

ADAPTATION, PUNCTUATION AND INFORMATION: A RATE-DISTORTION APPROACH TO NON-COGNITIVE 'LEARNING PLATEAUS' IN EVOLUTIONARY PROCESS

Rodrick Wallace

The New York State Psychiatric Institute and PISCS Inc.

Correspondence to: R Wallace, PISCS Inc., 549 W. 123 St., Suite 16F, New York, NY, 10027, USA. Telephone (212) 865-4766, email rdwall@ix.netcom.com

Received 30-X-2001

ABSTRACT

We extend recent information-theoretic phase transition approaches to evolutionary and cognitive process via the Rate Distortion and Joint Asymptotic Equipartition Theorems, in the circumstance of interaction with a highly structured environment. This suggests that learning plateaus in cognitive systems and punctuated equilibria in evolutionary process are formally analogous, even though evolution is not cognitive. Extending arguments by Adami *et al.* (2000), we argue that 'adaptation' is the process by which a distorted genetic image of a coherently structured environment is imposed upon a species.

Keywords: Adaptation, evolution, information theory, phase transition, punctuated equilibrium, rate distortion, renormalization, speciation.

1. INTRODUCTION

Recently Adami *et al.* (2000) applied an information theory approach to conclude that genomic complexity can be identified with the amount of information a gene sequence stores about its environment. Here we use a rate distortion argument, in the context of recent work on 'punctuation' of parametrized information sources, to obtain similar, but more wide-ranging, results.

Punctuation haunts evolutionary process. Ever since the benchmark paper by Gould and Eldredge (1977) describing evidence for 'punctuated equilibrium' - a result seemingly at some variance with purely adaptationist gene-centered views of evolution - a lively debate on the matter has raged at what Eldredge (1995) terms the 'high table of evolutionary theory.'

Direct extension of recent information theory approaches to cognition (Wallace, 2000a,b) suggests however, that even the most simplistic gene-centered view of evolution will give punctuation in a 'natural' manner.

The comparison of punctuated adaptation with cognitive learning plateaus is however, somewhat counterintuitive: Evolution is not a cognitive process, in the generally accepted sense. Cognition involves, at its foundation, an active selection of



one out of a complex repertory of possible responses to a sensory input, based on comparison with a learned internal representation of the outer world (e.g. Cohen, 1992, 2000, 2001; Atlan and Cohen, 1998). Although genes, or in the case of human biology, a composite of genes-and-culture (e.g. Richerson and Boyd, 1995, 1998; Durham, 1991), do indeed constitute a kind of 'memory' of past interaction with the world, response to selection pressure is not through direct comparison with that 'memory', but rather through the reproductive success of a random variation constrained by the path of evolutionary history. Even in the case of human biology, culture tends to be fairly rigid, and it can be argued that selection pressure usually dominates dynamics (e.g. Wallace and Wallace, 2000).

This is not cognition and there is no 'intelligent purpose' to adaptive or evolutionary process per se.

Nonetheless, selection pressures most often represent systematic patterns of interaction with an embedding and highly structured ecosystem in which each species is itself manifest: 'interpenetration,' to use the term made popular by Levins and Lewontin (2000). Adami *et al.* (2000) make an analogous argument for the influence of embedding ecology on gene complexity.

As a first and rather crude approximation, we take a 'gene-centered' view of reproduction as involving primarily the transmission of 'genetic information' within populations. Although, since it neglects factors of environment and development, this is a distorting oversimplification. Currently popular scientific paradigm involves examination of the 'genetic code,' and in the 1970s 'language' was taken as the underlying model for a Theory of General Biology by Waddington (1972) at the famous Villa Serbellion meetings. Our own evolutionary studies have been much in that direction, using a fairly straightforward extension of information theory methods (Wallace and Wallace, 1998, 1999).

While evolution is not a cognitive process, the critical roles of memory and 'language', in the largest sense, create a formal parallel with cognition which we will exploit to some effect in the exploration of punctuated evolutionary adaptation.

The point of intersection will prove to be the learning plateau.

Learning plateaus haunt cognitive systems. Successful immune response is predicated on sufficient exposure to antigen challenge to permit both mobilization and learning (Cohen, 1992, 2000; Atlan and Cohen, 1998): often fever and sickness ensue if the response is delayed - a plateau. Studying a new language, computer or human, is a frustrating experience as the learner fails to make progress until a 'breakthrough' occurs a 'learning plateau'. Learning to ride a bicycle is analogous. For those of us embedded in bureaucracies or active in community life, organizational learning seems glacial until some 'crisis' forces rapid adaptation and response a 'learning plateau'. Once learned however, the behavior becomes virtually permanent, and one, generally, never forgets an antigen, a language, nor how to pedal, balance, and steer.

Park *et al.* (2000) have explored a very general computer model of an artificial neural network: an array of feedforward multilayer perceptrons trained using a gradient descent error backpropagation algorithm. This is a system which inevitably suffers recalcitrant learning plateaus. They find that:

"Although there have been a lot of techniques for accelerating convergence [of network response to training pattern], most of them cannot solve the plateau problems..."

Park *et al.* go on to apply a steepest descent method to a loss function defined in the network tuning parameter space, based on an information geometry using a Riemannian metric defined in terms of the Fisher information matrix.

