Information resonance and pattern recognition in classical and quantum systems: toward a 'language model' of hierarchical neural structure and process

Rodrick Wallace, Ph.D.
The New York State Psychiatric Institute
and
PISCS Inc.*

May, 2000

Abstract

Recent applications of the Shannon-McMillan Theorem to arrays of nonlinear components undergoing what is effectively an 'information resonance' (R Wallace, 2000a) may be extended to include many neural models, both classical and quantum. Some consideration reduces the threefold interacting complex of sensory activity, ongoing activity, and nonlinear oscillator to a single object, a parametized ergodic information source. Invocation of the 'large deviations' program of applied probability that unifies treatment of dynamical fluctuations, statistical mechanics, and information theory allows a 'natural' transfer of thermodynamic and renormalization arguments from statistical physics to information theory, permitting a markedly simplified analysis of neural dynamics. This suggests an inherent language-based foundation, in a large sense, to neural structure and process, and implies that approaches without intimate relation to language may be seriously incomplete.

Key Words: Coevolution, information resonance, information theory, large deviations, multitasking, neural networks, Onsager relations, phase transition, quantum neural networks, renormalization, state space algebra, stochastic resonance.

Introduction

Researchers have begun to adopt an information-theory approach to the study of stochastic resonance, usually involving the maximization of noise-dependent 'mutual information' between input and output (Deco and Schurmann, 1998, Heneghan et al, 1996, Godivier and Chapeau-Blondeau, 1998, Nieman et al., 1996). Similarly, other groups are examining the neural code and neural networks from an information theory viewpoint, as is summarized by Rieke et al. (1997) and Deco and Obradovic (1996) respectively.

Recently Wallace (2000a, b) and Wallace and Fullilove (1999) effectively invoked an 'information resonance' combining both stochastic resonance and neural models under the intellectual umbrella of the 'large deviations program' of applied probability (e.g. Dembo and Zeitouni, 1998), permitting the transfer of renormalization methods and thermodynamic formalism from statistical mechanics to information theory as an expression of underlying 'architecture.'

Here we will describe these results as they apply to hierarchical neural structures and outline a research agenda based on the questions this effort raises.

*Address correspondence to R Wallace, PISCS Inc., 549 W. 123 St., Suite 16F, New York, NY, 10027. Tel. (212) 865-4766, rdwall@ix.netcom.com. Affiliations are for identification only.

The general context is given by Schurmann in his forward to the recent book by Deco and Obradovic (1996, p. vii). Schurmann writes:

> "Chronological milestones in the history of artificial neural networks are Hebb's book on the organization of behavior, Rosenblatt's book on principles of neurodynamics in which he defines the perceptrons, Hopfield's discovery of the analogy of certain types of neural networks to spin glasses and the exploitation of the associated energy function, the generalization of simple perceptrons to feedforward multi-layer perceptrons accompanied by the backpropagation learning algorithm of Rumelhart and others and its extension to multi-layer perceptrons with feedback accompanied by the recurrent backpropagation learning algorithm of Almeida, Pineda and others."

Schurmann sees the application of information theory methods as a next natural step in neural network theory, and finds "particularly high potential" if information theory treatments are explicitly "linked with the methods of nonlinear dynamics... [which] remains a topic for future research..."

Below we will outline such a linkage. Several particulars, however, distinguish our approach:

First, we attempt a draconian simplification, seeking to employ information theory concepts only as they directly relate to the basic limit theorems of the subject. That is, message uncertainty and information source uncertainty are interesting *only because they obey the Coding and Source Coding Theorems.* 'Information Theory' treatments which do not sufficiently center on these theorems are, from our view, off the mark. From this perspective most discussion of 'complexity,' 'entropy maximization,' other definitions of 'entropy,' and so forth, just does not appear on the horizon. In the words of William of Occam, "Entities ought not be multiplied without necessity."

The second matter is more complicated: Rojdestvenski and Cottam (2000, p.44), following Wallace and Wallace (1998), see the linkage between information theory and statistical mechanics in quite general terms as involving

> "...[Homological] mapping... between ... unrelated ... problems that share the same mathematical basis... [whose] similarities in mathematical formalisms... become powerful tools for [solving]... traditional problems."

We believe the relation of information theory to neural structure and process to be somewhat more sharply constrained, revolving about two homologies, in the above sense:

(1) a 'linguistic' equipartition of probable paths consistent with the Shannon-McMillan Theorem which serves as the formal connection with nonlinear mechanics and large fluctuation theory, and

(2) a correspondence between information source uncertainty and statistical mechanical free energy, not statistical mechanical entropy. Indeed, Bennett (1988), among others, long ago realized that "...[T]he value of a message is the amount of...work plausibly done by its originator, which the receiver is saved from having to repeat."

We discuss the first point in more detail below, and will invoke the second for exploration of the connection between neural architecture and dynamics to obtain deep results in a relatively elementary manner.

Information resonance

The central idea of stochastic resonance is that the addition of 'noise' to a weak input, usually taken as a sinusoidal or other repeated train of excitations, can raise the amplitude of the combined signal so as to exceed the triggering threshold of a powerful nonlinear oscillator, resulting in an amplified output train. Proper choice of noise amplitude can maximize the signal-to-noise ratio of the combined system (e.g. Gammaitoni et al., 1998): too little noise fails to reach threshold, while too much washes out the signal.

Using the 'prehistory probability density' concept (McClintock and Luchinsky, 1999) as the critical starting point, we have carried out a somewhat expanded development (R. Wallace, 2000a, b) leading to the more general concept of an 'information resonance.'

We consider a generalized ongoing activity 'noise,' which may in fact be very highly structured, and a sensory activity 'signal' mixed together, in some possibly complicated manner, to produce a compound intermediate which is then sent into the nonlinear oscillator to produce an amplified output.

Figure 1 of Wallace (2000a) shows this two-step process: The 'signal' and 'noise' are convoluted to produce a sequence of discrete states a_i, where i is a non-negative integer. A relatively small number of sequential patterns of these convoluted states having the form $a_0, a_1, ...a_n$, which we call paths x, lead to discontinuous observable events, a generalized 'information resonance' analogous to the enhanced clicking of a switch – the nonlinear oscillator – by a weak signal in the presence of noise.

That is, each path x has associated with it a discontinuous function $h(x)$ taking possible values $0, 1$. $h(x) = 1$ represents the triggering of the oscillator by the path x, $h(x) = 0$ implies that x did not trigger the oscillator.

The definition can be extended, under proper conditions, to a stochastic system in which $h(x)$ is the probability that the nonlinear oscillator fires, provided a disjunction can be made between paths x which have high and low probabilities of triggering the oscillator.

We make an application to the stochastic neuron: A series of inputs $y_i^j, i = 1...m$ from m nearby neurons at time j is convoluted with 'weights' $w_i^j, i = 1...m$, using an inner product (e.g. Deco and Obradovic, 1996, p. 24)

$$a_j = \mathbf{y^j} \cdot \mathbf{w^j} = \sum_{i=1}^{m} y_i^j w_i^j, \tag{1}$$

in the context of a 'transfer function' $f(\mathbf{y^j} \cdot \mathbf{w^j})$ such that the probability of the neuron firing and having a discrete output $z^j = 1$ is

$$P(z^j = 1) = f(\mathbf{y^j} \cdot \mathbf{w^j}).$$

The probability the neuron does not fire at time j is thus

$$P(z^j = 0) = 1 - P(z^j = 1) = 1 - f(\mathbf{y^j} \cdot \mathbf{w^j}).$$

From our viewpoint the m values y_i^j constitute the 'sensory activity' and the m weights w_i^j the 'ongoing activity' at time j, with $a_j = \mathbf{y^j} \cdot \mathbf{w^j}$ and $x = a_0, a_1, ...a_n$. The a_j will almost always be serially correlated, e.g. 'integrate-and-fire.'

It would appear that many neural models fall under what we have called an information resonance, although extension of the concept to address architecture and learning paradigms requires some work.

Given a fixed initial state a_0 such that $h(a_0) = 0$ we examine all possible subsequent paths x beginning with a_0 and leading exactly once to the event $h(x) = 1$. Thus $h(a_0, ..a_j) = 0$ for all $j < m$ but $h(a_0, ...a_m) = 1$.

For each positive integer n let $N(n)$ be the number of paths of length n which begin with some fixed a_0 having $h(a_0) = 0$, and lead to the condition $h = 1$, or, in a stochastic system, to a state a_n with a very high probability of firing the oscillator. We shall call such paths 'meaningful.' In general we assume $N(n)$ to be considerably less than the number of all possible paths of length n – information resonance transitions are comparatively rare – and in particular assume that the finite limit

$$H = \lim_{n \to \infty} \frac{\log[N(n)]}{n} \tag{2}$$

exists and is independent of the path x.

Thus meaningful paths are asymptotically equiprobable, with each of length n having probability $P(n) \propto \exp(-nH)$.

We shall, in accordance with the standard treatment of information theory (e.g. Khinchine, 1957), call an information resonance satisfying this condition *ergodic*.

It seems likely that the underlying space defined by the a_i can be partitioned into disjoint equivalence classes according to whether states can be connected by meaningful paths. This would be analogous to a partition into 'domains of attraction' for a chaotic system, and indeed much of the subsequent development, including questions of hierarchical structure, can be rephrased in terms of the algebraic structure of that space, a matter to which we will return repeatedly.

Such state space partitioning implies the possibility of a 'neural multitasking' where nonlinear oscillators participate in disjoint but (nearly) simultaneous information resonance processes. Imposition of an inverse algebraic relation results in a group of finite order. This implies the possibility of even more complicated multitasking structure, since the number of different groups of finite order is powerfully determined by the prime number factoring of that order. We can envision, then, not only a number of disjoint multitasking processes corresponding to a given group's equivalence class structure, but determined in total by the full number of different groups corresponding to that order.

By the Asymptotic Equipartition Theorem, otherwise known as the Shannon-McMillan Theorem (SMT), for a certain class of information resonances, as we have defined them, there will be an ergodic information source $\mathbf{X}$ associated with stochastic variates X_i taking the values a_i with joint and conditional probabilities $P[a_0, ..., a_n]$ and $P[a_n|a_0, a_1, ...a_{n-1}]$ such that appropriate joint and conditional Shannon uncertainties may be defined satisfying the relations

$$H[\mathbf{X}] = \lim_{n\to\infty} \frac{\log[N(n)]}{n}$$

$$= \lim_{n\to\infty} H(X_n|X_0...X_{n-1})$$

$$= \lim_{n\to\infty} \frac{H(X_0, ...X_n)}{n+1}$$

(3)

where $H(X|Y)$ and $H(X,Y)$ represent, respectively, the *conditional* and *joint* uncertainties of the variates X and Y.

