Distinguishing Discretization and Discrete Dynamics, with Application to Machine Learning, Ecology, and Atomic Physics

Karl Gustafson*

Abstract. The distinction between the discretizing of a continuous dynamical system, and an analogous discrete dynamical system, is examined. A number of critical conceptual misunderstandings are identified, in historical context. Implications for the internal structures in machine learning, ecological dynamics, and atomic wave systems, are discussed.

§1. Introduction. In a recent paper [Gus00] I promised "to analyze from a historical perspective how this rather fundamental finding was previously missed." The fundamental finding referred to was a basic connection I discovered (a dozen years ago) between widely used recently developed machine learning algorithms and the recently developing theory of chaotic discrete dynamical systems. My discovery moreover implied some critical conceptual misunderstandings within both the machine learning community and the dynamical system community. The first purpose of the present paper is to keep my promise of [Gus00]. In doing this I will go beyond [Gus 00, Gus90, Gus97, Gus98a, Gus98b,Gus98c, GS99], referring to those papers for convenience. A second purpose is to go beyond my previous work by discussing here also certain implications for the internal structures of population dynamics and quantum wave system dynamics.

§2. Machine Learning. In 1988 as a result of a successful interdisciplinary proposal for an NSF Engineering Research Center for Optoelectronic Computing Systems at the University of Colorado and Colorado State University, I found myself responsible for the mathematics and algorithm development to accompany an optical neural network being constructed in hardware. The original Perceptron machine learning algorithm, which was linear, had been by then superseded by the important Backpropagation algorithm. Backpropagation overcame many learning limitations of the linear Perceptron. This was accomplished in Backpropagation by introducing nonlinear thresholds, typically implemented by sigmoids $f(x) = (1+e^{-\beta x})^{-1}$. I refer the reader to the bibliography and in particular to [RM86] for a good discussion of Backpropagation (also called other names such as the δ-rule, multilayer perceptron, etc.). For the purposes of this paper, we may describe Backprop-

*Department of Mathematics, University of Colorado, Boulder, Colorado 80309–0395, USA

agation learning as the building of a learning surface (sometimes called the learning landscape) in multidimensional space on the basis of a number of repetitive training examples (input–output pairs).

To fix this idea, I show in Figure 1 such a learning surface from [GG92]. In that (unpublished) paper, we modified Backpropagation to an algorithm we called Anglelearning Backpropagation, or Angleprop for short. You may view Figure 1 as depicting conceptually the same type of landscapes the standard Backpropagation algorithm generates as a result of a large number of training pairs repetitively fed to it. Because Backpropagation learns this surface by a repetitive steepest descent procedure, convergence to the valleys (which carry smaller least squares error than mountains or plateaus) is often very slow, especially if the previous training iteration put you on one of the plateaus. Our idea in Angleprop was to just learn the angles between weights, rather than the weights themselves. I include Angleprop here to show you typical Backpropagation error surfaces, and because [GG92] was never published elsewhere and contains an interesting original idea. See also [WUG94] for some recent nice pictures of such learning surfaces from the Japanese engineering community.

When we went to implement Backpropagation on the hardware optical neural network, I learned that the optical devices were not available to us to implement the thresholdings. Therefore we just took this optical data out to a digital computer to do all thresholding and then went back into the optics for the next training epoch. At that point I learned that one reason the sigmoid thresholding was so popular in the machine learning community was that it had the nice property that its derivative is conveniently expressed in terms of itself: $f'(x) = \beta f(x)(1-f(x))$. In particular, in the Backpropagation algorithm, this permits the weight changes $\Delta\omega_{ij}$ to be calculated in terms of currently known network values. Although the Backpropagation update formulas become somewhat complicated due to lots of neural network interconnectivity and feedback, one can see that they take the form (at an output node, for simplicity) $\Delta\omega_{ij} = \eta f'(\text{net})(t_\ell - o_\ell)o_j = \eta(t_\ell - o_\ell)\beta o_\ell(1 - o_\ell)o_j$ where t_ℓ is a learning target value, o_ℓ is the net output at the current kth iteration, o_j is the transmitting node value, $\eta$ is a preassigned learning parameter, net is a linear combination of weighted inputs being fed to the nodes, and $\beta$ is the so-called gain. If we lump factors in we may see that the digitally implemented Backpropagation weight updates are *each* of the form $x_{n+1} = \mu x_n(1 - x_n)$, which is the discrete iterated quadratic map of dynamical systems theory.

