Variable Step-Size Sign Natural Gradient Algorithm for Sequential Blind Source Separation

Lianxi Yuan, Wenwu Wang, *Member, IEEE*, and Jonathon A. Chambers, *Senior Member, IEEE*

***Abstract*—A novel variable step-size sign natural gradient algorithm (VS-S-NGA) for online blind separation of independent sources is presented. A sign operator for the adaptation of the separation model is obtained from the derivation of a generalized dynamic separation model. A variable step size is also derived to better match the dynamics of the input signals and unmixing matrix. The proposed sign algorithm is appealing in practice due to its computational simplicity. Experimental results verify the superior convergence performance over conventional NGA in both stationary and nonstationary environments.**

***Index Terms*—Adaptive step size, natural gradient, sequential blind source separation (BSS), sign algorithm.**

I. Introduction

BLIND signal seperation (BSS) is a technique that aims to recover the underlying unknown source signals from their observed mixtures without prior knowledge of the mixing channels. This method has several applications in communications and signal processing [1]. Suppose n unknown statistically independent zero-mean source signals, with at most one having a Gaussian distribution, contained within $\mathbf{s} \in \mathbb{R}^n$ pass through an unknown mixing channel $\mathbf{A} \in \mathbb{R}^{m\times n}(m \geq n)$, such that m mixed signals $\mathbf{x} \in \mathbb{R}^m$ are, therefore, observed that can be modeled as $\mathbf{x} = \mathbf{A}\mathbf{s} + \mathbf{e}$, where $\mathbf{e} \in \mathbb{R}^m$ is the possible contaminating noise vector, ignored for simplicity in this letter. The objective of BSS is to recover the original sources given only the observed mixtures, using the separation model $\mathbf{y} = \mathbf{W}\mathbf{x}$, where $\mathbf{y} \in \mathbb{R}^n$ is an estimate of $\mathbf{s}$ to within the well-known permutation and scaling ambiguities, and $\mathbf{W} \in \mathbb{R}^{n\times m}$ is the separation matrix. The crucial assumption with conventional BSS is that the source signals are statistically independent. We further assume that the sources have unit variance and that the number of sources matches that of the number of mixtures, i.e., $m = n$, the exactly determined problem. To recover the source signals, it is frequently necessary to estimate an unmixing matrix that performs the inverse operation of the mixing process, as subsequently used in the separation model. In this letter, we are particularly concerned with a family of sequential BSS algorithms. Fig. 1 shows a block diagram of sequential BSS. The separating coefficients $\mathbf{W}(k)$ are updated iteratively according to some estimate of the independence between the estimated signal components in $\mathbf{y}(k)$. The sensor signal components in $\mathbf{x}(k)$ are fed into the algorithm in order to estimate iteratively the source signal components, i.e., $\mathbf{y}(k)$. Compared with block (batch)-based BSS algorithms, sequential approaches have particular practical advantage due to their computational simplicity and potentially improved performance in tracking a nonstationary environment [2]. The focus of this study is, therefore, the natural gradient algorithm (NGA) [1], where the discrete-time online updating equation of the separation matrix is denoted as

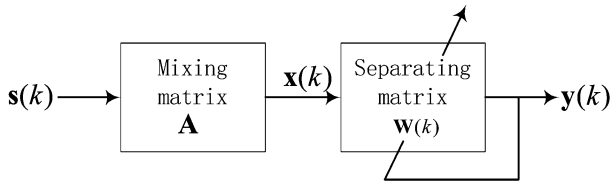

Fig. 1. Sequential blind source separation.

$$\mathbf{W}(k+1) = \mathbf{W}(k) + \mu[\mathbf{I} - \mathbf{Q}(k)]\mathbf{W}(k) \tag{1}$$

where k is the discrete-time index, μ is a positive parameter known generally as the step size, $\mathbf{I}$ is an identity matrix, and $\mathbf{Q}(k)$ is given by

$$\mathbf{Q}(k) = \mathbf{f}(\mathbf{y}(k))\mathbf{y}^T(k) \tag{2}$$

where $\mathbf{f}(\mathbf{y}(k))$ is an odd nonlinear function that acts element wise on the output vector $\mathbf{y}(k)$, and $(\cdot)^T$ is the vector transpose operator.

