AN ALTERNATIVE NATURAL GRADIENT APPROACH FOR ICA BASED LEARNING ALGORITHMS IN BLIND SOURCE SEPARATION

Andrea Arcangeli, Stefano Squartini, Francesco Piazza

DEIT, Università Politecnica delle Marche
Via Brecce Bianche 60131 Ancona, Italy (Europe)
Phone: +39 0712204453, fax: +390712204835, email: sts@deit.univpm.it

ABSTRACT

In this paper a new formula for natural gradient based learning in blind source separation (BSS) problem is derived. This represents a different gradient from the usual one in [1], but can still considered natural since it comes from the definition of a Riemannian metric in the matrix space of parameters. The new natural gradient consists on left multiplying the standard gradient for an adequate term depending on the parameter matrix to adapt, whereas the other one considers a right multiplication. The two natural gradients have been employed in two ICA based learning algorithms for BSS and it resulted they have identical behavior.

1. INTRODUCTION

Many optimization tasks in literature are based on the minimization (or maximization) of a cost function *l*, properly defined in relation to the addressed problem. The most used technique is the gradient descent (or ascent), that searches the global critical points of *l* by employing the gradient to the surface determined by *l*. A relevant improvement has been recently introduced in literature to speed such a searching up: it consists on substituting the standard gradient with a new one, namely the natural gradient. This is justified by the property of being Riemannian of the space of parameter interested to optimization, as shown in [1], [2]. Therefore a Riemannian metric must be known and in dependence of that the direction of gradient can be adjusted to make convergence faster. The measure of such an adjustment has been derived by Amari [1] in his steepest descent theorem, according to which, given a generic Riemannian space $\mathcal{S}=\{\boldsymbol{\omega}\in\mathbb{R}^n\}$ and a cost function *l* defined on it, the steepest descent direction is:

$$\tilde{\nabla} l(\boldsymbol{\omega}) = G^{-1}(\boldsymbol{\omega})\nabla l(\boldsymbol{\omega}) \tag{1}$$

where $\nabla l(\boldsymbol{\omega})$ is the conventional gradient, $\tilde{\nabla} l(\boldsymbol{\omega})$ is the here-defined natural gradient, and $G^{-1}(\boldsymbol{\omega})$ is the inverse of the metric tensor. It is obvious that derivation of suitable formulas for natural gradient is strictly linked to the problem under study, and it necessarily passes through the calculation of the metric.

Here blind source separation (BSS) is considered, and $G^{-1}(\boldsymbol{\omega})$ has been already derived in literature [1], [3] and corresponding natural gradient learning rules (relative to different algorithms) as well. Such rules have relevant properties, like Newton-like performances at gradient cost, and equivariance [4]. These also occur if relative gradient, defined by Cardoso and Laheld [5], is applied.

The same has been done in other relevant cases, as blind deconvolution problem and adaptation of multilayer neural network. Moreover, it has also been shown that natural gradient online learning is asymptotically Fisher efficient [1].

In all these considerations the Riemannian metric has been given without considering the chance of defining different ones that could lead to derivation of different natural gradients. This is what the authors have tried to show in the present paper, addressing the BSS problem. Moving from the same idea developed by Amari in [1] to define the Riemannian metric, a new tensor $G(\boldsymbol{\omega})$ is formulated and accordingly a new natural gradient formula. The usual one and the novel one have been implemented in two learning approaches for BSS and interesting results obtained.

