



Aitken-based acceleration Methods for assessing  
Convergence of Multilayer Neural Networks

By RAMANI S. PILLA, SAGAR V. KAMARTHI and BRUCE G. LINDSAY

Technical Report #01-5-09

2000

---

**Center for Likelihood Studies**  
DEPARTMENT OF STATISTICS  
THE PENNSYLVANIA STATE UNIVERSITY  
UNIVERSITY PARK, PA 16802

# Aitken-Based Acceleration Methods for Assessing Convergence of Multilayer Neural Networks

Ramani S. Pilla<sup>\*</sup>, Sagar V. Kamarthi<sup>†</sup> and Bruce G. Lindsay<sup>‡</sup>

1

## Abstract

Suppose a nonlinear and non quadratic objective function is being optimized over a high dimensional parameter space. Often a closed-form solution does not exist and iterative methods are employed to find a local optimum of the function. However, algorithms designed for such high-dimensional optimization problems tend to be very slow in order to ensure reliable convergence behaviors. This problem occurs frequently, for example, in training multilayer neural networks (NNs) using a gradient-descent (backpropagation) algorithm. Lack of measures of algorithmic convergence force one to use *ad hoc* criteria to stop the training process.

This paper first develops the ideas of Aitken  $\delta^2$  method to accelerate the rate of convergence of an error sequence (value of the objective function at each step) obtained by training a NN with a sigmoidal activation function via the backpropagation algorithm. The Aitken method is exact when the error sequence is exactly geometric. However, theoretical and empirical evidence suggests that the best possible rate of convergence obtainable for such an error sequence is log-geometric (an inverse power of the epoch  $n$ ). The current paper develops a new *Invariant Extended-Aitken* acceleration method for accelerating log-geometric sequences. The resulting accelerated sequence enables one to predict the final value of the error function. These predictions can in turn be used to assess the distance between the current and final solution and thereby provides a

---

<sup>\*</sup> Division of Epidemiology and Biostatistics, 2121 West Taylor Street, Room 503 (MC 922), University of Illinois, Chicago, IL-60612. E-mail: pillar@uic.edu. <sup>†</sup> Department of Mechanical, Industrial and Manufacturing Engineering, 334 Snell Engineering Center, Northeastern University, Boston, MA-02115. E-mail: sagar@coe.neu.edu. <sup>‡</sup> Department of Statistics, 326 Thomas Building, The Pennsylvania State University, University Park, PA-16802. E-mail: bgl@psu.edu.

stopping criterion for a desired accuracy. Each of the techniques described in the paper is applicable to a wide range of problems. The invariant extended-Aitken acceleration approach shows improved acceleration as well as outstanding prediction of the final error in the practical problems considered.

### Keywords

Acceleration of sequences; Aitken  $\delta^2$ ; Cross-validation; Extrapolation; Learning rate; Log-geometric convergence; Linear convergence; Multilayer neural networks; Minimization; Rate of convergence; Rate index; Shift invariance properties; Sublinear convergence; Stopping rule.

## I. INTRODUCTION

Suppose  $E(\mathbf{w}): \mathfrak{R}^\nu \rightarrow \mathfrak{R}$  is a nonlinear and non quadratic objective function that is being minimized over a high-dimensional parameter space  $\mathbf{w}$ . In many instances, analytical closed form solution does not exist and iterative methods are employed to find  $E(\hat{\mathbf{w}}) \equiv \hat{E}$ , a local minimum on  $\mathfrak{R}^\nu$ . In these techniques, starting with an initial parameter value  $\mathbf{w}_0$  and a corresponding function value  $E(\mathbf{w}_0) = E_0$ , an iterative process constructs a sequence of parameter values  $\{\mathbf{w}_n\}_{n=0}^\infty$  and corresponding sequence of function values  $\{E_n = E(\mathbf{w}_n)\}_{n=0}^\infty$  that converges to  $\hat{E}$ . In the context of neural networks (NNs), the function is the total error at the  $n$ th stage which is given by  $E(\mathbf{w}_n) = \sum_p E[\mathbf{w}_n(p)]$ , where  $E[\mathbf{w}_n(p)]$  is the error for the  $p$ th training pattern (in a set of  $P$  training patterns used for training a network). This paper introduces the ideas of Aitken  $\delta^2$  method and presents its new invariant extension to accelerate the rate of convergence of the sequence  $\{E_n\}_{n=0}^\infty$ .

### A. Types of Convergence

Let  $r_n = (E_n - \hat{E})$  be the *residual* for each  $n > 0$ . If positive constants  $\rho$  and  $\lambda$  exist such that

$$\lim_{n \rightarrow \infty} \frac{|r_n|}{|r_{n-1}|^\rho} = \lambda, \quad (1)$$

then  $\{E_n\}_{n=0}^\infty$  is said to have convergence of *order*  $\rho$  with *asymptotic rate*  $\lambda$ . The convergence is said to be of *first-order* if  $\rho = 1$ ; i.e., if  $|r_n| = O(|r_{n-1}|)$  and *second-order* or *quadratic* if  $|\rho| = 2$ ; i.e., if  $r_n = O(|r_{n-1}|^2)$ . The order of convergence is determined only by the properties of the sequence that hold as  $n \rightarrow \infty$ ; i.e., the *tail* properties. Larger values of order  $\rho$  imply, in a sense, faster convergence, since the distance from the limit  $\widehat{E}$  is reduced at least in the tail by the  $\rho$ th power in a single step. If the order of convergence is equal to unity ( $\rho = 1$ ), then depending on whether  $0 < \lambda < 1$ ,  $\lambda = 1$  or  $\lambda = 0$  the sequence is said to converge *linearly*, *sublinearly* or *super-linearly* respectively. In addition, order  $\rho > 1$  corresponds to super-linear convergence.

A linearly converging sequence, with convergence ratio  $\lambda$ , has remainders  $r_n$  that converge as fast as the geometric sequence  $a\lambda^n$  for some constant  $a$ . Thus linear convergence is sometimes referred to as *geometric convergence*. In the case of linear convergence, some constant number of iteration steps, say  $m$ , are asymptotically required to reduce the magnitude of  $r_n$  by one tenth. In other words, after  $m$  further iterations, another decimal place of the  $E_n$  is correct. From the relationship  $r_{n+m} \approx \lambda^m r_n$ , the condition for  $m$  is  $m \approx -1/[\log_{10}|\lambda|]$ . It is clear that the convergence factor  $|\lambda|$  is an important description of the convergent behavior. For example, for  $|\lambda| = 0.316$ ,  $0.750$  and  $0.9716$ , one requires  $m = 2$ ,  $8$  and  $80$  iterations respectively to gain an extra correct decimal place.

Delahaye (1980) provides precise definitions of algorithmic sequence transformations, an analysis of the study of power sequence transformations and identifies the families of sequences that are not accelerable. He shows that the family of linearly convergent sequences is one of the biggest families of accelerable sequences and certain types of alternating and oscillating sequences can be accelerated using the Aitken acceleration method (which is not a linear transformation). Delahaye defines the sequence as “logarithmically convergent” if  $\lambda = 1$  for  $\rho = 1$  in equation (1). Suppose  $\mathcal{F}$  is a family of logarithmically convergent sequences. Delahaye shows that it is difficult to accelerate the sequences in

$\mathcal{F}$  and in fact it is not possible to accelerate all of them with a single transformation. However, one can find a specific transformation that accelerates many sequences. In essence, the problem is no longer how to accelerate the sequences in  $\mathcal{F}$ , but identifying the accelerable subsets of the family and finding a transformation to accelerate it. Our goal in this paper is to identify one set of  $\mathcal{F}$  and provide a transformation to accelerate it (Sections IV and V).

### *B. Background*

The situation of linear or sublinear convergence occurs in one of the popular learning algorithms for NNs called the “backpropagation” (BP) algorithm (Rumelhart & McClelland, 1986). The BP algorithm has emerged as the most popular algorithm for the supervised training of multilayer networks for its numerical stability and computational efficiency (Haykin, 1999). It has long been noted that the major weakness of the BP algorithm is its extremely slow rate of convergence. However, faster methods are not necessarily as reliable. Towsey et al. (1995) showed that a network trained with the conjugate gradient algorithm, which gives super-linear convergence in some cases, often gets stuck in a bad local minimum from which it never escapes, and the resulting network can have a poor generalization ability. Additionally, their empirical results show that the local minimum achieved with the BP algorithm will in fact be a global minimum, or at least a solution that is close enough to the global minimum for most practical purposes.

Several researchers have tried other techniques to improve the convergence of the network itself. One of the more popular techniques is a dynamic adaptation of the learning-rate (Jacobs, 1988; Roy, 1993; Salomon & van Hemmen, 1996; Silva & Almeida, 1990; Weir, 1991; Pirez & Sarkar, 1993) and/or the dynamic adaptation of the momentum term (Fahlman, 1989; Rumelhart & McClelland, 1986; Kanda et al., 1994; Yu et al., 1993). These improvements have the disadvantage of being substantially more complex and must often be “tuned” to fit the particular application (see Pfister & Rojas (1993) for

a detailed comparison of these methods). In light of these limitations, the BP algorithm is still a widely used method, despite its slow convergence.

Shepherd (1997) compared several first-order methods and second-order methods in terms of the global reliability and convergence speed. In terms of speed, his recommendation is to use the Levenberg-Marquardt (LM) nonlinear least-squares method – a second-order method. However, speed is not the only issue as the storage requirements, localities of the required quantities (for a parallel or network implementation), reliability of convergence and the problem of local minima also play a major role when it comes to large-scale problems. The two main problems with the LM algorithm is its  $O(w^2)$  storage requirements and  $O(w^3)$  computational costs along with its sensitivity to the presence of residuals at the solution, where  $w$  is the total number of parameters. He recommends a first-order method to maximize the chance of avoiding local minima. Based on his arguments and recommendation it seems like the on-line BP algorithm is a good option for large-scale problems such as the ones considered in Section VII. We believe that there may not be a single best approach and an optimal choice must depend on the problem and on the design criteria.

We would like to emphasize that our goal in here is not speeding up the BP algorithm itself instead accelerating the error sequence  $\{E_n\}_{n=0}^{\infty}$  obtained by training the network through the BP algorithm and provide a stopping criterion with a desired accuracy.

The rest of the paper is organized as follows. Section II describes the motivation for developing several different measures of algorithmic convergence. Section III presents the concepts of Aitken  $\delta^2$  method for accelerating the rate of convergence of a geometric sequence. Section IV derives an extended-Aitken  $\delta^2$  estimator to accelerate a sequence that is converging more slowly than a geometric sequence. In Section V, we introduce shift invariance properties and develop an invariant extended-Aitken acceleration method. Section VI develops the measures of algorithmic convergence and Section VII discusses the results of our methods when applied to several NN training problems. In Section

VIII, we present the diagnostics to assess the reliability of different prediction methods. Lastly, Section IX presents the conclusions and future work.

## II. MOTIVATION AND OBJECTIVES

Finding a stopping rule for training a feedforward multilayer NN when using the BP method is an important problem in the field of artificial neural networks. With good generalization as the goal, it is very difficult to figure out when it is best to stop training if one were to look at the learning curve for training all by itself. It is possible for the network to end up over fitting the training data if the training session is not stopped at the right point (Fausett, 1994).

Amari et al. (1996) presented a statistical theory for the over fitting phenomenon which provides a word of caution on the use of an early stopping method of training based on cross-validation (Hecht-Nielson, 1990; Haykin, 1999, Section 4.14). The theory depends on batch learning and supported with detailed computer simulations involving multilayer neural net classifier. Two modes of behavior are identified depending on the size of the training set: *(i) Non asymptotic Mode* for  $N < w$ , where  $N$  is the size of the training set and  $w$  is the number of free parameters in the network and an *(ii) asymptotic Mode* for  $N > 30w$ . In the former case, the early stopping method does improve the generalization performance of the network over exhaustive training, whereas in the latter case exhaustive learning is preferable to the early stopping method of training. The latter situation occurs frequently in problems of data mining with NN (Bigus, 1996).