The joint uncertainty of stochastic variates X and Y, taking possible values x_i and y_i, is defined in terms of their joint probabilities as

$$H(X,Y) = -\sum_i \sum_j P(x_i, y_j) \log[P(x_i, y_j)].$$

(4)

The conditional uncertainty of X given Y is

$$H(X|Y) = -\sum_i \sum_j P(x_i, y_j) \log[P(y_j|x_i)].$$

(5)

See Khinchine (1957), pp. 117-120, Ash (1990) or Cover and Thomas (1991) for essential details.

We will define the information source $\mathbf{X}$, provided it exists, to be *dual* to the information resonance.

We have thus reduced three complex, synergistically interacting components – signal, noise and nonlinear oscillator – into a single object that can be appropriately parametized and on which we can impose important structure and symmetry.

The utility of this approach will become more apparent as we proceed.

Source uncertainty is a language function with an important heuristic interpretation (Ash, 1990, p. 206):

> "...[W]e may regard a portion of text in a particular language as being produced by an information source. The [conditional] probabilities $P[X_n = a_n|X_0 = a_0..., X_{n-1} = a_{n-1}]$ may be estimated from the available data about the language; in this way we can estimate the uncertainty associated with the language. A large uncertainty means... a large number of 'meaningful' sequences. Thus given two languages with uncertainties H_1 and H_2 respectively, if $H_1 > H_2$ then in the absence of noise it is easier to communicate in the first language; more can be said in the same amount of time. On the other hand, it will be easier to reconstruct a scrambled portion of text in the second language, since fewer of the possible sequences of length n are meaningful."

Languages are most fundamentally characterized by strict patterns of internal relationship, for example grammar, syntax, and higher levels of organization. Our development suggests that many information resonance phenomena, in this larger sense, are very highly structured and may be studied and perhaps predicted by understanding the 'metalanguage' in which they are embedded and indeed which they define.

According to this development, then, 'nonsense' paths $x = a_0, ..., a_n$ which violate the grammar and syntax of a particular information resonance cannot trigger it.

What we have done is, in the sense of Schurmann above, closely related to current research in nonlinear dynamics:

The condition $h(x) = 1$ represents, in this formulation, a 'large fluctuation' of the system in the sense of Dykman et al. (1996). To paraphrase that work, large fluctuations, although infrequent, are fundamental in a broad range of processes, and it was recognized by Onsager and Machlup (1953) that insight into the problem could be gained from studying the distribution of fluctuational paths along which the system moves to a given state. This distribution is a fundamental characteristic of the fluctuational dynamics, and its understanding leads toward control of fluctuations.

Fluctuational motion from the vicinity of a stable state may occur along different paths. For large fluctuations, the distribution of these paths peaks sharply along an optimal, i.e. most probable, path. In the theory of large fluctuations, the pattern of optimal paths plays a role similar to that of the phase portrait in nonlinear dynamics.

For our development the information-theoretic 'meaningful' statements $x = a_0, ..., a_n$ play the role of 'optimal' paths in the theory of large fluctuations, and we have given them an information theory treatment consistent with large deviation theory (Dembo and Zeitouni, 1998; Ellis, 1985).

The first real step in this direction was made more than sixty years ago by Cramer (1938). The ergodic theorem, in the context of recent generalizations of Cramer's results by Gartner and Ellis, permits derivation of the Shannon-McMillan Theorem as the 'zero error limit' under the rubric of 'rate distortion theory' (Dembo and Zeitouni, 1998). Our analysis suggests that neural phenomena and other generalizations may fit 'naturally' into this larger framework. See Appendix 1 for details of the most elementary large deviations argument.

Tuning an information resonance: learning paradigms and the Shannon Coding Theorem

Here we explore the extraordinary utility of an information resonance as a detector of subtle pattern. It is indeed this behavior which suggests characterization as a resonance. 'Learning paradigms' in neural networks evidently represent one systematic means of constructing such detectors, but application of a simple information theory argument appears to suggests room for improvement.

We now focus on the Shannon Coding Theorem, the other fundamental result of information theory, rather than on the Source Coding Theorem.

The properties of an information resonance combining 'signal,' ongoing activity 'noise,' convolution operation and nonlinear oscillators are the synergistic result of a subtle, multifactorial interaction. Effective control of such a system will likely involve a similarly synergistic tuning of more than one component.

In particular assume we have some complicated pattern represented by a stochastic variate of sensory input X, which we wish to detect. To reiterate, we feed that signal into a generalized information resonator by (1) convoluting it with the ongoing activity 'noise' which may itself be highly structured, and (2) feeding the combined result into a system of nonlinear oscillators, producing, under proper circumstances, an 'encoded' train of output spikes or more subtle coherent spatiotemporal patterns, which we characterize as a stochastic variate Y.

Proceeding somewhat schematically, let $H(X)$ be the Shannon uncertainty of the signal X, defined simply in terms of its probability distribution, $p_i = Pr[X = x_i]$ in the usual manner as

$$H(X) = -\sum_i p_i \log[p_i].$$

Let $H(X|Y)$ be the conditional uncertainty of X given the output pattern Y, again defined in terms of joint and conditional probabilities as

$$H(X|Y) = -\sum_i \sum_j p(x_i, y_j) \log[p(x_i|y_j)].$$

The information transmitted by the information resonance as a communication channel is, classically (e.g. Ash, 1990),

$$I(X|Y) \equiv H(X) - H(X|Y) = H(X) + H(Y) - H(X, Y), \tag{6}$$

where $H(X, Y)$ is the joint Shannon uncertainty of the stochastic variates X and Y.

Note that if there is no uncertainty in X given Y, then $H(X|Y) = 0$ and the information is transmitted without loss.

If we fix the ongoing activity 'noise,' convolution, and nonlinear oscillator properties, then we may vary the probability distribution of the signal variate X, which we write $P(X)$. The *capacity* of the channel is defined as

$$C \equiv \max_{P(X)} I(X|Y), \tag{7}$$

where we vary the probability distribution of X so as to maximize $I(X|Y)$.

The essential content of the Shannon Coding Theorem (Ash, 1990, Khinchine, 1957) is that for any rate of transmission of signals along the channel, R, such that $R < C$, it is possible to construct a 'coding scheme' for the signal X such that it may be sent with arbitrarily small error.

This is, in fact, one of the most singular and striking results of 20th Century applied mathematics, and a rigorous general proof is duly and suitably arduous (Khinchine, 1957).

The most direct application of the Coding Theorem to 'information resonance,' as we have characterized it, follows the arguments of Godivier and Chapeau-Blondeau (1998). We assume we have fixed the forms of the sensory input, the convolution and the nonlinear operator. Then variation of the parameters of the ongoing activity, i.e. the 'noise,' constitutes a kind of coding scheme such that the error-free throughput may be attained (or at least maximized) for any $R < C$ where C is the capacity of the information resonance as an information channel.

If the noise and convolution are simple, for example the additive white noise of simple stochastic resonance, there can be only a single noise parameter which can be tuned, its amplitude. Thus there will be a single amplitude which maximizes total throughput, i.e. mutual information or signal to noise ratio.

For neural structure of equation (1), obviously variation of the weight parameters, the w_i serves the same purpose as changing white noise amplitude in simple stochastic resonance. A systematic search for an optimum weight system constitutes the learning paradigm.

It may be possible to do considerably better if we generalize the tuning, i.e. if we allow both the convolution method and the properties of the nonlinear oscillator to vary along with the pattern of ongoing activity. Then we have more control over the 'coding scheme' and can more closely approach optimal coding, and hence error-free transmission at rates less than the channel capacity. Toward this end we invert, in a sense, the viewpoint.

The Shannon analysis traditionally has been used to study the transmission of a signal along a particular fixed channel, e.g. a noisy telephone line, optical waveguide, or the tenuous plasma through which a planetary probe transmits data.

Suppose we focus, rather than on the 'fixed' properties of the channel, on the signal X. Assume that signal is so important that we cannot vary it, but can, rather, *vary the probability distributions of the channel into which we encode that signal.* That is, we examine the *dual channel capacity*

$$C^* \equiv \max_{P(Y),P(Y|X)} I(X|Y). \tag{8}$$

But

$$C^* = \max_{P(Y),P(Y|X)} I(Y|X) \tag{9}$$

since $I(X|Y) = H(X) + H(Y) - H(X,Y) = I(Y|X)$. Thus, in a purely formal sense, *we may look at the message as transmitting the channel*, and there will be some channel distribution $P(Y)$ which maximizes C^* and, according to the Coding Theorem, some optimal coding scheme which permits attainment of error-free transmission at rates less than C^*.

This all becomes somewhat clearer when we examine the channel matrix $P(Y|X)$ more closely. The probability Y takes a value y_j, $P(y_j)$ may be written in terms of the probability X takes a value x_j and the joint probabilities $P(y_j, x_i)$ as

$$P(y_j) = \sum_i P(x_i)P(y_j|x_i). \tag{10}$$

Thus $P(Y)$ is defined by the channel matrix $P(Y|X)$ for a fixed distribution $P(X)$, and

$$C^* = \max_{P(Y),P(Y|X)} I(Y|X) = \max_{P(Y|X)} I(Y|X). \tag{11}$$

To calculate C^* in general we must maximize the complicated expression $H(X) + H(Y) - H(X,Y)$ which contains products of terms and their logs, subject to constraints that the sums of probabilities are 1 and each probability is itself between 0 and 1. For the case where the number of symbols in X and Y is the same, however, we can write the solution by inspection: Choosing

$$P(y_j|x_i) = \delta_{j,i}$$

where $\delta_{j,i} = 1$ when $i = j$ and 0 otherwise, gives

$$C^* = H(X).$$

The signal X is thus transformed into the coherently amplified pattern Y without error when the channel itself becomes 'typical' in the information-theoretic sense with respect to the fixed message probability distribution $P(X)$.

In reality, however, the best we can expect do is to adjust the structure of the ongoing activity 'noise,' the method of convolution, and the coupling strength or other parameters of the 'spatial' array of nonlinear oscillators and the properties of the oscillators themselves so as to maximize C^*.

This may often suffice:

Systems based on information resonance are inherently likely to be very sensitive and efficient detectors of subtle pattern when properly tuned. To be most fruitful, however, such tuning may necessitate far more than simply varying the neural weights w_i of equation (1), and will involve the systematic search for a more general efficient 'coding scheme' in the Shannon sense.

In conventional neural network terminology, the process of tuning, as we have described it, is a 'learning paradigm.' Taking a typical neural network configuration (e.g. a deterministic feed forward back propagation structure (Deco and Obradovic, p. 28), 'training' sessions compare actual with desired outputs for a series of fixed inputs. The weights w_i of equation (1) are varied in a systematic manner, e.g. mean square error minimization, until an acceptable error is attained. This process, as we described above, has meaning as a 'coding scheme' in conventional Shannon theory.

