Complex Systems **9** (1995) 349–374

Dynamic Properties of Neural Learning in the Information-theoretic Plane

Perambur S. Neelakanta
Salahalddin Abusalah
Raghavan Sudhakar
Dolores De Groff
Valentine Aalo
Department of Electrical Engineering,
Florida Atlantic University,
Boca Raton, FL 33431, USA

Joseph C. Park
Atlantic Undersea Test and Evaluation Center (AUTEC),
West Palm Beach, FL 33401, USA

Abstract. Learning in reference to the real neural complex depicts progressive modifications occurring at the synaptic levels of the interconnected neurons. The presence of intraneural disturbances (inherently present) or any extraneural noise in the input data or in the teacher values may affect such synaptic modifications as specified by the set of weighting vectors of the interconnections. Translated to artificial neurons, the noise considerations refer to inducing an offset in the convergence performance of the network in striving to reach the goal or objective value via the supervised learning procedure implemented. The dynamic response of a learning network when the target itself changes with time can be studied in the information-theoretic plane and the relevant nonlinear (stochastic) dynamics of the learning process can be specified by the Fokker–Planck equation, in terms of a conditional entropy– (or mutual information–) based error measure elucidated from the probabilities associated with the input and teacher (target) values. In this paper, the logistic growth (evolutionary aspects) and certain attractor features of the learning process are described and discussed in reference to neural manifolds using the mathematical foundations of statistical dynamics. Computer simulation studies on a test multilayer perceptron are presented, and the asymptotic behavior of accuracy and speed of learning vis-à-vis the convergence aspects of the test error measure(s) is elucidated.

1. Introduction

In relation to learning processes in the neural complex, it is well known that synaptic modifications (specified by a vector array of adjustable weighting parameters w_l) can be influenced by the inevitable presence of intraneural disturbances, which will affect the network's convergence toward equilibrium. Further, in the event that the input data or teacher values are themselves stochastic, the corresponding extraneural influence may also augment the entropy of the system (real or artificial), facilitating the eventual veering of the network's output from the equilibrium value/stable state. Relevant neurodynamic considerations governing the variable w_l in artificial neural networks (NNs) have been addressed in [1] in terms of a stochastic differential equation (of the Langevin or Fokker–Planck type). Also, the dynamic states of the architectures, such as the Hofield network subjected to white-noise (random) inputs, have been analyzed via Ito-type stochastic differential equations applied to the so-called "diffusion machine" [2].

In the present work, an alternative approach is presented to describe the learning dynamics of an artificial NN in the presence of destabilizing factors caused by intra- or extraneural influences. The stochastic variable considered to model the relevant nonlinear neural dynamics refers to an error-measure parameter evaluated in the information-theoretic plane. Although a limited extent of neural dynamics considerations have been addressed in the information-theoretic plane pertinent to biological neurons [3, 4], equitable study or considerations vis-à-vis artificial NN are rather sparse. [5] describes a basic neural manifold being embedded as a submanifold in the manifold of a general nonneural information-processing system, and have developed an "information geometry" method to study the information-theoretic approach to learning dynamics and pattern classification problems. Further, the dynamics of an ensemble of learning processes in a changing environment (which feeds the training inputs to the network) has been described in [6] via a continuous-time master equation.

In the present study, the approach is concerned with the logistic growth considerations pertinent to the network's learning process in the information-theoretic plane. Relevant to this proposed method, a cross-entropy– (or mutual information–) based distance measure (ε) is specified as a stochastic variable, the asymptotic behavior of which (with respect to time) is studied as a discourse of the learning process. It is given by the following relation:

$$\varepsilon = H_\varepsilon(p_l, q_l) = K \sum_{t=1}^{N} q_l \phi(p_l/q_l) \tag{1a}$$

or

$$\varepsilon = H_\varepsilon(q_l, p_l) = K \sum_{t=1}^{N} p_l \phi(q_l/p_l) \tag{1b}$$

where ϕ is a twice-differentiable convex function for which $\phi(1) = 0$ and K is a constant factor. This error measure is adopted to train a NN (depicted in

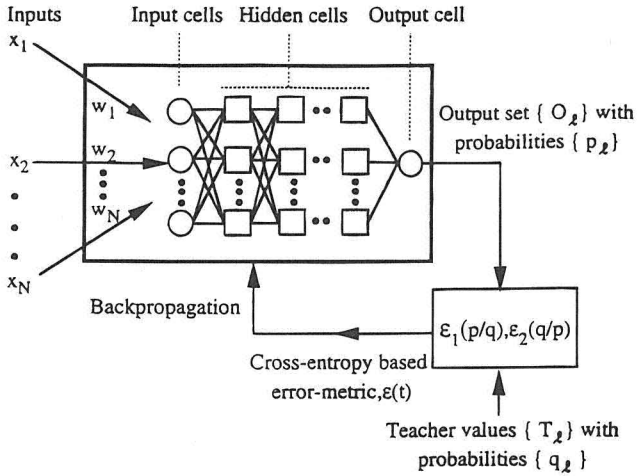

Figure 1a: An artificial neural network trained via a cross-entropy–based error metric in the information-theoretic plane using backpropagation mode.

Figure 1(a)) via a gradient-descent algorithm in the backpropagation mode [7]. The output of the network (Figure 1a), namely, O_l at the lth cell is assumed to have probabilities p_l, and q_l refers to a target set of probabilities, with $l = (1,\ 2,\ 3, \ldots,\ N)$ enumerating the number of the cells and thereby offering an index for the output units. The error function given by equation (1) is known as the Csiszár error measure [8] defined in the information-theoretic plane; and, when $\phi(y) = y\ \log(y)$ (with $y = p_l/q_l$ or q_l/p_l as appropriate), this measure is better known as the Kullback-Leibler measure [9].

The entropy attribution to the activities of the neural complex, and to the real neurons specifically, has been justified in [10] on the considerations of the principle of conservation of total "neural energy," its distribution, and an associated entropy. They have offered an operational definition of the macrostate of a neural system (in the same sense as in physical theromodynamic principles) and have associated it with the Shannon's concepts of information [11]. Disturbances in the real neural system caused by environment have been perceived in [10] as forces enhancing the associated entropy (or uncertainty) and correspondingly reducing the information content that would otherwise enable the physiological self-regulation.

These existing bases on real neural information processing offer a direction to extend the entropy- (or information theory–) based concepts to optimization algorithms used in artificial NNs.

The error measure indicated in equation (1) is a time-dependent stochastic variable specified over the epochs of iterations performed toward convergence and mediated through feedback strategies (such as the backpropaga-

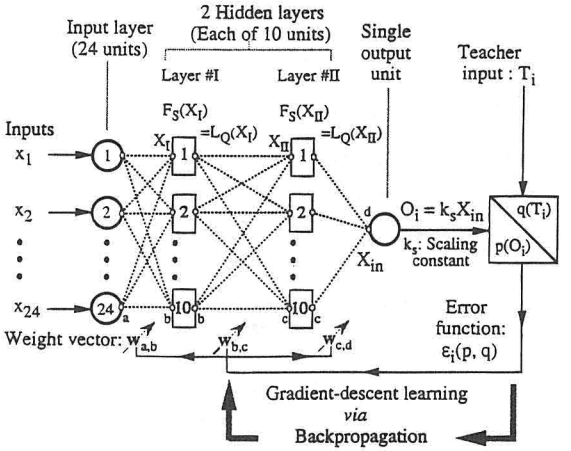

Figure 1b: Test neural network: A multilayered perceptron simulated for learning strategies in the information-theoretic plane.

tion algorithm) in the network. Due to the presence of any intra- and/or extracellular disturbances, the associated information flow in the neural system would, however, degrade with time; and the proliferation of information across the network may even become obsolete or nonpragmatic due to the asynchronous (random) synaptic delays between the internal state variable (being adjusted toward learning) and the adjusting influence (information) imparted (via the control loop) to the network by the error measure. That is, an aging of neural information (or degenerative negentropy) may occur that would lead to a devalued (or a value-weighted) knowledge with reduced utility (or pragmatics) being available to the converging efforts of the network striving toward the objective function. The degradation so perceived in the neural information plane depends on the extent of asynchronous delays encountered when the control loop (error) information arrives at the controlling section. That is, the asynchronously delayed error measure fed back will have no pragmatic value inasmuch as its asynchronous characteristics will not reflect the true (natural) output state (because the global state of the neural complex would have changed considerably by then) [12].

2. Stochastic neural dynamics

The trajectory of the time-dependent neural process pertinent to the learning endeavor represents the evolution of ε with time in a stochastic nonlinear dynamic system. Hence, in reference the variable $\varepsilon(t)$ describing the neural stochastic dynamics, a differential time evolution relation can be prescribed for $\varepsilon(t)$ as follows:

$$d\varepsilon(t)/dt = F_1[\varepsilon(t)]. \tag{2}$$

Or it can be specified by a recursive discrete-time process as follows:

$$\varepsilon(t+1) = F_2[\varepsilon(t)] \tag{3}$$

where F_1 (or F_2) is a differentiable function. It can be noted that ε, in general, is an N-dimensional vector affiliated with the phase space containing the time evolution of the underlying neural process spanned by the N-dimensional state vector of a dynamic system. A portrait of the corresponding time evolution of this system is therefore constituted by a set of trajectories in the N-dimensional phase-space. When the system reaches a state of permanent regime where the trajectories stay bounded, the corresponding invariant subset is termed an "attractor" specifying a state of stochastic equilibrium.

