A RANDOMLY PERTURBED INFOMAX ALGORITHM FOR BLIND SOURCE SEPARATION

Qi He and Jack Xin

Department of Mathematics,
University of California, Irvine.

ABSTRACT

We present a novel modification to the well-known infomax algorithm of blind source separation. Under natural gradient descent, the infomax algorithm converges to a stationary point of a limiting ordinary differential equation. However, due to the presence of saddle points or local minima of the corresponding likelihood function, the algorithm may be trapped around these "bad" stationary points for a long time, especially if the initial data are near them. To speed up convergence, we propose to add a sequence of random perturbations to the infomax algorithm to "shake" the iterating sequence so that it is "captured" by a path descending to a more stable stationary point. We analyze the convergence of the randomly perturbed algorithm, and illustrate its fast convergence through numerical examples on blind demixing of stochastic signals. The examples have analytical structures so that saddle points or local minima of the likelihood functions are explicit.

Index Terms— Blind source separation, unstable equilibria, randomly perturbed infomax method.

1. INTRODUCTION

Blind source separation (BSS) aims at recovering a set of independent source signals from the observations of their mixtures without knowledge of mixing. It has been an active area of research in signal and image processing literature [1–4] among others. For example, various algorithms have been developed based on minimizing mutual information (MMI) [1], information maximization (infomax) [2] and Maximum Likelihood (ML) approach [5]. For simplicity, we shall consider the instantaneous linear mixture model of d observations of d signals. The observed signals x can be represented by $x = A\,s$, where A is a $d \times d$ invertible mixing matrix, $s = [s_1, \ldots, s_d]^T$ is a source signal with mutually independent components. Assuming that the joint probability density function (pdf) of the source is known as $r(s) = \Pi_{i=1}^d r_i(s_i)$, infomax and ML approaches provide an estimator of A, by maximizing the likelihood function

$$L(A) = E[\log r(A^{-1}x)/|det(A)|].$$

In practice, the pdf of the source signals may not be known, so hypothetical pdf's $q_i(\cdot)$ are used as substitutes of $r_i(\cdot)$. Letting $W = A^{-1}$, the likelihood function becomes

$$J(W) = E[\log q(Wx)] + \log |det(W)|,$$

where $q(x) = (q_1(x_1), \ldots, q_d(x_d))$. Experience has shown that maximizing this alternative likelihood function still yields good demixing matrices as long as the true and hypothetical pdf's do not differ too much. Let $Y = W\,x = (y_1, \ldots, y_d)^T$ denote the recovered source vector and $x(i)$ denote the i-th sample of the mixture signal x. The associated algorithm is given by

$$W(n+1) = W(n) + \nu(I - F(n))W(n), \tag{1}$$

where $F(n) = \frac{1}{L}\sum_{i=nL+1}^{(n+1)L} f(Y^n(i))Y^n(i)^T$, $Y^n(i) = W(n)\,x(i)$, $f(Y) = (f_1(y_1), \ldots, f_d(y_d))^T$ and $f_j(u) = -q_j'(u)/q_j(u)$. We say that W is a demixing matrix if it is such that $WA = P\Lambda$, where P is a permutation matrix and Λ is an invertible diagonal matrix. An equilibrium W_{eq} of this learning rule satisfies the steady state equation:

$$E[f(Y)Y^T] - I = E[f(W\,x)(W\,x)^T] - I = 0, \tag{2}$$

where the expectation can be theoretically carried out with pdf function of the source signals, or approximated from data x. The left hand side of (2) is a function of W. We define the function:

$$g(W) := E\{I - f(W\,x)(W\,x)^T\}, \tag{3}$$

then (2) is just $g(W) = 0$.