2. BLIND SOURCE SEPARATION

As addressed in literature, BSS [6], [7] is the problem of recovering the original *m*-sources vector $\mathbf{u}(k)$ mixed by a non-singular *m*-by-*n* matrix $\mathbf{A}$, the mixing matrix, when both are unknown and the only available information is the mixed *n*-signals vector $\mathbf{x} = \mathbf{Au}$, a part from hypotheses of statistical independence and non-gaussianity of input sources and $n > m$. This is achieved by determining an *n*-by-*m* matrix $\mathbf{W}$, the demixing matrix, such that the resulting output $\mathbf{y} = \mathbf{Wx}$ is equal to $\mathbf{u}(k)$ up to permutation and scaling matrices, $\mathbf{P}$ and $\mathbf{D}$ respectively. That can be expressed as:

$$\mathbf{y} = \mathbf{Wx} = \mathbf{WAu} = \mathbf{DPu} \tag{2}$$

The overall structure of BSS is depicted in Figure 1, where the contribute of noise $\mathbf{v}(k)$ is also taken into account. We shall fix $\mathbf{v}(k)$ to zero in the following.

Different methods have been proposed in literature to solve this kind of problem, generally borrowed from Information Theory. We are going to consider here two ICA based approaches: one is based on minimization of mutual information whereas the other, proved to be equivalent to maximum likelihood approach, on maximization of output entropy (the Infomax approach). It has been shown that they yield the same solution under ideal condition of perfect reconstruction.

Both of them consist on adapting the de-mixing matrix by using gradient information deriving from derivatives of suitable cost function $l(\mathbf{x},\mathbf{W})$ respect with the elements of $\mathbf{W}$. The learning rule can be written as follows:

$$\Delta\mathbf{W} = \pm\eta\nabla l(\mathbf{x},\mathbf{W}) \tag{3}$$

where the sign indeterminacy allows to consider both maximization and minimization of $l(\mathbf{x},\mathbf{W})$. Equation (3) obviously assumes different forms in dependence of the cost function chosen. The first ICA based method addressed minimizes the Kullback-Leibler divergence $D_{f\|\tilde{f}}(\mathbf{W})$ between two proper distributions: the probability density function $f_{\mathbf{y}}(\mathbf{y},\mathbf{W})$ parameterized by $\mathbf{W}$, and the corresponding factorial distribution $\tilde{f}_{\mathbf{y}}(\mathbf{y},\mathbf{W}) = \prod_{i=1}^{m}\tilde{f}_{y_i}(y_i,\mathbf{W})$, that is the product of all marginal probability density functions of output $\mathbf{y}$. As derived in [6], [7], the final formula for the standard gradient based learning rule is the following:

$$\Delta\mathbf{W} = \eta\left[\mathbf{I} - \varphi(\mathbf{y})\mathbf{y}^T\right]\mathbf{W}^{-T} \tag{4}$$

where $\varphi(\mathbf{y})$ is the activation function and *T* stands for transposition. Infomax maximizes the entropy $H(\mathbf{z})$ where $\mathbf{z} = g(\mathbf{y})$ and g is the final nonlinearity whose shape depends on the knowledge of the prior distribution of sources. Under this hypothesis, the learning rule is:

$$\Delta\mathbf{W} = \eta\left[\mathbf{W}^{-T} + (1-2\mathbf{z})\mathbf{x}^T\right] \tag{5}$$

as derived in [6], [7], always valid in case of standard gradient. We are going to consider square dimensions for the involved matrices in the following. As already done, the time step will be omitted to simplify notation.

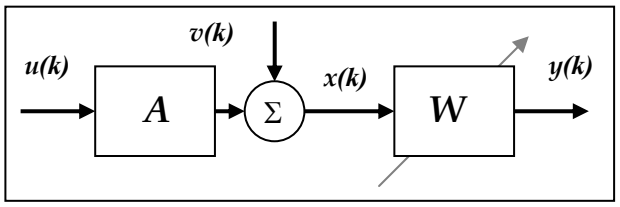

Figure 1. Structure of blind source separation problem.

3. NATURAL GRADIENTS IN MATRIX SPACES

As aforementioned, the natural gradient, as defined in (1), represents the steepest descent direction of cost function $l(\mathbf{x},\mathbf{W})$ to maximize (or minimize) in a Riemannian space. Consequently, this links the natural gradient definition to that one of Riemannian metric over the parameter space. Such a space is the space $\mathcal{W}$ of invertible matrix $\mathbf{W}_{m\times m}$ on $\mathbb{R}$ and it satisfies all properties of Lie groups, if we consider the usual matrix product as operation of multiplication between two elements of the group.

