Journal of Machine Learning Research 4 (2003) 1393-1410 Submitted 10/02; Published 12/03

Statistical Dynamics of On-line Independent Component Analysis

Gleb Basalyga BASALYGA@CS.MAN.AC.UK
Magnus Rattray MAGNUS@CS.MAN.AC.UK
Department of Computer Science
University of Manchester
Manchester, M13 9PL, UK

Editors: Te-Won Lee, Jean-François Cardoso, Erkki Oja and Shun-ichi Amari

Abstract

The learning dynamics of on-line independent component analysis is analysed in the limit of large data dimension. We study a simple Hebbian learning algorithm that can be used to separate out a small number of non-Gaussian components from a high-dimensional data set. The de-mixing matrix parameters are confined to a Stiefel manifold of tall, orthogonal matrices and we introduce a natural gradient variant of the algorithm which is appropriate to learning on this manifold. For large input dimension the parameter trajectory of both algorithms passes through a sequence of unstable fixed points, each described by a diffusion process in a polynomial potential. Choosing the learning rate too large increases the escape time from each of these fixed points, effectively trapping the learning in a sub-optimal state. In order to avoid these trapping states a very low learning rate must be chosen during the learning transient, resulting in learning time-scales of $O(N^2)$ or $O(N^3)$ iterations where N is the data dimension. Escape from each sub-optimal state results in a sequence of symmetry breaking events as the algorithm learns each source in turn. This is in marked contrast to the learning dynamics displayed by related on-line learning algorithms for multilayer neural networks and principal component analysis. Although the natural gradient variant of the algorithm has nice asymptotic convergence properties, it has an equivalent transient dynamics to the standard Hebbian algorithm.

Keywords: independent component analysis, statistical mechanics, Hebbian learning, diffusion, natural gradient

1. Introduction

On-line learning algorithms are often used for dealing with very large data sets or in dynamic situations in which data is changing according to a non-stationary process. Independent component analysis (ICA) is often applied under one or both of these conditions and a number of on-line ICA algorithms have been developed (see, e.g. Hyvärinen, 1999). In on-line learning the model parameters are updated after the presentation of each training example. Although there is good understanding of the asymptotic properties of on-line learning, much of the learning takes place far from the asymptotic regime and here a theoretical understanding of the process is often poor. Training methods are typically based on experimental observations in the absence of a successful theoretical analysis of on-line learning in this regime. The efficiency of on-line training is very sensitive to the choice of training parameters such as the learning rate and this dependence can slow down learning and influence the ability of learning to converge successfully to desired states. A deeper theoretical

understanding of the on-line learning process is needed to provide reliable methods for setting the parameters and optimising the training process.

Most classical theoretical results on on-line learning are from stochastic approximation theory (Kushner and Clark, 1978). For a review of recent advances and modern theoretical approaches, see e.g. Saad (1998). Theories describing on-line learning have mainly been developed in two different asymptotic regimes. Most work has been done in the limit of the long times, in which case the model parameters are close to a stable fixed point of the learning dynamics. Here one can work out the asymptotic distribution of the parameters for constant learning rate or study their convergence to a fixed point as the learning rate is reduced according to some annealing schedule. The conditions under which the parameters converge at an optimal rate are quite well understood (White, 1989). Work has also been carried out in the limit of small learning rates, in which case the dynamics can be shown to follow the mean gradient or flow globally (Kushner and Clark, 1978). Unifying these two strands to some extent one can study the long-time global behaviour under an annealed learning schedule (Kushner, 1987).

It is unclear how much practical relevance the classical asymptotic limits have, since often in practical applications learning is not asymptotic in the learning times or in the learning rate. In this work we pursue a different type of asymptotic analysis by considering the limit of large system size. This limit has been studied extensively by researchers applying statistical mechanics methods to the study of learning systems (Engel and Van den Broeck, 2001). For example, the dynamics of gradient decent in multilayer perceptrons (MLPs) (Biehl and Schwarze, 1995, Saad and Solla, 1995) and the dynamics of Sanger's principal component analysis (PCA) algorithm (Biehl, 1994, Biehl and Schlösser, 1998) have been studied in this limit. In these examples the dynamics displays interesting and non-trivial transient behaviour with unstable, sub-optimal fixed points appearing due to a symmetry in the parameter space of the models. The dynamics of natural gradient learning in MLPs has been studied using these techniques, showing that the transient fixed point becomes less serious and transient learning performance is improved compared to standard online gradient descent (Rattray et al., 1998, Rattray and Saad, 1999). Since a natural gradient algorithm has been shown to provide an asymptotically efficient on-line learning algorithm for ICA (Amari et al., 1996, Amari, 1998) we are interested in whether the statistical mechanics approach developed here can help understand its transient performance.

We have recently developed a novel theoretical framework for studying the dynamics of on-line ICA in the limit of large data dimension (Rattray, 2002, Rattray and Basalyga, 2002). We study a Hebbian learning rule, which is a simple and popular algorithm for on-line ICA with nice stability conditions (Hyvärinen and Oja, 1998) and is closely related to a popular fixed-point batch algorithm (Hyvärinen and Oja, 1997). The algorithm is particularly amenable to analysis in the limit of large input dimension because it can be used to extract a small number of independent components from high-dimensional data. We will see later that this allows the dynamics to be represented by a relatively small number of variables when the data dimension becomes large. As well as studying the standard version of the algorithm we also develop a natural gradient variant and study its dynamics. Because the parameter space is constrained to the Stiefel manifold of orthogonal rectangular matrices, the standard equivariant or natural gradient ICA algorithms due to Cardoso and Laheld (1996) and Amari et al. (1996) are not appropriate here. Instead it is possible to use the ideas developed by Edelman et al. (1999) in order to construct an algorithm which follows the gradient on a Stiefel manifold. Our variant uses the gradient defined by Amari (1999) but includes

the orthogonalisation term from Hyvärinen and Oja (1998) in order to keep the model parameters on the Stiefel manifold.

In order to obtain a tractable ICA model, we consider an idealised data set in which a small number of non-Gaussian sources are mixed into a large number of Gaussian sources (Rattray, 2002). By using the methods of statistical mechanics, we provide a solution to the dynamics of Hebbian ICA in the limit of large input dimension. This generalises on previous results (Rattray, 2002, Rattray and Basalyga, 2002), which were limited to the simplest single source case (except for the late time asymptotics, which were solved for the general case). We find that the transient dynamics of Hebbian ICA can be described as a stochastic process in which the system moves through a sequence of metastable fixed points, each of which can be described as a multi-dimensional diffusion in polynomial potential. By solving the dynamics of the multi-source case we can characterise the symmetry breaking process which is critical to performance of the learning process.