To ensure the convergence to a good demixing matrix, one usually examines the stability of the limiting demixing matrix. In the literature, there are many studies on the convergence properties of the algorithm near equilibria [6]. The stability condition in the neighborhood of an equilibrium is well studied, however, the analysis of the global convergence is much more complicated. Recently, [7] constructed examples where global maximizers are spurious equilibria which do not separate the signals at all. Even if the algorithm converges to the desired separating solution under the stability condition, there may exist an unstable equilibrium to cause slow convergence. In [8], an explicit formula of equilibria is found for the two source separation problem when f is a cubic nonlinearity. It shows that a set of saddle points always exists. Though the algorithm does not converge to these "bad" points, they tend to slow down the convergence dramatically. The reason is that the gradient descent as the driving force of the algorithm becomes very small when the iterates wander around these undesired equilibria points. To avoid such problem, one may increase the step size (learning rate) ν to help the algorithm leave these unstable points at the cost of introducing larger errors. To speed up convergence without sacrificing accuracy, some work has been done on variable step size [9]. However, this method is computationally more complex than our proposed remedy below of injecting random perturbations. Systematic expositions of random perturbation methods in the context of stochastic approximation theory can be found in [10]. The idea is to "shake the iterative sequence" until it is "captured" by a path descending to a more stable point, see Figure 1 for an illustration. We will show that a suitable random perturbation of the basic algorithm does not alter its convergence property and equilibria, however it exhibits more robustness to

The work was partially supported by NSF ATD grant DMS-0911277.
Author Emails: qhe.wsu@gmail.com, jxin@math.uci.edu.

initial data and fast convergence. The perturbed algorithm is found numerically to speed up convergence significantly in neighborhoods of unstable equilibria (e.g. saddle points) while maintaining the rate of convergence near stable equilibria.

The paper is organized as follows. Section 2 presents the mathematical analysis of the convergence of our randomly perturbed infomax algorithm. Section 3 gives examples of the unstable equilibrium points of the algorithm. Section 4 investigates the performance of this new algorithm via numerical examples. Concluding remarks are in section 5.

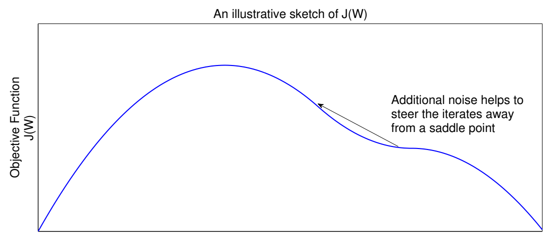

Fig. 1. An illustration of the random perturbation method near a saddle point.

2. ANALYSIS OF CONVERGENCE

The randomly perturbed infomax algorithm is:

$$
\begin{aligned}
&W(n+1) = W(n) + \nu(I - F(n))W(n), n \neq \varepsilon_k, \text{any } k,\\
&W(n+1) = W(n) + \nu(I - F(n))W(n) + b_k\chi_k, n = \varepsilon_k,\\
&F(n) = \tfrac{1}{L}\textstyle\sum_{i=nL}^{(n+1)L} f(Y^n(i))Y^n(i)^T,
\end{aligned}
\tag{4}
$$

where ν is the step size, ε_k is a sequence of integers going to infinity. The mixture signal is divided into blocks in time, each of which contains L data points. Define $t_0 = 0$, $t_n = \sum_{k=0}^{n-1}\nu_k$, $T_k = t_{\varepsilon_{k+1}} - t_{\varepsilon_k}$, such that $T_k \to \infty$. Let $\{\chi_k\}$ be a sequence of independent identically distributed (i.i.d) random matrices in $\mathbb{R}^{d\times d}$ whose entries are uniformly distributed on the interval $[-1, 1]$, and let $\{b_k\}$ be a sequence of positive numbers tending to 0. We assume that the input $x(i)$ is a well-mixed stochastic process and that $f(\cdot)$ is a continuous nonlinear function, the case of most commonly used nonlinearities [3].

To analyze the convergence of this perturbed algorithm, we consider a variant of (4) by replacing ν by ν_n and $F(n)$ by

$$
F(n) = \frac{1}{L_n}\sum_{i=S_n+1}^{S_{n+1}} f(Y^n(i))Y^n(i)^T,
$$

where ν_n is a sequence such that $\sum_{k=1}^{\infty}\nu_k = \infty$ and $\sum_{k=1}^{\infty}\nu_k^2 < \infty$, L_k is an increasing sequence of integers, $S_n = \sum_{i=1}^{n} L_k$. Here we choose decreasing step size ν_k and increasing block size L_k for the simplicity of convergence proof. In practice, we use constant ν and L in (4).

