Effects of Learning Parameters on Independent Component Analysis Learning Procedure

K. Chinnasarn and C. Lursinsap
Advanced Virtual and Intelligent Computing Center (AVIC)
Department of Mathematics
Chulalongkorn University
Bangkok 10330, Thailand
E-mail: krisana@buu.ac.th and lchidcha@chula.ac.th

Abstract

Although Independent Component Analysis (ICA) is effective for blind source separation of a set of unknown sources of signals, its convergent analysis time is rather lengthy due to the classical weight adjusting procedure. In this paper, we propose a technique to speed up this analysis time by introducing two learning parameters, a learning rate, η, and a momentum term, β. The values of these two parameters are dynamically adjusted. The success of this blind source separation is measured in terms of mutual information with the probability density functional approximation under Gram-Charlier Expansion. Our technique is tested on some benchmark examples. The separation outcomes are the same as the others's but our analysis time is significantly reduced.

1 Introduction

The application of Independent Component Analysis (ICA) or Blind Source Separation (BSS) covers several essential areas such as speech recognition, data communication, sensor signal processing, and medical science [1]. The problem of ICA or BSS concerns the techniques for separating a mixed source signals with priori unknown information related to their original occurences. The only known information provided are the number of signal sources and the statistical assumptions on their expected signal values at any time. Various separating algorithms based on the statistical cost functions such as Kullback-Leiber divergence and maximum likelihood estimator [1, 2, 4, 6, 7] are introduced. The performance on separability of these algorithms depends upon the selected activation function and, also, the appropriate cost function. The minimization of a cost function can be effectively acheived by using a supervised neural network. However, the convergent time, in some cases, cannot be tolerated. Yu [8] and Dai [3] used the generalized delta rule with a learning rate η and a momentum β to speed up the convergence. The approproate values of η and β depend on the applications, experiments, and researcher's experience. A large η can accelerate the learning procedure but can cause a local minima solution or a divergence. The momentum rate parameter is designed to smooth the error oscillation and reduce the number of iterations for convergence. The values of the learning rate and the momentum are fixed throughtout the learning period. Amari [1] and Pun [7] improved the convergent speed by varying the learning rate η from 0.1 to 0.9 during the learning period. The value of the learning rate at the current iteration step is computed by dividing the learning rate from the previous iteration step.

In this paper, we improve the convergent speed of Amari's by appropriately selecting the divisor of each learning rate and momentum. We organize our paper into term of problem of BSS, How to test an Independence of Signal, BSS Learning algorithms, our purposed algorithm, Simulation, and conclusion.

2 Blind Source Separation Problem

Let us consider a set of m unknown random source signals, $\mathbf{s}(t)$, which are mutually independent at a time t. They are defined by

$$\mathbf{s}(t) = [s_1, s_2, ..., s_m]^T \tag{1}$$

These independent signals has zero mean, $E(\mathbf{s}(t)) = 0$, and unit variance, σ^2. The signal $\mathbf{s}(t)$ is applied to a linear system by non-singular m-by-m matrix $\mathbf{A}$, called *mixing matrix*. The result is an $m - by - 1$ observation signal $\mathbf{x}(t)$ related to $\mathbf{s}(t)$ as follows

$$\mathbf{x}(t) = \mathbf{A}\mathbf{s}(t) + \mathbf{n}(t) \tag{2}$$

where

$$\begin{array}{ll} \mathbf{x}(t) & = [x_1(t), \cdots, x_m(t)]^T \\ \mathbf{n}(t) & \textit{is the additive noise} \end{array}$$

The source signal $\mathbf{s}(t)$ and mixing matrix $\mathbf{A}$ are unknown. The only prior information is the number of source signals and the number of the observation vector $\mathbf{x}(t)$. The elements of vector $\mathbf{x}(t)$'s are now dependent sources because of the mixing matrix $\mathbf{A}$. The problem is how to find a de-mixing matrix $\mathbf{W}$, which is sometime called an inversion matrix of $\mathbf{A}$, for estimating the source signals $\mathbf{s}(t)$. Let $\mathbf{y}(t)$ be the estimated signals of $\mathbf{s}(t)$. The value of $\mathbf{y}(t)$ can be written as follows

$$\begin{aligned} \mathbf{y}(t) &= \mathbf{W}\mathbf{x}(t) \\ \textit{or} & \\ \mathbf{y}(t) &= \mathbf{W}\mathbf{A}\mathbf{s}(t) = \mathbf{A}^{-1}\mathbf{A}\mathbf{s}(t) = \mathbf{I}\mathbf{s}(t) = \mathbf{s}(t) \end{aligned} \tag{3}$$

where $\mathbf{y}(t) = [y_1(t), \cdots, y_m(t)]^T$, $\mathbf{A}^{-1}$ is the inverse matrix of $\mathbf{A}$, and $\mathbf{I}$ is an identity matrix. The procedure for estimating signal set $\mathbf{s}(t)$ is illustrated in Figure 1.