We are going to show two different but related ways to define a Riemannian metric on $\mathcal{W}$, i.e. the tensor field $G(\mathbf{W})$ determining a $\mathcal{W}$-point wise inner product for the tangent space $\mathrm{T}_{\mathbf{W}}\mathcal{W}$ [8]. Both of them will get started from definition of such an inner product by means of a two-step procedure:

1) Definition of the inner product relative to a point in $\mathcal{W}$ (the identity matrix $\mathbf{I}$, the neutral element of the group).
2) Imposing that the $\mathcal{W}$-point wise inner product be invariant to translations in $\mathcal{W}$.

Such a translation operation is nothing but a function that allows to move in the parameter space. It is defined as follows, taking into account that $\mathcal{W}$ is a non-commutative group:

$$\mathfrak{T}_{\mathbf{V}} : \mathcal{W}\to\mathcal{W} \qquad {}_{\mathbf{V}}\mathfrak{T} : \mathcal{W}\to\mathcal{W}$$
$$\mathfrak{T}_{\mathbf{V}}(\mathbf{W}) = \mathbf{W}\cdot\mathbf{V} \qquad {}_{\mathbf{V}}\mathfrak{T}(\mathbf{W}) = \mathbf{V}\cdot\mathbf{W} \qquad \forall\mathbf{V}\in\mathcal{W}$$
Right Translation *Left Translation*

Hence, given a curve $\gamma(t):[-1,1]\to\mathcal{W},\ \gamma(0)=\mathbf{V}$, we can also apply the translation operation to the curve $\gamma(t)$ and to its derivative:

$$\dot{\gamma}(0) = \left.\frac{d\gamma(t)}{dt}\right|_{t=0} \in \mathrm{T}_{\mathbf{V}}\mathcal{W}$$

Canonical inner product has been chosen as inner product at the "starting point" $\mathbf{I}$, with $\mathbf{A}=\dot{\gamma}(0)$, $\mathbf{B}=\dot{\gamma}'(0)$ tangent vectors belonging to $\mathrm{T}_{\mathbf{I}}\mathcal{W}$:

$$\langle\mathbf{A},\mathbf{B}\rangle_{\mathbf{I}} = \sum_{i=1}^{m}\sum_{j=1}^{m}a_{ij}b_{ij}$$

Now, we can proceed with the second step of aforementioned procedure: it requires to distinguish between the two possible translation modes.

3.1. Right Translation

Here we require that:

$$\left\langle\tilde{\mathbf{A}},\tilde{\mathbf{B}}\right\rangle_{\mathbf{W}} = \left\langle\mathfrak{T}_{\mathbf{W}}(\mathbf{A}),\mathfrak{T}_{\mathbf{W}}(\mathbf{B})\right\rangle_{\mathbf{W}} = \left\langle\mathbf{A}\cdot\mathbf{W},\mathbf{B}\cdot\mathbf{W}\right\rangle_{\mathbf{W}} \doteq \langle\mathbf{A},\mathbf{B}\rangle_{\mathbf{I}} \tag{6}$$

The bi-linearity property of inner product and easy calculations let the following hold:

$$\left\langle\tilde{\mathbf{A}},\tilde{\mathbf{B}}\right\rangle_{\mathbf{W}} = \sum_{i,j,k,l}\tilde{a}_{ij}\tilde{b}_{kl}\tilde{G}_{ij,kl}(W) \tag{7}$$

where $\tilde{G}_{ij,kl}(\mathbf{W}) = \left\langle\mathbf{E}_{ij},\mathbf{E}_{kl}\right\rangle_{W}$ and the only $\mathbf{E}_{ij}$ entry not equal to zero is in location (i,j). Observing that (6) must be valid also for $\mathcal{W}$-element like $\mathbf{E}_{ij}$, the following equation