Let us define the associated interpolated process:

$$
\begin{aligned}
&W^0(t) = W(k), \quad \text{for } t_k < t < t_{k+1},\\
&W^n(t) = W^0(t_n + t), \quad t \in (-\infty, \infty),
\end{aligned}
$$

where we define $t_0 = 0$, $t_n = \sum_{k=0}^{n-1}\nu_k$ and $m(t)$ to be the unique value of n such that $t_n \le t < t_{n+1}$. Then $W^n(t)$ can be written as

$$
\begin{aligned}
W^n(t) &= W(n) + \sum_{i=n}^{m(t_n+t)-1}\nu_i Z_i\\
&\quad + \sum_{k:n\le\varepsilon_k\le m(t_n+t)} b_k\chi_k = W(n) + Z^n(t) + p^n(t),
\end{aligned}
$$

where

$$
\begin{aligned}
&p^n(t) = \textstyle\sum_{k:n\le\varepsilon_k\le m(t_n+t)} b_k\chi_k,\\
&Z_n = (I - \tfrac{1}{L_n}\textstyle\sum_{i=S_n+1}^{S_{n+1}} f(Y^n(i))Y^n(i)^T)W(n),\\
&Z^0(t) = \textstyle\sum_{i=1}^{m(t)-1}\nu_i Z_i,\\
&Z^n(t) = Z^0(t_n+t) - Z^0(t_n) = \textstyle\sum_{i=n}^{m(t_n+t)-1}\nu_i Z_i, t \ge 0.
\end{aligned}
$$

We have the following

Lemma 2.1 *Denote* $\beta_k = (I - \frac{1}{L_k}\sum_{i=S_n+1}^{S_{n+1}} f(Y^k(i))Y^k(i)^T) - g(W(k))$, *then* $\sum_k \nu_k|\beta_k| < \infty$ *almost surely. The perturbation process* $p^n(t)$ *converges to zero.*

Proof. Let us first consider that the samples of the mixture signal, $x(i), i = 1, 2, \ldots,$ are i.i.d. Hence $Y^k(i), i = 1, 2, \ldots,$ and $f(Y^k(i)), i = 1, 2, \ldots,$ are also i.i.d sequences with bounded variance. By the uniform law of large numbers [11] and the continuity of f, $\lim_{k\to\infty}(I - \frac{1}{L_k}\sum_{i=S_n+1}^{S_{n+1}} f(Y^k(i))Y^k(i)^T) - g(W(k)) \to 0$. Moreover, by the large deviation theory [12], this convergence is exponentially fast. Hence, $\sum_k \nu_k|\beta_k| < \infty$. We have $p^n(t) \to 0$, since $T_k \to \infty$ and $b_k \to 0$. The results also hold if $x(i)$'s satisfy suitable mixing condition in lieu of i.i.d, see [13], we skip the details here. $\Box$.

With this lemma, we prove the following

Theorem 2.2 *There is a set* N *of probability zero such that for* $\omega \notin N$, *the set of functions* $\{W^n(\omega,\cdot), n < \infty\}$ *is equicontinuous. Let* $W(\omega,\cdot)$ *be the limit of a convergent subsequence. Then it satisfies the ordinary differential equation (ODE)*

$$
\dot{W} = g(W)W. \tag{5}
$$

The iterates $W_n(\omega)$ *converge to the stationary set* S *of ODE (5). The set* S *is a union of finite disjoint compact subsets* $S_1, \ldots, S_N$. *Moreover,* $W_n(\omega)$ *converge to a unique stationary point set* S_i, *consisting of stationary solutions of* (2).

Proof. First, we show the equicontinuity of the interpolated process. Note that

$$
\begin{aligned}
W^n(t) &= W(n) + \sum_{i=n}^{m(t+t_n)-1}\nu_i Z_i + p^n(t)\\
&= W(n) + \sum_{i=n}^{m(t+t_n)-1}\nu_i(g(W(i)) + \beta(i))W(i) + p^n(t)\\
&= W(n) + \sum_{i=n}^{m(t+t_n)-1}\nu_i g(W(i))W(i)\\
&\quad + \sum_{i=n}^{m(t+t_n)-1}\nu_i\beta(i)W(i) + p^n(t)
\end{aligned}
\tag{6}
$$