Relevant to an error measure ε adopted in network training (to assist the neural complex to learn from the environmental inputs), the neural dynamics can be described by a stochastic differential equation (of the general types given by equation (2)). Both the conventional types of error measure (such as the quadratic error measure) as well as the error measure that can be evaluated in the information-theoretic plane (on a cross-entropy basis as given by equation (1)) can be regarded to follow the paradigm of stochastic dynamics along the temporal passage of iterative epochs facilitated via feedback methods (until the error value is minimized). To assess the approach characteristics of the error parameter $\varepsilon(t)$ toward an equilibrium value (attractor) over a period of time (i.e., over the iterations of learning epochs) in order to develop an explicit dynamic model, the following valid assumptions can be made.

1. The parameters that decide the stochastic aspects of $\varepsilon(t)$ are confined within the basin of attraction.
2. The initial conditions of the stochastic process $(\varepsilon_0,\ t_0)$ involved should be specified appropriately.
3. The process is likely to be attracted to a stationary stochastic process whose probability density function (pdf) can be uniquely determined by the parameters of the original system variable, namely, ε.
4. In view of the preceding assumption, in the terminal attractor regime the pdf of $\varepsilon(t)$ does not vary as $t \to \infty$. That is, $\varepsilon_\infty(t) = \varepsilon(t \to \infty)$ is a stationary process.
5. As a first-order approximation, the stochasticity of the dynamics of $\varepsilon(t)$ is influenced only at fixed times corresponding to each onset of iterative epochs facilitated by the feedback.
6. The epochal iteration times are much larger than the periods of any fluctuations associated with $\varepsilon(t)$.
7. At the terminal stage, convergence of $\varepsilon(t)$ toward an equilibrium value of ε_∞ is ensured only if the network's optimization efforts are constituted favorably by reinforcement error information.

8. On the contrary, in the presence of overwhelming degenerating (or annihilating) error information, the dynamics of $\varepsilon(t)$ will be divergent.

3. Stochastic dynamics of the error measure (ε): General considerations

The dynamics of $\varepsilon(t)$, in general, can be equated to a random walk process by virtue of the aforementioned assumptions and in view of the following considerations.

Specifically, the endeavor of the network toward convergence when conceived in the information-theoretic plane refers to an adaptation process wherein the progressive acquisition of information leads to minimization of disorganization or eradication of uncertainty (entropy) of the network output vis-à-vis the teacher function.

When the network has learned (or adapted itself to the environmental inputs) to the fullest extent, it does not need any more information inasmuch as it retains no further uncertainties about the output against the teacher values; that is, a fully trained network may not perceive any further information since the output is maximally certain against the teacher value with which it is compared.

A heuristic time-dependent model of the goal-oriented, converging aspect of the neural complex versus time expressed in terms of $\varepsilon(t)$ as described previously, can be depicted qualitatively in terms of the variance of the teacher function σ_T^2 and that of the network output σ_O^2. That is, the evolution of error entropy $\varepsilon(t)$ can be specified by an envelope profile given by [13]

$$\varepsilon(t) = (k/2)\log(1 + \sigma_O^2/m(t)\sigma_T^2) \tag{4}$$

where $m(t)$ is the number of iterations over time (t), which can be modeled as a simple case by depicting $m = \alpha t$, where α is the number of iterations per unit time; and k is a constant as determined by the base of the logarithm. Hence,

$$\varepsilon = (k/2)\log(1 + \sigma_O^2/\alpha\sigma_T^2 t). \tag{5}$$

In the initial time frame, that is, at the commencement of network learning, the error information pertinent to the output (in reference to the teacher value) could be significantly different, and thus the network has a large potential to receive information in tending toward the objective function. Therefore, the initial error information $\varepsilon(t_0 \to 0) = \varepsilon_0$ can be designated as the potential error information.

4. Random walk paradigm of $\varepsilon(t)$ dynamics

As discussed earlier, the error measure $\varepsilon(t)$ when specified in the information-theoretic plane can be written in the form of equation (1). More generally, it

can be constructed by combining linearly two weighted parts of the Csiszár metric, given by equations (1a) and (1b). That is:

$$\begin{aligned} \varepsilon &= K_1 \sum_i q_i \phi(p_i/q_i) + K_2 \sum_i p_i \phi(q_i/p_i) \\ &= \sum_i (\varepsilon_{1i} + \varepsilon_{2i}), \quad i = 1,\ 2,\ \ldots,\ m,\ \ldots,\ n \quad (n \to \infty,\ t \to \infty) \end{aligned} \tag{6}$$

where K_1 and K_2 are weighting factors. If $K_1 = K_2$, equation (6) can be considered as symmetrized and balanced.

Designating each iterative epoch as of duration $\Delta\tau$, the total time involved in reaching the terminal state of dynamics (with $n \to \infty$) is taken as an integral multiple $n\Delta\tau = T_\infty$ (say). Suppose the potential energy associated with the system (which is being minimized) is taken as E. For each epoch of iteration, there is a corresponding energy configuration, the ensemble of which can be represented by a canonical Gibb's distribution given by [12]

$$\mathcal{P}_i(\varepsilon_{1i}) = C_1 \exp(-\Delta E_{1i}/E_R) \tag{7a}$$

or by

$$\mathcal{P}_i(\varepsilon_{2i}) = C_2 \exp(-\Delta E_{2i}/E_R) \tag{7b}$$

where E_R is a reference energy level, and the normalization constants (partition functions) C_1 and C_2 are determined from the requirement $\sum_i \mathcal{P}_i(\varepsilon_{1i}) = \sum_i \mathcal{P}_i(\varepsilon_{2i}) = 1$. Hence, $C_1 = C_2 = 1/M(T_\infty)$, where M is the total number of energy levels configured over the time T_∞. The corresponding configurational entropy associated with ε_{1i} or ε_{2i} is

$$\begin{aligned} S_T(\varepsilon) &= -K \sum \mathcal{P}_i \log \mathcal{P}_i \\ &= -K \sum [1/M(T_\infty)] \log[1/M(T_\infty)] \\ &= K \log[M(T_\infty)] \end{aligned} \tag{8}$$

where, again, K is a constant specified by the base of the logarithm. The number of ways (or realizations) the ensemble $M(T_\infty)$ can be divided into two groups of $\boldsymbol{m}_1$ and $\boldsymbol{m}_2$ (corresponding to ε_1 and ε_2, respectively, without regard to order) is given by the binomial coefficient, namely,

$$\begin{pmatrix} M(T_\infty) \\ \boldsymbol{m}_1 \end{pmatrix} = M(T_\infty)!/\boldsymbol{m}_1!\boldsymbol{m}_2! \tag{9}$$

where $M(T_\infty) = (\boldsymbol{m}_1 + \boldsymbol{m}_2)$.

Inasmuch as the statistics concerning state transitions associated with $\varepsilon(t)$ are governed by Gibb's distribution (equation 7), the discourse of ε with time represents a time-homogeneous Markov chain. Further, the transitional epochal state of $\varepsilon(t)$ is determined by the configurational energy level ΔE_{1i} and/or ΔE_{2i} and can be modeled by the concept of one-dimensional random walk. Starting at $t = 0$ and taking steps of length $\Delta\tau$ each, let $\Delta\varepsilon^+$ and

$\Delta\varepsilon^-$ be the reinforcing and annihilating information, respectively, imparted by the error feedback via the control loop (with the probability of each being equal). The random walk model enables the computation of the probability of achieving a specific information state at $t = m\Delta\tau$ after m iterative steps. That is, by considering $\Delta\varepsilon^+$ as the reinforcement information and $\Delta\varepsilon^-$ as the degenerating counterpart, the corresponding (proportionate) contributions occurring randomly (with equal probabilities) refer to the evolution process depicting the excursion of $\varepsilon(t)$ about the equilibrium value (ε_∞) versus the iteration of epochs (Figure 2) performed.

The transitional probability associated with the excursion of $\varepsilon(t)$ by $\Delta\varepsilon^\pm$ in the aforementioned one-dimensional random walk process commencing at an initial state depicted by $\varepsilon_0(t \to t_0) = (\varepsilon_0, t_0)$ is given by

$$\begin{aligned}
&\mathcal{Z}[(\varepsilon + \Delta\varepsilon^\pm, t + m\Delta\tau) \mid (\varepsilon, t)] \\
&\quad = \text{Transitional probability of } \varepsilon(t) \text{ assuming the values} \\
&\qquad \left\{ \begin{array}{c} \varepsilon + \Delta\varepsilon^+ \\ \text{or} \\ \varepsilon + \Delta\varepsilon^- \end{array} \right\} \text{ at } mth \text{ epoch or time} \\
&\quad = 1/[1 + \exp[(\Delta E_{1m} - \Delta E_{2m})]/E_R] \qquad (10)
\end{aligned}$$

In this random walk process, the current value of $\varepsilon(t)$ is determined by the potential level ΔE, and therefore the corresponding probabilities of the state of $\varepsilon(t)$ as given by equation (10) also depend on the current value of $\varepsilon(t)$. This (energy-dictated) random walk process (as opposed to the free diffusion process) is a force field–dependent diffusion process and therefore corresponds to the Ornstein-Uhlenbeck process [14].

