USING PIECEWISE LINEAR NONLINEARITIES IN THE NATURAL GRADIENT AND FASTICA ALGORITHMS FOR BLIND SOURCE SEPARATION

Jih-Cheng Chao

Semiconductor Group
Texas Instruments
Dallas, Texas 75243 USA

Scott C. Douglas

Department of Electrical Engineering
Southern Methodist University
Dallas, Texas 75275 USA

ABSTRACT

In both the natural gradient algorithm and the FastICA algorithm for blind source separation (BSS), output nonlinearities for each extracted source must be selected, and the performance of each approach can be sensitive to the chosen output nonlinearity. In this paper, we propose to use simple piecewise-linear output nonlinearities for these algorithms and obtain a number of useful properties with such a choice. For the natural gradient BSS algorithm, nonlinearity-switching is easily achieved through a common stability criterion that guarantees local stability for all source distributions. For the FastICA algorithm, the chosen nonlinearities can be very close to linear, suggesting that simple (*e.g.* μ-law) output companding is sufficiently nonlinear to allow separation when used with this algorithm. Simulations are provided to verify the theoretical results.

Index Terms— Separation, multidimensional signal processing, adaptive systems, piecewise linear approximation

1. INTRODUCTION

The goal of blind source separation (BSS) is to estimate m distinct sources as observed in m linear signal mixtures. In the noise-free case, a linear demixing system is sufficient for this task, such that the problem reduces to computing m linear combinations of the m signal mixtures to estimate each source. Numerous procedures for adjusting the demixing coefficients have been developed; however, two of the most-popular methods are the natural gradient or INFOMAX algorithm [1, 2] and the FastICA algorithm [3, 4]. Both of these procedures assume that the sources are statistically-independent and non-Gaussian and rely on output nonlinearities $g_i(y_i)$ for each extracted source $y_i(k)$ to obtain separation. These nonlinearities are used to compute the updated coefficients in the separation algorithm, which require cross-correlations of $g_i(y_i(k))$ with the estimated output signals or the prewhitened input signal mixtures. Thus, an important design consideration in each of these algorithms is the choice of $g_i(y_i)$ used to obtain each estimated output $y_i(k)$.

The FastICA algorithm is more powerful than the natural gradient algorithm in terms of its separating capability when the $g_i(y_i)$ nonlinearities are fixed. The FastICA algorithm can separate almost any mixture of source types, although the estimated source quality suffers for some combinations of output nonlinearity and source type (*e.g.* using $g_i(y_i) = y_i^3$ for impulsive source distributions like speech). The natural gradient algorithm cannot separate all source mixtures with the output nonlinearities are fixed; as such, procedures for adapting the output nonlinearities according to the observed output signals may be required [5, 6, 7, 8]. While such an approach can work, it must guarantee local stability for all possible source distributions for some output nonlinearity within the design family to avoid algorithm failure[7].

In this paper, we propose a simple piecewise-linear family of output nonlinearities that is useful for both the natural gradient BSS algorithm and the FastICA algorithm. Besides being simple to implement, they have the following features:

1. For the natural gradient algorithm, local stability about a separation solution largely depends on a single parameter that can be easily selected from a single statistical quantity calculated from the estimated outputs of the system. All source distributions are separable locally with this method, thus avoiding a chief limitation of this approach.

2. For the FastICA algorithm, the deviation of the output nonlinearity from linearity can be extremely small – so small, in fact, that the signal distortion created by the nonlinearity can be inconsequential relative to the linear source estimate in most applications. Thus, $g_i(y_i(k))$ can be used in place of $y_i(k)$ wherever the latter signal is needed.

As the piecewise-linear family shares similar features to simple (μ-law) companding procedures common to certain signal coding schemes, we also explore the performance of these algorithms using companding nonlinearities, showing that they can be useful in BSS contexts. Simulations are used to verify the theoretical results derived.