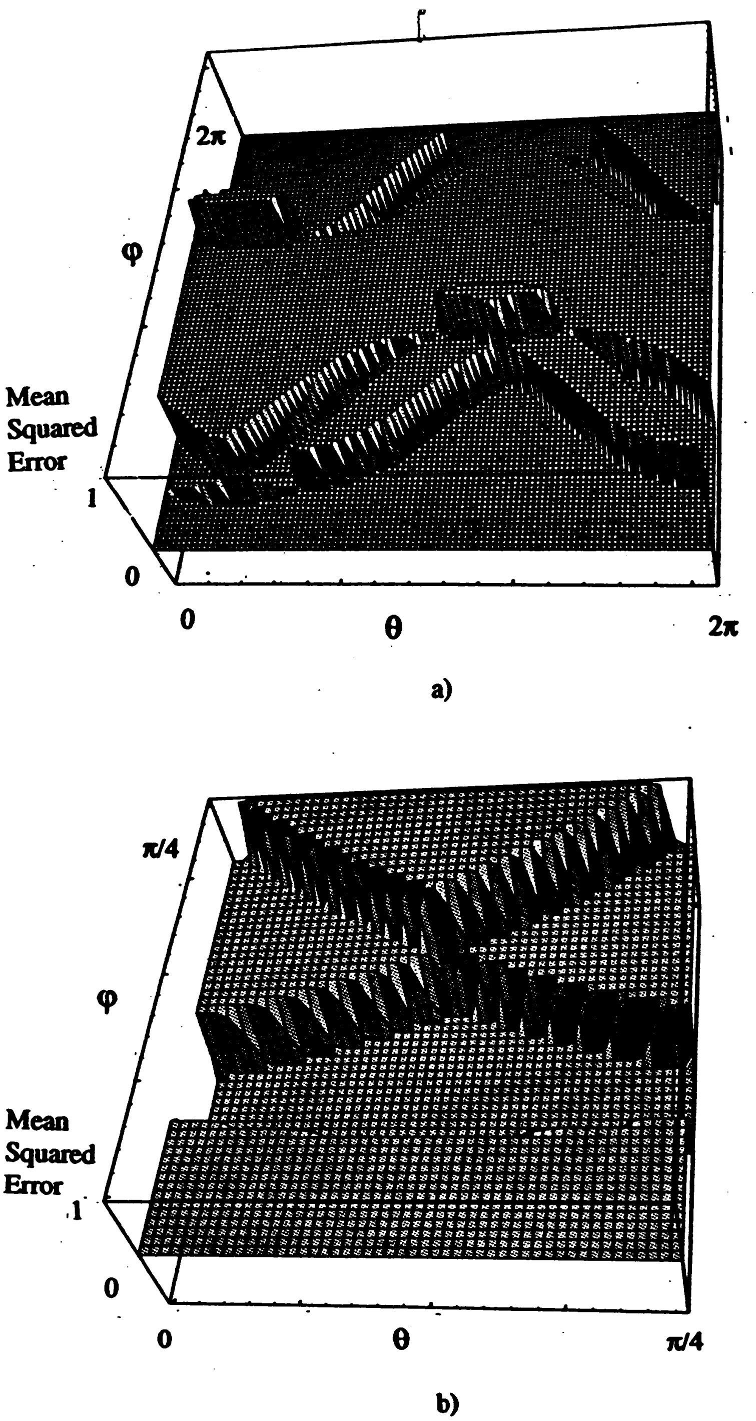

Figure 1: A plot of the error for the XOR problem versus the angles of the weights relative to the bias axis and one of the weight axes. a) A global view of the error surface. b) An enlarged view of solution in lower lefthand corner. Note constrained location of solution.