The development of this section suggests, however, that, since – formally – we may envision the fixed 'message' as transmitting the variable 'channel,' in addition the higher efficiency which may be attained by simply tuning more factors than just neural weights, there might well be a 'tuning theorem' converse to the Shannon Coding Theorem for appropriate input.

That is, under certain circumstances, for any fixed pattern X and a 'system rate' R^*, in some sense, such that $R^* < C^*$ there may exist an appropriate (and possibly elaborate) 'generalized tuning scheme' analogous to the optimal Shannon coding such that pattern detection can be carried out with very small error indeed – possibly zero error. For rates of input greater than C^* there may also be clever tuning schemes incorporating changes in 'noise' structure, the convolution method and the properties of the nonlinear oscillator which might reduce error to an absolute minimum. Transmission at rates above this minimum probably involves some kind of 'rate distortion' whose error one seeks to minimize in much the conventional manner.

At rates below that maximum, however, as is the case for the Shannon Coding Theorem, the system is relying on internal signal structures – the 'language' property – to enable optimal coding. This would, presumably, give better tuning schemes/learning paradigms than variation of the 'noise' alone.

As is also, however, usual with information theory existence arguments in general, establishment of a 'learning paradigm theorem' is unlikely to provide a clue to an engineering strategy: the development of efficient 'turbo codes' for straightforward signal coding in communication theory took until the early 1990's.

A simple neural architecture and some of its dynamics

Thus far we have restricted our attention to a single nonlinear oscillator/neuron and its properties when two generalized input signals are mixed. Next we attempt to expand the development to systems of oscillators coupled 'spatially,' in the largest sense, i.e. a neural network having an inherent architecture.

We assume, most simply, that there are two kinds of coupling between individual simple oscillator/neurons, defining a two-stage hierarchy of organization. More complicated architectures are treated in the next section.

The first coupling we will assume is reflexive, symmetric and transitive, permitting the division of the array into disjoint equivalence classes (Wallace and Wallace, 1998). We will call this a 'strong' tie, in the tradition of sociology (Granovetter, 1973). The second coupling, which we characterize as 'weak,' operates across all possible subdivisions of the array, and does not permit identification of disjoint equivalence classes.

The physicists might characterize these as 'local' and 'mean field' couplings.

We assume the index of 'strong' ties remains constant, and permit the index of 'weak' tie coupling, which we will call T, to vary and to characterize the array as a whole.

We assume the array, 'signal,' and 'noise' depend on three parameters, two explicit and one implicit.

The explicit are an 'external field strength' analog J which gives a 'direction' to the phenomenon, in addition to the inverse 'disorder' parameter K defined as $K = 1/T$, where T is a characteristic index of the strength of the 'weak' ties above, which, to reiterate, are those which couple elements of the array without disjointly partitioning them.

We may, in the limit, set $J = 0$. Other explicit parameters can be added, of course, at the expense of complicating the analysis.

The implicit parameter, which we call r, is an inherent 'generalized length' on which the phenomenon – including its direction and 'temperature' – is defined. That is, we can write J and K as functions of averages of the parameter r, which may be quite complex, having nothing to do with conventional ideas of space, although it may, of course, include a spatial component. For example Wallace and Wallace (1998) examine the role of social as well as of spatial separation within a population in determining the probability of weak ties between individuals or subgroups.

Rather than specify complicated patterns of individual dependence or interaction for signal, noise and elements of the array, we instead work entirely within the domain of the uncertainty of the ergodic information source dual to the large-scale information resonance defined by the entirety of the coupled array, which we write as $H[K, J, \mathbf{X}]$. This draconian simplification enables us to directly obtain certain general results.

Taking only K as significant for the moment, the relation

$$H(K) = \lim_{n \to \infty} \frac{\log[N(K)]}{n}$$

has the same form as the free energy density of a physical system. If $Z(K)$ is the partition function defined by the system's energy distribution, then the free energy density $F(K)$ is defined as

$$F(K) = \lim_{V \to \infty} \frac{\log[Z(K)]}{V}, \tag{12}$$

where V is the system volume. This is precisely the 'homology,' to use the description of Rojdestvenski and Cottam (2000), between free energy and information source uncertainty, in the context of the similarity between 'meaningful' statements and 'most probable paths' described above.

Imposition of invariance under a renormalization transform in the implicit parameter r on the dual information source of the ergodic information resonance characterizing the array as a whole leads to expectation of both a critical point in K, K_C, reflecting a phase transition to or from collective behavior across the entire array, and of power laws for system behavior near K_C.

See Wilson (1971) for calculational details, which are standard, and Wallace and Wallace (1998, 1999) for a more complete application to information sources.

We are thus suggesting that the mathematical 'unification' of fluctuations, statistical mechanics and information theory can be extended by generalized application of the renormalization methods which have proven so useful in statistical mechanics, and indeed in the analysis of deterministic chaos (e.g. McCauley, 1993).

Let $\kappa = (K_C - K)/K_C$ and take χ as the 'correlation length' defining the average domain in r-space for which the dual information source is primarily characterized by 'strong' ties. We begin averaging across r-space in terms of 'clumps' of length R, defining J_R, K_R as J, K for $R = 1$. Then, following the physical analog of Wilson (1971), we choose the renormalization symmetry relations as

$$H[K_R, J_R, \mathbf{X}] = R^D H[K, J, \mathbf{X}]$$

$$\chi(K_R, J_R) = \frac{\chi(K, J)}{R}, \tag{13}$$

where D is a nonnegative real constant, possibly reflecting a fractal network structure.

The first of these equations states that 'processing capacity,' as indexed by the source uncertainty of the system which represents the 'richness' of the inherent language, grows as R^D, while the second just says that the correlation length simply scales with R.

Other, markedly different, symmetry relations – not necessarily based on simple physical analogs – may well be possible, perhaps permitting a classification of information resonances.

For example renormalization analysis of 'deterministic chaos' leads to symmetry relations of the form (McCauley. 1993, p.168)

$$f(x) = \alpha^2 f(\alpha^{-1} f(\alpha^{-1} x))$$

and

$$f(x) = \alpha f(\alpha f(\alpha^{-2} x))$$

.

'Information' systems probably require complicated renormalization symmetries connected to their neural architectures, in the largest sense, to describe their behavior at phase transition, particularly if they are hierarchically nested.

Near the critical point K_C, for $J = 0$, some clever development and a simple series expansion of equation (13) (Wilson, 1971; Binney et al., 1995) gives

$$H = H_0 \kappa^{sD}$$

$$\chi = \chi_0 \kappa^{-s}$$

(14)

where s is a positive real constant.

Again, H is the dual source uncertainty of the entire array as an information resonance and χ the average size the region dominated by 'strong' ties, and $\kappa \equiv (K_C - K)/K_C$.

Further from the critical point matters are more complicated (Wilson, 1971).

We next attempt to estimate the size of the disjoint partition of the array into 'strongly' interacting subsets when the system undergoes a phase transition.

Assume $K < K_C$, where K_C is the critical point, and that the rate of change of κ remains constant as $K \to K_C$, so that $|d\kappa/dt| = 1/\tau_K$ for some fixed τ_K.

Further analogs with physical theory suggest there is a characteristic time constant for the phase transition, $\tau \equiv \tau_0/\kappa$, such that if changes in κ take place on a timescale longer than τ for any given κ, we may expect the correlation length to remain in equilibrium with internal changes, resulting in very large fragment sizes in r-space.

Following Zurek (1985, 1996), we argue that the 'critical' freezeout time $\hat{t}$, will occur at a 'system time' $\hat{t} = \chi/|d\chi/dt|$ such that $\hat{t} = \tau$. Taking the derivative $d\chi/dt$, remembering that by definition $d\kappa/dt = 1/\tau_K$, gives

$$\frac{\chi}{|d\chi/dt|} = \frac{\kappa \tau_K}{s} = \frac{\tau_0}{\kappa}$$

so that

$$\kappa = \sqrt{s\tau_0/\tau_K}.$$

Substituting this value of κ into the equation for correlation length, the expected size of fragments of the 'spatially distributed information resonance' in r-space, $d(\hat{t})$, becomes

$$d \approx \chi_0 (\frac{\tau_K}{s\tau_0})^{s/2}.$$

(15)

The more rapid the changes, the smaller τ_K and the smaller, and more numerous, on average, the resulting fragments.

Different renormalization symmetry relations than equation (13) would, of course, lead to different 'universal' power laws.

It is clear from this development that *sudden critical point transition is possible in the opposite direction*, that is, from a number of independent, isolated and spatially fragmented assembly of 'strongly' interacting resonators of some average 'diameter,' d, firing more or less at random, into a single large, tightly interlinked and 'spatially' extended resonator firing coherently with much enhanced properties. This result, which is quite general for parametized information sources and is not restricted to those dual to information resonance arrays, has obvious and significant implications for a number of disciplines (Wallace and Wallace, 1998, 1999; R Wallace, 2000a, b; Wallace and Fullilove, 1999).

Again, qualitatively different renormalization symmetries would lead to markedly different forms of coherence.

The first part of equation (14) gives

$$H = H_0(\frac{K_C - K}{K_C})^{Ds}$$

while the second gives

$$(\frac{K_C - K}{K_C})^s = \frac{\chi_0}{\chi}.$$

Thus

$$H = H_0(\frac{\chi_0}{\chi})^D,$$

or

$$H[\mathbf{X}] \propto \frac{1}{\chi^D}.$$

(16)

This result states that the 'richness' of the language defined by the spatial array of stochastic resonators is inversely related to the domain dominated by disjointly partitioning 'strong' ties. As the non-disjunctive 'weak ties' coupling decreases, the efficiency of the coupled array as an information channel declines precipitously near the transition point.

In a neural setting we might envision large χ, and hence small $H[\mathbf{X}]$, as representing a 'hypersynchronous' state of low complexity analogous to sleep or generalized epilepsy (Wallace, 2000a; Wallace and Fullilove, 1999).

We have implicitly assumed that renormalization symmetry is related to underlying system 'architecture,' in some sense. While making that relation formally explicit appears no small task, it seems likely that specification of renormalization symmetry may serve to specify fundamental properties of the underlying architecture of neural systems, and of the relation of architecture to certain kinds of dynamics.

Interacting information sources and structural hierarchy

The development above suggests that critical point transition can be reversed and may serve to unite disparate fragments into a single coherent entity. Here we describe two qualitatively different mechanisms.

We know that disparate information sources interact. The currently predominant view of human evolution focuses heavily on gene-culture interaction. As Durham (1991) puts it,

> "...[G]enes and culture constitute two distinct but interacting systems of information inheritance within human populations... [and] information of both kinds has influence, actual or potential, over ... behaviors, [which] creates a real and unambiguous symmetry between genes and phenotypes on the one hand, and culture and phenotypes on the other...
>
> [G]enes and culture are best represented as two parallel lines or 'tracks' of hereditary influence on phenotypes..."

