Proceedings of the 1991 International Joint conference on Neural Networks,
IJCNN-91, Seattle, Washington, July 8-12, 1991

The Kanerva Memory Is Stable

Tzi-Dar Chiueh* and Rodney M. Goodman**

*Department of Electrical Engineering, National Taiwan University, Taipei, Taiwan 10764.
**Department of Electrical Engineering, 116-81, Caltech, Pasadena, CA 91125, USA.

ABSTRACT

The Kanerva memory is a simple, yet important model of the cerebellar cortex. Its power has been demonstrated by its huge storage capacity as an associative memory. In this paper, the Kanerva memory is briefly introduced; then the Kanerva memory is shown to be asymptotically stable in both the parallel update and sequential update modes.

I. Kanerva Memory

Kanerva proposed an associative memory model that is essentially a parallel processing system consisting of address decoders, counters, binary adders, and threshold circuits [1,2]. He claimed that his model is similar to the human cerebellar model of Marr's [3] and Albus' [4]. This model is significant because of its high storage capacity that reaches the theoretical limit for artificial associative memories [5]. One can formulate the Kanerva memory in terms of a two-layer feed-forward neural network. There are N input nodes, L hidden neurons, and N output neurons. As suggested by Kanerva in [1], the Kanerva memory can act as an autoassociative memory if a feedback connection between the input and the output ends is incorporated, as shown in Figure 1.

All neurons in the Kanerva memory are hard-limiter neurons. The hidden neurons have a Heaviside step-function response

$$H(t) \equiv \begin{cases} 1 & t \geq s \\ 0 & t < s \end{cases},$$

while the output neurons have a *sgn*-function response

$$sgn(t) \equiv \begin{cases} +1 & t \geq 0 \\ -1 & t < 0 \end{cases}.$$

The first connection matrix $\mathbf{T}$ is an $L \times N$ matrix randomly populated with $+1$ and -1 with equal probability. Assume that $\mathbf{u}^{(k)}, k = 1,2,\ldots,M$ are the M N-bit bipolar ($+1$ or -1) memory patterns; then the corresponding M L-bit binary (1 or 0) internal representation codewords, $\mathbf{v}^{(k)}, k = 1,2,\ldots,M$ are defined as

$$\mathbf{v}^{(k)} \equiv H\left(\mathbf{T}\,\mathbf{u}^{(k)}\right), \qquad k = 1,2,\ldots,M. \tag{1}$$

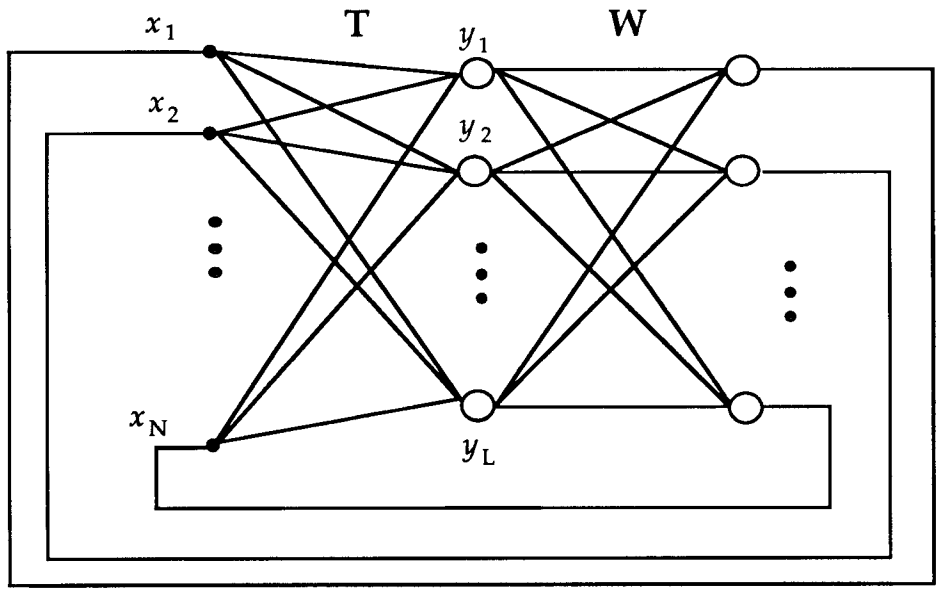

Figure 1: Configuration of the Kanerva memory with feedback.

