JOURNAL OF MATHEMATICAL PSYCHOLOGY 30, 73–80 (1986)

The "Brain-State-in-a-Box" Neural Model Is a Gradient Descent Algorithm

RICHARD M. GOLDEN

Brown University

The Brain-State-in-a-Box (BSB) neural model (J. A. Anderson, J. W. Silverstein, S. A. Ritz, & R. S. Jones, 1977, *Psychological Review*, **84**, 413–451) is a pattern categorization device inspired by neurophysiological considerations. This model has additionally been applied to a fairly diverse range of psychological phenomena. In this paper, the BSB model is demonstrated to be a deterministic constrained gradient descent algorithm that minimizes a quadratic cost function. A formal proof that all trajectories of the BSB algorithm in state vector space approach the set of system equilibrium points, under certain specific conditions, is presented. Some conditions regarding the existence of global energy minima are also briefly discussed.

THE BSB MODEL

The Brain-State-in-a-Box (BSB) model (Anderson, 1983; Anderson, Silverstein, Ritz, & Jones, 1977) is an abstract nervous system model that has been used to study various psychological behaviors. The model was inspired by the following physical situation. Consider a group of N neurons which are highly interconnected and feed back upon themselves. Information in this system is represented by an N-dimensional state vector representing the pattern of neural firing frequencies across this neuronal set. The system operates by accepting a pattern of neural activity and amplifying that neural activation pattern using positive feedback. Eventually all the neurons in the system will saturate. Thus, although the initial activation pattern is analog since each neuron's firing rate is a continuous real valued variable, the final activation pattern is digital since each neuron has saturated. Psychologically, the model is a categorization mechanism. In this paper the BSB model will be considered from a purely formal perspective. The model has been applied, however, to a wide range of psychological phenomena, including problems in categorical perception and probability learning (Anderson *et al.*, 1977), the development of visual

This research was funded in part by Grant BNS-82-14728 from the National Science Foundation, Memory and Cognitive Processes Section to J. A. Anderson. I thank Geoffrey Hinton for suggesting that mathematical programming methods might be useful for the analysis of neural models. I am also grateful for comments made by Stuart Geman, Ralph Golden, and two anonymous reviewers on an earlier draft of this paper. One anonymous reviewer's comments, in particular, significantly improved both the proof of Proposition 1 and the overall quality of this paper. I am also grateful to Jim Anderson for his generous support and encouragement. Requests for reprints should be sent to Richard M. Golden, Brown University, Psychology Department, Box 1853, Providence, RI 02912.

0022-2496/86 $3.00

word perception (Golden, 1985), multistable perception (Kawamoto & Anderson, 1985), and lexical ambiguity (Kawamoto, 1985).

The BSB model is defined mathematically as follows. We have an initial state vector $\mathbf{X}_0$ and a symmetric matrix $\mathbf{A}$. The eigenvalues of the $\mathbf{A}$ matrix of largest magnitude possess positive real components. The initial state vector represents some initial neural activation pattern. The ith component of the vector $\mathbf{X}_0$, $\mathbf{X}_0(i)$, represents the initial firing rate of the ith neuron in the system. If there are N neurons in the system, then the system state vector has dimension N. The direction of the vector represents the pattern's identity. The magnitude of the vector represents the relative strength of the pattern. The algorithm is then completely specified by the following equations:

$$\mathbf{Y}_k = \mathbf{X}_k + \gamma \mathbf{A}\mathbf{X}_k \tag{1}$$

$$\mathbf{X}_{k+1} = S(\mathbf{Y}_k) \tag{2}$$

where γ is a small positive scalar constant, and $\mathbf{X}_k(i)$ is the firing rate of the ith neuron in the pattern of neural activity at time k. The S function, a piecewise linear sigmoidal function, operates upon the ith vector element $\mathbf{Y}_k(i)$ in the following manner:

$$\mathbf{X}_{k+1}(i) = S(\mathbf{Y}_k(i)) = \begin{cases} +F & \mathbf{Y}_k(i) > +F \\ \mathbf{Y}_k(i) & |\mathbf{Y}_k(i)| \leqslant +F \\ -F & \mathbf{Y}_k(i) < -F \end{cases} \tag{3}$$

where F is the maximum magnitude of the firing rate of any neuron in the system. Geometrically, Eq. (3) constrains the system state vector to lie within an origin-centered hypercube.