A simple model of such coevolution envisions two information sources as *becoming each other's primary environments*, the coevolutionary Red Queen.

The two information sources are, respectively, characterized by information sources $\mathbf{X}$ and $\mathbf{Y}$, whose uncertainties are parametized (1) by inverse measures of several parameters, J and Q, and, most critically, (2) by each others inverse uncertainties, $\mathcal{H}_X \equiv 1/H[\mathbf{X}]$ and $\mathcal{H}_Y \equiv 1/H[\mathbf{Y}]$, i.e.

$$H[\mathbf{X}] = H[Q, J, \mathcal{H}_Y, \mathbf{X}]$$

$$H[\mathbf{Y}] = H[Q, J, \mathcal{H}_X, \mathbf{Y}].$$

(17)

Assume, for simplicity, that $J = Q = 0$ and that $H[\mathbf{X}]$ follows a *reverse S-shaped curve* with $K \equiv \mathcal{H}_Y$, and similarly $H[\mathbf{Y}]$ depends on $\mathcal{H}_X$. That is, increase or decline in the source uncertainty of the first information source leads to increase or decline in the uncertainty of the second, and vice versa. The 'richness' of the different languages are interlinked.

Start at the right of the K axis for $H[\mathbf{X}]$, the first source uncertainty, and indeed to the right of the critical point K_C, at the beginning of interaction. Assume interaction affects the other information source: $H[\mathbf{Y}]$ increases, reflecting the ability to 'say more things,' so $\mathcal{H}_Y$ decreases, and thus $H[\mathbf{X}]$ increases, walking up the reverse-S curve from the right: the richness of the first 'language' increases.

The increase of $H[\mathbf{X}]$ leads, in turn, to an increased capacity of the second 'language,' thus a decline in $\mathcal{H}_X$ triggers an increase of $H[\mathbf{Y}]$, whose increase leads to a further increase of $H[\mathbf{X}]$ and vice versa: The famous coevolutionary Red Queen, taking the system from the right of the K-axis, through a sudden phase transition and condensation, far up the curve to the left.

The process can go the other way. Start with an initially strongly linked coevolutionary composite, with $H[\mathbf{X}]$ at the far left of the reverse S curve, near the origin.

Suppose $H[\mathbf{Y}]$ begins to decline: $\mathcal{H}_Y$ increases and $H[\mathbf{X}]$ causing a further fall in $H[\mathbf{Y}]$, triggering another rapid decline in $H[\mathbf{X}]$, which triggers a further decline in $H[\mathbf{Y}]$, and so on until the system reaches the critical point K_C, and suddenly shatters.

This is the Red Queen's Ratchet, leading to a high-speed punctuated fragmentation (Wallace and Ullmann, 1999).

Other 'joining' mechanisms are possible, in addition to, or parallel with, the one above: Earlier we mentioned the possibility that 'state space' might be partitioned according to which points within it can be linked by 'meaningful' paths, much like the division of the underlying space of a nonlinear dynamic system into 'domains of attraction.' It seems clear that the hierarchical stacking or nesting of information sources, as we have discussed it, must be exactly mirrored in the algebraic structure of state space, since it is precisely the division of state space into a small number 'meaningful' and a large number of 'other' paths which defines the existence of an information source.

One 'natural' approach is related to *functional* hierarchy, characterized by some strong order relation $\prec$, so that each disjoint regime of the state space has its own characteristic information source $H_1, H_2, H_3, ...$ with the inherent ordering

$$H_1 \prec H_2 \prec H_3 \prec ...$$

(18)

The relation $\prec$ will typically indicate hierarchy of function rather than simple mathematical magnitude.

We thus assume a disjoint partitioning of state space in our 'language-based' neural network such that a 'large' number of (nearly) simultaneous information sources engage *each* nonlinear oscillator. That is, an oscillator participates in H_j disjoint information sources representing a 'neural domain,' each having a distinct 'path'

$$x^j = a_1^j, a_2^j, ..., a_n^j, ...$$

We will ultimately want to arrange a set of *meaningful* paths x^j of the H_j in a 'natural' manner *so as to constitute a message sent by an enveloping information source*. In general, however, there is no unique or 'natural' way of arranging the set of meaningful paths x^j in the same way the set of a_k^j constituting an individual path is ordered by the passage of time: we must impose an order relation in a 'natural' manner: functional hierarchy, based on equation (18), provides one answer, giving a well defined sequence of paths

$$x \equiv x^1, x^2, x^3, ..., x^m, ...$$

(19)

Other 'natural' ordering relations may be possible, and may indeed operate simultaneously.

We will require joint and conditional probabilities to be associated with each x across a hierarchy 'large enough' to define an approximate ergodic information source $\mathbf{H}$ which we characterize as being *induced* by the order relation $\prec$.

Provided the hierarchy is 'deep' enough, we can take the asymptotic equipartition property to permit division of the set of possible 'hierarchical paths' x of the induced ergodic source $\mathbf{H}$ into a small set with high probability and a much larger set with vanishingly small probability. A meaningful path x of the hierarchical information source $\mathbf{H}$ represents a collective action of the nested hierarchical structure defined by the imposed order relation.

Extension of this definition to partial ordering as above (i.e. $\preceq$ rather than $\prec$) seems nontrivial.

A standard inequality of information theory (e.g. Khinchine, 1957; Ash, 1990; Cover and Thomas, 1991) for any sequence of probability spaces X_j suggests

$$\mathbf{H} \leq \sum_{j=1}^{m} H[\mathbf{X}_j]$$

(20)

where the $H[\mathbf{X}_j]$ represent the ergodic source uncertainties of the m disjoint multitasking components.

If the multitasking information sources are *uncorrelated* across the hierarchy, we have reduced the system to a kind of product information channel, whose capacity C is the sum of the capacities of the components independently (Ash, 1990, p. 85; Shannon, 1957);

$$C = \sum C_j.$$

(21)

where C_j is the capacity of the channel corresponding to $H[\mathbf{X}_j]$.

Thus uncorrelated multitasking structures can have very high capacities as channels for pattern recognition, a matter of some interest in its own right.

For our case, however, we use the order relation implied by functional hierarchy to impose an approximate ergodic information source uncertainty on the entire hierarchical structure, allowing definition of an ergodic source uncertainty across the hierarchy itself,

$$\mathbf{H} \approx \frac{1}{m} H[\mathbf{X}_1, ..., \mathbf{X}_m] \leq \frac{1}{m} \sum_{j=1}^{m} H[\mathbf{X}_j],$$

(22)

where $m \gg 1$ represents the depth of the hierarchy. If $C_{\mathbf{H}}$ is the capacity of the ordered hierarchy of depth m as an information channel, then this suggests

$$C_{\mathbf{H}} \leq \frac{C}{m},$$

(23)

where C is the capacity of the uncorrelated product channel from equation (21).

The next iteration of the argument is a natural extension of previous developments, and is one way to formalize the speculation following equation (15) regarding the coagulation of uncorrelated fragments into a coherent object: we take the linkage between levels of hierarchy to be determined by an average inverse 'strength of weak ties' index L, while the average strength of the weak ties linking individual oscillators/neurons *at each level of hierarchy* is again parametized by K. L serves to provide an inherent 'direction' to the process, much like an external field strength in a physical system. The critical point K_C then becomes a critical line $K_C(L)$.

Let $\mathcal{D}$ be the real exponent defining the renormalization symmetry of the imposed ordered structure so that, not far from criticality,

$$\mathbf{H} \propto [\chi(K, L)]^{-\mathcal{D}}$$

(24)

where χ is the minimum 'correlation length' of the linkage across the hierarchy.

While equation (21) suggests enhanced capacity in an unstructured multitasking neural network for pattern recognition, the imposition of ergodicity through equation (22) permits introduction of the parameter L, allowing for much more complex system dynamics.

This suggests an inherent, and possibly very general, trade-off between channel capacity and complexity of system dynamics for hierarchical neural structures.

These two examples suggest we can 'infer the general from the particular,' and imply the possibility of a plethora of qualitatively different coupling schemes for building larger information sources from smaller ones.

Next we explore the matter of dynamic complexity more fully.

Dynamic behavior far from the critical surface

We have thus far focused on the dynamic properties of a parametized information source representing an array undergoing 'information resonance' near a critical surface. We now ask how a parametized information source behaves 'normally,' i.e. far from such a surface.

According to the Shannon-McMillan Theorem, the number $N(n)$ of meaningful paths of length n emitted by an information source satisfies the subadditive relation

$$H[\mathbf{X}] = \lim_{n \to \infty} \frac{\log[N(n)]}{n},$$

where $H[\mathbf{X}]$ is the source uncertainty defined from the joint and conditional probabilities of the paths x. Indeed, this kind of relation serves as the starting point for most modern treatments of large deviations (e.g. Dembo and Zeitouni, 1998).

For a physical system the free energy density is defined by an analogous relation:

$$F(K_1, ..., K_m) = \lim_{V \to \infty} \frac{\log[Z(K_1, ..., K_m)]}{V}$$

(25)

where V is the system volume, $K_1, ..., K_m$ are other system-wide parameters, and $Z(K_1, ..., K_m)$ is the partition function defined from system energy states. K_1 is usually taken as an inverse temperature.

Above we used this homology to impose renormalization symmetries relating 'phase change' to underlying architecture for parametized information sources. Here we make further use of it, far from critical points or surfaces, to connect architecture and dynamics, in the large sense.

We parametize information source uncertainty in a similar manner, so that

$$H = H[K_1, ..., K_m, \mathbf{X}]$$

(26)

where now we take K_1 as the inverse of the 'strength of weak ties' across the array.

For a physical system we can obtain the equation of state which describes the macroscopic behavior of the system through imposition of a *Legendre transform* on equation (25). The Legendre transform of a well-behaved function $f(K_1, ..., K_m)$ is defined by

$$g = f - \sum_{i=1}^{w} K_i \partial f / \partial K_i$$

$$\equiv f - \sum_{i=1}^{w} K_i Q_i,$$

(27)

so that $Q_i = \partial f / \partial K_i$, and is invertable provided $\partial f / \partial K$ is well behaved. Then we can write

$$f = g - \sum_{i=1}^{w} Q_i \partial g / \partial Q_i.$$

(28)

The generalization when f is not well-behaved is through a variational argument (Beck and Scholgl, 1993; Griffiths, 1972; Fredlin and Wentzell, 1998; Dembo and Zeitouni, 1998) rather than this tangent plane argument.