For a given m, the possible values of ε (especially for large values of t) would differ from each other by multiples of $2\Delta\varepsilon^\pm$ since changing ε (by $\Delta\varepsilon^+$ or $\Delta\varepsilon^-$) at any single step changes the final value of $\varepsilon(t)$ by that amount. Or a probability $\mathcal{W}(\varepsilon;\ m)$ can be defined such that $2\Delta\varepsilon^\pm\mathcal{W}(\varepsilon;\ m)$ refers to the probability of reaching ε after m excursions. That is, $2\Delta\epsilon^\pm\mathcal{W}(\varepsilon;\ m)$ is the probability reached in the interval $(t = m\Delta\tau)] \leq t \leq [(t = m\Delta\tau + \Delta\tau)]$ after m steps. The relation between $\mathcal{W}(\varepsilon;\ m)$ and $M(T_m)$ is therefore $2\Delta\varepsilon^\pm\mathcal{W}(\varepsilon) = M(T_m)(1/2)^m$, using $\mathcal{W}(\varepsilon)$ for $\mathcal{W}(\epsilon;\ m)$ and $T_m = (m\Delta\tau)$ for convenience.

It may be noted that any particular set $\{\Delta\varepsilon_i^+\}$ or $\{\Delta\varepsilon_i^-\}$ (regarded now as defining a particular ensemble sequence of increments or decrements in ε with respect to each step in a random walk) has probability $(1/2)^m$ and there are $M(T_m)$ such sets that lead to the mth epoch at $t = T_m$. Inasmuch as $\mathcal{W}(\varepsilon)$ and $M(T_M)$ differ only by a coefficient (independent of T_m), the corresponding configurational entropy can be written as

$$S_{T_m}(\varepsilon) = K \log \mathcal{W}(\varepsilon) \qquad (11)$$

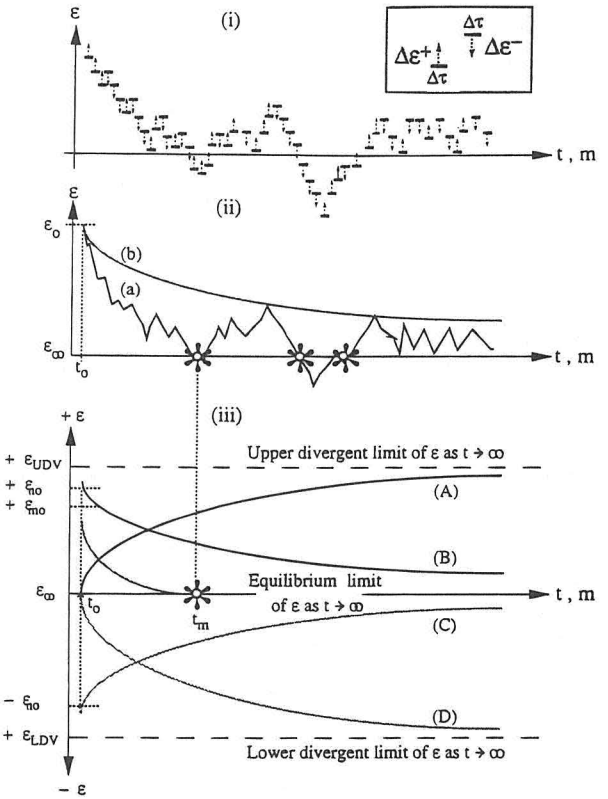

Figure 2: Convergent and divergent modes of ε as a function of time (t) or number of epochal iterations (m). (i) Random walk representation of $\Delta\varepsilon^+$ and $\Delta\varepsilon^-$ versus t or m. (ii) Temporal trajectory of ε. (ii)(a) Actual trajectory crossing the equilibrium value of ε (ε_∞) at specific attractors ($*$). (ii)(b) Envelope of the trajectory ε showing the asymptotic trend of ε at its terminal dynamics as t or $m \rightarrow \infty$. (iii) Divergent and convergent profiles of ε versus time. (iii)(A) & (D): Diverging envelopes directed from positive and negative sides, respectively. (iii)(B) & (C): Converging envelopes directed from positive and negative sides, respectively.

Further, for $m \gg 1$ and $n \to \infty$ and $(T_m/n\Delta\tau) = (\varepsilon/n\Delta\varepsilon^{\pm}) \ll 1$, the following approximation is valid:

$$\log[(m \pm \varepsilon/n\Delta\varepsilon^{\pm})/2] \approx (\varepsilon/n\Delta\varepsilon^{\pm}) - [(\varepsilon^2/2n^2(\Delta\varepsilon^{\pm})^2] + \log(m/2) \quad (12)$$

Using the preceding approximation and applying Sterling's formula[1] to $\log[M(T_m)]$, the following result is obtained:

$$\mathcal{W}(\varepsilon;\ m) \simeq [1/\{2\pi m(\Delta\varepsilon^{\pm})^2\}^{1/2}] \exp[-\varepsilon^2/2m(\Delta\varepsilon^{\pm})^2] \quad (13)$$

That is, for $m \gg 1$ and $n \to \infty$ and $(\varepsilon/n\Delta\varepsilon^{\pm} \ll 1)$, the pdf describing the statistics of $\varepsilon(t)$ at the mth epoch is gaussian with a mean $\langle\varepsilon\rangle = 0$ and a variance $\langle\varepsilon^2\rangle = m(\Delta\varepsilon^{\pm})^2$. (In the considerations presented earlier, the equilibrium value is taken as ε_∞, in reference to which, if ε is presumed to fluctuate, then $\langle\varepsilon\rangle = 0$.) Thus, the probability of the temporal statistics of ε, in a broad sense, refers to a superposition of m independent random variables and approaches a gaussian distribution with zero mean (about the equilibrium value) and of a finite variance in the limiting stage of m approaching n. This is in concordance with the central limit theorem.

5. Evolution of $\varepsilon(t)$: Representation via the Fokker–Planck equation

Pertinent to a given environment from which the network learns, if the time between learning steps is drawn from a Poisson distribution, the dynamics of an ensemble of learning processes has been described in [6] by a continuous-time master equation. Presently, the evolution of ε versus time (or number of iterative epochs) can be modeled as a forward equation (or the master equation) of diffusion process (Fokker–Planck equation). This refers to the description of the transition probabilities of ε changing by $\Delta\varepsilon^{+}$ or $\Delta\varepsilon^{-}$ at each step, with the conditions at the commencement of the iterative epochs being $(\varepsilon_0,\ t_0)$. Such a description satisfies a stochastic differential equation given by [14]

$$d\varepsilon(t)/dt = \mu(\varepsilon, t) + \sigma(\varepsilon, t)\zeta(t) \quad (14)$$

where $\zeta(t)$ is a random function such that $\int_0^t \zeta(s)ds$ imposes the attributes of a random walk to the variable $\varepsilon(t)$. In general, $\zeta(t)$ is a stationary, gaussian white noise, suggesting that the dynamics specified by equation (14) is driven by a stationary gaussian process. The evolution of $\varepsilon(t)$ models a continuous brownian motion. It has a pseudoderivative, namely, a time derivative in a mean-squared sense. This pseudoderivative random process $\{d\varepsilon(t)/dt\}$ equals an ideal gaussian white noise $\zeta(t)$ as given by equation (14), where $\zeta(t)$ is zero mean and uncorrelated in time, but has a finite variance $\langle\zeta^2\rangle < \infty$ for all t. Further, in the interval $(t_2 - t_1)$, the entity $n\Delta\tau\{\varepsilon, t_1 \le t \le t_2\}$ is a sample-continuous, second-order markovian process with $\mu(\varepsilon, t)$ and $\sigma(\varepsilon, t)$

[1]Sterling's formula: $\log(x!) = (x + 1/2)\log(x) + \log(\sqrt{2\pi} - x)$.

being the Borel functions of $\varepsilon(t)$ specified within certain bounds. Let the transition probability density function of $\varepsilon(t)$ be denoted by $\mathcal{Z}(\varepsilon, t \mid \varepsilon_0, t_0)$; $s(\varepsilon)$, $-\infty < \varepsilon < \infty$, is a Schwartz function of rapid descent. (That is, s is infinitely differentiable and, for any κ and λ, $\mid \varepsilon \mid^{\kappa} \mid f^{(\lambda)}(\varepsilon) \mid \to 0$ as $\mid \varepsilon \mid \to \infty$.) Suppose an initial condition is imposed such that $\mathcal{Z}(\varepsilon, t \mid \varepsilon_0, t_0) = \delta(\varepsilon - \varepsilon_0)$, where ε_0 refers to the initial value of ε at the onset of the iterative process commencing at $t = t_0$, and

$$\int_{\infty}^{+\infty} s(\varepsilon)\mathcal{Z}(\varepsilon, t) \mid \varepsilon_0, t_0) d\varepsilon \to s(\varepsilon_0) \quad t \to t_0, \quad \forall s \in S \tag{15}$$

