order. Neuronal analogues in such networks usually employ sigmoidal activation functions, which in some respects may be particularly suitable: they possess limited ranges, as do probability functions, and when class distributions have monotonic overlapping tails, posterior probability for a class c will in fact be sigmoidal over paths of increasing D_c and decreasing $\overline{D}_c$ in the input space. However, in a real problem, hypersurfaces of level probability in X may possess curvature, while for "neural" elements which pass an inner product of weights and inputs to their sigmoid activation functions, hypersurfaces of level output are hyperplanes in the input space. Formation of nonplane decision boundaries requires multilayer networks of such elements, or elements whose net inputs to their activation functions are higher order functions of their individual inputs. Horne and Hush [1] have shown that a "neuron" with a logistic activation function and with a net input composed of a quadratic form as well as linear and constant terms is capable of learning (by minimization of sum-square error) exact representation of posterior probability in a two-class problem, in which both class probability densities are Gaussian. Interestingly, in the case where the covariance matrices for the two classes are identical, the hypersurface of equal probability becomes a hyperplane, and the common "neuron" with linear and constant input terms and a logistic activation function is capable of exact representation of posterior probability of one or the other of the classes.

Even if a network or other model is capable, for a particular choice of its parameters, of good approximation to posterior probabilities in some classification problem, there is no guarantee that other, poorer local minima in the weighted sum of squares of the deviations δf_c do not exist. Such minima may prove to be obstacles to practical application of learning techniques such as gradient descent (e.g., error back-propagation [10]), just as may be the case with deterministic problems.

A number of authors have suggested alternative objective functions for training based upon information-theory and statistical considerations [5], [6], [12], [13], which may offer particular advantages over square-error, and which are generally compatible with back-propagation. The efficacy of such functions, the presence and avoidance of local suboptimal minima, and the overall performance of neural network models in learning to approximate posterior probabilities are promising subjects for further investigation.

Acknowledgment

The author thanks H. White, M. Carlin, and R. Shimabukuro for helpful comments.

References

[1] B. Horne and D. Hush, "On the optimality of the sigmoid perceptron," in *Proc. Int. Joint Conf. Neural Networks* (Washington), Jan. 1990, pp. 269-272.

[2] W. Y. Huang and R. P. Lippmann, "Comparisons between neural net and conventional classifiers," in *Proc. IEEE First Int. Conf. Neural Networks*, 1987, pp. 485-493.

[3] D. R. Hush and J. M. Salas, "Classification with neural networks: A comparison," in *Proc. Ideas in Science and Electronics Symp.* (Albuquerque), May 1989, pp. 107-114.

[4] L. Niles, H. Silverman, G. Tajchman, and M. Bush, "How limited training data can allow a neural network to outperform an 'optimal' statistical classifier," in *Proc. 1989 Int. Conf. Acoust., Speech, Signal Process.*, 1989, pp. 17-20.

[5] E. B. Baum, "Supervised learning of probability distributions by neural networks," *Neural Information Processing Systems*, D. Z. Anderson, Ed. New York: American Institute of Physics, 1988, pp. 53-61.

[6] H. Gish, "A probabilistic approach to the understanding and training of neural network classifiers," in *Proc. Int. Conf. Acoust., Speech, Signal Process.*, 1990, pp. 1361-1364.

[7] E. Levin, N. Tishby, and S. A. Solla, "A statistical approach to learning and generalization in layered neural networks," in *Proc. Second Annual Workshop on Computational Learning Theory*, 1989, pp. 245-260.

[8] H. White, "Learning in artificial neural networks: A statistical perspective," *Neural Computation*, vol. 1, pp. 425-464, 1989.

[9] H. White, "Consequences and detection of misspecified nonlinear regression models," *J. Amer. Statist. Assoc.*, vol. 76, pp. 419-433, 1981.

[10] D. E. Rumelhart, G. E. Hinton, and R. J. Williams, "Learning internal representations by error propagation," in *Parallel Distributed Processing, Explorations in the Microstructure of Cognition*, D. E. Rumelhart and J. L. McClelland, Eds., vol. 1. Cambridge: MIT Press, 1986, pp. 318-362.