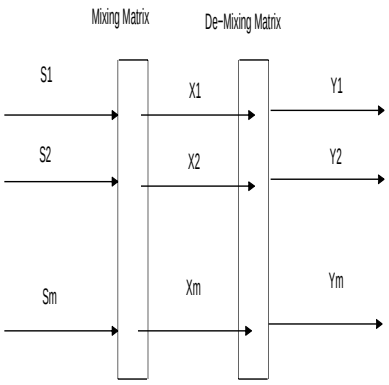

Figure 1: ICA Structure.

3 Independence of Signals

To obtain the completely separated any two signals y_i and y_j, the values of y_i and y_j must be statistically independent at all times. There are various statistical independence tests that can be used in this context. Two random variables y_i and y_j are said to be statistically independent if the value of y_i does not effect on the value of y_j, and vice versa [5]. The independence of sources can be considered in terms of probability density function. We denote by $p(y_i, y_j)$ the joint probability density function of y_i and y_j, and $p_i(y_i)$ the marginal probability density function of y_i as follows:

$$p_i(y_i) = \int_0^\infty p(y_i, y_j) dy_j$$

$$\begin{aligned} & \textit{and} \\ p(y_i, y_j) &= p_i(y_i) p_j(y_j) \end{aligned} \tag{4}$$

Practically, it is not easy to test whether two signals y_i and y_j are independent by using $p(y_i, y_j)$, $p_i(y_i)$, and $p_j(y_j)$. The easier testing is by considering their correlation. Two random variables y_i and y_j are said to be uncorrelated if their covariance is zero. The covariance can be computed in terms of the correlated expected values and the multiplication of the expected values of y_i and y_j as follows

$$E[y_i y_j] - E[y_i]E[y_j] = 0 \tag{5}$$

If the variables are independent, they are also uncorrelated. On the other hand, uncorrelatedness does not imply independence.

4 Learning Algorithm

The original frame work of ICA is to minimize the dependency among the estimated signals y_i, $i = 1, \ldots, n$. The dependency is measured by the Kullback-Leibler divergence between the joint probability density function of $\mathbf{y}(t)$ $p(\mathbf{y})$ and the factorial of marginal distribution of outputs $\prod p_i(y_i)$:

$$D(\mathbf{W}) = \int p(\mathbf{y}) \log \frac{p(\mathbf{y})}{\prod p_i y_i} d\mathbf{y} \tag{6}$$

where $p_i(y_i)$ is the marginal probability density function of output $\mathbf{y}(t)$. Amari [1] showed that Kullback-Liebler divergence $D(\mathbf{W})$ can be calculated from the average Mutual Information (MI) of y_i as follows:

$$D(\mathbf{W}) = -h(\mathbf{y}) + \sum_{i=1}^{n} h(y_i) \tag{7}$$

where

$$\begin{aligned} h(\mathbf{y}) &= -\int p(\mathbf{y}) \log p(\mathbf{y}) dy \\ h(y_i) &= -\int p(\mathbf{y}) \log p_i y_i dy_i \end{aligned}$$

Again, our target is to minimize the Kullback-Leibler divergence between the output y_i by estimating the de-mixing matrix $\mathbf{W}$. The problem is how to estimate the matrix $\mathbf{W}$ without an information about the mizing matrix. An MI is calculated from the differential entropy of $\mathbf{y}(t)$ and their marginal entropy. We assume the output signals have maximal differential entropy, hence they are Gaussian or Normal distribution. The marginal entropy $h(y_i)$ is computed by applying the Gram-Charlier Expansion to approximate the probability density function $p_i(y_i)$ as

follows

$$p_i(y_i) \approx \alpha(y_i)\{1 + \frac{k_3^i}{3!}H_3(y_i) + \frac{k_4^i}{4!}H_4(y_i)\} \quad (8)$$

