Memory Convergence Analysis for Different Iterative Procedures in RNN

Subhash Chandra Pandey, *Member, IAENG* , G.C.Nandi

Abstract— **Recurrent neural networks (RNN) are a class of densely connected single layer nonlinear networks of perceptrons. The network's energy function is defined through a learning procedure so that its minima coincide with states from a predefined set. . However, because of the network's nonlinearity a number of undesirable local energy minima emerge from the learning procedure. This has shown to significantly effect the network's performance. In this work we analyze the rate of convergence for three iterative procedures namely- Mann, Ishikawa and J-iterations in recurrent network and many important results have been worked out for decreasing as well as increasing functions. The results obtained are very useful for designing of inner product kernel of support vector machine with faster convergence rate.**

Index Terms—**Stable States, Mann Iteration, Ishikawa Iterations, J-Iteration, Convergence**

I. INTRODUCTION

Neural networks are a class of non-linear function approximators. It is originated by McCulloch and Pitt [1], Hebb[2], and Rosenblatt[3][4]. Hopfield defined a single-layer network consisting of interconnected individual perceptrons and modified perceptrons (with sigmoidal non-linearties) [5][6][7]. The basis for network operation as a *content addressable memory* is the Hebbian learning algorithm. The idea is to choose network connections in a way that the energy function associated with the network can be minimized for a set of desired network states. Unfortunately, because of its nonlinear character, the network has also exhibited non-desirable, local minima. This has shown to affect the network performance, both in its capacity and its ability to address its content [8][9][10]. Several approaches based on *simulated annealing* and other techniques have been proposed that deal with the problem of local minima [11][12][13][14][15][16]. In these approaches, an inherent assumption of the final network state (Gibbs) distribution is proposed. The motivation for these assumptions is that the Gibbs distribution provides a mechanism for the characterization of the global minima. In many applications, such as neural networks, however, the

desired final network state distribution corresponds to particular local minima, and not necessarily to the global minima. The use of Gibbs distribution is thus undesirable in many applications. A modification of this approach can be used to enhance the performance of neural networks.

In this work we have compared the number of iterations required to achieve the stable state in recurrent auto associative neural networks for the three iterative processes i.e. Mann, Ishikawa and J-iteration. The paper is organized as follows: In section II we introduced some preliminaries and definitions regarding the Mann, Ishikawa and J-iterative processes. In section III we relate the memory convergence concept in recurrent autoassociative neural network. Section IV describes the detailed experimental output obtained pertaining to these iterative processes. This is followed by applications in section V and concluding remarks are given in VI.

II. PRELIMINARIES

In this section, we will discuss the basic concepts related to Mann, Ishikawa and J-iterations.

Let $T : X \rightarrow X$ be a self mapping and (X, d) be a metric linear space. The three iterative processes are defined as:

Definition2.1: Let A be a lower triangular matrix with nonnegative entries and it is defined that $Z_{n+1} = T(V_n)$, where $V_n = \sum a_{nk} Z_k$ .The Mann iterative process is obtained by choosing a sequence $\{\alpha_n\}$ which satisfy (1) $\alpha_0 = 1$ (2) $0 \le \alpha_n < 1$ for $n > 0$ and (3) $\sum \alpha_n = \infty$.

Then the entries of A become $a_{nn} = \alpha_n$, $a_{nk} = \alpha_k \prod(1 - \alpha_i), k < n$. The above representation of A leads the following form: $Z_{n+1} = (1 - \alpha_n)Z_n + \alpha_n T Z_n$. It should be noted that for $\alpha_n = 1$, this form reduces to the Picard iterative process i.e. $Z_{n+1} = T Z_n$.

Definition2.2: The Ishikawa iterative procedure is defined as- Let $x_0 \in X$, $X_{n+1} = (1 - \alpha_n)X_n + \alpha_n T y_n, y_n = (1 - \beta_n)X_n + \beta_n T X_n$, $n > 0$, where $\{\alpha_n\}, \{\beta_n\}$ are sequences of positive numbers. They also satisfy the conditions (1) $0 \le \alpha_n \le \beta_n \le 1$ (2) $\lim \beta_n = 0$ and (3) $\sum \alpha_n \beta_n = \infty$. In [18] the inequality

condition has been replaced by $0 \le \alpha_n$ and $\beta_n \le 1$ and thus broaden the class of Ishikawa process. It also helps to reduce it to the Mann process by setting $\beta_n = 0$. In spite of this similarity, however, there is no any resemblance in convergence of these methods.