In a physical system for which F is the free energy, the Legendre transform defines the macroscopic entropy as

$$S \equiv F - \sum_i K_i \partial F / \partial K_i.$$

(29)

The associated generalized forces constituting the equation(s) of state are then

$$Q_i \equiv \partial F / \partial K_i$$

We propose as a macroscopic equation of state for a system characterized by a parametized information source uncertainty $H[K_1, ..., K_m; \mathbf{X}]$ the tautological relations

$$S \equiv H - \sum_i K_i \partial H / \partial K_i.$$

$$Q_i \equiv \partial H / \partial K_i$$

(30)

where we will now call S the macroscopic *disorder*.

Note in particular that each hierarchical ordering relation will add a new 'strength of weak ties' parameter to the thermodynamics.

For physical systems the 'Onsager relations' define the system's response to entropy. These assume that time rates of change of the defining parameters, the K_i, are in direct proportion to 'thermodynamic forces' defined as gradients in the entropy with respect to the characteristic parameters:

$$dK_i/dt = \sum_j L_{i,j} \partial S/\partial K_j.$$

Our initial approach to this essential question for information systems is not through analogs with simple physical constructs, but rather by means of Ives' (1995) theory of resilience for complex ecological systems, which are almost always plagued by multiple feedback loops. See Wallace (2000b) for the physics analog.

We will assume that complicated 'behavioral' systems are indeed affected by the entropy construct we have called the 'disorder,' but in a very general way, and seek feedbacks in the equation of state of the information system, with the form

$$S(t) = H[(S(t))] - \sum_i K_i(S(t)) \partial H/\partial K. \tag{31}$$

Like the conventional Onsager relations, we start with a first order effect, near some reference configuration (S^0, K_i^0). Writing deviations in S and the K_i from that configuration on the same footing as variates x_j, we will seek empirical 'Onsager-like' regression relations in the deviations themselves rather than in $\partial S/\partial K_i$ and dK_i/dt:

$$x_j(t) = \sum_{k \neq j}^{m} b_{j,k} x_k(t) + b_{j,0} + \epsilon(t, x_1(t), ...x_m(t)). \tag{32}$$

The ϵ terms include both 'noise' and nonlinearities, and are not necessarily small.

Note that we may well work with 'environmental indices' of S and the K_i as our variates x_j. These may well be more familiar psychological or indeed social constructs using available empirical information.

In matrix notation,

$$X(t) = \mathbf{B}X(t) + U(t), \tag{33}$$

where $\mathbf{B} = \mathbf{B}|_{S_0,K_0}$ is a fixed $m \times m$ matrix of regression coefficients having a zero diagonal and $U(t)$ is an m-dimensional vector containing both the constant and 'error' terms.

'Error' terms are taken as including 'shocks' outside the internal feedback loops. We are thus assuming that the system is operating 'normally' and we ask about the effect of perturbations from this 'normal' state (S_0, K_0).

We begin by rewriting the matrix equation as

$$[\mathbf{I} - \mathbf{B}]X(t) = U(t) \tag{34}$$

where $\mathbf{I}$ is the $m \times m$ identity matrix and, to reiterate, $\mathbf{B}$ has a zero diagonal.

We now reexpress matters *in terms of the eigenstructure of* $\mathbf{B}$.

Let $\mathbf{Q}$ be the matrix of eigenvectors which diagonalizes $\mathbf{B}$. Take $\mathbf{Q}Y(t) = X(t)$ and $\mathbf{Q}W(t) = U(t)$. Let $\mathbf{J}$ be the diagonal matrix of eigenvalues of $\mathbf{B}$, so that $\mathbf{B} = \mathbf{QJQ}^{-1}$.

In R. Wallace et al. (1997) we show how, after some nontrivial development, the treatment can be reduced to canonical correlation, so that the eigenvalues of $\mathbf{B}$ are all real. Then, for the eigenvectors Y_k of $\mathbf{B}$, corresponding to the eigenvalues λ_k,

$$Y_k(t) = \mathbf{J}Y_k(t) + W_k(t).$$

(35)

Using a term-by-term shorthand for the components of the non-orthogonal Y_k, this becomes

$$y_k(t) = \lambda_k y_k(t) + w_k(t).$$

(36)

In the usual manner define the mean of a time-dependent function $f(t)$ over the interval ΔT as

$$E[f(t)] \equiv \frac{1}{\Delta T}\int_0^{\Delta T} f(t)dt,$$

(37)

and the variance as $Var[f] = E[(f - E[f])^2]$. We assume an appropriately rational structure as $\Delta T \to \infty$.

Again taking matters term-by-term, we get the variance in the y_k as $Var[(1-\lambda_k)y_k] = Var[w_k]$, so that

$$Var[y_k] = \frac{Var[w_k]}{(1-\lambda_k)^2},$$

or, taking $\sigma = \sqrt{Var}$,

$$\sigma(y_k) = \frac{\sigma(w_k)}{|1-\lambda_k|}.$$

(38)

.

The y_k are the components of the eigentransformed deviation variates x_i, and the w_k are the similarly transformed variates of the driving externalities $u_i(t)$.

The eigenvectors Y_k are characteristic but non-orthogonal combinations of the original variates x_i whose standard deviation is that of the particular patterns of driving externalities W_k, but synergistically amplified by the term $1/|1-\lambda_k|$, a function of the eigenvalues of the matrix of regression coefficients $\mathbf{B}$. This kind of amplification is typical in ecosystem theoretics (e.g. Caswell, 1999).

For a two dimensional system involving normalized variates it is easy to show that λ is just their correlation.

The non-orthogonal nature of the excited eigenstates Y_k, however, means that this response will 'leak' into other characteristic tuned eigenmodes.

Wallace (2000b) uses this treatment in the design of a system to detect even slight deviations from a reference configuration in the parameter space of the K_i. Clearly proper tuning of the system to a particular pattern of 'sensory activity' perturbations from that reference would result in high amplitude Y-signals.

Note that, just as we can conceive of many different kinds of grammar and syntax for the inherent 'neural language,' different algebras of 'state space partition,' and different forms of renormalization symmetries characterizing behavior at phase transition, so too we might envision many possible 'generalized Onsager relations' defining the response of the generalized neural structure to the 'disorder' variate, connecting architecture and dynamics.

That is, the formal study of neural structure and process, in the large sense, is not necessarily constrained to examination of models taken from physics, or even of empirical studies using results from a particular line of successful evolutionary development: Von Neumann computer architecture has little to do with the structure of the human brain, it turns out, but is nonetheless a foundation of much interesting technology.

Quantum neural networks

The 'classical' development above suggests that a formal theory of quantum neural networks (QNN) can be based on the quantum generalizations of the fundamental theorems of information theory, provided we can construct the proper information resonance associated with the QNN, in the context of the appropriate 'quantum learning theorem.' Parametization of the resulting composite quantum information source by an index of structural hierarchy – i.e. disjoint partitioning vs. non-disjoint interaction – along with imposition of appropriate *information structural* renormalization and generalized Onsager relations, determines pattern recognition behaviors. To reiterate, these latter are properties of the quantum information source which may be quite unlike simple physical models.

That is, we can argue from information theoretic existence theorems for quantum versions of our classical pictures, without the necessity of detailed structural specifications. However, to reiterate, information theory existence arguments are notorious for providing general results but completely failing to suggest engineering strategies.

In spite of much physics literature *Sturm und Drang*, there are few mathematically rigorous quantum generalizations of information theory. One such is the extension of the Shannon-McMillan Theorem by King and Lesniewski (1995). We paraphrase that paper closely, and begin with some standard material on quantum information sources.

The fundamental object of interest is a quantum system whose state space is a tensor product of many copies of one fundamental space M. The source produces a signal which is encoded by a state in M, and the ensemble of possible states is represented by a density operator ρ on M. An extended source corresponds to a sequence of such states, which is interpreted as a message. The probabilistic nature of the message is contained in the density operator on the tensor product of copies of M. If that operator is the product $\rho \otimes ... \otimes \rho$, there are no correlations between signals in the message, the 'quantum Bernoulli' source which was the focus of Schumacher's work (Schumacher, 1995). A more useful quantum information source is one in which there are correlations on all time scales between signals in the message. The density operator then becomes a much more complex object.

For purely classical signals, the Shannon-McMillan Theorem permits the splitting of all possible signals into two classes, a relatively small number of 'meaningful' ones with significant probabilities, and a much larger number with vanishingly small probability. The criterion for splitting is the uncertainty of the information source. The quantum result is a splitting of the state space into relevant and irrelevant subspaces, with the Von Neumann entropy as the criterion.

King and Lesniewsky (1995) derive an estimate for the dimension of the relevant subspace by computing the entropy of a classical source obtained by taking measurements on the quantum system. For the case of a quantum source emitting orthogonal states, the Von Neumann entropy is the same as the uncertainty of the associated classical information source, since the density operators all commute, and the Shannon-McMillan Theorem is recovered exactly. Non-orthogonal sources are more complicated, and their result only provides somewhat loose limits on the dimensionality of the relevant space.

To give the flavor of these matters, we again paraphrase King and Lesniewski (1995).

The quantum source sends a series of signals, each of which is a vector in a finite dimensional Hilbert space $\mathcal{H}$. The source is taken as discrete, with each signal an element of a finite set $\mathcal{S} = |\psi_1 >, ..., |\psi_s >$ of normalized vectors in $\mathcal{H}$. We take $\mathcal{H}$ as spanned by $\mathcal{S}$, so that $\mathcal{H}$ is of dimension $d \leq s$. Unlike the classical case, this system can entertain a superposition of states. Let p_j be the given probability of the state $|\psi_j >$ being sent. The density matrix corresponding to the ensemble of signals $\mathcal{S}$ is then

$$\rho = \sum_{1 \leq j \leq s} p_j |\psi_j >< \psi_j|,$$

(39)

with $tr(\rho) \equiv 1$. While $\mathcal{S}$ and the distribution of p_j uniquely determine the density matrix, each such matrix corresponds to an infinite number of possible sets of states.

The observables associated with quantum signals are $d \times d$ hermitian matrices, the elements of a C^*-algebra $\mathcal{A} = \mathcal{L}(\mathcal{H})$ of linear observables on $\mathcal{H}$. The state on the algebra of observables $\mathcal{A}$ associated with the density matrix ρ is, for any given $A \in \mathcal{A}$,

$$\tau_1(A) \equiv tr(A\rho) = \sum_{1 \leq j \leq s} p_j < \psi_j | A | \psi_j > .$$

(40)

Appropriate generalizations can be given for infinite dimensional tensor products, and ergodic quantum information sources can be defined.