2. PIECEWISE-LINEAR OUTPUT NONLINEARITIES

The output nonlinearities considered in this paper are

$$g_i(y) = \begin{cases} y & |y| < \theta_i \\ [a_i(|y| - \theta_i) + \theta_i]\,\mathrm{sgn}(y) & |y| \geq \theta_i \end{cases} \quad (1)$$

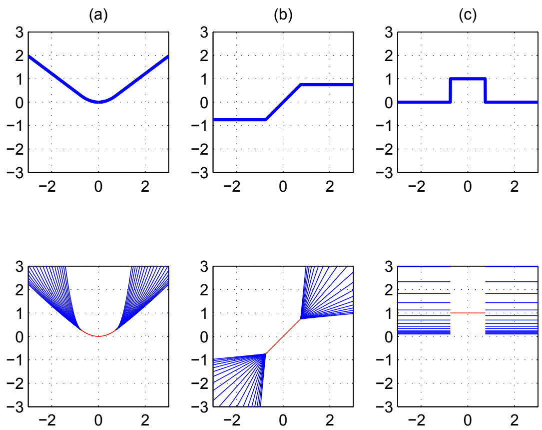

Fig. 1. $G_i(y)$, $g_i(y)$, and $g_i'(y)$ for the proposed nonlinearity with various values of a_i (bottom) in comparison to the Huber M-estimator cost function (top), $\theta_i = 0.7$.

where θ_i and a_i are parameters to be chosen based on the ith extracted source output from the separation system. For completeness, we also provide the integral (cost function) $G_i(y)$ and derivative $dg_i(y)/dy$ of $g_i(y)$ that are important for analyzing local stability of the separation algorithms:

$$G_i(y) = \begin{cases} \dfrac{y^2}{2} & |y| < \theta_i \\ \dfrac{1}{2}a_i y^2 + \theta_i(1-a_i)\left[|y| - \dfrac{\theta_i}{2}\right] & |y| \geq \theta_i \end{cases} \tag{2}$$

$$\frac{dg_i(y)}{dy} = g_i'(y) = \begin{cases} 1 & |y| < \theta_i \\ a_i & |y| \geq \theta_i \end{cases} \tag{3}$$

Fig. 1 shows the shapes of $G_i(y)$, $g_i(y)$, and $g_i'(y)$ for $\theta_i = 0.7$ and various values of a_i. The top row shows the case where $a_i = 0$, in which case $g_i(y)$ becomes the influence function of the Huber M-estimator cost [9, 10]. The bottom row shows a family of functions for values of a_i over the interval $[0, \infty]$ not including $a_i = 1$ for which $G_i(y)$ would be quadratic, $g_i(y)$ would be linear, and $g_i'(y) = 1$.

In the sequel, it will be important for analytical reasons to define the nonlinearities $\widehat{g}_\theta(y) = g_i(y)$ and $\widehat{g}'_\theta(y) = g_i'(y)$ for the special cases of $a_i = 0$ and $\theta_i = \theta$, corresponding to the Huber M-estimator influence function [9, 10]

3. NONLINEARITY DESIGN FOR NATURAL GRADIENT BSS

We now explain why $g_i(y)$ in (1) is useful for the natural gradient BSS algorithm. It is well-known that, for a source $s_i(k)$ with p.d.f. $p_i(s_i)$, a sufficient condition to guarantee that the natural gradient algorithm in [1, 11] is locally-stable about a separating solution for the ith extracted source is

$$E\{s_i(k)g_i(s_i(k))\} - E\{s_i^2(k)\}E\{g_i'(s_i(k))\} < 0 \tag{4}$$

where $E\{\cdot\}$ denotes statistical expectation. This inequality is only satisfied for certain nonlinearities $g_i(y)$, for any given source p.d.f. $p_i(s_i)$, which in our case corresponds to certain parameter choices θ_i and a_i. The following theorem shows how these parameters can be designed.