Definition2.3: The Jungck iterative procedure is commonly known as J-iterative procedure. It is defined as.

For $S : X \to X, T(X) \subseteq S(X), x_0 \in X, SX_{n+1} = TX_n, n = 0,1,....$

It is worthwhile to mention that if S is the identity map on X, this procedure will be reduce to the Picard iteration. In [17], various applications of J-iterative procedures have been discussed in numerical praxis.

Theorem2.4: Let $S, T[a,b] \to [a,b]$ be differentiable with $Sa \ne Sb$. Let $T'X$ and $S'X$ both are equal to zero for any $x \in]a,b[$. Then the pair (S,T) is said to be J-contraction of $[a,b]$ if and only if there exists a positive number $q < 1$ such that $|T'y| \le q|S'y|$ for all $y \in]a,b[$.

Let $\{X_n\}$ be a sequence generated by an iterative procedure $f(T, X_n)$ of the neural network that converges to a stable state of T. It is important to mention that for any map T, the initial choice x0 determines the stable state of T where the sequence $\{X_n\}$ will converge. Thus, if T is non-decreasing with three distinct stable states p, q, r and satisfies the conditions $0 \le p < q < r \le 1$, then $X_o \in [0,q)$ implies that $X_n \to p$ where as $X_0 \in (q,1]$ implies $X_n \to r$. The stable states p and r are known as the attractive global minima whereas q is a repulsive global minima. The sequence $\{X_n\}$ will not converge to the unstable state q unless and until $X_0 = q$ [18].

If $\{X_n\}$ and $\{Z_n\}$ are two iteration schemes which converge to the same global minima p, then $\{X_n\}$ is better than $\{Z_n\}$ if $|X_n - p| \le |Z_n - p|$ for all n i.e. $\{X_n\}$ converges to p faster than $\{Z_n\}$. This allows us to compare the rate of convergence of two iterative schemes.

III. MEMORY CONVERGENCE CONCEPT

We now look in more detail at the convergence of recurrent auto associative networks. Since the energy function $E(X)$ is bounded from below, the network evolves under the asynchronous dynamics toward $E(X^{n+1})$ such that

$$E(X^{n+1}) \le E(X^n) \tag{1}$$

It should be noted that vector X^{n+1} and X^n differ during the memory transitions by at most a single component, the stabilization of $E(X^{n+1})$ means that

$$E(X^{n+1}) = E(X^n) \quad \text{For } n > n_0 \tag{2}$$

The transition stop at the energy minimum, which also implies that

$$X^{n+1} = X^n \tag{3}$$

Thus the network reaches its stable state X^n, at the energy minimum $E(X^n)$. It is also obvious from energy function study that the only attractors of the discussed network are its stable states.

Let us begin with a one-dimensional recurrent system. For the output at $t = n+1$ we have

$$X^{n+1} = f(w, X^n) \tag{4}$$

The stable state is defined at X^* if the following relationship holds:

$$X^* = f(X^*) \tag{5}$$

Where it has been assumed that network parameter w in equation (4) is constant. In geometrical terms the global minima is found at the intersection of function $f(X)$ and X. In terms of the recursive formula (4), the global minima is said to be stable if

$$\lim_{n \to \infty} X^n = X^*$$

(6)

Or

$$\lim_{n \to \infty} e^{n+1} = 0 \quad . \tag{7}$$

Where e is the recursion error which can be defined as

$$e^{n+1} \cong |X^{n+1} - X^*|$$

The recursion error can be expressed using equation (4) as

$$e^{n+1} = |f(X^n) - X^*| \tag{8}$$

and further rearranged to the form

$$e^{n+1} = |[f\{X^* + (X^n - X^*)\} - f(X^*)]/(X^n - X^*|e^n \tag{9}$$

It is obvious that for a differentiable function f responsible for network's feed forward processing, this further reduces to

$$e^{n+1} \cong |f'(X^*)|e^n$$

(10)

Now the condition (10) translates to the form

$$|f'(X^*)| < 1 \tag{11}$$

This is sufficient condition for the existence of stable global minima in the neighborhood where condition (11) holds [19].