I learned in the period 1988–1990 that virtually everyone in the machine learning community was implementing Backpropagation, or similar thresholding multilayer perceptron algorithms, digitally. Yet they all were also viewing the thresholding as it appears in Figure 1, they spoke of it that way, they viewed it that way, in terms of a continuous steepest descent minimizing path down to a least squares cost function surface. I hinted at my discovery of local discrete quadratic map dynamics due to digital implementation in [Gus90] but it was only some years later after I had ascertained to the best of my ability that no one else shared my discovery, that I presented this finding rather completely in [GUS98c].

Recall that the quadratic map is well-known to map the interval $0 \leqq x \leqq 1$ to zero independent of initial guess x_0, when $\mu < 1$. For $1 < \mu < 3$ one converges to a nonzero stationary point. For $\mu > 3$ the quadratic map may exhibit periodic orbits, aperiodic orbits, or chaos. As the simulations of [Gus98c, Gus00] show, neural networks implementing Backpropagation do exhibit the same three qualitative behaviors. Although network connectivity, input and target values; initial weight choices, learning parameter, etc., all the complexity of the learning network architecture and data, may affect which of these basic behaviors you see, a main point is that this quadratic map behavior within a neural network is completely local, i.e., it applies to each individual node in the network. I illustrate this here in Figure 2. This Figure is the detail of the fourth column of [Gus98c, Figure 2]. As gain β increases through values $\beta = 3, 4, 5, 6, 7, 8$, one sees weight change behavior varying from rapid convergence to zero to intermittent oscillations to oscillatory nonconvergence.

§3. Historical Comment. Above I have pointed out how the machine learning community missed the fact that even the local node-specific learning dynamics was that (when implemented digitally) of the quadratic map of discrete dynamical systems theory. This fact was obscurred by the strong historical and cultural dogma influenced by the conceptual transition from the linear Perceptron to the nonlinear, smoothly thresholded multilayer perceptron, with its smooth slopes and ravines, upon which you can do gradient descent. That is not to say that others in the machine learning community had not happened onto notions or experience of chaos within neural network theory or practice. But (to my knowledge), all were confused by failing to distinguish the fundamental differences between continuous and discrete dynamical systems, or they were influenced too much by analogy, or there was confusion about onset of chaos being caused by high connectivity or large scale. Rather than repeat my discussions of these things already in [Gus00] and [Gus98c], let me just