The density matrix of order n becomes, in terms of the states ψ_j which span $\mathcal{S}$,

$$\Pi_n =$$

$$\sum_{1 \leq j_1, \ldots, j_n \leq s} p_{j_1, \ldots, j_n} |\psi_{j_1} >< \psi_{j_1}| \otimes \ldots \otimes |\psi_{j_n} >< \psi_{j_n}|$$

(41)

The entropy associated with a sequence of n signals is defined as

$$H_n(\Pi) \equiv -tr_{\mathcal{H}^{\otimes n}}(\Pi_n \log \Pi_n)$$

(42)

Some development gives

$$H_{m+n}(\Pi) \leq H_m(\Pi) + H_n(\Pi)$$

so that the limit

$$h(\Pi) = \lim_{n \to \infty} \frac{H_n(\Pi)}{n}$$

(43)

exists. We call $h(\Pi)$ the entropy of the quantum source. For a Bernoulli source $\Pi_n = \rho \otimes \ldots \otimes \rho$ and $h(\Pi) = -tr_{\mathcal{H}}(\rho \log \rho)$. General sources with internal serial correlations have far more complex expressions for h.

Let $\mathbf{A} = [A_1, \ldots, A_r], r < \infty$ be a family of observables on $\mathcal{H}$ such that $A_j \geq 0$ for all j, and

$$A_1 + \ldots + A_r = I$$

(44)

where I is the identity. We call the set $\chi_{\mathbf{A}} = [1, \ldots, r]$ the classical alphabet associated with $\mathbf{A}$, and denote by $\chi_{\mathbf{A}}^{\infty}$ the space of all infinite messages over the alphabet $\chi_{\mathbf{A}}$. In this way we can associate a classical information source with each quantum information source.

Let $\mathcal{H}^{\otimes n}$ be the space of all signals of length n for an ergodic quantum information source. According to the quantum Shannon-McMillan Theorem, it can be factored into two orthogonal subspaces

$$\mathcal{H}^{\otimes n} = \mathcal{S}_n \otimes \mathcal{S}_n^{\perp}, \tag{45}$$

whose relative dimensions are constrained by the uncertainty of the classical information source $h_{\mathbf{A}}$ associated with the quantum source in a precise manner. If the $|\psi_j>$ are orthogonal, $h_{\mathbf{A}}$ is just the Von Neumann entropy of the source, since the density operators all commute.

Let $P_{\mathcal{S}_n}$ be the orthogonal projection onto the relatively small subspace $\mathcal{S}_n$.

Let C be an observable $C \in \mathcal{L}(\mathcal{H}^{\otimes n})$, where the signal is of length n. Then, according to the quantum Shannon-McMillan Theorem, the difference

$$|\tau(CP_{\mathcal{S}_n}) - \tau(C)|$$

can be made arbitrarily small as n increases without limit. Here τ is an appropriate infinite-dimensional generalization of τ_1 above, in terms of the complicated density matrices Π.

Again paraphrasing King and Lesniewski (1995), in the case where the $|\psi_j>$ are orthogonal, there is a direct correspondence with the classical Shannon-McMillan theorem, and the quantum theory is simply a restatement of the classical result, with the associated classical source uncertainty $h_{\mathbf{A}}$, (which constrains the dimensionality of the significant space $\mathcal{S}_n$), given by the Von Neumann entropy.

Although we do not have equivalently full quantum forms of the Shannon Coding Theorem and its 'Learning Theorem' variant, these considerations nonetheless suggest a possible 'correspondence principle' generalization of the classical neural network results given in the earlier sections: Parametization of the quantum information source corresponding to a QNN must reflect the underlying structural hierarchy of the system, incorporated in the renormalization symmetry and other inherent properties of the information source. Measurement must give an appropriately parametized classical information source with appropriate renormalization and generalized Onsager properties.

The parametization of the quantum information source might well be complicated, for example simultaneously involving both quantized and unquantized physical quantities: one imagines simultaneously macroscopic external signals and an array of quantum oscillators coupled by some kind of quantized field - phonons, photons, etc.

Since a quantum information source is still a 'language,' in the sense of the earlier sections of this work, its renormalization and generalized Onsager properties may not be simple extensions or reflections of commonly understood physical systems, but characterize, in no small part, the patterns of internal correlations defining that language – the jointly defined grammar and syntax of the coupling of sensory signal, neural weights and array of nonlinear oscillators constituting the system: Neural networks, quantum or classical, are defined by their 'meaning' even more than by their physical structure.

Considerations of the various possible 'natural' relations between neural architecture, learning paradigms, renormalization symmetry and generalized Onsager relations which applied to classical systems would seem appropriate to the pure quantum case as well.

Density matrix and path integral

Rojdestvenski and Cottam (2000), in their explicit extension of Wallace and Wallace (1998) to physical processes, end with the following 'simple' observation:

> "If one takes an 'evolution' equation of any system..., it may always be written in the following differential form
>
> $$\psi(t + dt) = (1 + \mathbf{E}dt)\psi(t)$$
>
> where $\mathbf{E}$ is called the 'evolution operator.' If the evolution has different 'channels,' i.e.

$$\mathbf{E} = \sum_{i=1}^{N_0} \mathbf{E}_i,$$

then [the first equation] takes the following recursive form:

$$\psi(t + mdt) =$$

$$(1+\mathbf{E}dt(...(1+\mathbf{E}dt(1+\mathbf{E}dt(1+\mathbf{E}dt)))...)\psi(t)$$

$$= \sum_{r=1}^{m} (dt)^m \sum_{C_r} K(C_r)[\mathbf{E}_{i_1}...\mathbf{E}_{i_r}]\psi(t)$$

and again we deal with the 'sentence' representation. In a certain sense, any temporal evolution, if only it is describable by equations, is a message [from some information source] in its own right."

Behrman et al. (1996) open their description of a quantum dot neural network in a similar manner:

> "In most artificial neural network implementations, the neurons receive inputs from other processors *via* weighted connections and calculate an output which is passed on to other neurons. The calculated output... of the i^{th} neuron [is determined from] the signals from the other neurons in the network... Similarly we can write the expression for the time evolution of the quantum mechanical state of a system:
>
> $$|\psi(x_f, T) >= G(x_f, T; x_0, 0)|\psi(x_0, 0) > ...$$
>
> Here $|\psi(x_0, 0) >$ is the input state, the initial state of the quantum system. $|\psi(x_f, T) >$ is the output state, the state of the system at $t = T$. G is the Green's function, which propagates the system forward in time, from initial position x_0 at time $t = 0$ to final position x_f at time $t = T$. [G can be expressed] in the Feynman path integral formulation of quantum mechanics (Feynman, 1965), in which G is thought of as the infinite sum over all possible paths that the system could possibly take to get from x_0 to x_f... Each path is weighted by the complex exponential of the phase contributed by that path, given by the classical action for that path;... Each of the N [quantum] neurons' different possible states contribute to the final measured state; the amount it contributes can be adjusted by changing the potential energy..."

Those paths with higher weighting thus have higher probability – are 'meaningful,' in our terminology – than the others. For an 'ergodic' information source such paths would be equiprobable.

Using this formalism, Behrman et al. (1996) conclude that

> "Potentially, a quantum neural network would be an extremely powerful computational tool... capable, at least in principle, of performing computations that cannot be done, classically... an actual working quantum neural net would likely want to take advantage of the greater multiplicity and connectivity inherent in an entire array of quantum dot molecules, by placing molecules physically close enough to each other that nearest neighbors can interact directly..."

The path integral formulation of quantum density matrices (Feynman, 1998) thus seems to form the natural linkage between quantum mechanics and quantum information theory in much the same way that the Large Deviations Program of applied probability connects statistical mechanics, fluctuations and information theory in classical systems. Imposition of appropriate renormalization symmetry on the ergodic quantum information source dual to the QNN, in the context of a similarly appropriate 'generalized Onsager relation' and associated algebras, would indeed seem to be the most natural means of expressing the unique architecture of the network, hierarchical or otherwise.

By analogy, it seems that a Landau-like 'two fluid' model of superconductivity and superfluidity is likely to apply to the general QNN, with a classical information source uncertainty playing the role of a 'phonon gas excitation' of the purely quantum QNN (Feynman, 1998). It is difficult, at this point, to imagine any other outcome.

We find that the work of King and Lesniewski (1995), a rigorous extension of the Shannon-McMillan Theorem to quantum systems, in conjunction with the material described in the first part of this paper, suggests a direction for development of a purely quantum neural network formalism, in contrast, for example, with the quasi-classical results of Toth et al. (1996). Quantum neural networks, like their classical counterparts, should be reducible to the convolution of sensory activity, ongoing activity 'neural weights' and an array of nonlinear components into a single quantum information source parametized by continuous or quantized variates. 'Tuning' the parameters and the 'ongoing activity' should, as for classical systems, result in highly efficient pattern recognition, depending on the inherent grammar and syntax of the associated quantum information source: data consistent with the system's linguistic rules are recognized, others are not.

The inherently parallel nature of pure quantum computation should provide some significant advantages over classical neural network pattern recognition. Quantum neural architecture should, as in the classical case, express itself in the renormalization symmetry of the dual quantum information source, its 'generalized Onsager relations,' and the algebraic structure of the underlying state space. Thus, for a certain class of QNN, high probability paths will define a quantum information source having grammar, syntax and higher order structures which will define the characteristics of the system for pattern recognition.

We speculate that a quantum linguistics – the extended algebra of Π operators corresponding to quantized neural networks – will be a principal growth technology for the 21st century.

Discussion and conclusions

From an information theory base, we have created a very general phenomenological picture of hierarchical neural structure and process with several distinct pieces which can be modified and assembled in different ways. The approach is recognizably similar to the macroscopic spring-weight-and-dashpot models which 19th century physicists used successfully to predict a surprisingly large part of the subtle viscoelastic behaviors of materials without necessity of a detailed reductionist understanding of their microscopic structure.

That picture can be summarized as follows:

(1) Systems undergoing 'information resonance,' as we have described it, including but not limited to neural process, are characterized by an inherent information source and its associated language. Paths in state space which are not consistent with the grammar and syntax of that language do not trigger 'fundamental events.'

(2) The underlying space of the information source may well have a structure much like the division of the state space of a nonlinear dynamic system into domains of attraction: states connectable by 'meaningful' paths may form disjoint equivalence classes, a reflexive, symmetric and transitive algebraic relation between them. Existence of an inverse mapping imposes a group structure of finite order, possibly leading to exceedingly complex 'multitasking' ability, related to the prime number partitioning of the group size. Existence of an order relation between regimes imposes a 'natural' hierarchical multitasking structure in which there is a tradeoff between pattern recognition capacity and complexity of functional dynamics.

(3) Distributed systems subject to information resonance, including neural structures, are likely to undergo precipitate phase transitions. Indeed, neural architecture – in a large sense – may well be, at least in part, characterizable by the renormalization symmetry which describes the behavior of the system near such phase transition. Renormalization symmetries of such structures are not necessarily those of simple physical systems.