where $E[\mathbf{y}(t)] = E[\mathbf{W}\mathbf{x}(t)] = E[\mathbf{W}\mathbf{A}\mathbf{s}(t)] = 0$, $E[y_i] = 0$, $m_2^i = 1$, $k_3^i = m_3^i$, $k_4^i = m_4^i - 3$, $m_k^i = E[(y_i)^k]$ is the k^{th} order moment of y_i, $\alpha(y_i) = \frac{1}{\sqrt{2\pi}}e^{-\frac{y_i^2}{2}}$ and $H_k(y_i)$ are Chebyshev-Hermite Polynomials defined by the identity

$$(-1)^k \frac{d^k \alpha(y_i)}{dy_i^k} = H_k(y_i)\alpha(y_i) \quad (9)$$

Amari [1] and Haykin [4] have been proven that

$$\begin{aligned} h(y_i) \approx & \frac{1}{2}\log 2\pi e \\ & -\frac{(k_3^i)^2}{2*3!} - \frac{(k_4^i)^2}{2*4!} \\ & +\frac{5}{8}(k_3^i)^2 k_4^i + \frac{1}{16}(k_4^i)^3 \end{aligned} \quad (10)$$

It can be calculated by

$$-\int \alpha(y_i)\log\alpha(y_i)dy_i = \frac{1}{2}\log 2\pi e \quad (11)$$

From $\mathbf{y}(t) = \mathbf{W}\mathbf{x}(t)$, we get $H(\mathbf{y}(t)) = H(\mathbf{x}(t)) + \log[\det(\mathbf{W})]$. Applying (10) and (11) to (7), we get

$$\begin{aligned} D(\mathbf{W}) \approx & -H(\mathbf{x}) - \log[\det(\mathbf{W})] \\ & +\frac{n}{2}\log 2\pi e \\ & -\sum_{i=1}^{n}[-\frac{(k_3^i)^2}{2*3!} - \frac{(k_4^i)^2}{2*4!} \\ & +\frac{5}{8}(k_3^i)^2 k_4^i + \frac{1}{16}(k_4^i)^3] \end{aligned} \quad (12)$$

To find $\mathbf{W}$ to minimize $D(\mathbf{W})$, we differentiate $D(\mathbf{W})$ with respect to $\mathbf{W}$ as follows

$$\frac{\partial D(\mathbf{W})}{\partial(\mathbf{W})} = \eta(t)(I - f(\mathbf{y})\mathbf{y}^T)\mathbf{W}^{-T} \quad (13)$$

$\mathbf{W}$ at time $k+1$ is adjusted by the following cconstructive step

$$\mathbf{W}_{k+1} = \mathbf{W}_k + \eta_k(I - f(y_k)y_k^T)\mathbf{W}_k \quad (14)$$

where the activation function $f(\mathbf{y})$ can be defined as:

$$\begin{aligned} f(\mathbf{y}) = & \frac{3}{4}y^{11} + \frac{25}{4}y^9 - \frac{14}{3}y^7 \\ & -\frac{47}{4}y^5 + \frac{29}{4}y^3 \end{aligned} \quad (15)$$

5 Purposed Algorithm

In this paper, we revise an important experimental result of Haykin [4] and Amari [1]. Our improvement is based on these few observations of Amari's and Haykin's results. Firstly, only a small fixed step of learning rate value can make the separation of signals $\mathbf{y}$ converged but a larger learning rate values in the range of $0.5, \leq \eta \leq 0.9$ cause output signals $\mathbf{y}$ diverged. Secondly, the convergence speed can be increased by gradually reducing the learning rate until it is equal to zero. The learning rate may be initially set to any value. Thirdly, the reduction of learning rate in the current iteration step is done by dividing the learning rate from the previous iteration step, namely $\eta_{t+1} = \eta_t/1.005$. However, we find that this simple approach works well when $0.1 \leq \eta \leq 0.5$, but when $0.6 \leq \eta \leq 0.9$ the convergence speed is reduced and more iterations are required. Figures 4 and 5 shows different convergence speeds.

Instead of using a fixed divisor throughtout the learning period, we use different divisors for different learning rate. The learning rate should be divided proportionally to its value. If the learning rate is large then it should be divided by a large divisor. In addition, at each iteration k, a momentum term $\delta\mathbf{M}$ and a momentum rate β are added to adjust the weight $\mathbf{W}$. Let $\mathbf{W}_k$ be the weight $\mathbf{W}$ at iteration k and $\delta\mathbf{M}_k$ the momentum term at iteration k. The momentum term $\delta\mathbf{M}$ is adjusted by using this rule $\delta\mathbf{M} = \beta\delta\mathbf{W}$ and $\delta\mathbf{W} = \eta\mathbf{W}$. The stopping condition is defined in terms of the difference between $D(\mathbf{W}_t)$ at time t and $D(\mathbf{W}_{t-1})$ at time $t-1$.