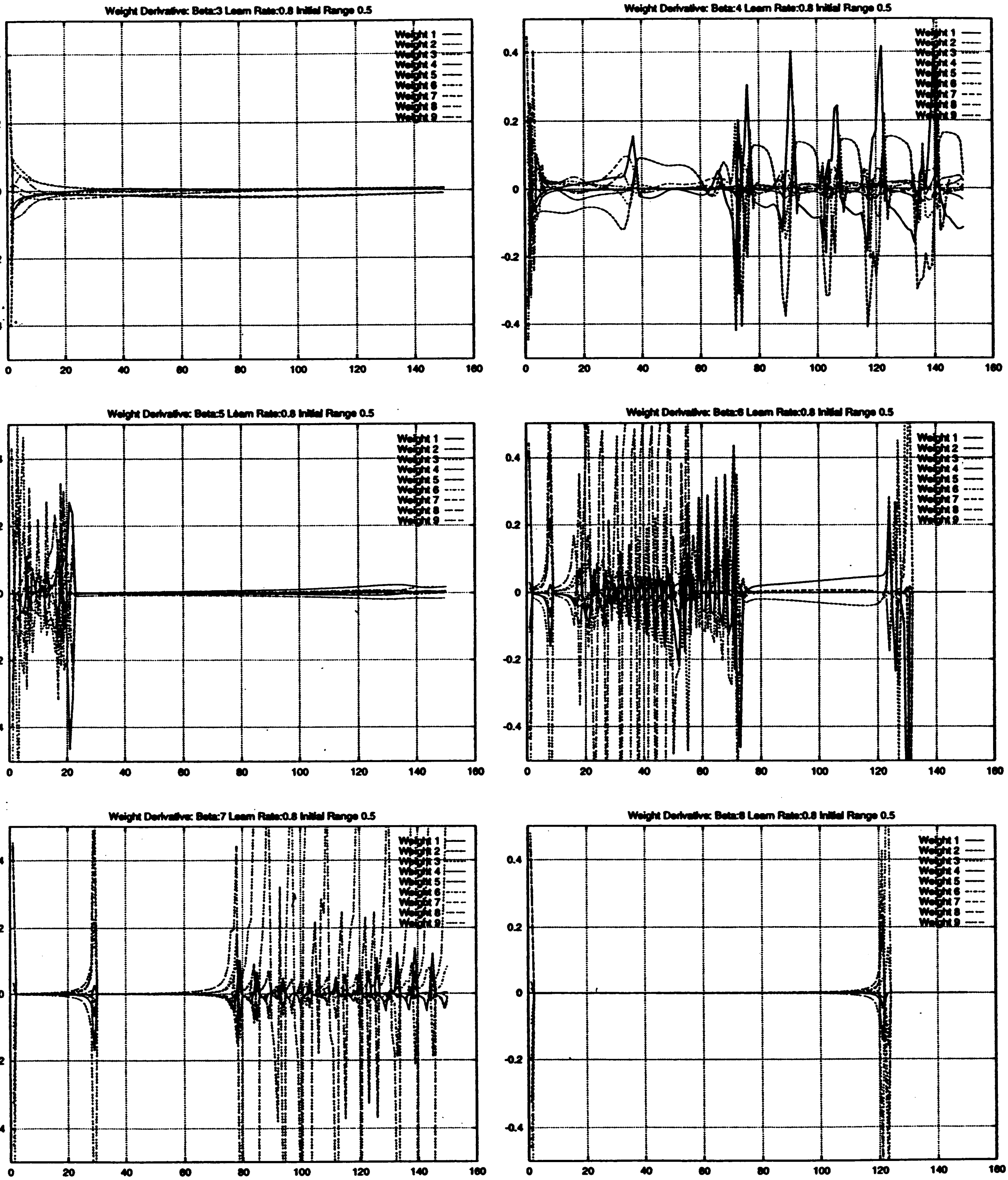

Figure 2: Discrete quadratic map irregularities in weight change dynamics. Learning parameter $\eta = 0.8$ and the initial weights chosen randomly in $(-0.5, 0.5)$, the same initial weights then used for each gain parameter $\beta = 3, 4, 5, 6, 7, 8$. Note the network internal nonlinear waves, which often appear to be coupled.

refer the reader to those papers, with the accompanying summarizing remark here that in [Gus00] and [Gus98c] you will find specific reference to, citation to, quotes from, the books of Devaney, Strogatz, Ott from discrete dynamical systems community, the books of Rumelhart–McClelland, Hertz, Levine, Wiegand–Gershenfeld, Kosko from the machine learning community, all excellent books and outstanding scientists. I also cite in [Gus00] and [Gus98c] a significant number of papers dealing with chaos in machine learning. Let me add a few more here which have become known to me and which may be helpful from the historical or conceptual perspectives to anyone who wishes to further pursue this work.

In [SF87] an early attempt is made to bridge the recently developed "connectionist" models (e.g., machine learning algorithms and neural network architectures) to actual brain function. To quote from the abstract: "Special emphasis is placed in our model on chaotic activity. We hypothesize that chaotic behavior serves as the essential ground state for the neural peceptual apparatus." However, the point of view of [SF87] is that the role of chaos in the brain is that of a source of background white noise as observed in EEG studies. Then they adopt the Grassberger–Procaccia et al. model of low-dimensional deterministic (continuous) dynamical system "strange attractor" chaos. Also they [SF87, p. 190] "we insist repeatedly that behaviorally relevant neural information is to be found in the average activity of ensembles (as manifested in the EEG) and not in the activity of single neurons." And [SF87, p. 171] "Connectionist models can certainly be modified to produce chaotic and oscillatory behavior, but current theorists have not included these behaviors in their models,.... Another [reason] is that engineers have traditionally viewed oscillatory and chaotic behavior as undesirable and something to be eliminated." [SF87] also contains an interesting adjoined Open Peer Commentary by many of the eminent researchers of the time, so the whole paper is interesting reading. My final point about it is the following. [SF87], representing the brain-science community, wants to get rid of the digital computer metaphor that the rival connectionist community employs. In doing so, [SF87] goes to the continuous dynamical system chaos models. Thus they have missed my finding that in the digital connectionist models, single neuron chaos was already present. The connectionist community missed this fact too.