Two important issues affecting the performance of sequential algorithms such as (1) are the *convergence rate* and the *misadjustment* in steady state [3]. A fixed step size can restrict the convergence rate and can lead to poor tracking performance [2]. In contrast, an adaptive step size can exploit the online measurements of the state of the separation system from the outputs and the parameter updates. This means that the step size can be increased for a higher convergence rate but can be systematically decreased for reducing any misadjustment of the parameters away from their optimum settings. To improve the convergence rate, we consider using a normalization technique (leading to a sign algorithm) together with gradient-based time-varying step size (leading to a variable step-size algorithm) in the updating

Manuscript received January 5, 2005; revised March 3, 2005. The work of L. Yuan was supported by the China Scholarship Council, and the work of W. Wang was supported by the Engineering and Physical Science Research Council, U.K. The associate editor coordinating the review of this manuscript and approving it for publication was Prof. Yue (Joseph) Wang.

L. Yuan was with the Centre of Digital Signal Processing, Cardiff School of Engineering, Cardiff University, Cardiff CF24 0YF, U.K., and is now with the Underwater Acoustics Research Institute, Harbin Engineering University, Harbin 15001, China (e-mail: yuan_lianxi@hotmail.com).

W. Wang was with the Centre of Digital Signal Processing, Cardiff School of Engineering, Cardiff University, Cardiff, CF24 0YF, U.K. and is now with the Tao Group Limited, Reading, RG6 1AZ, U.K. (e-mail: wangw2@cf.ac.uk).

J. A. Chambers is with the Centre of Digital Signal Processing, Cardiff School of Engineering, Cardiff University, Cardiff, CF24 0YF, U.K. (e-mail: chambersj@cf.ac.uk).

Digital Object Identifier 10.1109/LSP.2005.851264

process. Both techniques are shown to increase the convergence speed of the algorithm, and the sign operation can simultaneously reduce the computational complexity of the whole algorithm, additionally introduced by the adaptive step size. The remainder of this letter is organized as follows. A sign algorithm using a normalization technique based on the standard NGA algorithm is proposed in Section II. Section III is dedicated to deriving a variable step-size algorithm for NGA, where the step size is estimated from the input data and the separation matrix. Numerical experiments are presented in Section IV to compare the convergence performance of the proposed algorithms with that of the conventional NGA. Finally, Section V concludes the letter.

II. Sign NGA (S-NGA)

In this section, we consider using normalization of the output vector $\mathbf{y}(k)$ for the off-diagonal terms of $\mathbf{Q}(k)$. This, thereby, results in a sign operation on the elements of $\mathbf{Q}(k)$, which restricts the norm of the matrix $\mathbf{W}(k)$. Our expectation is that this will lead to better robustness in the adaptation. For mathematical formulation, let us consider a continuous matrix dynamic system

$$\frac{d}{dt}\mathbf{W}(t) = -\mu\frac{\partial J(\mathbf{y}(t),\mathbf{W}(t))}{\partial \mathbf{W}(t)}\mathbf{W}^T(t)\mathbf{\Pi}(\mathbf{y}(t))\mathbf{W}(t) \quad (3)$$

where $J(\cdot)$ is a cost function from which NGA is derived, and $\mathbf{\Pi}(\mathbf{y})$ is a diagonal matrix with positive elements. Equation (3) can be deemed as an extension of the standard NGA [4], since (1) is a result of $\mathbf{\Pi}(\mathbf{y}) = \mathbf{I}$. By a straightforward differential matrix calculation as in [1], we obtain