Very similar work has been published by Rose (1998), who addressed optimization problems using a deterministic annealing method in which the annealing process is formally equivalent to computation of the Shannon rate-distortion function. The annealing temperature is inversely proportional to the slope of the curve.

Elsewhere (Wallace, 2000a,b) have shown how cognitive pattern recognition-and-response can be characterized by a ‘dual information source,’ permitting, in a fashion recognizably similar to the Park and Rose papers, application of extremely general information theory arguments to the cognitive learning plateau problem. The approach is based on a canonical importation of renormalization methods from statistical mechanics to information theory which is much in the spirit of the Large Deviations Program of applied probability (e.g. Dembo and Zeitouni, 1998). Imposition of renormalization symmetry on the mutual information in the Rate Distortion Theorem gives a general learning plateau result equivalent to phase transformation in a highly ‘natural’ manner. For an evolutionary system this is equivalent to punctuated adaptation.

Some preliminary development is required.

2. ERGODIC INFORMATION SOURCES, THE SHANNON-MCMILLAN THEOREM, AND THE RATE DISTORTION THEOREM

Suppose we have an ordered set of random variables, X_k , at ‘times’ $k = 1, 2, \dots$, which we call \mathbf{X} , that emits sequences taken from some fixed alphabet of possible outcomes. Thus an output sequence of length n , x_n , termed a path, will have the form

$$x_n = (\alpha_0, \alpha_1, \dots, \alpha_{n-1})$$

where α_k is the value at step k of the stochastic variate X_k ,

$$X_k = \alpha_k.$$

A particular sequence x_n will have the probability

$$P(X_0 = \alpha_0, X_1 = \alpha_1, \dots, X_{n-1} = \alpha_{n-1}), \quad (1)$$

with associated conditional probabilities

$$P(X_n = \alpha_n \mid X_{n-1} = \alpha_{n-1}, \dots, X_0 = \alpha_0). \quad (2)$$

Thus substrings of x_n are not, in general, stochastically independent. That is, there may be powerful serial correlations along the x_n . We call \mathbf{X} an information source, and are particularly interested in sources for which the long run frequencies of strings converge stochastically to their time-independent probabilities, generalizing the law of large numbers. These we call *ergodic* (Ash, 1990; Cover and Thomas, 1991; Khinchine, 1957). If the probabilities of strings do not change in time, the source is called *memoryless*. We shall be interested in sources which can be parametrized and that are, with respect to those parameters, adiabatically *piecewise memoryless*, i.e. probabilities closely track parameter changes within a ‘piece’, but may change suddenly between pieces. This allows us to apply the simplest results from

information theory, and to use renormalization methods to examine transitions between ‘pieces’. Learning plateaus represent regions where, with respect to the parameter or parameters, the system is, to first approximation, memoryless in this sense. In what follows we use the term ‘ergodic,’ to mean ‘adiabatically piecewise memoryless ergodic’.

For any ergodic information source it is possible to divide all possible sequences of output, in the limit of large n , into two sets, S_1 and S_2 , having, respectively, very high and very low probabilities of occurrence. Sequences in S_1 we call *meaningful*.

The content of information theory’s Shannon-McMillan Theorem is twofold:

First, if there are $N(n)$ meaningful sequences of length n , where $N(n) \ll$ than the number of all possible sequences of length n , then, for each ergodic information source \mathbf{X} , there is a unique, path-independent number $H[\mathbf{X}]$ such that

$$\lim_{n \rightarrow \infty} \frac{\log[N(n)]}{n} = H[\mathbf{X}] \quad (3a)$$

See Ash (1990), Cover and Thomas (1991) or Khinchine (1957) for details.

Thus, for large n , the probability of *any* meaningful path of length $n \gg 1$, independent of path, is approximately

$$P(x_n \in S_1) \propto \exp(-nH[\mathbf{X}]) \propto 1/N(n). \quad (3b)$$

This is the *asymptotic equipartition property* and the Shannon-McMillan Theorem is often called the Asymptotic Equipartition Theorem (AEPT).

$H[\mathbf{X}]$ is the *splitting criterion* between the two sets S_1 and S_2 , and the second part of the Shannon-McMillan Theorem involves its calculation. This requires introduction of some nomenclature.

Suppose we have stochastic variables X and Y which take the values x_j and y_k with probability distributions

$$P(X = x_j) = P_j$$

$$P(Y = y_k) = P_k.$$

Let the joint and conditional probability distributions of X and Y be given, respectively, as

$$P(X = x_j, Y = y_k) = P_{j,k}$$

$$P(Y = y_k | X = x_j) = P(y_k | x_j).$$

The *Shannon uncertainties* of X and of Y are, respectively

$$\begin{aligned} H(X) &= - \sum_j P_j \log(P_j) \\ H(Y) &= - \sum_k P_k \log(P_k). \end{aligned} \quad (4)$$

The *joint uncertainty* of X and Y is defined as

$$H(X, Y) = - \sum_{j,k} P_{j,k} \log(P_{j,k}). \quad (5)$$

The *conditional uncertainty* of Y given X is defined as

$$H(Y|X) = - \sum_{j,k} P_{j,k} \log[P(y_k | x_j)]. \quad (6)$$

Note that by expanding $P(y_k | x_j)$ we obtain

$$H(X|Y) = H(X,Y) - H(Y).$$

The second part of the Shannon-McMillan Theorem states that the (path independent) splitting criterion, $H[\mathbf{X}]$, of the ergodic information source \mathbf{X} , which divides high from low probability paths, is given in terms of the sequence probabilities of equations (1) and (2) as

$$H[X] = \lim_{n \rightarrow \infty} H(X_n | X_0, X_1, \dots, X_{n-1}) = \lim_{n \rightarrow \infty} \frac{H(X_0, \dots, X_n)}{n+1}. \quad (7)$$

The AEPT is one of the most profound probability limit theorems of 20th Century applied mathematics.