can be derived to describe the tensor defining the Riemannian metric:

$$\tilde{G}_{ij,kl}(\mathbf{W}) = \left\langle \mathbf{E}_{ij}\cdot\mathbf{W}^{-1}, \mathbf{E}_{kl}\cdot\mathbf{W}^{-1}\right\rangle_{\mathbf{I}} = \\ = \delta_{ik}\sum_{c=1}^{m}\mathbf{W}^{-1}{}_{jc}\,\mathbf{W}^{-1}{}_{lc} = \delta_{ik}\tilde{h}^{\mathbf{W}}_{jl} \tag{8}$$

where $\tilde{\mathbf{H}}^{\mathbf{W}} = \left(\tilde{h}^{\mathbf{W}}_{ij}\right) = \mathbf{W}^{-1}\left(\mathbf{W}^{-1}\right)^{T}$.

A part from different notation, what just described is nothing but what Amari derived in [1], as (8) states. This let us anticipate the final formula for natural gradient associated with the aforementioned metric, namely *right natural gradient*:

$$\tilde{\nabla} l(\mathbf{W}) = \nabla l(\mathbf{W})\,\mathbf{W}^{T}\mathbf{W} \tag{9}$$

Derivation of (9) gets started from substituting (8) in (7) and operating as follows:

$$\left\langle \tilde{\mathbf{A}}, \tilde{\mathbf{B}}\right\rangle_{\mathbf{W}} = \sum_{i=1}^{m}\sum_{j=1}^{m}\sum_{k=1}^{m}\sum_{l=1}^{m}\tilde{a}_{ij}\tilde{b}_{kl}\delta_{ik}\tilde{h}^{\mathbf{W}}_{jl} = \sum_{r=1}^{m}\sum_{j=1}^{m}\sum_{l=1}^{m}\tilde{a}_{ij}\tilde{b}_{kl}\tilde{h}^{\mathbf{W}}_{jl}$$

that can be further reduced as

$$\left\langle \tilde{\mathbf{A}}, \tilde{\mathbf{B}}\right\rangle_{\mathbf{W}} = \sum_{r=1}^{m}\left(\left(\tilde{\mathbf{A}}_{r}\right)\tilde{\mathbf{H}}^{\mathbf{W}}\left(\tilde{\mathbf{B}}_{r}\right)^{T}\right).$$

This authorizes us to apply the theorem of steepest descent (1) separately for each row of the gradient $\nabla l(\mathbf{W})$ calculated at $\mathbf{W}$. In such a way we can identify the inverse of $G(\mathbf{W})$ as the inverse of $\tilde{\mathbf{H}}^{\mathbf{W}}$ and write:

$$\left(\tilde{\nabla}_{r} l(\mathbf{W})\right)^{T} = \left(\tilde{\mathbf{H}}^{\mathbf{W}}\right)^{-1}\left(\nabla_{r} l(\mathbf{W})\right)^{T} \tag{10}$$

It can be straightforward derived that $\tilde{\mathbf{H}}^{\mathbf{W}} = \mathbf{W}\mathbf{W}^{T}$ by $\left(\mathbf{W}^{-1}\left(\mathbf{W}^{-1}\right)^{T}\right)^{-1} = \left(\left(\mathbf{W}^{-1}\right)^{T}\right)^{-1}\left(\mathbf{W}^{-1}\right)^{-1} = \left(\left(\mathbf{W}^{-1}\right)^{-1}\right)^{T}\mathbf{W}$.

Substituting this result in (10) and addressing all gradient rows, it results that $\left(\tilde{\nabla} l(\mathbf{W})\right)^{T} = \mathbf{W}^{T}\mathbf{W}\left(\nabla l(\mathbf{W})\right)^{T}$ and consequently (9).