(4) Systems subject to information resonance are likely to have a 'thermodynamics,' an 'equation of state' derived from the source uncertainty of the defining language by imposition of a Legendre trans-

form, and associated 'generalized Onsager relations' describing the role of architecture, though the disorder construct, in driving system dynamics. The generalized Onsager relations are, again, unlikely to be constrained to simple physical analogs.

(5) We have described learning paradigms for arrays undergoing information resonance in terms of an information-theoretic 'tuning' which may permit more efficient pattern detection than simple least-squares or 'infomax' treatments which do not utilize the internal syntactic structures of the incoming signal constituting the pattern to be recognized. It is this tuning, in fact, which allows us to speak of information resonance.

(6) We have explored 'coevolutionary,' in the large sense and order-relation-based interactions between disparate information sources which may permit a 'natural' hierarchical and/or multitasking nesting of function in arrays undergoing information resonance, including neural networks. Many other linking mechanisms seem possible for building larger from smaller parametized information sources.

A number of research questions emerge from these discussions, several in particular concerning the relations between the points above. These are, in essence, a search for 'natural' arrangements of our phenomenological building blocks.

First, the algebra of state space partitioning seems intimately related to important 'language' structures, perhaps even the question of coevolutionary (in the large sense) hierarchy and nesting. If that algebra goes beyond a single reflexive, symmetric, and transitive relation, then hierarchy seems implicit: envision the additive structure of the integers supplemented first by multiplication and division to give fractions, and then extended to real and complex numbers. Any integer is at the same time a rational, a real and a complex number. On the other hand, an integer is either even or odd, and the integers modulo two form a group of order two.

That is, there is evidently some relation between state space partitioning and algebra, and language grammar and hierarchy for systems undergoing information resonance, including neural structures. Explicating that relation seems an important topic for further research, particularly in view of the possibility of multitasking inherent in any disjoint partition. Imposition of an inverse mapping to give a group structure further opens the multitasking vista since markedly different groups may have the same order, depending, roughly, on its prime partition.

This set of questions is related to a second point:

What are the relations between renormalization properties, order relations, hierarchy and neural architecture? The simple renormalization relation we chose for a two-stage hierarchy had 'language richness,' and presumably computing capacity, growing as some power of the clumping parameter, $\propto R^D$ according to equation (13). What happens to order-relation-nested hierarchical systems undergoing phase transition? What kind of renormalization is appropriate for higher iterations of hierarchy? Can renormalization symmetries be nested along with structural hierarchy? Recent work in landscape ecology (Richie and Olff, 1999; Milne, 1992) suggests that hierarchically nested structures may have nested scaling laws. That is, if the structures are nested and follow respective scaling laws $\propto R^D, R^F, R^Q$, then necessarily

$$D \geq F \geq Q.$$

This suggests that renormalization properties and state space algebra may both constrain architecture.

What other forms of renormalization symmetry are appropriate for information sources, in particular information resonances? Equation (13) is taken by 'abduction' from physical systems. Other qualitatively different expressions may be possible and necessary.

How are the two points above related to the generalized Onsager relations which define the response of the system the Legendre transform of the source uncertainty of the underlying language? Onsager relations are serious business for technological applications; how are they influenced by neural architecture, as we have characterized it, in particular hierarchy and coevolutionary condensation or fragmentation?

Specification of state space algebra, renormalization symmetry and generalized Onsager relations may, in fact, be equivalent to specification of architecture.

Is there indeed a formal 'tuning theorem' inverse to the Shannon Coding Theorem for systems undergoing information resonance which would allow use of the internal correlations or other structures of scanned input signals to greatly improve the efficiency of learning paradigms for neural networks? Does 'error-free' pattern recognition loom for appropriate input rates?

The discussion thus far has been in terms of fairly abstract structures. Can explicit application be made to some of the current neural models? For example, what is the relation between this work and currently popular spin glass and other models of neural networks? If, however, it is not possible to identify explicit 'language' structures underlying or associated with these models, does that not, perhaps, suggest a serious weakness to those approaches?

This conjecture leads to the next point:

How can the development be generalized to nonergodic information sources? The discussion in Appendix 2 indicates that the primary difference between the two kinds of information sources is that the limit $H[\mathbf{X}]$ exists for all sources, but is independent of path $x = a_0, a_1, ...a_n, ...$ only for ergodic sources. What can be done otherwise? Three lines of attack seem obvious.

If the underlying state space can indeed be partitioned into disjoint equivalence classes of meaningful paths, then we may break the system into mutually disjoint information sources, and proceed as above, much like working with separate domains of attraction in a nonlinear system. We will call such a system 'disjointly' ergodic. Imposition of an order relation gives a 'natural' induced information source.

A second way of proceeding is 'locally,' i.e. imposing a 'manifold' structure on the underlying state space, the collection of paths x. That is, we assume a topology for the state space such that each path x has an open neighborhood which can be mapped by a homeomorphism onto a reference state space which has an ergodic information source. With appropriate topology, each open covering of the state space has a finite sub covering which patches the thing together in the standard manner (e.g. Sternberg, 1964; Thirring, 1992). This differential geometry approach is recognizably similar to, but would seem to generalize, use of an 'information metric' to derive asymptotic statistical results (Amari, 1982; Kass, 1989).

A somewhat different attack is to suppose that, given one particular highly probable path, x_0, we can reasonably define the source uncertainty associated with nearby paths in terms of their distance from x_0. Let $x = x_0 + \delta x$, where δx represents a 'small variation,' and make the usual formal series expansion near $H(x_0) \equiv H_0$ in terms of a generalized derivative:

$$H(x) = H_0 + \delta H(x) \approx H_0 + \frac{\delta H}{\delta x}\delta x,$$

where we assume $\delta H \ll H_0$.

We might well call such a system 'nearly' ergodic.

Extension of our results to 'slightly less than' ergodic information sources appears direct. It may well be, however, to expand the suggestion above, that arrays undergoing 'ergodic' information resonance, i.e. in duality with an ergodic information source, are of great interest precisely because of their intimate relation to 'language,' in the large sense, and if the most popular neural network and information resonance array models do not have an inherent association with language, this may well be a serious deficiency of current approaches. We suggest that language is utterly fundamental to any realistic understanding of neural process, and that treatments without direct involvement of language are missing the forest and, indeed, most of the trees.

The next question is implied by those previous: Suppose we specify the grammar and syntax of the underlying language, the associated state-space algebra, the nested renormalization or hierarchical order symmetry defining behavior at phase transition, and the generalized Onsager relations giving more subtle behaviors. Do these, then, specify an optimal architecture and learning paradigm? That is, can we create an algorithmic 'neural compiler' which will spit out 'optimal' circuit diagrams, given an appropriate specification of desired behaviors?

A principal outcome of our 'ordered hierarchy' analysis is the inference of a tradeoff between the capacity of an uncorrelated parallel multitasking structure for pattern recognition, and the more complicated behavioral dynamics possible for a hierarchically ordered system. This may well be a far deeper result than our specialized treatment might indicate.

Pastor et al. (2000) have recently proposed a purely algebraic phenomenological model for information processing in large-scale cerebral networks. It seems likely our results have some relation to theirs.

Finally, the quantum generalization we have proposed seems worthy of further exploration, particularly the search for a 'two fluid' model, although the current state of quantum information theory remains a serious constraint.

While some of these matters can be addressed, if not quite really answered, using the kinds of specific case-history models popular with physicists, more formal treatments seem necessary, if not precisely our program, then something recognizably similar.

In sum, the information resonance approach represents, in our view, a theoretical advance which could well translate into a broad spectrum of significant technology.

Appendix 1: 'Large Deviations' and entropy

We can place our development in the context of 'large deviations' as follows (Dembo and Zeitouni, 1998, p.2):

Let $X_1, X_2, ...X_n$ be a sequence of independent, standard Normal, real-valued random variables and let

$$S_n = \frac{1}{n}\sum_{j=1}^{n} X_j.$$

(46)

Since S_n is again a Normal random variable with zero mean and variance $1/n$, for all $\delta > 0$

$$\lim_{n\to\infty} P(|S_n| \geq \delta) = 0,$$

(47)

where P is the probability that the absolute value of S_n is greater or equal to δ. Some manipulation, however, gives

$$P(|S_n| \geq \delta) = 1 - \frac{1}{\sqrt{2\pi}}\int_{-\delta\sqrt{n}}^{\delta\sqrt{n}} \exp(-x^2/2)dx,$$

(48)

so that

$$\lim_{n\to\infty} \frac{\log P(|S_n| \geq \delta)}{n} = -\delta^2/2$$

(49)

We can rewrite this for large n as

$$P(|S_n| \geq \delta) \approx \exp(-n\delta^2/2).$$

(50)

That is, for large n, the probability of a large deviation in S_n follows something much like what follows from equation (2), i.e. that meaningful paths of length n all have approximately the same probability $P(n) \propto \exp(-nH[\mathbf{X}])$.

Our questions about 'meaningful paths' thus appear suddenly as formally isomorphic to one of the central developments in an emerging sector of applied probability termed 'large deviation theory,' which encompasses statistical mechanics, what the physicists call fluctuation theory, and information theory into a single structure (Dembo and Zeitouni, 1998).

A cardinal tenet of large deviation theory is that the 'rate function' $-\delta^2/2$ in equation (49) can often be expressed as a mathematical 'entropy' having the form

$$-\sum_k p_k \log p_k, \tag{51}$$

for some set of probabilities p_k. This result goes under various names at various levels of approximation – Sanov's Theorem, Cramer's Theorem, the Gartner-Ellis Theorem, the Shannon-McMillan Theorem, and so on (Dembo and Zeitouni, 1998).

Appendix 2: Ergodic and non-ergodic information sources

Following the treatment of Cover and Thomas, (1991, p. 474), the Shannon-McMillan Theorem on which we have based our analysis is predicated on having a stationary ergodic information source – one whose long-time pattern of emitted symbols follows the strong law of large numbers. An ergodic source is defined on some probability space $(\Omega, \mathcal{B}, \mu)$, where $\mathcal{B}$ is a sigma algebra of subsets of the space Ω and μ is a probability measure. A random variable X is defined as a function $X(\omega), \omega \in \Omega$, on the probability space. There is also a time translation operator, $T : \Omega \to \Omega$. Let μ be the probability measure of a set $A \in \mathcal{B}$. Then the transformation is *stationary* if $\mu(TA) = \mu(A)$ for all $A \in \Omega$. The transformation is *ergodic* if every set A such that $TA = A$ almost everywhere satisfies $\mu(A) = 0$ or 1. That is, almost everything flows.

If the transformation T is stationary and ergodic, we call the process defined by $X_n(\omega) = X(T^n\omega)$ stationary and ergodic.

For a stationary ergodic source with a finite expected value, the Ergodic Theorem concludes that

$$\frac{1}{n}\sum_{i=1}^{n} X_i(\omega) \to E(X) = \int X d\mu$$

with probability 1. This is the generalized law of large numbers for ergodic processes: the arithmetic mean in time converges to the mathematical expectation in 'space.'