[11] H. White, "Connectionist nonparameteric regression: Multilayer feedforward networks can learn arbitrary mappings," *Neural Networks*, to be published.

[12] A. El-Jaroudi and J. Makhoul, "A new error criterion for posterior probability estimation with neural nets," in *Proc. Int. Joint Conf. Neural Networks* (San Diego, CA), June 1990, pp. 185-192.

[13] J. Movellan, "Error functions to improve noise resistance and generalization in backpropagation networks," in *Proc. Int. Joint Conf. Neural Networks* (Washington, DC), Jan. 1990, pp. 557-560.

Dispersive Propagation Skew Effects in Iterative Neural Networks

Seho Oh and Robert J. Marks II

***Abstract*—During communication between neurons in a continuous-time analog neural network, propagation skew typically varies from neuron pair to neuron pair. For no dispersion, we have previously demonstrated that the steady-state performance of an iterative neural network is not affected if the combination of the neural network's weights and neural nonlinearity is contractive. In this letter, this result is extended to the case of dispersive skew. We show that, under nearly the same conditions, the same steady-state result will occur in the neural network in the presence of dispersive skew.**

In high-speed iterative neural processors, the propagation delay between neurons and the response time of neurons to stimuli can be strong factors affecting the performance of the neural network. This is true in high-speed analog electronic neural networks, where the transmission line characteristics of interconnects must be considered, and in optical neural networks, where the interconnects are affected by the physics of the optics [1]. In this letter, we show that dispersive propagation skew between neurons in an iterative neural processor does not affect the steady-state solution when the neural network's weights and nonlinearities meet certain contractive criteria.

Consider analog neuron i communicating its state to neuron j through weight a_{ij}. In the case of nondispersive skew, the delay time between these neurons, possibly proportional to their optical or physical separation, is τ_{ij}. If the interconnect displays dispersive skew, on the other hand, the received signal at neuron j will be temporally spread, rather than localized.

We have established conditions for the convergence of iterative neural networks in the presence nondispersive skew for the case of linear operations and have generalized the result to include sigmoidal nonlinearities [2]. In practice, however, most processors have dispersive characteristics. In this letter, we deal with the skew effects for the same operation as before in the presence of dispersive skew. We demonstrate that dispersive skew, under the same conditions as in [2], does not affect the steady-state solution of the neural network if the dispersion function has unit area and the causal forcing functions satisfying a continuity criterion at the origin.

Considering the nonskewed and nondispersive iteration of neural

Manuscript received April 27, 1990.

The authors are with the Department of Electrical Engineering, University of Washington, Seattle, WA 98195.

IEEE Log Number 9040686.

S. Oh and R.J. Marks II, "Dispersive propagation skew effects in iterative neural networks",
IEEE Transactions on Neural Networks, vol.2, pp.160-162, (1991).

states $\{x_i(t) \mid 1 \le i \le N\}$

$$x_i(t) = \eta_i\left[\sum_{j=1}^{N} a_{ij}x_j(t-\tau) + f_i(t)\right] + g_i(t), \quad i = 1, 2, \cdots, N \tag{1}$$

where τ is the common time delay for communication between neuron pairs, a_{ij} is the weight between neurons i and j, η_i is a memoryless nonlinearity (e.g., a sigmoid), and $f_i(t)$ and $g_i(t)$ are forcing functions. We assume $f_i(t) \to f_i(\infty) =$ Constant and $g_i(t) \to g_i(\infty) =$ Constant for all i. Neural networks to which this model is applicable include those of Hopfield and the alternating projection neural network [3]-[7].

For nondispersive skew, the delay τ in this expression in simply replaced by τ_{ij}, which is the time delay between neurons i and j. The feedback operation with dispersive skew can be written as

$$x_i(t) = \eta_i\left[\sum_{j=1}^{N} a_{ij}\int_0^t h_{ij}(t-\xi)x_j(\xi-\tau_{ij})\,d\xi + f_i(t)\right] + g_i(t) \tag{2}$$

where $h_{ij}(t)$ is the temporal dispersion spread from neuron i to j.