Purposed ICA Algorithm
Input : Observed signal, $\mathbf{x}$
Output : Output signal $\mathbf{y}$ and a de-mixing weight matrix $\mathbf{W}$
begin
get an observed signal $\mathbf{x}$
i=0;
while i $\leq$ NumberOfIterations
Randomly initialize weight matrix $\mathbf{W}$
appropriate divisor $= 1.0 + 10^{-2}\eta$
Compute $\mathbf{y} = \mathbf{W}\mathbf{x}$
Compute Kullback-Liebler Divergence $D(\mathbf{W}_0)$
$\delta\mathbf{M}_0 = \mathbf{0}$
Set $t = 0$
repeat
$t = t + 1$
Compute $\delta\mathbf{W}_{t-1} = \eta(\mathbf{I} - f(\mathbf{y})\mathbf{y}^T)\mathbf{W}_{t-1}$
Compute $\mathbf{W}_t = \mathbf{W}_{t-1} + \delta\mathbf{W}_{t-1} + \delta\mathbf{M}_{t-1}$
Compute $\delta\mathbf{M}_t = \beta\delta\mathbf{W}_{t-1}$

$\eta = \frac{\eta}{appropriate\ divisor}$
Compute Kullback-Liebler Divergence $D(\mathbf{W}_t)$
until $D(\mathbf{W}_(t)) - D(\mathbf{W}_(t-1)) > \Delta KL$
$i = i + 1$
End While
end.

6 Simulation

In this paper, we simulate our algorithm on the computer using three-synthesis signal, a random mixing matrix $\mathbf{A}$, and a initial random de-mixing matrix $\mathbf{W}$. Each signal contains 2500 data points. The convergent test is set as $\Delta KL \leq 0.000001$. We simulated five iterations for each step with the learning rate values of $0.1 \leq \eta \leq 0.9$ and step size of 0.1. The system is simulated by using Matlab Application.

1.$\mathbf{S1}(t)$ $= 0.1\sin(400t)\cos(30t)$
2.$\mathbf{S2}(t)$ $= 0.01\text{sign}[\sin(500t + 9\cos(40t))]$
3.$\mathbf{S3}(t)$ $=$ uniform noise in range [-1,1]

7 Experimental Results

Computer simulations for ICA problem with appropriate learning rate and momentum rate are presented in this section. Six types of examples are provided to measure an algorithm's efficiency.

1. Fixed learning rate value: $0.1 \leq \eta \leq 0.9$
2. Approach Learning rate to 0 by
$$\eta_t = \eta_{t-1}/1.005$$
3. Approach Learning rate to 0 by
$$\eta_t = \frac{\eta_{t-1}}{appropriate\ divisor}$$
4. Approach Learning rate to 0 by
$$\eta_t = \frac{\eta_{t-1}}{appropriate\ divisor}$$
and 0.01 momentum rate.
5. Approach Learning rate to 0 by
$$\eta_t = \frac{\eta_{t-1}}{appropriate\ divisor}$$
and 0.10 momentum rate.
6. Approach Learning rate to 0 by
$$\eta_t = \frac{\eta_{t-1}}{appropriate\ divisor}$$
and 0.20 momentum rate.

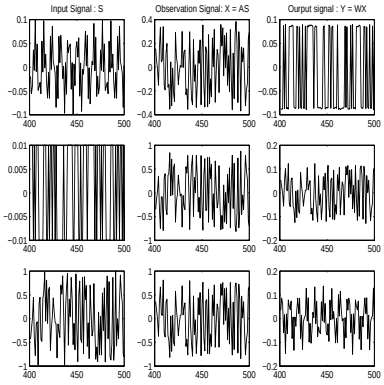

Figure 2: Successful separation of ICA Examples.

In Figure 2, three original signals $\mathbf{S}_i(t)$ were generated and passed to mixing non-singular matrix $\mathbf{A}$, whose information were unknown. An observation signal $\mathbf{x}_i(t)$,which are mixed, dependent, and unknown information signal, are the input of ICA system. Figure 2 shows a successful separation $\mathbf{y}_i(t)$ and Figure 3 displays a correlated output or unsuccessful separation by using ICA with a fixed learning rate of 0.9.