Beginning here, after some considerable mathematical travail, the Shannon-McMillan Theorem, as we have described it, follows (Khinchine, 1957; Petersen, 1995; Cover and Thomas, 1991).

The essential point is that for a stationary, ergodic information source the limit

$$H[\mathbf{X}] = \lim_{n\to\infty} \frac{H[X_0, ...X_n]}{n+1}$$

not only exists, but *is independent of path.* That is, as $x = a_0, ..., a_n$ gets longer and longer, all paths converge to the same value of $H[\mathbf{X}]$ regardless of their origin or meandering. This is the fundamental information theory simplification, onto which we have imposed parametization and on which we have further grafted invariance under renormalization at phase transition as an expression of architecture.

A careful reading of the proof to the Shannon-McMillan Theorem (Khinchine, 1957; Petersen, 1995) shows that non-ergodic information sources still converge to some value $\lim_{n\to\infty} H(x)$, where x is a path of increasing length, but the value $H(x)$ is now *path dependent.* That is, each increasing path x converges to its own value of $H(x)$, depending, thus, on both the overall 'language' and on the particular path chosen.

Acknowledgments

This work benefited from support under an Investigator Award in Health Policy Research given by the Robert Wood Johnson Foundation and under NIEHS Grant 1-P50-ES09600-02.

References

Amari S, 1982, "Differential geometry of curved exponential families – curvature and information loss," *Annals of Statistics*, **10**, 357-387.

Ash R, 1990, *Information Theory*, Dover, New York.

Behrman E, J Niemel, J Steck and S Skinner, 1996, "A quantum dot neural network," *Proceedings of the Workshop on Physics of Computation*, New England Complex Systems Institute, Cambridge, MA, pp. 22-24.

Bennett C, 1988, "Logical depth and physical complexity." In *The Universal Turing Machine: A Half-Century Survey*, R Herkin ed., pp. 227-257, Oxford University Press, Oxford.

Binney J, N Dowrick, A Fisher and M Newman, 1995, *The theory of critical phenomena; An introduction to the renormalization group*, Oxford Science Publications, Oxford.

Boyd R and P Richerson, 1985, *Culture and Evolutionary Theory*, University of Chicago Press, Chicago.

Braiman Y, J Linder and W Ditto, 1995, "Taming spatiotemporal chaos with disorder," *Nature*, **378**, 465.

Caswell H, 1999, *Matrix Population Models*, Sinaur Associates, New York.

Cavalli-Sforza L and M Feldman, 1981, *Cultural Transmission and Evolution: A Quantitative Approach*, Monographs in Population Biology, 16, Princeton University Press, Princeton, NJ.

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

Cramer H, 1938, "Sur un nouveau theoreme-limite de la theorie des probabilities," in *Actualities Scientifiques et Industrielles*, No. 736 in Colloque consacre al la theorie des probabilities, pp. 5-23, Hermann, Paris.

Deco G and D Obradovic, 1996, *An Information-Theoretic Approach to Neural Computing*, Springer-Verlag, New York.

Deco G and B Schurmann, 1998, "Stochastic resonance in the mutual information between input and output spike trains of noisy central neurons," *Physica D*, **117**, 276-282.

Dembo A, O Zeitouni, 1998, *Large Deviations: Techniques and Applications, 2nd. Ed.*, Springer-Verlag, New York.

Durham W, 1991, *Coevolution: Genes, Culture and Human Diversity*, Stanford University Press, Palo Alto, CA.

Dykman M, D Luchinsky, P McClintock and V Smelyansky, 1996, "Corrals and critical behavior of the distribution of fluctuational paths," *Physical Review Letters*, **77**, 5229-5232.

Ellis R, 1985, *Large Deviations and Statistical Mechanics*, Springer-Verlag, New York.

Feller W, 1977, *An Introduction to Probability Theory and its Applications, Vol. II*, second edition, John Wiley and Sons, New York.

Feynman R and A Hibbs, 1965, *Quantum Mechanics and Path Integrals*, McGraw-Hill, New York, NY.

Feynman R, 1998, *Statistical Mechanics*, Perseus Books, Reading, MA.

Freidlin M and A Wentzell, 1998, *Random Perturbations of Dynamical Systems*, Springer-Verlag, New York.

Gammaitoni A, P Hanggi, P Jung and F Marchesoni, 1998, "Stochastic resonance," *Reviews of Modern Physics*, **70**, 223-287.

Godivier X and F Chapeau-Blondeau, 1998, "Stochastic resonance in the information capacity of a nonlinear dynamic system," *International Journal of Bifurcation and Chaos*, **8**, 581-589.

Granovetter M, 1973, "The strength of weak ties," *American Journal of Sociology*, **78**, 1360-1380.

Griffiths R, 1972, "Rigorous results and theorems" in *Phase Transitions and Critical Phenomena*, C Domb and M Green, eds., Academic Press, London.

Heneghan C, C Chow, J Collins, T Imhoff, S Lowen and M Teich, 1996, "Information measures quantifying aperiodic stochastic resonance," *Physical Review A*, **54**, 2366-2377.

Holevo A, 1973, "Some estimates for information quantity transmitted by quantum communication channels," *Problems of Information Transmission*, **9**, 177-183.

Holevo A, 1998, "Coding theorems for Quantum Channels," xyz.lanl.gov/quant-ph/9809023.

Ives A, 1995, "Measuring resilience in stochastic systems," *Ecological Monographs*, **65** 217-233.

Kass R, 1989, "The geometry of asymptotic inference," *Statistical Science*, **4**, 188-234.

Khinchine A, 1957, *The Mathematical Foundations of Information Theory*, Dover, New York.

King C and A Lesniewski, 1995, "Quantum sources and a quantum coding theorem," xxx.LANL.gov quant-phy 9511019.

Kadtke J and A Bulsara, 1997, *Applied Nonlinear Dynamics and Stochastic Systems Near the Millenium*, AIP Conference Proceedings, American Institute of Physics, New York.

Linder J, B Meadows and W Ditto, 1995, "Array enhanced stochastic resonance and spatiotemporal synchronization," *Physical Review Letters*, **75**, 3-6.

Linder J, B Meadows, W Ditto, M Inchiosa and A Bulsara, 1996, "Scaling laws for spatiotemporal synchronization and array enhanced stochastic resonance," *Physical Review A*, **53**, 2081-2086.

Luchnisky D, 1997, "On the nature of large fluctuations in equilibrium systems: observations of an optimal force," *Journal of Physics A Letters*, **30**, L577-583.

McCauley L, 1993, *Chaos, Dynamics and Fractals: an algorithmic approach to deterministic chaos*, Cambridge University Press, Cambridge.

McClintock and D Luchinsky, 1999, "Glorious noise," *The New Scientist*, **161**, No. 2168, January, 36-39.

Milne B, 1992, "Spatial aggregation and neutral models in fractal landscapes," *American Naturalist*, **139**, 32-57.

Neiman A, B Shulgin, V Anishchenko, W Ebeling, L Schimansky-Geier and J Freund, "Dynamical entropies applied to stochastic resonance," *Physical Review Letters*, **76**, 4299-4302.

Nieman A, B Shulgin, V Anishchenko, W Ebeling, L Schimansky-Gier and J Freund, "Correction," *Physical Review Letters*, **77**, 4851.

Onsager L and S Machlup, 1953, "Fluctuations and irreversible processes," *Physical Review*, **91**, 1501-1512.

Pastor J, M Lafon, L Trave-Massuyes, J Demonet, B Doyon and P Celsis, 2000, "Information processing in large-scale cerebral networks: the causal connectivity approach, *Biological Cybernetics*, **82**, 49-59.

Petersen K, 1995, *Ergodic Theory*, Cambridge University Press, Cambridge, UK.

Ritchie M and H Olff, 1999, "Spatial scaling laws yield a synthetic theory of biodiversity," *Nature*, **400**, 557-560.

Rojdestvenski I and M Cottam, 2000, "Mapping of statistical physics to information theory with applications to biological systems," *J. Theor. Biol.*, **202**, 43-54.

Schimansky-Gier L, J Freund, U Siewert and A Nieman, 1996, "Stochastic resonance: informational aspects and distributed systems," ICND-96 book of abstracts.

Schumacher B, 1995, "Quantum Coding," *Physical Review A*, **51**, 2738-2747.

Schumacher B, 1996, "Sending entanglement through noisy quantum channels," *Physical Review A*, **55**, 2614-2628.

Sternberg S, 1964, *Lectures on Differential Geometry*, Prentice-Hall, New York.

Thrring W, 1991, *Classical Dynamical Systems and Classical Field Theory*, 2nd. ed., Springer-Verlag, New York.

Toth G, C Lent, P Tougaw, Y Brazhnik, W Weng, W Porod, R Liu and Y Huang, 1996, "Quantum cellular neural networks," *Superlattices and Microstructures*, **20**, 473-478.

Wallace R, Y Huang, P Gould and D Wallace, 1997, "The hierarchical diffusion of AIDS and violent crime among US metropolitan regions: inner-city decay, stochastic resonance and reversal of the mortality transition," *Social Science and Medicine*, **44** 935-947.

Wallace R and RG Wallace, 1998, "Information theory, scaling laws and the thermodynamics of evolution," *Journal of Theoretical Biology*, **192**, 545-559.

Wallace R and RG Wallace, 1999, "Organisms, organizations and interactions: An information theory approach to biocultural evolution," *BioSystems*, **51**, 101-119.

Wallace R and J Ullmann, 1999, "Pentagon capitalism and the killing of the Red Queen: How the US lost the coevolutionary arms race between civilian firms and technology." Submitted.

Wallace R, 2000a, "Language and coherent neural amplification in hierarchical systems: Renormalization and the dual information source of a generalized spatiotemporal stochastic resonance," *International Journal of Bifurcation and Chaos*, **10**, 493-502..

Wallace R, 2000b , "Language and coherent neural amplification in hierarchical systems: 'Thermodynamics,' generalized Onsager relations, and the detection of 'abnormal' pattern." Submitted.

Wallace R, 2000c, "Quantum linguistics: information theory and quantum neural networks." Submitted.

Wallace R and M Fullilove, 1999, "Culturally-dependent canonical patterns of mental disorder in the context of socioeconomic stress: the mathematical epidemiology of madness." Submitted.

Wilson K, 1971, "Renormalization group and critical phenomena. I Renormalization group and the Kadanoff scaling picture," *Physical Review B*, **4**, 3174-3183.

Zurek W, 1985, "Cosmological experiments in superfluid helium?" *Nature*, **317**, 505-508.

Zurek W, 1996, "The shards of broken symmetry," *Nature*, **382**, 296-298.