PRELIMINARY DRAFT

Do not cite or circulate without permission

Adaptation, punctuation, and rate distortion: non-cognitive 'learning plateaus' in evolutionary process

Rodrick Wallace, PhD

The New York State Psychiatric Institute

and

PISCS Inc.*

May, 2001

Abstract

Extending recent information-theoretic phase transition approaches to evolutionary and cognitive process via the Rate Distortion Theorem in the circumstance of interaction with a structured environment suggests that learning plateaus in cognitive systems and punctuated equilibria in evolutionary process are formally analogous, even though evolution is most certainly not cognitive. The result is curiously direct, and implies that evolutionary theories which do not produce punctuation are likely to be seriously incomplete.

KEY WORDS Evolution, information theory, phase transition, punctuated equilibrium, renormalization, speciation.

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

Introduction

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 learning plateaus in cognitive systems (R Wallace and R Fullilove, 2001; R Wallace, 2000a, b) suggests, however, that even the most simplistic gene-centered view of evolution will give punctuation in a 'natural' manner, and that theories without such behavior are grossly incomplete.

The comparison with cognition is counterintuitive: Evolution is not a cognitive process. Cognition involves, at its foundation, a 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. Boyd and Richerson, 1995, 1998; Durham, 1991), do indeed constitute a kind of 'memory' of past interaction with the world, response to selection pressure largely involves the reproductive success of more or less random variation. Even in the case of human biology, culture tends to be fairly rigid, and selection pressure usually dominates dynamics (e.g. D Wallace and R Wallace, 2000).

This is very far indeed from cognition. There is, thus, no 'intelligent purpose' to 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.

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 behavioral factors, this is a distorting oversimplification, currently popular scientific paradigm involves examination of the 'genetic code,' and in the 1970,s '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 (R Wallace and RG 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 process.

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.

Recently Park, Amari and Fukumizu (2000) examined a computer model of an artificial neural network – an array of feedforward multilayer perceptrons trained using a gradient descent error backpropagation algorithm, a system which inevitably suffers recalcitrant learning plateaus. Park et al. state 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, and the annealing temperature is inversely proportional to the slope of the curve.

Elsewhere (R Wallace, 2000a, b; R Wallace and R Fullilove, 2001) we have shown how cognitive pattern recognition-and-response can be characterized by a 'dual information source,' permitting us, in a fashion recognizably similar to the Park and Rose papers, to apply 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 punctuation.

Some preliminary development is required.

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' $k = 1, 2, ...$ – 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

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

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

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

with associated conditional probabilities

$$P(X_n = \alpha_n | X_{n-1} = \alpha_{n-1}, ..., X_0 = \alpha_0).$$

(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 parametized and that are, with respect to that parameter, *piecewise memoryless*, i.e. probabilities do not change markedly within a 'piece,' but may do so 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, the system is, to first approximation, memoryless in this sense. In what follows we use the term 'ergodic,' to mean '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

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

(3)

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

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

(3)

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

$$H(X) = -\sum_j P_j \log(P_j)$$

$$H(Y) = -\sum_k P_k \log(P_j)$$

(4)

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

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

(5)

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

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

(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

$$H[\mathbf{X}] = \lim_{n\to\infty} H(X_n|X_0, X_1, ..., X_{n-1}) =$$

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

(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 $P[X_n = \alpha_n|X_0 = \alpha_0, ..., X_{n-1} = \alpha_{n-1})$ 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 IR 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 within a cognitive system 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. We attempt a formal treatment, from which it will appear that the learning process is – almost inevitably – highly punctuated by 'learning plateaus'. Application to non-cognitive evolutionary process will be direct.

Suppose we have a 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

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

Suppose that chain elicits a corresponding chain of responses from the system of interest, producing another path $b^n = (b_1, ..., b_n)$, 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

$$d(y, \hat{y}) = 1, y \neq \hat{y}$$

$$d(y, \hat{y}) = 0, y = \hat{y}.$$

(8)

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

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

(9)

Possibilities abound.

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

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

(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

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

(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:

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

(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

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

(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 $n \to \infty$ the smaller set approaches unit probability, and we have for those pairs the condition

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

(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.

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

Phase transition and coevolutionary condensation

The essential homology relating information theory to statistical mechanics and nonlinear dynamics is twofold (R Wallace and RG Wallace, 1998, 1999, 2001; Rojdestevnski and Cottam, 2000):

(1) A 'linguistic' equipartition of probable paths consistent with the Shannon-McMillan 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 R Wallace and RG 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.

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

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

(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

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

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

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

(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 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

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

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

(17)

where $\mathcal{D}$ 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 $R^{\mathcal{D}}$, 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; R Wallace and RG Wallace, 1998) gives

$$H = H_0 \kappa^{s\mathcal{D}}$$

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

(18)

where s is a positive constant. Some rearrangement produces, near K_C,

$$H \propto \frac{1}{\chi^{\mathcal{D}}}$$

(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 pattern recognition or other '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:

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 parametization 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.

R Wallace and RG 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.

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 an 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

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

(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 parametized 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 (R Wallace and RG Wallace, 1998, 1999).

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.

Discussion and conclusions

Before proceeding further we are obligated to raise the red flag of a standard methodological caution regarding the use of mathematical models – like ours – to address complex ecosystem phenomena: The Word is Not the Thing, or, as the noted mathematical ecologist EC Pielou put it (Pielou, 1977, p. 106),

> "...[M]athematical models are easy to devise; even though the assumptions on 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."

Precisely in the spirit of Pielou's warning, our mathematical modeling exercise raises the speculation that empirical study of the non-cognitive learning plateaus of evolutionary punctuation may permit identification of at least a local and temporary 'universality' in such transitions. After transition, of course, the system has changed markedly. We thus speculate that, at best, with respect to the parameter K, such systems are 'piecewise memoryless,' in the strict information theory sense that the associated information sources do not change probabilities (much) between punctuation events.

These events would seem to include both speciation and coevolution, taken as inverse phenomena in the sense of R Wallace and RG Wallace (1998), i.e. splitting vs. coagulation of information sources.

We conclude that, 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.

References

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

Atlan H and IR 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, Dowrick N, Fisher A, Newman M, 1986, *The theory of critical phenomena*, Clarenden Press, Oxford.

Cohen IR, 1992, "The cognitive principle challenges clonal selection," *Immunology Today*, **13**, 441-444.

Cohen IR, 2000, *Tending Adam's Garden: evolving the cognitive immune self*, Academic Press, New York.

Cohen IR, 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.

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.

Levin S, 1989, “Ecology in theory and application,” in *Applied Mathematical Ecology*, S Levin, T Hallam and L Gross (eds.), Springer-Verlag, New York.

Levin S, 1990, *Mathematics in Biology: The interface-challenges and opportunities*, Cornell University, New York.

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.

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 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 RG Wallace, 2001, *The New Information Theory: applications in biology, medicine and social science*, Submitted book manuscript.

Wallace R and R Fullilove, 2001, "Learning plateaus in generalized cognitive condensations: phase transition, path dependence and health disparities." Submitted.

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