IV. EXPERIMENTS

Experiments have been conducted for studying the convergence rate of three iterative processes namely Mann, Ishikawa and J-iteration and it has been observed that for decreasing energy functions on [0, 1], computational analysis

of Mann and Ishikawa iteration exhibit interesting results. We programmed both schemes for the functions $f(X) = (1 - X)^m$, for m =7, 8... 29, taking initial choice $X_0 = 0.9$ and $\alpha_n = \beta_n = (n+1)^{-1/2}$. It is observed that in each case, the Mann process converges to a stable state (accurate to eight places) in 9-12 iterations whereas the Ishikawa method requires 38-42 iterations for the same accuracy.

In this section, we will continue this computational study and find the stable states of h and p given below.

Problem1. $h(X) = (1 - X)^m, m = 7,8,\ldots\ldots,29$

Problem2. $p(X) = 2X^3 - 7X^2 + 8X - 2$

First, we consider problem 1 and developed a computer program for Mann iterative procedure whose inputs are the initial guess X_0 and a value of m. The execution of program for m =7, $X_0 = 0.9$ and $\alpha_n = (n+1)^{-1/2}$ yield the results listed in table I.

Further, a program is developed to solve the same problem using Ishikawa iterative procedure. The program is executed taking m=7, $x_0 = 0.9$ and $\alpha_n = \beta_n = (n+1)^{-1/2}$ and results are listed in the table II.

Now, in order to solve the equation using J-iterative process, it is rewritten in the form of $f(X) = g(X)$ as $f(X) = X$ and $g(X) = (1 - X)^7$. It should be noted that the image of g is contained in the image f. Further, $|g'(X)| = 7|(1 - X)^6| \le 1 = |f'(X)|$ for $X \in [0.2, 0.9]$. Considering these f and g, again program is developed whose input is the initial guess X_0. The program is executed taking $X_0 = 0.9$ and the reading are recorded in the table III.

Table I: The Mann Process

n	fx$_n$	x$_{n+1}$
0	1e-07	1e-07
1	0.999999	0.70706
2	0.000185	0.298965
3	0.08321	91088
4	0.226626	0.206981
.	.	.
.	.	.
17	0.203456	0.203456

Table II: The Ishikawa Process

n	fx$_n$	x$_{n+1}$
0	1e-07	0.999999
1	8.235413e-44	0.355393
2	0.046245	0.2979776
3	0.084036	0.262381
4	0.118802	0.24034
.	.	.
.	.	.
29	0.203456	0.203456

Table III: The J-iterative Process

n	gx$_n$	fx$_{n+1}$	x$_{n+1}$
0	0.9	0.9	0.014939
1	0.014939	0.014939	0.451486
2	0.451486	0.451486	0.107387
3	0.107387	0.107387	0.27295
4	0.27295	0.27295	0.169307
.	.	.	.
.	.	.	.
29	0.203456	0.203456	0.203456

Table IV: The Mann Process

n	fx$_n$	x$_{n+1}$
0	0.988	0.988
1	0.999853	0.996381
2	0.999987	0.998463
3	0.999998	0.99923
4	0.999999	0.999574
5	1	0.999748
6	1	0.999843
.	.	.
.	.	.
27	1	1

Table V: The Ishikawa Process

n	fx$_n$	x$_{n+1}$
0	0.988	0.999853
1	1	0.999957
2	1	0.999982
3	1	0.9999991
4	1	0.999995
5	1	0.999997
6	1	0.999998
.	.	.
.	.	.
12	1	1

Following the same method, programs are developed to solve the problem 2 i.e. $p(X) = 2X^3 - 7X^2 + 8X - 2$, X_0 =0.9 for Mann and Ishikawa processes. The readings are listed in table IV and V.