Another interesting earlier paper is [Ha83]. The emphasis there is on "higher" neural processing accomplished by incorporating not only current information but also time-delayed connected information. However there is also some discussion of individual "trajectories for the netlet," which in turn leads to discussion of attractors and the low dimensional (continuous) dynamical system theory of chaos so popular in the 1980s. What interests me about this early paper is that it actually presents [Ha83, p. 786] "the parabola $F(x) = 4bx(1-x)$, $0 < b \leq 1$" as an example of a map which " will be chaotic for certain ranges of values of b and for certain 'seed' values of x." However again (in my opinion) a better understanding was obscured by not distinguishing and even more delineating the continuous chaos paradigm from the discrete chaos paradigm.

§4. **Ecological Dynamics.** In [Gus00] I state: "To May through his influential 1976 paper [Ma76] belongs much credit for the resurgence of recent interest in simple discrete maps, including the quadratic map. However, in our opinion, the use of the words 'analogous, corresponding' in the transition from population dynamics $P'(t) = aP(t) - bP^2(t)$ to map equations $y_{n+1} = ay_n - by_n^2$ is misleading. There is no way that you can discretize the former to obtain the latter in the sense of differential equations going to consistent difference equations, e.g. see [Gus99]. Succeeding treatises of discrete dynamical systems continue to fall into this (in our opinion) trap." I would like to elaborate this statement here, beyond what I said in [Gus00] and [Gus98c].

The ordinary differential equation initial value problem $\frac{dp}{dt} = ap - bp^2$, $p(t_0) = p_0$ occurs widely in science. In mathematics it falls within the category of Riccati equations. Its solution is easily found to be $p(t) = \frac{ap_0}{bp_0 + (a - bp_0)e^{-a(t-t_0)}} = \frac{1}{1+e^{-\beta t}}$ where in the last equality I took $a = b = \beta$, $p_0 = 1/2$, $t = 0$.

Discretization of a differential equation, e.g., to render a discrete version of it for numerical solution methods, usually carries the desired requirement of *consistency*: the true solution, when substituted into the discrete version, has truncation error which goes to zero as the discretization size becomes arbitrarily small. The obvious finite difference discretization of the initial value problem is the forward difference $\frac{p(t+\Delta t)-p(t)}{\Delta t} = \beta p(t)(1-p(t))$ which is just the difference quotient approximation to the first derivative. In this simple instance, consistency just reduces to the fact that the difference quotient rule of elementary calculus is a consistent one and the sigmoid function is differentiable.

However, now note that discrete quadratic map has virtually no connection to the continuous parameter differential equation from the point of view of discretizing the former to get the latter. Nor can you work back from the latter, as if it were a discretization, to get the former. If you try a few obvious discretization schemes, you will see no natural connections between the two equations.

Therefore I think a good word is *incommensurate*: the population dynamics equation and the quadratic map equation, although analogous, are incommensurate.