3.2. Left Translation

Here we require that:

$$\left\langle \underset{\sim}{\mathbf{A}}, \underset{\sim}{\mathbf{B}}\right\rangle_{\mathbf{W}} = \left\langle {}_{\mathbf{W}}\mathfrak{T}(\mathbf{A}), {}_{\mathbf{W}}\mathfrak{T}(\mathbf{B})\right\rangle_{\mathbf{W}} = \left\langle \mathbf{W}\cdot\mathbf{A}, \mathbf{W}\cdot\mathbf{B}\right\rangle_{\mathbf{W}} \doteq \left\langle \mathbf{A}, \mathbf{B}\right\rangle_{\mathbf{I}}$$

The following equalities can be derived through similar considerations as before:

$$\left\langle \underset{\sim}{\mathbf{A}}, \underset{\sim}{\mathbf{B}}\right\rangle_{\mathbf{W}} = \sum_{i,j,k,l}\underset{\sim}{a}_{ij}\underset{\sim}{b}_{kl}\underset{\sim}{G}_{ij,kl}(W)$$

$$\underset{\sim}{G}_{ij,kl}(\mathbf{W}) = \delta_{jl}\sum_{c=1}^{m}\mathbf{W}^{-1}{}_{ic}\,\mathbf{W}^{-1}{}_{ck} = \delta_{jl}\underset{\sim}{h}^{\mathbf{W}}_{ik}$$

where $\underset{\sim}{\mathbf{H}}^{\mathbf{W}} = \left(\underset{\sim}{h}^{\mathbf{W}}_{ij}\right) = \left(\mathbf{W}^{-1}\right)^{T}\mathbf{W}^{-1}$ and the new tensor of the new metric is $\underset{\sim}{G}_{ij,kl}(\mathbf{W}) = \left\langle \mathbf{W}^{-1}\cdot\mathbf{E}_{ij}, \mathbf{W}^{-1}\cdot\mathbf{E}_{kl}\right\rangle_{\mathbf{I}}$.

Now we can proceed as done before in right natural gradient case, observing first that:

$$\left\langle \underset{\sim}{\mathbf{A}}, \underset{\sim}{\mathbf{B}}\right\rangle_{\mathbf{W}} = \sum_{c=1}^{m}\left(\left(\underset{\sim}{\mathbf{A}}_{c}\right)^{T}\underset{\sim}{\mathbf{H}}^{\mathbf{W}}\underset{\sim}{\mathbf{B}}_{c}\right)$$

and then applying the steepest descent theorem separately for each column of the gradient $\nabla l(\mathbf{W})$ calculated at $\mathbf{W}$. Hence, we can move from $\underset{\sim}{\nabla}_{c} l(\mathbf{W}) = \left(\underset{\sim}{\mathbf{H}}^{\mathbf{W}}\right)^{-1}\nabla_{c} l(\mathbf{W})$ and calculate $\left(\underset{\sim}{\mathbf{H}}^{\mathbf{W}}\right)^{-1} = \mathbf{W}\mathbf{W}^{T}$, to derive finally the expression of *left natural gradient*:

$$\underset{\sim}{\nabla} l(\mathbf{W}) = \mathbf{W}\mathbf{W}^{T}\nabla l(\mathbf{W}) \tag{11}$$

4. EXPERIMENTAL RESULTS

In this section performances of the two ICA based learning algorithms are analyzed both when natural gradients and standard gradient are applied. Application of (9) and (11) to (4) and (5) leads to the following natural gradient based learning rules:

$$\begin{aligned} &\Delta\mathbf{W} = \eta\left[\mathbf{I} - \varphi(\mathbf{y})\mathbf{y}^{T}\right]\mathbf{W} && \textit{right} \\ &\Delta\mathbf{W} = \eta\mathbf{W}\left[\mathbf{I} - \mathbf{W}^{T}\varphi(\mathbf{y})\mathbf{x}^{T}\right] && \textit{left} \end{aligned} \tag{12}$$

$$\begin{aligned} &\Delta\mathbf{W} = \eta\left[\mathbf{I} + (1-2\mathbf{z})\mathbf{y}^{T}\right]\mathbf{W} && \textit{right} \\ &\Delta\mathbf{W} = \eta\mathbf{W}\left[\mathbf{I} + \mathbf{W}^{T}(1-2\mathbf{z})\mathbf{x}^{T}\right] && \textit{left} \end{aligned} \tag{13}$$