Now a program is developed to solve problem 2 using J-iterative process taking

$$f(X) = 2X^3 - 7X^2 + 8X - 2, \text{ and } g(X) = 7X^2.$$

The outputs of program for $X_0 = 0.9$ are given in table VI.

In order to provide a detailed study, these programs are again executed changing the parameters such as $X_0 = 0.2$, $m = 29$ and $\alpha_n = \beta_n = (n+1)^{-1/4}$ in problem 1 and $X_0 = 0.6$ and $\alpha_n = \beta_n = (n+1)^{-1/4}$ in problem 2.

Table VI: The J-iterative Process

n	fx$_n$	fy$_n$	x$_{n+1}$
0	5.758	5.758	0.9069 57
1	5.84077 5	5.84077 5	0.9134 53
2	5.91853 5	5.91853 5	0.9195 13
3	5.99150 1	5.99150 1	0.9251 64
4	6.05989 8	6.05989 8	0.9304 3
.	.	.	.
.	.	.	.
165	6.99999 4	6.99999 4	1

The experiments have been repeated by taking different parameters and it is observed that for the decreasing function $h(X) = (1 - X)^m$, for $m = 7$ and $X_0 = 0.9$, the Mann process converges to a stable state in 17th iteration, the Ishikawa in 29th iteration and J-iterative in 29th iteration (Table I, II, III). Similarly, for $m = 29$ and $X_0 = 0.9$, the Mann and Ishikawa iterative processes take 10 and 32 iterations respectively whereas J-iteration takes only15 iterations to converge to a stable state. It is also noted that when $X_0 = 0.2$ i.e. nearer to stable state, the number of iterations for Mann, Ishikawa and J-processes are 7, 30 and 16 respectively. Moreover, when $\alpha_n = \beta_n = (n+1)^{-1/4}$. We obtain the stable state in 17 and 45 iterations for Mann and Ishikawa. It has also been observed that when $\alpha_n = 1$ i.e. Picard process the result is an oscillating sequence consisting of 0 and 1.

The experiments revealed interesting facts for increasing function $p(X) = 2X^3 - 7X^2 + 8X - 2$ when $X_0 = 0.9$, the number of iterations required for Mann and Ishikawa are 27 and 12 respectively whereas J-iteration requires 165 iterations (Table IV, V, VI). It is also observed that when the initial guess is away from the stable states e.g. $X_0 = 0.6$, we get the stable states in just 5 iterations for all the three processes. Similarly, for $\alpha_n = \beta_n = (n+1)^{-1/4}$, the Mann process takes 9 iterations while the Ishikawa process takes only 5 iterations to convergence to a stable state. It is worth mentioning that for $X_0 = 0.5$ or 2, all the three processes converge to X_0 itself, but when $0.5 < X_0 < 2$, then $\{X_n\}$ in all the three cases converge to 1.Further, when $\alpha_n = 1$ in algorithm, the result is obtained just in two iterations.

V. APPLICATIONS

In this section, we will describe the engineering application domain of our work. The results obtained in this work possess multifaceted real-life applications but here emphasis is given on support vector machine. Support vector machine is a learning machine and was initially designed to deal with binary classification with their linear decision functions. A support vector machine first maps the input points in to a high dimensional feature space and then finds a separating hyperplane that maximizes the margin between two classes in this feature space. Without any knowledge of mapping, the SVM uses kernels as the dot product functions in feature space. The solution of the optimal hyperplane can be written as a combination of a few input points called support vectors.