It is interesting to observe that the dynamics of ICA is very different from the dynamics of learning in MLPs and in PCA studied previously (Saad and Solla, 1995, Biehl and Schlösser, 1998). In these cases the dynamics was observed to be "self-averaging" so that it followed a smooth trajectory in the limit of large data dimension N and the learning happened on an $O(N)$ time-scale. The ICA dynamics displays significant fluctuations even in this limit and learning occurs on a much slower time-scale, typically requiring of the order of N^2 or N^3 iterations depending on the details. We find that the natural gradient variant of the Hebbian algorithm has very nice asymptotic convergence properties with uniform convergence in the case of equal source statistics. However, the algorithm is shown to have equivalent transient performance to the standard algorithm.

This paper is organised as follows. The data model is described in Section 2. The on-line Hebbian ICA algorithm is introduced in Section 3. In Section 4 we introduce the macroscopic variables which provide a compact description of the dynamics for large N. In Section 5 we show that the learning dynamics near a sub-optimal fixed point close to the initial conditions can be considered as a diffusion process. The transient dynamics through a sequence of metastable states is analysed in Section 6. In Section 7 we introduce the natural gradient version of Hebbian ICA algorithm and study its dynamics. General conclusions are made in Section 8.

2. Data Model

We consider the following idealised linear data model introduced by Rattray (2002). The N-dimensional data x is generated from a noiseless linear mixture of a small number M of non-Gaussian sources s and a large number $N-M$ of uncorrelated Gaussian components, $n \sim \mathcal{N}(0, I_{N-M})$,

$$x = A \begin{bmatrix} s \\ n \end{bmatrix} = A_s s + A_n n,$$

where $A = [A_s\ A_n]$ is the $N \times N$ mixing matrix, I_N denotes an $N \times N$ identity matrix and $\mathcal{N}(a, \Sigma)$ denotes a Gaussian distribution with mean a and covariance matrix Σ. Without loss of generality it can be assumed that the sources each have unit variance. In order to apply the Hebbian ICA algorithm we assume also that the data is already sphered, i.e. the data has zero mean and an identity covariance matrix. This can be achieved for on-line learning by an adaptive sphering algorithm, such as the one introduced by Cardoso and Laheld (1996). These model assumptions lead to the

constraint that A must be an orthogonal matrix, e.g.

$$[A_s\, A_n] \begin{bmatrix} A_s^{\mathrm{T}} \\ A_n^{\mathrm{T}} \end{bmatrix} = A_s A_s^{\mathrm{T}} + A_n A_n^{\mathrm{T}} = I\,, \tag{1}$$

$$\begin{bmatrix} A_s^{\mathrm{T}} \\ A_n^{\mathrm{T}} \end{bmatrix} [A_s\, A_n] = \begin{bmatrix} A_s^{\mathrm{T}} A_s & A_s^{\mathrm{T}} A_n \\ A_n^{\mathrm{T}} A_s & A_n^{\mathrm{T}} A_n \end{bmatrix} = \begin{bmatrix} I & 0 \\ 0 & I \end{bmatrix}. \tag{2}$$

3. On-line Hebbian ICA Learning Rule

The goal of ICA is to find the de-mixing matrix W such that the projections,

$$y \equiv W^{\mathrm{T}} x\,,$$

will coincide with the non-Gaussian sources s up to scaling and permutations. The best possible solution is one in which the K projections will learn as many as possible of the M non-Gaussian sources. Note that specialisation of each projection to a particular source mostly depends on the details of the initial conditions.

We consider a simple Hebbian learning rule (Hyvärinen and Oja, 1998), which extracts non-Gaussian sources from the data by maximising some measure of non-Gaussianity of the projections. The change of the $N \times K$ de-mixing matrix W at the each learning step is given by,

$$\Delta W = \eta\, x \phi(y)^{\mathrm{T}} \sigma + \alpha W (I - W^{\mathrm{T}} W)\,. \tag{3}$$

Here, η is the learning rate and σ is a $K \times K$ diagonal matrix with elements

$$\sigma_{ii} = \mathrm{sign}\left(\mathrm{E}_{s_i}[s_i \phi(s_i) - \phi'(s_i)]\right), \tag{4}$$

which ensures stability of the correct solution as $y_i \to s_i$ (Hyvärinen and Oja, 1998). The first term on the right of (3) maximises some measure of non-Gaussianity of the projections. The second term provides orthogonalisation of the de-mixing matrix. The choice of learning rate η greatly influences the performance of this algorithm. Choosing too large a learning rate results in slow and inefficient learning but choosing too high a value may result in the learning dynamics becoming trapped in a poor solution, as we will see later. The learning dynamics is less sensitive to the choice of the parameter α and we set $\alpha = 0.5$ in simulations. The function $\phi(y) = [\phi(y_i)]$ is some smooth non-linear function which is applied to every component of the vector y. An even non-linearity, e.g. $\phi(y) = y^2$, is usually used to detect asymmetric non-Gaussian signals, while an odd non-linearity, e.g. $\phi(y) = y^3$ or $\phi(y) = \tanh(y)$, is used to extract symmetric non-Gaussian signals.

4. Learning Dynamics for Large Data Dimension

In the case of high-dimensional data (when N becomes very big) it is difficult to analyse the dynamics of the $N \times K$ de-mixing matrix. In order to provide a compact description of the system dynamics in the limit $N \to \infty$, we introduce new macroscopic variables

$$R \equiv W^{\mathrm{T}} A_s \quad \text{and} \quad Q \equiv W^{\mathrm{T}} W\,,$$

the dimension of which are $K \times M$ and $K \times K$ respectively. These overlap matrices contain all necessary information about the relationship between the projections y and the sources s. Therefore

the system can be described by a small number of macroscopic quantities in the limit of large N as long as K and M remain small. We will usually order the indices retrospectively so that the dynamics approaches the optimal solution with $R_{ij} = \delta_{ij}$ (assuming orthogonal W such that $Q = I$). However, it should be remembered that there are equivalent optima related to this solution by a permutation of indices or changes in sign of the components of R. In order to learn successfully the algorithm has to break symmetry and specialise to a particular solution.

Using Equation (1) one can show that,

$$y = W^{\mathrm{T}}(A_s s + A_n n) = Rs + z \; ,$$

where $z \sim \mathcal{N}(0, C)$ and $C = Q - RR^{\mathrm{T}}$. In order to analyse the dynamics we will have to compute expectations with respect to the distribution of y. The covariance matrix C is symmetric and positive definite so that we can always write z in the form $z = L\mu$ where $\mu \sim \mathcal{N}(0, I)$ are Gaussian variables and the matrix L can be found by special decomposition. A particularly useful decomposition for our purposes is the Cholesky decomposition (see, e.g. Press et al., 1992) in which case $C = LL^{\mathrm{T}}$ where L is lower-triangular with non-zero components satisfying the following recurrence relations

$$L_{kk} = \sqrt{C_{kk} - \sum_{j=1}^{k-1} L_{kj}^2} \tag{5}$$

and for the k-th row of L we have

$$L_{ki} = \frac{C_{ki} - \sum_{j=1}^{i-1} L_{ij} L_{kj}}{L_{ii}} \; , \tag{6}$$

for $i = 1, 2, ..., k-1$.