The procedure for applying Eqs. (1) and (2) may now be described. An initial pattern of neural activity $\mathbf{X}_0$ is first inserted in Eq. (1) to obtain the vector $\mathbf{Y}_0$. Each element of the vector $\mathbf{Y}_0$ is then truncated according to Eq. (2) to obtain the updated vector $\mathbf{X}_1$. The updated vector $\mathbf{X}_1$ is then cycled through Eqs. (1) and (2) to obtain $\mathbf{X}_2$. Each cycle through both Eqs. (1) and (2) is defined as one iteration. When the state vector $\mathbf{X}_{k+1}$ equals $\mathbf{X}_k$, then the system is said to have "categorized" the vector $\mathbf{X}_0$ as the category vector $\mathbf{X}_k$. When the category vector $\mathbf{X}_k$ is a vector possessing only saturated elements, then the category vector will be referred to as a "corner" vector. The number of iterations required to "categorize" a vector will be referred to as the reaction time of the system.

Hopfield's Analysis

A Liapunov cost function for a particular dynamic system is a scalar-valued function with the property that, as the state of the system evolves, the value of the

cost function decreases. Thus, the equilibrium points of the dynamic system correspond to the minima, maxima, and saddlepoints of the cost function. Hopfield (1984) found a family of cost functions for a large class of neural network models that are essentially continuous time versions of the BSB model. Each neuron in this large class could be represented as a linear integrator followed by a fairly general continuous sigmoidal nonlinearity.

Unfortunately, however, Hopfield's analysis is not directly applicable to the BSB model for several reasons. First, Hopfield's derivation for the continuous time system cannot be directly applied to the discrete-time system since the step size in the discrete-time system is usually large. Second, Eq. (2) of the BSB model equations is a sigmoidal nonlinearity that is piecewise linear. Hopfield's analysis requires the existence of the derivative of the inverse of the neural model's sigmoidal function. In the BSB model this condition is not satisfied when the ith element of the system state vector is either $+F$ or $-F$. And finally, the cost function proposed for this class of neural network models was fairly complex since Hopfield considered a large class of deterministic neural models. A less complex cost function may be used if one takes advantage of the BSB model's simplicity.

A Liapunov Function for the BSB Model

In this paper, the BSB neural model is demonstrated to be a gradient descent algorithm that minimizes the following Liapunov (or energy) function:

$$E(\mathbf{X}) = -(1/2)\,\mathbf{X}^T\mathbf{A}\mathbf{X}. \tag{4}$$

Also note that if $\mathbf{A}$ is symmetric, then the gradient (Duda & Hart, 1973, p. 47) of $E(\mathbf{X}_k)$ is easily calculated to be:

$$\mathbf{g}_k = \nabla E(\mathbf{X}_k) = -\mathbf{A}\mathbf{X}_k. \tag{5}$$

Lemma 1. *Let* $\mathbf{A}$ *be a symmetric matrix. Then*

$$\lambda_{\min} |\mathbf{X}|^2 \leqslant \mathbf{X}^T\mathbf{A}\mathbf{X} \leqslant \lambda_{\max} |\mathbf{X}|^2$$

where $\lambda_{\min}$ *is the minimum eigenvalue of* $\mathbf{A}$, *and* $\lambda_{\max}$ *is the maximum eigenvalue of* $\mathbf{A}$.

Proof. Since $\mathbf{A}$ is symmetric, $\mathbf{A}$ can be diagonalized. That is, there is an orthonormal basis $\{\mathbf{e}_i\}$ with: $\mathbf{A}\mathbf{e}_i = \lambda_i\mathbf{e}_i$. Thus,

$$\mathbf{X}^T\mathbf{A}\mathbf{X} = \sum_i \lambda_i[\mathbf{Y}(i)]^2, \qquad \text{where } \mathbf{Y}(i) = \mathbf{e}_i^T\mathbf{X}.$$