Ash (1990) describes the uncertainty of an ergodic information source as follows:

“...[W]e may regard a portion of text in a particular language as being produced by an information source. The probabilities $P[X_n = \alpha_n | X_0 = \alpha_0, \dots, X_{n-1} = \alpha_{n-1}]$ may be estimated from the available data about the language. A large uncertainty means, by the AEPT, 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 can affect each other, or, equivalently, systems can translate from one language to another, usually with error. The Rate Distortion Theorem, which generalized the SMT, describes how this can take place. As Cohen (2001) has put it, in the context of the cognitive immune system:

“An immune response is like a key to a particular lock; each immune response amounts to a functional image of the stimulus that elicited the response. Just as a key encodes a functional image of its lock, an effective [immune] response encodes a functional image of its stimulus; the stimulus and the response fit each other. The immune system, for example, has to deploy different types of inflammation to heal a broken bone, repair an infarction, effect neuroprotection, cure hepatitis, or contain tuberculosis. Each aspect of the response is a functional representation of the challenge.

Self-organization allows a system to adapt, to update itself in the image of the world it must respond to... The immune system, like the brain... aim[s] at representing a part of the world.”

These considerations suggest that the degree of possible back-translation between the world and its image represents the profound and systematic coupling between a biological system and its environment, a coupling which may particularly express the way in which the system has ‘learned’ the environment.

To reiterate, in an evolutionary context, Adami *et al.* (2000) make a roughly analogous argument that:

“A recent information-theoretic (but intuitively evident) definition identifies genomic complexity with the amount of information a [gene] sequence stores about its environment.”

We attempt a formal treatment, from which it will appear that such a process is, almost inevitably, highly punctuated by ‘learning plateaus’.

Suppose we have an ergodic information source \mathbf{Y} , a generalized language having grammar and syntax, with a source uncertainty $H[\mathbf{Y}]$ that ‘perturbs’ a system of interest. A chain of length n , a path of perturbations, has the form:

$$y^n = y_1, \dots, y_n.$$

Suppose that chain elicits a corresponding chain of responses from the system of interest, producing another path $b^n = (b_1, \dots, b_n)$, which has some ‘natural’ translation into the language of the perturbations, although not, generally, in a one-to-one manner. The image is of a continuous analog audio signal which has been ‘digitized’ into a discrete set of voltage values. Thus, there may well be several different y^n corresponding to a given ‘digitized’ b^n . Consequently, in translating back from the b-language into the y-language, there will generally be information loss.

Suppose however, that with each path b^n we specify an inverse code which identifies exactly one path \hat{y}^n . We assume further there is a measure of distortion which compares the real path y^n with the inferred inverse \hat{y}^n . Below we follow the nomenclature of Cover and Thomas (1991).

The *Hamming distortion* is defined as

$$\begin{aligned} d(y, \hat{y}) &= 1, y \neq \hat{y} \\ d(y, \hat{y}) &= 0, y = \hat{y}. \end{aligned} \quad (8)$$

For continuous variates the *Squared error distortion* is defined as

$$d(y, \hat{y}) = (y - \hat{y})^2. \quad (9)$$

Possibilities abound.

The distortion between paths y^n and \hat{y}^n is defined as

$$d(y^n, \hat{y}^n) = (1/n) \sum_{j=1}^n d(y_j, \hat{y}_j). \quad (10)$$

We suppose that with each path y^n and b^n -path translation into the y-language, denoted \hat{y}^n , there are associated individual, joint and conditional probability distributions $p(y^n)$, $p(\hat{y}^n)$, $p(y^n, \hat{y}^n)$ and $p(y^n | \hat{y}^n)$.

The *average distortion* is defined as

$$D = \sum_{y^n} p(y^n) d(y^n, \hat{y}^n). \quad (11)$$

It is possible, using the distributions given above, to define the information transmitted from the incoming Y to the outgoing \hat{Y} process in the usual manner, using the appropriate Shannon uncertainties:

$$I(Y, \hat{Y}) \equiv H(Y) - H(Y | \hat{Y}) = H(Y) + H(\hat{Y}) - H(Y, \hat{Y}). \quad (12)$$

If there is no uncertainty in Y given \hat{Y} , then no information is lost. In general, this will not be true.

The *information rate distortion* function $R(D)$ for a source Y with a distortion measure $d(y, \hat{y})$ is defined as

$$R(D) = \min_{p(y|\hat{y}); \sum_{(y,\hat{y})} p(y)p(y|\hat{y})d(y,\hat{y}) \leq D} I(Y, \hat{Y}) \quad (13)$$

where the minimization is over all conditional distributions $p(y|\hat{y})$ for which the joint distribution $p(y, \hat{y}) = p(y)p(y|\hat{y})$ satisfies the average distortion constraint.