From Equation (3) we can calculate the changes in R and Q after a single learning step,

$$\Delta R = \eta\sigma\phi(y)s^{\mathrm{T}} + \alpha(I - Q)R \; , \tag{7}$$

$$\begin{aligned} \Delta Q &= \eta\sigma(I + \alpha(I - Q))\phi(y)y^{\mathrm{T}} + \alpha^2(I - Q)^2 Q \\ &+ \eta\sigma y\phi(y)^{\mathrm{T}}(I + \alpha(I - Q)) + 2\alpha(I - Q)Q \\ &+ \eta^2\phi(y)x^{\mathrm{T}}x\phi(y)^{\mathrm{T}} \; , \end{aligned} \tag{8}$$

where we used the constraint in Equation (2) to set $x^{\mathrm{T}}A_s = s^{\mathrm{T}}$.

The dynamics is not very sensitive to the exact value of α as long as $\alpha \gg \eta$. We will see later that the learning rate must be chosen very small for large N so that typically we will have this situation in practice. As α increases relative to η, Q approaches I since the orthogonalisation term in Equation (3) dominates. If one defines $Q - I \equiv q/\alpha$ and sets α large relative to η then the fixed point of Equation (8) to leading order is,

$$q = \tfrac{1}{2}\left[\eta\sigma\left(\phi(y)y^{\mathrm{T}} + y\phi(y)^{\mathrm{T}}\right) + \eta^2 N\phi(y)\phi(y)^{\mathrm{T}}\right] \; , \tag{9}$$

where we have dropped terms lower than $O(\eta^2 N)$ and $O(\eta)$. Substituting Equation (9) into Equation (7) leads to the following update equation for R,

$$\Delta R = \eta\sigma\left[\phi(y)s^{\mathrm{T}} - \tfrac{1}{2}\left(\phi(y)y^{\mathrm{T}} + y\phi(y)^{\mathrm{T}}\right)R\right] - \tfrac{1}{2}\,\eta^2 N\phi(y)\phi(y)^{\mathrm{T}}R \; . \tag{10}$$

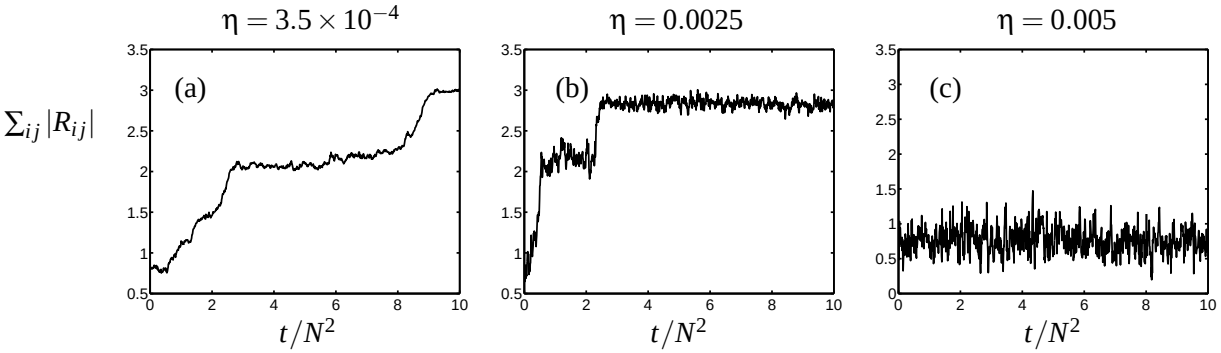

Figure 1: Typical learning dynamics of the Hebbian ICA algorithm for different learning rates. We used the non-linearity $\phi(y) = y^2$ to extract three asymmetrical binary sources with skewness $\kappa_3 = 1.5$ from 100-dimensional data ($K = M = 3$, $N = 100$). Each picture shows the quantity $\sum_{ij} |R_{ij}|$ as it changes over time. For the smallest learning rate ($\eta = 3.5 \times 10^{-4}$) the dynamics looks relatively smooth but it takes time to learn all three sources and the dynamics is localised at a sub-optimal metastable state near $\sum_{ij} |R_{ij}| = 2$ for a significant period of time. With a larger learning rate ($\eta = 0.0025$) the dynamics appears more stochastic and the system is again localised in the same metastable state. For the largest learning rate ($\eta = 0.005$) the fluctuations are so strong that the system remains trapped in a sub-optimal state close to the initial conditions for the entire simulation time.

This simplification procedure is an example of adiabatic elimination of fast variables (see, for example, Gardiner, 1985). In a more rigorous treatment one should consider the mean and covariance of ΔR and ΔQ using the appropriate large N scalings which are described in the next sections. One finds that fluctuations in Q are negligible in the limit of large N and therefore the dynamics of Q can be described by a differential equation in this limit with stable fixed point given by Equation (9).

In Figure 1 we show some typical dynamical trajectories. We observe the following types of fixed points in the R-dynamics:

- Optimal fixed points (see Figure 1a and b)

  Asymptotically, when $y_i \to s_i$, the optimal solution is given by $R^*_{ij} = \delta_{ij}$, which is a fixed point of Equation (10) as $\eta \to 0$. Note that all other possible solutions can be obtained by a trivial permutation of indices and/or changes in sign. Asymptotically the learning rate should be annealed in order to approach this fixed point at an optimal rate. For a detailed account of the asymptotic dynamics of the Hebbian ICA algorithm under an annealed learning rate, see Rattray (2002).

- Sub-optimal fixed points causing trapping near the initial conditions (see Figure 1c)

  Due to the $O(\eta^2)$ fluctuation term in Equation (10), the algorithm has a special class of sub-optimal fixed points near $R = 0$ which causes the presence of a stochastic trapping state near the initial conditions. We will discuss this situation in detail in the next section.

- Transient fixed points (see Figure 1a and b)

  When $T < M$ non-Gaussian sources have been learned, the dynamics can become localised in metastable states somewhere between the initial conditions and the final, optimal fixed point.

In this case the fixed points of Equation (10) are

$$R^*_{ij} = \delta_{ij}\, \mathrm{I}[i \leq T]\,, \tag{11}$$

where

$$\mathrm{I}[\text{predicate}] = \begin{cases} 1 & \text{if predicate is true,} \\ 0 & \text{if predicate is false.} \end{cases} \tag{12}$$

Again, similar fixed points can be obtained from (11) by a simple permutation of indices. In this case the system has to leave such metastable states in order to complete learning. We will analyse the dynamics near these metastable states in Section 6.

