ESANN'2003 proceedings - European Symposium on Artificial Neural Networks
Bruges (Belgium), 23-25 April 2003, d-side publi., ISBN 2-930307-03-X, pp. 137-142

Neural Net with Two Hidden Layers for Non-Linear Blind Source Separation

Rubén Martín-Clemente$^{\diamond}$, Susana Hornillo-Mellado$^{\diamond}$, José I. Acha$^{\diamond}$,
Fernando Rojas†, Carlos G. Puntonet†

$^{\diamond}$Teoría de la Señal y Com.,†Dpto. de Arquit. y Tecnol. de Comp.
Universities of Sevilla$^{\diamond}$ and Granada† — (SPAIN)
E-mails: {ruben,susanah,acha}@us.es, carlos@atc.ugr.es

Abstract. In this paper, we present an algorithm that minimizes the mutual information between the outputs of a perceptron with two hidden layers. The neural network is then used as separating system in the NonLinear Blind Source Separation problem.

1 Introduction.

Various signal processing applications involve estimating signals of interest from distorted observations. The so-called Blind Source Separation (BSS) is one that consists of retrieving unobserved source signals $s_1(t)$, ..., $s_N(t)$, assumed to be mutually statistically independent, from only N observed signals $x_1(t)$, ..., $x_N(t)$ which are unknown functions or mixtures of the sources. In a vector form, it reads:

$$\mathbf{x}(t) = \mathcal{F}(\mathbf{s}(t)) \tag{1}$$

where $\mathcal{F} : \mathbb{R}^N \to \mathbb{R}^N$ is an unknown reversible mapping, $\mathbf{s}(t) = [s_1(t), \ldots, s_N(t)]^T$ is the source vector and $\mathbf{x}(t) = [x_1(t), \ldots, x_N(t)]^T$ is the so-called observation vector, being the only available data. The task of BSS is that of recovering the sources from the observations.

Starting from the seminal work [3], this problem has been intensively studied over the last decade due to its wide range of applications, from array signal processing to biomedical engineering. In the *nonlinear* mixture case, locally linear BSS methods have been recently explored by Karhunen *et al* [5] using a K-means-clustering-based method. The post-nonlinear case has been dealt by Taleb and Jutten [8], who propose to minimize the mutual information between the estimated sources using a nonlinear system that precedes a linear separating stage. Puntonet *et al* [6] used simulated annealing to avoid undesired minima in the training of a modified Kohonen's network. In addition, Rojas *et al* [7]

ESANN'2003 proceedings - European Symposium on Artificial Neural Networks
Bruges (Belgium), 23-25 April 2003, d-side publi., ISBN 2-930307-03-X, pp. 137-142

proposed a separating system which approximates the nonlinearities of the post-nonlinear mixture model by means of odd polynomials and made use of genetic algorithms for the optimization of the system.

However, with the exception of [4, 9], contributors do not explore the capabilities of multilayer perceptrons (MLPs) to approximate nonlinear mappings. In [9], a two-layer perceptron is used as system to separate the sources. It is ensured that the network is *invertible* by setting the number of neurons in the hidden layer to the number of sources. *This is a very severe constraint that endangers the approximation capabilities of the net.* Nevertheless, if such a constraint were eliminated, one would meet serious mathematical difficulties. Hence, *rather than increasing the number of neurons in the first hidden layer, the solution may be the use of two of more hidden layers.* The purpose of this paper is to present learning rules for approximating the inverse of $\mathcal{F}$ by using MLPs with two hidden-layers. From this point of view, our work complements the one in [9]. The paper is organized as follows: in Sections 2 the basic idea is developed into a practical proposal. In Section 3, learning rules for the MLP are given. Section 4 is devoted to experiments. Section 5 contains our main conclusions.

2 Source Separation

2.1 Basic Assumptions and Notations

The following assumptions hold throughout the paper:

(A1) The sources $s_i(t)$ are mutually statistically independent. That is, at each time t, the elements of $\mathbf{s}(t)$ are independent.

(A2) The mixing mapping $\mathcal{F} : \mathbb{R}^N \to \mathbb{R}^N$ is memoryless, differentiable and bijective.

