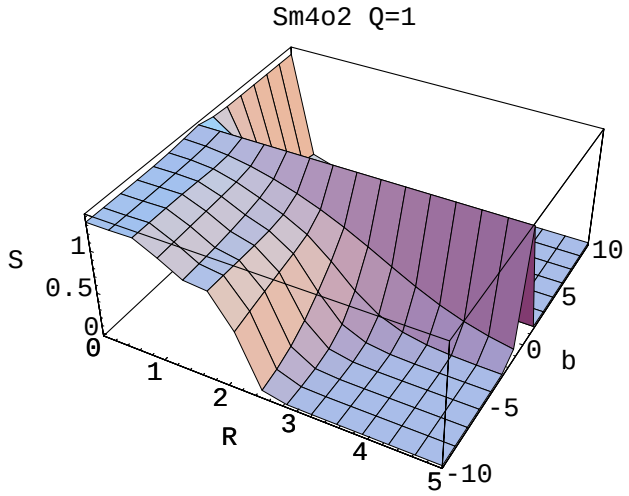

Figure 8.11: Second order entropy S_M for the 4 spin case.

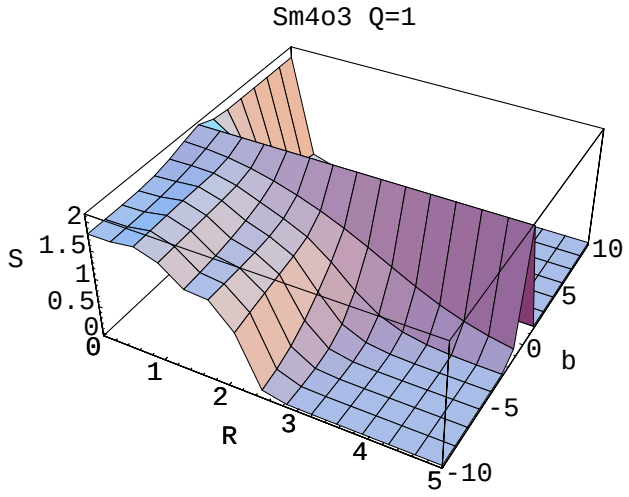

Figure 8.12: Third order entropy S_M for the 4 spin case.

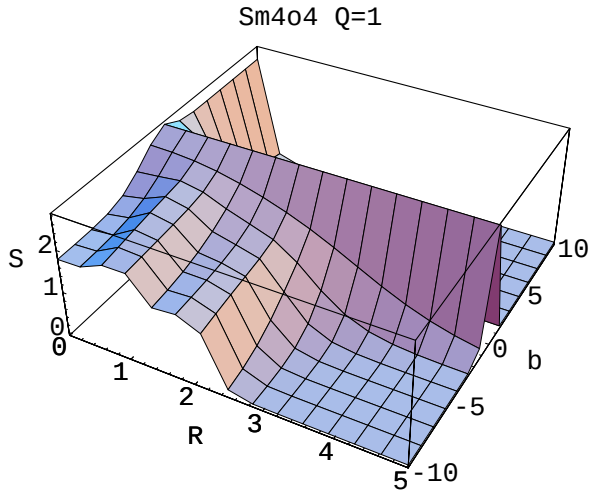

Figure 8.13: Fourth order entropy S_M for the 4 spin case.

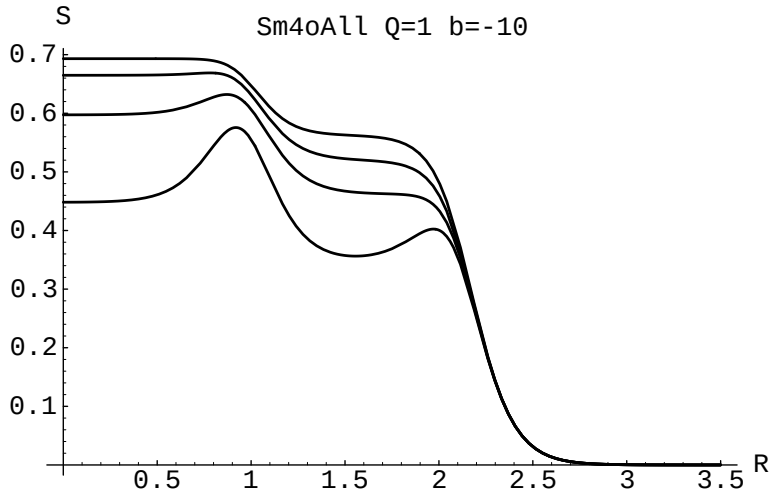

Figure 8.14: Entropies per spin for the 4 spin case of orders one thru four. Note the two entropy maxima, occurring at transitions between dominant energy eigenstates. The top graph is the first order entropy, the bottom is the fourth order entropy. Note that the full transition structure does not make itself apparent until the higher orders have been explored. Note also the ordering of the entropies per degree of freedom here, consistent with the entropy reduction per degree of freedom theorem.

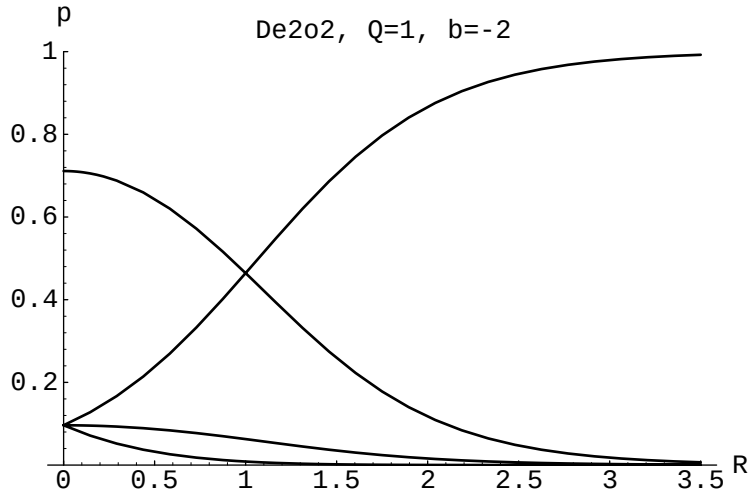

Figure 8.15: Probabilities of the two spin energy eigenstates. Note the transition between the eigenstates occurring at $R = 1$.

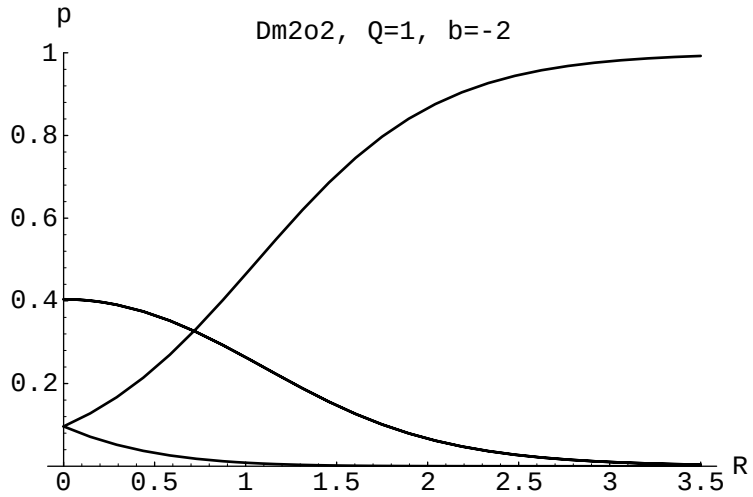

Figure 8.16: Probabilities of the two spin z-comonent states. Note the transition between the eigenstates occurring at $R = 1$. Two of the eigenstates are degenerate.

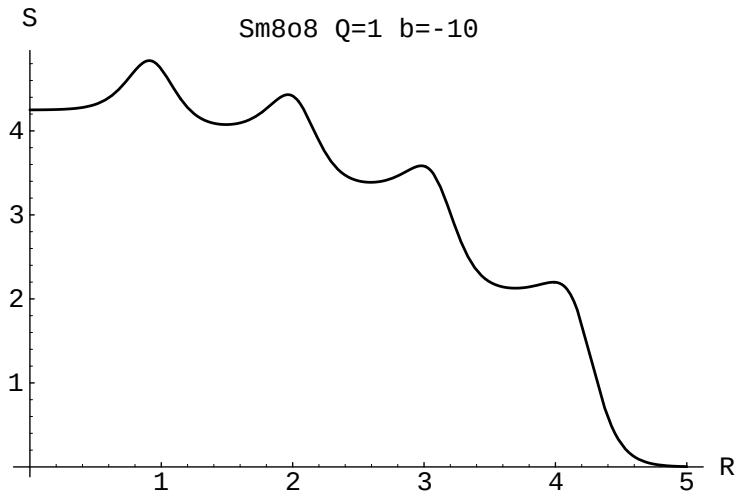

Figure 8.17: Cross section of the eighth order entropy of eight Heisenberg spins. Note that there are eight regions where the slope of the region has a sign different from the neighboring regions.

carrying capacity of the system at these fields is greater than it is at others. These indications of complex state behavior open the the possibility that significant applications of these systems will arise. Systems like these might serve as a way to prepare information for another system or an interface to another system, perhaps a system performing quantum computations directly based on the states of magnetic clusters like these. In section 8.12 we will see that the entropy changes are associated with changes in the magnetic moment of the cluster.

Finally, the amount of structure in the spin system of order k is reflected in the fact that there are k regions where the slope has a sign different from the neighboring regions, as seen in the graph of the eighth order entropy for eight spins, figure 8.17, i.e. there are $k/2$ bumps.

8.11 Information correlation functions

The information correlation functions C_k, according to the interpretation of section 3.13, may be seen as the information between these spins. The graph of figure 8.18 for the two spin system, figures 8.19–8.20 for the three spin system, and figures 8.21–8.23 for the four spin system indicate regions of temperature and external field where the distribution of subsets of spins in these systems carries new information not contained in smaller subsets, and where the subset distributions are redundant to smaller subset distributions. (Note that the logarithms are base e, and that β appears as b in the axis label.) Comparing figures 8.21–8.23, it is clear that the full system distribution is needed to describe the low temperature and small field antiferromagnetic regions. In these regions at each order lower than the number of spins there is structure having to do with the transitions that cannot be captured by still lower order distributions. The ferromagnetic side of the graphs appears far simpler, with almost all of the structure being captured by the lower order distributions (the first order information correlation is the first order negentropy). In a single system there are thus two regions where the behavior of the information correlation functions is vastly different, and where the complexity of the description needed is therefore vastly different.

8.12 Moments, correlations, cumulants

The graphs for the moments, correlation functions, and cumulants for the four spin system are given in figures 8.24–8.31. The moments show transitions at the same values of the external field that gave rise to entropy maxima. The moment graphs indicate that the antiferromegnatic region is where all of the interesting behavior occurs. The stairstep moment increase to spin alignment in the antiferromagnetic region as the external field is increased (recall that negative temperature is being used to create a negative coupling constant, and

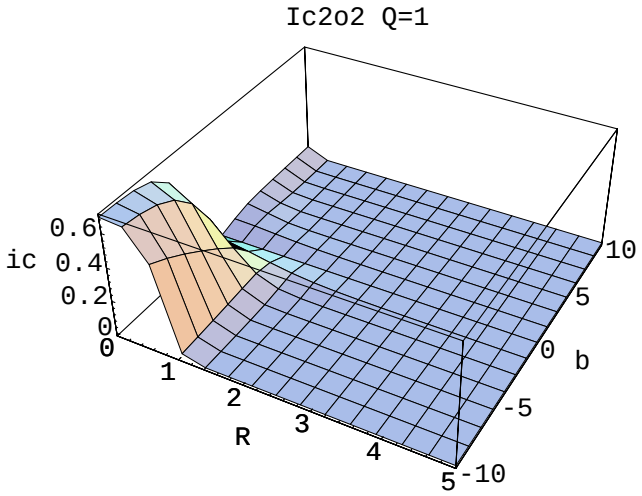

Figure 8.18: Second order information correlation for the 2 spin case.

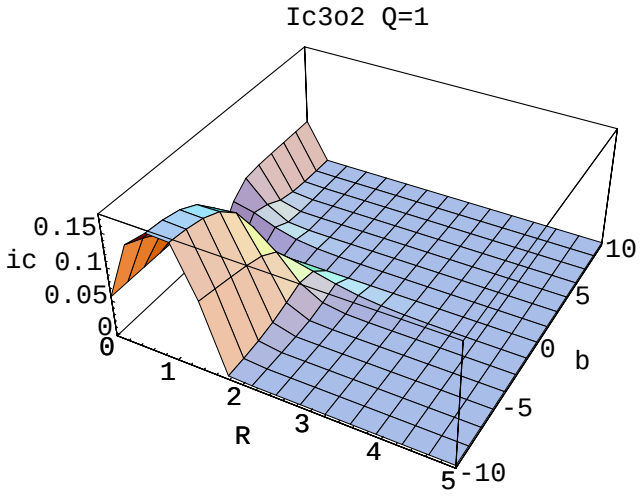

Figure 8.19: Second order information correlation for the 3 spin case.

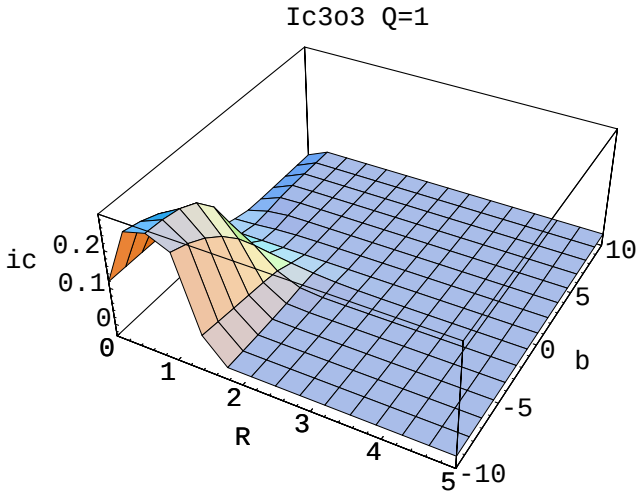

Figure 8.20: Third order information correlation for the 3 spin case.

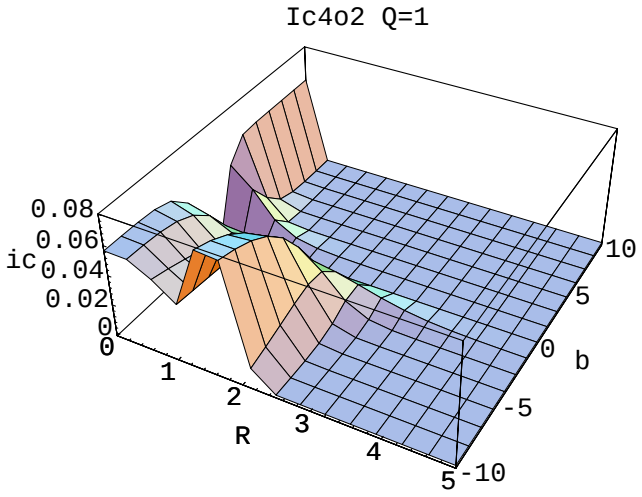

Figure 8.21: Second order information correlation for the 4 spin case.

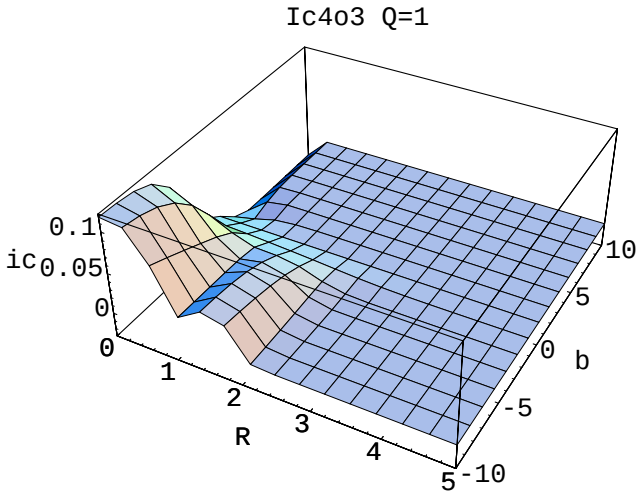

Figure 8.22: Third order information correlation for the 4 spin case.

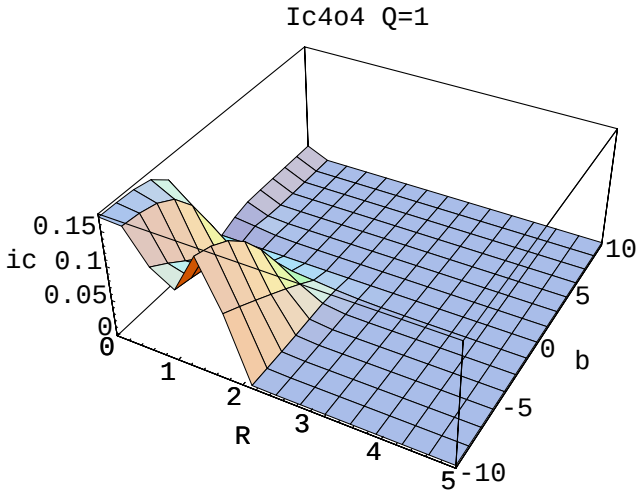

Figure 8.23: Fourth order information correlation for the 4 spin case.

that this affects the alignment in that it makes anti-alignment with the external field more favorable; the external field has thus also effectively changed in sign.) indicates that each transition is where a subset of the energy eigenstates with increasingly aligned spins becomes dominant. The plots of the correlations and cumulants (the first three cumulant functions and correlation functions are equal by order) indicate much of the same interesting behavior as the information correlation functions. The information they convey is about average values of products of deviations, and so they have the ability to indicate this aspect of the complex structure of the antiferromagnetic region. In the ferromagnetic region the third order correlation appears to dip in the region of small external field and high temperature. This region is not brought sharply to attention by any structure in the information correlation functions in this region.

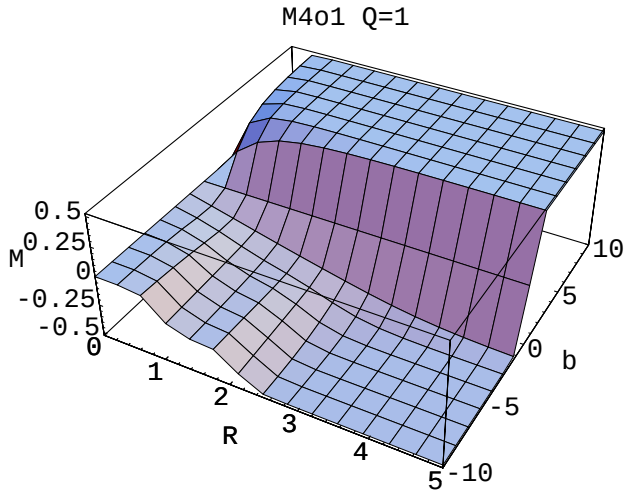

Figure 8.24: First order moment for the 4 spin case.

8.13 Mathematica for the symmetric spin Hamiltonian

See appendix B for the Mathematica code used to solve the quantum Heisenberg coupled spin model presented.

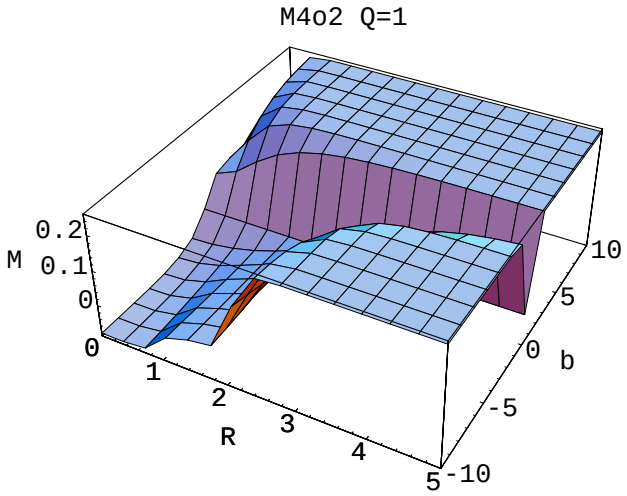

Figure 8.25: Second order moment for the 4 spin case.

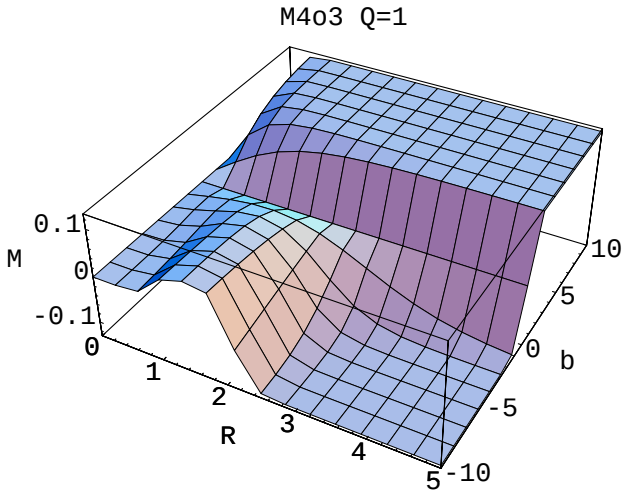

Figure 8.26: Third order moment for the 4 spin case.

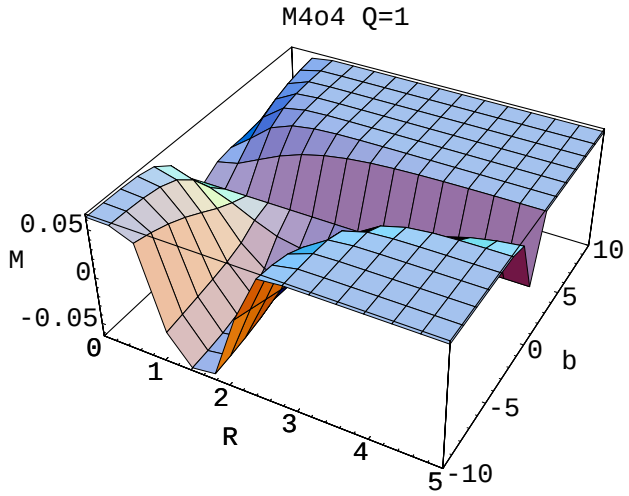

Figure 8.27: Fourth order moment for the 4 spin case.

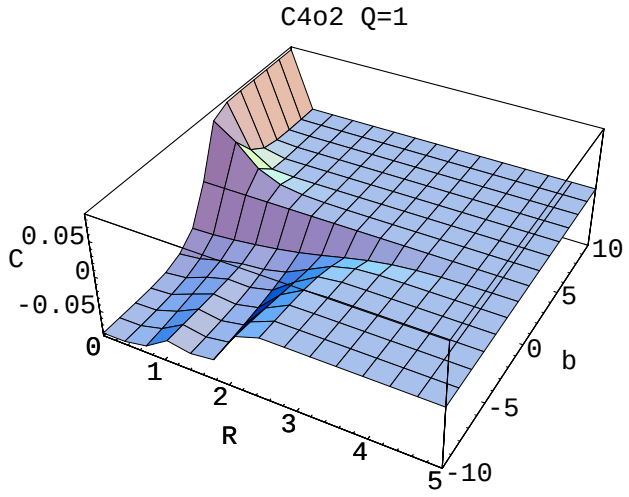

Figure 8.28: Second order correlation for the 4 spin case.

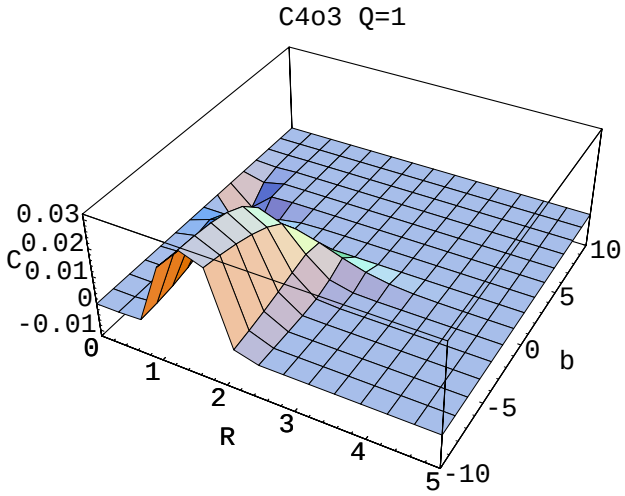

Figure 8.29: Third order correlation for the 4 spin case.