In situations when either the early stopping method is prohibitively expensive – occurs when the validation subset is large – or practically impossible to have the validation subset (Twomey & Smith, 1995) and/or the early stopping rule is ineffective, the following *ad hoc* criterion is used to stop the training: “stop if the absolute rate of change in the  $E_n$  is sufficiently small” (typically it is considered to be small enough if it lies in the range of 0.1 to 1 percent per epoch). That is, if  $\sum_{i=1}^k |E_{n-i} - E_{n-1-i}| / (kE_{n-i})$  for a pre-determined

integer  $k$ . Unfortunately this criterion may result in a premature termination of the learning process (Haykin, 1999).

Often it is suggested to stop the iterative process when  $e_n = |E_n - E_{n-1}| \leq \tau$  is satisfied, where  $\tau$  is the tolerance. However, this sequence which serves more as a *lack-of-progress* criterion than a useful stopping rule, gives a misleading picture of the attained accuracy. Another simple criterion is to stop the iterations when  $|E_m - E_n| \leq \tau$  for  $m = n + 1, n + 2, \dots, n + k$  and for a predefined integer  $k$ . However, this criterion is also not satisfactory for an exponentially declining error-at-iteration function such as the one shown in Fig. 1. In this case, the algorithm may be stopped prematurely because a considerable increase in iterations causes only small changes in  $E$  even though the minimum,  $\hat{E}$ , is far from  $E_m$ . In essence, the stopping criterion can be satisfied even if the potential change in the error  $|E_m - \hat{E}|$  is many orders of magnitude larger than the tolerance.

Our goal in here is to develop an “ideal stopping criterion” based on the closeness of  $E_n$  to  $\hat{E}$  rather than based on the closeness of  $\mathbf{w}$  to  $\hat{\mathbf{w}}$ . The gradient-based methods such as the BP algorithm are more useful when the goal is to obtain accuracy in the objective function – which is what is needed in the neural network problems, as opposed to obtaining accuracy in parameter estimates. Note that if the goal is to obtain accuracy in parameter estimates, gradient-based rules are not particularly useful.

The stopping rules that depend on the residual sums of squares, calibrated as fractional sums of squares remaining, defined as  $[E(\mathbf{w}) - \hat{E}]/[E(\mathbf{w}_0) - \hat{E}]$  can be thought of as a measure of the degree of “shrinkage” of initial parameter estimates towards the solution parameter estimators. Future work would relate this shrinkage factor to the prediction sums of squares as in Stein Shrinkage and other methods used to reduce prediction squared errors in statistics (Draper & Van Nostrand, 1979; Gunst & Mason, 1980). We believe that the methods developed in this article could be used in a more systematic fashion to correct the overfitting problem – reducing the sums of squares to a minimum,

encountered in regression (Stone, 1974).

The Expectation-Maximization (EM) algorithm (Dempster et al., 1977) is a type of gradient-ascent algorithm. In a potentially slow algorithm like EM, a far weaker convergence criterion based on successive differences in the value of the objective function or changes in the parameter estimates between iterations can be very misleading (Titterton et al., 1985, p. 90). In addition, inherently slower algorithms are stopped at smaller objective function values as the stepwise changes are smaller (Lindsay, 1995, pp. 62–63). A natural summary of the speed of a linearly convergent algorithm like the EM is the rate of convergence of the function to its optimum value. The value of the objective function is an important measurement of the convergence of an algorithm, as it provides information about the accuracy of the parameter estimates on a confidence interval scale (Lindsay, 1995, pp. 131–132). Suppose we define  $(E_n - \hat{E})$ , as the “residual” of the error function at the  $n$ th iteration, then this residual can be used to assess the asymptotic rate of convergence of the BP algorithm (see Pilla & Lindsay, 2001). In light of these properties, a stopping rule based on an objective function is more desirable.

The current paper evaluates the performance of a theoretical Aitken  $\delta^2$  model and its new invariant extension to (i) predict the value of the function  $E(\mathbf{w})$  at any future iteration  $m$ , including the final value at  $m = \infty$ , (ii) predict the number of iterations needed to obtain a targeted function value, (iii) construct an algorithmic stopping rule that provides a reasonable guarantee that one is near the solution and (iv) assess the degree of convergence if the algorithm is stopped after a fixed number of iterations.

### III. ACCELERATING THE CONVERGENCE OF A SEQUENCE: AITKEN $\delta^2$ METHOD

This section describes the principles of the Aitken  $\delta^2$  method (Aitken, 1926). Suppose the sequence  $\{E_n\}_{n=0}^{\infty}$  is converging linearly to  $\hat{E}$  such that

$$[E_n - \hat{E}] \cong \lambda[E_{n-1} - \hat{E}] \quad \text{for all } n \text{ and } 0 < \lambda < 1 \quad (2)$$

or equivalently

$$[E_n - E_{n-1}] \cong (1 - \lambda)[\widehat{E} - E_{n-1}], \quad (3)$$

where “ $\cong$ ” means equation (1) holds. We assume  $n$  to be sufficiently large so that (3) holds in the sense of a good approximation for two consecutive index values and one can develop an acceleration technique called the *Aitken*  $\delta^2$  method. From (3), it is clear that if  $\lambda$  is very close to 1, then a very small change in the error need not mean that  $E_n$  is close to the minimum,  $\widehat{E}$ . If the constant  $\lambda$  can be estimated by  $\widehat{\lambda}_n$ , one can predict  $\widehat{E}$  from (3) as

$$\widehat{E} \cong E_{n-1} + \frac{1}{(1 - \widehat{\lambda}_n)}[E_n - E_{n-1}] \quad \text{for } n > 2. \quad (4)$$

The next two sections offer two derivations of the Aitken  $\delta^2$  method.

#### A. Estimating the Value of a Geometric Sequence

Given the sequence  $\{E_n\}_{n=0}^{\infty}$ , define the *backward difference*  $\nabla E_n = (E_n - E_{n-1})$  for  $n > 1$ . Higher powers are defined recursively by  $\nabla^k E_n = \nabla^{k-1}(\nabla E_n)$  for  $k > 2$ . From this it follows that,  $\nabla^2 E_n = \nabla(E_n - E_{n-1}) = \nabla E_n - \nabla E_{n-1} = (E_n - 2E_{n-1} + E_{n-2})$  for  $n > 2$ . From equation (2) one can show that  $(E_n - E_{n-1}) \cong \lambda(E_{n-1} - E_{n-2})$ , where  $\lambda$  is estimated via

$$\widehat{\lambda}_n = \frac{\nabla E_n}{\nabla E_{n-1}} \quad \text{for } n > 2. \quad (5)$$

We rewrite equation (4) to obtain the *Aitken*  $\delta^2$  estimator of  $\widehat{E}$ , denoted  $\widehat{E}_n^a$ , as

$$\widehat{E}_n^a = E_{n-1} - \frac{\nabla E_n \nabla E_{n-1}}{\nabla^2 E_n} \quad \text{for } n > 2. \quad (6)$$

The above formula corresponds to the continuous case best and this will be used to motivate the derivation of an extended-Aitken  $\delta^2$  estimator in Section IV.

The Aitken methodology is based on the following assumption on the sequence  $E_n$ :

**A1.** Sequence  $\{E_n\}_{n=0}^{\infty}$  converges linearly to a limit  $\widehat{E}$  such that  $r_n = (\lambda + \delta_{n-1})r_{n-1}$  with  $|\lambda| < 1$  and  $\lim_{n \rightarrow \infty} \delta_{n-1} = 0$  for  $n > 1$ , where  $r_n = (E_n - \widehat{E})$ .

Aitken  $\delta^2$  method converts any convergent sequence – no matter how generated – that satisfies assumption A1 into a more rapidly convergent sequence  $\{\widehat{E}_n^a\}_{n=0}^\infty$ , in accordance with the following general result called the *Aitken Acceleration Theorem*.

*Theorem III.1:* If the assumption A1 holds, then the Aitken sequence  $\{\widehat{E}_n^a\}_{n=0}^\infty$  given in equation (6) converges to  $\widehat{E}$  faster in the sense that

$$\left(\frac{\widehat{E}_n^a - \widehat{E}}{E_n - \widehat{E}}\right) \rightarrow 0 \text{ as } n \rightarrow \infty.$$

*Proof:* See Kincaid & Cheney (1990), p. 232.

*Remarks III.1:* (i) In general, accelerated sequence  $\{\widehat{E}_n^a\}_{n=0}^\infty$  might itself form a linearly convergent sequence, so it is possible to apply the Aitken  $\delta^2$  process again to obtain an even more rapidly convergent sequence. However, this procedure is limited for numerical reasons as the computation of the first and second differences together with the terms of the new sequence suffer from rounding errors. (ii) By rewriting  $E_{n-1} = (E_{n-2} + \nabla E_{n-1})$  and  $E_{n-1} = (E_n - \nabla E_n)$  in formula (6) one obtains a number of algebraically equivalent Aitken  $\delta^2$  formulae:  $\widehat{E}_n^a = E_{n-2} - (\nabla E_{n-1})^2 / \nabla^2 E_n$  and  $\widehat{E}_n^a = E_n - (\nabla E_n)^2 / \nabla^2 E_n$  respectively.

### B. Extrapolating the Differences to Zero

Aitken  $\delta^2$  can be given a second interpretation that allows a graphical analysis of its properties. In Aitken acceleration, what one wishes to find is the value of  $E$  when the change in  $e_n = (E_n - E_{n-1})$  is nearly zero. In the  $(E_n, e_n)$  plane let the line segment joining the points  $(E_{n-1}, e_{n-1})$  and  $(E_n, e_n)$  be extended to intersect the  $x$ -axis. It can be shown that  $x$ -value at the point of intersection, which corresponds to  $e = 0$ , equals the above derived Aitken  $\delta^2$  estimate. In other words, Aitken acceleration corresponds to approximating  $(E, e)$  by a linear function  $e = \beta E + \gamma$  passing through the points  $(E_{n-1}, e_{n-1})$  and  $(E_n, e_n)$ , and solving  $\beta E + \gamma = 0$  for  $E$  to obtain a next approximation to the desired root. (The plot would be exactly linear for an exactly geometric series.)

Depending upon the location of the points in the plane, this procedure may be an interpolation or an extrapolation at each stage. Thus overall linearity of the plot of  $(E_n, e_n)$  gives a diagnostic for the adequacy of using the Aitken acceleration estimate. Thus one is applying the *secant method* to the sequence  $(E_n, e_n)$ . Therefore, if the original sequence was linearly convergent to  $\hat{E}$ , then the resulting sequence  $\{\hat{E}_n^a\}_{n=0}^\infty$  is super-linearly convergent (Böhning, 1993).

#### IV. EXTENDED-AITKEN $\delta^2$ METHOD

In this section, we derive a correction factor for the original Aitken accelerated estimator given in (6). To this end, we start with the geometric case to motivate the non-geometric case.

*Definition IV.1:* A transformation  $\mathcal{T}$  defined on a sequence  $\{E_n\}_{n=0}^\infty$  is said to be *exact* if there exists an  $N < \infty$  such that  $\mathcal{T}(E_n) = \hat{E}$  if  $n > N$ , where  $\hat{E} = \lim_{n \rightarrow \infty} E_n$ .

##### A. Geometric Case

We note that the Aitken accelerated estimator is “exact” when the original error sequence  $\{E_n\}_{n=0}^\infty$  exhibits “exactly geometric convergence”. That is, if

$$E_n = \kappa e^{-\alpha n} + \mathcal{K}, \quad (7)$$

where  $\kappa, \alpha$  and  $\mathcal{K}$  are some constants. The constant  $\mathcal{K}$  is the limit of the sequence and  $\xi_n = \kappa e^{-\alpha n}$  is a *residual error* tending to zero as  $n \rightarrow \infty$ .

*Theorem IV.1:* The Aitken transformation  $\hat{E}_n^a$  is exact when the sequence  $\{E_n\}_{n=0}^\infty$  exhibits geometric convergence.