The second connection matrix $\mathbf{W}$ is constructed by the sum of the outer products of the M memory patterns and the M internal representation codewords; i. e.,

$$\mathbf{W} = \sum_{k=1}^{M} \mathbf{u}^{(k)} \mathbf{v}^{(k)^t}. \tag{2}$$

The motion equation of the Kanerva memory with feedback takes the form of

$$\mathbf{x}' = sgn\{\mathbf{W}\,\mathbf{y}\} = sgn\{\mathbf{W} \cdot H(\mathbf{T}\,\mathbf{x})\}, \tag{3}$$

where $\mathbf{x}$ and $\mathbf{x}'$ are the current and the next state patterns. Substituting Equation (2) in (3) yields

$$\begin{aligned} \mathbf{x}' &= sgn\left\{\sum_{k=1}^{M} \left\langle \mathbf{v}^{(k)}, \mathbf{y} \right\rangle \mathbf{u}^{(k)}\right\} \\ &= sgn\left\{\sum_{k=1}^{M} \left\langle H\left(\mathbf{T}\,\mathbf{u}^{(k)}\right), H(\mathbf{T}\,\mathbf{x}) \right\rangle \mathbf{u}^{(k)}\right\}. \end{aligned} \tag{4}$$

The hardware complexity needed for the Kanerva memory is $L+N$ hard-limiter neurons, LN binary ($+1$ or -1) connection weights (matrix $\mathbf{T}$), and LN discrete ($-M$ to M) connection weights (matrix $\mathbf{W}$). The storage capacity of the Kanerva memory with zero attraction radius was shown by Chou [5] to be as large as $M = 2^{\alpha N}$, where α is a parameter depending on s (the threshold in the hidden neurons) and L. It is also shown that M can be no more than L — the number of hidden neurons. This implies that in order to get exponential capacity, exponential hardware complexity is necessary. To be more specific,

$$LN + LN\log_2(2M) \geq 2^{\alpha N}(\alpha N^2 + 2N)$$

bits are needed to store $2^{\alpha N}$ N-bit memory patterns.

II. Proof of Stability

There has been virtually no treatment on the stability issue of the Kanerva memory since it is originally proposed as a feed-forward network. In this section, we can see that under a certain condition the Kanerva memory is asymptotically stable in both the synchronous (parallel) update and the asynchronous (sequential) update modes. The method used to prove the asymptotic stability of the Kanerva memory is as follow: A Liapunov (energy) function is defined and it is shown that after each timestep the Liapunov function will decrease or stay the same. This fact, together with the boundedness of the Liapunov function and that the Liapunov function cannot stay at the same level forever, proves that the Kanerva memory is asymptotically stable.

Assumption : Let $\mathbf{w}$ and $\mathbf{x}$ be two bipolar state patterns. Then

$$\langle H(\mathbf{T}\,\mathbf{w}),\, H(\mathbf{T}\,\mathbf{x})\rangle \;=\; f(d(\mathbf{w},\mathbf{x})),$$

where $f(\cdot)$ is a strictly nonincreasing function and $d(\mathbf{w},\mathbf{x})$ is the Hamming distance between $\mathbf{w}$ and $\mathbf{x}$.

Note that the original interpretation of $\langle H(\mathbf{T}\,\mathbf{w}),\, H(\mathbf{T}\,\mathbf{x})\rangle$ given by Kanerva is the number of rows in $\mathbf{T}$ (here each row can be looked upon as a vertex in the N-dimensional hypercube) that are no more than $(N-s)/2$ bits away from both $\mathbf{w}$ and $\mathbf{x}$. Recall that the matrix $\mathbf{T}$ is an L-vertex sample of the N-dimensional hypercube. Therefore, if L is equal to 2^N, then $\langle H(\mathbf{T}\,\mathbf{w}),\, H(\mathbf{T}\,\mathbf{x})\rangle$ is the number of vertices inside the intersection of two N-dimensional hyperspheres both with radius $(N-s)/2$. One of the hypersphere is centered at $\mathbf{w}$ and the other at $\mathbf{x}$. In this case. it is obvious that $\langle H(\mathbf{T}\,\mathbf{w}),\, H(\mathbf{T}\,\mathbf{x})\rangle$ depends monotonically on $d(\mathbf{w},\mathbf{x})$, and equivalently on $\langle \mathbf{w},\, \mathbf{x}\rangle$. When L is less than 2^N, the assumption does not hold forever. Nonetheless if the rows in $\mathbf{T}$ is distributed among the N-dimensional hypercube in a uniform fashion, the above assumption should be a reasonable one.