Looking at (12) and (13), it seems that that the equivariance property is not satisfied by the new natural gradient. Indeed, being $\mathbf{C}(k) = \mathbf{W}(k)\mathbf{A}$ the global matrix describing the overall system, we can not specify the updating rule for the global matrix only in function of itself, as conversely it does in case of right natural gradient [3], [7]:

$$\begin{aligned} &\Delta\mathbf{C} = \eta\left[\mathbf{C} - \mathbf{W}\mathbf{W}^{T}\varphi(\mathbf{y})\mathbf{x}^{T}\mathbf{A}\right] && \textit{ICA based method 1} \\ &\Delta\mathbf{C} = \eta\left[\mathbf{C} + \mathbf{W}\mathbf{W}^{T}(1-2\mathbf{z})\mathbf{x}^{T}\mathbf{A}\right] && \textit{ICA based method 2} \end{aligned}$$

However, such property is satisfied again if we describe the system in Figure 1, interpreting $\mathbf{u}(k)$ as a row-vector, instead of a column vector, as done till now. Equation (2) can be now written as:

$$\mathbf{y} = \mathbf{x}\mathbf{W} = \mathbf{u}\mathbf{A}\mathbf{W} = \mathbf{u}\mathbf{D}\mathbf{P}$$

while, the learning rules (4) and (5) become:

$$\begin{aligned} &\Delta\mathbf{W} = \eta\mathbf{W}^{-T}\left[\mathbf{I} - \mathbf{y}^{T}\varphi(\mathbf{y})\right] && \textit{ICA based method 1} \\ &\Delta\mathbf{W} = \eta\left[\mathbf{W}^{-T} + \mathbf{x}^{T}(1-2\mathbf{z})\right] && \textit{ICA based method 2} \end{aligned}$$

allowing us to derive the following ones in global matrix notation (now $\mathbf{C}(k) = \mathbf{AW}(k)$), in case of left natural gradient:

$$\Delta\mathbf{C} = \eta\mathbf{C}\left[\mathbf{I} - \mathbf{y}^T\varphi(\mathbf{y})\right] \quad ICA\ based\ method\ 1$$

$$\Delta\mathbf{C} = \eta\mathbf{C}\left[\mathbf{I} + \mathbf{y}^T(1 - 2\mathbf{z})\right] \quad ICA\ based\ method\ 2$$

The example dealt with considers a system involving $u_1 = \sin(400k)\cos^2(30K)$ $u_2 = \text{sign}(\sin(150k+15\cos(30k)))$ as the two independent sources, while the mixing matrix is $\mathbf{A} = [0.56, 0.79; -0.75, 0.65]$. Simulations have been performed by using the batch version of considered learning algorithms, leaving unchanged the parameter values: number of iterations, learning rate η, number of signal samples and epoch size. Figures 2-3 and Table 1 show how the employment of natural gradient allows to get a relevant improvement respect with the standard gradient, while right and left translation based versions have basically identical behaviour.

Learning algorithms	1st channel S/N [dB]	2nd channel S/N [dB]
Standard gradient Infomax	12.94	9.89
Right natural gradient Infomax	93.39	94.34
Left natural gradient Infomax	96.20	89.77
Standard gradient ICA	14.52	10.22
Right natural gradient ICA	118.93	120.41
Left natural gradient ICA	124.20	102.04

Table 1. S/N ratios in all cases addressed. They are relative to the number of iterations at which natural gradient based algorithms get convergence.