Here, the basic idea is to approximate the inverse of $\mathcal{F}$ by using the neural network shown in Figure 1, as MLPs have the universal approximation property for smooth continuous mappings. Such a network is described by the equations:

$$\mathcal{F}^{-1}(\mathbf{x}(t)) \approx \mathbf{y}(t) = \mathbf{W}_1\mathbf{g}(\mathbf{u}(t) + \mathbf{b}_1) \tag{2a}$$

being

$$\mathbf{u}(t) = \mathbf{W}_2\mathbf{f}(\mathbf{w}(t) + \mathbf{b}_2) \tag{2b}$$

and

$$\mathbf{w}(t) = \mathbf{W}_3\mathbf{x}(t) \tag{2c}$$

where $\mathbf{W}_1$, $\mathbf{W}_2$ and $\mathbf{W}_3$ are square matrices, $\mathbf{g}(\mathbf{a}) = [g_1(a_1), \ldots, g_N(a_N)]^T$ and $\mathbf{f}(\mathbf{a}) = [f_1(a_1), \ldots, f_N(a_N)]^T$ are any continuous sigmoid-type functions and both $\mathbf{b}_1$ and $\mathbf{b}_2$ are $N \times 1$ vectors. Since the system is memoryless, notice that *we will drop time index t* in the following.

ESANN'2003 proceedings - European Symposium on Artificial Neural Networks
Bruges (Belgium), 23-25 April 2003, d-side publi., ISBN 2-930307-03-X, pp. 137-142

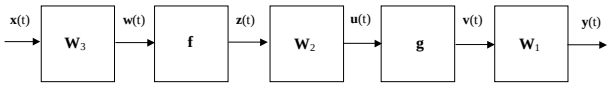

Figure 1: Neural Network Architecture.

2.2 Information-Theoretic Criterion

The guiding principle of *unsupervised* source separation is, in most approaches, to transform the observed data so that the transformed variables are as mutually independent as possible. Even thought that this transformation is not *unique* in the non-linear mixture case, *numerous* experiments show that source separation is feasible and one *separates* the sources (see, for example, [4, 5, 9]), provided that the nonlinear mixture of the sources is smooth and can be undone through a smooth transformation.

The degree of dependence between the outputs is commonly quantified by their *mutual information*, which is defined as:

$$I(\mathbf{y}) = -H(\mathbf{y}) + \sum_{i=1}^{N} H(y_i) \tag{3}$$

where $H(\cdot)$ is the Shannon differential entropy. The key property is that $I(\mathbf{y}) \geq 0$ with equality if and only if the outputs are independent. In practice, usual methods to estimate mutual information, that is, histogram-based estimators and kernel-based estimators are computationally expensive. On the contrary, we can easily calculate $I(\mathbf{y})$ as follows: by using (2), $-H(\mathbf{y})$ is expanded as:

$$-H(\mathbf{y}) = -H(\mathbf{x}) - \sum_{i=1}^{3} \log \mid \mathbf{W}_i \mid - \sum_{i=1}^{N} E[\log \mid g_i'(u_i + b_i^1)\, f_i'(w_i + b_i^2) \mid] \tag{4}$$

where $H(\mathbf{x})$ is the joint entropy of the observed signals, $\mid \mathbf{W}_i \mid = \mid \det(\mathbf{W}_i) \mid$, b_i^j stands for the i-th component of vector $\mathbf{b}_j$ and $g_i'(\cdot)$, $f_i'(\cdot)$ are the first-order derivatives of $g_i(\cdot)$ and $f_i(\cdot)$, respectively. Next, by assuming that the outputs are *pseudo-sphered* (*i.e.*, they are zero-mean unit-variance signals), the marginal entropies $H(y_i)$ can be approximated as (see [2], chapter 5 and [9]):

$$H(y_i) \approx \frac{1}{2}\log(2\pi e) - \frac{(\kappa_3^i)^2}{12} - \frac{(\kappa_4^i)^2}{48} + \frac{3}{8}(\kappa_3^i)^2\,\kappa_4^i + \frac{(\kappa_4^i)^3}{16} \tag{5}$$

where $\kappa_3^i = E[(y_i)^3]$ is the skewness measure of y_i and $\kappa_4^i = E[(y_i)^4] - 3$ equals its kurtosis. To encourage the pseudo-sphering, Tikhonov regularization terms are added to (3). The final cost function then reads:

$$\mathcal{J}(\mathbf{y}) = I(\mathbf{y}) + \lambda_1 \sum_{i=1}^{N} (E[y_i])^2 + \lambda_2 \sum_{i=1}^{N} (E[y_i^2] - 1)^2 \tag{6}$$

ESANN'2003 proceedings - European Symposium on Artificial Neural Networks
Bruges (Belgium), 23-25 April 2003, d-side publi., ISBN 2-930307-03-X, pp. 137-142