Since $W^n(\cdot)$ is defined as a piecewise constant function, (6) can be rewritten as

$$
W^n(t) = W(n) + \int_0^t g(W^n(s))W^n(s)ds + p^n(t) + B^n(t) + E^n(t),
\tag{7}
$$

where $E^n(t)$ is the error due to the replacement of the first sum by an integral. Note that $E^n(t) = 0$ at time points $t = t_k - t_n, k > n$, at which the interpolated processes have jumps, and $E^n(t) \to 0$ uniformly in t as $n \to \infty$. In (7), $B^n(t)$ is defined similarly as $Z^n(t)$ by

$$B_n = \sum_{i=n}^{m(t+t_n)-1} \nu_i \beta(i) W(i),$$
$$B^0(t) = \sum_{i=1}^{m(t)-1} \nu_i B_i,$$
$$B^k(t) = B^0(t_k + t) - B^0(t_k) = \sum_{i=k}^{m(t_k+t)-1} \nu_i B_i, t \geq 0.$$

By Lemma 2.1, there is a null set N such that for $\omega \notin N$, $E^n(\omega, \cdot)$ goes to zero uniformly on any bounded interval as $n \to \infty$, also $p^n(t) \to 0$ as $n \to \infty$. Let $\omega \notin N$, then the functions on the right hand side of (7) are equicontinuous in n with the limits of $B^n(\cdot)$ and $E^n(\cdot)$ being zero. By the Arzela- Ascoli Theorem [10], there is a convergent subsequence $\{W^{n_k}(\omega, \cdot\}$ for $\omega \notin N$. We denote the limit by $W(\omega, \cdot)$. It is easily seen that the limit must satisfy the following equation:

$$W(\omega, t) = W(\omega, 0) + \int_0^t g(W(\omega, s))W(\omega, s)ds.$$

Define the cost function $G(W) = -J(W)$. By the derivation of the natural gradient [3], $g(W)$ can be written as the gradient of the objective function $J(W) = -G(W)$, i.e. $g(W) = -\partial G(W)/\partial W$, where $\partial G/\partial W$ is a $d \times d$ matrix whose entries are $\partial G/\partial w_{ij}$. Then the set of stationary points of (5) can be divided into disjoint compact and connected subsets $S_i, i = 0, \ldots, N$, see [10, Sec 5.2]. The derivative of $G(W(\cdot))$ along the solution W of (5) is $-(\partial G/\partial W)^T(\partial G/\partial W) \leq 0$. Using $G(\cdot)$ as a Liapunov function, we can show that $W^n(t)$ must converge to some stationary point. It follows that $W^n(t)$ converge to a unique S_i, otherwise the resulting path would oscillate between distinct S_i's, implying the existence of limit points outside of the stationary points. $\square$

Remark 2.3 *Our proposed algorithm can also be applied to the iterations under the whiteness constraint [14], i.e.,* $E[Y\,Y^T] = I$ *and* $E[f(Y)\,Y^T - Y\,f(Y)^T] = 0$. *In this case, we just need to replace* $F(n)$ *in (4) by* $F(n) = \frac{1}{L}\sum_{i=nL+1}^{(n+1)L}(f(Y^n(i))Y^n(i)^T - Y^n(i)f(Y^n(i))^T + Y^n(i)Y^n(i)^T)$. *The convergence proof remains the same.*

3. EXAMPLES OF UNSTABLE STATIONARY POINTS

In this section, we give examples of unstable equilibria. We know that an equilibrium of the infomax algorithm is a solution of $E[I - f(Y)^T\,f(Y)] = 0$, or a solution of $E[I - Y\,Y^T + f(Y)\,Y^T - Y\,f(Y)^T] = 0$ under whiteness constraint. The stability conditions of W are given in [6] by

$$\begin{array}{lll} 1 + \kappa_i & > 0, & \text{for} 1 \leq i \leq r \\ (1 + \kappa_i)(1 + \kappa_j) & > 1, & \text{for} 1 \leq i < j \leq n, \end{array}$$

or under whiteness constraint

$$\kappa_i + \kappa_j > 0, \;\; \text{for} 1 \leq i < j \leq 0,$$

where $\kappa_i = E[f_i'(y_i)]\,E[y_i^2] - E[f_i(y_i)\,y_i]$. In the following examples, we consider 2 uniformly distributed source signals under cubic nonlinearity. It easily check that demixing W is a stable equilibrium. However, there also exist unstable equilibria.