*Proof:* For mathematical simplification, without loss of generality, we assume that the limit of the sequence  $\mathcal{K} = \hat{E} = 0$ . By substituting  $E_n = \kappa e^{-\alpha n}$  in the Aitken transformation formula (6), we obtain  $\hat{E}_n^a = 0$ , the limit of the sequence. ■

For some problems that we considered (see Pilla & Lindsay (2000) and Section VII), the error sequence systematically deviated from geometric convergence. In fact the Aitken sequence  $\{\hat{E}_n^a\}_{n=0}^\infty$  was very close to the sequence  $\{d \cdot E_n\}_{n=0}^\infty$  for some constant  $d$  as shown in Fig. 2(a). A subsequent plot of  $(n, E_n^{-1})$  turned out to be nearly linear as shown in Fig. 2(b); indicating that one could predict future errors extremely well from any point in the sequence, even though the Aitken procedure was itself not working well. Thus the underlying assumption, A1, for the Aitken  $\delta^2$  method fails here. Pilla & Lindsay (2000) showed that this behavior would occur if the residual error  $(E_n - \mathcal{K})$  is proportional to  $n^{-\alpha}$  instead of  $e^{-\alpha n}$ . That is, an Aitken  $\delta^2$  method that accelerates a linearly convergent sequence may not work very well in the sublinear (or “close” to sublinear) case as the Aitken will not be exact in this case (follows from Theorem IV.1).

Tesauro et al. (1989) examined the analytical rates of convergence of the BP algorithm as the training was approaching to final stages. We summarize their findings. (i) For networks without hidden units that use the standard quadratic error function and a sigmoidal transfer function for output units: (a) the error decreases as  $n^{-1}$  for large  $n$ , (b) it is possible to obtain a different convergence rate for certain error and transfer functions, but the convergence can never be faster than  $n^{-1}$  and (c) the above results are unaffected by a momentum term in the learning algorithm, but one could potentially obtain  $n^{-2}$  error convergence by an adaptive learning-rate scheme. (ii) For networks with hidden units, the rate of convergence is expected to be same as the single layer case; however under certain circumstances one can obtain a slight polynomial speed-up for non sigmoidal units, or a logarithmic speed-up for sigmoidal units. (iii) Lastly, the the sigmoidal function provides the maximum possible convergence rate, and is therefore a recommended choice of transfer function.

### B. Non-geometric Case

The goal here is to extend the Aitken estimator to predict the error at infinity when indeed the error sequence is decreasing proportional to an inverse power of  $n$ . To motivate our derivation for the non-geometric case, we present another derivation for the geometric case by considering the continuous problem. Taking the first and second derivatives of  $E_n = \kappa e^{-\alpha n} + \mathcal{K}$  with respect to  $n$  we obtain  $E'_n = -\alpha \kappa e^{-\alpha n}$  and  $E''_n = \alpha^2 \kappa e^{-\alpha n}$  respectively. If we define the *Aitken acceleration factor* to be

$$\mathcal{A}_{n-1} = \frac{E'_n E'_{n-1}}{E''_n}, \quad (8)$$

then for an exact geometric sequence  $\mathcal{A}_{n-1} = \kappa e^{-\alpha(n-1)}$  predicts the residual error  $\xi_{n-1}$  as well as the value of  $\mathcal{K}$  as  $\widehat{\mathcal{K}} = (E_{n-1} - \widehat{\mathcal{A}}_n)$ . This prediction rule is exactly the Aitken acceleration method. If analytical solutions for  $E'_n$  and  $E''_n$  are not available, one can use symmetric differences:  $E'_n \approx \nabla E_n$  and  $E''_n \approx \nabla^2 E_n$  respectively.

Suppose the error sequence  $\{E_n\}_{n=0}^\infty$  is not geometric, instead it behaves as

$$E_n = \kappa n^{-\alpha} + \mathcal{K} = \kappa e^{-\alpha \log_e n} + \mathcal{K}. \quad (9)$$

In this case, the residual error  $\xi_n = \kappa e^{-\alpha \log_e n}$  approaches to zero much slower than geometrically. In fact, the error sequence is converging geometrically on the  $\log_e n$  scale, and hence we will call it an *exact log-geometric sequence* with a *rate index parameter*  $\alpha$ .

### C. Correction Factor for the Aitken Accelerated Estimator

In order to find a transformation that is more effective on a sequence that behaves as in equation (9), we need to require a transformation to be exact on a sequence which closely resembles an exact log-geometric sequence than a geometric sequence. To accomplish this, we consider estimation of the parameters from the derivatives of an exact log-geometric sequence. One obtains from (8)

$$\mathcal{A}_{n-1} = \frac{\alpha}{(\alpha + 1)} \frac{n}{(n - 1)} \cdot \xi_{n-1}.$$

That is, the Aitken factor no longer estimates the residual error, but rather a constant times the residual error. Thus an *extended-Aitken acceleration factor*

$$\widehat{\mathcal{A}}_{n-1}^\varepsilon = \left( \frac{\widehat{\alpha}_n + 1}{\widehat{\alpha}_n} \right) \left( \frac{n-1}{n} \right) \cdot \widehat{\mathcal{A}}_{n-1}$$

can be used to estimate the residual error  $\xi_{n-1}$  leading to the extended-Aitken accelerated estimator  $\widehat{\mathcal{K}} = E_{n-1} - \widehat{\mathcal{A}}_{n-1}^\varepsilon$ , denoted  $\widehat{E}_n^\varepsilon$ . A simpler equivalent formula arises by replacing  $(n-1)/n$  with 1 in which case  $\widehat{E}_n^\varepsilon$  would always be less than  $\widehat{E}_n^a$  if  $\widehat{\alpha}_n > 0$ . (In Appendix A, we show that  $\widehat{\mathcal{A}}_{n-1}^\varepsilon/\xi_{n-1} \rightarrow 1$  as  $n \rightarrow \infty$ .)

Suppose we use ordinary Aitken on the sequence  $\{E_n\}_{n=0}^\infty$  which is converging as an exact log-geometric sequence. Then the change in Aitken estimate becomes

$$\begin{aligned} \nabla \widehat{E}_n^a &= \widehat{E}_n^a - \widehat{E}_{n-1}^a = \nabla E_{n-1} - \left[ \frac{\widehat{\alpha}_n}{\widehat{\alpha}_n + 1} \cdot \xi_{n-1} - \frac{\widehat{\alpha}_n}{\widehat{\alpha}_n + 1} \cdot \xi_{n-2} \right] \\ &= \nabla \xi_{n-1} - \frac{\widehat{\alpha}_n}{\widehat{\alpha}_n + 1} \cdot \nabla \xi_{n-1} \quad (\text{since } \mathcal{K} \text{ is fixed}) \\ &= \frac{1}{\widehat{\alpha}_n + 1} \cdot \nabla E_{n-1}. \end{aligned}$$

Thus  $\alpha$  can be estimated from the sequences  $\{\widehat{E}_n^a\}_{n=0}^\infty$  and  $\{E_n\}_{n=0}^\infty$  as

$$\widehat{\alpha}_n = \frac{\nabla E_{n-1}}{\nabla \widehat{E}_n^a} - 1 \quad \text{for } n > 2. \quad (10)$$

In Appendix A (Lemma A.2), we show that  $\widehat{\alpha}_n \rightarrow \alpha$  as  $n \rightarrow \infty$ . Hence the *extended-Aitken accelerated estimator* for the error at infinity becomes

$$\widehat{E}_n^\varepsilon = E_{n-1} - \widehat{\mathcal{A}}_{n-1}^\varepsilon \quad \text{for } n > 2. \quad (11)$$

*Theorem IV.2:* The Extended-Aitken transformation  $\widehat{E}_n^\varepsilon$  is exact when the sequence  $\{E_n\}_{n=0}^\infty$  exhibits log-geometric convergence.

*Proof:* For mathematical simplification, without loss of generality, once again, we assume that the limit of the sequence  $\mathcal{K} = \widehat{E} = 0$ . By substituting  $E_n = \kappa e^{-\alpha \log_e n}$  in (11), we obtain  $\widehat{E}_n^\varepsilon = 0$ , the limit of the sequence. ■

Extended-Aitken is based on the following assumption of “asymptotic log-geometric convergence”:

**A2.** Sequence  $\{E_n\}_{n=0}^{\infty}$  converges to  $\widehat{E}$  as  $n \rightarrow \infty$  and the residuals  $r_n$  are proportional to the inverse power of  $n$ . That is,  $r_n = (E_n - \widehat{E}) = kn^{-\alpha}(1 + o(1))$  for  $\alpha > 0$ .

It is clear that, a sequence satisfying assumption A2 has convergence of order one and is converging sublinearly since  $\lambda = 1$  in equation (1). The rate at which the sequence  $\{E_n\}_{n=0}^{\infty}$  is tending to  $\widehat{E}$  is governed by  $\alpha$ ; the smaller the  $\alpha$  is, the slower the sequence is approaching to the solution. In other words, the sequence with  $0 < \alpha < 1$  converges slower than a sequence with  $\alpha \geq 1$ .

Some properties of the extended Aitken accelerated estimator are established in the following two theorems whose proofs are relegated to Appendix A.

*Theorem IV.3:* If assumption A2 holds, then the Aitken sequence  $\{\widehat{E}_n^a\}_{n=0}^{\infty}$  does not converge to  $\widehat{E}$  faster than the original sequence  $\{E_n\}_{n=0}^{\infty}$ . Indeed

$$\lim_{n \rightarrow \infty} \left( \frac{\widehat{E}_n^a - \widehat{E}}{E_n - \widehat{E}} \right) = \frac{1}{(\alpha + 1)} \quad \text{for any } \alpha > 0.$$

From the above theorem it follows that the underlying assumption for the Aitken  $\delta^2$  method fails and hence it may not work very well when the sequence has sublinear (or close to sublinear) convergence. In fact, under the log-geometric convergence, we have  $\alpha = 1$  giving  $\widehat{E}_n^a \approx (1/2)E_n$  (see Fig. 3 for illustration).

*Theorem IV.4:* If assumption A2 holds, then the extended-Aitken sequence  $\{\widehat{E}_n^\varepsilon\}_{n=0}^{\infty}$  converges to  $\widehat{E}$  faster than the sequence (i)  $\{E_n\}_{n=0}^{\infty}$  in the sense that

$$\left( \frac{\widehat{E}_n^\varepsilon - \widehat{E}}{E_n - \widehat{E}} \right) \rightarrow 0 \quad \text{as } n \rightarrow \infty$$

and (ii)  $\{\widehat{E}_n^a\}_{n=0}^{\infty}$  in the sense that

$$\left( \frac{\widehat{E}_n^\varepsilon - \widehat{E}}{\widehat{E}_n^a - \widehat{E}} \right) \rightarrow 0 \quad \text{as } n \rightarrow \infty.$$

Bjørstad et al. (1981) considered a nonlinear acceleration formula reminiscent of the Aitken method – called a modified Aitken  $\delta^2$  formula – to accelerate the convergence of the sequences  $\{t_n\}_{n=1}^\infty$  satisfying

$$t_n = t_\infty + n^{-\theta}(\mu_0 + \mu_1 n^{-1} + \mu_2 n^{-2} + \dots) \quad (12)$$

with  $\mu_i \neq 0$  and positive for  $i = 0, 1, \dots$  and for possibly non-integral  $\theta > 0$ . They considered several examples to assess its performance when the limit  $t_\infty$  is known exactly. Note that we have  $E_n = \widehat{E} + n^{-\alpha}(\mu_0 + o(n^{-1}))$  with  $\alpha > 0$  and  $\mu \neq 0$ . Our extended Aitken formula is equivalent to the formula (2.6) provided by Bjørstad et al. (1981) with  $i = 0$ . We provide a new extension which is similar in spirit to, but more general than, the modified Aitken acceleration method given by Bjørstad et al. (1981) and relies on the idea of a “shift invariance” (defined next) property.

## V. INVARIANT EXTENDED-AITKEN ACCELERATION METHOD

In this section we define an extended log-geometric family of sequences with one additional parameter. Since the log-geometric convergence is a tail behavior, this additional parameter will enable us to shift the tail of the sequence to allow for any uncertainty about where the tail behavior starts. The “exact shifted-log-geometric” sequences have the form  $E_n = \kappa e^{-\alpha \log_e(n+s)} + \mathcal{K}$ . Note that if  $\{E_n\}_{n=0}^\infty$  is exact shifted-log-geometric sequence with a set of parameters  $(\mathcal{K}, \alpha, s)$ , then  $\{E_{n+m}\}_{n=0}^\infty$  is also such a sequence with the set of parameters  $(\mathcal{K}, \alpha, s + m)$ .