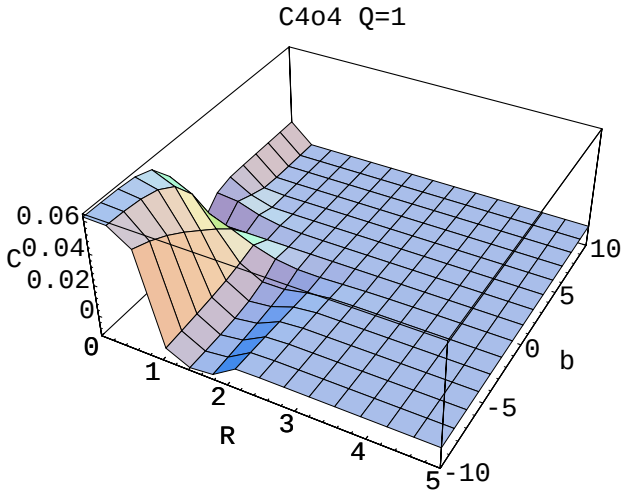

Figure 8.30: Fourth order correlation for the 4 spin case.

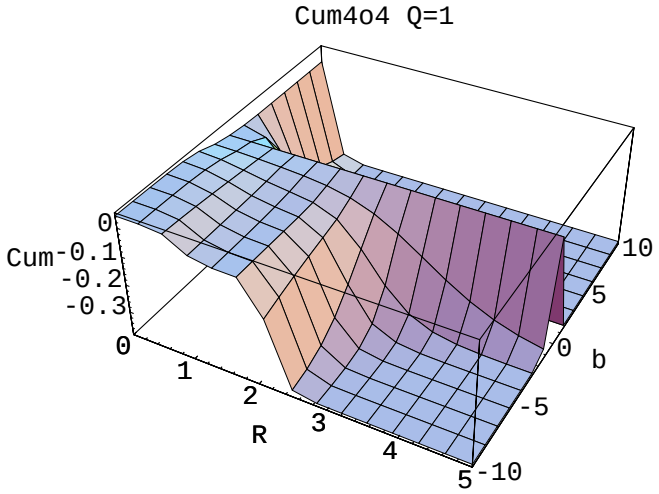

Figure 8.31: Fourth order cumulant for the 4 spin case. Recall that the first three correlations and cumulants are equal by order.

Chapter 9

Estimating information and correlation from finite data

9.1 Inferring unknown parameters from data

What is given in general is data, not the true data generating system, nor the true parameters for the data generating system. What is usually somewhat understood is the class of possible systems for the data generating system, and less well understood usually is what the values of the parameters are for any particular data generating system. For instance, the system generating the data may be known to give gaussian distributed data. But the mean and width parameters of the gaussian are unknown. The data does not determine these parameters, but it indicates them, and if we proceed intelligently we may infer them. With a sufficient amount of data the parameters are determined as well as is possible. In this case (large data) almost any method for estimating the parameters will suffice. What is crucial is the method used to estimate the parameters when the data does not strongly determine the parameters. For instance, how would you estimate the probability that the next bit will be a one or a zero after seeing only three examples of bits generated by a random bit generator which could be generating ones with any probability $0 \leq p \leq 1$? How would you estimate the uncertainty of the value of p guessed?

The nuts and bolts of estimating something unknown from data lies within Bayes' theorem. Bayes' theorem is simply two ways of writing the joint

probability distribution of data and parameters, but most importantly in our application it relates the probability of the *unknown* parameters to the *known* data. Bayesian estimation methodology differs widely from likelihood based methods in that it is this distribution of parameters given data, $P(\boldsymbol{x} \mid \boldsymbol{d})$, that is used in making guesses about the parameters, rather than the likelihood $P(\boldsymbol{d} \mid \boldsymbol{x})$. The likelihood is simply what we use to generate simulated data *after* the unknown parameters are guessed. Bayes' theorem may be written

$$P(\boldsymbol{x} \mid \boldsymbol{d}) = \frac{P(\boldsymbol{d} \mid \boldsymbol{x})P(\boldsymbol{x})}{P(\boldsymbol{d})} \tag{9.1}$$

where $P(\boldsymbol{x})$ is called the *prior* probability of the parameters $\boldsymbol{x}$. In all Bayesian methods it is the choice of prior that reflects everything that was known about the distribution of the parameters before the data was seen. For any given data set, $P(\boldsymbol{d})$ is a constant and can be found by summing over the parameters as in $P(\boldsymbol{d}) = \int P(\boldsymbol{d} \mid \boldsymbol{x})\, P(\boldsymbol{x})\ dx$.

Referring back to likelihood methods mentioned before, if you have a demon who is generating data using a known type of generating system but with the parameters for the system unknown to you, and the demon states the probability distribution from which it chose the parameters for the data generating system, then the Bayesian methodology is provably optimal when the prior is taken to be the demon's parameter choice distribution. Likelihood methods will tend to be too guided by the data, overfitting the parameters to match the data too closely, and they are provably not optimal in general.

In the next section the question of whether data always provides an increasing amount of information about the unknown parameters is answered. The answer is no. Sometimes more data leads to more confusion. However, on the average more data does provide more information about the unknown parameters. What is meant here by information, confusion, and average is quantified, and the theorem showing that on the average information about the unknown parameters increases upon seeing data is proved.

Later sections apply Bayesian methodology to the task at hand of estimating entropy, information correlation functions, chi-squared, moments, and correlations, from finite data. Finite data really is the only interesting case in problems of inference. Large data allows the use of almost any random provably suboptimal methodology to find good estimates. Developing methods of inference that make use of all available information is the elegant side of statistical inference. Practically, data is expensive. Thus the large data case is ignored.

9.2 Inference increases information about the true distribution on average

Let the data be given one sample point at a time. The sample data are thus ordered and will be indicated by $\boldsymbol{d}_n = (d_1, \ldots d_n)$. After each new sample is given an inference of the unknown parameters using Bayes' theorem, taking the form of the distribution of these parameters, will be made. The information in the inferred parameter distribution will be found, and surprisingly it will be shown that this information does not generally increase upon seeing new data. However, when this inferred information is averaged over possible data samples, the average information about the unknown parameters does increase when new data is seen. Thus, on average, more is learned about the unknown parameters when additional data is presented, but additional data can sometimes lead to more confusion.

The measure of the uncertainty of (and confusion about) the unknown parameters is the entropy of the distribution of the unknown parameters. Working in the probability density framework the entropy of the distribution of parameters after seeing n data samples $\boldsymbol{d}_n$ is given by

$$S_n(\boldsymbol{d}_n) = -\int P(\boldsymbol{x} \mid \boldsymbol{d}_n)\ log(P(\boldsymbol{x} \mid \boldsymbol{d}_n))\ d\boldsymbol{x} \tag{9.2}$$

The change in the uncertainty of the parameters upon seeing the nth data sample is then given by $\Delta S_n(\boldsymbol{d}_n) := S_n(\boldsymbol{d}_n) - S_{n-1}(\boldsymbol{d}_{n-1})$. At first it might be thought that this uncertainty should always decrease. However, this is not the case. Suppose, for example, that the mean of a gaussian distribution is to be inferred, while the width of the gaussian is known. Suppose further that $n = 2$, and that d_1 and d_2 happen to lie very far apart from each other, which can happen by chance for gaussian distributed data. In this case, $P(\boldsymbol{x} \mid \boldsymbol{d}_1)$ is going to be more sharply peaked than $P(\boldsymbol{x} \mid \boldsymbol{d}_2)$. This is because the $\boldsymbol{d}_1 = (d_1)$ inference puts density at d_1, while the two sample $\boldsymbol{d}_2 = (d_1, d_2)$ inference puts density at both of the data locations, which are far apart. In fact, taking the uniform distribution for the parameter prior, the one sample inference is a single gaussian bump centered on the data, while, if the two data samples are very far apart, the two sample inference is two identical gaussian bumps (having half the height and the same width as the one sample inference gaussian bump), and the entropy of the two sample inference will be one bit greater than that of the one sample inference. Sometimes new data leads to increased confusion about the parameters.

Now, consider what happens when the average over data sets $\boldsymbol{d}_n$ of the change in uncertainty (confusion) is taken. The average of interest is the average change in confusion when a new data sample is seen, given by

$$\langle \Delta S_n \rangle := \int \Delta S_n(\boldsymbol{d}_n)\, P(\boldsymbol{d}_n)\; d\boldsymbol{d}_n \tag{9.3}$$

The next step is to show that the average change in the confusion about the parameters is negative. To do this, note that for any function $f(\boldsymbol{x}, \boldsymbol{d}_{n-1})$

$$\begin{aligned}
&\int f(\boldsymbol{x}, \boldsymbol{d}_{n-1})\; P(\boldsymbol{x} \mid \boldsymbol{d}_{n-1})\; P(\boldsymbol{d}_n)\; d\boldsymbol{d}_n \\
&= \int f(\boldsymbol{x}, \boldsymbol{d}_{n-1})\; P(\boldsymbol{x} \mid \boldsymbol{d}_{n-1})\; P(\boldsymbol{d}_{n-1})\; d\boldsymbol{d}_{n-1} \\
&= \int f(\boldsymbol{x}, \boldsymbol{d}_{n-1})\; P(\boldsymbol{x}, \boldsymbol{d}_{n-1})\; d\boldsymbol{d}_{n-1} \\
&= \int f(\boldsymbol{x}, \boldsymbol{d}_{n-1})\; P(\boldsymbol{x}, \boldsymbol{d}_n)\; d\boldsymbol{d}_n
\end{aligned} \tag{9.4}$$

Now, expand the average change in the confusion written in equation9.3 as

$$\begin{aligned}\langle \Delta S_n \rangle &= -\int\int P(\boldsymbol{x} \mid \boldsymbol{d}_n)\ log(P(\boldsymbol{x} \mid \boldsymbol{d}_n))\ P(\boldsymbol{d}_n)\, d\boldsymbol{d}_n d\boldsymbol{x} \\ &\quad + \int\int P(\boldsymbol{x} \mid \boldsymbol{d}_{n-1})\ log(P(\boldsymbol{x} \mid \boldsymbol{d}_{n-1}))\ P(\boldsymbol{d}_n)\, d\boldsymbol{d}_n d\boldsymbol{x} \qquad (9.5)\end{aligned}$$

and simplify the inner integral of the second term on the right side using the identity proven in equation 9.4 to find

$$\begin{aligned}\langle \Delta S_n \rangle &= -\int\int P(\boldsymbol{x}, \boldsymbol{d}_n)\ log(P(\boldsymbol{x} \mid \boldsymbol{d}_n))\, d\boldsymbol{d}_n d\boldsymbol{x} \\ &\quad + \int\int P(\boldsymbol{x}, \boldsymbol{d}_n)\ log(P(\boldsymbol{x} \mid \boldsymbol{d}_{n-1}))\ d\boldsymbol{d}_n d\boldsymbol{x} \qquad (9.6)\end{aligned}$$

Collect the logarithms and note that the integral over $\boldsymbol{x}$ is a Kullback-Leibler distance, and allows us to apply the information identity (for probability densities $p_1(x)$ and $p_2(x)$, $KL(p_1, p_2) := \int p_1 log(p_1/p_2) dx \geq 0$) to find that the average change in the confusion about the parameters is negative

$$\begin{aligned}\langle \Delta S_n \rangle &= -\int\int P(\boldsymbol{x}, \boldsymbol{d}_n)\ log(\frac{P(\boldsymbol{x} \mid \boldsymbol{d}_n)}{P(\boldsymbol{x} \mid \boldsymbol{d}_{n-1})}) d\boldsymbol{x}\, d\boldsymbol{d}_n \\ &= -\int P(\boldsymbol{d}_n)\ \left(\int P(\boldsymbol{x} \mid \boldsymbol{d}_n)\ log(\frac{P(\boldsymbol{x} \mid \boldsymbol{d}_n)}{P(\boldsymbol{x} \mid \boldsymbol{d}_{n-1})}) d\boldsymbol{x} \right) d\boldsymbol{d}_n \\ &\leq 0 \qquad (9.7)\end{aligned}$$

Because negative uncertainty is information, a negative change in confusion corresponds to a positive change in information. Thus we have proven the following theorems:

Theorem: Information increases on the average. *Although in particular data the information about the parameters may decrease upon seeing a new data sample, on the average the information about the parameters increases upon seeing a new data sample.*

Theorem: Average information increase is the Kullback-Leibler distance. *The average increase in the information about the parameters is the average of the Kullback-Leibler distance between the parameter distributions conditioned on the data after and before the new sample is seen.*

9.3 Estimating functions of probability distributions from finite samples.

9.3.1 Statement of the problem solved here

Consider a system with m possible states and an associated m-vector of probabilities of those states, $\boldsymbol{p} = (p_i)$, $1 \leq i \leq m$, $(\sum_{i=1}^{m} p_i = 1)$. The system is repeatedly and independently sampled according to the distribution $\boldsymbol{p}$. Let the total number of samples be N and denote the associated vector of counts of states by $\boldsymbol{n} = (n_i)$, $1 \leq i \leq m$, $(\sum_{i=1}^{m} n_i = N)$. The problem is to estimate a given function $Q(\boldsymbol{p})$ from $\boldsymbol{n}$, the samples. The functions considered are the entropy, mutual information, moments, average, variance, covariance and other correlations, and chi-squared.

Some previous work on estimating $Q(\boldsymbol{p})$ from $\boldsymbol{n}$, using frequency methods to generate correction terms, appears in [4, 21, 22, 31, 32, 33, 34, 51, 52, 61, 77].

Fully formal justifications of the manipulations carried out in this paper can be found as appendices 9.4–9.10.

9.3.2 Estimating from finite data is ubiquitous

The problem of estimating a function of an unknown distribution from finite samples of that distribution is ubiquitous in physics, arising for example in dimension estimation and in estimating correlations from data.

For example, in information dimension estimation, [31, 63] we imagine a discretization of a space containing an attractor. The attractor constitutes a probability density function across the space, and therefore a probability distribution across the bins of the discretization. We are interested in how the Renyi entropy of the distribution across the bins changes as the discretization changes. This behavior gives us the information dimension of the attractor,

which is useful in non-linear time-series analysis, in connection with estimating the embedding dimension, see [21, 63]. It turns out that to accurately measure the information dimension we would like to make accurate estimates of the Renyi entropy for as wide a range of granularities of the discretization as possible. In particular, we would like to make accurate estimates when the discretization is quite fine. In such a regime, the number of counts per bin - i.e., the values n_i - will be quite small. Accordingly, we are unavoidably faced with the small sample statistics problem of how to meaningfully perform inference with small samples. This is precisely the regime in which Bayesian techniques excel.

9.3.3 The distribution of data

By the assumptions of section 9.3.1 the data is multinomially distributed as

$$P(\boldsymbol{n} \mid \boldsymbol{p}) = \binom{N}{\boldsymbol{n}} \prod_{i=1}^{m} p_i^{n_i} \tag{9.8}$$

9.3.4 The Bayes' estimator

The Bayes' estimator of $Q(\boldsymbol{p})$ will be denoted $Q(\boldsymbol{n})$ and taken as the posterior average

$$Q(\boldsymbol{n}) = \int d\boldsymbol{p}\ P(\boldsymbol{p} \mid \boldsymbol{n})\ Q(\boldsymbol{p}) \tag{9.9}$$

The reason for this is explained in the next section.

Note that it may also be of interest to compute the distribution of the estimator. This is given by

$$P(Q(\boldsymbol{p}) = q \mid \boldsymbol{n}) = \int d\boldsymbol{p}\ \delta(Q(\boldsymbol{p}) - q)\ P(\boldsymbol{p} \mid \boldsymbol{n}) \tag{9.10}$$

In most cases here only moments of Q are considered. When moments are considered the notation $Q^{(k)}(\boldsymbol{n})$ will denote the estimator of $Q(\boldsymbol{p})^k$.

9.3.5 The Bayes' estimator minimizes posterior mean square error

If $\boldsymbol{p}$ is known, then the estimator of $Q(\boldsymbol{p})$ that minimizes the mean square error (data average) is directly $Q(\boldsymbol{p})$, regardless of the data. More generally, when $\boldsymbol{p}$ is not fixed then the mean-squared error is given by

$$\int d\boldsymbol{p} P(\boldsymbol{p}) \sum_{\boldsymbol{n}} P(\boldsymbol{n} \mid \boldsymbol{p}) \ (Q(\boldsymbol{n}) - Q(\boldsymbol{p}))^2 \tag{9.11}$$

Varying with respect to $Q(\boldsymbol{n})$ and applying Bayes' theorem (see equation 9.1) gives the result that the estimator is the posterior average of the function being estimated [18], equation 9.9. Thus, assuming that the prior for $\boldsymbol{p}$ is known, the posterior average of $Q(\boldsymbol{p})$ is the average mean square error optimal estimator. Note that in general other error measures than mean square error may be considered, and these lead to different estimators.

9.3.6 Form of the prior

The priors considered here are sums of terms of the form

$$P(\boldsymbol{p}) \propto \prod_{i=1}^{m} p_i^{k_i} \Delta(\boldsymbol{p}) \Theta(\boldsymbol{p}) \tag{9.12}$$

where $\Delta(\boldsymbol{p}) = \delta(\sum_{i=1}^{m} p_i - 1)$ and $\Theta(\boldsymbol{p}) = \prod_{i=1}^{m} \theta(p_i)$ with $\theta(x) = 1$ for $x \geq 0$ and zero otherwise. The proportionality constant is set by normalization of $P(\boldsymbol{p})$.

It should be noted that a prior of the form above may be used to approximate any prior of several variables because the polynomials form a complete set of basis functions for function approximation.

9.3.7 Form of the integrals giving the estimator

From equation 9.9 and by Bayes' theorem, equation 9.1, the estimator may be written as the ratio of integrals

$$Q(\boldsymbol{n}) = \frac{\int d\boldsymbol{p} P(\boldsymbol{n} \mid \boldsymbol{p}) P(\boldsymbol{p}) \, Q(\boldsymbol{p})}{\int d\boldsymbol{p} P(\boldsymbol{n} \mid \boldsymbol{p}) P(\boldsymbol{p})} \tag{9.13}$$

and considering a prior of the form in equation 9.12 and cancelling proportionality constants this may then be written as

$$Q(\boldsymbol{n}) = \frac{\int d\boldsymbol{p} \prod_{i=1}^{m} p_i^{n_i+k_i} \Delta(\boldsymbol{p}) \Theta(\boldsymbol{p}) Q(\boldsymbol{p})}{\int d\boldsymbol{p} \prod_{i=1}^{m} p_i^{n_i+k_i} \Delta(\boldsymbol{p}) \Theta(\boldsymbol{p})} \tag{9.14}$$

Clear from this expression is that the exponents $\boldsymbol{k}$ of the prior are simply added to $\boldsymbol{n}$; from now on they will be ignored and may be replaced at the end of the calculation by substituting $\boldsymbol{n} + \boldsymbol{k}$ for $\boldsymbol{n}$ within the range of validity of the expression. Thus, defining

$$I[Q(\boldsymbol{p}), \boldsymbol{n}] = \int d\boldsymbol{p} \prod_{i=1}^{m} p_i^{n_i} \Delta(\boldsymbol{p}) \Theta(\boldsymbol{p}) Q(\boldsymbol{p}) \tag{9.15}$$

the estimator is given by the ratio

$$Q(\boldsymbol{n}) = \frac{I[Q(\boldsymbol{p}), \boldsymbol{n}]}{I[1, \boldsymbol{n}]} \tag{9.16}$$

When the prior is the sum of terms of the form in equation 9.12, the numerator and denominator both contain respective summations. In each term of each summation though, the respective $\boldsymbol{k}$ are simply added to $\boldsymbol{n}$.

9.3.8 Integrating

The results of this section are presented as a series of theorems leading quickly to the estimators of the moments of the entropy, presented in the next section. Note that the conditions of existence of the results are given in full generality in the complex domain, although the actual event counts are non-negative integers.

Define the Laplace convolution operator $\otimes$ by its operation on two functions

$$(f \otimes g)(\tau) := \int_0^{\tau} dx f(x) g(\tau - x) \tag{9.17}$$

Theorem 1. *If* $F(\boldsymbol{p}) = \prod_{i=1}^{m} f_i(p_i)$ *then*

$$\int d\boldsymbol{p} \Delta(\boldsymbol{p}) \Theta(\boldsymbol{p}) F(\boldsymbol{p}) = (\otimes_{i=1}^{m} f_i)(1) \tag{9.18}$$

Proof: Define $\tau_k = 1 - \sum_{i=1}^{k} p_i$. Define $F_k = \prod_{i=1}^{k} f_i(p_i)\theta(p_i)$. Define $d\boldsymbol{p}_k = dp_1 \ldots dp_k$. The integral in the statement of the theorem, after substituting for $\Delta(\boldsymbol{p})$ and $\Theta(\boldsymbol{p})$, becomes

$$\begin{aligned} &\int d\boldsymbol{p}_m \delta(\tau_m) F_m \\ &\quad = \int d\boldsymbol{p}_{m-1} \left(\int dp_m \delta(\tau_{m-1} - p_m)\theta(p_m) f_m(p_m) \right) F_{m-1} \end{aligned} \tag{9.19}$$

Integrating over p_m the integral takes the form

$$\begin{aligned} &\int d\boldsymbol{p}_{m-1} \theta(\tau_{m-1}) f_m(\tau_{m-1}) F_{m-1} \\ &\quad = \int d\boldsymbol{p}_{m-2} \left(\int dp_{m-1} \theta(\tau_{m-2} - p_{m-1}) f_m(\tau_{m-2} - p_{m-1}) \right. \\ &\qquad \left. \times \theta(p_{m-1}) f_{m-1}(p_{m-1}) \right) F_{m-2} \end{aligned} \tag{9.20}$$

The inner integral may be written as a convolution, and noting that τ_{m-2} must be positive we find

$$\begin{aligned} &\int d\boldsymbol{p}_{m-1} \theta(\tau_{m-1}) f_m(\tau_{m-1}) F_{m-1} = \\ &\quad = \int d\boldsymbol{p}_{m-2} \theta(\tau_{m-2}) (f_m \otimes f_{m-1})(\tau_{m-2}) F_{m-2} \end{aligned} \tag{9.21}$$

Clearly, this can be extended by induction and the associativity of the convolution operator to give the desired result. QED.

For the following, the Laplace transform operator L and its inverse are discussed in appendix 9.6.3.

Theorem 2. (Laplace convolution) *If $L[f_i]$ exists for $i = 1, \ldots, m$, then*

$$L[\otimes_{i=1}^{m} f_i] = \prod_{i=1}^{m} L[f_i] \tag{9.22}$$

Proof: The proof is lengthy, but may be found in [36]. The result is mentioned in [50, 54].

Define the Gamma function $\Gamma(z)$, $Re(z) \geq -1$ by

$$\Gamma(z) = \int_0^\infty t^{z-1} e^{-t} dt \tag{9.23}$$

Theorem 3. *If* $Re(n_i) \geq -1$, $i = 1, \ldots, m$, *then*

$$I[1, \boldsymbol{n}] = \frac{\prod_{i=1}^m \Gamma(n_i + 1)}{\Gamma(N + m)} \tag{9.24}$$

Proof: By theorem 1

$$I[1, \boldsymbol{n}] = (\otimes_{i=1}^m p_i^{n_i})(1) \tag{9.25}$$

By theorem 2

$$(\otimes_{i=1}^m p_i^{n_i})(\tau) = L^{-1}[\prod_{i=1}^m L[p^{n_i}](s)](\tau) \tag{9.26}$$

Note that

$$L[p^n](s) = \frac{\Gamma(n+1)}{s^{n+1}} \ , \ Re(n) > -1 \tag{9.27}$$

Substitute from equation 9.27 in the right side of equation 9.26. Finally use equation 9.27 again to take the inverse. Setting $\tau = 1$ finishes the proof. QED.

Theorem 4. *If* $Re(n_i) > -1$, $i = 1, \ldots, m$, *then*

$$I[\prod_{i=1}^m log(p_i)^{r_i}, \boldsymbol{n}] = \left(\prod_{i=1}^m \partial_{n_i}^{r_i} \right) I[1, \boldsymbol{n}] \tag{9.28}$$

Proof: Note that $\partial_n p^n = p^n log(p)$. The interchange of derivative and integral is given in appendix 9.7.

In using theorem 4 we will take the derivatives in the expression on the right side of equation 9.28 where N appears in $I[1, \boldsymbol{n}]$ (see equation 9.24); thus note that since $N = \sum_{i=1}^m n_i$, we have $\partial_{n_i} N = 1$.

Define the polygamma function

$$\Psi^{(n)}(z) = \partial_z^{n+1} log(\Gamma(z)) \tag{9.29}$$

(see [3] for properties of the polygamma function), and define

$$\Phi^{(n)}(z) = \Psi^{(n-1)}(z) = \partial_z^n log(\Gamma(z)) \tag{9.30}$$

Introduce

$$\Delta\Phi^{(n)}(z_1, z_2) = \Phi^{(n)}(z_1) - \Phi^{(n)}(z_2) \tag{9.31}$$

Theorem 5. *For* $Re(n_i) > -1$, $i = 1, \ldots, m$,

$$I[log(p_u), \boldsymbol{n}] = \frac{\prod_{i=1}^m \Gamma(n_i + 1)}{\Gamma(N + m)} \Delta\Phi^{(1)}(n_u + 1, N + m) \tag{9.32}$$

Proof: By theorems 3 and 4

$$\begin{aligned} I[log(p_u), \boldsymbol{n}] &= \partial_{n_u} I[1, \boldsymbol{n}] \\ &= \partial_{n_u} \frac{\prod_{i=1}^m \Gamma(n_i + 1)}{\Gamma(N + m)} \end{aligned} \tag{9.33}$$

Taking the derivative and substituting for the Φ function yields the result. QED.