5. Stochastic Trapping State near the Initial Conditions

In similar studies of on-line learning, macroscopic quantities like the overlap R usually have a "self-averaging" property such that the variance of these macroscopic quantities tends to zero in the limit $N \to \infty$ (see e.g. Saad and Solla, 1995, Biehl and Schlösser, 1998). A random and uncorrelated choice for A and the initial entries of de-mixing matrix W leads us to expect $R = O(N^{-1/2})$ initially. Larger initial values of R could only be obtained with some prior knowledge of the mixing matrix which we will not suppose. In this case one can no longer assume that fluctuations are negligible as $N \to \infty$. Moreover, as we will see below, the mean and variance of the change in R at each iteration are of the same order. That means that the overlap R does not self-average and the fluctuations have to be considered even in the limit. Therefore, it is more natural to model the on-line learning dynamics near the initial conditions as a diffusion process (see, for example, Gardiner, 1985, van Kampen, 1992). In order to establish a clear picture of the dynamics we have to choose an appropriate scaling for macroscopic quantities and learning parameters. In the following discussion we set $r \equiv R\sqrt{N}$ where r is assumed to be an $O(1)$ quantity.

5.1 Diffusion in a Potential

For a diffusion process the probability density $p(r,t)$ of a random variable $r = [r_{ij}]$ at time t obeys the Fokker-Planck equation

$$\frac{\partial p(r,t)}{\partial t} = -\sum_{ij} \frac{\partial}{\partial r_{ij}} \left(A_{ij}\; p(r,t)\right) + \frac{1}{2} \sum_{ijkl} \frac{\partial^2}{\partial r_{ij} \partial r_{kl}} \left(B_{ijkl}\; p(r,t)\right)\,,$$

where the coefficients A_{ij}, which are called "drift" coefficients, represent the expectation of the change of variable r_{ij} with respect to the stochastic process in question. They can be written as

$$A_{ij} = \mathrm{E}[\Delta r_{ij}] = -\frac{\partial U(r)}{\partial r_{ij}} N^{-p}\,,$$

where N and p are the input dimension and the scaling order for our system and $U(r)$ is some differentiable function of r. This function is analogous to the potential function for the case of a particle undergoing a diffusion in a potential. The coefficients B_{ijkl} are the covariance of the change of variable r_{ij},

$$B_{ijkl} = \mathrm{Cov}[\Delta r_{ij}, \Delta r_{kl}] = \mathrm{E}[\Delta r_{ij} \Delta r_{kl}] - \mathrm{E}[\Delta r_{ij}]\mathrm{E}[\Delta r_{kl}] = D_{ijkl} N^{-p}\,,$$

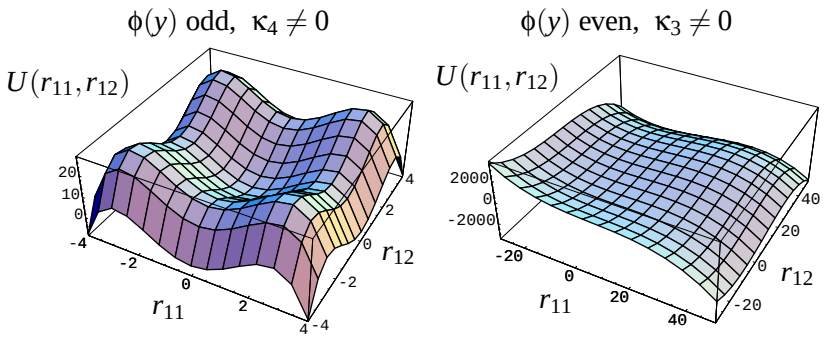

Figure 2: Close to the initial conditions the learning dynamics is equivalent to diffusion in a polynomial potential. For symmetrical source distributions with non-zero kurtosis κ_4 we should use an odd non-linearity in which case the potential is quartic, as shown on the left (for $K = 1$, $M = 2$). For asymmetrical source distributions with skewness κ_3 we can use an even non-linearity in which case the potential is cubic, as shown on the right. Escaping over the minimal barriers results in symmetry breaking, i.e. specialisation to a particular source.

where D_{ijkl} is called the "diffusion term". Usually this would be a matrix but notice that in our case the dynamical variables are in a matrix and therefore the diffusion term has four indices. However, we can think of each pair as a single index in a vectorised system. If we do so then for our case the diffusion term can be considered a diagonal matrix of magnitude D,

$$D_{ijkl} = D\,\delta_{ik}\delta_{jl}\ ,$$

where D is called the "diffusion coefficient". It is typical for the diffusion process that the mean and covariance of the random variable are of the same order $\sim O(N^{-p})$. This means that, for large N, the system is equivalent to diffusion in the potential $U(r)$ with a characteristic time-scale $(\delta t)^{-1} = N^p$.

In our case the potential $U(r)$ has a minimum at $r_{ij} = 0$ surrounded by potential barriers of differing heights. Examples are shown in Figure 2 for an odd and even non-linearity on the left and right respectively. The escape time (the mean first passage time) from the minimum of the potential at $r = 0$ is mainly determined by the effective size of the minimal potential barrier ΔU (see, for example, Gardiner, 1985, van Kampen, 1992),

$$T_{\text{escape}} = A \exp\left(\frac{\Delta U}{D}\right)\ ,$$

where prefactor A is proportional to the characteristic time-scale and depends on the curvature of the potential. Because diffusion through a higher potential barrier is exponentially less likely as the difference between barrier heights increases, the escape time will effectively be inversely proportional to the number of escape points (the number of barriers with the same minimal height ΔU) when the Arrhenius factor $\Delta U/D$ is large.

In the next two sections we will find the form of potential barrier and estimate the escape time from the trapping state for the two main classes of non-linear function $\phi(y)$.

5.2 Odd Non-linearity

If we need to extract symmetrical non-Gaussian signals from the data then we have to use an odd non-linearity, e.g. $\phi(y) = y^3$ or $\phi(y) = \tanh(y)$ are common choices. In this case the appropriate scaling for the learning rate will be $\eta = \nu N^{-2}$, where ν is an $O(1)$ scaled learning rate parameter. After expanding Equation (10) near $r = 0$ we obtain the following expressions for the mean and covariance of the change in r at each iteration,

$$\mathrm{E}[\Delta r_{ij}] = \left(-\tfrac{1}{2}\langle\phi^2(\mu)\rangle\nu^2 r_{ij} + \tfrac{1}{6}\kappa_4^j\langle\phi'''(\mu)\rangle\sigma_{ii}\nu\, r_{ij}^3 \right) N^{-3} + O(N^{-4}) \;, \tag{13}$$

$$\mathrm{Cov}[\Delta r_{ij},\, \Delta r_{kl}] = \langle\phi^2(\mu)\rangle\nu^2\, \delta_{ik}\delta_{jl}\, N^{-3} + O(N^{-4}) \;, \tag{14}$$

where κ_4^j is the fourth cumulant of the j-th source distribution (measuring kurtosis) and brackets denote averages over a Gaussian variable $\mu \sim \mathcal{N}(0, I)$. In this case the system can be described by a Fokker-Planck equation for large N with a characteristic time-scale $(\delta t)^{-1} = N^3$. To compute the expectations we have made use of the Cholesky decomposition of y as described in Equations (5) and (6). This allows us to remove the dependence of the parameters in the Gaussian averages by writing $y = L\mu$ where the lower diagonal (i.e. non-zero) elements of L are found to be,

$$L_{ij} = \delta_{ij} - \left(\tfrac{1}{2}\delta_{ij}\sum_m^M r_{im}^2 + \sum_m^M r_{im} r_{jm} \right) N^{-1} + O(N^{-2}) \quad \text{for} \quad i \geq j \;.$$