We first discuss the asymptotic stability and convergence for the special case of the linear operation:

$$x_i(t) = \sum_{j=1}^{N} a_{ij}\int_0^t h_{ij}(t-\xi)x_j(\xi-\tau_{ij})\,d\xi + f_i(t). \tag{3}$$

Assuming stability, the steady-state solution of the corresponding linear version of (1) is [2]

$$\vec{x}(\infty) = [\boldsymbol{I} - \boldsymbol{A}]^{-1}\vec{f} \tag{4}$$

where $\vec{x}(t) = [x_1(t), x_2(t), \cdots, x_N(t)]^T$, $\vec{f} = [f_1(\infty), f_2(\infty), \cdots, f_N(\infty)]^T$, and $\boldsymbol{A} = (a_{ij})$.

Let $H_{ij}(s)$, $F_i(s)$, and $G_i(s)$ denote the (unilateral) Laplace transform of $h_{ij}(t)$, $f_i(t)$, and $g_i(t)$ respectively. We assume that $F_i(s)$ is analytic for Re $(s) \ge 0$ except $s = 0$ and

$$\lim_{|s|\to\infty}\left|sF_i(s)\right| = 0 \quad \text{and} \quad \lim_{|s|\to\infty}\left|sG_i(s)\right| = 0, \quad i = 1, 2, \cdots, N \tag{5}$$

when Re $(s) \ge 0$. From the initial value theorem of Laplace transforms, this condition requires that the forcing functions be continuous at the origin. In addition, we require that $H_{ij}(s)$ be analytic in Re $(s) \ge 0$. The following lemma establishes basic conditions for convergence of the linear iteration in the presence of dispersive skew.

Lemma 1: Let $\boldsymbol{A}(s) = (a_{ij}H_{ij}(s)e^{-s\tau_{ij}})$. If $\|\boldsymbol{A}(s)\| < 1$ for Re $(s) \ge 0$ and $H_{ij}(0) = 1$ for all i, j then (3) is stable and convergences to the result in (4).

Note that the criterion that $H_{ij}(0) = 1$ is identical to requiring that $h_{ij}(t)$ have unit area. A proof of Lemma 1 is given in the Appendix. Lemma 1 provides the foundation for the two following important special lemmas.

Lemma 2: Let $\boldsymbol{B} = (|a_{ij}|)$. If $\|\boldsymbol{B}\| < 1$, $|H_{ij}(s)| \le 1$ for all Re $(s) \ge 0$ and $H_{ij}(0) = 1$, then (3) is stable and converges to the result in (4).

Lemma 3: Assume that the nondispersive version of (3) converges to the result in (4). Let the dispersion be separable in the sense that $H_{ij}(s) = H_i^{(1)}(s) \cdot H_j^{(2)}(s)$. Let $H_{ij}(0) = 1$. If $|H_i^{(1)}(s)| \le 1$ and $|H_j^{(2)}(s)| \le 1$ for all Re $(s) \ge 0$, then (3) is stable and converges to the result in (4).

Lemma 3 is a generalization of the result in [2] for nondispersive skew, which states stability and accuracy of the iteration are ensured if we can write the nondispersive skew as $\tau_{ij} = u_i + v_j$. Proofs of Lemmas 2 and 3 are given in the Appendix.

We can now state our main result. (The pointwise vector operator $\boldsymbol{\eta}$ is comprised of the η_i nonlinearity.)

Theorem 1: For a given matrix $\boldsymbol{A}$, time delays $\{\tau_{ij}\}$ and dispersions $\{h_{ij}(t)\}$, if (3) converges for every $\vec{f}(t)$ which satisfies (5) and $\boldsymbol{\eta}$ is a nonexpansive operation, then (2) is stable and converges to the same steady-state solution as that obtained in the absence of any skew.