Theorem 6. *For* $Re(n_i) > -1$, $i = 1, \ldots, m$,

(6.1)

$$\begin{aligned} I[log(p_u) log(p_v), \boldsymbol{n}] &= \frac{\prod_{i=1}^m \Gamma(n_i + 1)}{\Gamma(N + m)} \times \\ &\{\Delta\Phi^{(1)}(n_u + 1, N + m)\Delta\Phi^{(1)}(n_v + 1, N + m) - \Phi^{(2)}(N + m)\} \\ u &\neq v \end{aligned} \tag{9.34}$$

(6.2)

$$\begin{aligned} I[log(p_u)^2, \boldsymbol{n}] &= \frac{\prod_{i=1}^m \Gamma(n_i + 1)}{\Gamma(N + m)} \times \\ &\{\Delta\Phi^{(1)}(n_u + 1, N + m)^2 - \Delta\Phi^{(2)}(n_u + 1, N + m)\} \end{aligned} \tag{9.35}$$

Proof: Apply theorems 3 and 4 in a manner similar to the proof of theorem 5. Here two derivatives are needed.

9.3.9 Estimators for the first and second moments of the entropy

Define $\boldsymbol{e}_j = (\delta(i,j))$, the vector of all zeroes, except for a single one in the jth position.

Theorem 7. *For* $Re(n_i) > -1$, $i = 1, \ldots, m$,

$$S^{(1)}(\boldsymbol{n}) = -\sum_{i=1}^{m} \frac{n_i + 1}{N+m} \Delta\Phi^{(1)}(n_i + 2, N + m + 1) \tag{9.36}$$

Proof: Applying theorems 3, 4, and 5,

$$\begin{aligned} S^{(1)}(\boldsymbol{n}) &= \frac{I[-\sum_{i=1}^{m} p_i log(p_i), \boldsymbol{n}]}{I[1, \boldsymbol{n}]} \\ &= -\sum_{i=1}^{m} \frac{I[log(p_i), \boldsymbol{n} + \boldsymbol{e}_i]}{I[1, \boldsymbol{n}]} \\ &= -\sum_{i=1}^{m} \frac{\partial_{n_i} I[1, \boldsymbol{n} + \boldsymbol{e}_i]}{I[1, \boldsymbol{n}]} \end{aligned}$$

Theorem 8. *For* $Re(n_i) > -1$, $i = 1, \ldots, m$,

$$\begin{aligned} S^{(2)}(\boldsymbol{n}) &= \sum_{i \neq j=1}^{m} \frac{(n_i+1)(n_j+1)}{(N+m)(N+m+1)} \times \\ &\quad \{\Delta\Phi^{(1)}(n_i + 2, N + m + 2)\Delta\Phi^{(1)}(n_j + 2, N + m + 2) \\ &\quad -\Phi^{(2)}(N + m + 2)\} \\ &+ \sum_{i=1}^{m} \frac{(n_i+1)(n_i+2)}{(N+m)(N+m+1)} \times \\ &\quad \{\Delta\Phi^{(1)}(n_i + 3, N + m + 2)^2 + \Delta\Phi^{(2)}(n_i + 3, N + m + 2)\} \end{aligned} \tag{9.37}$$

Proof: Similar to proof of theorem 7 noting that

$$S^{(2)}(\boldsymbol{n}) = \frac{\sum_{i,j=1}^{m} \partial_{n_i}\partial_{n_j} I[1, \boldsymbol{n} + \boldsymbol{e}_i + \boldsymbol{e}_j]}{I[1, \boldsymbol{n}]} \tag{9.38}$$

In a similar manner, all higher moments of the entropy may be calculated by differentiation, leading to the observation that the posterior moment

generating function of the entropy is given by

$$mgf[S](x) = \int d\boldsymbol{p}\ P(\boldsymbol{p} \mid \boldsymbol{n})\ e^{S(\boldsymbol{p})x} \tag{9.39}$$

$$= \sum_{k=0}^{\infty} \frac{\left[\Pi_{i=i}^{k}(\sum_{\alpha_i=1}^{m} \partial_{n_{\alpha_i}})\right] I[1, \boldsymbol{n} + \sum_{i=i}^{k} \boldsymbol{e}_{\alpha_i}]}{I[1, \boldsymbol{n}]} \frac{(-x)^k}{k!} \tag{9.40}$$

Note that for $\boldsymbol{n} = 0$ we have $S^{(1)}(\boldsymbol{n}) = -\Delta\Phi^{(1)}(2, m+1) = \sum_{i=1}^{m} i^{-1} - 1$. Also note that $\lim_{N->\infty} S^{(1)}(\boldsymbol{n}) = -\sum_{i=1}^{m}(n_i/N)log(n_i/N)$, the frequency counts estimator.

9.3.10 Entropy estimator comparison

The graphs appearing in figures 9.1–9.4 depict several comparisons of the Bayes' and frequency-counts estimators for entropy. In all cases the solid line represents the Bayes' estimator, the dash-dot line represents the frequency-counts estimator, and the dotted line represents the true value of the entropy, where applicable. Figure 9.5 depicts the pdf of the Bayes' estimator for a fixed ratio of counts as the number of counts increases. The graphs are the result of exact numerical computations of the various quantities represented.

Figure 9.1a is a demonstration of the fact that the Bayes' estimator is better than any other estimator in the least mean-square sense; in particular it is better than the frequency counts estimator. Figure 9.1a shows the mean-squared error of the Bayes' estimator and the frequency counts estimator when $m = 2$ and the prior is uniform, where the mean-squared error for any estimator $Q(\boldsymbol{n})$ is given by expression 9.11. As is immediately seen the Bayes' estimator has a smaller mean-squared error than the frequency-counts estimator for all N (consistent with section 9.3.5).

Figure 9.1b depicts the average sample variance, that is

$$\int d\boldsymbol{p}\ P(\boldsymbol{p}) \sum_{\boldsymbol{n}} P(\boldsymbol{n} \mid \boldsymbol{p}) \left(S(\boldsymbol{n}) - \sum_{\boldsymbol{n}'} P(\boldsymbol{n}' \mid \boldsymbol{p}) S(\boldsymbol{n}') \right)^2 \tag{9.41}$$

of the Bayes' and frequency counts estimators. For a particular sample size N, the Bayes' estimator has a smaller sample variance.

Figure 9.2 shows the sample averages of the estimators as functions of the sample size N for various values of fixed $\boldsymbol{p}$, that is

$$\sum_{\boldsymbol{n}} P(\boldsymbol{n} \mid \boldsymbol{p}) S(\boldsymbol{n}) \tag{9.42}$$

Figure 9.3 shows the same sample averages (equation 9.42) of the estimators, but now as functions of the true $\boldsymbol{p}$ for various values of the sample size N.

It is of interest to note that for a particular range of $\boldsymbol{p}$ values and sufficiently large N, the sample average of the frequency-counts estimator actually comes closer to the true entropy than does the sample average of the Bayes' estimator (see figures 9.2d–f and 9.3d–f]). To see how this is possible in light of the fact that the Bayes' estimator has lower mean-squared error, first note that

$$\begin{aligned} \int d\boldsymbol{p} P(\boldsymbol{p}) \sum_{\boldsymbol{n}} P(\boldsymbol{n} \mid \boldsymbol{p}) \ (S(\boldsymbol{n}) - S(\boldsymbol{p}))^2 = \\ \int d\boldsymbol{p} \ P(\boldsymbol{p}) \sum_{\boldsymbol{n}} P(\boldsymbol{n} \mid \boldsymbol{p}) \left(S(\boldsymbol{n}) - \sum_{\boldsymbol{n}'} P(\boldsymbol{n}' \mid \boldsymbol{p}) S(\boldsymbol{n}') \right)^2 \\ + \int d\boldsymbol{p} \ P(\boldsymbol{p}) \left(\sum_{\boldsymbol{n}} P(\boldsymbol{n} \mid \boldsymbol{p}) S(\boldsymbol{n}) - S(\boldsymbol{p}) \right)^2 \end{aligned} \tag{9.43}$$

so that the mean-squared error is the sum of the mean sample variance and the mean-squared bias. The left hand side of equation 9.43 is depicted in figure 9.1a. The first integral on the right hand side is depicted in figure 9.1b. The integrand of the last integral on the right hand side (excluding the prior) appears in figure 9.2 as the square of the difference between the curve for the estimator, and the value of $\boldsymbol{p}$ being estimated. This quantity favors the frequency-counts estimator for some values of $\boldsymbol{p}$ for sufficiently large N; however

the first integral on the right more than compensates to give a result favoring the Bayes' estimator.

Figure 9.4 depicts the sample averages of the estimators' square differences from true as a function of r for various sample sizes N,

$$\sum_{\boldsymbol{n}} P(\boldsymbol{n} \mid \boldsymbol{p})(S(\boldsymbol{n}) - S(\boldsymbol{p}))^2 \tag{9.44}$$

The integral of expression 9.44 multiplied by the prior (here uniform), depicted for various N, is shown in 9.1a.

Finally, figure 9.5 shows

$$P(S(\boldsymbol{p}) = s \mid \boldsymbol{n}) = \int d\boldsymbol{p}\ \delta(S(\boldsymbol{p}) - s) P(\boldsymbol{p} \mid \boldsymbol{n}) \tag{9.45}$$

for a fixed ratio $(1 : 15)$ of observed counts $n_1 : n_2$, as a function of the number of counts $N = n_1 + n_2$. Note the increasing density placed upon the entropy $S(1/16, 15/16)$ as the counts N increase. The average of s according to this density is the Bayes' estimator given the observations.

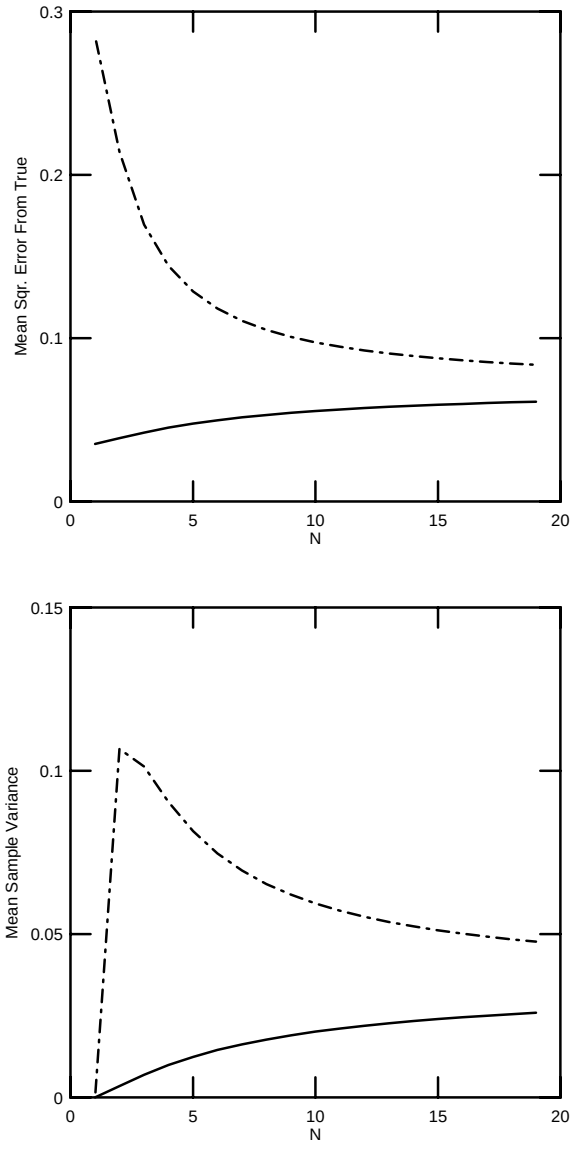

Figure 9.1: Top. Mean square error. Bot. Sample variance.

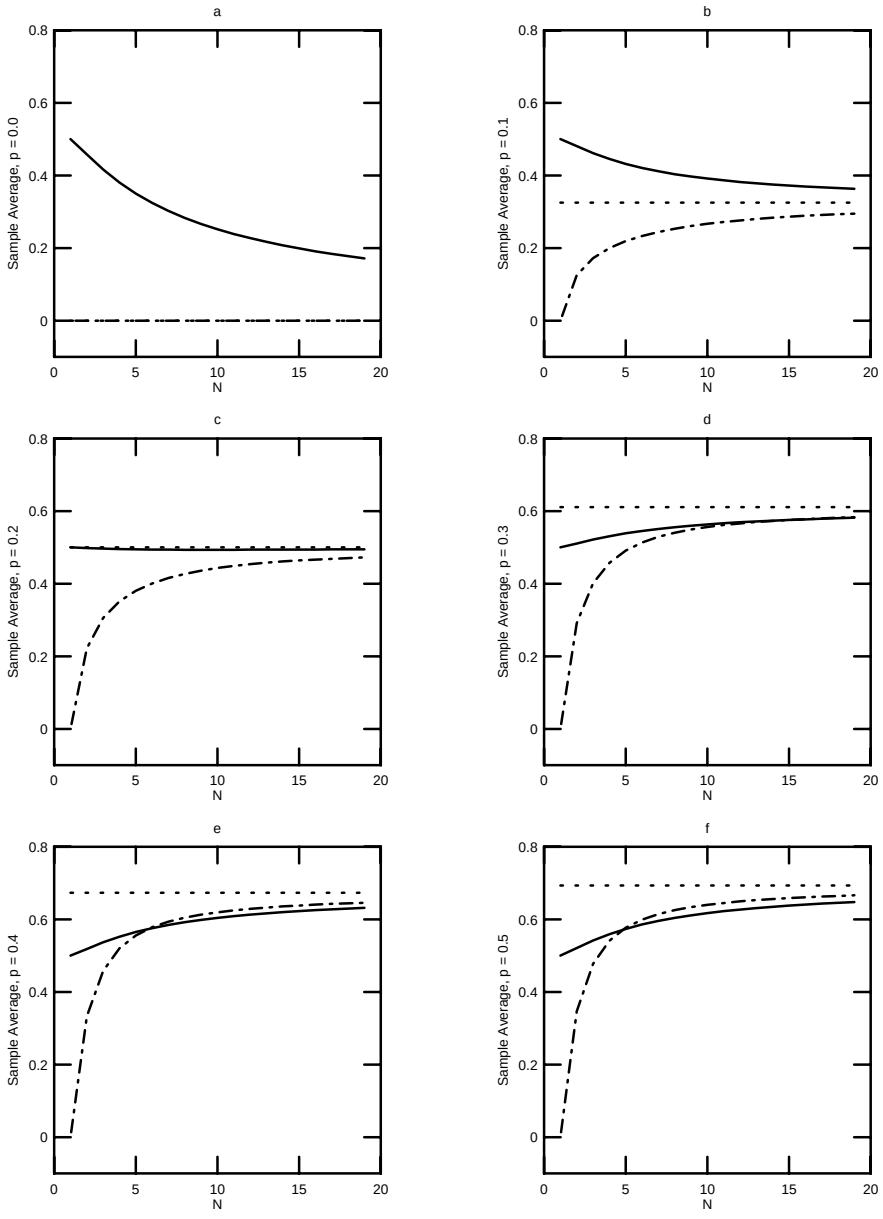

Figure 9.2: Sample average.

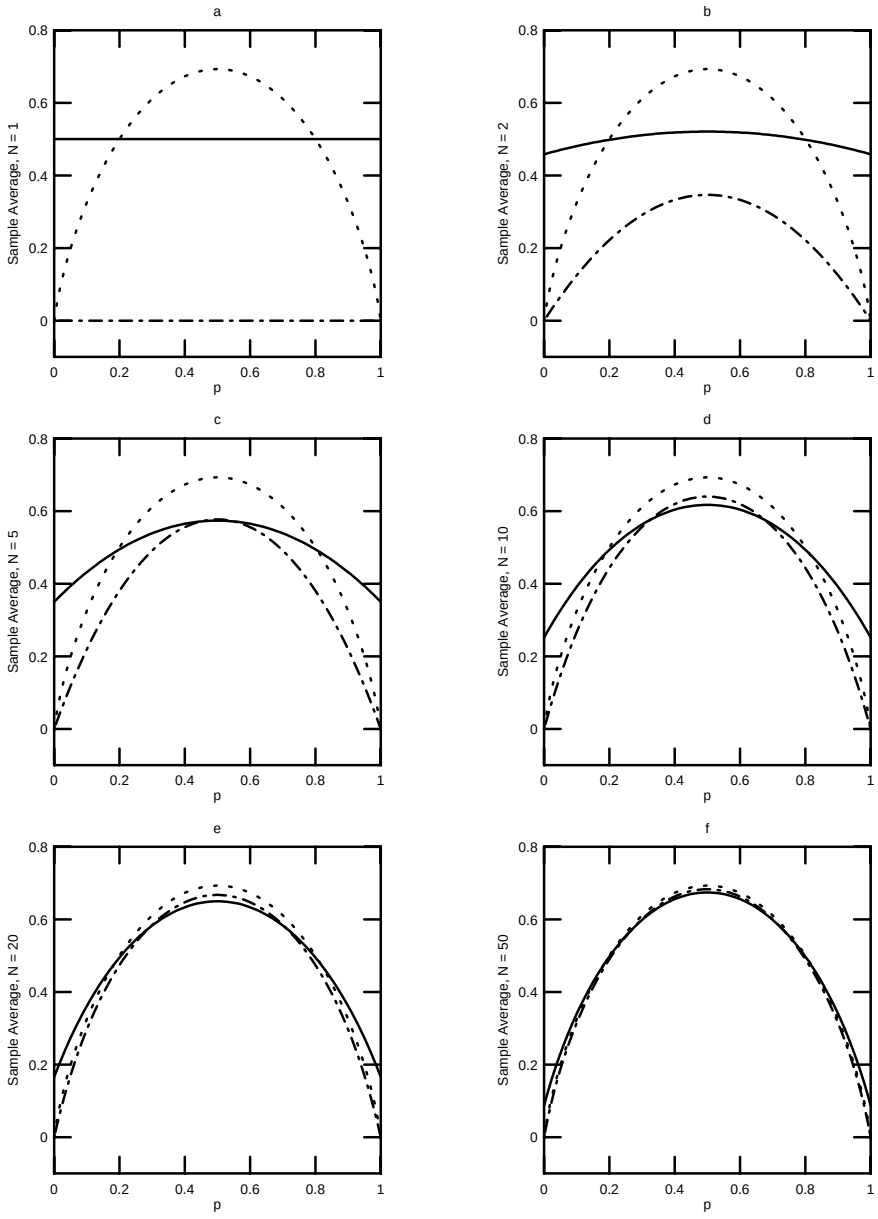

Figure 9.3: Sample average.

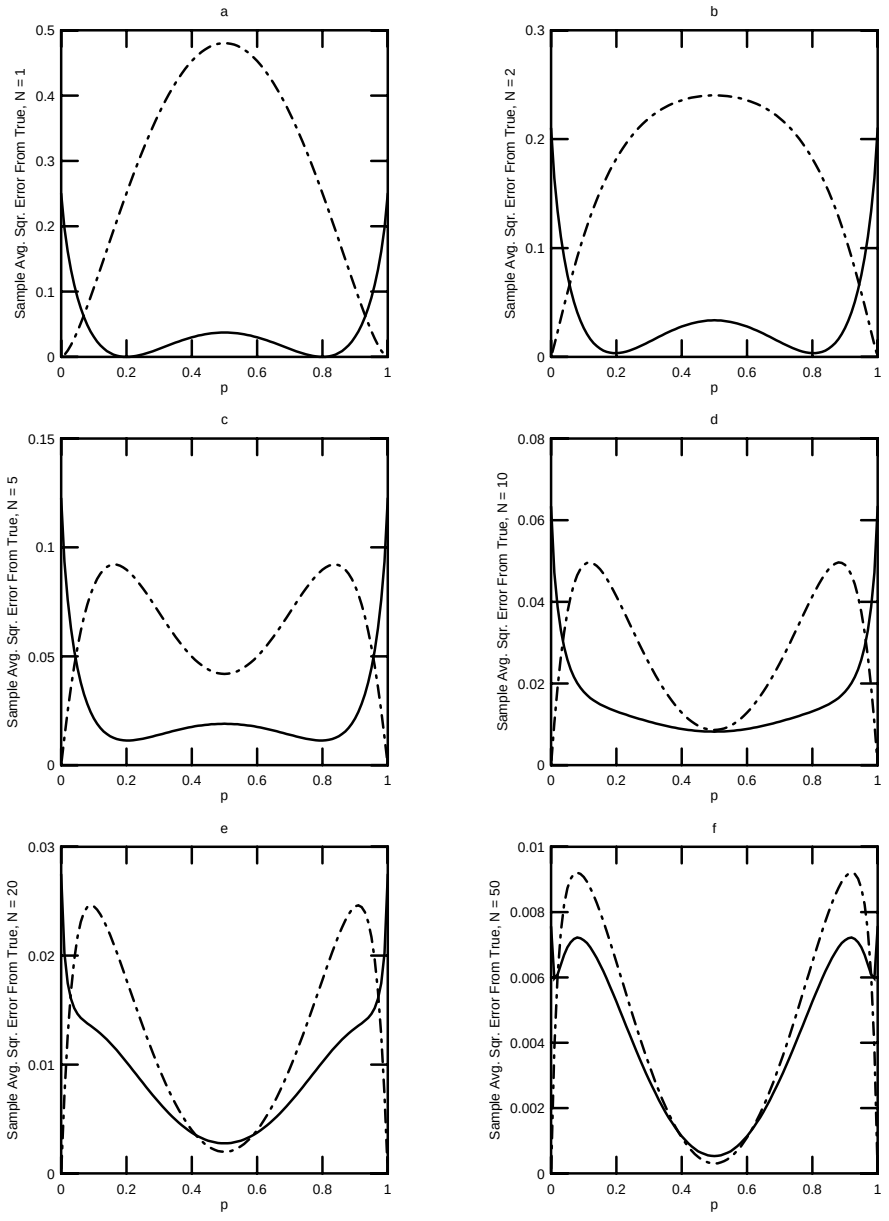

Figure 9.4: Average square error from true.

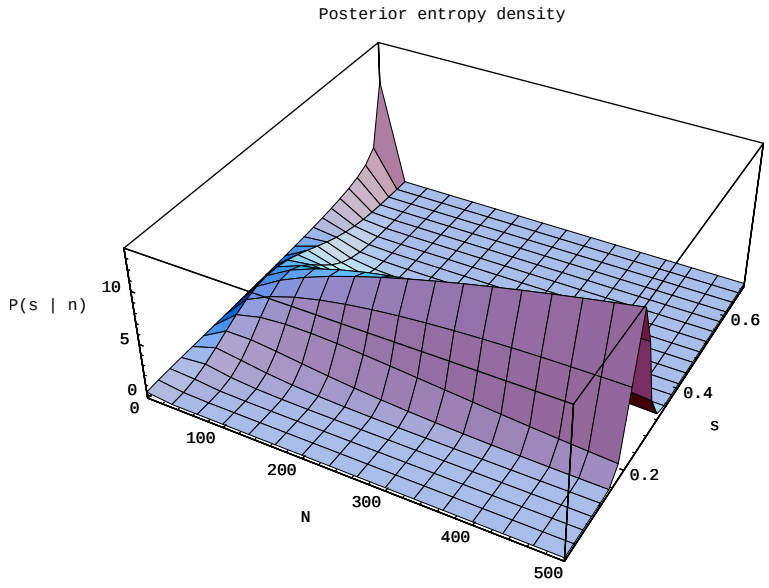

Figure 9.5: Posterior density of entropy.

9.3.11 Estimators for moments, correlations, cumulants.

Consider the function $A(\boldsymbol{p}) = \sum_{\boldsymbol{q}} a(\boldsymbol{q}) \prod_{i=1}^{m} p_i^{q_i}$. For instance, special cases of this function include all of the moments of a set of random variables, the average

$$Avg(\boldsymbol{a}, \boldsymbol{p}) = \sum_{i=1}^{m} a_i p_i \tag{9.46}$$

the variance

$$Var(\boldsymbol{a}, \boldsymbol{p}) = \sum_{i=1}^{m} (a_i - Avg(\boldsymbol{a}, \boldsymbol{p}))^2 p_i = Avg((\boldsymbol{a} - Avg(\boldsymbol{a}, \boldsymbol{p}))^2, \boldsymbol{p}) \tag{9.47}$$

and the other correlations and cumulants, etc. defined within chapter 3.