The dynamics is equivalent to diffusion in the following potential

$$U(r) = \sum_{i=1}^{K}\sum_{j=1}^{M} \left(\tfrac{1}{4}\langle\phi^2(\mu)\rangle\, \nu^2 r_{ij}^2 - \tfrac{1}{24}\kappa_4^j\langle\phi'''(\mu)\rangle\sigma_{ii}\nu\, r_{ij}^4 \right)$$

with a diagonal diffusion matrix of magnitude $D = \langle\phi^2(\mu)\rangle\nu^2$. Notice that the potential is a sum of contributions $U(r) = \sum_{ij} U(r_{ij})$ which depend only on a single element in r. Since the diffusion matrix is diagonal, this means that each element of r undergoes an independent diffusion process equivalent to the one-dimensional case described by Rattray (2002). The effective size of the potential barriers for each element r_{ij} is given by,

$$\frac{\Delta U(r_{ij})}{D} = \frac{3\langle\phi^2(\mu)\rangle\nu}{8|\kappa_4^j\langle\phi'''(\mu)\rangle|} \;. \tag{15}$$

Escape over one such potential barrier results in the corresponding source j being learned by projection i. This breaks the symmetry of the system as one projection specialises to a particular source. Once this happens the system can again become trapped in a metastable state with other sources remaining unlearned. The dynamics in the neighbourhood of this more general class of fixed point is described in Section 6.

The shape of the potential for an example with two sources (with equal kurtosis) and one projection ($K = 1$, $M = 2$) is shown on the left of Figure 2. In this case we have four minimal potential barriers which the system has to overcome. For the special case when all non-Gaussian sources have the same kurtosis $\kappa_4^i = \kappa_4$ ($i = 1, 2, \ldots, M$), the escape time, i.e. the mean first passage time for a single source to be learned, is given by,

$$T_{\text{escape}}^{\text{odd}} \propto \frac{N^3}{2MK} \exp\left[\frac{3\,\langle\phi^2(\mu)\rangle\,\nu}{8\,|\kappa_4\langle\phi'''(\mu)\rangle|} \right],$$

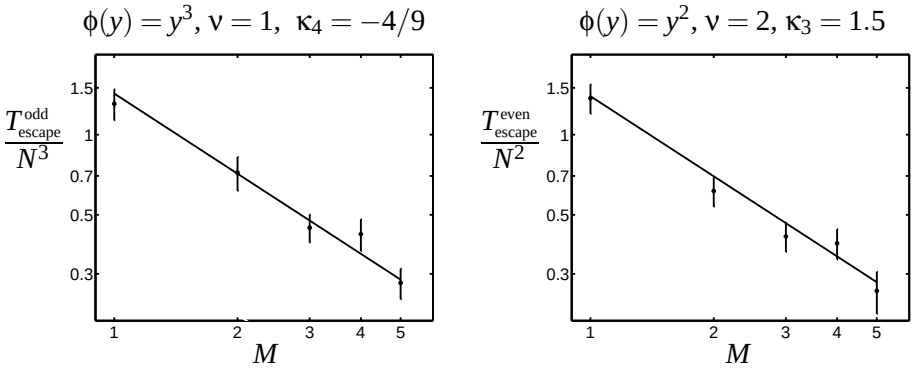

Figure 3: Dependence of the escape time (time to learn at least one source signal) on the number of non-Gaussian sources M for Hebbian ICA with one projection ($K = 1$). The solid line shows the slope predicted by the theory for comparison. The points with error bars denote the simulation results for $N = 50$ averaged over 50 experiments.

where $\sigma = \text{sign}(\langle \phi'''(\mu) \rangle \kappa_4)$ is a necessary condition for successful learning. Notice that the time-scale for escape diverges with the learning rate. On the left of Figure 3 we show numerical simulations for this situation. We have generated data of dimension $N = 50$ with M non-Gaussian sources each with a uniform distribution and we extract a single projection $K = 1$. The figure shows the dependence of the escape time on the number of non-Gaussian sources, on a log-log plot. The solid line shows the inverse scaling with M predicted by the theory and this is consistent with the simulation results. A source was considered learned when the associated overlap R_{ij} was observed to be greater than a threshold magnitude (0.1 greater than the position of the potential barrier).

If the sources have different kurtosis then the algorithm will be most likely to learn the source with highest kurtosis first since this will correspond to the escape point with the lowest potential barrier. The mean first passage time will be dominated by the contribution from this barrier for large learning rates and the escape time will not depend on the number of sources.

5.3 Even Non-linearity

If the non-Gaussian signals are asymmetrical, then we can use an even non-linearity, for example $\phi(y) = y^2$. In this case the appropriate scaling for the learning rate is $\eta = \nu N^{-3/2}$. After expanding Equation (10) near $r = 0$ we find that the mean and covariance of the change in r at each iteration are given by (to leading order in N^{-1}),

$$\mathrm{E}[\Delta r_{ij}] = (-\tfrac{1}{2}\langle \phi^2(\mu) \rangle \nu^2 r_{ij} + \tfrac{1}{2}\kappa_3^j \langle \phi''(\mu) \rangle \sigma_{ii} \nu r_{ij}^2 - \tfrac{1}{2}\langle \phi(\mu) \rangle^2 \nu^2 \sum_{l \neq i}^{K} r_{lj}) N^{-2}, \tag{16}$$

$$\mathrm{Cov}[\Delta r_{ij},\, \Delta r_{kl}] = \langle \phi^2(\mu) \rangle \nu^2 \, \delta_{ik} \delta_{jl} N^{-2}, \tag{17}$$

where κ_3^j is the third cumulant of the j-th source distribution (third central moment), which measures skewness, and brackets denote averages over Gaussian variables $\mu \sim \mathcal{N}(0, I)$. Again the system can be described by a Fokker-Planck equation for large N but now with shorter characteristic time-scale