Subject to the preceding initial conditions, $\mathcal{Z}(\varepsilon, t) \mid \varepsilon_0, t_0)$ satisfies the Fokker–Planck equation given by [14]

$$\partial\mathcal{Z}/\partial t = (1/2)\partial^2[\sigma^2\mathcal{Z}]/\partial\varepsilon^2 - \partial[\mu\mathcal{Z}]/\partial\varepsilon \tag{16}$$

where $\mathcal{Z} \equiv \mathcal{Z}(\varepsilon, t) \mid \varepsilon_0, t_0)$, $\sigma \equiv \sigma(\varepsilon, t)$, $\mu \equiv \mu(\varepsilon, t)$, and $t_2 > t > t_0 > t_1$. Suppose $\sigma^2(\varepsilon, t) = \lfloor(t)$ and $\mu(\varepsilon, t) = \dashv(t)\varepsilon$. Specific to these prescriptions, the Fokker–Planck equation (equation (16)) reduces to

$$\partial\mathcal{Z}/\partial t = (1/2)\lfloor(t)\partial^2\mathcal{Z}/\partial\varepsilon^2 - \partial\dashv(t)\partial(\varepsilon\mathcal{Z})/\partial\varepsilon \tag{17}$$

which has a solution given by [6]

$$\mathcal{Z}(\varepsilon, t \mid \varepsilon_0, t_0) = (1/\sqrt{2\pi V^2}) \exp[-(\varepsilon - U\varepsilon_0)^2/2V^2] \tag{18}$$

This expression depicts again the gaussian nature of ε with mean and variance being U and V^2, respectively. They can be obtained explicitly by solving [14]: $d(V^2)\Delta = \lfloor + 2\dashv V^2$ and $dU\Delta = \dashv U$ with the initial conditions $V^2(t_0) = 0$ and $U(t_0) = 1$.

The time-dependent wandering of ε under random force as given by equation (14) can also be described by the following Langevin equation [15]:

$$d\varepsilon\Delta + B_L\varepsilon = A_L(t) \tag{19}$$

where $B_L = (2/\Delta\tau)$ and $A_L(t)$ is the random force function. The initial condition, namely, $\varepsilon(t \to t_0 \to 0) = \varepsilon_0$ specifies the solution of equation (19) as

$$\varepsilon(t) - \varepsilon_0 \exp(-2t/\Delta\tau) = \int_0^t [\exp[2(x-t)/\Delta\tau]]A(x)dx$$

The corresponding solution for transition probability $\mathcal{Z}(\varepsilon - \varepsilon_0 e^{-2t/\Delta\tau})$ is given by

$$\begin{aligned} \mathcal{Z}(\varepsilon, t \mid \varepsilon_0, t_0) &= 1/\{2\pi[1 - \exp(-4t/\Delta\tau)]\}^{1/2} \\ &\quad \exp\{-[\varepsilon - \varepsilon_0 \exp(-2t/\Delta\tau)]^2/2(1 - \exp(-4t/\Delta\tau)\} \\ &= [1/(2\pi\sigma_{\mathcal{Z}}^2)^{1/2}] \exp[-(\varepsilon - \bar{\varepsilon}(t)^2/2\sigma_{\mathcal{Z}}^2(t)] \end{aligned} \tag{20}$$

which approaches a delta-dirac function as $t \to t_0 \to 0$ (Figure 3), and $\bar{\varepsilon}(t)$ depicts $\varepsilon_0 e^{-2t/\Delta\tau}$ and $\sigma_{\mathcal{Z}}^2 = [1 - \exp(-4t/\Delta\tau)]$.

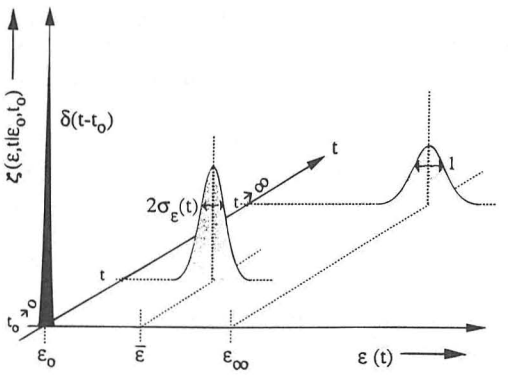

Figure 3: Stochsticity of ε (expressed in terms of the transitional probability of $\Delta\varepsilon^{\pm}$) versus time, t.

Thus, the heuristic discussion on the random walk model $\varepsilon(t)$ as presented earlier has led to the per-unit-time probabilities of the excursions of ε (namely, $\mathcal{W}(\varepsilon;\ m)$ along the temporal framework of the iterative epochs) being modeled as a gaussian process. It is in conformity with the solution of stochastic differential equation representation of the transition probabilities of $\varepsilon(t)$ given by equation (20). Again, as dictated by the initial conditions ($\varepsilon_0,\ t_0$) at the onset of the iterative epochs, it can be observed that the stochasticity of $\varepsilon(t)$ is determined by a gaussian process along the passage of time in the attractor basin.

6. Logistic growth model of $\varepsilon(t)$

The temporal evolution of ε can also be modeled as a logistic growth with an oscillatory trend around its equilibrium (terminal) value (ε_∞), depending on the initial state of the evolutionary process. Suppose at the early stages of iterations of the error feedback imposed on the network, the value of ε is slightly less than the equilibrium value. This reduced value offers a niche to receive reinforcing information that augments the network's efforts toward optimization, leading to a further reduction in the cross-information–based error at the output. However, after a few iterations, the stochasticity associated with the network inputs may yield an output that could lead to negative information, with the result that an enhancement of the error measure (causing an excursion of ε to drift away from the equilibrium value) may be perceived. The causative actions changing ε toward or away from the equilibrium value could also, in general, be asynchronous in reference to the current state of $\varepsilon(t)$ and may be separated to the extent of a few iterative epochs, depicting a delayed arrival in reference to t. Hence, the rate of change of ε caused by $\Delta\varepsilon^{\pm}$ is not a function of the present value $\varepsilon(t)$, but is due to a past value $\varepsilon(t - t_d)$, where t_d is the delay time involved. The corresponding divergent and/or convergent aspect of $\varepsilon(t)$ from the equilibrium value can

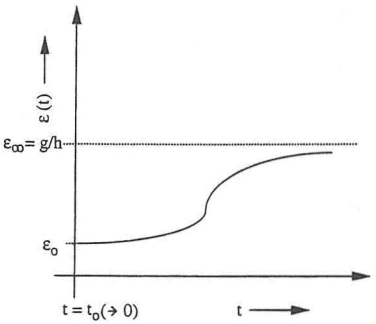

Figure 4: Logistic growth of ε as a function of time (t) seeking the equilibrium value ε_∞ as $t \to \infty$.

therefore be specified by the following growth model, in which the rate of change of $\varepsilon(t)$ is functionally dependent on a past value, $\varepsilon(t - t_d)$:

$$[1/\varepsilon(t)]d\varepsilon(t)/dt = \Gamma(t) \tag{21}$$

where $\Gamma(t)$ depicts the growth rate function, which can be denoted by a simple linear form, as follows:

$$\Gamma(t) = [g - h\varepsilon(t - t_d)] \tag{22}$$

where g is the growth rate of ε without external influences and h represents the effect of augmentation in the values of ε.

Hence, the differential delayed equation depicting the logistic growth of $\varepsilon(t)$ is written as

$$d\varepsilon(t)/dt = \varepsilon(t)[g - h\varepsilon(t - t_d)] \tag{23}$$

For incremental changes in time (Δt), the preceding equation can be written as a second-order difference equation of the form

$$\varepsilon(t + \Delta t) - \varepsilon(t) = (\Delta t)\varepsilon(t)[g - h\varepsilon(t - \Delta t)] \tag{24}$$

With the change of notations given by $t = m\Delta t$, $\varepsilon(t) = \varepsilon(m\Delta t) = \varepsilon_m$, $G = g\Delta t$, and $H = h\Delta t$, the discrete logistic equation at the mth epoch of iteration becomes

$$\varepsilon_{m+1} - \varepsilon_m = \varepsilon_m(G - H\varepsilon_{m-1}) \tag{25}$$

Any delay involved in the preceding process could be asynchronous with respect to the current value of time, namely, t. Further, assuming that the displacements of ε (namely, $\Delta\varepsilon^{\pm}$) are small and are confined to the proximity of the equilibrium value, a linear approximation of the discrete logistic equation is valid. That is, as ε approaches its equilibrium value, $\varepsilon \to \varepsilon_\infty = g/h$ (Figure 4) and

$$\varepsilon(t + \Delta t) - \varepsilon(t) = -G\varepsilon(t - \Delta t). \tag{26}$$

Or, if ε_m is denoted as $(G/H)+\Delta\varepsilon^{\pm}(t)$, where $\mid \Delta\varepsilon_m^{\pm}(t) \mid$ is much smaller than the equilibrium value G/H, the following linearized (approximate) growth relation can be written:

$$y_{m+1} - y_m \simeq -Gy_{m-1} \tag{27}$$

where $y_m = \Delta\varepsilon_m^{\pm}(t)$ is the displacement ε from the terminal value ε_∞ at the mth epochal instant of iteration. Equation (27) can be recast in a more general form as

$$y_{m+1} + P_0 y_m + Q_0 y_{m-1} = 0 \tag{28}$$

which is known as a constant coefficient difference equation. Its solution is analogous to the solution of constant coefficient second-order differential equations, with the necessary conditions for a unique solution being the initial conditions of the first two values of y_0 and y_1 available [16]. Such a solution can be written as

$$y_m = C_1 r_1^m + C_2 r_2^m \tag{29}$$

where the values of r are determined by substitution in equation (28). Hence, it follows that

$$r^{m+1} + P_0 r^m + Q_0 r^{m-1} = 0 \tag{30}$$

Upon division by r^{m-1}, the following quadratic equation is obtained:

$$r^2 + P_0 r + Q_0 = 0 \tag{31}$$

Its two roots are $r_{1,2} = [-P_0 \pm (P_0^2 - 4Q_0)^{1/2}]/2$.