The Rate Distortion Theorem states that $R(D)$, as we have defined it, is the maximum achievable rate of information transmission which does not exceed distortion D . Note that the result is *independent of the exact form of the distortion measure* $d(y, \hat{y})$.

More to the point however, is the following: Pairs of sequences (y^n, \hat{y}^n) can be defined as *distortion typical*, that is, for a given average distortion D , pairs of sequences can be divided into two sets, a high probability one containing a relatively small number of (matched) pairs with $d(y^n, \hat{y}^n) \leq D$, and a low probability one containing most pairs. As $n \rightarrow \infty$ the smaller set approaches unit probability, and we have for those pairs the condition

$$p(\hat{y}^n) \geq p(\hat{y}^n | y^n) \exp[-nI(Y, \hat{Y})]. \quad (14)$$

Thus, roughly speaking, $I(Y, \hat{Y})$ embodies the splitting criterion between high and low probability pairs of paths. These pairs are, again, the input ‘training’ paths and corresponding output path. Dimitrov and Miller (2001) make extensive use of this formalism in their treatment of neural networks.

For the theory we will explore later, then, $I(Y, \hat{Y})$ plays the role of H in the formalism of the next section.

Note that, in the absence of a distortion measure, which is a very specific relation associated with a particular learning paradigm, a more general result much like equation (14) remains true for two interacting information sources, involving their mutual information as the splitting criterion. This is the principal content of the Joint Asymptotic Equipartition Theorem (JAEPT). See Cover and Thomas, (1991), Theorem 8.6.1, for details. Thus the imposition of a distortion measure results in a limitation in the number of possible *jointly typical* sequences defined by that theorem to those satisfying the distortion criterion, the basis of our strong analogy with neural network learning. Extension of our results is direct using the JAEPT.

3. PHASE TRANSITION AND COEVOLUTIONARY CONDENSATION

The essential homology relating information theory to statistical mechanics and nonlinear dynamics is twofold (Wallace and Wallace, 1998, 1999; Rojdestvenski and Cottam, 2000):

(1) A ‘linguistic’ equipartition of probable paths consistent with the Shannon-McMillan, JAEP, and Rate Distortion Theorems serves as the formal connection with

nonlinear mechanics and fluctuation theory, a matter we will not fully explore here, and

(2) A correspondence between information source uncertainty and statistical mechanical free energy density, rather than entropy. See Wallace and Wallace (1998, 2000) for a fuller discussion of the formal justification for this assumption, described by Bennett (1988) as follows:

“...[T]he value of a message is the amount of mathematical or other work plausibly done by the originator, which the receiver is saved from having to repeat.”

This is a central insight, consistent, we believe, with the conclusion of Adami *et al.* (2000) that

“Perhaps a key aspect of information theory is that information cannot exist in a vacuum; that is, information is physical... This statement implies that information must have an instantiation (be it ink on paper, bits in a computer memory, or even the neurons in a brain)... [an] arrangement of symbols... acquires the status of information only when its correspondence, or correlation, to other physical objects is revealed.”

The definition of the free energy density for a parametrized physical system is

$$F(K_1, \dots, K_m) = \lim_{V \rightarrow \infty} \frac{\log[Z(K_1, \dots, K_m)]}{V} \quad (15)$$

where the K_j are parameters, V is the system volume and Z is the ‘partition function’ defined from the energy function, the Hamiltonian, of the system.

For an ergodic information source the equivalent relation associates source uncertainty with the number of ‘meaningful’ sequences $N(n)$ of length n , in the limit

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

We will *parametrize* the information source to obtain the crucial expression on which our version of information dynamics will be constructed:

$$H[K_1, \dots, K_m, \mathbf{X}] = \lim_{n \rightarrow \infty} \frac{\log[N(K_1, \dots, K_m)]}{n}. \quad (16)$$

The essential point is that while information systems do not have ‘Hamiltonians’ allowing definition of a ‘partition function’ and a free energy density, they may have a source uncertainty obeying a limiting relation like that of free energy density. Importing ‘renormalization’ symmetry gives phase transitions at critical points (or surfaces), and importing a Legendre transform in a ‘natural’ manner gives dynamic behavior far from criticality. Only the first will be needed to solve the problems we wish to address here.

As neural networks demonstrate so well, it is possible to build larger pattern recognition systems from assemblages of smaller ones. We abstract this process in terms of a generalized linked array of subcomponents which ‘talk’ to each other in two different ways. These we take to be ‘strong’ and ‘weak’ ties between subassemblies. ‘Strong’ ties are, following arguments from sociology (Granovetter, 1973), those which permit disjoint partition of the system into equivalence classes. Thus the strong ties are associated with some reflexive, symmetric, and transitive relation between

components. ‘Weak’ ties do not permit such disjoint partition. In a physical system these might be viewed, respectively, as ‘local’ and ‘mean field’ coupling.

We are, thus, concerned with languages ‘spoken’ on an underlying network, be it chemical, neural, social, ecological, or some mix of these. The network will be manifest in the properties of any language ‘spoken’ on it, and vice versa, if language process can affect network properties. It is this composite, interactive, phenomenon we wish to model.