Example 1. Consider the two dimensional independent source $s = (s_1, s_2)^T$ drawn from uniform distribution on $[-1, 1]$ and the nonlinearity of the learning rule being $f(x) = x^3$. Let the mixing matrix be $A = \begin{bmatrix} 1 & 1 \\ -1 & 1 \end{bmatrix}$. Denote the demixing matrix by $W = \begin{bmatrix} w_{11} & w_{12} \\ w_{21} & w_{22} \end{bmatrix}$. The recovered source is $Y = [y_1 y_2]^T = WAs$.

Plugging these expressions into equation (2) or

$$E\begin{bmatrix} 1 - y_1^4 & y_1^3 y_2 \\ y_2^3 y_1 & 1 - y_2^4 \end{bmatrix} = 0, \tag{8}$$

carrying out the expectations, one solves for W [8]. There are 16 equilibria points listed explicitly in the following table. We see that

Table 1. Equilibrium Points

	w_{11}	w_{12}	w_{21}	w_{22}
A1-8	0	$15^{1/4}/2$	$15^{1/4}/2$	0
unstable	$15^{1/4}/2$	0	0	$15^{1/4}/2$
equilibrium	0	$15^{1/4}/2$	$-15^{1/4}/2$	0
points	$15^{1/4}/2$	0	0	$-15^{1/4}/2$
	$-15^{1/4}/2$	0	0	$15^{1/4}/2$
	0	$-15^{1/4}/2$	$15^{1/4}/2$	0
	$-15^{1/4}/2$	0	0	$-15^{1/4}/2$
	0	$-15^{1/4}/2$	$-15^{1/4}/2$	0
B1- 8	$-5^{1/4}/2$	$5^{1/4}/2$	$5^{1/4}/2$	$5^{1/4}/2$
stable	$5^{1/4}/2$	$-5^{1/4}/2$	$5^{1/4}/2$	$5^{1/4}/2$
equilibrium	$5^{1/4}/2$	$5^{1/4}/2$	$-5^{1/4}/2$	$5^{1/4}/2$
points	$5^{1/4}/2$	$5^{1/4}/2$	$5^{1/4}/2$	$-5^{1/4}/2$
	$-5^{1/4}/2$	$-5^{1/4}/2$	$-5^{1/4}/2$	$5^{1/4}/2$
	$5^{1/4}/2$	$-5^{1/4}/2$	$-5^{1/4}/2$	$-5^{1/4}/2$
	$-5^{1/4}/2$	$5^{1/4}/2$	$-5^{1/4}/2$	$-5^{1/4}/2$
	$-5^{1/4}/2$	$-5^{1/4}/2$	$5^{1/4}/2$	$-5^{1/4}/2$

there are two sets of equilibria. Set A consists of solutions which are unstable equilibria and are not separating matrices. Set B consists of stable equilibria which are demixing matrices up to scaling and permutation. We also calculated the eigenvalues of the Jacobian from linearization of the ODE (5) of $(w_{11}, w_{12}, w_{21}, w_{22})^T$ at each equilibrium point. We observe that the equilibria in set A are saddle

Table 2. Eigenvalues of Jacobian Matrix at Equilibria of Table 1

	λ_1	λ_2	λ_3	λ_4
Eigenvalues at A's	-4.000	-4.000	-1.500	0.500
Eigenvalues at B's	-4.000	-4.000	-2.667	-0.667

points since one of the four eigenvalues is positive. Note that equilibria A still attract the iterates in three directions except the fourth direction along which there is only a weak force to keep the iteration off. That is why the iterates can be trapped around set A for a long time.

Next, we give an example of the algorithm under whiteness constraint.