Suppose we use our algorithm with an initial value of  $\mathbf{w}_0$ , then at epoch  $m$  we have a current value of  $\mathbf{w}_m$ . If we were to restart the algorithm from an initial value  $\mathbf{w}_m$ , then  $E_n^*$ , the error at  $n$ , is equal to the error  $E_{(n+m)}$  at epoch  $n^* = (n + m)$  of the original sequence. If we wish to have an approximation that works equally well for the original sequence  $\{E_n\}_{n=0}^\infty$  and the shifted sequence  $\{E_n^* = E_{n+m}\}_{n=0}^\infty$ , then it should have *shift invariance* properties in the sense that the predictions from  $E_n$  and  $E_n^*$  should

be identical. To this end, Pilla & Lindsay (2000) found that the Aitken  $\delta^2$  formula (7) is shift invariant; however, the extended-Aitken formula (9) is not independent of this shift on the time scale. That is, if  $n$  is replaced by  $n^*$  in the exact geometric and log-geometric sequences respectively, the shifted geometric sequence is still geometric on the new scale  $n^*$  but the log-geometric sequence is no longer exactly log-geometric. We will verify these statements below. However, if we derive a method of acceleration for all exact shifted-log-geometric sequences, we will have a shift invariant methodology.

*Lemma V.1:* A geometric sequence is invariant under a transformation whereas a log-geometric sequence is not.

*Proof:* Consider a shift in the geometric sequence by writing  $E_n = \kappa e^{-\alpha(n+s)} + \mathcal{K} = \kappa e^{-\alpha s} e^{-\alpha n} + \mathcal{K} = \kappa^* e^{-\alpha n} + \mathcal{K}$ , where  $\kappa^* = \kappa e^{-\alpha s}$  is a constant independent of  $n$ ; indicating that the *shift parameter*  $s$  only effects the constant  $\kappa$ . Next, we will rewrite the log-geometric sequence with a shift parameter as

$$\begin{aligned} E_n &= \kappa (n+s)^{-\alpha} + \mathcal{K} = \kappa n^{-\alpha} \left(1 + \frac{s}{n}\right)^{-\alpha} + \mathcal{K} \\ &= \kappa n^{-\alpha} \left(1 - \frac{\alpha s}{n} + R^*\right) + \mathcal{K} \quad (\text{follows from Lemma A.1}), \end{aligned} \tag{13}$$

where  $R^* = \sum_{i=2}^{\infty} R_i(s/n) \rightarrow 0$  since  $|R_i(s/n)|$  approaches to zero as  $n \rightarrow \infty$  for a fixed  $s$  and  $\alpha$ . We will only consider the leading terms and write

$$\begin{aligned} E_n &= \kappa n^{-\alpha} \left[1 - \frac{\alpha s}{n} + o\left(\frac{1}{n}\right)\right] + \mathcal{K} \\ &= \kappa n^{-\alpha} + \alpha s n^{-(\alpha+1)} + o\left(n^{-(\alpha+1)}\right) + \mathcal{K}. \end{aligned}$$

Thus the shift on a time scale effects the log-geometric sequence. ■

More generally, truncated versions of expression (12) are not shift-invariant, so predictions based on them will change with shifts on the sequence.

We consider a parametric family of exact sequences: (i)  $\mathcal{K} + \kappa e^{-\alpha n}$  and (ii)  $\mathcal{K} + \kappa e^{-\alpha \log_e(n+s)}$  called geometric and log-geometric sequences respectively. We consider a

larger family of sequences generated by these families, which we call “asymptotic geometric” and “asymptotic log-geometric” respectively. There exists a set of parameters  $\mathcal{K}, \alpha$  and  $\kappa$  such that in the case of geometric sequence  $(E_n - \mathcal{K})/e^{-\alpha n} \rightarrow \kappa$  and in the case of log-geometric sequence  $(E_n - \mathcal{K})/e^{-\alpha \log_e(n+s)} \rightarrow \kappa$  respectively. Thus both the log-geometric sequence and its shifted version are in the asymptotic log-geometric family. However, the latter is a richer family of sequences and provides a shift-transformation – since the sequences in the family of asymptotic sequences are invariant to the starting position.

#### A. Estimator Based on the Shifted-log-geometric Sequence

Pilla & Lindsay (2000) derived a correction factor for the original Aitken accelerated estimator by considering the log-geometric sequence with a shift parameter  $s$  as in equation (13). They called this sequence a *shifted-log-geometric* sequence since  $E_n \propto \kappa e^{-\alpha \log_e(n+s)} = \xi_n$ . (We will see in Section VII that this model fits our error sequences much better than log-geometric without shift.) Following Pilla & Lindsay, we examine in detail the Aitken factor when the sequence  $\{E_n\}_{n=0}^{\infty}$  is an exact shifted-log-geometric. From  $E'_n = -\kappa \alpha (n+s)^{-(\alpha+1)}$  and  $E''_n = \kappa \alpha (\alpha+1)(n+s)^{-(\alpha+2)}$ , we obtain

$$\mathcal{A}_{n-1} = \left( \frac{\alpha}{\alpha+1} \right) \cdot \left( \frac{n+s}{n-1+s} \right) \cdot \xi_{n-1},$$

where  $\xi_{n-1} = \kappa/(n-1+s)^\alpha$ . If one can estimate  $\alpha$ , one can use an *invariant extended-Aitken acceleration factor*

$$\hat{\mathcal{A}}_{n-1}^i = \left( \frac{\hat{\alpha}_n + 1}{\hat{\alpha}_n} \right) \left( \frac{n-1+\hat{s}_n}{n+\hat{s}_n} \right) \cdot \hat{\mathcal{A}}_{n-1}$$

to predict the residual error at infinity, leading to the *invariant extended-Aitken estimator*  $\hat{\mathcal{K}} = E_{n-1} - \hat{\mathcal{A}}_{n-1}^i$ , denoted  $\hat{E}_n^i$ . By replacing the *shift adjustment factor*  $(n-1+\hat{s}_n)(n+\hat{s}_n)^{-1}$  with 1 it follows that  $\hat{E}_n^i$  would always be less than  $\hat{E}_n^a$  if  $\hat{\alpha}_n > 0$ .

### B. Estimation of the Rate Index from the Aitken Sequence

In this section, we will derive the estimator for the rate index parameter, given by Pilla & Lindsay (2000). From  $E_n''/E' = -(\alpha + 1)(n + s)^{-1}$  and  $E_n'''/E'' = -(\alpha + 2)(n + s)^{-1}$  we obtain  $(\alpha + 1)/(\alpha + 2) = (E_n'')^2/(E_n'E_n''')$ . Solving for  $\alpha$  yields the estimator of the rate index parameter as

$$\hat{\alpha}_n^i = \left[ \frac{E_n'''E_n' - 2(E_n'')^2}{(E_n'')^2 - E_n'''E_n'} \right] \quad \text{for } n > 3$$

which is independent of the shift parameter or its estimator;  $\hat{\alpha}_n^i$  is used to distinguish from the  $\hat{\alpha}_n$  in the extended case.

### C. Estimating the Shift Parameter from the Aitken Sequence

In what follows, we will derive the estimator for the shift parameter given by Pilla & Lindsay (2000). Solving the equation  $E_n''/E_n' = -(\alpha + 1)(n + s)^{-1}$  for  $s$  we obtain

$$\hat{s}_n = -[(\hat{\alpha}_n^i + 1) \cdot E_n' (E_n'')^{-1} + n] \quad \text{for } n > 3.$$

Hence the *invariant extended-Aitken accelerated estimator* for the error at infinity becomes

$$\hat{E}_n^i = E_{n-1} - \hat{\mathcal{A}}_{n-1}^i \quad \text{for } n > 3. \quad (14)$$

Similar to Theorem IV.4, one can show that the invariant extended-Aitken accelerated sequence  $\{\hat{E}_n^i\}_{n=0}^\infty$  converges to  $\hat{E}$  faster than the sequences  $\{E_n\}_{n=0}^\infty$  and  $\{\hat{E}_n^a\}_{n=0}^\infty$  as  $n \rightarrow \infty$ . The details will not be presented here.

Just as in the ordinary Aitken case, one can find other algebraically equivalent formulae for an invariant extended-Aitken accelerated estimator by rewriting equation (14) using  $E_{n-1} = (E_{n-2} + \nabla E_{n-1})$  and  $E_{n-1} = (E_n - \nabla E_n)$  respectively.

## VI. ASSESSING CONVERGENCE VIA THE ACCELERATION SCHEMES

This section describes how the Aitken  $\delta^2$ , including the invariant extended-Aitken and the related concepts can be applied to make predictions as well as to find a stopping rule.

### A. Prediction of Error

One can use the geometric or shifted-log-geometric parameter estimates to predict not just the final error, but also at any future stage  $m$ . For the geometric case, consider the following set of equations

$$\begin{aligned} \nabla E_{n+1} &\cong \lambda \nabla E_n \\ &\vdots \\ \nabla E_{n+s} &\cong \lambda \nabla E_{n-1+s} \cong \lambda^s \nabla E_n. \end{aligned}$$

Summing the right and left hand sides of the above equations results in

$$E_{n+s} - E_n \cong \sum_{i=2}^s \lambda^i \nabla E_n \cong \lambda \left[ \frac{1 - \lambda^s}{1 - \lambda} \right] \nabla E_n.$$

We predict the error  $E$  at any future stage  $m = (n + s)$  using the Aitken  $\delta^2$  estimator (Pilla et al., 1995) as

$$\hat{E}_{n,m}^a = E_n + \hat{\lambda}_n \left[ \frac{1 - \hat{\lambda}_n^{\{m-n\}}}{1 - \hat{\lambda}_n} \right] \nabla E_n, \quad (15)$$

where the subscripts  $n$  and  $m$  in  $\hat{E}_{n,m}^a$  refer to the current and future stages respectively. Recall that for  $m = \infty$ , one can predict the error using equation (6).

For the invariant log-geometric case, we have  $E_m = \xi_m + \mathcal{K}$ . After simplification, we have

$$\hat{E}_{n,m} = \frac{\hat{\kappa}}{(m + \hat{s}_n)^{\hat{\alpha}_n^i}} + (E_{n-1} - \hat{\xi}_{n-1}) = \frac{\hat{\kappa}}{(m + \hat{s}_n)^{\hat{\alpha}_n^i}} + \hat{E}_n^i.$$

From  $\hat{\xi}_{n-1} = \hat{\kappa}/(n-1 + \hat{s}_n)^{\hat{\alpha}_n^i}$ , it follows that  $\hat{\kappa} = (n-1 + \hat{s}_n)^{\hat{\alpha}_n^i} \hat{\xi}_{n-1} = (n-1 + \hat{s}_n)^{\hat{\alpha}_n^i} \hat{\mathcal{A}}_{n-1}^i$ . Hence the predicted error at stage  $m$  based on the invariant extended-Aitken  $\delta^2$  estimator

(Pilla & Lindsay, 2000; Pilla et al., 2000) is

$$\begin{aligned}\widehat{E}_{n,m}^i &= \left(\frac{n-1+\widehat{s}_n}{m+\widehat{s}_n}\right)^{\widehat{\alpha}_n^i} \cdot \widehat{\mathcal{A}}_{n-1}^i + \widehat{E}_n^i \\ &= E_{n-1} + \left[\left(\frac{n-1+\widehat{s}_n}{m+\widehat{s}_n}\right)^{\widehat{\alpha}_n^i} - 1\right] \cdot \widehat{\mathcal{A}}_{n-1}^i.\end{aligned}\quad (16)$$

### B. Prediction of the Number of Stages

Our next goal is to predict the number of stages  $m$  required to reach a targeted error,  $E_{\text{tar}}$ , when the error sequence exhibits geometric or shifted-log-geometric convergence. By rearranging equations (15) and (16) and replacing  $\widehat{E}_{n,m}^a$  and  $\widehat{E}_{n,m}^i$  with  $E_{\text{tar}}$  one can predict the number of stages required to reach the targeted final error  $E_{\text{tar}}$  as

$$\widehat{m}^a = n + \frac{1}{\log \widehat{\lambda}_n} \log \left[ 1 - \left( \frac{1 - \widehat{\lambda}_n}{\widehat{\lambda}_n} \right) \frac{(E_{\text{tar}} - E_n)}{\nabla E_n} \right]$$

and

$$\widehat{m}^i = (n - 1 + \widehat{s}_n) \left[ \frac{(E_{\text{tar}} - E_{n-1})}{\widehat{\alpha}_n^i} + 1 \right]^{-1/\widehat{\alpha}_n^i} - \widehat{s}_n.$$

respectively. Thus  $\widehat{s}^a = \widehat{m}^a - n$  and  $\widehat{s}^i = \widehat{m}^i - n$  are the estimated additional number of stages required to reach the desired error. We note that if the targeted error is too small, namely smaller than the predicted final error, the prediction is that the targeted error is impossible to achieve even in an infinite number of epochs.