The arbitrary constants C_1 and C_2 in equation (29) can be determined uniquely by the initial conditions, leading to two independent solvable equations,

$$y_0 = C_1 + C_2 \tag{32a}$$

$$y_1 = C_1 r_1 + C_2 r_2 \tag{32b}$$

assuming that the two roots r_1 and r_2 are distinct—that is, $(P_0^2 - 4Q_0) \neq 0$.

The corresponding solution to equation (27) refers to that of a constant coefficient linear difference equation, namely, $r^2 - r + G = 0$ with the roots given by $r_{1,2} = [1 \pm (1 - 4G)^{1/2}]/2$.

The equilibrium state of ε (at the terminal stage) is stable if the solution of y_m does not grow as $t \to \infty$ (i.e., when $m \to \infty$) for any initial conditions $(\varepsilon_0,\ t_0)$. If $0 < G < 1/4$, the roots are real, positive, and less than 1 (i.e., $0 < r_{1,2} < 1$). Consequently, if $0 < G < 1/4$, $y_m = \Delta\varepsilon_m^{\pm}(t) \to 0$ as $t \to \infty$, or the excursion of ε vanishes and $\varepsilon(t)$ will approach the equilibrium value ε_∞ asymptotically. When $G < 1/4$, the two corresponding roots are complex conjugates and hence the solution for $y_m(t)$ is given by

$$y_m(t) = |r|^m (C_3 \cos m\theta + C_4 \sin m\theta) \tag{33}$$

where $|r_1| = |r_2| = |r| = G^{1/2}$, $\theta = \arctan[(4G-1)^{1/2}]$, and $m = (t/\Delta t)$.

The solution of equation (33) grows or decays as it oscillates, depending on $\mid r \mid = G^{1/2}$. When $1/4 < G < 1$, the solution is a decaying oscillation. That is, $y_m \to 0$ and $\varepsilon(t)$ is convergent toward its equilibrium value. When $G > 1$, the equilibrium value may be reached quickly, but the unstability renders an oscillatory growth around the attracted value, leading to a divergent solution.

The divergent growth stems from the asynchronous occurrences of ε_{1i} and ε_{2i}, namely, the reinforcement information ($\Delta\varepsilon^+$) and the degenerating (annihilating) information ($\Delta\varepsilon^-$) fed back via the cross-information error metric. Considering equation (22), suppose g is a positive growth rate (without any external constraints) and h depicts a constraint stipulated by the network as a limiting factor. Let Δt_d refer to the time delay (due to the asynchronous attributes of ε_{1i} or ε_{2i}), which can vary without any limit imposed. Then $g\Delta t_d < 1/4$ would represent an extremely small delay permitting ε_{1i} or ε_{2i} to yield an entity amounting to $\Delta\varepsilon_m^+$, which constitutes reinforcement information by canceling any divergent trend in the current value of ε, namely, $\varepsilon_m(t)$ at the mth epochal iteration. This is possible because $\varepsilon_m(t)$ and $\Delta\varepsilon^+$ occur synchronously due to the negligible delay involved; it would also guarantee an eventual stabilization of $\varepsilon(t)$ at an equilibrium value ε_∞. If the delay is specified by $(1/4 < g\Delta t_d < 1)$, the function $\varepsilon(t)$ would oscillate with larger excursions, but would ultimately seek the equilibrium value with the passage of time.

However, in the case of $g\Delta t_d > 1$, the oscillation would become divergent, destabilizing the optimization effort. That is, an entity $\Delta\varepsilon_m^\pm$ encountered by $\varepsilon_m(t)$ at the mth epochal instant of iteration would predominantly augment any divergent trend in the current value of $\varepsilon_m(t)$. This can happen when $\Delta t_d \to \infty$, meaning that either ε_{1i} or ε_{2i} is absent or disproportionately unbalanced and dissimilar (asymmetric) so that the degenerating information component, namely $\Delta\varepsilon^-$, dominates. Hence, if the reinforcement information contributed by ε_{1i} or ε_{2i} is absent, the chances of $\varepsilon(t)$ to diverge are increased. In other words, for an asymmetric (one-sided/unbalanced) error metric represented via cross-information measures (by either ε_{1i} or ε_{2i} alone), the cumulative augmentation of $\Delta\varepsilon^+$ or $\Delta\varepsilon^-$ renders $\Delta\varepsilon_m^\pm$ to take over the dynamics of $\varepsilon_m(t)$. Therefore, the convergence of the network output toward the teacher value in not guaranteed.

The dynamics of $\varepsilon(t)$ with a logistic functional growth characteristic as discussed previously may cause the function $\varepsilon(t)$ to cross the equilibrium value (Figure 2) at several instants of time. These crossings indicate the bottom of the attractor basins being attained repetitively during the iteration of epochs aimed at the convergence of the network's output toward the teacher values.

7. Convergence considerations

7.1 Stochastical equilibrium

The crossings of the $\varepsilon(t)$ trajectory at the equilibrium value (Figure 6) represent conditions of instantaneous stochastic equilibrium states attained by the vector ε_j. They depict a set of fixed-state point attractors (corresponding to steady-state conditions). Implicitly, at these points the stochastic equilibrium is specified by $\partial \boldsymbol{w}/\partial t = \mathbf{o}$, where $\mathbf{o}$ is the null synaptic matrix (that is, the $\mathcal{R}^m$ null vector $\{\mathbf{o}\}$) and $\boldsymbol{w}$ is the coupling matrix adjusted through feedback via epochal iterations of the error function. In the sample space of the vector $\boldsymbol{\varepsilon}$, $\partial \boldsymbol{\varepsilon}/\partial \boldsymbol{t} = \mathbf{o}$ denotes stability or neural equilibrium, with $\mathbf{o}$ representing the null vector of the changes in error activity. Globally, the neuronal stochastic stability is dictated by the steady-state conditions in the neuronal field, namely

$$\partial \boldsymbol{w}/\partial t = \boldsymbol{\chi}_t \tag{34}$$

and

$$\partial \boldsymbol{x}/\partial t = \boldsymbol{\xi}_t \tag{35}$$

where χ_t is a random vector from a gaussian white random process $\{\boldsymbol{\chi}_t\}$ that can be related to the random vector ζ_t used in equation (14) to model the stochasticity of $\varepsilon(t)$. The neural state vector $\mathbf{x}$ has an associated (independent) gaussian white noise process denoted by $\boldsymbol{\xi}_t$. Equations (34) and (35) represent the stochastic equilibrium conditions vis-à-vis the neuronal state vector $\boldsymbol{x}$ and the synaptic state matrix $\boldsymbol{w}$. Both $\boldsymbol{x}$ and $\boldsymbol{w}$ hover in a brownian motion about (fixed/deterministic) equilibrium, or terminal attractor value, as m (or t) $\rightarrow \infty$, and they reach the state of stochastic equilibrium only when the random vectors $\boldsymbol{\chi}$ and $\boldsymbol{\xi}$ alter them temporally.

As mentioned earlier, the dynamics of neural parameters pursued here correspond to the fluctuations of the error metric $\varepsilon(t)$, which is computed presently in the information-theoretic plane. It offers competitive feedback information that either reinforces or destroys the current informational status at time t (or at the mth iterative epoch) of the adjustments imparted to the network via the weight matrix $\boldsymbol{w}$. Therefore, the relation given by equation (34) can be written as

$$\partial(\Delta \boldsymbol{\varepsilon})/\partial t = \boldsymbol{\chi}_t \tag{36}$$

where $\Delta\boldsymbol{\varepsilon}$ is the fluctuating vector component of the error metric vector set $\{\boldsymbol{\varepsilon}_i\}$. That is, in elucidating the stochastic equilibrium of the NNs, equation (36) can be adopted in lieu of equation (34).