Theorem 1.1: *For $g_i(y)$ and $g_i'(y)$ in (1) and (3), let $a_i \neq 1$. Furthermore, assume that $s_i(k)$ is continuous and non-Gaussian-distributed. Then, there always exists a value of θ_i such that*

$$E\{s_i(k)g_i(s_i(k))\} - E\{s_i^2(k)\}E\{g_i'(s_i(k))\} \neq 0 \tag{5}$$

Theorem 1.2: *For $g_i(y)$ and $g_i'(y)$ in (1) and (3), let θ_i be chosen such that (5) is satisfied. Then, the local stability condition in (4) is satisfied if*

$$0 \leq a_i < 1 \qquad \text{for } \widehat{f}_i(\theta_i) < 0 \tag{6}$$

$$a_i > 1 \qquad \text{for } \widehat{f}_i(\theta_i) > 0, \tag{7}$$

where

$$\widehat{f}_i(\theta_i) = E\{s_i(k)\widehat{g}_{\theta_i}(s_i(k))\} - E\{s_i^2(k)\}E\{\widehat{g}'_{\theta_i}(s_i(k))\} \tag{8}$$

Proof: Assume without loss of generality that $s_i(k)$ has an even-symmetric p.d.f. with unit variance. Using straightforward calculus similar to that used in the proofs in [9, 10], one can show that

$$\begin{aligned} & E\{s_i(k)g_i(s_i(k))\} - E\{s_i^2(k)\}E\{g_i'(s_i(k))\} \\ &= 2(1-a_i)\int_\theta^\infty (1+\theta s - s^2)p_i(s)ds \qquad (9) \\ &= (1-a_i)\widehat{f}_i(\theta_i). \qquad (10) \end{aligned}$$

We now leverage a result from [9, 10]: for any continuous non-Gaussian distribution $p_i(s)$, there always exists a $\theta_i \geq 0$ such that $\widehat{f}_i(\theta_i) \neq 0$. Therefore, if $a_i \neq 1$, Theorem 1.1 follows. Moreover, substituting (10) into (4) yields the inequality

$$(1-a_i)\widehat{f}_i(\theta_i) < 0. \tag{11}$$

Knowing the value of $\widehat{f}_i(\theta_i)$, we can guarantee (11) using the strategy in (6)–(7). Thus, Theorem 1.2 follows.

Remark #1: The above theorem suggests a straightforward technique for using piecewise-linear nonlinearities for the natural gradient algorithm. In this technique, each nonlinearity $g_i(y_i)$ would be chosen according to the output statistics of the ith extracted source $y_i(k)$. These methods require the calculation of the following quantity, where sample averages are used in place of expectations:

$$f_i(\theta_i) = E\{y_i(k)\widehat{g}_{\theta_i}(y_i(k))\} - E\{y_i^2(k)\}E\{\widehat{g}'_{\theta_i}(y_i(k))\}. \tag{12}$$

Note that if $y_i(k) = s_i(k)$, $f_i(\theta_i) = \widehat{f}_i(\theta_i)$.

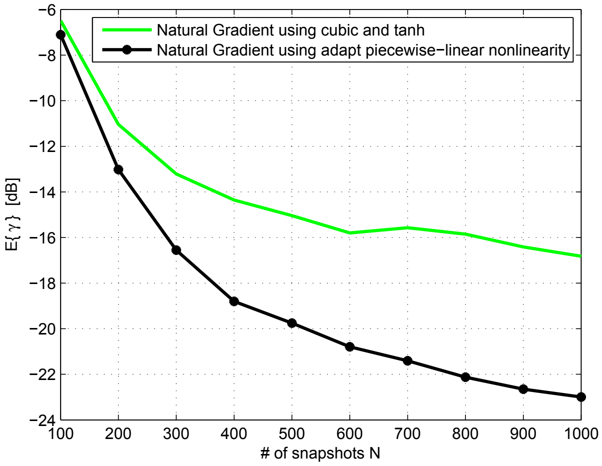

Fig. 2. $E\{\gamma\}$ vs. number of snapshots N for the adaptive piecewise-linear nonlinearity algorithm with $\{a_-, a_+\} = \{0, 5\}$ and for the algorithm in [5] using y_i^3 and $\tanh(y_i)$.

Step 1: Select θ_i such that $f_i(\theta_i)$ is non-zero. From our experience with the Huber M-estimator influence function, θ_i need not be carefully chosen; typical values for θ_i are in the range $0.5 \le \theta_i \le 1$.