$$\frac{d}{dt}\mathbf{W}(t) = \mu\mathbf{\Pi}(\mathbf{y}(t))[\mathbf{I} - \mathbf{\Pi}^{-1}(\mathbf{y}(t))\tilde{\mathbf{f}}(\mathbf{y}(t))\mathbf{y}^T(t)\mathbf{\Pi}(\mathbf{y}(t))] \times \mathbf{W}(t) \quad (4)$$

where $\tilde{\mathbf{f}}(\mathbf{y})$ is a vector of nonlinear activation functions. Defining $\mathbf{\Pi}^{-1}(\mathbf{y}(t))\tilde{\mathbf{f}}(\mathbf{y}(t)) = \mathbf{f}(\mathbf{y}(t))$ and $\mu\mathbf{\Pi}(\mathbf{y}(t)) = \mu(t)$, we have

$$\frac{d}{dt}\mathbf{W}(t) = \mu(t)[\mathbf{I} - \mathbf{f}(\mathbf{y}(t))\mathbf{y}^T(t)\mathbf{\Pi}(\mathbf{y}(t))]\mathbf{W}(t). \quad (5)$$

In parallel with (1), from (5), we have

$$\mathbf{Q}(t) \triangleq \mathbf{f}(\mathbf{y})\mathbf{y}^T(t)\mathbf{\Pi}(\mathbf{y}(t)). \quad (6)$$

Denote by $f_i(y_i)$ and $y_i, i = 1, \ldots, n$ the entries of $\mathbf{f}(\mathbf{y})$ and $\mathbf{y}$, and by π_{ij}, the elements of $\mathbf{\Pi}, \mathbf{Q}(t)$ can be rewritten element wise as

$$q_{ij}(t) = f_i(y_i)y_j\pi_{jj}. \quad (7)$$

If π_{jj} takes the form of the normalization by y_j, i.e., $\pi_{jj} = |y_j|^{-1}$, then (6) is reduced to

$$\mathbf{Q}(t) \triangleq \mathbf{f}(\mathbf{y})[\text{sign}(\mathbf{y}(t))]^T \quad (8)$$

where $\text{sign}(\mathbf{y}(t)) = [\text{sign}(y_1(t)), \ldots, \text{sign}(y_n(t))]^T$, and

$$\text{sign}(z) = \begin{cases} 1, & z > 0 \\ -1, & z < 0 \\ 0, & z = 0. \end{cases} \quad (9)$$

Note that (8) could be deemed as a degenerate form of the median learning rule discussed in [4]. The introduced normalization could, moreover, potentially lead to a faster convergence rate because of the resulting sign activation function of the output data $\mathbf{y}$ increasing the magnitude of small output values, i.e., values of $|y_i| \in (0, 1)$ are set to ± 1. On the other hand, however, this could reduce the accuracy of statistics within the adaptation process, leading to relatively inaccurate separation and increased misadjustment. However, this effect is not obvious, as shown in our simulations. To achieve a better tradeoff between the convergence rate (also computational complexity) and the separation performance in simulation, we suggest to use different normalization schemes for the elements of $\mathbf{Q}(t)$. Particularly, $\mathbf{\Pi}$ does not hold fixed values at its diagonal elements, but these change according to the association between $\mathbf{f}(\mathbf{y}(t))$ and $\mathbf{y}(t)$. That is, (7) is rewritten in the discrete-time form as

$$q_{ij}(k) = \begin{cases} f_i(y_i(k))y_i(k), & i = j \\ f_i(y_i(k))\text{sign}(y_j(k)), & i \neq j. \end{cases} \quad (10)$$