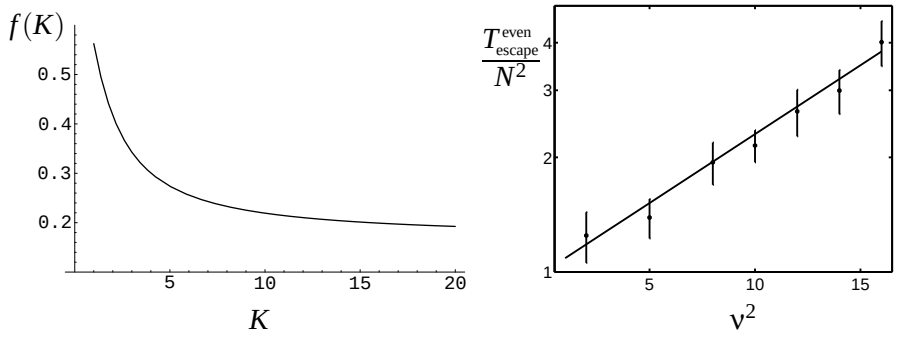

Figure 4: Dependence of effective size of barrier of number of projections K for the specific choice of the even non-linear function $\phi(\mu) = \mu^2$ is shown on the left. Dependence of the escape time on learning rate ν for the one-dimensional Hebbian ICA ($K = M = 1$) is shown on the right. The solid line shows the trend predicted by the theory. The points with error bars denote the simulation results for $\kappa_3 = 1.5$ and $N = 50$ averaged over 50 simulations.

$(\delta t)^{-1} = N^2$. The system is locally equivalent to a diffusion process in the cubic potential

$$U(r) = \sum_{i=1}^{K}\sum_{j=1}^{M}(\tfrac{1}{4}\langle\phi^2(\mu)\rangle\nu^2 r_{ij}^2 - \tfrac{1}{6}\kappa_3^j\langle\phi''(\mu)\rangle\sigma_{ii}\nu r_{ij}^3 + \tfrac{1}{2}\langle\phi(\mu)\rangle^2\nu^2\sum_{l\neq i}^{K} r_{lj}r_{ij}) \ ,$$

with a diagonal diffusion matrix of magnitude $D = \langle\phi^2(\mu)\rangle\nu^2$ as before. In this case the sum in the potential contains "cross-terms" which depend on more than one element in r. The dynamics is therefore not equivalent to the one-dimensional case and features of the potential will depend on the particular value of K and M considered, making analysis less straightforward than for the odd non-linearity.

The shape of the potential for an example with two sources of equal skewness and one projection ($K = 1$, $M = 2$) is shown on the right of Figure 2. In this case we have a ledge in the potential with two points of minimum height ΔU (escape points). In the general case we find that the effective size of the minimal barriers is given by,

$$\frac{\Delta U}{D} = \frac{f(K)\,\nu^2}{\langle\phi^2(\mu)\rangle\ (\kappa_3^i)^2} \ ,$$

where the function $f(K)$ has a complex form which depends on the choice of non-linear function $\phi(\mu)$. For example, the shape of this function for $\phi(\mu) = \mu^2$ is shown on the left in Figure 4. We see that the size of potential barrier decreases with increasing number of projections K and this appears to be a general feature of the function. This suggests that parallel algorithms for extracting asymmetrical signals may prove more efficient than deflationary ones which separate one signal at a time.

For the case of a single projection ($K = 1$) the potential does decompose into a sum of independent terms and each component of r evolves independently. For the case of sources with equal skewness $\kappa_3^i = \kappa_3$ ($i = 1, 2, \ldots, M$), the mean first passage time, i.e. the time until one of the source signals is learned, is then given by,

$$T_{\text{escape}}^{\text{even}} \propto \frac{N^2}{M}\exp\left[\frac{1}{12}\left(\frac{\langle\phi^2(\mu)\rangle\nu}{\kappa_3\langle\phi''(\mu)\rangle}\right)^2\right]. \tag{18}$$

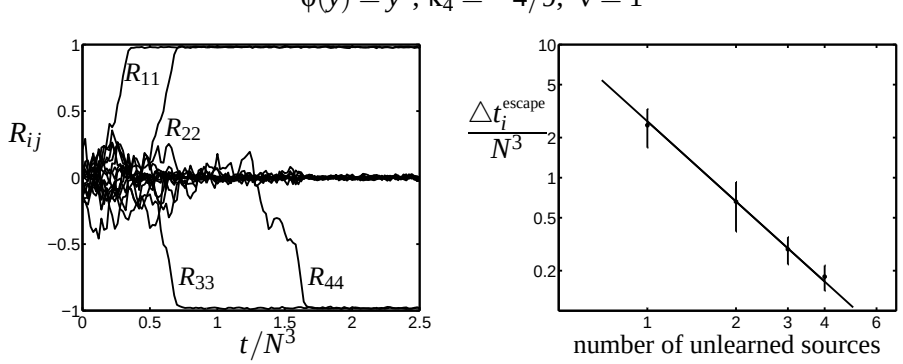

Figure 5: Transient dynamics of the Hebbian ICA algorithm for $K = M = 4$ and $N = 50$. The left plot shows the simulation results for a single run. Dependence of time required to learn next source signal on the number of unlearned sources is shown on the right plot using a log-log scale. The solid line shows the trend predicted by the theory. The points with error bars denote the simulation results averaged over 10 simulations.

Numerical results from simulations of this scenario with $N = 50$ are shown on the right of Figures 3 and 4. The asymmetrical sources used in the simulations are binary and each have the same skewness with $\kappa_3 = 1.5$. In Figure 3 we show the escape time as a function of the number of sources M on a log-log plot. The solid line shows the inverse scaling predicted by the theory and is consistent with the experimental results. On the right of Figure 4 we show how the escape time (on a log scale) depends on the learning rate parameter ν. The slope of the solid line is the theoretical prediction from Equation (18) and is consistent with the simulation results.

6. Transient Dynamics

Consider the more general situation when $T < M$ non-Gaussian sources have already been learned by the system. The corresponding fixed points of Equation (10) are $R^*_{ij} = \delta_{ij}\, \mathrm{I}\,[i \leq T]$, where $\mathrm{I}\,[i \leq T]$ is defined by (12). We have already considered the special case when $T = 0$ which is appropriate close to the initial conditions when no sources have yet been learned. A typical learning dynamics proceeds by passing through these states one by one until all the sources are learned. We show such a dynamical trajectory on the left of Figure 5. The indices have been labeled retrospectively so that the sources are learned in order although this labeling is clearly arbitrary and only chosen for notational convenience. It appears that the typical time to learn each source increases as more sources are learned, a phenomenon which is explained below.

We introduce new $O(1)$ scaled variables,

$$\nu = \eta\, N^d, \;\; v = (R - R^*)\sqrt{N},$$

where N is the input dimension and d is the scaling order for the learning rate. For the case of an even non-linearity we set $d = \frac{3}{2}$ while for the case of an odd non-linearity we choose $d = 2$. In our new variables the fixed point is $v = 0$. We can compute the mean and covariance of these variables as we did for the r variables close to the initial conditions in the previous Section. In the present case it is convenient to consider four categories of variables separately.