Step 2: As $f_i(\theta_i) \approx \widehat{f}_i(\theta_i)$ near algorithm convergence, select a_i according to the rule in (6)–(7) using $f_i(\theta_i)$ as computed from sample averages in place of $\widehat{f}(\theta_i)$. Typically, only two possible values of a_i are allowed based on the sign of $f_i(\theta_i)$, such as $a_i \in \{0, 5\}$ or $a_i \in \{0.5, 2\}$. In the sequel, we refer to these values as a_- and a_+, respectively.

Remark #2: The practical approach described in the above remark assumes that $\mathrm{sgn}(f_i(\theta_i)) = \mathrm{sgn}(\widehat{f}_i(\theta_i))$, which is only true if the algorithm is close to a separating solution. Thus, the technique only assures local stability; separation of mixtures from an arbitrary initial demixing matrix is not guaranteed. This difficulty is inherent to the parallel structure of the natural gradient algorithm no matter how the output nonlinearities are selected, a problem that is avoided by the FastICA algorithm when sequential extraction is used.

4. SIMULATIONS

We now explore the performance of the natural gradient algorithm using piecewise-linear nonlinearities. In our first example, we consider mixtures of 10 sources: two binary-$\{\pm 1\}$, two four-level $\{-3/\sqrt{5},\ -1/\sqrt{5},\ 1/\sqrt{5},\ 3/\sqrt{5}\}$, two uniform-$[-\sqrt{3}, \sqrt{3}]$, two Laplacian, and two generated from a particularly-difficult distribution described in [7]. This particular distribution is four-level symmetric with symbols having the values $[\pm A_1, \pm A_2]$ with $A_1 = 0.718$, $A_2 = 2.7284$, $Pr(|s_i| = A_1) = 0.465$ and $Pr(|s_i| = A_2) = 0.035$. This last distribution is not separable using the natural gradient algorithm for either $g_i(y_i) = y_i^3$ or $g_i(y_i) = \tanh(y_i)$, two common choices for output nonlinearities. Each mixture was first separated by the FastICA algorithm to achieve local convergence, at which point the natural gradient algorithm with piecewise-linear nonlinearities was used. For comparison, we also attempted a similar separation procedure using the algorithm in [5], which switches between the output nonlinearities y_i^3 and $\tanh(y_i)$ depending on their local stability conditions as measured by $y_i(k)$ for each output. In these simulations, the scaled natural gradient algorithm from [12] was used in batch mode, where $\mu = 0.35$. The average interchannel interference (ICI) was computed at convergence for each algorithm, as given by

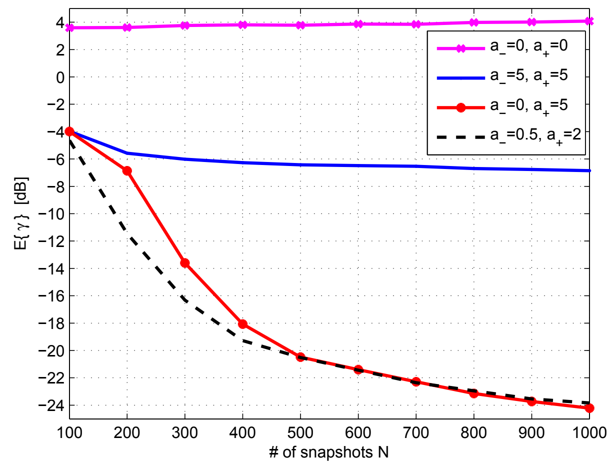

Fig. 3. $E\{\gamma\}$ vs. number of snapshots N for the adaptive piecewise-linear nonlinearity algorithm with various values of a_- and a_+.

$$\gamma = \frac{1}{2m}\left(\sum_{i=1}^{m}\sum_{l=1}^{m}\frac{|c_{il}|^2}{\max\limits_{1\le i\le m}|c_{il}|^2} + \frac{|c_{il}|^2}{\max\limits_{1\le l\le m}|c_{li}|^2}\right) - 1 \quad (13)$$

where $\mathbf{C}$ is the combined (mixing times demixing) matrix. One hundred simulations were averaged to obtain each data point shown.