Using the Kronecker dot product $\odot$ (element-wise product of matrices, i.e., each entry is a product of the corresponding entries from two individual matrices), we have the following concise expression:

$$\mathbf{Q}(k) \triangleq \mathbf{f}(\mathbf{y})\mathbf{y}^T(k) \odot \mathbf{\Phi}(\mathbf{y}(k)) \quad (11)$$

where $\mathbf{\Phi}(\mathbf{y}k))$ is derived from $\mathbf{\Pi}$ and (10), i.e., the entries of $\mathbf{\Phi}$ are denoted as

$$\varphi_{ij} = \begin{cases} 1, & i = j \\ |y_j|^{-1}, & i \neq j. \end{cases} \quad (12)$$

Note that (11) can also be written as

$$\mathbf{Q}(k) \triangleq \text{diag}[\mathbf{f}(\mathbf{y}(k))\mathbf{y}^T(k)] + \text{off}[\mathbf{f}(\mathbf{y}(k))\text{sign}(\mathbf{y}^T(k))] \quad (13)$$

where $\text{diag}[\cdot]$ and $\text{off}[\cdot]$ denote the operation of taking the diagonal elements and off-diagonal elements of a matrix, respectively.

We call the adaptation procedure of using (11) and (1) the sign natural gradient algorithm (S-NGA). Compared with the NGA using (2), the sign algorithm (SA) has reduced computational complexity, i.e., $n(n-1)$ multiplications in (2) are replaced with simple sign tests that are easily implementable. However, for each k, the off-diagonal elements of $\mathbf{Q}(k)$ are not continuous [see (10)], this makes the analysis of such an algorithm more difficult than that of (1). For the ease of analysis, we assume the elements to have fixed values. It is, therefore, straightforward to show that the algorithm is Lyapunov stable. Noticing that $\mathbf{W}^T\mathbf{\Pi}\mathbf{W} = (\sqrt{\mathbf{\Pi}}\mathbf{W})^T(\sqrt{\mathbf{\Pi}}\mathbf{W})$ in (3), where $\sqrt{\mathbf{\Pi}}$ represents a diagonal matrix whose diagonal entries are the square root of the corresponding diagonal elements of Π, and denoting by

w_{ij}, γ_{ij}, and $\psi_{ij}, i, j = 1, \ldots, n$, the elements of $\mathbf{W}$, $\sqrt{\Pi}\mathbf{W}$, and $(\partial J)/(\partial \mathbf{W})(\sqrt{\Pi}\mathbf{W})^T$, we obtain from (3) that

$$\begin{aligned}\frac{d}{dt}J(\mathbf{y}(t), \mathbf{W}(t)) &= \sum_{i,j}\frac{\partial J}{\partial w_{ij}}\frac{dw_{ij}}{dt} \\ &= -\sum_{i,j}\frac{\partial J}{\partial w_{ij}}\sum_{k}\psi_{ik}\gamma_{kj} \\ &= -\sum_{i,k}\psi_{ik}^2 \leqslant 0 \end{aligned} \tag{14}$$

where zero is obtained if and only if $d\mathbf{W}(t)/dt = 0$, which means the solution to $\mathbf{W}$ is an equilibrium of (3). The subsequent simulation studies and the well-known advantage of normalization suggest that the stability of S-NGA follows from the included Lyapunov analysis.

III. Variable Step-Size Sign NGA (VS-S-NGA)

It has been shown [2] that, as compared with using a fixed step-size, which would restrict convergence rate, the algorithm with an adaptive step-size has an improved tracking performance for a nonstationary environment, i.e., the value is adjusted according to the time-varying dynamics of the input signals and the separating matrix. As another contribution, we, therefore, derive a gradient adaptive step-size algorithm for the NGA algorithm, which adapts the step size in the form of

$$\mu(k) = \mu(k-1) - \rho\nabla_{\mu}J(k)|_{\mu=\mu(k-1)} \tag{15}$$

where ρ is a small constant, and $J(k)$ is an instantaneous estimate of the cost function from which the NGA algorithm is derived. To proceed, we use an inner product of matrices defined as [2]

$$\langle \mathbf{C}, \mathbf{D}\rangle = \text{tr}(\mathbf{C}^T\mathbf{D}) \tag{16}$$