Here we demonstrate that the posterior moments of A are available in closed form with the results already available. This leads to the observation that the posterior moments are available for any function which may be expressed as a series in the form of A, and this includes all of the functions we consider in this chapter. However, great simplifications occur for many of these functions, as has already been seen in the case of the entropy, and as will be seen in

the cases of the mutual information, chi-squared, and other functions to be considered in the next sections. For now, simply note that

$$A^{(1)}(\boldsymbol{n}) = \frac{\sum_{\boldsymbol{q}} a(\boldsymbol{q}) I[\prod_{i=1}^{m} p_i^{q_i}, \boldsymbol{n}]}{I[1, \boldsymbol{n}]} \tag{9.48}$$

$$= \frac{\sum_{\boldsymbol{q}} a(\boldsymbol{q}) I[1, \boldsymbol{n} + \boldsymbol{q}]}{I[1, \boldsymbol{n}]} \tag{9.49}$$

If the components of $\boldsymbol{q}$ range over the non-negative integers then the powers of A may be written as

$$A(\boldsymbol{p})^k = \left(\sum_{\boldsymbol{q}} a(\boldsymbol{q}) \prod_{i=1}^{m} p_i^{q_i} \right)^k \tag{9.50}$$

$$= \Pi_{\alpha=1}^{k} \sum_{\boldsymbol{q}^\alpha} a(\boldsymbol{q}^\alpha) \prod_{i=1}^{m} p_i^{q_i^\alpha} \tag{9.51}$$

$$= \sum_{\boldsymbol{q}} c(\boldsymbol{a}, \boldsymbol{q}, k) \prod_{i=1}^{m} p_i^{q_i} \tag{9.52}$$

where

$$c(\boldsymbol{a}, \boldsymbol{q}, k) = \sum_{\boldsymbol{q}^1, \ldots, \boldsymbol{q}^k, \sum_{\alpha=1}^{k} q_i^\alpha = q_i} a(\boldsymbol{q}^1) \ldots a(\boldsymbol{q}^k) \tag{9.53}$$

Thus the posterior moments are available as

$$A^{(k)}(\boldsymbol{n}) = \frac{\sum_{\boldsymbol{q}} c(\boldsymbol{a}, \boldsymbol{q}, k) I[1, \boldsymbol{n} + \boldsymbol{q}]}{I[1, \boldsymbol{n}]} \tag{9.54}$$

When $A(\boldsymbol{p}) = \boldsymbol{p}$ then not only are all of the posterior moments available, but the complete posterior distribution is simply expressed (see equations 9.8, 9.12, and 9.10. Equation 9.10 becomes a vector equation, the delta-functions selecting a particular value of $\boldsymbol{p}$).

9.3.12 Extended notation for more complicated functions of probability distributions

Because estimators like that for the mutual information and chi-squared involve what is best described as subset-sums, notation is now introduced to handle this situation.

The convention for indices is i, j, ij, etc. The generalization to two dimensional indices is transparent for the following notation.

Subsets of indices will be denoted using σ_u and σ_v, while the full set of indices will be denoted by σ. Union and intersection are denoted by $\sigma_{u+v} = \sigma_u \cup \sigma_v$ and $\sigma_{uv} = \sigma_u \cap \sigma_v$. Set subtraction is denoted by $\sigma_{u-uv} = \sigma_u \cap \overline{\sigma_{uv}}$ indicating the elements in σ_u not in σ_v. Two sets of indices having non-empty intersection will be called pairwise overlapping. More complicated index set intersection structure also appears. The pairwise overlap case may also occur where one index set is contained in the other, and this will be denoted the contained overlap case.

There are two basic entities that have to be summed over subsets of indices, the counts, $\boldsymbol{n}$, and the probabilities, $\boldsymbol{p}$. Define $\rho_u = \sum_{i \in \sigma_u} p_i$, $\beta_u = \sum_{i \in \sigma_u} \nu_i$, where $\nu_i = n_i + 1$. Also needed are $p_{i\cdot} = \sum_j p_{ij}$, $p_{\cdot j} = \sum_i p_{ij}$, $\nu_{i\cdot} = \sum_j \nu_{ij}$ and $\nu_{\cdot j} = \sum_i \nu_{ij}$. The sum of all counts in the form $\beta = \nu := \sum_i \nu_i$ is useful.

Exponents will involve the variables α_i or η_u, with η_u being variables to be differentiated with respect to, and η_u being associated with the subset σ_u. The sum of these variables will be denoted by $\eta = \sum_u \eta_u$.

The variables of transformation will be taken to be s, t, η_u (associated with σ_u), and τ.

Hypergeometric functions ${}_pF_q$ and ${}_{p_1,p_2,p_{12}}F_{q_1,q_2,q_{12}}$ are given in appendix 9.4, along with the definition of the Pockhammer symbol $(a)_b$.

The product of gamma functions $\gamma_{\boldsymbol{n}} = \prod_{i=1}^{m} \Gamma(\nu_i)$ is useful.

The prior is suppressed, The subscript Δ_P on an integral indicates that the surface of integration and weight of integration are those set by the prior, see section 9.3.6.

9.3.13 More integration techniques

In this section the reader is led quickly via a series of integration theorems to the the next section where the theorems are applied.

Theorem 9. *If* $Re(\alpha_i) > 0$, $i = 1, 2$, *and* $\alpha = \alpha_1 + \alpha_2$ *then*

(9.1)

$$\left(p^{\alpha_1 - 1} \otimes p^{\alpha_2 - 1}\right)(\tau) = \frac{\Gamma(\alpha_1)\Gamma(\alpha_2)}{\Gamma(\alpha)} \times \tau^{\alpha - 1} \tag{9.55}$$

(9.2)

$$\left(p^{\alpha_1 - 1} e^{-pt} \otimes p^{\alpha_2 - 1} e^{-pt}\right)(\tau) = \frac{\Gamma(\alpha_1)\Gamma(\alpha_2)}{\Gamma(\alpha)} \times \tau^{\alpha - 1} \times e^{-\tau t} \tag{9.56}$$

Proof: (9.1) Apply theorem 2, noting that $L[p^{\alpha-1}](s) = \frac{\Gamma(\alpha)}{s^\alpha}$. (9.2) Apply theorem 2 noting that $L[p^{\alpha-1}e^{-pt}](s) = \frac{\Gamma(\alpha)}{(s+t)^\alpha}$. Both for $Re(\alpha) > 0$.

Theorem 10. *If* $Re(\alpha_i) > 0$, $i = 1, 2$, *and* $\alpha = \alpha_1 + \alpha_2$ *then*

(10.1)

$$\begin{aligned}\left(p^{\alpha_1 - 1} e^{-pt_1} \otimes p^{\alpha_2 - 1} e^{-p(t_1 + t_2)}\right)(\tau) &= \frac{\Gamma(\alpha_1)\Gamma(\alpha_2)}{\Gamma(\alpha)} \times \tau^{\alpha - 1} \times e^{-\tau(t_1 + t_2)} \\ &\quad \times {}_1F_1(\alpha_1; \alpha; t_2\tau)\end{aligned} \tag{9.57}$$

(10.2)

$$\begin{aligned}\left(p^{\alpha_1 - 1} \otimes p^{\alpha_2 - 1} e^{-pt}\right)(\tau) &= \frac{\Gamma(\alpha_1)\Gamma(\alpha_2)}{\Gamma(\alpha)} \times \tau^{\alpha - 1} \times e^{-\tau(t)} \\ &\quad \times {}_1F_1(\alpha_1; \alpha; t\tau)\end{aligned} \tag{9.58}$$

Proof: (10.1) Write the convolution in its integral form and compare with equation 9.117 of appendix 9.4. (10.2) Substitute $t_1 = 0$ in the result for (10.1).

Theorem 11. *If* $Re(\alpha_1) > 0, Re(\alpha_{12}) > 0, Re(\alpha_2) > 0$, *and* $\alpha = \alpha_1 + \alpha_{12} + \alpha_2$ *then*

$$\left(p^{\alpha_1 - 1} e^{-pt_1} \otimes p^{\alpha_{12} - 1} e^{-p(t_1 + t_2)} \otimes p^{\alpha_2 - 1} e^{-pt_2}\right)(\tau) =$$

$$\frac{\Gamma(\alpha_1)\Gamma(\alpha_{12})\Gamma(\alpha_2)}{\Gamma(\alpha)} \times \tau^{\alpha-1} \times e^{-\tau(t_1+t_2)}$$
$$\times {}_{1,1,0}F_{0,0,1}(\alpha_1, \alpha_2; \alpha; t_2\tau, t_1\tau) \tag{9.59}$$

Proof: Apply theorem 10.1 to find the result for the first convolution, then use the series representation for ${}_1F_1$ for the second convolution, comparing the result with the definition of ${}_{1,1,0}F_{0,0,1}$ in appendix 9.4. Convolving the series term-by-term is ok since the series is uniformly convergent on $[0,1]$.

Theorem 12 presents preliminary results for the non-overlapping case.

Theorem 12. *If the subsets σ_u defined for $u = 1, \ldots, k$, satisfy $\sigma_{uv} = \emptyset$ for all $u \neq v$ and if $Re(\beta_u + \eta_u) > 0$ for $u = 1, \ldots, k$, and if $Re(\nu_i) > 0$ for $i = 1, \ldots, m$ then*

$$I[\prod_{u=1}^{k} \rho_u^{\eta_u}, \boldsymbol{n}] = \frac{\gamma \boldsymbol{n}}{\Gamma(\beta + \eta)} \prod_{u=1}^{k} \frac{\Gamma(\beta_u + \eta_u)}{\Gamma(\beta_u)} \tag{9.60}$$

Proof: Assume $k = 1$ and $Re(\eta_1) < 0$ to begin. Apply $T^{-1}T$ with respect to η_1 to the integral (see appendix 9.6.1 for the definition of the T transform) and evaluate the inner T transform. Noting $e^{-\rho_u t} = \prod_{i \in \sigma_u} e^{-p_i t}$ and that $T[z^\eta](t) = e^{-zt}$ for $\eta_1 < 0$, find

$$I[\rho_1^{\eta_1}, \boldsymbol{n}] = T^{-1}\left[\int_{\Delta_P} d\boldsymbol{p} \left(\prod_{i \in \sigma_1} p_i^{n_i} e^{-p_i t}\right) \times \left(\prod_{i \notin \sigma_1} p_i^{n_i}\right)\right] \tag{9.61}$$

(See appendix 9.7.1 for the justification of the interchange the integral over $\boldsymbol{p}$ and the T transform.) Now, write the transformed integral above as the convolution

$$I[\rho_1^{\eta_1}, \boldsymbol{n}] = T^{-1}\left(\left(\otimes_{i \in \sigma_1} p_i^{n_i} e^{-p_i t}\right) \otimes \left(\otimes_{i \notin \sigma_1} p_i^{n_i}\right)\right)(1) \tag{9.62}$$

Use theorem 9.1 9.55 and induction to find (with $\beta_{\overline{1}} = \beta - \beta_1$)

$$\left(\otimes_{i \notin \sigma_1} p_i^{n_i}\right)(\tau) = \frac{\prod_{i \notin \sigma_1} \Gamma(\nu_i)}{\Gamma(\beta_{\overline{1}})} \times \tau^{\beta_{\overline{1}}-1} \tag{9.63}$$

Similarly, use theorem 9.2 and induction to find

$$\left(\otimes_{i\in\sigma_1} p_i^{n_i} e^{-p_i t}\right)(\tau) = \frac{\prod_{i\in\sigma_1}\Gamma(\nu_i)}{\Gamma(\beta_1)} \times \tau^{\beta_1 - 1} \tag{9.64}$$

Substituting the last two expressions into $I[\rho_1^{\eta_1}, \boldsymbol{n}]$ yields

$$I[\rho_1^{\eta_1}, \boldsymbol{n}] = \frac{\gamma \boldsymbol{n}}{\Gamma(\beta_1)\Gamma(\beta_{\overline{1}})} T^{-1}\left(\tau^{\beta_1 - 1} e^{-\tau t} \otimes \tau^{\beta_{\overline{1}} - 1}\right)(1) \tag{9.65}$$

The T^{-1} transform may now be taken to find (see appendix 9.4,9.7)

$$I[\rho_1^{\eta_1}, \boldsymbol{n}] = \frac{\gamma \boldsymbol{n}}{\Gamma(\beta_1)\Gamma(\beta_{\overline{1}})}\left(\tau^{\beta_1 - \eta_1 - 1} \otimes \tau^{\beta_{\overline{1}} - 1}\right)(1) \tag{9.66}$$

Now apply theorem 9.1 in this expression to find for $Re(\eta_1) < 0$

$$I[\rho_1^{\eta_1}, \boldsymbol{n}] = \frac{\gamma \boldsymbol{n}\Gamma(\beta_1 + \eta_1)}{\Gamma(\beta_1)\Gamma(\beta + \eta)} \tag{9.67}$$

Refer to appendix 9.8 for the continuation to $Re(\eta_1) \geq 0$. Refer to appendix 9.9 for the existence conditions. Now, for $k > 1$ apply the identity operator $T^{-1}T$ k times (with respect to $\eta_1, \ldots, \eta_k$ respectively) and evaluate only the T transforms initially. Since $\sigma_{uv} = \emptyset$ for $u \neq v$, the convolution form of the transformed integral becomes

$$I[\prod_{i=1}^{k} \rho_i^{\eta_i}, \boldsymbol{n}] = T_1^{-1} \ldots T_k^{-1}\left(\left(\otimes_{u=1}^{k}\left(\otimes_{i\in\sigma_u} p_i^{n_i} e^{-p_i t}\right)\right) \otimes \left(\otimes_{i\notin\cup_{u=1}^{k}\sigma_u} p_i^{n_i}\right)\right)(1) \tag{9.68}$$

Now extend the application of theorem 9.2 to the k convolution products $(\otimes_{i\in\sigma_u} p_i^{n_i} e^{-p_i t})$ for $u = 1, \ldots, k$. Do the substitutions and take the inverse T transforms to find the result. QED.

Appendix 9.5 contains a derivation of an interesting identity based on an alternate form of the result in theorem 12.

Theorem 13 applies theorem 12 to find non-overlap results needed specifically for the expression of Bayes' estimators for the first two moments of the entropy, mutual information, and various other functions.

Theorem 13. *If the subsets σ_u defined for $u = 1, \ldots, k$, satisfy $\sigma_{uv} = \emptyset$ for all $u \neq v$ and if $Re(\beta_u + \eta_u) > 0$ for $u = 1, \ldots, k$, and if $Re(\nu_i) > 0$ for $i = 1, \ldots, m$ then the following hold*

(13.1) *One logarithm subset sum.*

$$I[\prod_{u=1}^{k} \rho_u^{\eta_u} \times log(\rho_u), \boldsymbol{n}] = \frac{\gamma \boldsymbol{n}}{\Gamma(\beta + \eta)} \prod_{w=1}^{k} \frac{\Gamma(\beta_w + \eta_w)}{\Gamma(\beta_u)} \times \Delta\Phi^{(1)}(\beta_u + \eta_u, \beta + \eta) \tag{9.69}$$

(13.2) *Two logarithms of subset sums, different subsets.*

$$I[\prod_{u=1}^{k} \rho_u^{\eta_u} \times log(\rho_u) \times log(\rho_v), \boldsymbol{n}] = \frac{\gamma \boldsymbol{n}}{\Gamma(\beta + \eta)} \prod_{w=1}^{k} \frac{\Gamma(\beta_w + \eta_w)}{\Gamma(\beta_u)} \times \Big(\Delta\Phi^{(1)}(\beta_u + \eta_u, \beta + \eta) \Delta\Phi^{(1)}(\beta_v + \eta_v, \beta + \eta) - \Phi^{(2)}(\beta + \eta) \Big) \tag{9.70}$$

(13.3) *Squared logarithm of a subset sum.*

$$I[\prod_{u=1}^{k} \rho_u^{\eta_u} \times log(\rho_u) \times log(\rho_v), \boldsymbol{n}] = \frac{\gamma \boldsymbol{n}}{\Gamma(\beta + \eta)} \prod_{w=1}^{k} \frac{\Gamma(\beta_w + \eta_w)}{\Gamma(\beta_u)} \times \Big(\Delta\Phi^{(1)}(\beta_u + \eta_u, \beta + \eta)^2 + \Delta\Phi^{(2)}(\beta_u + \eta_u, \beta + \eta) \Big) \tag{9.71}$$

Proof: The proof is done for theorem 13.1; theorem 13.2 and 13.3 follow in a similar manner. Differentiate both sides of the formula for $I[\prod_{u=1}^{k} \rho_u^{\eta_u}, \boldsymbol{n}]$ given in theorem 12 with respect to η_u using the fact that $\partial_\eta \rho^\eta = \rho^\eta log(\rho)$. (See appendix 9.7.2 for justification of the interchange the integral and derivative.) Doing this gives the desired result. QED.

Theorems 12 and 13 dealt with non-overlap sums. Theorems 14 and 15 below discuss pair-wise overlap sums. In theorem 14a the non-contained overlap

case is discussed. In theorem 14b the contained overlap case is discussed. See appendix 9.4 for the definition of the hypergeometric function ${}_{2,2,0}F_{0,0,1}$.

Theorem 14a. *If the subsets* σ_1 *and* σ_2 *satisfy* $\sigma_{12} \neq \emptyset$, $\sigma_1 \neq \sigma_{12} \neq \sigma_2$, $Re(\beta_1 + \eta_1) > 0$, $Re(\beta_2 + \eta_2) > 0$, *and* $Re(\nu_i) > 0$, $i = 1, \ldots, m$, *then*

$$I[\rho_1^{\eta_1}\rho_2^{\eta_2}, \boldsymbol{n}] = \frac{\gamma \boldsymbol{n}}{\Gamma(\beta+\eta)} \times \frac{\Gamma(\beta_{1+2}+\eta)}{\Gamma(\beta_{1+2})} \times {}_{2,2,0}F_{0,0,1}((\beta_{1-12}, -\eta_2), (\beta_{2-12}, -\eta_1); \beta_{1+2}; 1, 1) \tag{9.72}$$

Proof: To begin, assume that $Re(\eta_i) < 0$, $i = 1, 2$, and that the η_i are not integers. Apply $T_1^{-1}T_2^{-1}T_1T_2$ (T_i is with respect to η_i, see appendix 9.6.1) to the integral $I[\rho_1^{\eta_1}\rho_2^{\eta_2}, \boldsymbol{n}]$. Evaluating the (non-inverse) T transforms yields the convolution (see theorem 1)

$$\begin{aligned} I[\rho_1^{\eta_1}\rho_2^{\eta_2}, \boldsymbol{n}] \;=\; & T_1^{-1}T_2^{-1}\left(\left(\otimes_{i\in\sigma_{1-12}} p_i^{n_i} e^{-p_i t_1}\right) \otimes \left(\otimes_{i\in\sigma_{12}} p_i^{n_i} e^{-p_i(t_1+t_2)}\right)\right. \\ & \left. \otimes \left(\otimes_{i\in\sigma_{2-12}} p_i^{n_i} e^{-p_i t_2}\right) \otimes \left(\otimes_{i\notin\sigma_{1+2}} p_i^{n_i}\right)\right)(1) \end{aligned} \tag{9.73}$$

(See appendix 9.7.1 for justification of the interchange of the integral over $\boldsymbol{p}$ and the T transform.) Apply theorem 9.1 and induction to find (where $\beta_{\overline{1+2}} = \beta - \beta_{1+2}$)

$$\left(\otimes_{i\notin\sigma_{1+2}} p_i^{n_i}\right) = \frac{\prod_{i\notin\sigma_{1+2}} \Gamma(\nu_i)}{\Gamma(\beta_{\overline{1+2}})} \times \tau^{\beta_{\overline{1+2}}-1} \tag{9.74}$$

Similarly, use theorem 9.2 and induction to find

$$\begin{aligned} &\left(\left(\otimes_{i\in\sigma_{1-12}} p_i^{n_i} e^{-p_i t_1}\right) \otimes \left(\otimes_{i\in\sigma_{12}} p_i^{n_i} e^{-p_i(t_1+t_2)}\right) \otimes \left(\otimes_{i\in\sigma_{2-12}} p_i^{n_i} e^{-p_i t_2}\right)\right)(\tau) = \\ &\frac{\prod_{i\in\sigma_{1+2}} \Gamma(\nu_i)}{\Gamma(\beta_{1-12})\Gamma(\beta_{12})\Gamma(\beta_{2-12})} \left(p^{\beta_{1-12}-1} e^{-pt_1} \otimes p^{\beta_{12}-1} e^{-p(t_1+t_2)} \otimes p^{\beta_{2-12}-1} e^{-pt_2}\right)(\tau) \end{aligned} \tag{9.75}$$

Substitute the result for theorem 11 into the triple convolution above, and substitute the last two expressions into the convolution form of the transformed

integral to find

$$I[\rho_1^{\eta_1}\rho_2^{\eta_2},\boldsymbol{n}] = T_1^{-1}T_2^{-1}[\frac{\gamma\boldsymbol{n}}{\Gamma(\beta_{1+2})\Gamma(\beta_{\overline{1+2}})}$$
$$\times\Big(\Big(p^{\beta_{1+2}-1}e^{-p(t_1+t_2)}\times{}_{1,1,0}F_{0,0,1}(\beta_{1-12},\beta_{2-12};\beta_{1+2};t_2p,t_1p)\Big)\otimes p^{\beta_{\overline{1+2}}-1}\Big)](1) \tag{9.76}$$

Now, take inverse T transforms and apply theorem 9.1 to find the desired result. Refer to appendix 9.9 to determine the conditions for the existence of the identity. Refer to appendix 9.8 for the continuation of the result to $Re(\eta_i)\geq 0$. Finally, for values of $\eta\geq 0$, refer to appendix 9.10. QED.

See appendix 9.5 for a derivation of two interesting identities resulting from alternate forms of this proof. When $\sigma_2\subset\sigma_1$ the above result simplifies as in theorem 14b below.

Theorem 14b. *If the subsets σ_1,σ_2 satisfy $\sigma_{12}\neq\emptyset$, $\sigma_2\subset\sigma_1$, $Re(\beta_1+\eta)>0$, $Re(\beta_{12}+\beta_2)>0$, and $Re(\nu_i)>0$, $i=1,\ldots,m$, then*

14b.1

$$I[\rho_1^{\eta_1}\rho_2^{\eta_2},\boldsymbol{n}] = \frac{\gamma\boldsymbol{n}}{\Gamma(\beta+\eta)}\times\frac{\Gamma(\beta_1+\eta)}{\Gamma(\beta_1)}\times{}_2F_1((\beta_{1-12},-\eta_2);\beta_1;1) \tag{9.77}$$

14b.2

$$I[\rho_1^{\eta_1}\rho_2^{\eta_2},\boldsymbol{n}] = \frac{\gamma\boldsymbol{n}}{\Gamma(\beta_{12})}\times\frac{\Gamma(\beta_{12}+\eta_2)}{\Gamma(\beta_1+\eta_2)}\times\frac{\Gamma(\beta_1+\eta)}{\Gamma(\beta+\eta)} \tag{9.78}$$

Proof: Similar to proof of theorem 14a, but apply theorem 10.1 instead of theorem 11. The second form (theorem 14b.2) of the result is derived by applying Gauss's identity (see appendix 9.5) to the first form of the result above. QED.

Theorems 15a and 15b build upon the results of theorems 14a and 14b respectively and state results needed to express specific terms of the various Bayes' estimators. Theorem 15a contains results for the non-contained overlap

case. Theorem 15b contains results for the contained overlap case. Since we are most directly interested in non-negative integer η's and because simplification occurs at those η's, Theorem 15a is stated only for non-negative integer η's.