1. $i \leq T,\ j \leq T$.

$$\begin{aligned} \mathrm{E}[\Delta v_{ij}] &= \left[-\left(\xi_i + \tfrac{1}{2}\xi_j\right) v_{ij} - \tfrac{1}{2}\,\xi_i v_{ji}\right] \nu N^{1-p} + O(N^{-p})\,, \\ \mathrm{Cov}[\Delta v_{ij},\, \Delta v_{kl}] &= O(N^{-p})\,. \end{aligned} \tag{19}$$

2. $i > T,\ j \leq T$

$$\mathrm{E}[\Delta v_{ij}] = -\tfrac{1}{2}\xi_j v_{ij} \nu N^{1-p} + O(N^{-p})\,, \quad \mathrm{Cov}[\Delta v_{ij},\, \Delta v_{kl}] = O(N^{-p})\,. \tag{20}$$

3. $i \leq T,\ j > T$

$$\mathrm{E}[\Delta v_{ij}] = -\xi_i v_{ij} \nu N^{1-p} + O(N^{-p})\,, \quad \mathrm{Cov}[\Delta v_{ij},\, \Delta v_{kl}] = O(N^{-p})\,. \tag{21}$$

4. $i > T,\ j > T$

$$\mathrm{E}[\Delta v_{ij}] = -\frac{\partial U(v)}{\partial v_{ij}} N^{-p} + O(N^{-p-1})\,, \quad \mathrm{Cov}[\Delta v_{ij},\, \Delta v_{kl}] = D\,\delta_{ik}\delta_{jl} N^{-p} + O(N^{-p-1})\,. \tag{22}$$

In these equations, p is the scaling order of our system with $p = 2$ for the even non-linearity and $p = 3$ for the odd non-linearity. We define,

$$\xi_i = \sigma_{ii} \left(\mathrm{E}_{s_i}[s_i \phi(s_i) - \phi'(s_i)]\right)\,.$$

In Equation (22) the exact expressions for the potential $U(v)$ and for the diffusion coefficient D have the same form as those which were found in Section 5 given by Equations (13), (14), (16) and (17).

In the first three groups of variable we observe that the fluctuations are of negligible order, so that the dynamics can be described by linear differential equations in the large N limit with a relatively fast time-scale of $\delta t^{-1} = N^{p-1}$. Equations (19), (20) and (21) therefore converge exponentially to the fixed point as long as the condition in Equation (4) is met. However, the variables in the fourth group display a similar diffusive dynamics to that considered in the previous Section. The dynamics for these variables is completely equivalent to the r-dynamics close to the initial conditions. Therefore we observe the same behaviour in these variables, with localisation at the fixed point until one component escapes over the potential barrier resulting in another source being learned. Once all the sources are learned we effectively have $T = M$ and only the first three groups of variables above remain. In this case we can increase the learning rate to an $O(N^{-1})$ quantity in principle without the stochastic effects dominating, but then the learning rate should be annealed as described by Rattray (2002) in order to converge asymptotically to the optimal solution.

The picture on the left in Figure 5 is a good illustration of the transient dynamics. We show numerical simulations for the typical dynamics of a Hebbian ICA algorithm extracting four ($K = M = 4$) uniformly distributed non-Gaussian sources from an $N = 50$ dimensional data set. On the right we show a log-log plot of the time $\triangle t_i^{\text{escape}}$ required to learn the next source signal in the case when i non-Gaussian sources have already been learned by the system and we see that it is consistent with the expected trend shown by the solid line. The total learning time for extracting all the non-Gaussian sources in this case will be

$$T_{\text{escape}}^{\text{total}} = \sum_{i=0}^{M-1} \triangle t_i^{\text{escape}} \propto \exp\left[\frac{\Delta U}{D}\right] \sum_{i=0}^{M-1} \frac{1}{2(M-i)^2}\,,$$

where $\Delta U/D$ is given by Equation (15).

7. Natural Gradient ICA

Natural gradient algorithms have been developed for ICA which use the structure of the parameter space to define a Riemannian gradient descent direction (Amari et al., 1996). Along with closely related relative gradient algorithms (Cardoso and Laheld, 1996) these methods provide some advantages over standard gradient descent methods, such as greater simplicity, robustness and asymptotic efficiency (Amari, 1998). However, these algorithms have mainly been defined for the special case where the mixing matrix is square and invertible.

The algorithm we use here searches the space of tall thin orthogonal matrices. This allows it to extract a relatively small number of independent components from a high-dimensional data set possibly containing Gaussian components. Standard natural gradient ICA algorithms are not appropriate in this case and we therefore need a different approach. One possibility would be to use a parameterisation of the set of orthogonal matrices. This approach is considered by Moon and Gunther (2002) who provide an interesting reinterpretation of natural gradient as a pullback. This allows them to define natural gradient algorithms for various structured matrices. Although they restrict their attention to square, invertible matrices their ideas could be extended to tall thin matrices. However, the available parameterisations appear to be quite complex in this case and computing the gradient even more so.

The approach of Moon and Gunther (2002) is to use a set of coordinates which are intrinsic to the manifold. An alternative approach is to use the original variables subject to constraints, i.e. work in the space of tall thin matrices W but impose the orthogonality constraint,

$$W^{\mathrm{T}}W = I \; .$$

This is the approach taken by Edelman et al. (1999) and it leads to a much more straightforward gradient definition for ICA which is described by Amari (1999). The constraint surface is known as a Stiefel manifold and for a function $F(W)$ defined on the Stiefel manifold, the "natural" gradient of F at the point W of the manifold is defined by

$$\tilde{\nabla}_W F = \nabla_W F - W \nabla_W F^{\mathrm{T}} W, \tag{23}$$

where the standard gradient $\nabla_W F$ is the K-by-M matrix of partial derivatives of F with respect to the elements of W. The loss function used in Hebbian ICA is some non-quadratic function of the projections $F(y)$ and the standard gradient of this function is given by

$$\nabla_W F = x\phi(y)^{\mathrm{T}}\sigma \; .$$

Then, according to Equation (23), the natural gradient of this function on the Stiefel manifold will be (Amari, 1999),

$$\tilde{\nabla}_W F = x\phi(y)^{\mathrm{T}}\sigma - W\sigma\phi(y)y^{\mathrm{T}}.$$

A disadvantage of using non-intrinsic variables is that the algorithm is not guaranteed to stay on the manifold. This is especially problematic for stochastic gradient algorithms which only approximately follow the gradient direction. We therefore add the same orthogonalisation term used in the standard Hebbian algorithm. The natural gradient algorithm then has the following form,

$$\Delta W = \eta \left[x\phi(y)^{\mathrm{T}}\sigma - W\sigma\phi(y)y^{\mathrm{T}} \right] + \alpha W(I - W^{\mathrm{T}}W) \; .$$