where $\langle \cdot \rangle$ denotes the inner product, $\text{tr}(\cdot)$ is the trace operator, and $\mathbf{C}, \mathbf{D} \in \mathbb{R}^{m\times n}$. Therefore, exploiting (16), the gradient term on the right-hand side of (15) can be evaluated as

$$\begin{aligned}&\nabla_{\mu}J(k)|_{\mu=\mu(k-1)} \\ &\quad= \langle \partial J(k)/\partial \mathbf{W}(k), \partial \mathbf{W}(k)/\partial\mu(k-1)\rangle \\ &\quad= \text{tr}(\partial J(k)/\partial \mathbf{W}(k)^T \times \partial \mathbf{W}(k)/\partial\mu(k-1))\end{aligned} \tag{17}$$

where

$$\partial J(k)/\partial \mathbf{W}(k) = -[\mathbf{I} - \mathbf{f}(\mathbf{y}(k))\mathbf{y}^T(k)]\mathbf{W}(k) \tag{18}$$

which is the instantaneous estimate of the natural gradient of the cost function of $J(k)$. From (1), the separating matrix $\mathbf{W}$ at time k is obtained as

$$\begin{aligned}\mathbf{W}(k) = \mathbf{W}(k-1) + \mu(k-1)[\mathbf{I} - \mathbf{f}(\mathbf{y}(k-1))\mathbf{y}^T(k-1)] \\ \times \mathbf{W}(k-1).\end{aligned} \tag{19}$$

Following the approach from [2] and [5], from the above equation, we have

$$\partial \mathbf{W}(k)/\partial\mu(k-1) = [\mathbf{I} - \mathbf{f}(\mathbf{y}(k-1))\mathbf{y}^T(k-1)]\mathbf{W}(k-1). \tag{20}$$

TABLE I
Comparision of Computational Complexities of the Four Algorithms

Algorithms	NGA	S-NGA	VS-NGA	VS-S-NGA
Equations required	(1) and (2)	(1) and (11)	(1), (2), (21) and (23)	(1), (11), (21) and (23)
Operations required	$2n^3 + 3n^2$	$2n^3 + 2n^2 + n$	$6n^3 + 3n^2 + n + 1$	$6n^3 + 2n^2 + 2n + 1$

Using the notation of (2) for $\mathbf{Q}(k)$ in the standard NGA algorithm and denoting

$$\mathbf{\Gamma}(k) \triangleq [\mathbf{I} - \mathbf{Q}(k)]\mathbf{W}(k) \tag{21}$$

we have

$$\nabla_{\mu}J(k)|_{\mu=\mu(k-1)} = -\text{tr}(\mathbf{\Gamma}^T(k)\mathbf{\Gamma}(k-1)). \tag{22}$$

Hence, an adaptive step size with the form of (15) can be written as

$$\mu(k) = \mu(k-1) + \rho\,\text{tr}(\mathbf{\Gamma}^T(k)\mathbf{\Gamma}(k-1)) \tag{23}$$

which can be estimated from the input signals and the separation matrix. It is worth noting that (21) has a similar form as [2, (7)], which was derived for an equivariant adaptive source separation via independence (EASI) algorithm [6]. However, due to the different formulations between NGA and EASI algorithms, the expressions of (18), (20), and (21) essentially take distinctive forms from those in [2]. For the purpose of readability, we, therefore, incorporate the explicit derivations as detailed above, rather than simply refer to [2]. The separation procedure using (1), (2), (21), and (23) represents the VS-NGA. Following a similar procedure as in Section II, see (6) and (11), and as in this section, see (18) and (20). It is straightforward to derive an adaptive step-size algorithm using different normalization for the off-diagonal elements of $\mathbf{Q}(k)$. In this case, $\mathbf{Q}(k)$ takes the form of (11). We represent (1), (11), (21), and (23) as the sign version of the variable step-size NGA algorithm, i.e., VS-S-NGA for notational simplicity.