### C. Stopping Rule and Assessment of Convergence Accuracy

Ensuring the convergence to a local minimum before stopping an iterative process is important in many minimization problems. Construction of a rule for stopping an iterative process is a somewhat *ad hoc* process that cannot be perfect for every problem, yet it calls for considerable thought. The decision of when one is “close enough to the minimum error” usually has two parts: “Have we approximately solved the minimization problem?” and “Have we reached the targeted accuracy?” Our goal is to stop once the

targeted accuracy is attained, but also not to be wasteful of computing time. A criterion to stop the iterative process should be designed in such a way that it will not terminate too early due to local properties of the error surface.

Suppose the error sequence has geometric convergence asymptotically. In this case, one simple algorithmic stopping rule would require that the predicted distance to the solution,  $|\widehat{E}_n^a - E_{n-1}|$ , be smaller than tolerance  $\tau$  for some specified number of successive stages  $s$ . That is, we stop the iterative process at the  $(n + s)$ th stage if the following set of equations is satisfied:

$$\begin{aligned} |\widehat{E}_n^a - E_{n-1}| &< \tau, \\ |\widehat{E}_{n+1}^a - E_n| &< \tau, \\ &\vdots \\ \text{and } |\widehat{E}_{n+s}^a - E_{n-1+s}| &< \tau. \end{aligned}$$

The tolerance  $\tau$  would be chosen to reflect the user's idea of being close enough to zero for the problem at hand. We check the above conditions for  $s$  consecutive stages mainly to maintain the regularity of the iterative process; i.e., to see if the rates are holding steady and the predictions are good.

If the error sequence has log-geometric convergence asymptotically, then the stopping rule based on the invariant extended-Aitken  $\delta^2$  estimator would give a better accuracy. A possible stopping rule along the above lines would be to replace  $\widehat{E}_n^a$  with  $\widehat{E}_n^i$  in the above set of equations.

## VII. EXPERIMENTAL RESULTS

In this section the convergence measures are demonstrated while training twolayer feedforward NNs through the BP algorithm in the alphanumeric and car problems. Recall from Section II that the performance of the cross-validation method for early stopping will be poor in the case of the alphanumeric problem. In the case of car problem, the

number of training patterns is very large compared to the network parameters (see Table I) so that the cross-validation method may not prove useful. In addition, compared to the number of network parameters, the amount of available training data is limited in which case it is not a good practice to break the training data into cross-validation dataset and training dataset. In this scenario, one needs a more sophisticated convergence criterion to stop the training process.

Table I gives the network architecture for the problems considered in the study. In general, a small learning-rate parameter  $\eta$  results in a slower convergence. However, it can locate “deeper” local minima in the error surface than a larger  $\eta$  (Haykin, 1999, p. 194). We chose an  $\eta$  value of 0.01 to avoid oscillations in the error during the learning as well as to achieve a small value for the final error at convergence. The proposed methodology does not restrict the learning rate parameter  $\eta$  to be any value smaller than what is usually chosen for the BP algorithm.

The thirty six binary patterns in the alphanumeric problem were created by representing the twenty six alphabet and ten digits on a  $7 \times 5$  grid of binary values. The objective of the second problem is to classify cars as unacceptable, acceptable, good or very good based on several attributes (Bohanec & Rajkovic, 1988). The performance of our new methodology in the thyroid decease problem (Coomans et al., 1983) which were conducted to determine whether the thyroid glands are normal, hypo or hyper and tic-tac-toe problem (Aha, 1991) are very similar to those of the above problems and hence are not presented here.

In each problem, presentation of the complete set of training patterns was considered as an epoch and the network was trained for 60,000 epochs; the error was recorded only at every 100th epoch which we refer to as one stage. This decimation of the error sequence was done as a smoothing technique to improve the regularity of the iterative process. The hidden and output layer nodes of the networks used bipolar sigmoidal,  $f_h(x) = (1 - e^{-x}) / (1 + e^{-x})$ , and binary sigmoidal,  $f_o(x) = 1 / (1 + e^{-x})$ , activation

functions respectively. Pilla et al. (1995a, 1995b) present the results for the XOR and alphanumeric problems when the binary sigmoidal activation function was used for both the hidden and output layer nodes.

#### A. Prediction of Error

The prediction results are summarized in Table 2. In the alphanumeric problem, the Aitken acceleration and its invariant extension predict the error at  $m = 250$  (from the current stage  $n = 150$  with  $E_{150} = 0.4888$ ) as  $\hat{E}_{150,250}^a = 0.2889$  and  $\hat{E}_{150,250}^i = 0.2686$  respectively ( $E_{250} = 0.2692$ ). That is, the invariant one predicts the change in the error function over 100 stages with less than 1% relative error whereas the Aitken has 9% relative error, where the “relative error” is defined as

$$\hat{\psi}_m^g = \frac{|\hat{E}_{n,m}^g - E_n|}{|E_m - E_n|}$$

with  $g = a$  or  $i$  corresponding to the Aitken or its invariant one respectively.

Fig. 4 gives the plots of the number of stages versus (i) the actual error  $E_n$ , (ii) the Aitken predicted error  $\hat{E}_{n,m}^a$  and (iii) the invariant extended-Aitken predicted error  $\hat{E}_{n,m}^i$  for the alphanumeric (left panel) and car (right panel) problems respectively. For each plot in the figure, the current stage  $n$  was chosen and then varied the future stages from  $m = n$  to  $m = (n + s)$ . From these plots it is clear that the invariant extended-Aitken predicted sequence follows closely the actual error sequence where as the Aitken predicted sequence moves further away from it as the training proceeds farther away from the current stage.

#### B. Prediction of Epochs

Goal is to predict the additional number of stages  $s = (m - n)$  required to obtain a target error of  $E_{\text{tar}}$  from the current stage  $n$ . By choosing the target error  $E_{\text{tar}}$  to be equal to  $E_m$ , the error at future stage  $m$ , one can better assess the effectiveness of the

Aitken-based predictions. In the car problem, Aitken-based methods predict that the network needs to be trained for  $\widehat{s}^a = 235$  and  $\widehat{s}^i = 156$  additional number of stages to reach the desired target error of  $E_{\text{tar}} = 3.6576 = E_{450}$  from the current stage of  $n = 300$  ( $E_{300} = 4.3365$ ). The last two columns of Table 3 give the percent ratio of predicted decrease in errors to actual decrease in error defined as

$$\widehat{\phi}_s^g = 100 \times \frac{|\widehat{s}^g - s|}{s}$$

with  $g = a$  or  $i$  corresponding to the Aitken  $\delta^2$  or its invariant one respectively. For this problem,  $\widehat{\phi}_s^a = 56.7\%$  and  $\widehat{\phi}_s^i = 9.3\%$ . It is clear from Table 3 that the number of stages needed for target error reduction,  $\widehat{m}^a$ , can be predicted more accurately at advanced stages of the network training than at moderate stages of the training. However, the predictions based on the invariant extended-Aitken are more accurate over the whole range of training.

Note that if the chosen  $E_{\text{tar}}$  value is less than the Aitken predicted final error, then naturally one cannot predict the number of stages using the Aitken formula (see Section VI.B) even in an infinite number of epochs. For example, in the alphanumeric problem the chosen  $E_{\text{tar}} = 0.2183$ , which is less than  $\widehat{E}_{150}^a = 0.2304$ , is outside the Aitken prediction range and hence not possible to predict the desired number of stages. This is indicated by a “–” in Table 3. For example, one can never reach the target error of 0.2 since the Aitken predicted error at infinity is greater than the final predicted error of  $\widehat{E}_n^a = 0.2$ .

### C. Prediction of Error at Infinity

The training error at infinity can be predicted using either of the Aitken-based methods. Note that  $\widehat{\alpha}_n^i$  is strictly greater than zero is a necessary condition for the invariant extension to be stable. (See Section VIII for the conditions under which one can use the invariant extended-Aitken method to make predictions.) It is clear from Table 4 that the predicted training error at infinity via the invariant acceleration method is approach-

ing to the local minimum faster than the corresponding one via the Aitken acceleration method.

#### *D. Comparison of Error with Aitken-based Acceleration Schemes*

In this section, we will compare the behavior of the Aitken and its invariant extension in the alphanumeric problem. To this end, we plot  $\log_{10}$  error versus  $\log_{10} n$  to examine the numerical convergence of the Aitken sequences, where the error is  $E_n$ ,  $\hat{E}_n^a$  or  $\hat{E}_n^i$ . The  $\log_{10}$  error represents accuracy reached by a method in a given number of epochs, if the minimum to which the sequence appears to be converging is zero. It is clear from Fig. 3 that an invariant extended-Aitken accelerated estimator has attained an accuracy of 0.01 in approximately 400 stages, whereas to reach the same accuracy, Aitken sequence needs much more than 600 stages.

A second feature of these plots is the near linearity of the original error sequence on this scale. Note that a true log-geometric sequence converging to zero is linear on this scale, with slope equal to  $-\alpha$ , the negative of the rate index. Thus the linearity of these plots is diagnostic for the sequence being log-geometric.

#### *E. Stopping Rule and Assessment of Accuracy via the Aitken-based Acceleration Schemes*

We now demonstrate, through an example, that the rule based on the invariant extended-Aitken acceleration generally provides better accuracy for a slowly converging sequence such as the log-geometric sequence. Table 5 presents the number of stages required (and corresponding error value) by the network to reach the desired accuracy of  $\tau$  based on several stopping criteria. From the alphanumeric problem, it is clear that if one uses the “naive” stopping criterion then one would interpret that one has achieved accuracy  $\tau$  which indeed is misleading as the invariant extended-Aitken based stopping rule suggests that the network needs to be trained for approximately fifteen times as many stages to reach the same accuracy. Similarly with the Aitken-based stopping rule

one would falsely stop the training process at stage  $n = 333$ . This is because, the Aitken sequence – exact only for geometric sequence and hence is a poor fit – predicts that one is near the solution. The invariant extended-Aitken sequence which is exact for log-geometric sequence gives a better fit to the original error sequence.

*Remark VII.1:* From our empirical evidence, we conclude that the on-line version of the BP algorithm with a fixed or varying learning rate (not shown here) seems to generate an error sequence that satisfies the conditions required for the application of Aitken-based methods.

## VIII. DIAGNOSTICS

In this section, we present some diagnostics to assess the reliability of predictions. One should hold off making predictions using the Aitken-based methods until the network has finished the initial stages of training. As evident from the examples, this initial stage is problem dependent. If the number of training patterns are small to moderate in size, one achieves stability after about 10,000 epochs, whereas in the case of large size stability is reached after about 30,000 epochs.

One can evaluate the fit of the acceleration methods via the respective prediction formulae in a “backward prediction” framework. If the approximations in (7) or (13) are accurate then formulae (15) and (16) should be accurate for prediction at  $m$ ; for any  $m$  larger or smaller than the current stage  $n$ . We recommend an acceleration method that has the least error in backward prediction. Perhaps a histogram of backward error predictions (or forward ones) over time could be used to provide a “standard error” for forward predictions, and even a bias correction.

*Choosing an Acceleration Scheme:* As a first step, make a short-term *backward prediction of error* using both the Aitken and its invariant one. Compare the resulting predictions with that of the original error values obtained during the earlier part of training. If Aitken gives better backward predictions than the invariant one, then declare the former

as a winner for making long term forward predictions.