Also note $\sum_i [\mathbf{Y}(i)]^2 = |\mathbf{Y}|^2 = |\mathbf{X}|^2$ since the basis is orthonormal. Therefore,

$$\lambda_{\min} |\mathbf{X}|^2 \leqslant \mathbf{X}^T\mathbf{A}\mathbf{X} = \sum \lambda_i[\mathbf{Y}(i)]^2 \leqslant \lambda_{\max} |\mathbf{X}|^2. \qquad \text{Q.E.D.}$$

LEMMA 2. *If* $\mathbf{g}_k(i) \neq 0$, *then* $0 \leqslant \alpha(i, k) \leqslant 1$ *where*

$$\alpha(i, k) = [S(\mathbf{Y}_k(i)) - \mathbf{X}_k(i)]/[-\gamma \mathbf{g}_k(i)]$$

and $S(\mathbf{Y}_k(i))$, $\mathbf{X}_k(i)$, $\mathbf{g}_k(i)$, *and* γ *are defined in Eqs.* (1), (2), (3), *and* (5).

Proof. Note that for $\mathbf{g}_k(i) \neq 0$, Eqs. (1) and (2) may be rewritten as:

$$\mathbf{X}_{k+1}(i) = \mathbf{X}_k(i) - \gamma \alpha(i, k)\, \mathbf{g}_k(i).$$

Case 1: If $|\mathbf{X}_k(i) - \gamma \mathbf{g}_k(i)| \leqslant F$, then $\alpha(i, k) = 1$.

Case 2: If $\mathbf{X}_k(i) - \gamma \mathbf{g}_k(i) > F$, then $\mathbf{X}_{k+1}(i) = F$ and $-\gamma \mathbf{g}_k(i) > 0$. Thus, $1 > \alpha(i, k) = [F - \mathbf{X}_k(i)]/[-\gamma \mathbf{g}_k(i)] \geqslant 0$.

Case 3: If $\mathbf{X}_k(i) - \gamma \mathbf{g}_k(i) < -F$, then $\mathbf{X}_{k+1}(i) = -F$ and $-\gamma \mathbf{g}_k(i) < 0$. Thus, $1 > \alpha(i, k) = [-F - \mathbf{X}_k(i)]/[-\gamma \mathbf{g}_k(i)] \geqslant 0$. Q.E.D.

LEMMA 3. *Assume* $\mathbf{d}_k = \mathbf{X}_{k+1} - \mathbf{X}_k \neq \mathbf{0}$ *where* $\mathbf{X}_{k+1}$ *and* $\mathbf{X}_k$ *are defined as in* (1) *and* (2), *and* $\mathbf{0}$ *is a vector of zeros. If* (*a*) $\mathbf{A}$ *is symmetric, and* (*b*) *either* $\mathbf{A}$ *is positive semidefinite or* $\gamma < 2/|\lambda_{\min}|$ *where* $\lambda_{\min}$ *is the minimum eigenvalue of* $\mathbf{A}$, *then* $\mathbf{d}_k^T \mathbf{g}_k < (1/2)\, \mathbf{d}_k^T \mathbf{A} \mathbf{d}_k$.

Proof. First note that since, for at least one value of i, $\mathbf{d}_k(i) = -\gamma \alpha(i, k)\, \mathbf{g}_k(i) \neq 0$,

$$\sum_i \alpha(i, k)[\mathbf{g}_k(i)]^2 \neq 0. \tag{6}$$

Now note that, from (6) and Lemma 2,

$$\mathbf{d}_k^T \mathbf{g}_k = -\sum_i \gamma \alpha(i, k)[\mathbf{g}_k(i)]^2 < 0. \tag{7}$$

Also, from Lemma 1, we have

$$(1/2)\, \lambda_{\min} |\mathbf{d}_k|^2 \leqslant (1/2)\, \mathbf{d}_k^T \mathbf{A} \mathbf{d}_k. \tag{8}$$

If $\mathbf{A}$ is positive semidefinite (i.e., $\lambda_{\min} \geqslant 0$), then (7) and (8) yield:

$$\mathbf{d}_k^T \mathbf{g}_k < 0 \leqslant (1/2)\, \mathbf{d}_k^T \mathbf{A} \mathbf{d}_k. \tag{9}$$