Theorem 15a. *If $\eta_1 \geq 0$ and $\eta_2 \geq 0$ are integers and the conditions for theorem 14a hold then*

15a.1

$$I[\rho_1^{\eta_1}\rho_2^{\eta_2}, \boldsymbol{n}] = C^{(0)}F^{(00)} \tag{9.79}$$

15a.2

$$I[\rho_1^{\eta_1} log(\rho_1)\rho_2^{\eta_2}, \boldsymbol{n}] = C^{(1)}F^{(00)} + C^{(0)}F^{(10)} \tag{9.80}$$

15a.3

$$I[\rho_1^{\eta_1} log(\rho_1)\rho_2^{\eta_2} log(\rho_2), \boldsymbol{n}] = C^{(2)}F^{(00)} + C^{(1)}[F^{(10)} + F^{(01)}] + C^{(0)}F^{(11)} \tag{9.81}$$

15a.4

$$I[\rho_1^{\eta_1} log(\rho_1)^2\rho_2^{\eta_2}, \boldsymbol{n}] = C^{(2)}F^{(00)} + 2C^{(1)}F^{(10)} + C^{(0)}F^{(20)} \tag{9.82}$$

where

$$C^{(0)} := \frac{\gamma \boldsymbol{n}}{\Gamma(\beta_{1+2})} \times \frac{\Gamma(\beta_{1+2}+\eta)}{\Gamma(\beta+\eta)} \tag{9.83}$$

$$C^{(1)} := C^{(0)} \times \Delta\Phi^{(1)}(\beta_{1+2}+\eta, \beta+\eta) \tag{9.84}$$

$$C^{(1)} := C^{(0)} \times \left(\Delta\Phi^{(1)}(\beta_{1+2}+\eta, \beta+\eta)^2 + \Delta\Phi^{(2)}(\beta_{1+2}+\eta, \beta+\eta)\right) \tag{9.85}$$

and

(a)

$$F^{(00)} := \eta_1!\eta_2! \sum_{i=0}^{\eta_2}\sum_{j=0}^{\eta_1} \frac{(\beta_{1-12})_i(\beta_{1-12})_i}{(\beta_{1+2})_{i+j}} \frac{1}{(\eta_1-j)!(\eta_2-i)!} \frac{(-1)^{i+j}}{i!j!} \tag{9.86}$$

(b)

$$F^{(10)} := \eta_1!\eta_2! \sum_{i=0}^{\eta_2} \sum_{j=0}^{\infty} \frac{(\beta_{1-12})_i (\beta_{2-12})_j}{(\beta_{1+2})_{i+j}} \frac{1}{(\eta_2 - i)!} \frac{(-1)^{i+j}}{i!j!} Q_1(j, \eta_1) \tag{9.87}$$

with Q_1 *given by*

$$\begin{aligned} Q_1(j, \eta_1) \quad := \quad & (1 - \theta(j - \eta_1 - 1)) \frac{(-1)^j}{(\eta_1 - j)!} \sum_{r=0}^{j-1} \frac{1}{\eta_1 - r} \\ & + \theta(j - \eta_1 - 1)(-1)^{\eta_1 + 1} \Gamma(j - \eta_1) \end{aligned} \tag{9.88}$$

(c) $F^{(01)}$ *is the same as (b) with* $i \leftrightarrow j$ *and* $\eta_1 \leftrightarrow \eta_2$.

(d)

$$F^{(11)} := \eta_1!\eta_2! \sum_{i=0}^{\infty} \sum_{j=0}^{\infty} \frac{(\beta_{1-12})_i (\beta_{2-12})_j}{(\beta_{1+2})_{i+j}} \frac{1}{i!j!} Q_1(j, \eta_1) Q_1(i, \eta_2) \tag{9.89}$$

(e)

$$F^{(20)} := \eta_1!\eta_2! \sum_{i=0}^{\eta_2} \sum_{j=0}^{\infty} \frac{(\beta_{1-12})_i (\beta_{2-12})_j}{(\beta_{1+2})_{i+j}} \frac{1}{(\eta_2 - i)!} \frac{(-1)^{i+j}}{i!j!} Q_2(j, \eta_1) \tag{9.90}$$

with Q_2 *given by*

$$\begin{aligned} Q_2(j, \eta_1) \quad := \quad & (1 - \theta(j - \eta_1 - 1)) \frac{(-1)^j}{(\eta_1 - j)!} \sum_{r,s=0, r \neq s}^{j-1} \frac{1}{(\eta_1 - r)(\eta_1 - s)} \\ & + \theta(j - \eta_1 - 1)(-1)^{\eta_1 + 1} 2\Gamma(j - \eta_1) \sum_{r=0, r \neq \eta_1}^{j-1} \frac{1}{(\eta_1 - r)} \end{aligned} \tag{9.91}$$

Proof: The proof is done for theorem 15a.2. The other cases have similar proofs. Differentiate both sides of the expression for $I[\rho_1^{\eta_1} \rho_2^{\eta_2}, \boldsymbol{n}]$ given in theorem 14a with respect to η_1. After differentiating, the left-hand side is given by $I[\rho_1^{\eta_1} log(\rho_1) \rho_2^{\eta_2}, \boldsymbol{n}]$. (The justification of the interchange of the integral and derivative is given in appendix 9.7.1.) Write the differentiated right-hand side as

$$\partial_{\eta_1} \left(C^{(0)} \times {}_{2,2,0}F_{0,0,1}((\beta_{1-12}, -\eta_2), (\beta_{2-12}, -\eta_1); \beta_{1+2}; 1, 1) \right) \tag{9.92}$$

This expands to $\partial_{\eta_1}(C^{(0)}) \times {}_{2,2,0}F_{0,0,1}(\ldots) + C^{(0)}\partial_{\eta_1}({}_{2,2,0}F_{0,0,1}(\ldots))$. The derivative of $C^{(0)}$ is given by $\partial_{\eta_1} C^{(0)} = C^{(0)} \times \Delta\Phi^{(1)}(\beta_{1+2} + \eta, \beta + \eta) = C^{(1)}$. The undifferentiated hypergeometric is evaluated at η_1 and η_2 using the results in appendix 9.10, cases 1 and 2. This evaluates to $F^{(00)}$ defined in (a) above. The derivative of the hypergeometric may be taken term-by-term (this is justified below). Use the results in appendix 9.10, cases 1 and 2, equations 9.144 and 9.147, to evaluate this derivative η_1 and η_2. Doing this gives the expression $F^{(10)}$ defined in (b). With these derivatives and evaluations, theorem 15a.2 follows immediately. Now consider the validity of term-by-term differentiation of the hypergeometric. There exists a closed neighborhood N containing the integer η_1 with $Re(\beta_1 + x) \geq 0, \forall x \in N$. The results of appendix 9.10 show that any truncation (in j) of the series for ${}_{2,2,0}F_{0,0,1}((\beta_{1-12}, -\eta_2), (\beta_{2-12}, -x); \beta_{1+2}; 1, 1)$ (see appendix 9.4) may be differentiated with respect to x on N. The sequence of derivatives of the increasing order truncations converges uniformly on N. (To see this, note that $S_i(x) := \sum_{j=\eta_1+1}^{\infty} \frac{\Gamma(\beta_{2-12}+j)}{\Gamma(\beta_{1+2}+i+j)} \frac{\Gamma(-x+j)}{j!}$ is convergent for each i, and $Re(\beta_1 + x) > 0$. Now, note that $S_i(x)$ is a series of terms each monotonic on N with the same monotonicity in x holding for each term, and that the summation over i in (b) is finite. These observations and the convergence just established demonstrate the claim of uniform convergence.) Finally, by theorem 7.17 of [73], the sequence of derivatives of the increasing order truncations converges to the derivative of the limit of the series on N, justifying the term-by-term differentiation of the infinite series. QED.

See appendix 9.5 for some comments regarding alternate forms for the results given above in theorem 15a. Theorem 15b builds on theorem 14b and states the results for the case in which there are two subset sums, with the indices of one subset completely contained in the other. Here, unlike in theorem 15a, there is no hypergeometric function to consider, so the presentation of these results is much shorter. Further, unlike theorem 15a, the expressions given are valid for all η's in the range specified (not just at nonnegative integers as in

theorem 15a) because there are no poles in the expressions being considered at the integers and therefore no further simplification occurs at these points.

Theorem 15b. *If the conditions for theorem 14b hold, then*

15b.1

$$I[\rho_1^{\eta_1} log(\rho_1)\rho_2^{\eta_2}, \boldsymbol{n}] = C^{(00)} \times \Delta\Phi^{(1)}(\beta_1 + \eta, \beta + \eta) \tag{9.93}$$

15b.2

$$I[\rho_1^{\eta_1}\rho_2^{\eta_2} log(\rho_2), \boldsymbol{n}] = C^{(00)} \times \Big(\Delta\Phi^{(1)}(\beta_{12} + \eta, \beta_1 + \eta_2) + \Delta\Phi^{(1)}(\beta_1 + \eta, \beta + \eta)\Big) \tag{9.94}$$

15b.3

$$\begin{aligned} I[\rho_1^{\eta_1} log(\rho_1)\rho_2^{\eta_2} log(\rho_2), \boldsymbol{n}] = C^{(00)} \times \Big(&\Delta\Phi^{(1)}(\beta_1 + \eta, \beta + \eta)^2 \\ &+\Delta\Phi^{(1)}(\beta_{12} + \eta_2, \beta_1 + \eta_2)\Delta\Phi^{(1)}(\beta_1 + \eta, \beta + \eta) \\ &+\Delta\Phi^{(2)}(\beta_1 + \eta, \beta + \eta)\Big) \end{aligned} \tag{9.95}$$

15b.4

$$I[\rho_1^{\eta_1} log(\rho_1)\rho_2^{\eta_2}, \boldsymbol{n}] = C^{(00)} \times \Big(\Delta\Phi^{(1)}(\beta_1 + \eta, \beta + \eta)^2 + \Delta\Phi^{(2)}(\beta_1 + \eta, \beta + \eta)\Big) \tag{9.96}$$

15b.5

$$\begin{aligned} I[\rho_1^{\eta_1}\rho_2^{\eta_2} log(\rho_2)^2, \boldsymbol{n}] = &C^{(00)} \times \Big(\Big(\Delta\Phi^{(1)}(\beta_{12} + \eta_2, \beta_1 + \eta_2) + \Delta\Phi^{(1)}(\beta_1 + \eta, \beta + \eta)\Big)^2 \\ &+ \Delta\Phi^{(1)}(\beta_{12} + \eta_2, \beta_1 + \eta) + \Delta\Phi^{(1)}(\beta_1 + \eta, \beta + \eta)\Big) \end{aligned} \tag{9.97}$$

where

$$C^{(00)} = \frac{\gamma \boldsymbol{n}}{\Gamma(\beta_{12})} \frac{\Gamma(\beta_{12}+\eta_2)}{\Gamma(\beta_1+\eta_2)} \frac{\Gamma(\beta_1+\eta)}{\Gamma(\beta+\eta)} \tag{9.98}$$

Proof: The proof is done for theorem 15b.2. The proofs of the other results follow in a similar manner. The result of theorem 14b is

$$I[\rho_1^{\eta_1}\rho_2^{\eta_2}, \boldsymbol{n}] = C^{(00)} = \frac{\gamma \boldsymbol{n}}{\Gamma(\beta_{12})} \frac{\Gamma(\beta_{12}+\eta_2)}{\Gamma(\beta_1+\eta_2)} \frac{\Gamma(\beta_1+\eta)}{\Gamma(\beta+\eta)} \tag{9.99}$$

Differentiate both sides of this with respect to η_2. The left-hand side of the differentiated expression is $I[\rho_1^{\eta_1}\rho_2^{\eta_2} log(\rho_2), \boldsymbol{n}]$. (The justification of the interchange of the integral and derivative is given in appendix 9.7.2.) The derivative of $\frac{\Gamma(\beta_{12}+\eta_2)}{\Gamma(\beta_1+\eta_2)}$ is given by $\frac{\Gamma(\beta_{12}+\eta_2)}{\Gamma(\beta_1+\eta_2)} \times \Delta\Phi^{(1)}(\beta_{12}+\eta_2, \beta+\eta_2)$. The derivative of $\frac{\Gamma(\beta_1+\eta)}{\Gamma(\beta+\eta)}$ is given by $\frac{\Gamma(\beta_1+\eta)}{\Gamma(\beta+\eta)} \times \Delta\Phi^{(1)}(\beta_1+\eta, \beta+\eta)$. Substituting these expressions for the appropriate derivatives in the overall derivative of the right-hand side of the equality above for $C^{(00)}$ gives the claimed result. QED.

9.3.14 Estimators for functions involving up-to-pairwise overlap integrals

Results for the estimators of the entropy (first and second moments), mutual information (first and second moments), average (first and second moments), covariance (first moment), and chi-squared (first moment) are given next.

Entropy $S(\boldsymbol{p}) = -\sum_{i=1}^{m} p_i log(p_i)$. These results appear in a different form in equations 9.36 and 9.37.

Theorem 16. *If* $Re(\nu_i) > 0$, $i = 1, \ldots, m$, *then*

16.1

$$S^{(1)}(\boldsymbol{n}) = \sum_{i=1}^{m} \frac{\nu_i}{\nu} \Delta\Phi^{(1)}(\nu_i+1, \nu+1) \tag{9.100}$$

16.2

$$
\begin{aligned}
&S^{(2)}(\boldsymbol{n}) = \\
&-\sum_{i\neq j=1}^{m} \frac{\nu_i\nu_j}{\nu(\nu+1)} \times \\
&\qquad \Big(\Delta\Phi^{(1)}(\nu_i+1,\nu+2)\Delta\Phi^{(1)}(\nu_j+1,\nu+2) - \Phi^{(2)}(\nu+2)\Big) \\
&+\sum_{i=1}^{m} \frac{\nu_i(\nu_i+1)}{\nu(\nu+1)} \times \Big(\Delta\Phi^{(1)}(\nu_i+2,\nu+2)^2 + \Delta\Phi^{(2)}(\nu_i+2,\nu+2)\Big)
\end{aligned}
\tag{9.101}
$$

Mutual information $M(\boldsymbol{p}) = \sum_{i,j=1}^{m,n} p_{ij} log(\frac{p_{ij}}{p_{i\cdot}p_{\cdot j}})$. Define $\nu_{\overline{ij}} = \nu_{i\cdot} + \nu_{\cdot j} - \nu_{ij}$.

Theorem 17. *If the ν_{ij} are non-negative integers (the integer condition is used only in the simplification of the term in theorem 17.2; for the other terms it may be relaxed) then*

17.1

$$M^{(1)}(\boldsymbol{n}) = \overline{IJ} - \overline{I} - \overline{J} \tag{9.102}$$

where

$$\overline{IJ} = \frac{I[\sum_{i,j=1}^{m,n} p_{ij} log(p_{ij}), \boldsymbol{n}]}{I[1,\boldsymbol{n}]} = \sum_{i,j=1}^{m,n} \frac{\nu_{ij}}{\nu} \Delta\Phi^{(1)}(\nu_{ij}+1,\nu+1) \tag{9.103}$$

$$\overline{I} = \frac{I[\sum_{i=1}^{m} p_{i\cdot} log(p_{i\cdot}), \boldsymbol{n}]}{I[1,\boldsymbol{n}]} = \sum_{i=1}^{m} \frac{\nu_{i\cdot}}{\nu} \Delta\Phi^{(1)}(\nu_{i\cdot}+1,\nu+1) \tag{9.104}$$

$$\overline{J} = \frac{I[\sum_{j=1}^{n} p_{\cdot j} log(p_{\cdot j}), \boldsymbol{n}]}{I[1,\boldsymbol{n}]} = \sum_{j=1}^{n} \frac{\nu_{\cdot j}}{\nu} \Delta\Phi^{(1)}(\nu_{\cdot j}+1,\nu+1) \tag{9.105}$$

17.2

$$M^{(2)}(\boldsymbol{n}) = \overline{IJRS} + \overline{IR} + \overline{JS} - 2(\overline{IJR} + \overline{IJS} - \overline{IS}) \tag{9.106}$$

where

$$\overline{IJRS} = \frac{\sum_{i,j=1}^{m,n}\sum_{r,s=1}^{m,n} I[p_{ij}log(p_{ij})p_{rs}log(p_{rs}),\boldsymbol{n}]}{I[1,\boldsymbol{n}]} =$$

$$\sum_{i,j=1}^{m,n}\sum_{r,s\neq i,j=1}^{m,n}\frac{\nu_{ij}\nu_{rs}}{\nu(\nu+1)}\times$$

$$\Big(\Delta\Phi^{(1)}(\nu_{ij}+1,\nu+2)\Delta\Phi^{(1)}(\nu_{rs}+1,\nu+2)-\Phi^{(2)}(\nu+2)\Big)$$

$$+\quad\sum_{i,j=1}^{m,n}\frac{\nu_{ij}(\nu_{ij}+1)}{\nu(\nu+1)}\times\Big(\Delta\Phi^{(1)}(\nu_{ij}+2,\nu+2)^2+\Delta\Phi^{(2)}(\nu_{ij}+2,\nu+2)\Big) \tag{9.107}$$

$$\overline{IR} = \frac{\sum_{i,r=1}^{m} I[p_{i\cdot}log(p_{i\cdot})p_{r\cdot}log(p_{r\cdot}),\boldsymbol{n}]}{I[1,\boldsymbol{n}]}$$

$$\sum_{i=1}^{m}\sum_{r=1,r\neq i}^{m}\frac{\nu_{i\cdot}\nu_{r\cdot}}{\nu(\nu+1)}\times$$

$$\Big(\Delta\Phi^{(1)}(\nu_{i\cdot}+1,\nu+2)\Delta\Phi^{(1)}(\nu_{r\cdot}+1,\nu+2)-\Phi^{(2)}(\nu+2)\Big)$$

$$+\quad\sum_{i=1}^{m}\frac{\nu_{i\cdot}(\nu_{i\cdot}+1)}{\nu(\nu+1)}\times\Big(\Delta\Phi^{(1)}(\nu_{i\cdot}+2,\nu+2)^2+\Delta\Phi^{(2)}(\nu_{i\cdot}+2,\nu+2)\Big) \tag{9.108}$$

To find $\overline{JS}$ *substitute* $\nu_{\cdot j}$ *for* $\nu_{i\cdot}$ *and* $\nu_{\cdot s}$ *for* $\nu_{r\cdot}$ *in the expression for* $\overline{IR}$*, letting* $i \leftrightarrow j$ *and* $r \leftrightarrow s$ *and let* $m \leftrightarrow n$*.*

$$\overline{IJR} = \frac{I[\sum_{i,j=1}^{m,n}\sum_{r=1}^{m} p_{ij}log(p_{ij})p_{r\cdot}log(p_{r\cdot}),\boldsymbol{n}]}{I[1,\boldsymbol{n}]} =$$

$$\sum_{i,j=1}^{m,n}\sum_{r=1,r\neq i}^{m}\frac{\nu_{ij}\nu_{r\cdot}}{\nu(\nu+1)}\times$$

$$\Big(\Delta\Phi^{(1)}(\nu_{ij}+1,\nu+2)\Delta\Phi^{(1)}(\nu_{r\cdot}+1,\nu+2)-\Phi^{(2)}(\nu+2)\Big)$$

$$+\quad\sum_{i,j=1}^{m,n}\frac{\nu_{ij}(\nu_{i\cdot}+1)}{\nu(\nu+1)}\times\Big(\Delta\Phi^{(1)}(\nu_{i\cdot}+2,\nu+2)^2$$

$$+\Delta\Phi^{(1)}(\nu_{ij}+1,\nu_{i\cdot}+1)\Delta\Phi^{(1)}(\nu_{i\cdot}+2,\nu+2)+\Delta\Phi^{(2)}(\nu_{i\cdot}+2,\nu+2)\Big) \tag{9.109}$$

To find $\overline{IJS}$ *substitute* $\nu_{\cdot s}$ *for* $\nu_{r\cdot}$*, let* $r \leftrightarrow s$*, and note that the range on the changed summation affected changes from* m *to* n*, and substitute* $\nu_{\cdot j}$ *for* $\nu_{i\cdot}$*, all*

of this in the expression for $\overline{IJR}$.

$$\overline{IS} = \frac{I[\sum_{i=1}^{m}\sum_{s=1}^{n} p_{i.}log(p_{i.})p_{.s}log(p_{.s}), \boldsymbol{n}]}{I[1, \boldsymbol{n}]} =$$

$$\sum_{i=1}^{m}\sum_{s=1}^{n} \frac{\nu_{is}(\nu_{is}+1)}{\nu(\nu+1)} \times \left\{ \left[\Delta\Phi^{(1)}(\nu_{\overline{is}}+2, \nu+2)^2 + \Delta\Phi^{(2)}(\nu_{\overline{is}}+2, \nu+2)\right] \right.$$

$$\left. \times \left[1 - \frac{\nu_{i.}+\nu_{.s}-2\nu_{is}}{\nu} + \frac{(\nu_{i.}-\nu_{is})(\nu_{.s}-\nu_{is})}{\nu(\nu+1)}\right] \right.$$

$$+\Delta\Phi^{(1)}(\nu_{\overline{is}}+2, \nu+2)\times$$

$$\left. \sum_{a=0}^{\infty} Q_1(a,1) \left[\frac{(\nu_{.s}-\nu_{is})_a}{(\nu_{\overline{is}})_a}\left(1+\frac{\nu_{i.}-\nu_{is}}{\nu_{\overline{is}}+a}\right) + \frac{(\nu_{i.}-\nu_{is})_a}{(\nu_{\overline{is}})_a}\left(1+\frac{\nu_{.s}-\nu_{is}}{\nu_{\overline{is}}+a}\right)\right] \right.$$

$$\left. + \sum_{a,b=0}^{\infty} \frac{(\nu_{i.}-\nu_{is})_a(\nu_{.s}-\nu_{is})_b}{(\nu_{\overline{is}})_{a+b}} \frac{Q_1(a,1)}{a!}\frac{Q_1(b,1)}{b!} \right\} \tag{9.110}$$

where Q_1 *is defined in theorem 15a.*

Proof: (17.1) Write the mutual information as the sum of three entropies $S(\boldsymbol{p})$, $S((p_{i.}))$, and $S((p_{.j}))$ and apply theorem 13.1. (17.2) Square the mutual information written as the sum of three entropies and make the obvious identifications necessary for each term. Apply theorems 13.1, 13.2, 13.3, 15a.3, and 15b.3 as needed. QED.

Average $Avg(\boldsymbol{p}) = \sum_{i=1}^{m} p_i X_i$.

Theorem 18. *If* $Re(\nu_i) > 0$, $i = 1, \ldots, m$, *then*

18.1

$$Avg^{(1)}(\boldsymbol{n}) = \sum_{i=1}^{m} \frac{\nu_i}{\nu} X_i \tag{9.111}$$

18.2

$$Avg^{(2)}(\boldsymbol{n}) = \sum_{i\neq j=1}^{m} \frac{\nu_i\nu_j}{\nu(\nu+1)} X_iX_j + \sum_{i=1}^{m} \frac{\nu_i(\nu_i+1)}{\nu(\nu+1)} X_i^2 \tag{9.112}$$

Variance $Var(\boldsymbol{p}) = \sum_{i=1}^{m} p_i (X_i - \mu_X)^2$. Note that $V^2(\boldsymbol{n}) \neq ((\Delta A)^2)^{(1)}(\boldsymbol{n})$. where $(\Delta A)(\boldsymbol{p}) = A(\boldsymbol{p}) - A^{(1)}(\boldsymbol{p})$; $V(\boldsymbol{p})$ is not the variance of the estimator $A^{(1)}$.

Theorem 19. *If* $Re(\nu_i) > 0$, $i = 1, \ldots, m$, *then*

19.1

$$Var^{(1)}(\boldsymbol{n}) = \sum_{i=1}^{m} \frac{\nu_i(\nu - \nu_i)}{\nu(\nu+1)} X_i^2 - \sum_{i \neq j=1}^{m} \frac{\nu_i \nu_j}{\nu(\nu+1)} X_i X_j \tag{9.113}$$

19.2

$$\begin{aligned} Var^{(2)}(\boldsymbol{n}) &= I[1, \boldsymbol{n}]^{-1} \left(\sum_{i,j=1}^{m} I[p_i p_j, \boldsymbol{n}] X_i^2 X_j^2 - 2 \sum_{i,j,k=1}^{m} I[p_i p_j p_k, \boldsymbol{n}] X_i^2 X_j X_k \right. \\ &\quad \left. + \sum_{i,j,k,l=1}^{m} I[p_i p_j p_k p_l, \boldsymbol{n}] X_i X_j X_k X_l \right) \end{aligned} \tag{9.114}$$

Proof: The integrals may be found by applying theorem 12.

Covariance $Cov(\boldsymbol{p}) = \sum_{i,j=1}^{m,n} p_{ij}(X_i - \mu_X)(Y_i - \mu_Y)$.

Theorem 20. *If* $Re(\nu_i) > 0$, $i = 1, \ldots, m$, *then*

$$Cov^{(1)}(\boldsymbol{n}) = \sum_{i,j=1}^{m,n} X_i Y_j \frac{\nu \nu_{ij} - \nu_{i\cdot} \nu_{\cdot j}}{\nu(\nu+1)} \tag{9.115}$$

Proof: Apply theorem 14a.

The second posterior mean, the estimator for the second moment, of the covariance depends on multiple overlap integrals and is given in reference [95], theorem 26.

Chi-squared $\chi^2(\boldsymbol{p}) = \sum_{i,j=1}^{m,n} \frac{(p_{ij} - p_{i\cdot} p_{\cdot j})^2}{p_{i\cdot} p_{\cdot j}}$.

Theorem 21. *If* $Re(\nu_{ij}) > 0$, $Re(\nu_{i\cdot}) > -1$, $Re(\nu_{\cdot j}) > -1$, $i = 1, \ldots, m$, $j = 1, \ldots, n$ *then*

$$(\chi^2)^{(1)}(\boldsymbol{n}) = -1 + \sum_{i,j=1}^{m,n} \frac{(\nu-1)(\nu-2)}{\nu_{\overline{ij}}(\nu_{\overline{ij}}+1)} \times \sum_{a=0}^{\infty} \sum_{b=0}^{\infty} \frac{(\nu_{i\cdot} - \nu_{ij})_a (\nu_{\cdot j} - \nu_{ij})_b}{(\nu_{\overline{ij}}+2)_{a+b}} \tag{9.116}$$

Proof: Apply theorem 14a.