For more substantial comparison of the aforementioned algorithms, we finally quantify their computational costs and summarize their required operations (multiplications and additions) in Table I. From this table, it is observed that S-NGA and VS-S-NGA are less complex than NGA and VS-NGA, respectively. Fairly speaking, VS-NGA, however, has actually increased the computational cost due to additional computations required for the variable step size. This indicates that the improved convergence rate of VS-NGA, as verified in our simulations, is obtained at the expense of additional computations.

IV. Numerical Experiments

In the first experiment, we mix a fixed sinusoidal signal with a randomly selected uniform source signal by using a 2-by-2 $(m = n = 2)$ matrix $\mathbf{A}_0$, i.e.,

$$\mathbf{A}_0 = \begin{bmatrix} 1.0 & 0.5 \\ -0.3 & 0.8 \end{bmatrix}. \tag{24}$$

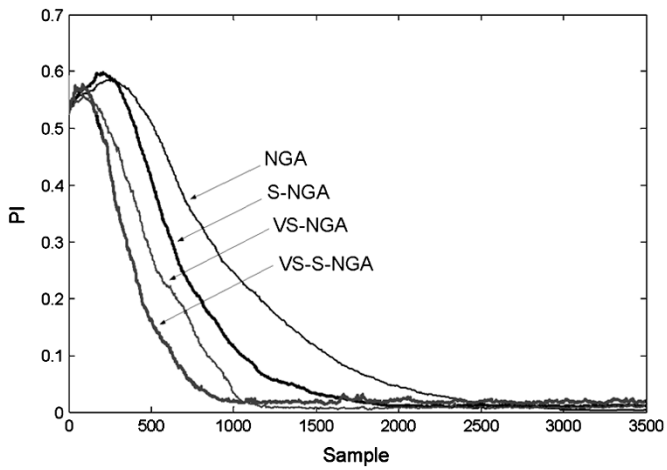

Fig. 2. Comparison of convergence rate by performance index in a stationary environment.

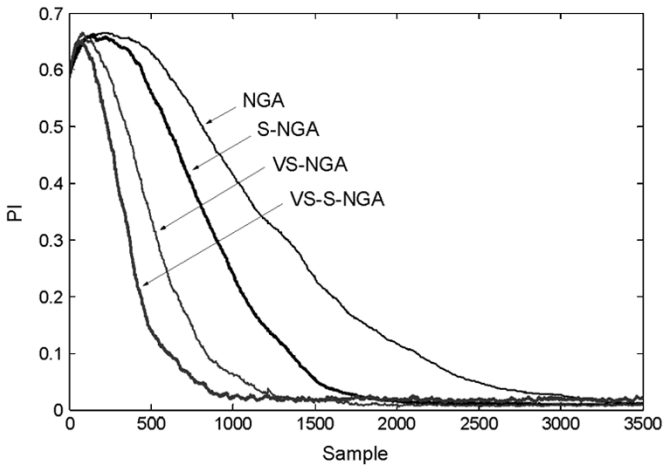

Fig. 3. Comparison of convergence rate by performance index in a nonstationary environment.

Zero-mean independent white Gaussian noise was added to the mixtures such that the signal-to-noise ratio (SNR) equals 20 dB. A cubic nonlinearity $f(\cdot)$ was used as the activation function. The performance index (PI) [1], as a function of the system matrix $\mathbf{G} = \mathbf{WA}$, was used to evaluate the proposed algorithm

$$\mathrm{PI}(\mathbf{G}) = \left[\frac{1}{n}\sum_{i=1}^{n}\left(\sum_{k=1}^{m}\frac{g_{ik}}{\max_k g_{ik}}\right) - 1\right] + \left[\frac{1}{m}\sum_{k=1}^{m}\left(\sum_{i=1}^{n}\frac{g_{ik}}{\max_i g_{ik}}\right) - 1\right] \quad (25)$$