Example 2. Consider two source signal distributed uniformly on $[-\sqrt{3}, \sqrt{3}]$ with variance 1. The nonlinearity for the learning rule is $f(x) = x^3$. In this case, the transfer matrix WA is a rotation or reflection given by

$$WA = \begin{pmatrix} \cos(\theta) & \sin(\theta) \\ \pm\sin(\theta) & \mp\cos(\theta) \end{pmatrix}.$$

The objective function can be written as

$$J(\theta) = E[\log q(y_1(\theta))] + E[\log q(y_2(\theta))].$$

The algorithm is separating successfully if it has local maxima only at $k\pi/2$. A sketch of the objective function is in Fig. 2. It can be seen that $\theta = 0, \pi/2$ are solutions corresponding to stable stationary points, while $\theta = \pi/4$ is a local minimum corresponding to an unstable stationary point.

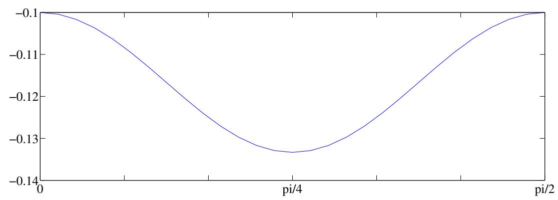

Fig. 2. Objective function for algorithm under whiteness constraint.

4. NUMERICAL SIMULATIONS

From the above section, we see that there are unstable equilibria which are saddle points or local minimizers of J. In the following, we simulate the original algorithm to see how it performs around these unstable equilibria. By comparing with the perturbed algorithm, we find that random perturbations improve the convergence very well. We shall measure the performance by ICI index defined as

$$ICI_k = \left(\sum_{i=1}^{n}\sum_{j=1}^{n} \frac{|w_{ij}|^2}{\max_l w_{il}(k)}\right) - n.$$

Example 1. To simulate Example 1 of section 3, we set step size $\nu = 0.001$ and $L = 10$. Let $\varepsilon_k = k(k-1)/2 + 1$, $b_k = 1/k^2$ and let χ_k be 2×2 random matrices whose elements are drawn independently from uniform distribution on $[-1, 1]$. Let the algorithm start from identity matrix I_2, which is common for most algorithms. However, this is not a good initial value in this example, since I_2 is close to the saddle points in table 3. Figure 3 shows the ICI index after 4000 iterations. The original algorithm does not appear to converge in 4000 iterations (or very slow convergence). The randomly perturbed algorithm converges more rapidly after about 500 iterations. Figure 4 shows the convergence path for each component of W under original algorithm and perturbed algorithm respectively.

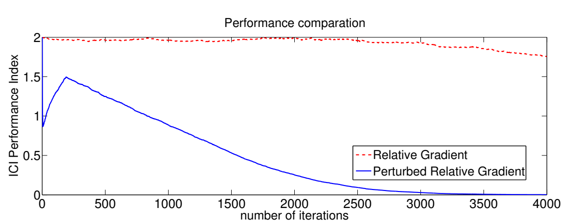

Fig. 3. Comparison of the original and randomly perturbed infomax algorithms built from relative (natural) gradients.

Example 2. We simulate Example 2 of section 3 with the infomax algorithms under whiteness constraint. The step size $\nu = 0.002$, $L = 10$ and the mixing matrix $A = \begin{bmatrix} 1 & -4 \\ 1 & 1 \end{bmatrix}$. We use perturbations $\varepsilon_k = k(k-1)/2 + 1$, $b_k = 1/k$ and χ_k the 2×2 random matrices whose elements are uniformly distributed on $[-1, 1]$. Setting the initial value $W(1) = I_2$. Figure 5 shows the comparison of the ICI index of the original and the randomly perturbed algorithms under whiteness constraint. The original algorithm almost does not converge, while the randomly perturbed one achieves convergence immediately after about 50 iterations.

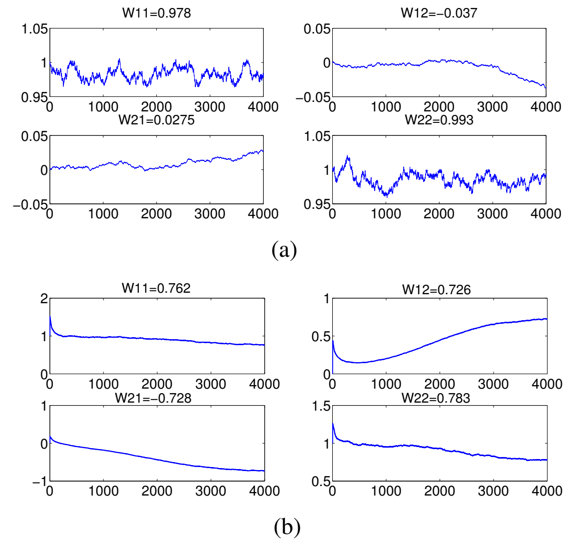

Fig. 4. Convergence paths for demixing matrix W under infomax algorithm (a) and randomly perturbed infomax algorithm (b) built from relative (natural) gradients.