We fix the magnitude of strong ties, but vary the index of weak ties between components, which we call P , taking $K = 1/P$.

We assume the ergodic information source depends on three parameters, two explicit and one implicit. The explicit are K as above and an ‘external field strength’ analog J , which gives a ‘direction’ to the system. We may, in the limit, set $J = 0$.

The implicit parameter, which we call r , is an inherent generalized ‘length’ on which the phenomenon, including J and K , are defined. That is, we can write J and K as functions of averages of the parameter r , which may be quite complex, having nothing at all to do with conventional ideas of space, for example degree of niche partitioning in ecosystems or separation in social structures.

For a given generalized language of interest with a well defined ergodic source uncertainty H we write:

$$H[K, J, \mathbf{X}].$$

Imposition of invariance of H under a renormalization transform in the implicit parameter r leads to expectation of both a critical point in K , which we call 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 . Addition of other parameters to the system, e.g. some Q , results in a ‘critical line’ or surface $K_C(Q)$.

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 dominated by ‘strong’ ties. We begin by averaging across r -space in terms of ‘clumps’ of length R , defining J_R, K_R as J, K for $R = 1$. Then, following Wilson’s (1971) physical analog, we choose the renormalization relations as:

$$H[K_R, J_R, \mathbf{X}] = R^{\mathbf{D}} H[K, J, \mathbf{X}], \quad \chi(K_R, J_R) = \frac{\chi(K, J)}{R} \quad (17)$$

where \mathbf{D} is a non-negative real constant, possibly reflecting 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 generalized language, grows as $R^{\mathbf{D}}$, while the second just states that the correlation length simply scales as R .

Other, very subtle, symmetry relations (not necessarily based on elementary physical analogs) may well be possible. For example McCauley, (1993, p.168) describes the counterintuitive renormalization relations needed to understand phase transition in simple ‘chaotic’ systems.

For K near K_C , if $J \rightarrow 0$, a simple series expansion and some clever algebra (e.g. Wilson, 1971; Binney *et al.*, 1995; Wallace and Wallace, 1998) gives

$$H = H_0 \kappa^{s\mathbf{D}}, \quad \chi = \chi_0 \kappa^{-s} \quad (18)$$

where s is a positive constant.

Note that the algebraic argument leading from our equations (17) to our equations (18) is very precisely the same as that in Wilson's (1971) famous paper which leads from his equations (4) and (5) to his (22) and (23). We provide details in a calculational appendix.

Some rearrangement produces, near K_C ,

$$H \propto \frac{1}{\chi^D}. \quad (19)$$

This relation implies that the 'richness' of the generalized language is inversely related to the domain dominated by disjointly partitioning strong ties near criticality. As the nondisjunctive weak ties coupling declines, the efficiency of the coupled system as an information channel declines precipitously near the transition point: see (e.g.) Ash (1990) for discussion of the relation between channel capacity and information source uncertainty.

Further from the critical point matters are more complicated, involving 'Generalized Onsager Relations' and a kind of thermodynamics associated with a Legendre transform. We do not pursue that discussion here, which would lead to a study of 'evolutionary dynamics' far from punctuation.

The essential insight is that *regardless of the particular renormalization symmetries involved, sudden critical point transition is possible in the opposite direction for this model*, that is, from a number of independent, isolated and fragmented systems operating individually and more or less at random, into a single large, interlocked, coherent structure, once the parameter K , the inverse strength of weak ties, falls below threshold, or, conversely, once the strength of weak ties parameter $P = 1/K$ becomes large enough.

Thus, increasing weak ties between them can bind several different network-based 'language' processes into a single, embedding hierarchical metalanguage which contains the different languages as linked subdialects.

This heuristic insight can be made exact using a rate distortion argument (or, more generally, using the JAEPT):

Suppose that two ergodic information sources \mathbf{Y} and \mathbf{B} begin to interact, to 'talk' to each other, i.e. to influence each other in some way so that it is possible, for example, to look at the output of \mathbf{B} (strings b) and infer something about the behavior of \mathbf{Y} from it (strings y). We suppose it possible to define a retranslation from the B-language into the Y-language through a deterministic code book, and call $\hat{\mathbf{Y}}$ the translated information source, as mirrored by \mathbf{B} .

Take some distortion measure d comparing paths y to paths \hat{y} , defining $d(y, \hat{y})$. We invoke the Rate Distortion Theorem's mutual information $I(Y, \hat{Y})$, which is the splitting criterion between high and low probability pairs of paths. Impose, now, a parametrization by an inverse coupling strength K , and a renormalization symmetry representing the global structure of the system coupling. This may be much different from the renormalization behavior of the individual components. If $K < K_C$, where K_C is a critical point (or surface), the two information sources will be closely coupled enough to be characterized as condensed.

Wallace and Wallace (1998, 1999) use this approach to address speciation, coevolution and group selection in a relatively unified fashion.

We have, however, now constructed enough machinery to obtain our principal results in a deceptively direct and ‘obvious’ manner.