We have tested the performance of this backward prediction rule in several problems. The results were very similar to Fig. 4 and hence are not reported here. In all the problems, the invariant extended-Aitken acceleration method gave better backward predictions compared to that of the Aitken and hence is in confirmatory with our forward predictions. From this empirical evidence, it appears that one can use the backward prediction methodology for choosing a reliable acceleration scheme to make forward predictions.

Other simple diagnostic checks for each acceleration scheme are outlined below:

*Aitken Case:* One should use the Aitken acceleration scheme to make predictions if the (i) number of epochs,  $n$ , is sufficiently large so that

$$\frac{E_n - E_{n-1}}{E_{n-1} - E_{n-2}} \approx \lambda \approx \frac{E_{n+1} - E_n}{E_n - E_{n-1}}$$

holds and (ii)  $\hat{\lambda}_n$  obtained from (5) is constant over a wide range of  $n$  so that the estimated asymptotic rate constant is stabilized.

*Invariant Extended-Aitken Case:* For sufficiently large  $n$ , one can make valid predictions via the invariant extended-Aitken acceleration method if (i)  $\hat{\alpha}_n^i$  is stable over a certain range; i.e., it is either steadily increasing or decreasing to a constant; (ii)  $\hat{\alpha}_n^i$  is strictly positive; and (iii)  $(\hat{E}_n^a - \hat{E}_{n-1}^a)/(E_n - E_{n-1})$  is small. In essence, the estimated  $\alpha$  provides a check of the underlying hypothesis. Successive estimates of  $\hat{\alpha}_n^i$  that steadily increase indicate an error sequence that may be exponential in  $(-n)$ , say, rather than polynomial in  $n^{-1}$ . The rate parameter was very stable through out the iterative process in our examples. The rate parameter information suggests that in both the examples, the original error sequence has  $\kappa/(n + s)^\alpha$  behavior as indicated in Section V.

## IX. DISCUSSION

The BP algorithm used in training the NNs is very slow compared to the conjugate-gradient or quasi-Newton methods, however, speed is not the ultimate issue to consider as the storage requirements, reliability of convergence and the problem of local minima are also important (see Section I.B). Probably there is no single best approach and an optimal choice must depend on the problem and on the design criteria (see also Saarinen et al. (1992)).

We introduced the ideas of Aitken  $\delta^2$  method (Section III) to accelerate the rate of convergence of an error sequence obtained via the BP algorithm. We further developed and investigated an extended-Aitken acceleration method and its invariant version, proposed by Pilla & Lindsay (2000), for accelerating the convergence of a log-geometric sequence while obtaining its shift invariance properties (Sections IV and V). In addition, we proposed different convergence measures including the stopping criterion based on the Aitken acceleration estimator and its invariant extension (Section VI). The latter technique is best suited for a sequence converging more slowly than a geometric convergence. For example, when one is marching through a plateau region of the error surface, then many epochs of the algorithm may be required to produce a significant reduction in the error performance of the network, which in turn may make it computationally excruciating. The different measures developed in this paper prove useful in reducing the computational burden in these situations. The assessment of the convergence was done while training multilayer NNs with the BP algorithm in several different problems (Section VII). We hope that the methods developed in this article will prove to be useful for assessing convergence of NNs in applications such as data mining, sensor-assisted monitoring, forecasting and diagnosis.

In applications such as, finite mixture problems, random effects models, latent class models, image analysis and missing data problems, it is a common practice to use the

Expectation-Maximization algorithm (Dempster et al., 1977) to draw the appropriate inferences but lack good convergence criteria. Future work involves further investigations to assess the performance of the invariant extended-Aitken acceleration technique in such types of problems.

#### APPENDIX A

*Proof of Theorem IV.3:* One can rewrite equation (6) using the residual  $r_n = (E_n - \widehat{E})$ , for  $n > 0$ , as

$$\widehat{E}_n^a = \frac{(r_n + \widehat{E})(r_{n-2} + \widehat{E}) - (r_{n-1} + \widehat{E})^2}{(r_n + \widehat{E}) - 2(r_{n-1} + \widehat{E}) + (r_{n-2} + \widehat{E})}.$$

After simplification, the above equation reduces to

$$\widehat{E}_n^a - \widehat{E} = \frac{(r_n r_{n-2} - r_{n-1}^2)}{(r_n - 2r_{n-1} + r_{n-2})}.$$

Since  $\xi_n = \rho_{n-1}\xi_{n-1}$ , where  $\rho_n = (1 + 1/n)^{-\alpha}$ , it follows from assumption A2 that  $r_n = \xi_n[1 + o(1)] = \rho_{n-1}\xi_{n-1}[1 + o(1)] = \rho_{n-1}r_{n-1}$  and  $r_{n-1} = \rho_{n-2}r_{n-2}$ . Thus the above equation simplifies to

$$\left( \frac{\widehat{E}_n^a - \widehat{E}}{E_n - \widehat{E}} \right) = \frac{(\rho_{n-1} - \rho_{n-2})}{\rho_{n-1}(\rho_{n-1}\rho_{n-2} - 2\rho_{n-2} + 1)}. \quad (17)$$

Using Lemma A.1 below, one can write

$$\rho_n = \left(1 + \frac{1}{n}\right)^{-\alpha} = 1 - \frac{\alpha}{n} + \frac{\alpha(\alpha + 1)}{2n^2} + o\left(\frac{1}{n^2}\right)$$

by considering the leading terms only since the remainder term  $\sum_{i=3}^{\infty} R_i(1/n) \rightarrow 0$  as  $n \rightarrow \infty$ . The numerator and denominator of (17) simplify to  $\alpha \cdot [(n-1)(n-2)]^{-1} + o(n^{-2})$  and  $\alpha(\alpha + 1) \cdot [(n-1)(n-2)]^{-1} + o(n^{-2})$  respectively. Therefore,

$$\left( \frac{\widehat{E}_n^a - \widehat{E}}{E_n - \widehat{E}} \right) = (\alpha + 1)^{-1} [1 + o(1)]$$

tends  $(\alpha + 1)^{-1}$  as  $n \rightarrow \infty$ . ■

*Lemma A.1:* One can write  $\rho_n$  as

$$\rho_n = \left(1 + \frac{1}{n}\right)^{-\alpha} = 1 - \frac{\alpha}{n} + \frac{\alpha(\alpha+1)}{2n^2} + \sum_{i=3}^{\infty} R_i \left(\frac{1}{n}\right),$$

where

$$\left| R_i \left(\frac{1}{n}\right) \right| = \left| (-1)^{i+1} \frac{\alpha}{1} \cdot \frac{(\alpha+1)}{2} \cdots \frac{(\alpha+i)}{(i+1)} \cdot \frac{1}{n^{i+1}} \right| \rightarrow 0 \quad \text{as } n \rightarrow \infty. \quad (18)$$

*Proof:* We have from the Taylor's series expansion

$$f(x) = f(0) + x \cdot f'(0) + \frac{x^2}{2!} \cdot f''(0) + \cdots + \frac{x^{i+1}}{(i+1)!} \cdot f^{(i+1)}(0) + \cdots$$

with the Lagrange remainder term as

$$|R_i(x)| = \left| \frac{x^{i+1}}{(i+1)!} \cdot f^{(i+1)}(x) \right|. \quad (19)$$

Equation (18) follows from (19) with  $x = n^{-1}$ . Since  $(\alpha+i)/(i+1) \rightarrow 1$  and  $n^{-(i+1)} \rightarrow 0$  as  $i \rightarrow \infty$ , we have  $|R_i(n^{-1})| \rightarrow 0$  as  $n \rightarrow \infty$  for any fixed  $i$ .  $\blacksquare$

*Lemma A.2:* Estimated rate index parameter  $\hat{\alpha}_n \rightarrow \alpha$  as  $n \rightarrow \infty$ .

*Proof:* One can rewrite equation (10) as

$$\hat{\alpha}_n = \left[ \frac{E_{n-1} - E_{n-2}}{\hat{\mathcal{A}}_{n-1} - \hat{\mathcal{A}}_{n-1}} - 1 \right]^{-1}. \quad (20)$$

One can rewrite  $(E_{n-1} - E_{n-2}) = (r_{n-1} - r_{n-2}) = r_{n-2}(\rho_{n-2} - 1)$ . By using Lemma A.1 and ignoring the leading terms, we obtain

$$\begin{aligned} \hat{\mathcal{A}}_{n-1} &= \frac{r_{n-2}(\rho_{n-1} - 1)(\rho_{n-2} - 1)\rho_{n-2}}{(\rho_{n-1}\rho_{n-2} - 2\rho_{n-2} + 1)} \\ &= r_{n-2} \left( \frac{\alpha}{\alpha+1} \right) [1 + o(1)] \end{aligned}$$

and similarly

$$\hat{\mathcal{A}}_{n-2} = r_{n-3} \left( \frac{\alpha}{\alpha+1} \right) [1 + o(1)].$$

After simplification via  $r_{n-2} = \rho_{n-3}r_{n-3}$ , we have

$$\widehat{\mathcal{A}}_{n-1} - \widehat{\mathcal{A}}_{n-2} = r_{n-2} \left( \frac{\alpha}{\alpha + 1} \right) \frac{(\rho_{n-3} - 1)}{\rho_{n-3}} [1 + o(1)].$$

Thus

$$\left( \frac{E_{n-1} - E_{n-2}}{\widehat{\mathcal{A}}_{n-1} - \widehat{\mathcal{A}}_{n-2}} \right) = \left( \frac{\alpha + 1}{\alpha} \right) \left( \frac{\rho_{n-2} - 1}{\rho_{n-3} - 1} \right) \rho_{n-3} [1 + o(1)]$$

tends to  $(\alpha + 1)/\alpha$  as  $n \rightarrow \infty$ , since  $(\rho_{n-2} - 1)/(\rho_{n-3} - 1) = (n - 3)/(n - 2)[1 + o(1)] \rightarrow 1$  and  $\rho_{n-3} \rightarrow 1$  as  $n \rightarrow \infty$ . Therefore, from equation (20) we have  $\widehat{\alpha}_n \rightarrow \alpha$  as  $n \rightarrow \infty$ . ■

*Proof of Theorem IV.4:* (i) One can rewrite equation (11) as

$$\left( \frac{\widehat{E}_n^\varepsilon - \widehat{E}}{E_n - \widehat{E}} \right) = \frac{r_{n-1}}{r_n} - \left( \frac{\widehat{\alpha}_n + 1}{\widehat{\alpha}_n} \right) \left( \frac{n - 1}{n} \right) \frac{\widehat{\mathcal{A}}_{n-1}}{r_n}.$$

We have  $r_{n-1}/r_n = r_{n-1}/(\rho_{n-1}r_{n-1}) = 1/\rho_{n-1}$ . Similarly, we obtain

$$\begin{aligned} \frac{\widehat{\mathcal{A}}_{n-1}}{r_n} &= \frac{(\rho_{n-1} - 1)(\rho_{n-2} - 1)}{\rho_{n-1}(\rho_{n-1}\rho_{n-2} - 2\rho_{n-2} + 1)} \\ &= \frac{\alpha}{(\alpha + 1)} [1 + o(1)] \end{aligned}$$

which tends to  $\alpha/(1 + \alpha)$  as  $n \rightarrow \infty$ . Thus

$$\lim_{n \rightarrow \infty} \left( \frac{\widehat{E}_n^\varepsilon - \widehat{E}}{E_n - \widehat{E}} \right) = 0 \quad \text{for any } \alpha > 0$$

since  $1/\rho_{n-1} \rightarrow 1$ ,  $\widehat{\alpha}_n \rightarrow \alpha$  and  $(n - 1)/n \rightarrow 1$  as  $n \rightarrow \infty$ .

Part (ii) follows from part (i) and Theorem IV.3 since

$$\lim_{n \rightarrow \infty} \left( \frac{\widehat{E}_n^\varepsilon - \widehat{E}}{\widehat{E}_n^a - \widehat{E}} \right) = \lim_{n \rightarrow \infty} \frac{(\widehat{E}_n^\varepsilon - \widehat{E})/(E_n - \widehat{E})}{(\widehat{E}_n^a - \widehat{E})/(E_n - \widehat{E})} = \frac{0}{(\alpha + 1)^{-1}} = 0. \quad \blacksquare$$

## ACKNOWLEDGEMENTS

The authors gratefully acknowledge the constructive comments of an associate editor and three referees that improved the presentation of the paper significantly. Lindsay's research was partially supported by the National Science Foundation Grant DMS-9870193.