The vector operator $\boldsymbol{\eta}$ is said to be nonexpansive if $\|\boldsymbol{\eta}(\vec{x}) - \boldsymbol{\eta}(\vec{y})\| \le \|\vec{x} - \vec{y}\|$ for all $\vec{x}$ and $\vec{y}$. For the sigmoidal operator, $\eta_i(x) = (1 + e^{-x})^{-1}$, for example, $|d\eta_i/dx| \le 1$ and $\boldsymbol{\eta}(\cdot)$ is nonexpansive. The proof of the theorem is given in the Appendix.

In summary, we have shown that a dispersively skewed nonlinear neural iteration is stable and properly converges if the neural nonlinearity is nonexpansive and the corresponding dispersively skewed linear iteration is stable. The stability and proper convergence of the dispersively skewed linear iteration can be ensured by adherence to the criteria in either of the three lemmas.

Appendix

Proof of Lemma 1

We take the Laplace transform of (3):

$$X_i(s) = \sum_{i=1}^{N} a_{ij}H_{ij}(s)\exp(-s\tau_{ij})X_j(s) + F_i(s)$$

or, equivalently, in matrix form,

$$\vec{X}(s) = \boldsymbol{A}(s)\vec{X}(s) + \vec{F}(s).$$

Since $\|\boldsymbol{A}(s)\| < 1$, $\det[\boldsymbol{I} - \boldsymbol{A}(s)] \ne 0$, the vector $\vec{X}(s)$ becomes

$$\vec{X}(s) = [\boldsymbol{I} - \boldsymbol{A}(s)]^{-1}\vec{F}(s).$$

From the assumption of $\vec{f}(t)$,

$$\lim_{|s|\to\infty}\left|sX_i(s)\right| = 0, \quad i = 1, 2, \cdots, N$$

for Re $(s) \ge 0$ and $\vec{X}(s)$ is analytic in Re $(s) \ge 0$ except $s = 0$. Using [8, theorem 4.13] $\vec{x}(t)$ is $\bigcirc$ (1). This requires that $\vec{x}(t)$ be bounded. By applying the final value theorem of Laplace transform theory, we obtain

$$\vec{x}(\infty) = \lim_{s\to 0} s\vec{X}(s) = \lim_{s\to 0}[\boldsymbol{I} - \boldsymbol{A}(s)]^{-1}[s\vec{F}(s)]$$
$$= [\boldsymbol{I} - \boldsymbol{A}(0)]^{-1}\vec{f} = [\boldsymbol{I} - \boldsymbol{A}]^{-1}\vec{f}$$

which is our desired result. Q.E.D.

Proof of Lemma 2

Let $\vec{y}$ be an N-dimensional vector and $\boldsymbol{A}(s) = (a_{ij}H_{ij}(s)\exp[-s\tau_{ij}])$ and $N \times N$ matrix. Then

$$\|\boldsymbol{A}(s)\vec{y}\|^2 = \vec{y}^*\boldsymbol{A}^*(s)\boldsymbol{A}(s)\vec{y}$$
$$= \sum_i\sum_j\sum_k y_i^* a_{ki}^* H_{ki}^*(s)\exp(-s^*\tau_{ki})a_{kj}H_{kj}(s)\cdot\exp(-s\tau_{kj})y_j$$
$$\le \sum_i\sum_j\sum_k |y_i|\cdot|a_{ki}H_{ki}(s)|\cdot|a_{kj}H_{kj}(s)|\cdot|y_j|$$

where the asterisks denote the complex conjugate for scalars and the complex conjugate transpose for matrices. Let $z_i = |y_i|$, $b_{ki} = |a_{ki}|$, and $\vec{z} = (z_i)$. Then

$$\|\boldsymbol{A}(s)\vec{y}\|^2 \le \sum_i\sum_j\sum_k z_i b_{ki}b_{kj}z_j = \|\boldsymbol{B}\vec{z}\|^2.$$

Since $\|\vec{y}\| = \|\vec{z}\|$, $\|\boldsymbol{A}(s)\| \le \|\boldsymbol{B}\| < 1$. From Lemma 1, we conclude that (3) converges to the desired result. Q.E.D.