where g_{ik} is the ikth element of $\mathbf{G}$. The initial value of μ for all the algorithms was set to 0.004, $\rho = 1 \times 10^{-5}$, and 100 Monte Carlo trials were run for an averaged performance. The same simulation conditions were used for all the algorithms to allow fair comparison. Fig. 2 shows convergence behavior of the various approaches. From Fig. 2, it is found that the proposed sign algorithms have much faster convergence speed. For example, for the fixed step size, S-NGA needs approximately 2000 samples to converge, whereas the conventional NGA needs approximately 3250 samples. Note that we mean the *convergence* by the PI reduced to 0.02 (corresponding to an approximately successful separation). For the adaptive step size, VS-S-NGA only requires approximately 1050 samples for convergence; however, VS-NGA requires approximately 1700 samples. It is clear that VS-S-NGA has the fastest convergence rate, which is a very promising property for sequential algorithms.

In the second experiment, the different approaches were examined for a nonstationary environment. To this end, we use the following time-varying mixing matrix:

$$\mathbf{A} = \mathbf{A}_0 + \hat{\mathbf{\Xi}} \quad (26)$$

where $\hat{\mathbf{\Xi}} = \alpha\hat{\mathbf{\Xi}} + \beta\,\mathrm{randn}(\mathrm{size}(\mathbf{A}), 1), \mathrm{randn}(\cdot)$ and $\mathrm{size}(\cdot)$ are MATLAB built-in functions, and the initial $\hat{\mathbf{\Xi}}$ is set to a null matrix. $\mathbf{A}_0$ is the same as in (24). Here, α is set to 0.9 and β to 0.001. Other parameters are the same as those in the first experiment. Again, their convergence performance is compared in Fig. 3. For this figure, we observed similar performance improvement gained for the proposed approaches in a nonstationary environment. This experiment also indicates that the proposed algorithms retained the *equivariant* property, as for NGA algorithm. Note that lower PI generally indicates a better separation performance. In both Figs. 2 and 3, although we have not observed much difference between the final separation performance by S-NGA and VS-S-NGA in terms of PI measurement, the key point is that with the proposed normalization techniques, faster convergence rates have been achieved.

V. Conclusion

A new sign and variable step-size natural gradient algorithm for online blind separation of independent sources has been presented. The derivation is based on the gradient calculation of a generalized dynamic equation. By applying the sign operation, the separation algorithm has been found to have much faster convergence rate as compared with the conventional natural gradient algorithm. The algorithm was shown to be Lyapunov stable. We also derived a variable step-size algorithm for the natural gradient learning that was also shown to have faster convergence rate and better tracking performance in a nonstationary environment than using a fixed step-size algorithm.

References

[1] A. Cichocki and S. Amari, *Adaptive Blind Signal and Image Processing: Learning Algorithms and Applications*. New York: Wiley, 2002.
[2] J. A. Chambers, M. G. Jafari, and S. McLaughlin, "Variable step-size EASI algorithm for sequential blind source separation," *Electon. Lett.*, pp. 393–394, 2004.
[3] S. C. Douglas and A. Cichocki, "On-line step-size selection for training of adaptive systems," *IEEE Signal Process. Mag.*, vol. 14, no. 6, pp. 45–46, Nov. 1997.
[4] P. Georgiev, A. Cichocki, and S. Amari, "On some extensions of the natural gradient algorithm," in *Proc. ICA*, 2001, pp. 581–585.
[5] V. J. Mathews and Z. Xie, "A stochastic gradient adaptive filter with gradient adaptive step size," *IEEE Trans. Signal Process.*, vol. 41, no. 6, pp. 2075–2087, Jun. 1993.
[6] J.-F. Cardoso and B. H. Laheld, "Equivariant adaptive source separation," *IEEE Trans. Signal Process.*, vol. 44, no. 12, pp. 3017–3030, Dec. 1996.