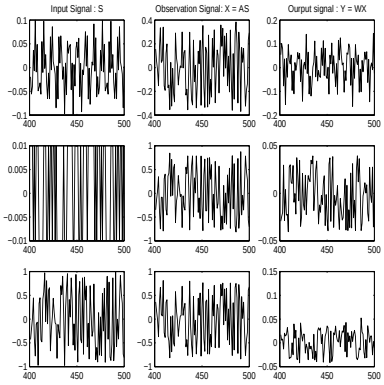

Figure 3: Unsuccessful separation of ICA Algorithm

Figure 4 shows the total number of epoches in Y-axis and the learning rate step in X-axis for six types of experiments. The comparison of all results is illustrated in Figure 5.

8 Conclusion

This paper introduce a new optimization technique for independent component analysis problem. The main problem of ICA or BSS, is working with unknown information of source and

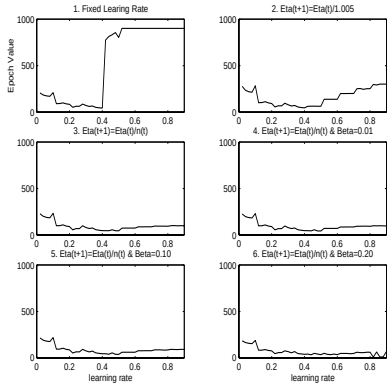

Figure 4: ICA Results: number of epoch for each experiment

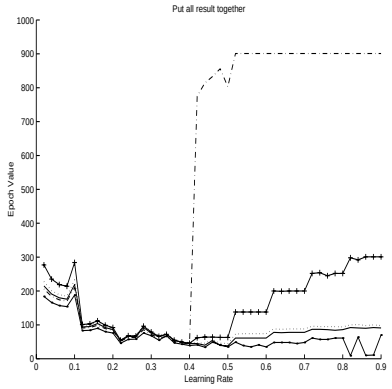

Figure 5: Comparison Experimental Results. Five types of lines are used to denote the results. The dashed-and-dotted line is for fixed η. The ticked line is for $\eta = \eta/1.005$. The dotted line is for $\eta = \eta/1.0+\eta 10^{-2}, \beta = 0.01$. The thick line is for $\eta = \eta/1.0 + \eta 10^{-2}, \beta = 0.10$. The thick-and-dotted line is for $\eta = \eta/1.0 + \eta 10^{-2}, \beta = 0.20$.

mixing matrix $\mathbf{A}$. We try to solve an inversion matrix of $\mathbf{A}$, called de-mixing matrix $\mathbf{W}$ here. The approximation of probability density function is selected as a maximization entropy, Gram-Charlier Expansion. The convergence test is measured by Kullback-Liebler divergence. This algorithm can be easily implemented on an unsupervised neural network model. Our approach provides a flexible and efficiency in selection of the learning rate and momentum factor. The momentum rate and the decreasing of learning rate by their proportion are newly significant parameter in the minimization of number of iterations for the convergence.

References

[1] Amari.S, Cichocki.S, and Yang.H.H. *A New Learning Algorithm for Blind Signal Separation*, MIT Press, pp.757-763, 1996.

[2] Comon.P *Independent Component Analysis: A New Concept?* Signal Processing, vol. 36. pp. 287-314, 1994.

[3] Dai.H. and Macbeth.C. *Effects of Learing Parameters on Learing Procedure and Performance of a BPNN*, Neural Network, Vol.10, No.8, pp.1505-1521, 1997.

[4] Haykin.S. Neural Network a Comprehensive foundation. 2nd, Prentice Hall,1999.

[5] Hyvarinen.A and Oja.E. *Independent Component analysis: algorithms and applications* Neural Networks. Vol. 13 pp. 411-430, 2000

[6] Lee.T.W., Lewicki.M.S and Sejnowski.T.J. *ICA Mixture Models for Unsupervised Classification on Non-Gaussian Classes and Automatic Context Switching in Blind Signal separation*, IEEE Transactions on Pattern analysis and Machine Intelligence, Vol. 22, No.10, Oct 2000.

[7] Pun.M.O. A Simple variable step algorithm for blind source separation(BSS), Master Thesis, university of Tsukuba. 1999.

[8] Yu.X.H. and Chen.G.A. *Efficient Backpropagation Learning Using Optimal Learing Rate and Momentum*, Neural Network, Vol.10, No.3, pp.517-527, 1997.