Proof of Lemma 3

Let $\boldsymbol{A}_1(s) = [a_{ij}\exp(-s\tau_{ij})]$. From the assumption, we have

$$\boldsymbol{A}(s) = [a_{ij}\exp(-s\tau_{ij})H_{ij}(s)]$$
$$= [a_{ij}\exp(-s\tau_{ij})H_i^{(1)}(s)H_j^{(2)}(s)].$$

By letting

$$\boldsymbol{D}_1 = \operatorname{diag}\left[H_1^{(1)}(s), H_2^{(1)}(s), \cdots, H_N^{(1)}(s)\right]$$

$$\boldsymbol{D}_2 = \operatorname{diag}\left[H_1^{(2)}(s), H_2^{(2)}(s), \cdots, H_N^{(2)}(s)\right]$$

$\boldsymbol{A}(s)$ becomes $\boldsymbol{A}(s) = \boldsymbol{H}^{(1)}(s)\boldsymbol{A}_1(s)\boldsymbol{H}^{(2)}(s)$. Since $\|\boldsymbol{H}^{(1)}(s)\| \le 1$ and $\|\boldsymbol{H}^{(2)}(s)\| \le 1$ for $\operatorname{Re}(s) \ge 0$, we have

$$\|\boldsymbol{A}(s)\| \le \|\boldsymbol{H}^{(1)}(s)\| \cdot \|\boldsymbol{A}_1(s)\| \cdot \|\boldsymbol{H}^{(2)}(s)\| \le \|\boldsymbol{A}_1(s)\| < 1.$$

From Lemma 1, we conclude that (3) converges to the desired result. Q.E.D.

Proof of Theorem 1

Rewrite (2) as

$$x_i(t) = \eta_i[\, y_i(t) + f_i(t)] + g_i(t)$$

where

$$y_i(t) = \sum_{j=1}^{N} a_{ij} \int_0^t h_{ij}(t - \xi) x_j(\xi - \tau_{ij})\, d\xi.$$

Let $\vec{y}(t)$ denote the vector containing the $y_i(t)$'s. Evaluate the norm of $\vec{x}(t_1) - \vec{x}(t_2)$ as

$$\begin{aligned}\|\vec{x}(t_1) - \vec{x}(t_2)\| &\le \|\boldsymbol{\eta}[\,\vec{y}(t_1) + \vec{f}(t_1)] - \boldsymbol{\eta}[\,\vec{y}(t_2) + \vec{f}(t_2)]\| \\ &\quad + \|\vec{g}(t_1) - \vec{g}(t_2)\| \\ &\le \|\vec{y}(t_1) - \vec{y}(t_2)\| + p(t_1, t_2)\end{aligned}$$

or

$$\|\vec{x}(t_1, t_2)\| \le \|\vec{y}(t_1, t_2)\| + p(t_1, t_2)$$

where

$$p(t_1, t_2) = \|\vec{f}(t_1) - \vec{f}(t_2)\| + \|\vec{g}(t_1) - \vec{g}(t_2)\|$$

$$\vec{x}(t_1, t_2) = \vec{x}(t_1) - \vec{x}(t_2).$$

The ith component of $\vec{y}(t_1, t_2)$ is

$$y_i(t_1, t_2) = \sum_{j=1}^{N} a_{ij}\left[\int_0^{t_1} h_{ij}(t_1 - \xi) x_i(\xi - \tau_{ij})\, d\xi - \int_0^{t_2} h_{ij}(t_2 - \xi) x_j(\xi - \tau_{ij})\, d\xi\right].$$

Substitute $\vec{Q}$ for $\vec{x}$ in the linear dispersive skewed iteration in (3) and take the difference between (3) at t_1 and t_2 to obtain

$$\vec{Q}(t_1, t_2) = \vec{z}(t_1, t_2) + \vec{f}(t_1, t_2)$$

where

$$\vec{Q}(t_1, t_2) = \vec{Q}(t_1) - \vec{Q}(t_2)$$

$$\vec{f}(t_1, t_2) = \vec{f}(t_1) - \vec{f}(t_2)$$

and

$$\vec{z}(t_1, t_2) = \sum_{j=1}^{N} a_{ij}\left[\int_0^{t_1} h_{ij}(t_1 - \xi) Q_j(\xi - \tau_{ij})\, d\xi - \int_0^{t_2} h_{ij}(t_2 - \xi) Q_j(\xi - \tau_{ij})\, d\xi\right].$$