Fig. 2 shows the average ICI as a function of data record length N, where $\{a_-, a_+\} = \{0, 5\}$ for the piecewise-linear nonlinearity algorithm version. As can be observed, the proposed method has better performance than that of the technique in [5]. This improved performance is due to the single stability condition that governs both nonlinearities used in the piecewise-linear algorithm. So long as $f_i(\theta_i) \neq 0$, we can always find a nonlinearity for $y_i(k)$ that performs separation, as determined by the local stability of the algorithm.

Fig. 3 shows the performance of the natural gradient algorithm with piecewise-linear nonlinearities for different values of a_- and a_+ on the same 10-source mixture data. Clearly, a single nonlinearity ($a_- = a_+$) is inadequate; whereas tuning the nonlinearities to the sources allows separation. Moreover, the proposed algorithm's performance is not very sensitive to the exact value of a_- and a_+ so long as $0 \le a_- < 1 < a_+$.

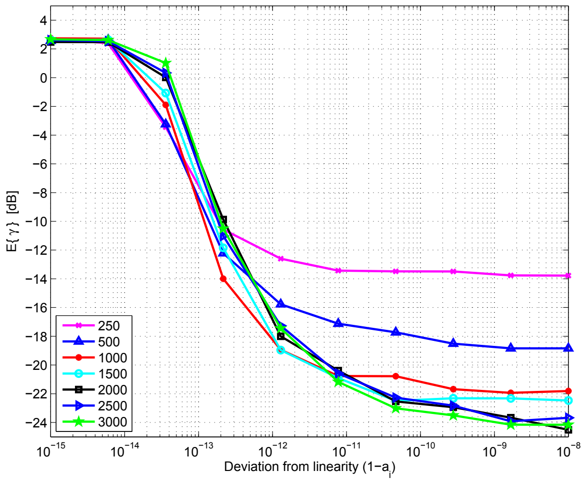

Fig. 4. $E\{\gamma\}$ vs. number of snapshots N for small $(1-a_i)$ values in the FastICA algorithm.

5. PIECEWISE-LINEAR NONLINEARITIES FOR THE FASTICA ALGORITHM

The FastICA algorithm is a useful procedure for blind source separation that requires an output nonlinearity to be chosen. Several possible nonlinearities have been proposed for this algorithm, and all have different local stability properties. The following theorem describes the local stability properties of this algorithm if the piecewise-linear nonlinearity in (1) is used, the proof of which is given in [13].

Theorem 2: *For $g_i(y)$ and $g_i'(y)$ in (1) and (3), let $a_i \neq 1$. Furthermore, assume that $s_i(k)$ is continuous and non-Gaussian-distributed. Then, the local stability properties of the single-unit FastICA algorithm with output nonlinearity $g_i(y)$ are identical to those of the FastICA algorithm with Huber M-estimator-based influence function analyzed in [9, 10].*

The above result indicates that the separating capabilities of FastICA with proposed nonlinearity are likely to be close to those of FastICA with the Huber M-estimator-based influence function where $a_i = 0$. In our exploration of this algorithm's performance, however, we found a rather surprising result: *the algorithm can separate mixtures even if $g_i(y_i)$ is very close to a linear function.*

Fig. 4 shows the average ICI for the same 10-mixture source used in previous simulations for different numbers of snapshots N as a function of the deviation of a_i away from unity, where $\theta_i = 1$ for all output nonlinearities. As can be seen, if $(1-a_i)$ is greater than 10^{-8}, the algorithm performs separation successfully. Moreover, the convergence speed of the FastICA procedure is not compromised; fewer than 10 iterations are often required to achieve good performance. The mean-squared error between $y_i(k)$ and $g_i(y_i(k))$ is $\mathcal{O}((1-a_i)^2)$, implying that $g_i(y_i(k))$ can be used in place of $y_i(k)$ in practical applications.