4. NON-COGNITIVE ‘LEARNING PLATEAUS’ IN EVOLUTIONARY PROCESS

We suppose a self-reproducing system (more specifically a linked, and in the large sense coevolutionary, condensation of several such systems) is exposed to a structured pattern of selective environmental pressures to which it must adapt if it is to survive. From that adaptive selection, changes in genotype and phenotype, we can infer, in a direct manner, something, but not everything, of the form of the structured system of selection pressures. We suppose the system of selection pressures to have sufficient grammar and syntax so as to itself constitute a piecewise ergodic information source Y whose probabilities are fixed on the timescale of analysis. The output of that system, B , is backtranslated into the ‘language’ of Y , and we call that translation \hat{Y} . The rate distortion behavior relating Y and \hat{Y} , is, according to the RDT, determined by the mutual information $I(Y, \hat{Y})$.

We take there to be a measure of the ‘strength’ of the selection pressure, P , which we use as an index of coupling with the species of interest, having an inverse $K = 1/P$, and write:

$$I(Y, \hat{Y}) = I[K]. \quad (20)$$

P might be measured by the rate of ‘cropping’ by predators, or the response to extreme environmental perturbation, and so on.

$I[K]$ thus defines the splitting criterion between high and low probability pairs of input and output paths for a specified average distortion D , and is analogous to the parameterized information source uncertainty upon which we imposed renormalization symmetry to obtain phase transition.

We thus interpret the sudden changes in the measured average distortion $D \equiv \sum p(y)d(y, \hat{y})$ which determines ‘mean error’ between pressure and response, i.e. the *ending* of a ‘learning plateau’, as representing onset of a phase transition in $I[K]$ at some critical K_C , consonant with our earlier developments.

Note that $I[K]$ constitutes an interaction between the species of interest and the impinging ecosystem’s selection pressure, so that its properties may be quite different from those of the individual or conjoined subcomponents.

From this viewpoint highly punctuated ‘non-cognitive learning plateaus’ are an inherently ‘natural’ phase transition behavior of evolutionary systems. While one may perhaps, in the sense of Park *et al.* (2000), find more efficient ‘gradient learning algorithms’, our development suggests plateaus will be both ubiquitous and highly characteristic of evolutionary process or pathway. Indeed, it seems likely that proper analysis of evolutionary plateaus, to the extent they can be observed or reconstructed, will give deep insight into the mechanisms underlying that system.

5. DISCUSSION AND CONCLUSIONS

As a reviewer has noted, Wilson’s theory deals formally with Hamiltonian physical systems and within certain ensembles of equilibrium thermodynamics. We are, in the

spirit of the Large Deviations Program of applied probability (e.g. Dembo and Zeitouni, 1998), extending that theory by imposing invariance under renormalization at ‘critical points’ of driving parameters for language-on-network structures. Such analogy, in the reviewer’s words, is a very useful instrument, but sometimes taking a certain mathematical technique out of context may cause transcendental artifacts. While we cannot promise that our particular development is without such artifacts, nonetheless, empirical studies in sociolinguistics have come to focus heavily on the role of Granovetter’s (1973) strong and weak ties, and the forms and dynamics of spoken language (commonalities, differences, and processes of fragmentation) can be almost entirely understood in terms of the weak and strong tie structure of the underlying social networks in a unified manner (e.g. Labov, 1966, 1980, 1986; Milroy, 1992; Milroy; 1987; Trudgill and Cheshire, 1998). We are thus led to suggest, via our mathematical model, a similar program for other language-and-network structures, including those subject to adaptive selection pressures.

More broadly however, the mathematical ecologist Pielou has noted particularly severe constraints on the utility of mathematical models in the study of complex ecosystem phenomena (Pielou, 1977, p. 106):

“...[Mathematical] models are easy to devise; even though the assumptions of which they are constructed may be hard to justify, the magic phrase ‘let us assume that...’ overrides objections temporarily. One is then confronted with a much harder task: How is such a model to be tested? The correspondence between a model’s predictions and observed events is sometimes gratifyingly close but this cannot be taken to imply the model’s simplifying assumptions are reasonable in the sense that neglected complications are indeed negligible in their effects...

In my opinion the usefulness of models is great... [however] it consists *not in answering questions but in raising them*. Models can be used to inspire new field investigations and these are the only source of new knowledge as opposed to new speculation.”

Our analysis thus creates a new body of model-based speculation that empirical study of the transition between plateaus of evolutionary or adaptive punctuation may permit identification of at least a local and temporary ‘universality’. This would seem to apply both to speciation and coevolution, taken as inverse phenomena in the sense of Wallace and Wallace (1998), i.e. splitting vs. coagulation of information sources.

Just as learning plateaus will always haunt theories of cognitive systems, so too their non-cognitive, highly punctuated analogs will continue to haunt theories of evolutionary process, as ecosystem regularities write themselves onto species’ genetic structure through ‘adaptation’ in much the manner that Adami *et al.* (2000) have proposed.

ACKNOWLEDGEMENTS

The author thanks a reviewer for comments useful in revision. This work benefited from support under NIEHS Grant I-P50-ES09600-03 and from previous monies under an Investigator Award in Health Policy Research from the Robert Wood Johnson Foundation.