§5. **Historical Comment.** May [Ma76] popularized the mathematics of iterated maps and imagined them to be a possible explanation of the oscillations he had observed in population dynamics. Gleick [Gl87] gives a good account of this story. Then [Gl87, p. 80]: "May realized that the astonishing structures he had barely begun to explore had no intrinsic connection to biology." But [Gl87] does not develop this latter statement. Let me do so here. If you look at [Ma81] you will find a lowered emphasis on the potential use of iterated maps for such ecological population modelling. May clearly [Ma81, Chapter 2] now restricts the conceptual use of one-dimensional quadratic maps to models for single populations. The assumption has to be [Ma81, p. 6] that "generations are nonoverlapping and growth is a discrete process (first order difference equations)." Then for continuous growth, differential equations are proposed [Ma81, section 2.2]. However, to fit population data, time delays are also allowed into these equations. This is okay, these continuous population models have served ecological dynamics well, but it should be noted at this point that mathematicians' know well that differential-delay equations can model a wide variety of dynamics. When we get to the main section 2.3 of [Ma81], Discrete growth (difference equations), some actual ecological examples are claimed. However, the discussion quickly slides away from the biology and into the iterated map mathematical lore. Without a single specific ecological data set yet, we come to the statement [Ma81, p. 17]: "To see mathematical ecology informing theoretical physics is a pleasing inversion of the usual order of things." Without any ecology yet, this would appear to me to be inverted logic. When we do get to population data, one resorts to time delay or the continuous dynamics models. In Fig. 2.6 [Ma81, p. 21], Fig. 6 [Ma76], we find only one data point in the "chaos" region, and that not for a quadratic equation, rather for a fit by $y_{t+1} \cong 60y_t(1+y_t)^{-10}$.

To this day, confusion persists about the appropriate roles of discrete and continuous chaos. By this I mean, respectively: chaotic iterative dynamics produced by an iterated discrete dynamical system, and chaotic dynamics produced by the continuous time evolution of a nonlinear system of ordinary differential equations. For example, beyond the popular quadratic map, another popular discrete dynamical system is the quadratic Henon–Heiles map, see [Ta89]. Perhaps the most popular example of a continuous chaotic dynamical system is the famous quadratic Lorenz system which was an extreme simplification of the partial differential equations of meteorology. See the discussions in [Gl87], [Ta89], and elsewhere. Both the Lorenz continuous dynamical system and the Henon discrete dynamical system possess strange attractors. What I am asserting above in Section 4 and throughout this paper is that the algebraic *similarities in form* between the discrete dynamical systems and the continuous dynamical systems equations can be critically misleading, and one cannot assert a priori any 'corresponding' behavior in their dynamics without a further analysis which would (unlikely in most instances) prove such.

How this confusion happens? I would like to identify two key factors, although there are certainly others. Let me briefly present these two factors: analogy and vocabulary. The point I would like to make here is that Analogy although a powerful mental function, should be regarded as a subjective reasoning. Always it should then be placed into an objective analysis. Subjective reasoning relies on experience-based intuition and can be very powerful but can also lead to serious errors unless checked by deductive systematic testing. Right brain and left brain cannot trust each other and their coexistence may be viewed as a valuable system of checks and balances. The second factor I wish to identify is vocabulary. For example, one finds the term *logistic* used in the literature for both the continuous population equation and discrete quadratic map. Also the same function is called both the logistic function and the *sigmoid* function. Such vocabulary failures of precision can translate into conceptual confusions.

§6. Atomic Physics. Next I would like to turn to a third field of scientific endeavor, Atomic Physics, where I believe caution should be exercised to avoid critical misunderstandings due to insufficient care in distinguishing continuous and discrete dynamical systems. I haven't discussed this situation previously. The problems arise in attempts to model quantum dynamics by classical Hamiltonians so that one can actually calculate approximate bound states and their energies as periodic orbitals as if they were in the old Bohr "solar system" quantum mechanics. Within the mathematics of quantum mechanics these theories and techniques go under the names Born–Oppenheimer approximation, WKB method, Bohr–Sommerfeld quantization. In our conference volume [GR81] you will find several articles on this topic. I also recommend the conference volume [Hi83] for a similar perspective. Also see [Ta89], to which I will refer below. I will also refer to [BR97], with all due respect and apologies to my colleague William P. Reinhardt with whom I put out the book [GR81] about twenty years ago.