The important types of support vector machines are: (1) Polynomial learning machine and the inner product kernel $K(X, X_i), i = 1, 2, ..., N$ for this SVM is $(X^T X_i + 1)$, where p is specified a *priori* by the user (2) Radial-basis function network , the kernel for RBF is defined as $y = e^{\left(-\frac{1}{2\sigma^2} \|x - x_i\|^2 \right)}$. The width σ^2 , common to all the kernels, is specified as *priori* by the user (3) Two-layer perceptron for which the inner product kernel is $y = \tanh(\beta_0 x^T x_i + \beta_1)$, and (4) Gaussian, for which the inner product kernel is $y = \frac{1}{\sqrt{2\pi\sigma^2}} e^{\frac{-(x-a)^2}{2\sigma^2}}$. The detail study of support vector machine is beyond the scope of this paper. For in-depth study of SVM, interested readers are requested to refer [20]. The results obtained in this work can be used for designing the inner product kernel with higher convergence rate. We can consider it as future work of this paper.
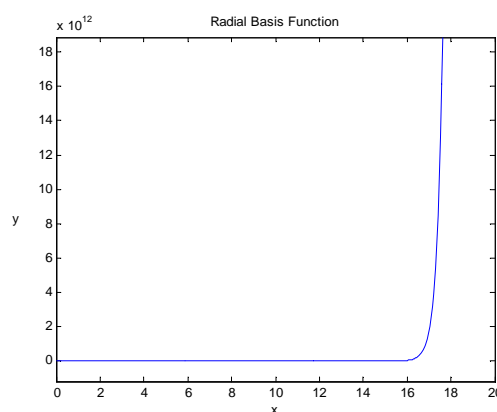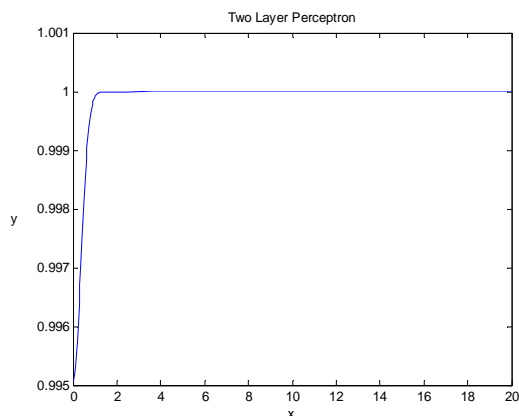

Fig (1) Learning curve for Radial-Basis Function

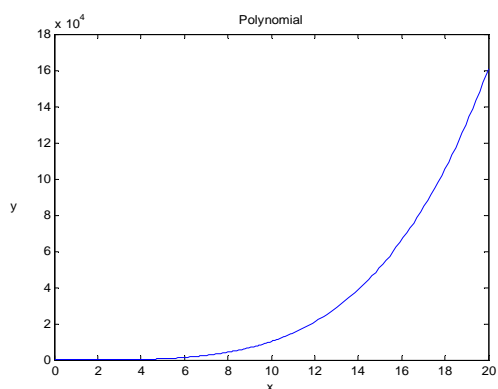Fig (2) Learning curve for Two-Layer Perceptron

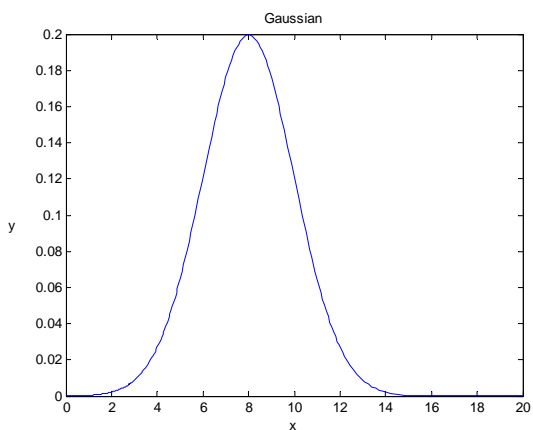Fig (3) Learning curve for Polynomial SVM

Fig (4) Learning curve for Gaussian SVM

VI. CONCLUSION

In this paper we analyze and compare the convergence rate for three iterative processes. For decreasing functions it has been concluded that for all combinations of X_0, m, α_n and β_n, the decreasing order of rate of convergence of iterative procedures is: Mann, J-iterations and Ishikawa. On increasing the value of m, the Mann and Ishikawa processes require more number of iterations while J-iterative process requires less number of iterations to locate the stable state. For the initial guess nearer to the stable state, the Ishikawa scheme shows an increasing tendency whereas the

J-iterations have decreasing tendency in number of iterations but the Mann process shows no change. Indeed, the speed of iterative procedures depend on the position of $\{\alpha_n\}$ and $\{\beta_n\}$ in the interval (0, 1). If α_n and β_n are larger, the stable state is obtained in more number of iterations for both Mann and Ishikawa scheme.