9.3.15 Multiple overlap integration

Multiple overlap integrals are necessary in order to compute the second moment estimators of chi-squared and covariance (strictly, the second moment estimator of the covariance may be computed by expanding the covariance as the sum of powers of the p_i's, but this ignores much of the symmetry that leads to simplifications.) Multiple overlap convolution integrations are rather tricky, and much development needs to be done in order to present the results, amounting to more than three pages just to describe the new notation needed, and another four to present the integration theorems. The interested reader is encouraged to look in other work by this author where the necessary multiple overlap convolution integrals are computed, see reference [95], theorems 22 through 27.

9.4 Appendix A. Hypergeometric functions

The hypergeometric functions used in this chapter are defined here. Paralleling Lebedev [50], let $\boldsymbol{a}$ and $\boldsymbol{b}$ be vectors of dimensions p and q respectively. Define $(\lambda)_k = \Gamma(\lambda + k)/\Gamma(\lambda)$. Define the single summation hypergeometrics F by

$$F(\boldsymbol{a}; \boldsymbol{b}; \tau) = \sum_{i=1}^{\infty} \left(\frac{\Pi_{\alpha=1}^{p} (a_\alpha)_i}{\Pi_{\beta=1}^{q} (b_\beta)_i} \right) \frac{\tau^i}{i!} \tag{9.117}$$

An example of a single summation hypergeometric is ${}_1F_1(a; b; \tau)$, which has the integral representation for $b > a > 0$ (see reference [50]

$${}_1F_1(a; b; \tau) = \frac{\Gamma(b)}{\Gamma(a)\Gamma(b-a)} \int_0^1 dx e^{\tau x} x^{a-1} (1-x)^{b-a-1} \tag{9.118}$$

Now, given vectors $\boldsymbol{a}^1, \boldsymbol{a}^2, \boldsymbol{a}^{12}, \boldsymbol{b}^1, \boldsymbol{b}^2, \boldsymbol{b}^{12}$ of dimensions $p_1, p_2, p_{12}, q_1, q_2, q_{12}$ respectively, define the double summation hypergeometrics

$$\begin{aligned} {}_{p_1,p_1,p_2}F_{p_{12},q_1,q_2} q_{12}(\boldsymbol{a}^1, \boldsymbol{a}^2, \boldsymbol{a}^{12}; \boldsymbol{b}^1, \boldsymbol{b}^2, \boldsymbol{b}^{12}; \tau_1, \tau_2) = \\ \sum_{i=0}^{\infty} \sum_{j=0}^{\infty} \frac{(\Pi_{\alpha_1=1}^{p_1} (a_{\alpha_1}^1)_i)(\Pi_{\alpha_2=1}^{p_2} (a_{\alpha_2}^2)_j)(\Pi_{\alpha_{12}=1}^{p_{12}} (a_{\alpha_{12}}^{12})_{i+j})}{(\Pi_{\beta_1=1}^{q_1} (b_{\beta_1}^1)_i)(\Pi_{\beta_2=1}^{q_2} (b_{\beta_2}^2)_j)(\Pi_{\beta_{12}=1}^{q_{12}} (b_{\beta_{12}}^{12})_{i+j})} \frac{\tau_1^i \tau_2^j}{i!j!} \end{aligned} \tag{9.119}$$

In writing the arguments of hypergeometrics, vectors will be denoted by e.g., $\boldsymbol{c} = (c_1, \ldots, c_k)$. However, when listing the components of a 1-dimensional vector the parentheses will be dropped. Further, when any of the or $\boldsymbol{p}$ or $\boldsymbol{q}$ subscripts are zero (which corresponds to an empty argument for that position), the empty vector argument of the hypergeometric will simply be omitted from the list of arguments.

9.5 Appendix B. Hypergeometric function identities

Gauss's identity for ${}_2F_1$. An alternate proof of theorem 12 leads to an identity. Where the inverse T transform is applied in the proof of theorem 12, instead find the convolution $(\tau^{\beta_1 - 1} e^{-\tau t} \otimes \tau^{\beta - 1})$ using theorem 10.2 and express it in terms of ${}_1F_1$. Now, do the inverse transform and equate this result to the result of theorem 12 to find Gauss's identity:

$$ {}_2F_1((a,b);c;1) = \frac{\Gamma(c)\Gamma(c-a-b)}{\Gamma(c-a)\Gamma(c-b)} \tag{9.120} $$

Identity for ${}_3F_2$. An alternate proof of theorem 14a leads to another identity. Instead of applying theorem 11, apply theorem 10.1 to the first two terms of the pairwise overlap convolution $p^{\beta_{1-12}-1}e^{-pt_1} \otimes p^{\beta_{12}-1}e^{-p(t_1+t_2)}$ and immediately take the inverse T_1 transform. Now, do the final convolution with $p^{\beta_{2-12}-1}e^{-pt_2}$ and take the inverse T_2 transform. Note that there is only a single summation in the result, whereas in the result in theorem 14a there are two summations. On the other hand, the convolution $p^{\beta_{12}-1}e^{-p(t_1+t_2} \otimes p^{\beta_{2-12}-1}e^{-pt_2}$ can be taken first, followed by a convolution with $p^{\beta_{1-12}-1}e^{-pt_1}$, effectively interchanging indices 1 and 2. Equating these two single-sum forms gives the identity

$$ \begin{aligned} &{}_3F_2((a_1,a_2,a_3);(b_1,b_2);1) = \\ &\qquad \frac{\Gamma(b_1)\Gamma(b_2)\Gamma(b_1+b_2-(a_1+a_2+a_3))}{\Gamma(a_2)\Gamma(b_1+b_2-(a_1+a_2))\Gamma(b_1+b_2-(a_2+a_3))} \\ &\qquad\quad \times {}_3F_2(b_1-a_2, b_1+b_2-(a_1+a_2+a_3), b_2-a_2; \\ &\qquad\qquad\qquad b_1+b_2-(a_1+a_2), b_1+b_2-(a_2+a_3);1) \end{aligned} $$

(9.121)

while equating either of the single-sum results just described to the original result of theorem 14a yields Gauss's identity equation 9.120, above. (To make the symmetry of equation 9.121 obvious, try substituting $a_i = (x_j - x_k)/2 - (b_1 + b_2)/3)$.

Further simplifications. Utilizing Gauss's identity (see equation 9.120 above and [3], equation 15.1.1) provides further simplification in theorem 15a for cases 15a.2 and 15a.4. These simplifications are due to simplifications appearing in $F^{(10)}$ and $F^{(01)}$ respectively. The choice of the form of the results presented was made considering the simplicity of the results and consistency between the results.

9.6 Appendix C. Transforms

Appendix 9.6.1 discusses the T transform. Appendix 9.6.2 discusses the Z transform.

9.6.1 Appendix C.1. The T transform

Start with the identity

$$\Gamma(-\eta) = \int_0^\infty u^{-\eta-1} e^{-u} du, \qquad Re(-\eta) > 0 \tag{9.122}$$

for the gamma function. Make the change of variables $u = \rho t$. With $\rho > 0$, independent of t, we find

$$\rho^\eta = \frac{1}{\Gamma(-\eta)} \int_0^\infty t^{-\eta-1} e^{-\rho t} dt, \qquad Re(-\eta) > 0, \qquad \rho > 0 \tag{9.123}$$

Define the operator T^{-1} by

$$T^{-1}[F](\eta) = \frac{1}{\Gamma(-\eta)} \int_0^\infty t^{-\eta-1} F(t) dt, \qquad Re(-\eta) > 0 \tag{9.124}$$

and define the transform T by $T[T^{-1}[F]] = F$. As defined, the transform T is closely related to the Mellin transform [54] (it is an inverse-Mellin transform) and this similarity may be used to establish the conditions for the existence of the transform and its inverse. Of interest in this work is the following: For ρ independent of t, the functions ρ^η and $e^{-\rho t}$ form a transform pair. That is,

$$T[\rho^\eta](t) = e^{-\rho t}, \qquad T^{-1}[e^{-\rho t}](\eta) = \rho^\eta \tag{9.125}$$

When using the T transform, the order of evaluation of the integral over t and other integrals will need to be interchanged. See appendix 9.7.1 for the justification of the commutation of the integrals.

9.6.2 Appendix C.2. The Z transform

Let $f(\boldsymbol{n})$ be any function that factors as $f(\boldsymbol{n}) = \prod_{i=1}^m f_i(n_i)$. For such functions, the Z transform $Z[f](z) = \sum_{n=0}^\infty f(n)z^n$ is useful in simplifying calculations involving sums $\sum_{\boldsymbol{n}} f(\boldsymbol{n})$, where the summation extends over all $\boldsymbol{n}$ having non-negative integer components and $\sum_i n_i = N$. Define the discrete convolution product of two functions g and h by $(g \otimes h)(n) = \sum_{i=0}^n g(i)h(n-i)$. (Note that $\otimes$ is both commutative and associative, so that the order that the convolutions are taken in is irrelevant, justifying the use of the above notation when several functions are involved.)

The Z transform convolution theorem may be thought of as a discretized form of the Laplace convolution theorem (see theorem 2).

Theorem 9.6.2.1: *If $F(N) = \sum_{\boldsymbol{n}} f(\boldsymbol{n})$ where the function $f(\boldsymbol{n}) := \prod_{i=1}^m f_i(n_i)$ then $F(N) = (\otimes_{i=1}^m f_i)(N)$ and $Z[F](z) = \prod_{i=1}^m Z[f_i](z)$, for all z such that $Z[f_i](z)$, $i = 1, \ldots, m$, converges.*

Proof: For $m = 2$ we have $f(\boldsymbol{n}) = (f_1 \otimes f_2)(N)$ and the Z transforms of f_1 and f_2 are given by $Z[f_i](z) = \sum_{n=0}^\infty f_i(n)z^n$, $i = 1, 2$, respectively. For z within the radii of convergence of both of these power series, we

have (after collecting terms having the same power of z) $Z[f_1](z) \times Z[f_2](z) = \sum_{n=0}^{\infty} z^n \sum_{i=0}^{n} f_1(i) f_2(n-i)$. The right-hand side is immediately seen to be $Z[F](z)$. The result for arbitrary m follows by induction. QED.

Note that due to the uniqueness of power series representations, inverses of Z transforms exist on the nonnegative integers.

9.6.3 The Laplace transform and inverse

The Laplace transform is given by

$$L[f](s) := \int_0^{\infty} f(t) e^{-st} dt \tag{9.126}$$

If $L[f]$ is the laplace transform of f then the inverse transform is given by

$$L^{-1}[L[f](s)](x) = f(x) = \frac{1}{2\pi i} \int_{c-i\infty}^{c+i\infty} L[f](s) ds \tag{9.127}$$

where c is a real number greater than all of the real parts of the singularities of $L[f](s)$ (see [50, 54]). The inverse is found by applying the Fourier-Mellin inversion theorem, [14].

9.7 Appendix D. Commuting linear operators

In appendix 9.7.1 we discuss the interchange of integrals. In appendix 9.7.2 we discuss the interchange of derivatives and integrals.

9.7.1 Appendix D.1. Commuting two integrals

Interchanging the integrals appearing in these papers as the $\boldsymbol{p}$ integral and the T transform integral is possible due to Fubini's theorem [49], which justifies the interchange of uncoupled integrations (region of integration of either integral does not depend on the other integral's parameters) when the double integral exists.

9.7.2 Appendix D.2. Commuting integrals and derivatives

Consider differentiating the integral $\int F(x,t)dx$ with respect to t Theorem D.2 generalizes theorem 9.42 of [73] and establishes conditions general enough to allow the commutation of the derivative and integral for the functions appearing in this paper. Define $D_2F(x,t)$ to be the partial derivative of F with respect to its second argument, evaluated at (x,t).

Theorem D.2: *If*

(1) $F(x,t)$ *and* $D_2D(x,t)$ *are defined for* $(x,t) \in \Delta_x \times \Delta_t$, *where* $\Delta_x = (0,\infty)$, *and where* Δ_t *is convex,*

(2) $\int_0^\infty F(x,t)dx$ *exists* $\forall t \in \Delta_t$,

(3) $\forall \epsilon > 0$ *and* $b > 0$, $\exists f(x)$ *with* $f(x) > 0$ *for* $x \in \Delta_t$, *and* $\delta > 0$ *such that* $\int_0^\infty f(x)dx < \epsilon$ *and* $\forall x \geq, t \in \Delta_t$, $\mid t-s \mid < \delta \Rightarrow \mid D_2F(x,t) - D_2F(x,s) \mid < f(x)$,

then $D_2 \int_0^\infty F(x,t)dx = \int_0^\infty D_2F(x,t)dx$ *on* $\Delta_x \times \Delta_t$.

Proof: Let $\phi(s,t) = \frac{F(x,t)-F(x,s)}{t-s}$ for $s \neq t$. By (1) and the mean value theorem, $\forall t > s$ with $t,s \in \Delta_t$, $\exists u(s,t) \in [s,t]$ such that $\phi(s,t) = D_2F(x,u(s,t))$. Using this and (3) we have that for any $\epsilon > 0$, $\forall b > 0$, $\exists \delta > 0$ and a nowhere-negative (in Δ_x) $f(x)$ obeying $\int_0^\infty f(x)dx < \epsilon$ such that if $\mid t-s \mid < \delta$ and $x \geq b$, then $\mid \phi(s,t) - D_2F(x,t) \mid < f(x)$. From this and (2) it follows that for all $b > 0$, $\exists \delta > 0$, and a nowhere-negative (in Δ_x) $f(x)$ obeying $\int_b^\infty f(x)dx < \epsilon$ such that if $\mid t-s \mid < \delta$, then $\mid \int_b^\infty \phi(s,t)dx - \int_b^\infty D_2F(x,t)dx \mid \leq \int_b^\infty f(x)dx < \epsilon$. Taking the limit $s \to t$, noting that $\lim_{s\to t} \int_b^\infty \phi(s,t)dx = D_2 \int_b^\infty F(x,t)dx$, and finally taking $\epsilon \to 0$ with $b = \epsilon$, we arrive at the desired result. QED.

The functions $F(x,t)$ of interest have the form $F(x,t) = x^t log(x)^m e^{-cx}$ with $Re(t) > -1$ and $c > 0$. For these functions it may be shown that the conditions of theorem D.2 hold.

9.8 Appendix E. Analytic continuation: Expanding η's domain

To apply the T transform, the assumption that all $\eta_i < 0$ had to be made. Here we present a simple theorem that expands the region of validity of the various expressions derived in this paper to the region where any of the η_i may be non-negative. We present the theorem for the single subset sum case only, although the multiple non-overlapping subset case and the contained overlap case may be handled in an almost identical manner.

Theorem 9.8.1: *If* $Re(\eta + \beta_1) > 0$, $Re(\eta) \geq 0$ *and* $Re(n_i) > -1$, $i = 1, \ldots, m$ *then*

$$I[\rho^{\eta}, \boldsymbol{n}] = \gamma_{\boldsymbol{n}} \frac{\Gamma(\eta + \beta_1)}{\Gamma(\beta_1)\Gamma(\eta + \beta)} \tag{9.128}$$

Proof: Note that $\eta > 0$ implies that there is an integer $q > 0$ and an $\overline{\eta} < 0$ such that $\eta = \overline{\eta} + q$. Thus $I[\rho^{\eta}, \boldsymbol{n}]$ may be rewritten as

$$I[\rho^{\eta}, \boldsymbol{n}] = \sum_{i \in \sigma} I[p_i \rho^{\overline{\eta}+q-1}, \boldsymbol{n}] = \sum_{i \in \sigma} I[\rho^{\overline{\eta}+q-1}, \boldsymbol{n} + \boldsymbol{e}_i] \tag{9.129}$$

where $[\boldsymbol{e}_i]_j = \delta_{ij}$, the Kronecker delta function. Iterate this operation q times (removing one power from ρ and summing with an increased count vector each time) to find

$$I[\rho^{\eta}, \boldsymbol{n}] = \sum_{i_1 \in \sigma} \cdots \sum_{i_q \in \sigma} I[\rho^{\overline{\eta}}, \boldsymbol{n} + \boldsymbol{e}_{i_1} + \ldots + \boldsymbol{e}_{i_q}] \tag{9.130}$$

Simplify this to yield

$$I[\rho^{\eta}, \boldsymbol{n}] = \sum_{\boldsymbol{q}, \Sigma_i q_i = q} \binom{q}{\boldsymbol{q}} I[\rho^{\overline{\eta}}, \boldsymbol{n} + \boldsymbol{q}] \tag{9.131}$$

where the vector $\boldsymbol{q}$ has nonnegative integer components summing to q with $q_i = 0$ for $i \notin \sigma$. Since $\overline{\eta} < 0$, evaluate the integral $I[\rho^{\overline{\eta}}, \boldsymbol{n} + \boldsymbol{q}]$ using theorem

12 with $k = 1$ (noting that β_1 and β increase by q due to $\boldsymbol{q}$ being added to $\boldsymbol{n}$) to find

$$(*) \qquad I[\rho^{\eta}, \boldsymbol{n}] = \frac{\Gamma(\beta_1 + q + \overline{\eta})}{\Gamma(\beta_1 + q)\Gamma(\beta + q + \overline{\eta})} \sum_{\boldsymbol{q}, \Sigma_i q_i = q} \binom{q}{\boldsymbol{q}} \gamma_{\boldsymbol{n}+\boldsymbol{q}} \tag{9.132}$$

Now, we put $\sum_{\boldsymbol{q}, \Sigma_i q_i = q} \binom{q}{\boldsymbol{q}} \gamma_{\boldsymbol{n}+\boldsymbol{q}}$ into closed form by noting that it is the discrete convolution product of the functions of $(q_i + 1)_{n_i}$ of q_i given by

$$\sum_{\boldsymbol{q}, \Sigma_i q_i = q} \binom{q}{\boldsymbol{q}} \gamma_{\boldsymbol{n}+\boldsymbol{q}} = \Gamma(q+1) \left[\otimes_{i=1}^{m} \frac{\Gamma(n_i + q_i + 1)}{\Gamma(q_i + 1)} \right] (q) \tag{9.133}$$

Apply the Z transform convolution theorem (see appendix 9.6.2) to find

$$\sum_{\boldsymbol{q}, \Sigma_i q_i = q} \binom{q}{\boldsymbol{q}} \gamma_{\boldsymbol{n}+\boldsymbol{q}} = \Gamma(q+1) Z^{-1} \left[\prod_{i=1}^{m} Z[(q_i + 1)_{n_i}](z) \right] (q) \tag{9.134}$$

Note that

$$Z[(q_i + 1)_{n_i}] = Z[\frac{\Gamma(n_i + q_i + 1)}{\Gamma(q_i + 1)}](z) = \Gamma(n_i + 1)(1 - z)^{-(n_i+1)} \tag{9.135}$$

for $\mid z \mid < 1$ and substitute for the Z transforms to find

$$\sum_{\boldsymbol{q}, \Sigma_i q_i = q} \binom{q}{\boldsymbol{q}} \gamma_{\boldsymbol{n}+\boldsymbol{q}} = \gamma_{\boldsymbol{n}} \frac{\Gamma(\beta_1 + q)}{\Gamma(\beta_1)} \tag{9.136}$$

Substituting this result in (*) and simplifying leads to the desired result. QED.

We resort to analytic continuation in the non-contained overlap case.

9.9 Appendix F. Existence conditions

An example of determining the conditions of existence of the various integrals. Existence of the integrals depends upon the behavior of the singularities appearing at the edges of the region of integration.

Consider the single pair overlap integral $I[\rho_1^{\eta_1} \rho_2^{\eta_2}, \boldsymbol{n}]$, where $\rho_1, \rho_2, \sigma_1, \sigma_2$ are as in the definitions for theorem 14a, with the minor change that σ_{1+2}

contains all m indices, which may be made without loss of generality. It will be shows that the conditions for existence of this integral are $Re(\nu_i) > 0$, $i = 1, \ldots, m$, $Re(\beta_1 + \eta_1) > 0$ and $Re(\beta_2 + \eta_2) > 0$. Write the integral as

$$I[\rho_1^{\eta_1}\rho_2^{\eta_2}, \boldsymbol{n}] = \int_{\Delta_P} d\boldsymbol{p}\rho_1^{\eta_1}\rho_2^{\eta_2}\prod_{i=1}^{m} p_i^{n_i} \tag{9.137}$$

That $Re(\nu_i) > 0$, $i = 1, \ldots, m$, follows immediately from the fact that $\int_{\Delta_P} p_i^{n_i}$ exists iff $Re(\nu_i) > 0$ and the fact that any p_i may independently be near zero for this particular overlap case. Now, either ρ_1 or ρ_2 may also be near zero. We consider the first case, in which ρ_1 is near zero, and use symmetry to supply the result for the second case. Letting $x = \sum_{i\in\sigma_1} p_i$ and $y = \sum_{i\in\sigma_{12}} p_i$, rewrite equation 9.137 in a form that isolates ρ_1 as

$$\begin{aligned} I[\rho_1^{\eta_1}\rho_2^{\eta_2}, \boldsymbol{n}] = &\int_0^1 dx x^{\eta_1}\int_0^x dy(1-x+y)^{\eta_2} \\ &\times\left(\int d\boldsymbol{p}_{1-12}\delta(\sum_{i\in\sigma_{1-12}} p_i - (x-y))\prod_{i\in\sigma_{1-12}} p_i^{n_i}\right. \\ &\quad\times\int d\boldsymbol{p}_{12}\delta(\sum_{i\in\sigma_{12}} p_i - y)\prod_{i\in\sigma_{12}} p_i^{n_i} \\ &\quad\left.\times\int d\boldsymbol{p}_{2-12}\delta(\sum_{i\in\sigma_{2-12}} p_i - (1-(x-y)))\prod_{i\in\sigma_{2-12}} p_i^{n_i}\right) \end{aligned} \tag{9.138}$$

Each of the three integrals over $\boldsymbol{p}$ in equation 9.138 may be done in closed form. Do these integrals using theorem 9a and induction to find

$$\begin{aligned} I[\rho_1^{\eta_1}\rho_2^{\eta_2}, \boldsymbol{n}] = &\frac{\gamma\boldsymbol{n}}{\Gamma(\beta_{1-12})\Gamma(\beta_{12})\Gamma(\beta_{2-12})}\times \\ &\int_0^1 dx x^{\eta_1}\int_0^x dy(x-y)^{\beta_{1-12}-1}y^{\beta_{12}-1}(1-(x-y))^{\eta_2+\beta_{2-12}-1} \end{aligned} \tag{9.139}$$

Apply the binomial theorem in equation 9.139 to expand two of the three factors in the integrand, $(x-y)^{\beta_{1-12}-1}$ and $(1-(x-y))^{\eta_2+\beta_{2-12}-1}$, in series. Using these series, note that each term in the series for $(x-y)^{\beta_{1-12}-1}$ will contribute the same power of x after the integration over y, while the terms of the series for

$(1-(x-y))^{\eta_2+\beta_{2-12}-1}$ contribute increasing powers of x after the integration over y. Note also that if the lowest-power-of-x term is integrable over x in a region containing 0, then all of the terms are. Thus, the worst case occurs with the constant term from the binomial series for $(1-(x-y))^{\eta_2+\beta_{2-12}-1}$. After integration over y with this constant term, and considering the small x region of integration, we are left with the integral over x given by

$$C\int_0^{\underline{x}} dx x^{\eta_1+\beta_1-1} \tag{9.140}$$

where $0<\underline{x}<1$, and C is a constant. This integral exists for $Re(\beta_1+\eta_1)>0$. This, symmetry, and the first condition (given by $Re(\nu_i)>0$, $i=1,\ldots,m$) establish the result. The method for more complicated overlap structures is also indicated by this discussion. The discussion above is of interest in another way: it provides a general method for finding multiple overlap integrals without the use of transform theory.

9.10 Appendix G. Derivatives of overlap convolutions: Poles

This appendix discusses derivatives with respect to η of expressions such as $(-\eta)_k$, where $k\in\{0,1,\ldots\}$, and η may be any number within the constraints of existence. We consider the various cases that arise when some combination of poles occurs and demonstrate the various simplified expressions for the derivatives in these cases.

When η is not an integer, there are no poles in either $\Gamma(k-\eta)$ or $\Gamma(-\eta)$ and the usual derivative expressions hold. When η is an integer and $\eta<0$, the usual expressions also hold since $k\geq 0$ and therefore $k-\eta\geq 0$. The case where the usual expressions hold will be denoted *Case* 0.

When η is an integer and $\eta\geq 0$, there are two cases are of interest. The first, *Case* 1, occurs when $\eta\geq 0$ and $k-\eta>0$, so that there are poles in the

denominator of only. The second, *Case* 2, occurs when $\eta \geq 0$ and $k - \eta \leq 0$, so that there are poles in both the numerator and the denominator.