Without getting into the details, given a quantum mechanical Hamiltonian H viewed semiclassically, the density of states per unit energy $\rho(E)$ is given by $\rho(E) = Tr[\delta(E-H)] = \sum_n \delta(E - E_n)$ where δ is the delta function and the E_n are the distinct eigenvalues of the Hamiltonian. By Fourier Transform $\delta(E - H) = \frac{1}{2\pi\hbar}\int_{-\infty}^{\infty} e^{iEt/\hbar}e^{-iHt/\hbar}dt$, the density becomes $\rho(E) = \int_{-\infty}^{\infty} e^{iEt/\hbar}Tr(e^{-iHt/\hbar})dt = \frac{1}{2\pi\hbar}\int_{-\infty}^{\infty} e^{iHt/\hbar}\int\langle q, e^{-iHt/\hbar}q\rangle dqdt$ where the last integral represents integration of the expectation values over all configuration states q at time t in the evolution. The semiclassical approximation then is achieved by approximating this expectation value integrand by $\langle q_1, e^{-iHt/\hbar}q_2\rangle \sim e^{i\phi(q_1,q_2)/\hbar}$ where $\phi(q_1, q_2)$ is the action integral along the classical trajectory connecting q_1 and q_2 in time interval $[0, t]$. In other words, the quantum averaging is replaced by a single frequency oscillation. This leads via the now classical Hamiltonian dynamics to a requirement that the values of q which actually contribute in this stationary phase sense to the integral must lie on a periodic trajectory. See [BR97] and [Ta89] for more details.

It is well-known that classical Hamiltonian systems may exhibit chaos. For example [Ta89] the Henon–Heiles Hamiltonian $H = \frac{1}{2}(p_x^2 + p_y^2 + x^2 + y^2) + x^2y - \frac{1}{3}y^3$ exhibits chaotic Poincaré cross sections. There are many other examples and it can be said that the thrust of quantum chaos studies are motivated by these classical Hamiltonian chaotic dynamical systems. [BR97, p. 83] are careful to point out that true quantum systems do not typically display chaos in the sense of exponential sensitivity to initial state. They are careful to distinguish (I) quantized chaos, (II) semi-quantum chaos, (III) quantum chaos. It is really (I) which dominates most current modelling. Thus one may consider not only periodic orbits from the stationary phase approximation I described above, but also aperiodic orbits and more irregular orbits in the same setting. [Ta89, p. 229] is also very careful to point out that the steps in the semiclassical approximation from the Schrödinger partial differential equation to a classical Hamilton–Jacobi equation, "is very subtle." As Planck's constant $\hbar$ is taken to zero, one is neglecting the $i\hbar\nabla^2 S/2m$ term in the limit. Moreover $\hbar \to 0$ corresponds to "ever more rapid oscillations in the wave function." Here $S = S(q, t)$ represents a single phase evolution in the wave function $\psi(q, t) = e^{iS/\hbar}$ resulting from a separation of variables. They go on to make clear that "It is completely wrong to think that someone can somehow write quantum mechanical quantities as classical quantities plus an expansion of corrections in power of $\hbar$."

§7. **Historical Comment.** My concern about the models presented above in Section 6 is the mix of discrete and continuous dynamical systems employed in this current research in atomic physics. This mix is found throughout [BR97] and [Ta89] and elsewhere. It is too easy to imagine that the discrete model's wave system dynamics somehow really depict what is happening in the continuous model's wave system, even when staying completely within the frame of classical Hamiltonian systems. Although [BR97] and [Ta89] are careful to put in qualifying provisos, there still is the inference that one really is modelling quantum chaos, i.e. stated more carefully, quantum evolutions which depend on an underlying chaos. The fact that underlying discrete nonlinear chaotic phase-space dynamics can be treated in terms of state space (e.g., probability distributions) functions over the phase space is demonstrably true in statistical mechanics, see e.g. [Gus97] or [GR81] and citations therein. But it only holds true in that situation for rather special 'Kolmogorov' dynamical systems and on compact phase spaces. Thus the "manifestations of chaos in atomic and molecular physics" [BR97] must be taken as experimentally or conceptually inspired rather than theoretically proven. Finally, is there really any need for such chaos models in atomic physics? The physical evidence presented in [BR97] and [Ta89] is meager at best. And a manifestation of chaos is not proof of true underlying physical chaos.