It has also been worked out that for increasing functions for all combinations of X_0, α_n and β_n, the decreasing order of rate of convergence of iterative processes are: Ishikawa, Mann and J-iterations. If the initial guess is away from the stable state, the number of iterations increases in each of the three processes. It explicitly indicates that if initial guess is closer to the stable state, the results will be obtained quickly. Similarly, larger values of α_n and β_n, produces the results quickly for the Mann and Ishikawa process. The Picard iterative procedure i.e. $\alpha_n = 1$ is the best process for the approximation of stable state in increasing function.

Most importantly, we may use the results obtained in this work for the design of inner-product kernel $K(X, X_i)$ of support vector machine to construct the optimal hyperplane in the feature space without having to consider the feature space itself in explicit form.

REFERENCES

[1] W.McCulloch and W.Pitts. "Alogical calculus of the ideas imminent in nervous activity," Bulletin of Mathematical Biophysics, vol 5, pp. 115-133, 1943.
[2] D.Hebb,The Organization of Behavior, New York, New York, John Wiley and Sons, 1949
[3] R. Rosenblatt, "The perceptron: A probabilistic model for information storage and organization in the brain, "Psychological Review", vol65, pp.386-408, 1958.
[4] R. Rosenblatt Principles of Neurodynamics, New York, New York: Spartan Books, 1959.
[5] J. Hopfield, "Neural Network and Physical Systems with Emergent Collective Computational Abilities," Proceedings of the Neural Academy of Science .USA, vol 79, pp. 2554-2558, April1982.
[6] J. Hopfield, "Neurons with Graded Response Have Collective Computational Properties like Those of Two-State Neurons," Proceedings of the Neural Academy of Science .USA, vol81, pp. 3088-3092.
[7] J. Hopfield and D. Tank, "Computing with Neural Circuits: A Model," Science, vol 233, pp. 625-633, 1986.
[8] S. Amari, "learning Pattern and Pattern sequence by Self-Organizing Nets of Threshold Elements,"IEEE Transaction on Computers, vol. C-21, no. 11, pp. 1197-1206.
[9] S. Amari and K. Maginu, "Statistical Neurodynamics of Associative Memory," Neural Networks,vol 1, pp. 63-73, 1988.
[10] R. Lippmann, "An introduction to Computing with Neural Net," IEEE Acoustics, Speech and Signal Processing Magazine,pp. 4-22,April 1987.
[11] S. Venkatesh and D. Psatis, "On Reliable Computations with Formal Neurons," IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 14, no. 1, pp. 87-91, January 1992.
[12] I. Arizono, A. Yamamoto, and H. Ohta, "Scheduling for Minimizing total actual flow time by neural networks," International Journal of Production Research, vol. 30, no.3, pp. 503-511, March 1992.
[13] B. Lee and B. Sheu, "Modified Hopfield Neural Networks for Retrieving the Optimal Solution," IEEE Transaction on Neural Networks. vol. 2, no. 1, pp. 137-142, January 1991.
[14] M. Lu, Y. Zhan, and G. Mu, "Bipolar Optical Neural Networks with Adaptive Threshold," Optik. Vol. 91, no 4, pp. 178-182, 1992.

[15] S.L. Lin, J.M. Wu, an C.Y. Liou, " Annealing by Perturbing Synapses," IEICE Transactions on Information and Systems, vol. E75-D, no. 2, pp. 210-218, March 1992.

[16] E. Wong, "Stochastic Neural Networks," Algorithmica, pp. 466-478, June 1991.

[17] S.L. Singh, "Anew approach in numerical praxis," Prog. Math. 32 (2), pp. 75-89, 1998.

[18] B. E. Rhoades, "Comment on two fixed point iteration methods,"J.Math. Anal. Appl.56, pp. 741-750.

[19] Jacek M. Zurada, "Introduction to Artificial Neural System",jaico publication, pp.349-350,2002.

[20]Simon Haykin, "Neural Networks-A Comprehensive Foundation",Pearson Education, Asia,pp.318-343, 2002.