At first, let us give some definitions that will be used in the proof. For each memory pattern $\mathbf{u}^{(k)}$, define a sequence of patterns that are vertices in the path from $\mathbf{u}^{(k)}$ to $-\mathbf{u}^{(k)}$, where the path traverses the hypercube in an orderly fashion — flipping from the leftmost bit to the rightmost bit. We denote these patterns $\mathbf{z}_0^{(k)} \equiv \mathbf{u}^{(k)}, \mathbf{z}_1^{(k)}\, \mathbf{z}_2^{(k)}, \cdots, \mathbf{z}_N^{(k)} \equiv -\mathbf{u}^{(k)}$. For example, if $N = 4$ and $\mathbf{u}^{(k)} = (-1\,+1\,+1\,-1)$, then

$$\begin{aligned}
\mathbf{z}_0^{(k)} &= (-1\,+1\,+1\,-1)\\
\mathbf{z}_1^{(k)} &= (+1\,+1\,+1\,-1)\\
\mathbf{z}_2^{(k)} &= (+1\,-1\,+1\,-1)\\
\mathbf{z}_3^{(k)} &= (+1\,-1\,-1\,-1)\\
\mathbf{z}_4^{(k)} &= (+1\,-1\,-1\,+1)
\end{aligned}$$

By the assumption,

$$\left\langle H\left(\mathbf{T}\,\mathbf{u}^{(k)}\right),\, H(\mathbf{T}\,\mathbf{x})\right\rangle \;=\; \left\langle H\left(\mathbf{T}\,\mathbf{u}^{(k)}\right),\, H\left(\mathbf{T}\,\mathbf{z}_r^{(k)}\right)\right\rangle$$

if $d(\mathbf{u}^{(k)},\mathbf{x}) = d(\mathbf{u}^{(k)},\mathbf{z}_r^{(k)}) = r$. We now define the energy of state pattern $\mathbf{x}$ associated with the k^{th} memory pattern, $\mathbf{u}^{(k)}$ as

$$E^{(k)}(\mathbf{x}) \;\equiv\; \sum_{i=0}^{d(\mathbf{u}^{(k)},\mathbf{x})} \left\langle H\left(\mathbf{T}\,\mathbf{u}^{(k)}\right),\, H\left(\mathbf{T}\,\mathbf{z}_i^{(k)}\right)\right\rangle \tag{5}$$

The Liapunov function of the system at state $\mathbf{x}$ is then given by

$$E(\mathbf{x}) \equiv \sum_{k=1}^{M} E^{(k)}(\mathbf{x}).$$

Suppose all neurons in the output layer update themselves according to Equation (4); then the difference in the Liapunov functions between the current and the next states is

$$\begin{aligned} \Delta E &= E(\mathbf{x}') - E(\mathbf{x}) \\ &= \sum_{k=1}^{M} \left\{ E^{(k)}(\mathbf{x}') - E^{(k)}(\mathbf{x}) \right\}. \end{aligned} \tag{6}$$

Let $d(\mathbf{u}^{(k)}, \mathbf{x}) = d_k$ and $d(\mathbf{u}^{(k)}, \mathbf{x}') = d'_k$. Now, considering only the k^{th} term of the sum in Equation (6) yields

(a) if $d_k \geq d'_k$:

$$\begin{aligned} E^{(k)}(\mathbf{x}') - E^{(k)}(\mathbf{x}) &= -\sum_{i=d'_k+1}^{d_k} \left\langle H\left(\mathbf{T}\,\mathbf{u}^{(k)}\right), H\left(\mathbf{T}\,\mathbf{z}_i^{(k)}\right) \right\rangle \\ &\leq -\left\langle H\left(\mathbf{T}\,\mathbf{u}^{(k)}\right), H\left(\mathbf{T}\,\mathbf{z}_{d_k}^{(k)}\right) \right\rangle \cdot (d_k - d'_k) \\ &= \left\langle H\left(\mathbf{T}\,\mathbf{u}^{(k)}\right), H(\mathbf{T}\,\mathbf{x}) \right\rangle \cdot (d'_k - d_k). \end{aligned}$$