The case where $\lambda_{\min} < 0$ is now considered. From (7) and (8), we require the following inequality to be true:

$$-\sum_i \gamma \alpha(i, k)[\mathbf{g}_k(i)]^2 < (1/2)\, \lambda_{\min} \sum_i [-\gamma \alpha(i, k)\, \mathbf{g}_k(i)]^2$$

or

$$\gamma < 2 \sum_i \alpha(i, k)[\mathbf{g}_k(i)]^2 / |\lambda_{\min}| \sum_i [\alpha(i, k)\, \mathbf{g}_k(i)]^2.$$

Also, from Lemma 2, we have $[\alpha(i,k)]^2 \leqslant \alpha(i,k)$. Thus, the following bound is obtained:

$$\gamma < 2 \sum_i \alpha(i,k)[\mathbf{g}_k(i)]^2 / |\lambda_{\min}| \sum_i \alpha(i,k)[\mathbf{g}_k(i)]^2$$

or

$$\gamma < 2/|\lambda_{\min}|$$

since (6) permits the cancellation of the summations. Q.E.D.

With the aid of Lemmas 1–3, some basic concepts from system analysis will now be used to characterize the behavior of Eqs. (1) and (2). A dynamic system may be considered as some transformation that maps the current state vector, $\mathbf{X}_k$, into the succeeding state vector, $\mathbf{X}_{k+1}$. Moreover, such system state vectors are most easily represented as "points" in a high dimensional space. A "trajectory" of a discrete-time system in state vector space is the sequence of state vectors that represent the evolution of the system's state as time increases. Suppose now that a state of the system exists such that when the system is in that state, the system remains in that state forever as time increases. Such a state will be referred to as a "system equilibrium point." The possible trajectories of the BSB algorithm through state vector space are now formally considered.

BSB Energy Minimization Theorem. Let $\mathbf{X}_{k+1}$ and $\mathbf{X}_k$ be defined as in Eqs. (1) and (2). If conditions (a) and (b) of Lemma 3 are satisfied, then

(i) $E(\mathbf{X}_{k+1}) < E(\mathbf{X}_k)$ if $\mathbf{X}_{k+1} \neq \mathbf{X}_k$

(ii) $E(\mathbf{X}_{k+1}) = E(\mathbf{X}_k)$ if and only if $\mathbf{X}_{k+1} = \mathbf{X}_k$

where $E(\mathbf{X}_k)$ is defined in (4).

Proof. Assume $\mathbf{d}_k = \mathbf{X}_{k+1} - \mathbf{X}_k \neq \mathbf{0}$. Also let

$$\Delta = 2(E(\mathbf{X}_{k+1}) - E(\mathbf{X}_k)) = 2(E(\mathbf{X}_k + \mathbf{d}_k) - E(\mathbf{X}_k))$$

$$\Delta = -(\mathbf{X}_k + \mathbf{d}_k)^T \mathbf{A}(\mathbf{X}_k + \mathbf{d}_k) + \mathbf{X}_k^T \mathbf{A}\mathbf{X}_k$$

$$\Delta = -\mathbf{X}_k^T \mathbf{A}\mathbf{d}_k - \mathbf{d}_k^T \mathbf{A}\mathbf{X}_k - \mathbf{d}_k^T \mathbf{A}\mathbf{d}_k.$$

Now, by the symmetry of $\mathbf{A}$, we have

$$\Delta = -2\mathbf{d}_k^T \mathbf{A}\mathbf{X}_k - \mathbf{d}_k^T \mathbf{A}\mathbf{d}_k = 2\mathbf{d}_k^T \mathbf{g}_k - \mathbf{d}_k^T \mathbf{A}\mathbf{d}_k.$$

We require that $\Delta < 0$ or that:

$$\mathbf{d}_k^T \mathbf{g}_k < (1/2)\, \mathbf{d}_k^T \mathbf{A}\mathbf{d}_k. \tag{10}$$

By Lemma 3, Eq. (10) is satisfied, thus establishing condition (i). Condition (ii) follows immediately from condition (i). Q.E.D.