We now construct a function $\vec{f}(t_1, t_2)$ to be collinear with $\vec{z}(t_1, t_2)$. Let

$$\vec{f}(t_1, t_2) = \begin{cases} \text{any vector} & \text{for } \vec{z}(t_1, t_2) = 0 \\ C(t_1, t_2)\vec{z}(t_1, t_2) & \text{otherwise} \end{cases}$$

where the proportionality constant $C(t_1, t_2)$ is chosen so that $\|\vec{f}(t_1, t_2)\| = p(t_1, t_2)$. Then, by construction,

$$\|\vec{Q}(t_1, t_2)\| = \|\vec{z}(t_1, t_2)\| + p(t_1, t_2).$$

Because $\|\vec{Q}(t_1, t_2)\|$ converges to 0 by assumption, (2) also converges:

$$\|\vec{x}(t_1, t_2)\| \to 0 \qquad \text{for } t_1 \to \infty,\ t_2 \to \infty$$

or, equivalently, since our vectors are in a finite dimensional space (compact),

$$\lim_{t \to \infty} \vec{x}(t) = \vec{x}(\infty)$$

and our proof is complete. Q.E.D.

References

[1] J. Shamir, "Fundamental speed limitation on parallel processing," *Appl. Opt.*, vol. 9, no. 26, p. 1567, 1987.
[2] S. Oh, D. C. Park, R. J. Marks II, and L. E. Atlas, "Nondispersive propagation skew in iterative neural networks and optical feedback processors," *Opt. Engineering*, vol. 28, pp. 526-532, 1989.
[3] J. J. Hopfield, "Neural networks and physical systems with emergent collective computational abilities," in *Proc. Nat. Acad. Sci. U.S.*, vol. 79, pp. 2554-2558, 1982.
[4] R. P. Lippman, "An introduction to computing neural nets," *IEEE ASSP Mag.*, pp. 7, 1987.
[5] J. J. Hopfield and D. W. Tank, "Neural computation of decisions in optimization problem," *Biol. Cybern.*, vol. 52, pp. 141-152, 1985.
[6] R. J. Marks II, "A class of continuous level associative memory neural nets," *Appl. Opt.*, vol. 26, no. 10, pp. 2005-2010, 1987.
[7] R. J. Marks II, S. Oh, and L. E. Atlas, "Alternating projection neural networks," *IEEE Trans. Circuits Syst.*, vol. 36, pp. 846-857, June 1989.
[8] L. C. Andrews and B. K. Shivamoggi, *Integral Transforms for Engineers and Applied Mathematicians*. New York: Macmillan, 1986.

Using the Karhunen-Loe've Transformation in the Back-Propagation Training Algorithm

H. A. Malki and A. Moghaddamjoo

***Abstract*—A new training approach based on the back-propagation algorithm is introduced. In the proposed approach, initially, a set of training vectors is obtained by applying the Karhunen-Loe've (K-L) transform on the training patterns. The training is first started in the direction of the major eigenvectors of the correlation matrix of the training patterns and then continues by gradually including the remaining components, in their order of significance. With this approach, the number of computations is significantly reduced and the learning rate is improved. The performance of this method is compared with the standard back-propagation algorithm in segmenting a synthetic noisy image.**

I. Introduction

We confine ourselves to the back-propagation algorithm [1] and modifications which improve its slow rate of convergence. The back-propagation algorithm is one of the most popular training algorithms; it has been applied extensively and shown very good performance. In practice, however, the algorithm encounters two main difficulties: (1) its rate of convergence is very slow and (2) it does

Manuscript received June 18, 1990.
The authors are with the Department of Electrical Engineering and Computer Science, University of Wisconsin-Milwaukee, P.O. Box 784, Milwaukee, WI 53201.
IEEE Log Number 9040688.