(b) if $d_k < d'_k$:

$$\begin{aligned} E^{(k)}(\mathbf{x}') - E^{(k)}(\mathbf{x}) &= \sum_{i=d_k+1}^{d'_k} \left\langle H\left(\mathbf{T}\,\mathbf{u}^{(k)}\right), H\left(\mathbf{T}\,\mathbf{z}_i^{(k)}\right) \right\rangle \\ &\leq \left\langle H\left(\mathbf{T}\,\mathbf{u}^{(k)}\right), H\left(\mathbf{T}\,\mathbf{z}_{d_k}^{(k)}\right) \right\rangle \cdot (d'_k - d_k) \\ &= \left\langle H\left(\mathbf{T}\,\mathbf{u}^{(k)}\right), H(\mathbf{T}\,\mathbf{x}) \right\rangle \cdot (d'_k - d_k). \end{aligned}$$

The two inequalities comes about because of the monotonicity of the function $f(\cdot)$ mentioned in the assumption. The last equalities in both cases are because $d(\mathbf{u}^{(k)}, \mathbf{x}) = d_k$. In both cases, the same inequality holds. Substituting the above inequality in Equation (6) yields

$$\begin{aligned} \Delta E &\leq \sum_{k=1}^{M} \left\langle H\left(\mathbf{T}\,\mathbf{u}^{(k)}\right), H(\mathbf{T}\,\mathbf{x}) \right\rangle \cdot (d'_k - d_k) \\ &= \frac{1}{2} \sum_{k=1}^{M} \left\langle H\left(\mathbf{T}\,\mathbf{u}^{(k)}\right), H(\mathbf{T}\,\mathbf{x}) \right\rangle \cdot \left(\left\langle \mathbf{u}^{(k)}, \mathbf{x} \right\rangle - \left\langle \mathbf{u}^{(k)}, \mathbf{x}' \right\rangle \right) \\ &= \frac{1}{2} \left\{ \sum_{k=1}^{M} \left\langle H\left(\mathbf{T}\,\mathbf{u}^{(k)}\right), H(\mathbf{T}\,\mathbf{x}) \right\rangle \cdot \left(\sum_{j=1}^{N} u_j^{(k)} (x_j - x'_j) \right) \right\} \end{aligned}$$

$$= \frac{1}{2}\sum_{j=1}^{N}\left\{\sum_{k=1}^{M}\left\langle H\left(\mathbf{T}\,\mathbf{u}^{(k)}\right)\,,\,H(\mathbf{T}\,\mathbf{x})\right\rangle\cdot\,u_j^{(k)}\right\}\,(x_j - x'_j)$$

$$\leq \quad 0. \tag{7}$$

The first equality is because $d(\mathbf{x},\mathbf{y}) = (N - \langle \mathbf{x},\mathbf{y}\rangle)/2$. The last inequality comes from the motion equation of the Kanerva memory in Equation (4). It can also be shown that the Kanerva memory can not stay at states with the same Liapunov function forever; i.e., the cases when $\Delta E = 0$ can not occur infinitely often. Also it is clear that the Liapunov function is bounded form below. Therefore, we conclude that the Kanerva memory is asymptotically stable in the parallel update mode. Similarly, the Kanerva memory is also asymptotically stable in the sequential update mode.

■

III. Conclusion

In this paper, the Kanerva memory is briefly described. The possibility of it being modified as a recurrent network associative memory is presented. Its asymptotic stability is proved by introducing a Liapunov function and showing that the function follows a descent trajectory as the Kanerva memory evolves.

Acknowledgement: Tzi-Dar Chiueh would like to thank the support given by the National Science Council, Taiwan, ROC under Grant NSC-79-0404-E-002-46.

References

[1] P. Kanerva, "Parallel Structure in Human and Computer Memory," in *Neural Networks for Computing*, J. S. Denker (editor), New York, NY : American Institute of Physics, 1986, pp. 247–258.

[2] P. Kanerva, *Sparse Distributed Memory,* Cambridge, MA: MIT Press, 1988.

[3] D. Marr, "A Theory of Cerebellar Cortex," *Journal of Physiology*, Vol. 202, 437–470, 1969.

[4] J. S. Albus, *Brain, Behavior, and Robotics.* Peterborough, NH : BYTE book of McGraw-Hill, 1981.

[5] P. A. Chou, "The Capacity of the Kanerva Associative Memory is Exponential," in *Neural Information Processing Systems*, D. Z. Anderson (editor), New York, NY : American Institute of Physics, 1988, pp. 184–191.