In essence, the error metric fed back and the corresponding corrective algorithms pursued in the information-theoretic plane can be regarded as those pertinent to a competitive learning strategy in the information-theoretic plane. It is also a differential learning pursuit. That is, learning takes place only when a change occurs in $\varepsilon(t)$, namely $\Delta\varepsilon^{\pm}$, and it does so according

to the competitive information provided by $\Delta\varepsilon^+$ or $\Delta\varepsilon^-$. The learning process is associated to an indicator function that flags whether the learning is augmentative/reinforcing as a result of $\Delta\varepsilon^+$ or whether the learning is degenerated due to the addition of annihilating information $\Delta\varepsilon^-$ facilitated via feedback.

For a guaranteed convergence, the weight adjustment requires that the error metric (a distance measure) specified in terms of the cross-entropy (mutual information) parameter of equation (1), computed terms of the pdf of the output (p_i) and that of the teacher values (q_i), should be a balanced (equally weighted) and symmetrized (two-sided) function. The characteristics of such functions are described in the following definitions and theorems and are verified by simulation studies presented later.

7.2 Definitions and theorems

Definition 1. Let $p = \{p_1, p_2, p_3, \ldots, p_k\}$ and $q = \{q_1, q_2, q_3, \ldots, q_N\}$ denotes two complete sets of probabilities $(\sum_i p_i = \sum_i q_i = 1,\ i = 1, 2, \ldots, N)$ representing the a priori probability distribution of the discrete random output $\{\boldsymbol{O}_i\}$ of the NN and that of the teacher source $\{\boldsymbol{T}_i\}$, respectively, with $(p_i,\ q_i)$ corresponding to the ith iterative epoch in the network training schedule. Or, when the network output and the teacher values are specified as continuous variables, p and q refer to the respective probability density functions such that $\int pdp = \int qdq = 1$.

Definition 2. The relative entropy, or cross/mutual, information $I(q \mid p)$ of O with respect to T is defined by the expression

$$I(q \mid p) = \sum_i q_i \log(q_i/p_1) = \varepsilon_{1\mathrm{KL}} \tag{37}$$

Likewise, $I(p \mid q)$ refers to the cross-entropy of T with respect to O. Hence,

$$I(p \mid q) = \sum_i p_i \log(p_i/q_i) = \varepsilon_{2\mathrm{KL}} \tag{38}$$

As indicated before, the transformations expressed by equations (37) and (38) are known as Kullback-Leibler measures [17], and they represent the amount of information contributed by T about O and the amount of information contained in O about T, respectively. That is, they refer to the mathematical expectation of the transinformation about (or directed divergence of) each outcome of T versus O and O versus T, respectively. Hence, it follows that

$$\begin{aligned} I(q \mid p) &= \langle(\text{prior uncertainty})_p - (\text{posterior uncertainty})_q\rangle \\ &= \sum_i q_i\{[-\log(p_i)] - [-\log(q_i)]\} \end{aligned} \tag{39}$$

Likewise,

$$\begin{aligned} I(p \mid q) &= \langle(\text{prior uncertainty})_q - (\text{posterior uncertainty})_p\rangle \\ &= \sum_i p_i\{[-\log(q_i)] - [-\log(p_i)]\} \end{aligned} \tag{40}$$

Definition 3. The distance between two probability distributions refers to a divergence measure between them and is given by Kullback-Leibler-Jensen metric [18] defined as follows:

$$J(p \mid q) = I(p \mid q) + I(q \mid p) = \varepsilon_{\mathrm{KLJ}} \tag{41}$$

The J measure refers to the divergence, or the discrimination, between the hypotheses $\mathcal{H}_{\boldsymbol{O}}$ and $\mathcal{H}_{\boldsymbol{T}}$ (constituted by O and T respectively), or between p and q, and it implicitly represents a measure of difficulty in discriminating between them. That is,

$$\begin{aligned} \varepsilon_{\mathrm{KLJ}} &= J(p \mid q) = I(q \mid p) - [-I(p \mid q)] \\ &= \sum_i p_i \log(p_i/q_i) - \sum_i q_i \log(p_i/q_i) = \varepsilon_{1\mathrm{KL}} + \varepsilon_{2\mathrm{KL}} \\ &= \sum_i (p_i - q_i) \log(p_i/q_i) \end{aligned} \tag{42}$$

Definition 4. If $\Phi(x)$ is a convex function for $x > 0$, with $\Phi(1) = 0$, then the f-divergence (f depicting the function Φ) of a distribution p or q is defined in a two-sided form of the Csiszár error measure [8] with weighting factors K_1 and K_2 as

$$\varepsilon_{\mathrm{cz}} = I^f(p \mid q) = K_1 \sum_i q_i \Phi(p_i/q_i) + K_2 \sum_i p_i \Phi(q_i/p_i) \tag{43}$$

where $\Phi(x) = x \ \log(x)$, $\varepsilon_{\mathrm{Cz}} \rightarrow \varepsilon_{\mathrm{KLJ}}$, and $\varepsilon_{\mathrm{Cz}}$ is a more generalized version than $\varepsilon_{\mathrm{KLJ}}$.

Theorem 1. *If the f-divergence as defined by equation (44) is to be considered as a feasible error metric ($\varepsilon_{\mathrm{Cz}}$) in training a multilayered NN, then the necessary condition is that $I^f = \varepsilon_{\mathrm{Cz}}$ must be two-sided and bounded. That is,*

$$\begin{aligned} (I^f = \varepsilon_{\mathrm{Cz}}) &= K_1 \sum_i q_i \Phi(p_i/q_i) + K_2 \sum_i p_i \Phi(q_i/p_i) \\ &= (\varepsilon_{1\mathrm{Cz}} + \varepsilon_{2\mathrm{Cz}}) \end{aligned} \tag{44}$$

In equation (45), the condition for balanced symmetrization is that the weighting factors be equal (i.e., when $K_1 = K_2$).

Proof. This theorem can be proved by the geometrical notions of Pythagoras as follows: The relative entropy-based error metric $\varepsilon(p|q,\ q \mid p)$ behaves intuitively like the square of the euclidean distance norm, although $\varepsilon(p \mid q)$ itself represents no geometrical measure. For a convex set ε in $\mathcal{R}^m$, let $\mathcal{A}$ be a point outside the set, $\mathcal{B}$ be the point in the set closest to $\mathcal{A}$, and $\mathcal{C}$ be any other point in the set. Then the angle between the lines $\mathcal{BA}$ and $\mathcal{BC}$ must be obtuse, which implies via the Pythagorean theorem that $\ell^2_{\mathcal{AC}} \geq \ell^2_{\mathcal{AB}} + \ell^2_{\mathcal{BC}}$, where ℓ represents the linear distance. Hence the convergence of ε (toward an infimum) in the ℓ norm refers to the minimum distance between the two distributions. That is, the infimum of $\ell^2_{\mathcal{AC}} = (\ell^2_{\mathcal{AB}} + \ell^2_{\mathcal{BC}})$, or $\varepsilon_{\mathrm{Cz}}(p \mid q,\ q \mid p) = (\varepsilon_{1\mathrm{Cz}} + \varepsilon_{2\mathrm{Cz}})$ where $\varepsilon_{1\mathrm{Cz}} = K_1 \sum q_i \phi(p_i/q_i)$ and $\varepsilon_{1\mathrm{Cz}} = K_1 \sum p_i \phi(q_i/p_i)$. ■

Theorem 2. *The sufficient condition for* $(I^f = \varepsilon_{Cz})$ *representing an error metric for training a NN is that both parts of* ε, *namely* ε_1 *and* ε_2, *be nonzero in their syntactic values so that the corresponding semantics imparted to the network (via feedback through the corrective algorithm) add meaningful information to the weight adjustments in the multilayered network during each iterative epoch and lead to an eventual convergence of the output error toward an equilibrium value* ε_∞.

Proof. The constituent part of ε_{Cz}, namely $\varepsilon_{1Cz} \in (0, \Delta\varepsilon^{\pm}]$ and $\varepsilon_{2Cz} \in (0, \Delta\varepsilon^{\pm}]$ carry messages of relative importance and are applied to the system dynamics (via feedback), which allows the state variable $\varepsilon_{Cz}(t)$ to converge toward an attractor (ε_∞) at a given kth instant of iteration.

Let ε_{1Cz} supply a "message of relative importance" given by $[\mathcal{M}_{1k}/\mathcal{M}]$ at the kth instant, and $\mathcal{M} = \sum(\mathcal{M}_{1k} + \mathcal{M}_{2k})$. At the same instant, the corresponding message of relative importance imparted by ε_{2Cz} is given by $[\mathcal{M}_{2k}/\mathcal{M}] = [1 - (\mathcal{M}_{1k}/\mathcal{M})]$ since the semantic aspects of ε_{1Cz} and ε_{2Cz} complement each other.