MATHEMATICAL APPENDIX

We have supposed an ergodic information source \mathbf{X} characterizes the reproduction and/or persistence of a population, organization, language or other structure defined ‘on’ a network. Assume now that the source uncertainty $H[K, J, Q, \mathbf{X}]$ depends explicitly on parameters J, K and Q , and implicitly on an embedding manifold metric r , which may be the ‘distance’ between network nodes in some abstract space, that might, in fact, be fractal.

We assume it possible to redefine characteristics of the information source \mathbf{X} as functions of averages of the metric r , which we write as R . We ‘renormalize’ by clustering the entire system in terms of blocks of different sized R .

Let $N(K, J, Q, n)$ be the number of ‘meaningful’ correlated sequences of length n across the entire community in the r -manifold, given parameter values K, J, Q . We study changes in

$$H[K, J, Q, \mathbf{X}] \equiv \lim_{n \rightarrow \infty} \frac{\log[N(K, J, Q, n)]}{n}$$

as $K \rightarrow K_C$ and/or $Q \rightarrow Q_C$ for critical values K_C, Q_C at which the community begins to undergo a marked transformation from one kind of structure to another.

Given the metric r , a correlation length, $\chi(K, J, Q)$, can be defined as the average length in r -space over which structures involving a particular phase dominate.

We begin by clumping the community into blocks of average size R in the multivariate r -manifold, the ‘space’ in which the reproducing structure is implicitly embedded.

Following Wilson (1971) we impose renormalization symmetry on the source uncertainty on H and χ by assuming at transition the relations of equation (17) hold:

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

$$\chi(K_R, J_R, Q_R) = \frac{\chi(K, J, Q)}{R}$$

K_R, J_R and Q_R are the transformed values of K, J and Q after the clumping of renormalization. We take $K_I, J_I, Q_I \equiv K, J, Q$ and permit the characteristic exponent \mathbf{D} to be nonintegral.

These equations are assumed to hold in a neighborhood of the transition values K_C and Q_C .

Differentiating these with respect to R gives complicated expressions for dK_R/dR , dJ_R/dR and dQ_R/dR depending simply on R which we write as

$$\begin{aligned} dK_R / dR &= \frac{u(K_R, J_R, Q_R)}{R} \\ dQ_R / dR &= \frac{w(K_R, J_R, Q_R)}{R} \\ dJ_R / dR &= \frac{v(K_R, J_R, Q_R)}{R} J_R. \end{aligned} \quad (21)$$

Solving these differential equations we obtain K_R, J_R and Q_R as functions of J, K, Q and R .

Substituting back into equations (17) and expanding in a first order Taylor series near the critical values K_C and Q_C gives power laws much like the Widom-Kadanoff relations for physical systems. For example, letting $J = Q = 0$ and taking $\kappa \equiv (K_C - K)/K_C$ we obtain, in first order near K_C :

$$H = \kappa^{D_s} H_0, \quad \chi = \kappa^{-s} \chi_0$$

where $s > 0$ is a constant arising from the series expansion.

The essential fact is that we have only two fundamental equations, (17), in $n \geq 2$ unknowns: The critical ‘point’ is, in this formulation, most likely to be a complicated implicitly defined ‘critical surface’ in J, K, Q, \dots -space. The ‘external field strength’ J remains distinguished in this treatment, but neither K, Q nor other parameters are, by themselves, fundamental, rather their joint interaction defines critical behavior along this surface.

That surface is a fundamental object, not the particular set of parameters (except for J) used to define it, which may be subject to any set of transformations which leave the surface invariant. Thus ‘weak ties’, or whatever other parameters may be identified as affecting reproduction, are inextricably intertwined and mutually interacting, according to the form of this ‘evolutionary surface.’ That surface, in turn, is unlikely to remain fixed, and should vary with time or other extrinsic parameters, including, but not likely limited to, J .

At the critical surface a Taylor expansion of the renormalization equations (17) gives a first order matrix of derivatives whose eigenstructure defines fundamental system behavior. For physical systems the surface is a ‘saddle point’ (Wilson, 1971), but more complicated behavior seems likely in what we study.

Taking, for the moment, the simplest formulation, ($Q = 0, J \rightarrow 0$), as K increases toward a threshold value K_C , the source uncertainty of the reproductive, behavioral or other language common across the network declines and, at K_C , the average regime dominated by strong ties expands without limit. That is, the system begins to ‘freeze’ into one having a large correlation length for strong ties. The two phenomena are linked at criticality in physical systems by the scaling exponent s which may be exceedingly difficult to observe in natural systems, but might be more easily seen in the organizational dynamics of the social or economic systems for which elaborate administrative data are kept.

Assume the rate of change of $\kappa \equiv (K_C - K)/K_C$ remains constant, $|d\kappa/dt| = 1/\tau_K$. Analogs with physical theory suggest there is a characteristic time constant for the transition to a system dominated by strong ties, $\tau \equiv \tau_0/\kappa$, such that if changes in κ take place on a timescale longer than τ for any given κ , we may expect the correlation length $\chi \equiv \chi_0/\kappa^{-s}$, will be in equilibrium with internal changes and result in a very large fragment in r -space dominated by ‘strong’ ties. Some algebra gives the average fragment size, d , as

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

with $s > 0$. The more rapidly K approaches K_C the smaller is τ_K and the smaller and more numerous are the resulting fragments dominated by strong ties. A larger number of fragments suggests a higher probability of subsequent speciation, but small

fragments seem more likely to risk extinction. Wallace and Wallace (1998) discuss these matters in more detail.