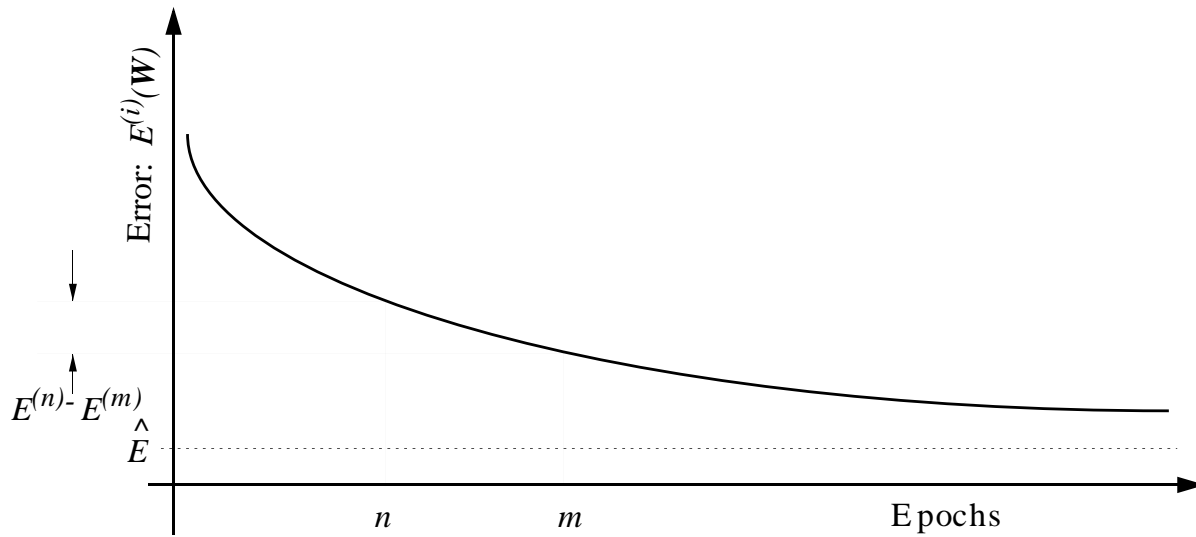
Part of the work was done while the first author was at the National Institutes of Health, Bethesda, MD, USA and was supported by the NIH fellowship.

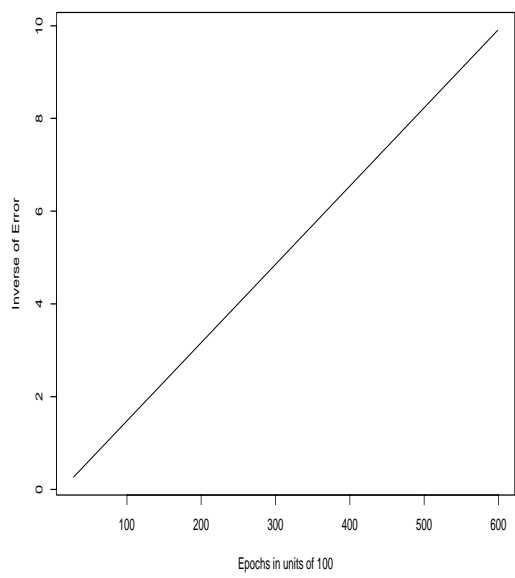
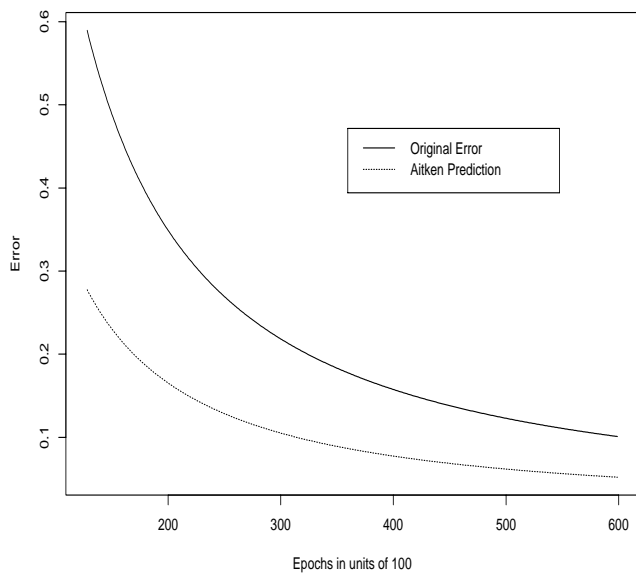
## REFERENCES

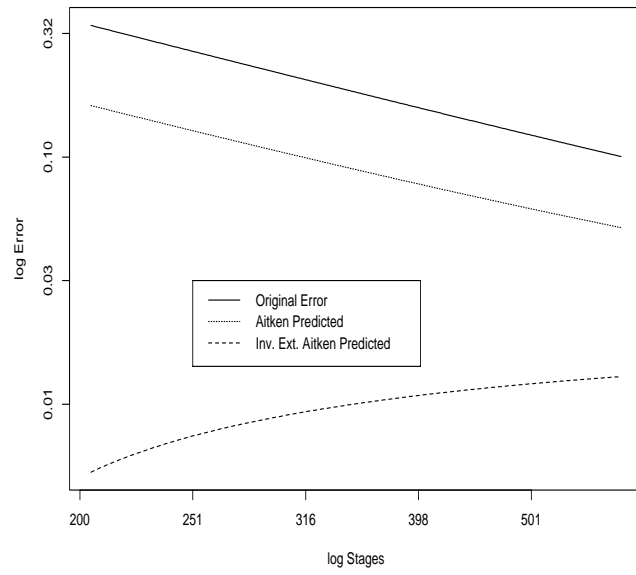
- [1] A.C. Aitken, "On Bernoulli's numerical solution of algebraic equations", *Proc. Roy. Soc. Edinburgh*, vol. 46, pp. 289–305, 1926.
- [2] D.W. Aha, "Incremental Constructive Induction: An Instance-Based Approach", *Proc. of 8th Intl. Workshop on Machine Learning*, pp. 117–121, Evanston, IL, 1991.
- [3] S. Amari, N. Murata, K.R. Muller, M. Finke and H. Yang, "Statistical Theory of Overtraining—Is Cross-Validation Asymptotically Effective?" *Advances in Neural Information Processing Systems*, vol. 8, pp. 176–182, Cambridge, MA: MIT Press, 1996.
- [4] J.P. Bigus, *Data Mining with Neural Networks*, New York: McGraw-Hill, 1996.
- [5] P. Bjørstad, G. Dahlquist and E. Grosse, "Extrapolation of Asymptotic Expansions by a Modified Aitken  $\delta^2$ -Formula", *BIT*, 21, pp. 56–65, 1981.
- [6] M. Bohanec and V. Rajkovic, "Knowledge Acquisition and Explanation for Multi-Attribute Decision Making", *Proc. of 8th Intl. Workshop on Expert Systems and Their Applications*, pp. 59–78. Avignon, France, 1988.
- [7] D.Böhning, "Acceleration Techniques in Fixed-Point Methods for Finding Percentage Points", *Statistics and Computing*, vol. 3, pp. 1–5, 1993.
- [8] D. Coomans, M. Broeckaert, M. Jonckheer D.L. and Massart, "Comparison of Multivariate Discriminant Techniques for Clinical Data – Application to the Thyroid Functional State", *Meth. Inform. Med.*, Vol. 22, pp. 93-101, 1983.
- [9] J.P. Delahaye, *Sequence transformations*, New York: Springer, 1980.
- [10] N.R. Draper and R.C. Van Nostrand, "Ridge regression and James Stein estimators: Review and comments", *Technometrics*, vol. 21, pp. 451–466, 1979.
- [11] S.E. Fahlman, "Faster-Learning Variations on Backpropagation: An Empirical Study", *Proc. of 1988 Connectionist Models Summer School*, pp. 38–51. Pittsburgh, PA, 1989.
- [12] L.V. Fausett, *Fundamentals of Neural Networks*, Englewood Cliffs, NJ: Prentice Hall, 1994.
- [13] R.F. Gunst and R.L. Mason, *Regression analysis and its applications: A data oriented approach*, New York: Marcel Decker, 1980.
- [14] S. Haykin, *Neural Networks: A Comprehensive Foundation*, Upper Saddle River, NJ: Prentice Hall, 1999.
- [15] R. Hecht-Nielsen, *Neurocomputing*, Reading, MA: Addison-Wesley, 1990.
- [16] R.A. Jacobs, "Increased Rates of Convergence Through Learning Rate Adaptation", *Neural Networks*, vol. 1, pp. 295–307, 1988.
- [17] B. Jones, "A Note on the  $T_{+m}$  Transformation", *Nonlinear Analysis, Theory, Methods and Applications*, vol. 6, pp. 303–305, 1982.
- [18] A. Kanda, S. Fujita and T. Ae, "Acceleration by Prediction for Error Backpropagation Algorithm of Neural Network", *Systems and Computers in Japan*, Vol. 25, pp. 78–87, 1994.
- [19] D. Kincaid and W. Cheney, *Numerical Analysis Mathematics of Scientific Computing*, Pacific Grove, CA: Brooks/Cole Publishing Company, 1990.
- [20] B.G. Lindsay, *Mixture Models: Theory, Geometry and Applications*. NSF-CBMS Regional Conference Series in Probability and Statistics, vol. 5. Hayward, CA: Institute of Mathematical Statistics, 1995.