With the aid of the BSB Energy Minimization Theorem, some statements regarding the trajectory of any path of the system in state vector space can be made. More specifically, every trajectory of the system will be shown to approach the set of system equilibrium points as time increases. First, however, the Invariant Set Theorem must be introduced. Luenberger (1979, pp. 345–346) presents a proof and a discussion of this theorem. Vidyasagar (1978, pp. 156–157) proves this theorem for the continuous case in a more rigorous manner.

Definition of an invariant set. A set G is an invariant set for a dynamic system if whenever a point $\mathbf{X}$ on a system trajectory is in G, the trajectory remains in G. In particular, any set of equilibrium points is invariant.

Invariant Set Theorem. Consider some operator $\mathbf{\Gamma}(\ )$ such that $\mathbf{\Gamma}(\mathbf{X}_k) = \mathbf{X}_{k+1}$. Also let $V(\mathbf{X})$ be a scalar-valued function with continuous first partial derivatives. Now suppose that for some $s > 0$, the set $W_s = \{\mathbf{X}: V(\mathbf{X}) < s\}$ is bounded. Also assume that

$$V(\mathbf{\Gamma}(\mathbf{X}_k)) \leqslant V(\mathbf{X}_k) \qquad \forall \mathbf{X} \in W_s.$$

Let S denote the subset of W_s defined by $S = \{\mathbf{X} \in W_s: \mathbf{\Gamma}(\mathbf{X}_k) = \mathbf{X}_k\}$ and let G be the largest invariant set within S. Then every trajectory in W_s tends to G as time increases.

To apply the Invariant Set Theorem, identify V with E and W_s as the system hypercube. The system hypercube is a bounded region of vector space, and a value of the energy function can always be found such that all values of the energy function over the hypercube are less than that value. We have also demonstrated, using the BSB Energy Minimization Theorem, that if the system matrix is positive semidefinite or $\gamma < 2/|\lambda_{\min}|$ where $\lambda_{\min}$ is the most negative eigenvalue of A, then $E(\mathbf{X}_{k+1}) \leqslant E(\mathbf{X}_k)$. Finally note that G is simply the entire set of system equilibrium points from (ii) of the BSB Energy Minimization Theorem. These conclusions immediately lead to the following key theorem.

BSB Asymptotic Stability Theorem. Consider the system of equations defined in Eqs. (1) and (2). If (a) $\mathbf{A}$ is symmetric, and (b) either $\mathbf{A}$ is positive semidefinite or $\gamma < 2/|\lambda_{\min}|$ where $\lambda_{\min}$ is the minimum eigenvalue of $\mathbf{A}$, then every trajectory of the system tends toward the set of system equilibrium points as time increases.

Geometric Interpretation

If all the eigenvalues of the $\mathbf{A}$ matrix are positive, then the set of vectors $\mathbf{X}$ such that $E(\mathbf{X}) = -c$, for some $c > 0$, forms a hyperellipsoid centered at the origin. That is, the contours of constant energy are hyperellipsoids. These hyperellipsoids are oriented in the same direction. Thus, the state vector is travelling down a unimodal function whose peak is at the origin. The directions of the principal axes of the

hyperellipsoids are indicated by the eigenvectors of the **A** matrix. The specific length of a given principal axis of each hyperellipsoid is inversely proportional to the square root of the appropriate eigenvalue of the matrix.

Some Comments Regarding Global Energy Minima

In typical simulations of the BSB model, the Widrow–Hoff learning rule (Duda & Hart, 1973, pp. 155–159) is used to adjust the synaptic weights of the **A** matrix using a set of learning stimuli that are vertices of the system hypercube. This learning rule is simply another gradient descent algorithm that modifies the **A** matrix's weights, and is connected with some interesting psychological and neurophysiological data (see Sutton & Barto, 1981 for a review). In general, the number of stimuli learned by the system is fairly small relative to the system's dimensionality. A point of particular interest, however, is that this learning rule biases the system matrix such that the hypercube vertices learned by the system become eigenvectors, with large positive eigenvalues, of the matrix. The following two propositions represent an attempt to provide some rationale for constructing the system matrix for the BSB model in this manner.