REFERENCES

- Adami, C., C. Ofria and T. Collier (2000). Evolution of biological complexity. *Proceedings of the National Academy of Sciences* 97: 4463-4468.
- Ash, R. (1990). *Information Theory*. Dover Publications, New York.
- Atlan, H. and I.R. Cohen (1998). Immune information, self-organization and meaning. *International Immunology* 10: 711-717.
- Bennett, C. (1988). Logical depth and physical complexity. In: Herkin, R. (Ed). *The Universal Turing Machine: A Half-Century Survey*. Oxford University Press, pp. 227-257.
- Binney, J., N. Dowrick, A. Fisher and M. Newman (1995). *The theory of critical phenomena*. Clarendon Press, Oxford.
- Cohen, I.R. (1992). The cognitive principle challenges clonal selection. *Immunology Today* 13: 441-444.
- Cohen, I.R. (2000). *Tending Adam's Garden: evolving the cognitive immune self*. Academic Press, New York.
- Cohen, I.R. (2001). Immunity, set points, reactive systems, and allograft rejection. To appear.
- Cover, T. and J. Thomas (1991). *Elements of Information Theory*. Wiley, New York.
- Dembo, A. and O. Zeitouni (1998). *Large Deviations: Techniques and Applications*, 2nd Ed.. Springer-Verlag, New York.
- Dimitrov, A. and J. Miller (2001). Neural coding and decoding: communication channels and quantization. *Network: Computation on Neural Systems* 12: 441-472.
- Durham, W. (1991). *Coevolution: Genes, Culture and Human Diversity*. Stanford University Press, Palo Alto, CA.
- Eldredge, N. (1995). *Reinventing Darwin: the great debate at the high table of evolutionary theory*. John Wiley and Sons, New York.
- Gould, S. and N. Eldredge (1977). Punctuated equilibria: the tempo and mode of evolutionary reconsidered. *Paleobiology* 3: 115-151.
- Granovetter, M. (1973). The strength of weak ties. *American Journal of Sociology* 78: 1360-1380.
- Khinchine, A. (1957). *The Mathematical Foundations of Information Theory*. Dover Publications, New York.
- Labov, W. (1966). *The Social Stratification of English in New York City*. Center for Applied Linguistics, Washington, DC.
- Labov, W. (ed.) (1980). *Locating Language in Time and Space*. Academic Press, New York.
- Labov, W. (1986). Language structure and social structure, in S. Lindenberg *et al.* (eds.), *Approaches to Social Theory*. Russell Sage, New York.
- Lewontin, R. (2000). *The Triple Helix: Gene, Organism and Environment*. Harvard University Press, Cambridge, MA.
- McCauley, L. (1993). *Chaos, Dynamics, and Fractals: an algorithmic approach to deterministic chaos*. Cambridge University Press, Cambridge, UK.
- Milroy, L. (1987). *Language and Social Networks*. second edition, Blackwell, Cambridge, USA.
- Milroy, J. (1992). *Linguistic Variation and Change*. Blackwell, Cambridge, USA.
- Park H., S. Amari and K. Fukumizu (2000). Adaptive natural gradient learning algorithms for various stochastic models. *Neural Networks* 13: 755-764.
- Pielou, E. (1977). *Mathematical Ecology*. John Wiley and Sons, New York.
- Richerson, P. and R. Boyd (1995). The evolution of human hypersociality. Paper for Ringberg Castle Symposium on Ideology, Warfare and Indoctrinability (January, 1995), and for HBES meeting, 1995.

- Richerson, P. and R. Boyd (1998). Complex societies: the evolutionary origins of a crude superorganism. to appear.
- Rojdestvenski, I. and M. Cottam (2000). Mapping of statistical physics to information theory with applications to biological systems. *Journal of Theoretical Biology* 202: 43-54.
- Rose, K. (1998). Deterministic annealing for clustering, compression, classification, regression and related optimization problems. *Proceedings of the IEEE* 86: 2210-2239.
- Trudgill, P. and J. Cheshire (eds.) (1998). *The Sociolinguistics Reader*. Arnold, New York.
- Waddington, C. (1972). Epilog. In Waddington C, (Ed.) *Toward a Theoretical Biology* 4: Essays. Aldin-Atherton, Chicago.
- Wallace, D. and R. Wallace (2000). Life and death in Upper Manhattan and the Bronx: toward an evolutionary perspective on catastrophic social change. *Environment and Planning A* 32: 1245-1266.
- 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). Information resonance and pattern recognition in classical and quantum systems: toward a 'language model' of hierarchical neural structure and process. www.ma.utexas.edu/mp_arc-bin/mpa?yn=00-190.
- Wallace, R. and R.G. Wallace (1998). Information theory, scaling laws and the thermodynamics of evolution. *Journal of Theoretical Biology* 192: 545-559.
- Wallace, R. and R.G. Wallace (1999). Organisms, organizations and interactions: an information theory approach to biocultural evolution. *BioSystems* 51: 101-119.
- Wilson, K. (1971). Renormalization group and critical phenomena. I Renormalization group and the Kadanoff scaling picture. *Physical Review B* 4: 3174-3183.