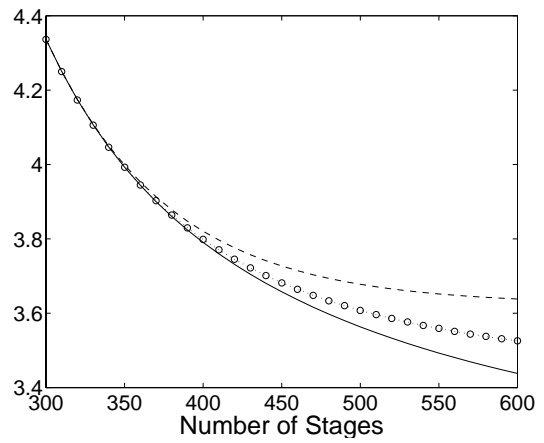
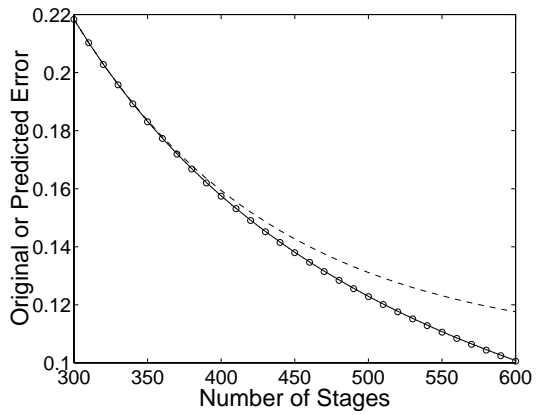
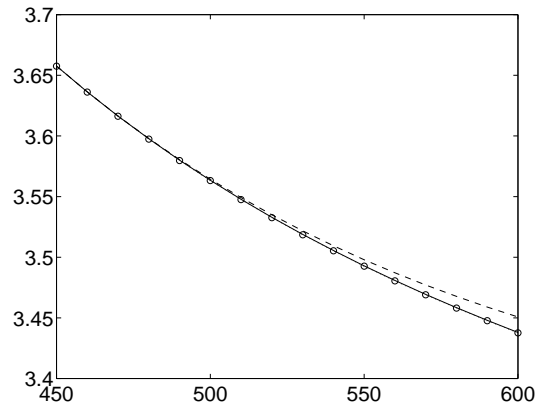
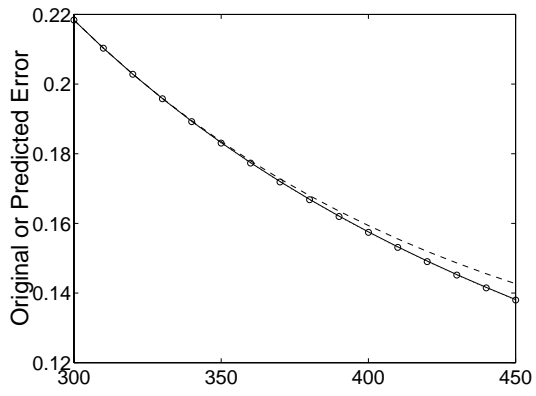
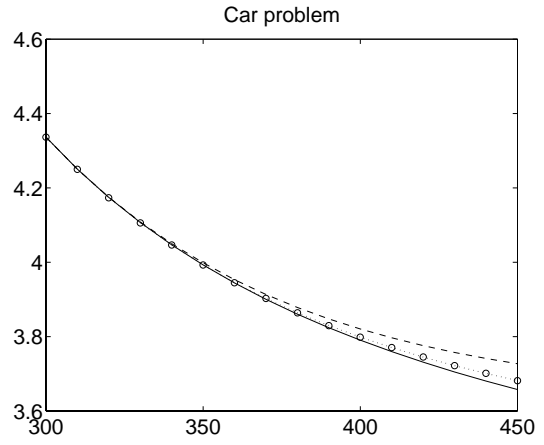
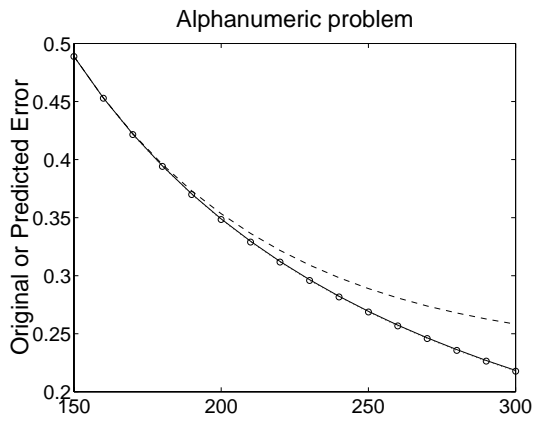
- [21] M. Pfister and R. Rojas “Speeding-Up Backpropagation — A Comparison of Orthogonal Techniques”, *Proc. of Intl. Joint Conf. on Neural Networks*, vol. 1, pp. 517–523. Japanese Neural Network Society, Nagoya, Japan, 1993.
- [22] R.S. Pilla and B.G. Lindsay, “Assessing Convergence in High-Dimensional Optimization Problems: Application to Neural Networks”, *Proc. of the Second International ICSC Symposium on Neural Computation (NC’2000)*, May 23–26, 2000 in Berlin, Germany.
- [23] R.S. Pilla and B.G. Lindsay, “Alternative EM methods for nonparametric finite mixture models”, *Biometrika*, to appear, 2001.
- [24] R.S. Pilla, S.V. Kamarthi and B.G. Lindsay, “Application of Aitken Acceleration Method to Neural Networks. *Computing Science and Statistics: Proc. of Twenty-Seventh Symposium on Interface*, vol. 27, pp. 332–336, 1995.
- [25] R.S. Pilla, S.V. Kamarthi and B.G. Lindsay, “Convergence Behavior of an Iterative Process: Application to Neural Networks,” *Intelligent Engineering Systems Through Artificial Neural Networks*, C.H. Dagli, M. Akay, C.L.P. Chen, B.R. Fernandez and J. Ghosh, (eds.), vol. 5, pp. 147–152. New York: ASME Press, 1995.
- [26] R.S. Pilla, S.V. Kamarthi and B.G. Lindsay, “An Extended Aitken Acceleration Method for Assessing Convergence of Multilayer Neural Networks,” *Intelligent Engineering Systems Through Artificial Neural Networks*, Dagli, Buczak, Ghosh, Embrechts, Ersoy and Kercel (eds.). In press. New York: ASME Press, 2000.
- [27] Y.M. Pirez and D. Sarkar, “Backpropagation Algorithm with Controlled Oscillation of Weights”, *Proc. of IEEE Intl. Conf. on Neural Networks*, vol. 1, pp. 21–26, IEEE, San Francisco, CA, 1993.
- [28] S. Roy, “Near-Optimal Dynamic Learning Rate for Training Backpropagation Neural Networks”, *Science of Artificial Neural Networks*, D.W. Ruck, eds., vol. II, pp. 277–283, Society of Industrial and Applied Mathematics, Bellingham, WA, 1993.
- [29] D.E. Rumelhart and J.L. McClelland, *Parallel Distributed Processing: Explorations in the Microstructure of Cognition*, vol. 1, Cambridge, MA. MIT Press, 1986.
- [30] R. Salomon and J.L. van Hemmen, “Accelerating Backpropagation Through Dynamic Self-Adaptation”, *Neural Networks*, vol. 9, pp. 589–601, 1996.
- [31] A.J. Shepherd, *Second-Order Methods for Neural Networks*, New York: Springer, 1997.
- [32] S. Saarinen, R.B. Bramley and G. Cybenko, “Neural Network, Backpropagation and Automatic Differentiation,” *Automatic Differentiation of Algorithms: Theory, Implementation, and Application*, A. Griewank and G.F. Corliss, eds., Philadelphia, SIAM, pp. 31–42, 1992.
- [33] F.M. Silva and L.B. Almeida, “Acceleration Techniques for the Backpropagation Algorithm”, *Neural Networks: EURASIP Workshop*, L.B. Almeida, C.J. Wellekens, eds., pp. 110–119. Berlin: Springer-Verlag, 1990.
- [34] M. Stone, “Cross-validatory choice and assessment of statistical predictions”, *Journal of the Royal Statistical Society, Series B*, vol. 36, pp. 111–147, 1974.
- [35] G. Tesauro, Y. He and S. Ahmad, “Asymptotic Convergence of Backpropagation”, *Neural Computation*, vol. 1, pp. 382–391, 1989.
- [36] D.M. Titterton, A.F.M. Smith and U.E. Makov, *Statistical Analysis of Finite Mixture Distributions*. New York: John Wiley & Sons, 1985.

- [37] M. Towsey, D. Alpsan and L. Sztriha, "Training a Neural Network with Conjugate Gradient Methods", *Proc. of IEEE Intl Conf. on Neural Networks*, vol. 1, pp. 373–378, IEEE, Perth, Australia, 1995.
- [38] J.M. Twomey and A.E. Smith, "Committee Networks by Resampling", *Intelligent Engineering Systems Through Artificial Neural Networks*, C.H. Dagli, M. Akay, C.L.P. Chen, B.R. Fernandez and J. Ghosh, (eds.), vol. 5, pp. 153–158. New York: ASME Press, 1995.
- [39] M.K. Weir, "A Method for Self-Determination of Adaptive Learning Rates in Backpropagation", *Neural Networks*, vol. 4, pp. 371–379, 1991.
- [40] X. Yu, N.K. Loh and W.C. Miller, "A New Acceleration Technique for The Backpropagation Algorithm", *Proc. of IEEE Intl. Conf. on Neural Networks*, vol. 3, pp. 1157–1161, IEEE, San Francisco, CA, 1993.









Original Error     
  Aitken Predicted     
 ○ Inv. Ext. Aitken Predicted

Fig. 1. An exponentially decreasing function.

Fig. 2. (a) Comparison of the original and Aitken predicted sequences and (b) Graph of  $E_n^{-1}$  versus  $n$ .

Fig. 3. Comparison of  $E_n$ ,  $\hat{E}_n^a$  and  $\hat{E}_n^i$  sequences.

Fig. 4. Predicted errors based on the Aitken  $\delta^2$  and its invariant extension.

<b>Problem</b>	<b>I - Nodes</b>	<b>H - Nodes</b>	<b>O - Nodes</b>	<b>Weights</b>	<b># Patterns</b>	<b><math>\eta</math> (Gain)</b>
<b>Alphanumeric</b>	35	12	35	887	36	0.01
<b>Car</b>	6	15	2	137	1728	0.01

Problem	$n$	$m$	$E_m$	$\hat{E}_{n,m}^a$	$\hat{E}_{n,m}^i$	$\hat{\psi}_m^a$	$\hat{\psi}_m^i$
Alphanumeric	150	200	0.3489	0.3533	0.3483	0.96	1.00
		250	0.2692	0.2889	0.2686	0.91	1.00
		300	0.2183	0.2582	0.2176	0.85	1.00
	300	350	0.1831	0.1835	0.1830	0.99	1.00
		400	0.1575	0.1593	0.1574	0.97	1.00
		450	0.1381	0.1426	0.1379	0.94	1.00
	300	500	0.1229	0.1311	0.1227	0.91	1.00
		550	0.1107	0.1231	0.1105	0.88	1.00
		600	0.1008	0.1175	0.1005	0.85	1.00
Car	300	350	3.9928	3.9984	3.9927	0.98	1.00
		400	3.7904	3.8206	3.7986	0.94	0.98
		450	3.6576	3.7270	3.6818	0.89	0.96
	450	500	3.5634	3.5643	3.5631	0.99	1.00
		550	3.4928	3.4979	3.4925	0.96	1.00
		600	3.4379	3.4507	3.4376	0.94	1.00
	300	500	3.5634	3.6778	3.6078	0.85	0.94
		550	3.4928	3.6519	3.5589	0.81	0.92
		600	3.4379	3.6382	3.5255	0.77	0.90

Problem	Current Error ( $E_n$ )	$m$	$E_{tar} \equiv E_m$	s	$\hat{s}^a$	$\hat{s}^i$	$\hat{\phi}_s^a$ (%)	$\hat{\phi}_s^i$ (%)	
Alphanumeric	0.4888 ( $E^{150}$ )	200	0.3489	50	52	49	4.0	2.0	
		250	0.2692	100	127	99	27.0	1.0	
		300	0.2183	150	–	149	–	0.6	
	0.2183 ( $E^{300}$ )	350	0.1831	50	50	49	0.0	2.0	
		400	0.1575	100	104	99	4.0	1.0	
		450	0.1381	150	167	149	11.3	0.6	
	0.2183 ( $E^{300}$ )	500	0.1229	200	251	199	25.5	0.5	
		550	0.1107	250	407	249	62.8	0.4	
		600	0.1008	300	–	298	–	0.6	
	Car	4.3365 ( $E^{300}$ )	350	3.9928	50	51	49	2.0	2.0
			400	3.7904	100	112	102	12.0	2.0
			450	3.6576	150	235	156	56.7	9.3
3.6576 ( $E^{450}$ )		500	3.5634	50	50	49	0.0	2.0	
		550	3.4928	100	104	99	4.0	1.0	
		600	3.4379	150	167	149	11.3	0.6	
4.3365 ( $E^{300}$ )		500	3.5634	200	–	244	–	22.0	
		550	3.4928	250	–	375	–	50.0	
		600	3.4379	300	–	788	–	162.6	

Stage $n$	Alphanumeric		Car	
	$\hat{E}_n^a$	$\hat{E}_n^i$	$\hat{E}_n^a$	$\hat{E}_n^i$
200	0.1653	0.0050	2.2114	4.1909
250	0.1285	0.0073	3.5880	4.0044
350	0.0890	0.0100	3.5062	3.1080
450	0.0686	0.0115	3.3346	2.9473
550	0.0563	0.0125	3.2457	2.9599

Problems	$\tau$	Stages $n$ and error $E_n$ to reach					
		$ E_n - E_{n-1}  < \tau$		$ \hat{E}_n^a - E_n  < \tau$		$ \hat{E}_n^i - E_n  < \tau$	
		$n$	$E_n$	$n$	$E_n$	$n$	$E_n$
<b>Alphanumeric</b>	0.1	36	2.7668	333	0.1938	543	0.1123
<b>Car</b>	0.5	46	23.2830	346	4.0137	575	3.4638

TABLE I

NETWORK ARCHITECTURE AND TRAINING PARAMETERS.

TABLE II

COMPARISON OF THE PREDICTED ERRORS.

TABLE III

COMPARISON OF THE PREDICTED NUMBER OF STAGES.

TABLE IV

COMPARISON OF THE FINAL PREDICTED ERRORS.

TABLE V

COMPARISON OF DIFFERENT STOPPING CRITERIA.