3 Learning Rules

In the following, let b_i^j denote the i-th entry of vector $\mathbf{b}_j$. Likewise, we define:

$$\Phi_g[\mathbf{u}] \stackrel{def}{=} -[\frac{g_1''(u_1+b_1^1)}{g_1'(u_1+b_1^1)}, \ldots, \frac{g_N''(u_N+b_N^1)}{g_N'(u_N+b_N^1)}]^T \tag{7}$$

$$\Phi_f[\mathbf{w}] \stackrel{def}{=} -[\frac{f_1''(w_1+b_1^2)}{f_1'(w_1+b_1^2)}, \ldots, \frac{f_N''(w_N+b_N^2)}{f_N'(w_N+b_N^2)}]^T \tag{8}$$

$$\mathbf{D}_g(\mathbf{u}) \stackrel{def}{=} diag(g_1'(u_1+b_1^1), \ldots, g_N'(u_N+b_N^1)) \tag{9}$$

$$\mathbf{D}_f(\mathbf{w}) \stackrel{def}{=} diag(f_1'(w_1+b_1^2), \ldots, f_N'(w_N+b_N^2)) \tag{10}$$

$$h_i = \{-\frac{\kappa_3^i}{2} + \frac{9}{4}\kappa_3^i\,\kappa_4^i\,\}y_i^2 + \{\frac{3}{4}(\kappa_4^i)^2 + \frac{3}{2}(\kappa_3^i)^2 - \frac{1}{6}\kappa_4^i\}\,y_i^3 \tag{11}$$

and

$$\hat{\mathbf{y}} = \mathbf{h} + 2\,\lambda_1\,E[\mathbf{y}] + 4\,\lambda_2\,E[\mathbf{y}\odot\mathbf{y} - \mathbf{1}]\odot\mathbf{y} \tag{12}$$

where $\mathbf{h} = [h_1, \ldots, h_N]$, $\odot$ stands for the Hadamard product and $\mathbf{1}$ is a vector of ones. In order to avoid inverse matrix operations, we use the *natural gradient* rule (see [1], chapter 1) to derive the *unsupervised* learning rule for minimizing the mutual information between the outputs of a perceptron with two hidden layers (see Table 1). It has been verified that the natural gradient adaptation is better than the conventional gradient in terms of computational complexity and theoretical appeal, since natural gradient takes into account the Riemannian metrics of the problem [1]. Due to the lack of space, proofs are left for an extended version of the paper; in any case, they are quite simple (although somewhat cumbersome).

1. $\frac{d}{dt}\mathbf{W}_1 = \{I - E[\hat{\mathbf{y}}\mathbf{y}^T]\}\mathbf{W}_1$
2. $\frac{d}{dt}\mathbf{W}_2 = \{I - E[\Phi_g\,\mathbf{u}^T - \mathbf{D}_g\,\mathbf{W}_1^T\hat{\mathbf{y}}\mathbf{u}^T]\}\mathbf{W}_2$
3. $\frac{d}{dt}\mathbf{W}_3 = \{I - E[\mathbf{D}_f\mathbf{W}_2^T\Phi_g\mathbf{w}^T + \Phi_f\mathbf{w}^T + +\mathbf{D}_f\mathbf{W}_2^T\mathbf{D}_g\mathbf{W}_1^T\hat{\mathbf{y}}\mathbf{w}^T]\}\mathbf{W}_3$
4. $\frac{d}{dt}\mathbf{b}_1 = -E[\Phi_g + \mathbf{D}_g\,\mathbf{W}_1^T\hat{\mathbf{y}}]$
5. $\frac{d}{dt}\mathbf{b}_2 = -E[\mathbf{D}_f\mathbf{W}_2^T\Phi_g + \Phi_f + \mathbf{D}_f\mathbf{W}_2^T\mathbf{D}_g\mathbf{W}_1^T\hat{\mathbf{y}}]$

Table 1: Learning Rules.

ESANN'2003 proceedings - European Symposium on Artificial Neural Networks
Bruges (Belgium), 23-25 April 2003, d-side publi., ISBN 2-930307-03-X, pp. 137-142

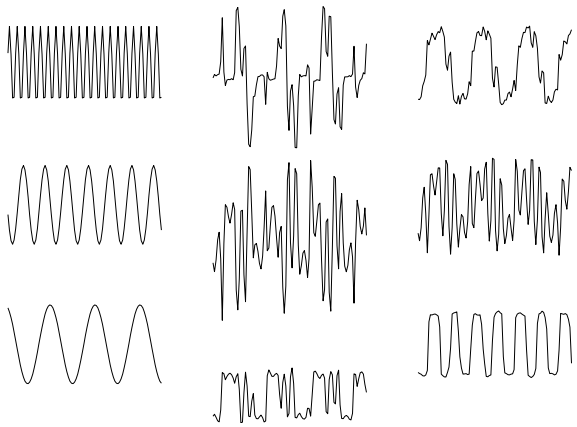

Figure 2: from the left to the right, *a*) source signals *b*) nonlinear mixtures *c*) estimated sources (after 20 sweeps).