This result suggests that other nonlinearities commonly used in signal processing applications can be leveraged for FastICA-based separation. For example, μ-law companding is a standard scheme for improving the signal-to-noise ratio (SNR) in telephone transmission systems employing PCM. The nonlinearity used in μ-law encoding is

$$g_\mu(y) \quad = \quad \mathrm{sgn}(y)\frac{\log(1+\mu|y|)}{\log(1+\mu)}, \tag{14}$$

where μ is the compression parameter. We have successfully used the above μ-law nonlinearity in both 8-bit compressive ($\mu = 255$) and decompressive ($\mu = 1/255$) modes within FastICA to separate signal mixtures; typically, performance of such schemes is more than 2dB better than that provided by the cubic nonlinearity and about 1dB worse than the Huber M-estimator-based methods [13].

6. CONCLUSIONS

In many blind source separation and independent component analysis algorithms, the cost function used to measure signal independence is a design parameter. In this paper, we have considered piecewise-linear nonlinearities for use within the natural gradient and FastICA algorithm. For continuous-valued non-Gaussian-distributed signals, the nonlinearity family provides for the local stability of the natural gradient algorithm with simple nonlinearity selection and a single stability-monitoring condition. For the FastICA algorithm, the piecewise-linear nonlinearity can be extremely close to linear without any loss in separation performance.

7. REFERENCES

[1] A. Bell and T. Sejnowski, "An information maximization approach to blind separation and blind deconvolution," *Neural Computation,* vol. 7, pp. 1129-1159, 1995.

[2] S. Amari, A. Cichocki, and H. H. Yang, "A new learning algorithm for blind signal separation," *Adv. Neural Inform. Proc. Syst.,* (The MIT Press, 1996), vol. 8, pp. 757-763, 1996.

[3] A. Hyvärinen and E. Oja, "A fast fixed-point algorithm for independent component analysis," *Neural Computation,* vol. 9, pp. 1483-1492, Oct. 1997.

[4] A. Hyvärinen, "Fast and robust fixed-point algorithms for independent component analysis," *IEEE Trans. Neural Networks,* vol. 10, pp. 626-634, May 1999.

[5] S.C. Douglas, A. Cichocki, and S. Amari, "Multichannel blind separation and deconvolution of sources with arbitrary distributions," *Proc. IEEE Workshop Neural Networks Signal Processing*, Amelia Island, FL, pp. 436-445, Sept. 1997.

[6] S. Choi, A. Cichocki, and S. Amari, "Flexible independent component analysis, *Neural Networks for Signal Processing VIII,* pp. 83-92, 1998.

[7] H. Mathis and S.C. Douglas, "On the existence of universal nonlinearities for blind source separation," *IEEE Trans. Signal Processing,* vol. 50, pp. 1007-1016, May 2002.

[8] K. Kokkinakis and A.K. Nandi, "Generalized gamma density-based score functions for fast and flexible ICA," *Signal Processing,* vol. 87, pp. 1156-1162, May 2007.

[9] J. Chao and S.C. Douglas, "A simple and robust fastICA algorithm using the Huber M-estimator cost function," *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing,* Toulouse, France, vol. 5, pp. 685-688, May 2006.

[10] S.C. Douglas and J. Chao, "Simple, robust, and memory-efficient fastICA algorithms using the Huber M-estimator cost function," *J. VLSI Signal Processing Syst.,* vol. 48, pp. 143-159, Aug. 2007.

[11] S. Amari, T. Chen, and A. Cichocki, "Stability analysis of learning algorithms for blind source separation," *Neural Networks,* vol. 8, pp. 1345-1351, 1997.

[12] S.C. Douglas and M. Gupta, "Scaled natural gradient algorithms for instantaneous and convolutive blind source separation," *IEEE Int. Conf. Acoust., Speech, Signal Processing,* Honolulu, HI, vol. 2, pp. 637-640, Apr. 2007.

[13] J. Chao, "On the design of robust criteria and algorithms for blind source separation," Ph.D. thesis, Southern Methodist University, Dallas, TX, Dec. 2007.