Proposition 1. *Consider the energy function defined in Eq.* (4). *Assume the* **A** *matrix in* (4) *is symmetric, and let* $\lambda_{\max}$ *be the maximum eigenvalue of* **A**. *If* (a) **C** *is a hypercube vertex, and* (b) $\mathbf{AC} = \lambda_{\max}\mathbf{C}$, *then* **C** *and* $-\mathbf{C}$ *are global minima of the energy function over the system hypercube defined by Eq.* (3).

Proof. From Lemma 1,

$$E(\mathbf{X}) = -(1/2)\,\mathbf{X}^T\mathbf{A}\mathbf{X} \geqslant -(\lambda_{\max}/2)\,|\mathbf{X}|^2.$$

Also, $|\mathbf{X}|^2 \leqslant |\mathbf{C}|^2$ where $\mathbf{X} \in$ hypercube. Thus,

$$E(\mathbf{X}) \geqslant (-\lambda_{\max}/2)\,|\mathbf{X}|^2 \geqslant (-\lambda_{\max}/2)\,|\mathbf{C}|^2 = -(1/2)\,\mathbf{C}^T\mathbf{A}\mathbf{C}$$

since $\mathbf{AC} = \lambda_{\max}\mathbf{C}$. Also note $E(\mathbf{C}) = E(-\mathbf{C})$. Q.E.D.

Proposition 2. *Consider the energy function defined in Eq.* (4), *and the hypercube defined by Eq.* (3). *If the matrix is positive semidefinite, and* **C** *is a global minimum of the energy function over the hypercube, then* **C** *is a hypercube vertex.*

Proof. A classical result of the multivariate theory of convex functions states that if f is a convex function defined on a bounded, closed convex set W that has a maximum over W, that maximum is obtained at an extreme point of W (Luenberger, 1984, p. 181). If the **A** matrix is positive semidefinite, then the energy function is a concave function since the Hessian matrix of the energy function is simply $-\mathbf{A}$ (Luenberger, 1984, p. 180). (Note that the statement that the Hessian matrix of the energy function is positive semidefinite is the multivariate equivalent of the statement that the second derivative of a scalar-valued function of a single

variable is always nonnegative.) Also, the hypercube is a convex set whose vertices are extreme points. Now, since the maximization of a convex function over a convex set is equivalent to the minimization of a concave function over a convex set, a global minimum must occur at a hypercube vertex. Q.E.D.

REFERENCES

ANDERSON, J. A. (1983). Cognitive and psychological computation with neural models. *IEEE Transactions on Systems, Man, and Cybernetics*, **SMC-13**, 799–815.

ANDERSON, J. A., SILVERSTEIN, J. W., RITZ, S. A., & JONES, R. S. (1977). Distinctive features, categorical perception, and probability learning: Some applications of a neural model. *Psychological Review*, **84**, 413–451.

DUDA, R. O., & HART, P. E. (1973). *Pattern classification and scene analysis*. New York: Wiley.

GOLDEN, R. M. (1985). A developmental neural model of word perception. *Proceedings of the Seventh Annual Conference of the Cognitive Science Society*, Irvine, CA.

HOPFIELD, J. J. (1984). Neurons with graded response have collective properties like those of two-state neurons. *Proceedings of the National Academy of Sciences, USA*, **81**, 3088–3092.

HUMMEL, R. A., & ZUCKER, S. W. (1983). On the foundations of relaxation labeling processes. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **5**, 267–287.

KAWAMOTO, A. (1985). *Dynamic processes in the (re)solution of lexical ambiguity*. Unpublished doctoral dissertation, Brown University, Providence, RI.

KAWAMOTO, A., & ANDERSON, J. A. (1985). A neural network model of multistable perception. *Acta Psychologica*, **59**, 35–65.

LUENBERGER, D. G. (1979). *Introduction to dynamic systems: Theory, models, and applications*. New York: Wiley.

LUENBERGER, D. G. (1984). *Linear and nonlinear programming*. London: Addison–Wesley.

SUTTON, R. S., & BARTO, A. G. (1981). Toward a modern theory of adaptive networks: Expectation and prediction. *Psychological Review*, **88**, 135–170.

VIDYASAGAR, M. (1978). *Nonlinear systems analysis*. Englewood Cliffs, NJ: Prentice–Hall.

RECEIVED: February 20, 1985