4 Computer Simulation

Due to limited space, we shall present in this paper only an illustrative example. The sources, nonlinear mixtures and separated signals are depicted in Figure 2. The mixtures were generated using the model: $\mathbf{x} = \mathbf{A}_1 \tanh(\mathbf{A}_2\,\mathbf{s})$, where

$$\mathbf{A}_1 = \begin{bmatrix} 1.2 & -0.1 & 1.0 \\ -1.2 & -1.6 & 1.4 \\ -0.1 & 0.2 & -0.8 \end{bmatrix}, \quad \mathbf{A}_2 = \begin{bmatrix} 0.5 & -2.2 & 0.6 \\ -0.2 & -0.1 & 0.5 \\ -0.9 & -1.0 & 1.7 \end{bmatrix}$$

The mixing function is strongly nonlinear (no algorithm originally devised for linear mixtures [2] was able to separate the sources). A 1000-sample training set was used for adjusting the network. We employed a batch version of the learning rule (block size and learning rate were set to 100 samples and 0.001 respectively). A little momentum term was also added to speed up the learning process. Both regularization parameters λ_1 and λ_2 were set to 10. The algorithm converges in less than 20 sweeps. The three sources are recognizable after the separation, specially in the Fourier domain. Nevertheless, the system connecting the sources to the estimated sources still exhibits some form of nonlinear behavior. Hence, the output of the separating system contains new frequency components which are not present in the sources. This illustrates the complexity of the nonlinear BSS problem. In fact, other algorithms [2, 9] do not obtain much better results, being only able to separate few sources in practice.

ESANN'2003 proceedings - European Symposium on Artificial Neural Networks
Bruges (Belgium), 23-25 April 2003, d-side publi., ISBN 2-930307-03-X, pp. 137-142

5 Conclusions and Future Research

We have presented learning rules for minimizing the mutual information between the outputs of a perceptron with two hidden layers. This neural network can be successfully applied to the nonlinear BSS problem. Our work was inspired by the fact that neural networks with several hidden-layers can enlarge the nonlinear mapping set that can be approximated. However, the experiments show that networks with two hidden layers are more prone to fall into bad local minima than networks with a single hidden layer, such as the one proposed in [9]. To avoid such undesired minima, we have obtained promising results by using a genetic algorithm [7]. The rough estimation (5) of the marginal entropies may be also responsible for this proliferation of local minima [8]. This point should be studied further in the future.

References

[1] S. I. Amari, "Natural Gradient Works Efficiently in Learning", *Neural Computation*, vol. 10, pp. 251-276, 1998.

[2] A. Hyvärinen, J. Karhunen and E. Oja, "Independent Component Analysis", *John Willey and Sons*, 2001.

[3] C. Jutten and J. Herault, "Blind Separation of Sources, Part I: an adaptive algorithm based on neuromimetic architecture", *Signal Processing*, vol. 24, pp. 1-10, 1991.

[4] J. Karhunen, "Nonlinear Independent Component Analysis". Everson and S. Roberts (Eds.), *"ICA: Principles and Practice"*, pp. 113-134 Cambridge University Press, 2001.

[5] J. Karhunen, S. Malaroiu and M. Ilmoniemi, "Local linear ICA based on clustering", *International Journal of Neural Systems*, vol. 10, no. 6, pp. 439-451, 2000.

[6] C. G. Puntonet, A. Mansour, C. Bauer and E. Lang, "Separation of Sources using Simulated Annealing and Competitive Learning", *Neurocomputing*, Vol. 49, pp. 39- 60, 2002.

[7] F. Rojas, I. Rojas, R. Martín-Clemente and C.G.Puntonet, "Nonlinear Blind Source Separation using Genetic Algorithms", *Proc. ICA 2001*, pp.771-774, San Diego, USA, 2001

[8] A. Taleb and C. Jutten " Source Separation in Post-Nonlinear Mixtures " *IEEE Trans. on Signal Proc.*, Vol. 47, No. 10, pp. 2807-2820, 1999.

[9] H.H. Yang, S.Amari and A. Cichocki, "Information-Theoretic Approach to Blind Separation of Sources in Non-Linear mixture", in *Signal Processing*, vol. 64, No. 3, pp. 291-300, 1998.