In order to find expressions for the derivatives in cases *Case* 1 and *Case* 2 we use the following three facts. 1) The only singularities of the gamma function $\Gamma(x)$ are simple poles at $x = -n, \quad n = 0, 1, \ldots$ with residues $(-1)^n/n!$ respectively. 2) $\Delta\Phi^{(n)}(k-\eta, -\eta) = (-1)^{n-1}\Gamma(n)\sum_{i=1}^{k} \frac{1}{(k-i-\eta)^n}$ whenever the expressions exist. (The identity $\Gamma(k-\eta)/\Gamma(-\eta) = \prod_{i=1}^{k}(k-i-\eta)$ may be used in deriving this.) 3) $\Gamma(k-\eta)/\Gamma(-\eta)$ is the representation (away from the poles in the gamma functions) of an analytic function (note that $k \geq 0$ is still assumed). Using these facts, the expressions for *Case* 1 and *Case* 2 are found by substituting $\xi = \eta + \epsilon$ for η (now restricted by the conditions of *Case* 1 and *Case* 2 to be a nonnegative integer) in the corresponding case *Case* 0 expressions and taking the limit $\epsilon \to 0$.

Case 0: η non-integer, or η an integer with $\eta < 0$ and $k - \eta > 0$. There are no poles in the numerator or denominator of $\Gamma(k-\eta)/\Gamma(-\eta)$.

The first derivative is given by

$$\partial_\eta^1 \left[\frac{\Gamma(k-\eta)}{\Gamma(-\eta)}\right] = -\frac{\Gamma(k-\eta)}{\Gamma(-\eta)}\Delta\Phi^{(1)}(k-\eta, -\eta) \tag{9.141}$$

The rth derivative may be found by iteration, using equation 9.141 and the recursion relation

$$\partial_\eta^1 \Phi^{(n)}(k-\eta) = -\Phi^{(n+1)}(k-\eta) \tag{9.142}$$

The second derivative is given by

$$\partial_\eta^2 \left[\frac{\Gamma(k-\eta)}{\Gamma(-\eta)}\right] = \frac{\Gamma(k-\eta)}{\Gamma(-\eta)}\left[\Delta\Phi^{(1)}(k-\eta, -\eta)^2 + \Delta\Phi^{(2)}(k-\eta, -\eta)^2\right] \tag{9.143}$$

Case 1: η an integer, $\eta \geq 0$ and $k - \eta > 0$. The denominator contains a pole.

The zeroth derivative is 0.

Taking the appropriate limit in equation 9.141 gives us the first derivative

$$(-1)^{\eta+1}\eta!\Gamma(k-\eta) \tag{9.144}$$

Taking the appropriate limit in equation 9.143 yields the second derivative

$$(-1)^{\eta+1}2\eta!\Gamma(k-\eta)\sum_{i=1,i\neq k-\eta}^{k}(k-i-\eta)^{-1} \tag{9.145}$$

Case 2: η *an integer,* $\eta \geq 0$ *and* $k-\eta \leq 0$. *Both the numerator and denominator contain poles.*

The zeroth derivative is simply

$$(-1)^{k}\frac{\eta!}{(\eta-k)!} \tag{9.146}$$

Taking the appropriate limit in equation 9.141 gives the first derivative

$$(-1)^{k+1}\frac{\eta!}{(\eta-k)!}\Delta\Phi^{(1)}(k-\eta,-\eta) \tag{9.147}$$

Taking the appropriate limit in equation 9.143 yields the second derivative

$$(-1)^{k}\frac{\eta!}{(\eta-k)!}\left[\Delta\Phi^{(1)}(k-\eta,-\eta)^2+\Delta\Phi^{(2)}(k-\eta,-\eta)^2\right] \tag{9.148}$$

Chapter 10

Derivatives of system functions, fluctuations, and correlations

10.1 Correlations and power spectra

The derivatives of some of the functions plotted in the various chapters with respect to the parameters of the Hamiltonian, external fields, coupling strengths, and temperature, are of interest because they are related to correlations of statistical variables in the systems described. These correlations are averages over the possible states of the system using the calculated equilibrium distribution. Such averages must therefore implicitly ignore any dynamics. However, for a system in thermodynamic equilibrium at a given temperature exhibiting stationary ergodic dynamics, the average of a system variable using the equilibrium distribution is equivalent to the time average of the system variable. We can go further. It is shown next that if there is a dynamics, the inverse Fourier transform of the product of the amplitude spectra of two system variables is the equilibrium distribution time-dependent correlation of those system variables. When the time between the two chosen system variables is taken to be zero, we immediately have that the average power in the spectrum of a system variable is equal to the equilibrium distribution correlation. In this manner it is possible to get a connection between the time average behavior of a system and equilibrium derivatives. The connection is best presented by way of the Weiner-Khinchin theorem [27], described next.

10.2 Relationship of equilibrium correlations to power spectrum of dynamics: Weiner-Khinchin theorem

Let $\tilde{X}(t), \tilde{Y}(t)$ be system variables for a system with stationary ergodic dynamics. The correlation function of $X = \tilde{X} - \langle x \rangle$ and $Y = \tilde{Y} - \langle y \rangle$, using the equilibrium time-dependent distribution $\rho_\tau(x, y)$, the probability that $X(t+\tau) = x$ and $Y(t) = y$ (independent of t because of stationarity), is given by

$$\begin{aligned} c_\tau(X, Y) &= \int xy\, \rho_\tau(x, y)\, dxdy \\ &= \langle xy \rangle_{\rho_\tau} \end{aligned} \tag{10.1}$$

Because of ergodicity, this correlation may be written as

$$c_\tau(X, Y) = \lim_{T \to \infty} \frac{1}{2T} \int_{-T}^{T} X(t + \tau)\, Y(t)\, dt \tag{10.2}$$

Now, the Fourier transform of the process $X(t)$ is given by

$$X(\omega) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{i\omega t} X(t)\, dt \tag{10.3}$$

with inverse

$$X(t) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{-i\omega t} X(\omega)\, d\omega \tag{10.4}$$

and similarly for Y. Substituting equation 10.4 into equation 10.2, and considering that X and Y are real, the correlation may be written

$$c_\tau(X, Y) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{-i\omega\tau} X(\omega) Y^*(\omega)\, d\omega \tag{10.5}$$

which shows that the time-dependent correlation is the inverse Fourier transform of a product of the amplitude spectra. This proves the Weiner-Khinchin theorem, equation 10.5. Equations 10.1 and 10.5 are two ways that the time dependent correlation function can be written as an inner product.

Correlations at zero time difference are the objects that may be calculated from the equilibrium distribution. Setting τ to zero in equation 10.5 we

see that these correlations are related to the amplitude spectra by

$$c_0(X,Y) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} X(\omega) Y^*(\omega)\, d\omega \tag{10.6}$$

Thus, the equilibrium correlations give directly the inner product of the amplitude spectra, and in the case that $X = Y$, the correlation gives the total power in the spectrum, i.e.

$$Pwr(X) = c_0(X,X) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} \mid X(\omega) \mid^2 \, d\omega \tag{10.7}$$

so that the *details* of the time progression and spectrum of a single system variable are not available from the equilibrium distribution, but the *total power* in the spectrum of the system variable is available.

As a final comment on a different use of equilibrium correlation functions, when a linear response for the system function to (small) perturbations, and a dissipative stochastic relaxation dynamics of the system function are assumed, the time response coefficient Γ of the system function to changes in the system perturbed from equilibrium is related to the temperature and the equilibrium fluctuation strength D (correlation of the fluctuations with themselves) by, $D = 2\Gamma/\beta$ [69, 30]. This relationship is called the fluctuation-dissipation theorem. It is a relationship between the equilibrium fluctuations of a system variable and the response of the system variable to perturbations from equilibrium.

In the next section the derivatives of the distribution of states, entropies, and moments with respect to the external field, coupling constant, and temperature in the coupled spin system are discussed.

10.3 Derivatives and correlations in spin systems

As usual, let an individual state of the system be denoted by σ, let the probability of σ be given by $P(\sigma) \propto Exp(-\beta E(\sigma))$, let averages using $P(\sigma)$ over σ

be denoted by angle brackets, and let the context be the spin Hamiltonians of equations 6.1 and 8.1.

Distribution function:

$$\frac{\partial P(\sigma)}{\partial \beta} = -P(\sigma)(E(\sigma) - \langle E \rangle) \tag{10.8}$$

$$\frac{\partial P(\sigma)}{\partial R} = \beta P(\sigma)(S_z(\sigma) - \langle S_z \rangle) \tag{10.9}$$

$$\frac{\partial P(\sigma)}{\partial Q} = \beta P(\sigma) \sum_{ij} (S_i(\sigma) S_j(\sigma) - \langle S_i S_j \rangle) \tag{10.10}$$

Average energy:

$$\frac{\partial \langle E \rangle}{\partial \beta} = -(\langle E^2 \rangle - \langle E \rangle^2) \tag{10.11}$$

$$\frac{\partial \langle E \rangle}{\partial R} = \beta(\langle E S_z \rangle - \langle E \rangle \langle S_z \rangle) \tag{10.12}$$

$$\frac{\partial \langle E \rangle}{\partial Q} = \beta \sum_{ij} (\langle E S_i S_j \rangle - \langle E \rangle \langle S_i S_j \rangle) \tag{10.13}$$

Average magnetic moment:

$$\frac{\partial \langle S_z \rangle}{\partial \beta} = -(\langle E S_z \rangle - \langle E \rangle \langle S_z \rangle) \tag{10.14}$$

$$\frac{\partial \langle S_z \rangle}{\partial R} = \beta(\langle S_z^2 \rangle - \langle S_z \rangle^2) \tag{10.15}$$

$$\frac{\partial \langle S_z \rangle}{\partial Q} = \beta \sum_{ij} (\langle S_z S_i S_j \rangle - \langle S_z \rangle \langle S_i S_j \rangle) \tag{10.16}$$

Entropy:

$$\frac{\partial S_E}{\partial \beta} = -\beta(\langle E^2 \rangle - \langle E \rangle^2) \tag{10.17}$$

$$\frac{\partial S_E}{\partial R} = \beta^2(\langle ES_z\rangle - \langle E\rangle\langle S_z\rangle) \tag{10.18}$$

$$\frac{\partial S_E}{\partial Q} = \beta^2\sum_{ij}(\langle ES_iS_j\rangle - \langle E\rangle\langle S_iS_j\rangle) \tag{10.19}$$

From the above note that

$$\frac{\partial S_E}{\partial(\beta,R,Q)} = \beta\frac{\partial\langle E\rangle}{\partial(\beta,R,Q)} \tag{10.20}$$

$$\frac{\partial\langle S_z\rangle}{\partial\beta} = -\frac{1}{\beta}\frac{\partial\langle E\rangle}{\partial R} \tag{10.21}$$

The first relationship is the familiar $dS = dQ/T$ from thermodynamics in different units, and indicates that the entropy is an increasing function of increasing average energy. In fact, combining equations 10.20 we have

$$\frac{\partial S_E}{\partial\langle E\rangle} = \beta \tag{10.22}$$

which does not hold true for the entropy of the reduced densities, for example look at the Ising model, equation 6.30 where the derivatives of the reduced entropy per spin are explicitly found to be larger than $\frac{\partial S_E}{\partial\langle E\rangle}$ at all orders. See also the discussion in chapter 2.

10.4 Derivatives of the reduced entropy

Briefly,

$$\frac{\partial S_r}{\partial\beta} = \langle fE\rangle - \langle f\rangle\langle E\rangle \tag{10.23}$$

where S_r is the reduced entropy based on the distribution P_r, a reduction of the full distribution, and $f = log(P_r)$. Thus the derivative above is the second order correlation between the logarithm of the reduced distribution and the energy. In the non-reduced case we have that $f = -\beta E$. Also,

$$\frac{\partial S_r}{\partial(R,Q)} = \beta\left(\langle f\frac{\partial E}{\partial(R,Q)}\rangle - \langle f\rangle\langle\frac{\partial E}{\partial(R,Q)}\rangle\right) \tag{10.24}$$

Note that for a probability distribution that factors, as for independent particles, as

$$P(\sigma) \propto \Pi_i e^{-\beta E(\sigma_i)} \tag{10.25}$$

the reduced entropy is simply the entropy of the subsystem not averaged over, and $f = -\beta E_r$, where E_r is the energy of the reduced system. All of the previous expressions equations 10.8– 10.19 hold for the subsystem, as expected.

10.5 Ordering properties of the reduced entropy derivatives

The derivatives with respect to β of the reduced entropy are not ordered. This can be seen by rewriting the expression of equation 10.23 as

$$\begin{aligned}
\frac{\partial S_r}{\partial \beta} &= \sum_\sigma \sum_{\overline{\sigma}} P(\sigma,\overline{\sigma}) log(\sum_{\overline{\sigma}} P(\sigma,\overline{\sigma}))(E(\sigma,\overline{\sigma}) - \langle E \rangle) \\
&= \sum_\sigma \sum_{\overline{\sigma}} P(\sigma,\overline{\sigma}) log(P(\sigma))(E(\sigma,\overline{\sigma}) - \langle E \rangle) \\
&= \sum_\sigma \sum_{\overline{\sigma}} P(\sigma,\overline{\sigma}) log(P(\sigma,\overline{\sigma}))(E(\sigma,\overline{\sigma}) - \langle E \rangle) \\
&\quad - \sum_\sigma \sum_{\overline{\sigma}} P(\sigma,\overline{\sigma}) log(P(\overline{\sigma} \mid \sigma))(E(\sigma,\overline{\sigma}) - \langle E \rangle) \\
&= \frac{\partial S_E}{\partial \beta} - \sum_\sigma \sum_{\overline{\sigma}} P(\sigma,\overline{\sigma}) log(P(\overline{\sigma} \mid \sigma))(E(\sigma,\overline{\sigma}) - \langle E \rangle)
\end{aligned} \tag{10.26}$$

Consider the second term of equation 10.26. Is it always the same sign? The answer to this is no, as may be seen more clearly by expanding the energy difference as

$$\begin{aligned}
E(\sigma,\overline{\sigma}) - \langle E \rangle &= (E(\sigma,\overline{\sigma}) - \sum_{\overline{\sigma}} E(\sigma,\overline{\sigma}) P(\overline{\sigma} \mid \sigma)) \\
&\quad + (\sum_{\overline{\sigma}} E(\sigma,\overline{\sigma}) P(\overline{\sigma} \mid \sigma) - \langle E \rangle)
\end{aligned} \tag{10.27}$$

The first term of equation 10.27 substituted into the second term of equation 10.26 is immediately positive for the same reason that the derivative of

the full entropy is negative. However, the second term of equation 10.27 substituted into the second term of equation 10.26 is guaranteed neither positive nor negative. Nor does its magnitude appear easily bounded. See figures 10.1 and 10.2 for an example of a region where the derivative of the first order entropy changes sign while the derivative of the second order entropy does not. Note that the same may hold true in general for derivatives of the reduced entropies with respect to the other parameters, as figures 8.4 and 8.17 show clearly. Thus, although the reduced entropies have an ordering, the derivatives of them do not.

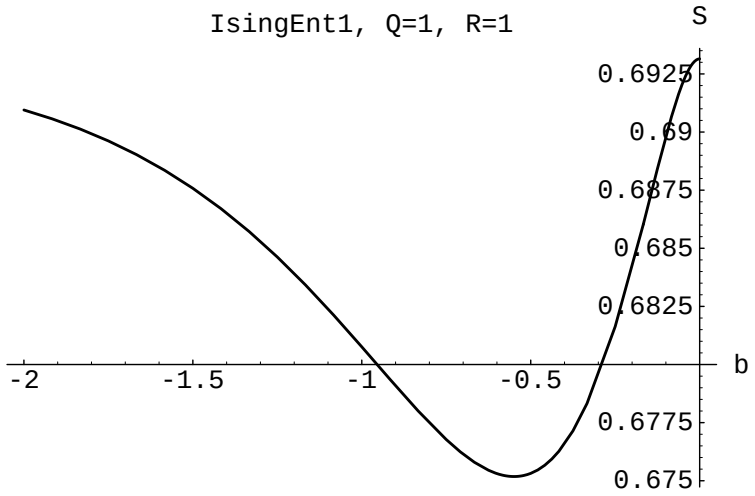

Figure 10.1: First order entropy of Ising system in cross section in the antiferromagnetic region. Note the change in the sign of the derivative with respect to β. Considering a neighboring pair of sites, this occurs as the temperature is lowered as the states $\uparrow\uparrow$, $\uparrow\downarrow$, $\downarrow\uparrow$, $\downarrow\downarrow$, which yield equal probabilities of $\uparrow$ and $\downarrow$ at a single site, get replaced by the smaller set of states $\uparrow\uparrow$, $\uparrow\downarrow$, $\downarrow\uparrow$, which yield a greater probability for $\uparrow$ than $\downarrow$ at a single site, and these in turn finally replaced by $\uparrow\downarrow$ and $\downarrow\uparrow$ alone, which again yield equal probabilities of $\uparrow$ and $\downarrow$ at a single site.

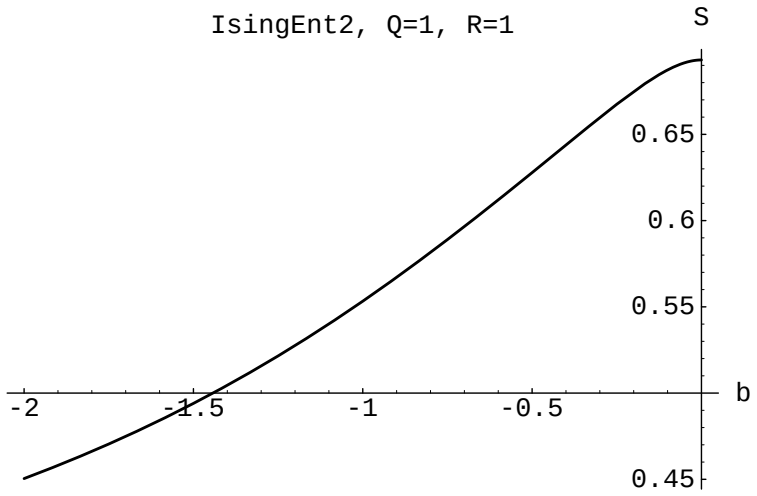

Figure 10.2: Second order entropy of Ising system in cross section in the antiferromagnetic region. Note that the derivative of the entropy with respect to β maintains the same sign.

Chapter 11

Conclusion

In this dissertation the topic of multivariable systems has been explored from the information theoretic point of view, with the focus on delineating the meaning of the information correlation functions.

The information correlation functions are the generalization of the entropy and mutual information. They have the interpretation of being the information between a set of random variables. They also may be seen in a certain approximation to be the information in the complete distribution that is not in the specified subset distributions. The connection between the multi-variable cumulant expansion, the Ursell functions, the linked cluster theorem and the information correlation function was explored. Here, all of these methods for approximating in various ways the moment generating function or density function were seen to be a consequence of a single theorem of structured partitions.

The dynamics of information flow in a reduced system of variables was described from both the classical and quantum points of view. The flow of information into a subsystem when the rest of the system is taken to be maximum entropy distributed was seen to be zero. It can be noted that the notion of dynamical irreversibility can be studied from the point of view of this analysis, although the notion of the increase of entropy from the point of view of ensembles of systems or the density matrix is itself in question: what is really being asked is why the second law holds for the particular instance of the universe

that we inhabit.

The Ising model and the Heisenberg long-range coupled spin systems were used as a testing area for the information correlation functions. In each case it was seen that the interesting behavior within these systems was due to high order correlations occurring, and that these were reflected in the fact that the high order information correlation functions. We are forced to conclude that the information correlation functions provide a powerful method for determining structure in the full distribution function that is not available from the lower order reduced distribution functions. There are several results concerning the entropies of the reduced spin systems. The first result is that in the Ising model at zero field, the reduced entropies all have a very simple analytical relationship: they are ordered by a factor independent of the system variables. This makes it possible to state that changes in the structure reflected in the entropies with respect to changes in any variable are bigger the higher the order of the entropy. The second result is that the information correlation functions for any set of spins on the Ising chain are given by the mutual information between the first and the last spin on the chain. This is to say that the information between this set of spins is given by the information between the first and last spins of the chain - the spins that interact with the environment that the set of spins is within. This result is generalized to other topologies. The third result, quite surprising, is that as the temperature *increases* the entropy of a subsystem may be *decreasing* while the entropy of the complete system is *increasing. Order may be created in a subsystem while it is being destroyed in the overall system.* The fourth result found concerns the relationship between the measurement entropy and the intrinsic $-Tr[\rho log(\rho)]$ entropy. The measurement entropy is larger. The development of structure in the reduced spin systems was noted. The ordering of the entropy per spin provides an easy way to compare the structures in a manner that is best described as emergent. The fifth result is that the entropy at sufficiently large order may both increase

and decrease as some system parameter is varied. This was seen clearly the external field was varied in both model systems studied.

The fact that quantum mechanics allows exactly one measurement of a particular state to occur without changing the state leads to an interesting quantity, the *mutual information between a measured variable and an unmeasured variable*. This result yields a measure proposed here as a quantity that should be considered when experiments are done: it measures how much of the information in the measurement is relevant to the desired information about the physical system. This should have some relevance to quantum computer systems proposed.

Information about a system is acquired from data, and from the data inferences of the underlying physical system are made. In the large data case many methods for determining the underlying system, including maximum likelihood methods, work reasonably well. Bayesian methodology is the tool of choice in the small data case. One new result is: *Information about the underlying system may decrease upon seeing new data, but on the average information about the underlying system will increase upon seeing new data.* The amount of the average increase in information upon seeing new data is quantified exactly as a Kullback-Leibler distance.

Estimation of the entropy and the mutual information functions is the subject of later sections of the chapter on estimation. Inferences of the values *and their uncertainties* of the functions of interest are developed in a consistent manner. Along with the estimators for the entropy and mutual information developed there, provably optimal for the specified priors, are available estimators for other functions such as chi-squared, covariance, and moments, correlations and cumulants of all orders.

Bibliography

[1] Daniel P. Aalberts and A. Nihat Berker. Spin-wave bound-state energies from an Ising model. *Phys. Rev. B*, 49(2):1073–1078, Jan 1994.

[2] Ryuzo Abe. *Statistical Mechanics*. University of Tokyo Press, Tokyo, 1975.

[3] Milton Abramowitz and Irene A. Stegun. *Handbook of Mathematical Functions*. Dover, New York, 1963.

[4] G. P. arin. On a statistical estimate for the entropy of a sequence of independent random variables. *Theory Prob. Appl.*, 4(3):333–336, 1959.

[5] R. Balescu. *Non-equilibrium statistical mechanics*, chapter 5, pages 209–271. In Biel and Rae [9], 1972.

[6] R. J. Baxter. *Exactly Solved Models in Statistical Mechanics*. Academic Press, London, 1982.

[7] R. J. Baxter. *Exactly Solved Models in Statistical Mechanics*, chapter 1, page 1. Volume 1 of D'Ariano et al. [17], 5–64.

[8] Shahar Ben-Menahem. Spin-measurement retrodiction. *Phys. Rev. A*, 39(4):1621–1627, Feb 1989.

[9] J. Biel and J. Rae, editors. *Irreversibility in the Many-Body Problem*. Sitges International School of Physics, May, 1972. Plenum, New York, 1972.

[10] K. Binder and A. P. Young. Spin Glasses: Experimental facts, theoretical concepts, and open questions. *Rev. Mod. Phys.*, 58(4):801–976, Oct 1986.

[11] J. De Boer and G. Uhlenbeck, editors. *Studies in Statistical Mechanics, Vol. 1.* Interscience Publishers, New York, 1961.

[12] N. Bogoliubov. *Problems of Dynamical Theory in Statistical Physics*, chapter 1, pages 5–118. In Boer and Uhlenbeck [11], 1961.

[13] R. Camassa, D. D. Holm, and J. M. Hyman. *Adv. Appl. Mech.*, 31:1, 1994.

[14] H. S. Carslaw and J. C. Jaeger. *Operational Methods in Applied Mathematics.* Oxford University Press, London, England, 1953.

[15] Peter V. Coveney. The second law of thermodynamics: entropy, irreversibility, and dynamics. *Nature*, 333(2):409–415, Jun 1988.

[16] Thomas M. Cover and Joy A. Thomas. *Elements of Information Theory.* John Wiley and Sons, New York, 1991.

[17] G. M. D'Ariano, A. Montorosi, and M. G. Rasetti, editors. *Integrable Systems in Statistical Mechanics*, volume 1 of *Series on Advances in Statistical Mechanics.* Institute for Scientific Interchange, Torino, Italy, 1985.

[18] Morris H. DeGroot. *Probability and Statistics.* Addison Wesley, Reading, MA, 1984.

[19] Richard O. Duda. *Pattern Classification and Scene Analysis.* Wiley, New York, 1973.

[20] A. Elkouraychi, M. Saber, and J. W. Tucker. On the theory of the transverse Ising model with arbitrary spin. *Physica A*, 213:576–586, 1995.

[21] S. Eubank and D. Farmer. *An Introduction to Chaos and Randomness.* 1989.