Considering the total messages delivered over k iterations, it is given by

$$\mathcal{M}_{\text{total}}/\mathcal{M} = \sum_k [\mathcal{M}_{1k}/\mathcal{M}] + \sum_k [1 - \mathcal{M}_{1k}/\mathcal{M}]. \tag{45}$$

Equation (45), specified in terms of a controlling (cybernetic) information parameter of the network, say C_ε (which results from the error feedback by the control loop), can be written in reference to the kth iteration as

$$C_{\varepsilon k} = (I^k_{\varepsilon^+} + I^k_{\varepsilon^-}) \Rightarrow \{\Delta\varepsilon^{\pm}\}. \tag{46}$$

Here, I_{ε^+} can be regarded as the reinforcing information (which directs the output error toward the equilibrium value ε_∞), and I_{ε^-} refers to annihilation information, which leads the system dynamics to diverge from the equilibrium. Dominance of I_{ε^-} implies an information deficiency, and the overwhelming influence of I_{ε^+} means that information augmentation is perceived by the system dynamics in pursuit of equilibrium or an attractor value. As in [19], a parameter such as $C_{\varepsilon k}$ either "sensitizes" or "desensitizes" the convergence process (depending on the dominance of the messages delivered by I_{ε^+} or I_{ε^-} at any given kth instant in the network optimization strategy).

Considering $\Delta\varepsilon^+$ and $\Delta\varepsilon^-$ (which determine I_{ε^+} and I_{ε^-}) as dichotomous events, their repeated occurrences constitute Bernoulli trials with binomial distribution. Suppose ε is single-sided (i.e., ε_{Cz} is assumed to be constituted by ε_{1Cz} or ε_{2Cz} alone). The corresponding number of occurrences of $\Delta\varepsilon^+$ and $\Delta\varepsilon^-$ will be unbalanced significantly over $n \to \infty$ iterations (trials). This unbalanced condition allows either $\Delta\varepsilon^+$ or $\Delta\varepsilon^-$ to dominate as $n \to \infty$, offering one-sided information to the control dynamics. Hence, the system will diverge positively or negatively from the equilibrium (ε_∞) depending on the dominance of $\Delta\varepsilon^+$ and $\Delta\varepsilon^-$, respectively (Figure 2). On the other hand, if both ε_{1Cz} and ε_{2Cz} are present (two-sided, symmetrized representation of

ε_{Cz}), the Bernoulli events of $\Delta\varepsilon^{\pm}$ are shared by ε_{1Cz} and ε_{2Cz} over n iterations. This leads to a balanced state with $\Delta\varepsilon^{+}$ and $\Delta\varepsilon^{-}$ sharing about $n/2$ iterations each. Because of the system's stochasticity, however, $\Delta\varepsilon^{+}$ may dominate marginally in number of occurrences ($= n/2 + \Delta$, $\Delta \ll n/2$) so that $\Delta\varepsilon^{-}$ occurs $(n/2 - \Delta)$ times. This unbalanced state of sharing the iterations with a marginal dominance in number by $\Delta\varepsilon^{+}$ will augment the necessary information cumulatively, as would be required for convergence (from the negative side). Likewise, if $\Delta\varepsilon^{-}$ occurs $(n/2+\Delta)$ times, the convergence will be directed from the positive side (Figure 2).

Existence of both ε_1 and ε_2 as constituent parts of ε is therefore a necessary condition for the network to converge. Hence, the competitive aspect of $I_{\varepsilon+}$ and $I_{\varepsilon-}$ facilitated by the dichotomous occurrences of $\Delta\varepsilon^{\pm}$ will ultimately decide the convergence toward attractor(s), and it can be realized by proper choice of network parameters (such as the learning coefficient) and by adopting symmetrically weighted ε_1 and ε_2. ■

7.3 Simulation and results

Shown in Figure 1(b) is a test NN (multilayer perception) with 24 input units, 2 hidden layers (each with 10 units), and a single output. It was trained to recognize a teacher function $|\sin(x)|$ using the error metrics given by equations (37), (38), and (41). Presented in Figures 5 and 6 are the results pertinent to (i) histograms of error-value distributions (Figure 5) and (ii) trajectories of the errors as functions of the number of epochal iterations performed (Figure 6). The actual and simulated test functions $|\sin(x)|$ versus the argument x are presented in Figure 7. The simulated function in Figure 7 corresponds to the symmetric error measure given by equation (41), which converged to the equilibrium status as depicted in Figure 6. The asymmetric error measures given by equations (37) and (38) failed in leading the network's performance toward convergence. Their trajectories, as can be seen from Figure 6, veered off from the equilibrium value with the discourse of iterations. Thus, this simulation study confirms the need for a symmetrized error measure for NN learning applications in the information-theoretic domain. In the simulation studies performed, the learning coefficient was taken as $\eta = 0.01$ and the nonlinear sigmoidal function was the Bernoulli function $L_Q(x)$ with $Q = 10$ [1]. They were chosen so as to realize minimum root–mean squared value of the deviations of the predicted function from the teacher function at 50 equally spaced arguments.

8. Further considerations on the dynamics of $\varepsilon(t)$

8.1 Competing augmentative and annihilative information species

As discussed in the previous section, the control dynamics of a NN are dictated by the competition of reinforcing information $I_{\varepsilon+}$ and by the annihilating counterpart $I_{\varepsilon-}$. The eventual convergence (or divergence) of network

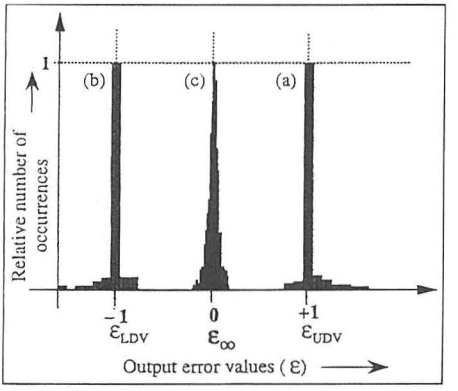

Figure 5: Histograms of output error value distributions corresponding to (a) one-sided Kullback-Leibler error measure, ε_{1KL}, equation (37); (b) one-sided Kullback-Leibler error measure, ε_{2KL}, equation (38); and (c) symmetrized Kullback-Leibler-Jensen error-measure, ε_{KLJ}, equation (41). (Note: Relative number of occurrences in each case refers to value normalized with respect to the maximum value; and as $t \rightarrow \infty$, the lower divergent value ε_{LDV} of ε_{2KL}, the equilibrium limit of ($\varepsilon_{1KL} = \varepsilon_{1KL} + \varepsilon_{2KL}$) ε_{∞}, and the upper divergent value ε_{UDV} of ε_1 are set at -1, 0, and $+1$, respectively.)

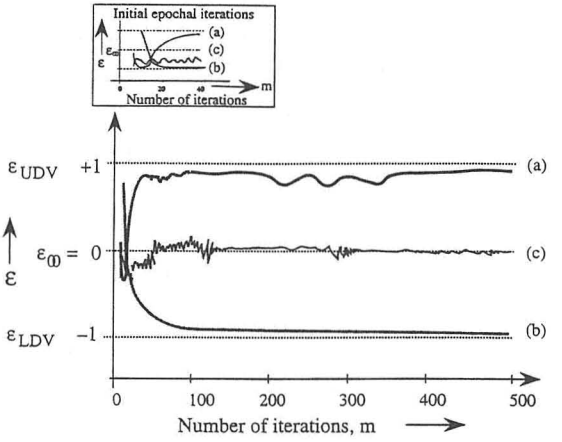

Figure 6: Computed trajectories of ε versus the number of iterations, m, with the test network of Figure 1(b) subjected to simulation studies: (a) for ε_{1KL} equation (37); (b) for ε_{2KL} equation (38); and (c) for ε_{KLJ} equation (41).

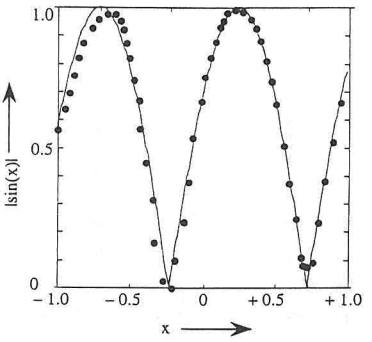

Figure 7: Actual (—) and simulated (· · ·) test functions.

performance is decided by the dominance of $\Delta\varepsilon^+$ or $\Delta\varepsilon^-$ facilitated by the ε_1 and ε_2 constituents of the error metric. The growth of the $\Delta\varepsilon^+$ and $\Delta\varepsilon^-$ species (which eventually decides the convergence or divergence of the network performance) can be presumed to depend on the population of both species. That is, if we depict the population of $\Delta\varepsilon^+$ species by n_1 and that of $\Delta\varepsilon^-$ species by n_2, the dynamics of n_1 and n_2 can be represented in terms of arbitrary functions $\mathcal{Y}$ and $\mathcal{Z}$ as follows:

$$dn_1\Delta = \mathcal{Y}(n_1, n_2) \tag{47a}$$
$$dn_2\Delta = \mathcal{Z}(n_1, n_2) \tag{47b}$$

Correspondingly, both populations may affect each other negatively, so that the interaction between the species is competitive. That is, the growth rate of each species will be retarded by the presence of the other. From equations (47), it follows (by eliminating the explicit dependence on the time factor t) that

$$dn_2/dn_1 = \mathcal{Y}(n_1, n_2)/\mathcal{Z}(n_1, n_2) \tag{48}$$

which represents the phase-plane representation of $\Delta\varepsilon^+$ and $\Delta\varepsilon^-$.