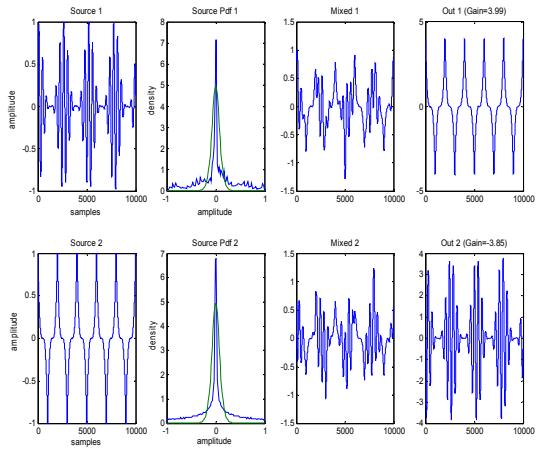

Figure 2. Signals involved in BSS problem: sources, mixed, and recovered. They are valid for all addressed algorithms (at convergence point).

5. CONCLUSION

A mathematical demonstration for derivation of a new natural gradient based learning rule for BSS problem has been provided. The definition of a different Riemannian metric from the usual one allowed to get an original natural gradient that behaves as well as that one proposed in [1], when applied to ICA and Infomax learning algorithms. This fact has relevant implications, since we can assume to derive other different natural gradients once we will be able to define proper Riemannian metrics in the parameter space. Consequently, we could be interested to compare and rate their performances when applied to BSS learning algorithms. These aspects are actually under study. Another point to investigate should be the calculation of original natural gradients in other spaces, as those ones of perceptrons and dynamical linear systems.

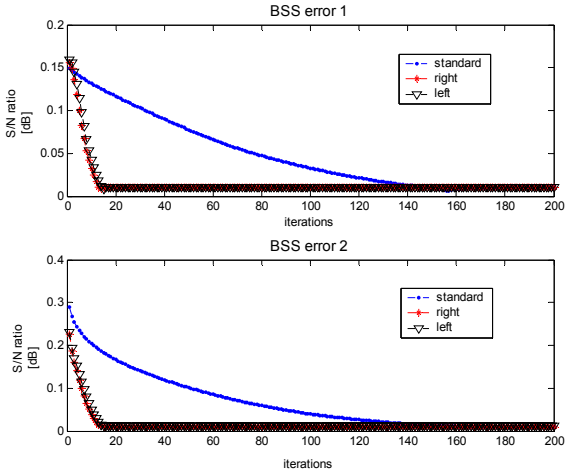

Figure 3. Decaying shapes of S/N ratios of two output signals for three cases under study.

REFERENCES

[1] S.I. Amari, "Natural Gradient works efficiently in learning", *Neural computation*, vol. 10, pp. 251-276,1998.

[2] S. Douglas and S. Amari, "Natural-gradient adaptation", Chapter 2 (pp. 13-61) in S. Haykin (ed.), *Unsupervised Adaptive Filtering*, vol. I, Wiley 2000.

[3] A. Cichocki, and S.I. Amari, *Adaptive Blind Signal and Image Processing*, Wiley&Sons, England, 2002.

[4] J.F. Cardoso, "Learning in Manifolds: the case of Source Separation", *Proceedings of IEEE SSAP*, USA, 1998.

[5] J.F. Cardoso and B.H. Laheld, "Equivariant adaptive source separation", *IEEE Trans. Signal Processing*, vol.44, pp. 3017-3030, December 1996.

[6] T.W.Lee, M. Girolami, A. J. Bell, and T.J.Sejnowski, "A unifying information-theoretic framework for indipendent component analysis", *International journal of computers and mathemathics with applications*, 1999.

[7] S.Haykin, *Neural Networks – A comprehensive Foundation*. Prentice Hall, February 1999.

[8] W. Boothby, *An Introduction to Differentiable Manifolds and Riemannian Geometry*. Academic Press, 2nd ed., 1986.