§8. **Conclusions.** The three 'stories' I have given here illustrate the force of fashion within the scientific enterprise. The machine learning community followed a fashion of smooth learning surfaces and did not see and did not want the chaos which I identified as inadvertently introduced through digital, i.e., discrete, implementation of nonlinear thresholdings. The ecological dynamics community became entranced with a fashion of chaos even though chaos was not in their population dynamics. The atomic physics community created a fashion of quantum chaos which no-one has yet seen.

References

[BR97] Blümel, R. and Reinhardt, W. P., *Chaos in Atomic Physics*, Cambridge University Press, Cambridge, 1997.

[Gl87] Gleick, J., *Chaos: Making a New Science*, Viking, New York, 1987.

[Gus00] Gustafson, K., Chaos in discrete learning systems, *Chaos, Solitons and Fractals* 11 (2000), 321–327.

[Gus90] ______________, Reversibility in neural processing systems, in: *Statistical Mechanics of Neural Networks*, (L. Garido, ed.), Lecture Notes in Physics 368, Springer, Berlin, 1990, 269–285.

[Gus97] ______________, *Lectures on Computational Fluid Dynamics, Mathematical Physics, and Linear Algebra*, World Scientific, Singapore, 1997.

[Gus98a] ______________. Internal dynamics of Backpropagation learning, in: *Proc. Ninth Australian Conference on Neural Networks*, (T. Downs, M. Frean, M. Gallagher, eds.), University of Queensland, Brisbane, 1998, 179–182.

[Gus98b] ______________, Ergodic learning algorithms, in: *Unconventional Models of Computation*, (C. Calude, J. Casti, M. Dinneen, eds.), Springer, Singapore, 1998, 228–242.

[Gus98c] ______________, Internal sigmoid dynamics in feedforward neural networks, *Connection Science* 10 (1998), 43–73.

[Gus99] , ______________, *Partial Differential Equations and Hilbert Space Methods*, Dover Publications, New York, 1999.

[GG92] , Gustafson, K. and Goggin, S., Anglelearning Backpropagation, (1992, unpublished).

[GR81] Gustafson, K. and Reinhardt, W. P., *Quantum Mechanics in Mathematics, Chemistry, and Physics*, Plenum Press, New York, 1981.

[GS99] Gustafson, K. and Sartoris, G., Assigning initial weights in feedforard neural networks, in: *Proc. 8th IFAC Symposium on Large Scale Systems*, (N. Koussoulas, P. Groumpos, eds.), Patras, Greece, July 1998, Pergamon Press, N.Y. 1999, 1108–1113.

[Ha83] Harth, E., Order and chaos in neural systems: an approach to the dynamics of higher brain functions, *IEEE Trans. on Systems, Man, and Cybernetics*, Sept./Oct. 1983, 782–789.

[Hi83] Hinze, J. (ed.), *Energy Storage and Redistribution in Molecules*, Plenum Press, New York, 1983.

[Ma76] May, R., Simple mathematical models with very complicated dynamics, *Nature* 261 (1976), 459–467.

[Ma81] ______________ (ed.), *Theoretical Ecology: Principles and Applications*, 2nd Ed., Blackwell Scientific Publications, Oxford, 1981.

[RM86] Rumelhart, D. and McClelland, J. et al., *Parallel Distributed Processing*, Vols. I, II, MIT Press, Cambridge, MA, 1986.

[SF87] Skarda, C. and Freeman, W., How brains make chaos in order to make sense of the world, *Behavioral and Brain Science* 10 (1987), 161–195.

[Ta89] Tabor, M., *Chaos and Integrability in Nonlinear Dynamics*, Wiley, New York, 1989.

[WUG94] Watanabe, T., Uchikawa. Y., and Gouhara, K., Experimental studies of memory surfaces and learning surfaces in recurrent neural networks, *Systems and Computers in Japan* 25, No. 8 (1994), 27–39, (Denshi Joho Gakkai Ronbunshi J76-D-II, No. 5, 1993, 1055–1065.