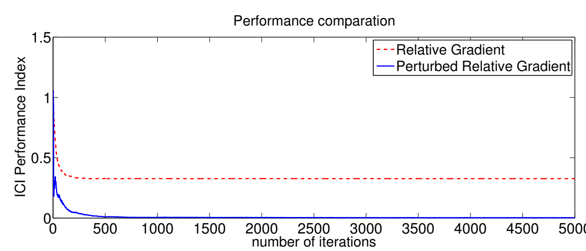

Fig. 5. Comparison of the original (relative/natural gradient) and the randomly perturbed infomax algorithms under whiteness constraint.

5. CONCLUSIONS

We showed that the existence of unstable equilibria may slow down the convergence of infomax learning algorithm tremendously. We analyzed convergence of a randomly perturbed infomax algorithm and showed by numerical simulations that it achieves fast convergence even when starting around unstable equilibria. The selection of noise types for fast convergence, and the extension of our work to convolutive mixtures [15, 16] will be left for future research.

6. REFERENCES

[1] S. Amari, A. Cichocki, and H. H. Yang, "A new learning algorithm for blind signal separation," in *Advances in Neural Information Processing Systems*. 1996, pp. 757–763, MIT Press.

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

[3] A. Cichocki and S. Amari, *Adaptive Blind Signal and Image Processing*, John Wiley and Sons, Ltd., 2002.

[4] J. Liu, J. Xin, and Y. Qi, "A soft-constrained dynamic iterative method of blind source separation," *Multiscale Modeling & Simulation*, vol. 7, no. 4, pp. 1795–1810, 2009.

[5] J.F. Cardoso, "Blind signal separation: statistical principles," *Proceedings of the IEEE*, vol. 86, no. 10, pp. 2009–2025, 1998.

[6] J.F. Cardoso, "On the stability of source separation algorithms," *The Journal of VLSI Signal Processing*, vol. 26, no. 1, pp. 7–14, 2000.

[7] F. Ge and J. Ma, "Spurious solution of the maximum likelihood approach to ICA," *Signal Processing Letters, IEEE*, vol. 17, no. 7, pp. 655–658, 2010.

[8] C.C. Cheung and L. Xu, "Separation of two independent sources by the information-theoretic approach with cubic nonlinearity," in *International Conference on Neural Networks*, 1997, vol. 4, pp. 2239–2244.

[9] T. P. von Hoff, *On the convergence of blind separation and deconvolution*, Ph.D. thesis, ETH, 2000.

[10] H.J. Kushner and G. Yin, *Stochastic Approximation Algorithms and Applications*, vol. 35, Springer, 1997.

[11] W.K. Newey and D. McFadden, "Large sample estimation and hypothesis testing," *Handbook of econometrics*, vol. 4, pp. 2111–2245, 1994.

[12] A. Dembo and O. Zeitouni, *Large deviations techniques and applications*, vol. 38, Springer, 2009.

[13] B.M. Pötscher and I.R. Prucha, "A uniform law of large numbers for dependent and heterogeneous data processes," *Econometrica: Journal of the Econometric Society*, pp. 675–683, 1989.

[14] J.F. Cardoso and B.H. Laheld, "Equivariant adaptive source separation," *IEEE Transactions on Signal Processing*, vol. 44, no. 12, pp. 3017–3030, 1996.

[15] S. C Douglas and M. Gupta, "Scaled natural gradient algorithms for instantaneous and convolutive blind source separation," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2007, vol. 2, pp. II–637.

[16] J. Xin, M. Yu, Y-Y Qi, H-I Yang, and F-G Zeng, "A nonlocally weighted soft-constrained natural gradient algorithm for blind separation of reverberant speech," in *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, 2009, pp. 81–84.