The two competing species, $\Delta\varepsilon^+$ and $\Delta\varepsilon^-$ are virtually identical in their information values; they differ only in dictating the convergence process to occur in opposite directions. As presumed earlier, attributing a marginal unbalance to the competition so that $\Delta\varepsilon^+$ is dominant (by letting $n_1 > n_2$), the following explicit equations can be specified in lieu of equations (47):

$$dn_1/dt = n_1(a_1 - b_1 n_1 - c_1 n_2) \tag{49a}$$
$$dn_2/dt = n_1(a_1 - b_1 n_2 - c_2 n_1) \tag{49b}$$

and the corresponding phase-plane equation is given by

$$dn_2/dn_1 = n_2(a_1 - b_1 n_2 - c_2 n_1)/n_1(a_1 - b_1 n_1 - c_2 n_2) \tag{50}$$

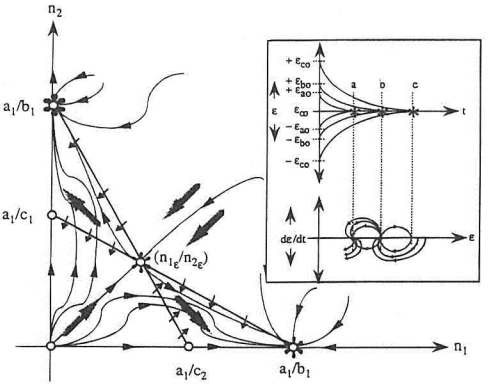

Figure 8: Phase-plane diagram in reference to equilibrium dynamics of interaction between two nearly identical, competing species of error metrics.

Inasmuch as $n_1 > n_2$, the interaction between $\Delta\varepsilon^+$ and $\Delta\varepsilon^-$ is strongly competitive in the sense that the interaction terms, $-c_1 n_1 n_2$ and $-c_2 n_1 n_2$, are greater (as $n_1 \to n_2 \to n/2$) than the self-interaction terms, $-b_1 n_1^2$ and $-b_1 n_2^2$. Thus, $c_1 > b_1$ and $c_2 > b_1$. Further, since $c_1 > c_2$, the resulting conditions lead to the inequality $c_1 > c_2 > b_1$, in which case an equilibrium state can be reached. Sketching the phase-plane diagram as shown in Figure 8, one can see that the isoclines are straight lines with positive n_1 and n_2 intercepts; the equilibrium states are each marked with an asterisk on the diagram. Designating the equilibrium population as n_{1E} and n_{2E},

$$n_{1E} = (a_1 c_1 - a_1 b_1)/(c_1 c_2 - b_1^2) \tag{51a}$$

$$n_{2E} = (a_1 c_2 - a_1 b_1)/(c_1 c_2 - b_1^2) \tag{51b}$$

Analysis pertinent to the stability of this problem [20] leads to the principle of competitive exclusion, meaning only one species can ultimately survive. The solution curves for this problem can be sketch as shown in Figure 8 by classifying the equilibrium points on the basis of the following considerations.

1. Coexistent equilibrium population is a saddle point (being always unstable).

2. A species that eliminates its competition is a stable node.

Thus, the unbalanced parts of I_{ε^+} and I_{ε^-} contributed by $\Delta\varepsilon^{\pm}$ constituents (of the coexisting ε_1 and ε_2 terms) in the error metric feedback toward network training can facilitate a stable control/dynamics with an eventual equilibrium of the system (or seeking the convergence toward the objective function).

Thus, optimization in NNs implies the convergence of the learning process mediated by a cost function such as $\varepsilon(t)$ to an attractor implicitly. The

locations of these attractors and their basins in the phase space are dictated by the weight modifications, that is, by the iterative adjustments of $\mathcal{W}_{ij}$ as a result of the supervised learning foreseen. The corresponding nonlinear dynamics follow a random walk paradigm–based information flow. Convergence toward the attractor also refers to the trend in the network's performance temporally approaching stored vectors/memory configurations.

9. Concluding remarks

The focus of this paper is twofold: (1) it portrays the dynamics of the learning process in NNs; and (2) the relevant portrayals are referred to the information-theoretic plane. Within the broad scope of the aforesaid considerations, the major inferences and conclusions that can be gathered from the analysis are as follows.

- The stochastic dynamics associated with the neural learning process can be comprehended in the information-theoretic plane (as it can be done in the parametic space plane).
- The relevant dynamics can be specified in terms of a class of error metrics of the network, which can be elucidated in the information-theoretic plane for the purpose of network learning optimization using the aforesaid error metrics (ε) as feedback entities. Hence, the relevant dynamics refer to ε versus time (t) over which the epochs of iterations of error feedback are performed to achieve the convergence.
- The associated stochasticity models the dynamics of $\varepsilon(t)$ in terms of a probability function versus time as governed by the Fokker–Planck diffusion equation.
- The dynamics of $\varepsilon(t)$ can be specified by a logistic growth model depicting equilibrium conditions.
- Learning dynamics analyzed indicate that in backpropagation mode, the network training follows the same type of gradient descent algorithm in the information-theoretic plane as in the parametic space plane.
- The convergence or divergence aspects of ε with the passage of time (or along the iterative epochs of error feedback) depend on the competitive role played by augmenting and annihilating information imparted to the system by the error information feedback.
- Corresponding values of $\Delta\varepsilon^+$ and $\Delta\varepsilon^-$ (deviatory measures of ε from the equilibrium value ε_∞) constitute dichotomous events repeated along the discourse of iterations performed. This Bernoulli process has binomial distribution on a discrete basis. As n (the number of events of $\Delta\varepsilon^\pm$) $\to 0$, this distribution becomes a gaussian process. Excessive unbalance between $\Delta\varepsilon^+$ and $\Delta\varepsilon^-$ leads to divergence in the network's performance. A near-balanced state, however, enables convergence.

- In the terminal region, that is, at $t \rightarrow \infty$, the convergence endeavor could set $\varepsilon(t)$ as a stationary process [21]. Apart from this terminal attractor status, during the discourse of ε versus t the error metric value may also cross the equilibrium value ε_∞ at several instants of time, each representing an attractor in the basin of convergence.

References

[1] Neelakanta, P. S., Sudhakar, R., and De Groff, D., "Langevin Machine: A Neural Network Based on Stochastically Justifiable Sigmoidal Function," *Bio. Cybern.*, **65** (1991) 331–338.

[2] Kesidis, W., "Analog Optimization with Wong's Stochastic Neural Network," *IEEE Trans. Neural Networks*, **6** (1995) 258–260.

[3] Usami, H., Masaki, S., and Sato, R., "The Stochastic Properties of the Basic Neuron Populations as Information Processing Systems," *Biol. Cybern.*, **29** (1978) 167–179.

[4] Bergström, R. M., "Neural Macrostates," *Synthese*, **17** (1967) 245–443.

[5] Amari, S., Kurata, K., and Nagaoka, H., "Information Geometry of boltzmann Machines," *IEEE Trans. Neural Networks.*, **3** (1992) 260–272.

[6] Heskes, T. M., and Kappen, B., "Learning Processes in Neural Networks," *Phy. Review A*, **44** (1991) 2718–2726.

[7] Park J. C., Neelakanta, P. S., Abusalah, S., De Groff, D., and Sudhakar, R., "Information-theoretic Based Error-metrics for Gradient-descent Learning in Neural Networks," *Complex Systems*, **9** (1995) 287–304.

[8] Csiszár, I., "A class of Measures of Informativity of Observation Channels," *Periodica Math. Hungarica*, **2** (1972) 191–213.

[9] Kullback, S., and Leibler, R. A., "On Information and Sufficiency," *Ann. Math. Stat.*, **22** (1951) 79–86.

[10] Bergström, R. M., and Nevanlinna, O., "An Entropy Model of Primitive Neural Systems," *Intern. J. Neuroscience*, **4** (1972) 171–173.

[11] Shannon, C. E., "A Mathematical Theory of Communication," *Bell Syst. Tech. J.*, **27** (1948) 623–659.

[12] Neelakanta, P. S., and De Groff, D., *Neural Network Modeling* (CRC Press, Boca Raton, FL, 1994).

[13] Norwich, K. H., *Information, Sensation, and Perception* (Academic Press, San Diego, CA, 1993).

[14] Wong, E., *Stochastic Process in Information and Dynamic Systems*, (Robert E. Krieger Publishing Co., Huntington, NY, 1979).

[15] Langevin, P., Sur la théorie du movement brownien, C. R. Hebd. *Séances Acad. Sci.*, **146** (1908) 503–533.

[16] Haberman, R., *Mathematical Models* (Prentice-Hall, Englewood Cliffs, NJ, 1977).

[17] Kullback, S., *Information Theory and Statistics* (John Wiley & Sons, New York, 1959).

[18] Lin, J., "Divergence Measures Based on Shannon Entropy," *IEEE Trans. Inform. Theory*, **17** (1991) 145–151.

[19] Haken, H., *Information and Self-Organization* (Springer-Verlag, Berlin, 1988).

[20] Haberman, R., *Mathematical Models* (Prentice-Hall, Englewood Cliffs, NJ, 1977).

[21] Zak, M., "Terminal Attractors in Neural Networks," *Neural Networks*, **2** (1989) 259–274.