[22] R. M. Fagen. Information measures: Statistical confidence limits and inference. *J. Theor. Biol.*, 73:61–79, 1978.

[23] J. E. G. Farina. An elementary approach to quantum probability. *American Journal of Physics*, 61(5):466–468, May 1993.

[24] K. H. Fischer and J. A. Hertz. *Spin Glasses.* Cambridge University Press, Cambridge, UK, 1991.

[25] Daniel S. Fisher, Jeffrey M. Grinstein, and Anil Khurana. Theory of Random Magnets. *Physics Today*, pages 56–67, Dec 1988.

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

[27] C. W. Gardiner. *Handbook of Stochastic Methods.* Springer-Verlag, New York, 1990.

[28] Apostolos Gerasoulis. On the granularity and clustering of directed acyclic task graphs. *IEEE Trans. Par. Dist. Sys.*, 4(6):686–702, Jun 1993.

[29] R. J. Glauber. *J. Math. Phys.*, 4:294, 1963.

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

[31] P. Grassberger. Finite sample corrections to entropy and dimension estimates. *Phys. Lett. A*, 128:369–373, 1988.

[32] B. Harris. *Colloq. Math. Societatis Janos Bolyai*, 16:323–355, 1975.

[33] H. Herzel. *Syst. Anal. Model. Simul.*, 5:435–444, 1988.

[34] H. Herzel, A. O. Schmitt, and W. Ebeling. Finite sample effects in sequence analysis. *Tech. Report.*

[35] Melvin J. Hinich. Higher Order Cumulants and Cumulant Spectra. *Circ. Syst. Sig. Proc.*, 13(4):391–402, 1994.

[36] D. L. Holl, C. G. Maple, and B. Vinograde. *Introduction to the Laplace Transform.* Aplleton, New York, 1959.

[37] Henry B. Hollinger and Michael John Zenzen. *The Nature of Irreversibility.* D. Reidel Pub. Co., Dordrecht, 1985.

[38] Kerson Huang. *Statistical Mechanics.* John-Wiley and Sons, New York, 1987.

[39] Setsuo Ichimaru. *Basic Principles of Plasma Physics. A Statistical Approach.* Benjamin/Cummings, Reading, MA, 1973.

[40] John David Jackson. *Classical Electrodynamics.* Wiley, New York, 1975.

[41] E. Jen, editor. *1989 Lectures in Complex Systems, SFI Studies in the Sciences of Complexity, Lect. Vol. II*, volume II of *SFI Studies in the Sciences of Complexity.* Addison-Wesley, 1990.

[42] Henrik Jeldtoft Jensen. $1/f$ Noise from diffusion in Lattice Gases. *Mod. Phys. Lett. B*, 5(9):625–634, 1991.

[43] Norman L. Johnson and Samuel Kotz. *Discrete Distributions.* John Wiley and Sons, New York, 1969.

[44] Shigetoshi Katsura. Statistical Mechanics of the Anisotropic Linear Heisenberg Model. *Phys. Rev.*, 127(5):1508–1518, Dec 1962.

[45] Scott Kirkpatrick and David Sherrington. Infinite-ranged models of spin-glasses. *Phys. Rev. B*, 17(11):4384–4403, Jun 1978.

[46] Y. Kishimoto, T. Tajima, D. L. Fisher, and K. Mima. *Cooling and Phase Control by Using laser-Undulator Beatwave.* In Tajima [85], 1996.

[47] Charles Kittel. *Introduction to Solid State Physics.* John Wiley and Sons, New York, fifth edition, 1976.

[48] Charles Kittel. *Quantum Theory of Solids.* John Wiley and Sons, New York, 2nd edition, 1987.

[49] A.N. Kolmogorov and S.V. Fomin. *Measure, Lebesgue Integrals, and Hilbert Space.* Academic Press, New York, 1961.

[50] N. N. Lebedev. *Special Functions and their Applications.* Dover - Republication of the translation by Richard A. Silverman. Prentice, 1965, New York, 1972.

[51] T. Leonard and J. Hsu. Bayesian inference for a covariance matrix. *Ann. Stat.*, 20:1669–1696, 1992.

[52] W. Li. Mutual information functions versus correlation functions. *J. Stat. Phys.*, 60:823–837, 1990.

[53] Seth Lloyd. Use of mutual information to decrease entropy: Implications for the second law of thermodynamics. *Phys. Rev. A*, 39(10):5378–5386, 1989.

[54] Jon Mathews and R. L. Walker. *Mathematical Methods of Physics.* Benjamin/Cummings, Reading, MA, 1970.

[55] Richard D. Mattuck. *A guide to Feynman diagrams in the many-body problem.* McGraw-Hill, New York, 2nd edition, 1976.

[56] Barry M. McCoy, Jacques H. H. Perk, and Robert E. Schrock. Time-dependent correlation functions of the transverse Ising chain at the critical magnetic field. *Nucl. Phys.*, B220([FS8]):35–47, 1983.

[57] Donald McQuarrie. *Statistical Mechanics.* Harper, 1973.

[58] Eugen Merzbacher. *Quantum Mechanics.* John Wiley and Sons, New York, 2nd edition, 1970.

[59] Albert Messiah. *Quantum Mechanics*, volume II. Wiley, New York, 1976.

[60] Marc Mezard, Giorgio Parisi, and Miguel Angel Virasoro. *Spin Glass Theory and Beyond.* World Scientific, New Jersey, 1987.

[61] G. A. Miller. *Note on the Bias of Information Estimates*, pages 95–100. 1955.

[62] Franz Mohling. *Statistical Mechanics, Methods and Applications.* John Wiley and Sons, New York, 1982.

[63] F. C. Moon. *Chaotic Vibrations.* Wiley, New York, 1987.

[64] Gerhard Muller and Robert E. Shrock. Dynamical correlation functions for one-dimensional quantum spin systems: New results based on a rigorous approach. *J. Appl. Phys.*, 55(6):1874–1876, Mar 1984.

[65] Chrysostomos L. Nikias and Jerry M. Mendel. Signal processing with higher-order spectra. *IEEE Signal Processing*, pages 10–37, Jul 1993.

[66] 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.

[67] V.S. Pugachev. *Probability Theory and Mathematical Statistics for Engineers.* Pergamon, New York, 1984.

[68] H. Quastler, editor. *Information Theory in Psychology.* Free Press, Glencoe, IL, 1955.

[69] L. E. Reichl. *A Modern Course in Statistical Physics.* University of Texas Press, Austin, Texas, 1980.

[70] H. Reiger and A. P. Young. Zero-Temperature Quantum Phase Transition of a Two-Dimensional Ising Spin Glass. *Phys. Rev. Lett.*, 27(26):4141–4144, Jun 1994.

[71] Fazlollah M. Reza. *An Introduction to Information Theory.* Dover, New York, 1961.

[72] M. A. Ruderman and C. Kittel. *Phys. Rev.*, 96:99, 1954.

[73] Walter Rudin. *Priciples of Mathematical Analysis.* McGraw-Hill, New York, 1976.

[74] Hidetsugu Sakaguchi. Renyi entropy and statistical mechanics. *Prog. Theor. Phys.*, 81(4):732–737, Apr 1989.

[75] Hidetsugu Sakaguchi. Kolmogorov-Sinai Entropy of the Ising Model. *Prog. Theor. Phys.*, 86(2):303–307, Aug 1991.

[76] J. J. Sakurai. *Modern Quantum Mechanics.* Benjamin/Cumings, Menlo Park, CA, 1985.

[77] A. O. Schmitt, H. Herzel, and W. Ebeling. A new method to calculate higher-order entropies from finite samples. *Europhysics Letters*, 23(5):303–309, 1993.

[78] Thomas D. Schneider. Theory of Molecular Machines. I. Channel Capacity of Molecular Machines. *J. Theor. Biol.*, 148:83–123, 1991.

[79] Thomas D. Schneider. Theory of Molecular Machines. II. Energy Dissipation from Molecular Machines. *J. Theor. Biol.*, 148:125–137, 1991.

[80] James B. Seaborn. *Hypergeometric functions and their applications*, volume 8 of *Texts in applied mathematics.* Springer-Verlag, New York, 1991.

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

[82] Barry Simon. Uniqueness of Phase Transition Temperature in Multidimensional Ising Models. *Physics Today*, pages s43–s44, Jan 86.

[83] H. Eugene Stanley. *Introduction to Phase Transitions and Critical Phenomenoa*. Oxford University Press, New York, 1971.

[84] T. Tajima. *Computational Plasma Physics with Applications to Fusion and Astrophysics*. Advanced Book Program. Addison-Wesley Pub. Co., Redwood City, CA, 1989.

[85] T. Tajima, editor. *The Future of Accelerator Physics: The Tamura Symposium Proceedings*. American Institute of Physics, New York, 1996.

[86] Minoru Takahashi. Analytical and Numerical Investigations of Spin Chains. *Prog. Theor. Phys.*, 91(1):1–15, Jan 1994.

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

[88] Lev Vaidman, Yakir Aharonov, and David Z. Albert. How to ascertain the values of σ_x, σ_y, and σ_z of a spin-$\frac{1}{2}$ particle. *Phys. Rev. Lett.*, 58(14):1385–1387, Apr 1987.

[89] P. van Ede van der Pals and P. Gaspard. Two-dimensional quantum spin Hamiltonians: Spectral properties. *Phys. Rev. E*, 49(1):79–98, Jan 1994.

[90] J. H. van Lint and R. M. Wilson. *A Course in Combinatorics*. Cambridge University Press, New York, 1992.

[91] Richard F. Voss. Evolution of Long-Range Fractal Correlations and $1/f$ Noise in DNA Base Sequences. *Phys. Rev. Lett.*, 68(25):3805–3808, Jun 1992.

[92] Gregory H. Wannier. *Statistical Physics.* Wiley, New York, 1966.

[93] Satoshi Watanabe. *Pattern Recognition: Human and Mechanical.* John Wiley and Sons, New York, 1985.

[94] Andreas S. Weigend and Neil A. Gershenfeld, editors. *Time Series Prediction: Forecasting the Future and Understanding the Past. Proceedings of the NATO Advanced Research Workshop on Comparative Time Series Analysis held in Santa Fe, New Mexico, May 14-17, 1992*, number XV in Santa Fe Institute Studies in the Sciences of Complexity. Addison-Wesley, 1994.

[95] D. R. Wolf and D. H. Wolpert. Estimating functions of probability distributions from finite samples, part 2: Mutual information, chi-squared. *Los Alamos Unclassified Report*, LA-UR-93-833(Send email to compgas@xyz.lanl.gov with subject "get 9403002"), 1993.

[96] D. H. Wolpert and D. R. Wolf. Estimating functions of probability distributions from finite samples. *Phys. Rev. E*, 52(6):6841–6854, Dec 1995.

[97] Hubert P. Yockey. *Information Theory and Molecular Biology.* Cambridge University Press, Cambridge, UK, 1992.

[98] Yi-Cheng Zhang. Complexity and $1/f$ noise. A phase space approach. *J. Phys. I France*, 1:971–977, Jul 1991.

Appendix

Appendix A

Clebsch-Gordon transformation of spin bases

A.1 Energy eigenbasis.

The coupled spin energy eigenbasis is generated in a direct manner with use of the following two theorems, both proven in [59].

Theorem 1. *Given two subsystems having total angular momenta s_1 and s_2 respectively, the system consisting of both subsystems may assume any total angular momentum value in the set* $\{s_1 + s_2, s_1 + s_2 - 1, \ldots, Abs(s_1 - s_2)\}$.

Theorem 2. *Given a system with total angular momentum value s, the z-component of the angular momentum may assume any value in the set* $\{s, s-1, \ldots, -s\}$.

The Mathematica shown in appendix B.1 recursively generates the coupled spin basis

$$\mid s_1, s_2, s_{12}, s_3, s_{13}, \ldots, s_{1N}, m\rangle \tag{A.1}$$

where the s_i are the spins of the individual particles, the s_{1k} are the total angular momentum values of the system consisting of particles 1 thru k, and m is the angular momentum z-component of the system of all the particles.

A.2 Measurement basis. Energy to measurement basis transformation.

The measurements of the z-components of the spins of the individual particles are made most easily when these energy eigenvectors are transformed to the basis $\mid (s_i, m_i)\rangle$. Doing this involves the use of transformation coefficients called Clebsch-Gordon coefficients, which are real, are the overlaps in the expansion of an energy eigenvector in terms of (S_{iz}) eigenvectors

$$\mid s_1, s_2, s_{12}, m\rangle = \sum_{m_1 m_2} \mid s_1, m_1, s_2, m_2\rangle\langle s_1, m_1, s_2, m_2 \mid s_1, s_2, s_{12}, m\rangle \qquad \text{(A.2)}$$

Defining $cg(s_1, m_1, s_2, m_2; s_{12}, m)$ as this coefficient, Racah demonstrated

$$\begin{aligned} & cg(s_1, m_1, s_2, m_2; s_{12}, m) = \\ & \sqrt{2s_{12}+1} \times \Delta_1(s_1, s_2, s_{12}) \times \Delta_2(s_1, m_1)\Delta_2(s_2, m_2)\Delta_2(s_{12}, m) \\ & \times \sum_t \frac{(-1)^t}{t!}[\ \Delta_3(s_2, m_1)\ \Delta_3(s_1, -m_2) \times (s_1 + s_2 - s_{12} - t)! \\ & \qquad \times \Delta_4(s_1, -m_1)\Delta_4(s_2, m_2)\ ]^{-1} \end{aligned} \qquad \text{(A.3)}$$

where

$$\Delta_1(a, b, c) = \sqrt{\frac{(a+b-c)!(b+c-a)!(c+a-b)!}{(a+b+c+1)!}} \qquad \text{(A.4)}$$

$$\Delta_2(a, b) = \sqrt{(a+b)!(a-b)!} \qquad \text{(A.5)}$$

$$\Delta_3(a, b) = \sqrt{(s_{12}+t-a+b)!} \qquad \text{(A.6)}$$

$$\Delta_4(a, b) = \sqrt{(a+b-t)!} \qquad \text{(A.7)}$$

The proedure used to transform the energy eigenbasis vectors to the z-component eigenbasis vactors is given in appendix B, section B.2. In this routine, each energy eigenvector is expanded recursively in terms of the S_{iz} eigenvectors, using equation A.2 repeatedly.

Appendix B

Mathematica for the Heisenberg spin system

B.1 Generating the coupled spin basis

The basis vectors of the coupled spin (energy eigenstate) basis $\mid s_1 s_2 s_{12} \ldots sm\rangle$ are generated using **genVecs**[].

```
genVecs[ {rep_,order_,scale_} ] :=
         Block[ { a={},n,j,m,q,newrep },
                n=Length[rep]-order;
                q=2(order+1);
                If[ order<(n-1) ,
                    For[ j=Abs[rep[[q]]-rep[[q-1]]],
                         j<=(rep[[q]]+rep[[q-1]]),
                         j++,
                         newrep=Insert[rep,j,q+1] ;
                         a=Join[a,
                           genVecs[{newrep,order+1,scale}]] ;
                       ] ;
                  ] ;
                If[ order==(n-1) ,
                    For[ m=-rep[[q-1]], m<=rep[[q-1]], m++,
                         a=Join[a,{{Join[rep,{m}],n-1,scale}}]
                       ] ;
                  ] ;
                Return[a] ;
              ]
```

The arguments of *genVecs* are *rep*, a list of the spin values, *order*, the order of the representation given, usually 0, and *scale*, the value of the

amplitude to be assigned to the vectors generated. For example,

$$genVecs[\{\{\frac{1}{2},\frac{1}{2}\},0,1\}] \tag{B.1}$$

produces the list of four vectors with their order and scale.

$$\{\{\{\frac{1}{2},\frac{1}{2},0,0\},1,1\},\{\{\frac{1}{2},\frac{1}{2},1,-1\},1,1\},\{\{\frac{1}{2},\frac{1}{2},1,0\},1,1\},\{\{\frac{1}{2},\frac{1}{2},1,1\},1,1\}\} \tag{B.2}$$

Here the order has become 1 because one coupling has occurred between the two spins given.

B.2 Transforming to the measurement basis

The coupled-spin basis vectors are transformed from the energy eigenbasis to the (S_{iz}) basis using **transVec[].**

```
transVec[{rep_,order_,scale_}] :=
  Block[ { vec={},n=Length[rep],q,j1,m1,j2,m2,j12,m12,newrep } ,
         q = 2 order ;
         If[ order!=0 ,
             j1  = rep[[q-1]] ;
             j2  = rep[[q]]   ;
             j12 = rep[[q+1]] ;
             m12 = rep[[q+2]] ;
             For[ m1=-j1,m1<=j1,m1++,
                  m2=m12-m1 ;
                  If[ ((-j2<=m2) && (m2<=j2)),
                      newrep=Delete[rep,q+1] ;
                      newrep=Delete[newrep,q+1] ;
                      newrep=Insert[newrep,m1,q] ;
                      newrep=Insert[newrep,m2,q+2] ;
                      vec=Join[vec,
                      transVec[{newrep,
                      order-1,
                      scale*cg[j1,j2,m1,m2,j12,m12]}] ] ;
                    ] ;
```

```
            ] ; ,
          Return[{{rep,order,scale}}] ;
        ] ;
      Return[vec] ;
    ] ;
```

The argument of $transVec$ is a length-three list consisting of rep, the representation of the basis vector in the basis produced by $genVecs$, $order$, the order of the representation given, usually $N-1$, where N is the number of spins (note that $transVec$ is recursive, $order$ keeps track of the depth), and $scale$, the value of the normalization of the vector generated. For example,

$$transVec[\{\{\frac{1}{2},\frac{1}{2},1,\frac{1}{2},\frac{1}{2},\frac{1}{2}\},2,1\}] \tag{B.3}$$

transforms the vector $\mid s_1 = 1/2, s_2 = 1/2, s_{12} = 1, s_3 = 1/2, s = 1/2, m = 1/2\rangle$ to the vector

$$\begin{aligned}&\{\{\{\frac{1}{2},-\frac{1}{2},\frac{1}{2},\frac{1}{2},\frac{1}{2},\frac{1}{2}\},0,-\frac{1}{\sqrt{6}}\},\\ &\{\{\frac{1}{2},\frac{1}{2},\frac{1}{2},-\frac{1}{2},\frac{1}{2},\frac{1}{2}\},0,-\frac{1}{\sqrt{6}}\},\\ &\{\{\frac{1}{2},\frac{1}{2},\frac{1}{2},\frac{1}{2},\frac{1}{2},-\frac{1}{2}\},0,\sqrt{\frac{2}{3}}\}\}\end{aligned} \tag{B.4}$$

which is a list of three lists, each of these being a vector in the $(s_i m_i)$ basis, along with an order and scale indicator, and the whole list to be interpreted as the sum of these vectors.

Along with $transVec$, the routine $transVecs$ transforms a list of many vectors, such as would be produced by the output of $genVecs$.

```
transVecs[x_] :=
  Block[ {n,vecs={},i} ,
         n=Length[x] ;
         For[ i=1,i<=n,i++,
```

```
        vecs=Join[vecs,{transVec[x[[i]]]}] ;
      ] ;
    Return[vecs] ;
  ] ;
```

For example

$$transVecs[genvecs[\{\{\frac{1}{2},\frac{1}{2}\},0,1\}]] \tag{B.5}$$

produces the output

$$\begin{aligned}\{\{\{\{\frac{1}{2},-\frac{1}{2},\frac{1}{2},\frac{1}{2}\},0,-\frac{1}{\sqrt{2}}\},\{\{\frac{1}{2},\frac{1}{2},\frac{1}{2},-\frac{1}{2}\},0,\frac{1}{\sqrt{2}}\}\},\\ \{\{\{\frac{1}{2},-\frac{1}{2},\frac{1}{2},-\frac{1}{2}\},0,1\}\},\\ \{\{\{\frac{1}{2},-\frac{1}{2},\frac{1}{2},\frac{1}{2}\},0,\frac{1}{\sqrt{2}}\},\{\{\frac{1}{2},\frac{1}{2},\frac{1}{2},-\frac{1}{2}\},0,\frac{1}{\sqrt{2}}\}\},\\ \{\{\{\frac{1}{2},\frac{1}{2},\frac{1}{2},\frac{1}{2}\},0,1\}\}\}\end{aligned} \tag{B.6}$$

which is a list of the four vectors in the (s_i, m_i) basis corresponding to the four vectors in the coupled spin basis produced by the $genVecs$ command in equation B.1.

B.3 Finding the coupled spin basis probability distribution

Finding the probabilities of the energy eigenstates is a matter of utilizing the quantum numbers in the eigenvectors in the coupled spin basis to find the energy of each, and using the Boltzmann distribution to assign the probability. This is done using the routine **getProbabilities**[].

```
getProbabilities[x_] :=
  Block[ {i,j,numVecs,y,rep,order,energy,
          e={},relP,absP=0,p={}} ,
```

```
numVecs=Length[x] ;
For[ i=1,i<=numVecs,i++,
     y=x[[i]] ;
     order=y[[2]] ;
     rep=y[[1]] ;
     energy=rep[[2 order+1]] (rep[[2 order+1]]+1) ;
     energy-=rep[[1]] (rep[[1]]+1) ;
     For[ j=1,j<=order,j++,
          energy-= rep[[2 j]] (rep[[2 j]]+1) ;
        ] ;
     e=Join[e,{energy}] ;
     relP=Exp[Q/2 b energy] *
          Exp[R b rep[[2 (order+1)]] ];
     absP+=relP ;
     p=Join[p,{relP}] ;
   ] ;
  Return[{absP,p}] ;
] ;
```

The argument of $getProbabilites$ is a list of vectors in the coupled spin basis, like that produced by $genVecs$. The output is a two element list, the first element being the normalization for the second element, which is proportional by the first element to the list of probabilies of the vectors input. For example, the command

$$getProbabilities[genVecs[\{\{\frac{1}{2},\frac{1}{2}\},0,1\}]] \tag{B.7}$$

produces the output

$$\{e^{\frac{-3\,b\,Q}{4}}+e^{\frac{b\,Q}{4}}+e^{\frac{b\,Q}{4}-b\,R}+e^{\frac{b\,Q}{4}+b\,R},\{e^{\frac{-3\,b\,Q}{4}},e^{\frac{b\,Q}{4}-b\,R},e^{\frac{b\,Q}{4}},e^{\frac{b\,Q}{4}+b\,R}\}\} \tag{B.8}$$

Note that the first element is the sum of the elements of the second element.

B.4 Distribution of measured states

The distribution of the measured states of (S_{iz}) is found using **getDistribution[].**

```
getDistribution[order_,vecs_,probs_] :=
  (* finds the distribution in the +/- basis *)
  Block[ {tbl,dist,numVecs,numBasis,phase,probVal,vecProb,
          index,i,j,k,vec} ,
          dist=Table[0,{2^order}] ;
          numVecs=Length[vecs] ;
          For[ i=1,i<=numVecs,i++,
               vec=vecs[[i]] ;
               numBasis=Length[vec] ;
               vecProb=probs[[2]][[i]]/probs[[1]] ;
               For[ j=1,j<=numBasis,j++,
                    phase=vec[[j]][[3]] ;
                    probVal=phase*phase ;
                    index=0 ;
                    For[ k=0,k<order,k++,
                         index=index+2^k*
                           If[vec[[j]][[1]][[2(k+1)]]<0,0,1] ;
                       ] ;
                    dist[[index+1]]+=probVal*vecProb ;
                  ] ;
             ] ;
          Return[dist] ;
       ] ;
```

The arguments are *order*, which must be less than or equal to the number of spins, a list of vectors *vecs* in the (s_i, m_i) basis like that produced by $transVecs$, and *probs*, which is the list of probabilites produced by the output of $getProbabilites$ of the energy eigenvectors produced by $genVecs$.

The same distribution of the measured states of (S_{iz}) can be produced in polynomial form using **getDistributionPoly[].**

```
getDistributionPoly[order_,vecs_,probs_] :=
  Block[ {tbl,dist,numVecs,numBasis,phase,probVal,vecProb,poly,
          spin,i,j,k,vec},
         numVecs=Length[vecs] ;
         dist=0 ;
         For[ i=1,i<=numVecs,i++,
```

```
        vec=vecs[[i]] ;
        numBasis=Length[vec] ;
        vecProb=probs[[2]][[i]]/probs[[1]] ;
        For[ j=1,j<=numBasis,j++,
             phase=vec[[j]][[3]] ;
             probVal=phase*phase ;
             poly=1 ;
             For[ k=1,k<=order,k++,
                  spin=vec[[j]][[1]][[2k]] ;
                  poly*=If[spin<0,(1/2-x[k]),(1/2+x[k])] ;
                ] ;
             dist+=probVal*vecProb*poly ;
           ] ;
      ] ;
  Return[Simplify[dist]] ;
] ;
```

The output is a polynomial in *order* variables, specific to spin one-half for the routine above. Inserting a vector of spins in the resulting expression produces the probability of that measurement.