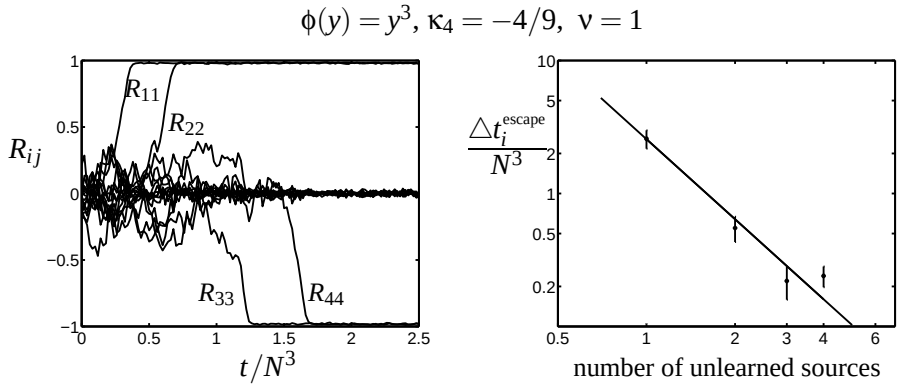

Figure 6: Transient dynamics of the natural gradient Hebbian ICA algorithm for $K = M = 4$ and $N = 50$ (compare with Figure 5). The left plot shows the simulation results for a single run. Dependence of time required to learn next source signal on the number of unlearned sources is shown on the right plot using a log-log scale. The solid line shows the trend predicted by the theory. The points with error bars denote the simulation results averaged over 10 simulations.

The update increment for the overlap matrices $R \equiv W^{\mathrm{T}} A_s$ and $Q \equiv W^{\mathrm{T}} W$ at every learning iteration is found to be,

$$\begin{aligned}
\Delta R &= \eta \left(\sigma \phi(y) s^{\mathrm{T}} - y \phi(y)^{\mathrm{T}} \sigma R \right) + \alpha (I - Q) R \,, \\
\Delta Q &= \eta \sigma (I - Q + \alpha (I - Q) - \alpha (I - Q) Q) \phi(y) y^{\mathrm{T}} + \alpha^2 (I - Q)^2 Q \\
&+ \eta \sigma y \phi(y)^{\mathrm{T}} (I - Q + \alpha (I - Q) - \alpha Q (I - Q)) + 2\alpha (I - Q) Q \\
&+ \eta^2 \phi(y) x^{\mathrm{T}} x \phi(y)^{\mathrm{T}} \,.
\end{aligned}$$

After adiabatic elimination of the Q variables by a similar procedure as we carried out for the case of Hebbian ICA (see Section 3) we have the following dynamical equations for the overlaps,

$$\Delta R = \eta \sigma \left(\phi(y) s^{\mathrm{T}} - y \phi(y)^{\mathrm{T}} R \right) - \tfrac{1}{2} \eta^2 N \phi(y) \phi(y)^{\mathrm{T}} R \,.$$

The typical learning dynamics of this natural gradient version of Hebbian algorithm is shown on the left of Figure 6, where we used the odd non-linearity $\phi(y) = y^3$ to extract four symmetrical sources with kurtosis $\kappa_4 = -4/9$ from 50-dimensional data ($K = M = 4$, $N = 50$).

Following the procedure outlined in Section 6 we can expand near the general fixed points with $T \leq M$ sources learned. Using the same variables we expand around $v = 0$ and find the following results

1. $i \leq T,\ j \leq T$.

$$\mathrm{E}[\Delta v_{ij}] = [-(\xi_i + \xi_j)]\, v_{ij} \nu N^{1-p} + O(N^{-p}) \,, \quad \mathrm{Cov}[\Delta v_{ij},\, \Delta v_{kl}] = O(N^{-p}) \,. \tag{24}$$

2. $i > T,\ j \leq T$

$$\mathrm{E}[\Delta v_{ij}] = -\xi_j v_{ij} \nu N^{1-p} + O(N^{-p}) \,, \qquad \mathrm{Cov}[\Delta v_{ij},\, \Delta v_{kl}] = O(N^{-p}) \,.$$

3. $i \leq T,\ j > T$

$$\mathrm{E}[\Delta v_{ij}] = -\xi_i v_{ij} \nu N^{1-p} + O(N^{-p}) , \quad \mathrm{Cov}[\Delta v_{ij},\ \Delta v_{kl}] = O(N^{-p}) .$$

4. $i > T,\ j > T$

$$\mathrm{E}[\Delta v_{ij}] = -\frac{\partial U(v)}{\partial v_{ij}} N^{-p} + O(N^{-p-1}) , \quad \mathrm{Cov}[\Delta v_{ij},\ \Delta v_{kl}] = D\ \delta_{ik}\delta_{jl} N^{-p} + O(N^{-p-1}) . \quad (25)$$

As before, $p = 2$ for the even non-linearity and $p = 3$ for the odd non-linearity and the potential in the last case is the same as for standard Hebbian ICA. We see that the equations for the first set of variables ($i \leq T,\ j \leq T$) in Equation (24) are simplified in comparison to Equation (19) and no longer contain cross-terms. This means, for example, that the algorithm will enjoy uniform asymptotic convergence if all sources have identical statistics and $M = K$. Generally speaking the eigenvalues determining the convergence of the variables in the first three groups have lower variance and the asymptotic convergence of the natural gradient algorithm will be faster than that of the Hebbian algorithm. However, Equations (25) and (22) are identical and therefore the transient dynamics of the algorithms will be very similar. These are the variables which provide the rate limiting factor and learning time-scale during the transient.

The plot on the right of Figure 6 shows the escape time from each of the transient fixed points encountered during the dynamics. These simulation results confirm that the transient dynamics is very similar to the standard Hebbian algorithm results (see Figure 5) as predicted by our theory.

8. Conclusion

The dynamics of on-line ICA learning has been studied in the limit of large data dimension. We have analysed a Hebbian learning algorithm which is appropriate for extracting a prescribed number of components from high dimensional data possibly containing Gaussian components. We also studied a natural gradient variant of the algorithm which uses the gradient defined on the Stiefel manifold of orthogonal matrices.

We find that the learning time-scale of both algorithms is mainly determined by the transient dynamics. Learning takes place by a sequence of symmetry breaking steps in which a new source is learned and these steps can be described as a diffusion and escape process. The learning time-scale is found to be longer than expected from the analysis of related algorithms such as on-line back-propagation and Sanger's PCA algorithm (e.g. Saad and Solla, 1995, Biehl and Schlösser, 1998). To learn each symmetric source typically requires of the order of N^3 learning iterations while to learn an asymmetrical source using an even non-linearity typically requires of the order of N^2 learning iterations. Both algorithms exhibit equivalent transient dynamics and we only find an advantage in using the natural gradient variant asymptotically.

Acknowledgments

This work was supported by an EPSRC award (ref. GR/M48123).

References

S.-I. Amari. Natural gradient works efficiently in learning. *Neural Computation*, 10:251–276, 1998.

S.-I. Amari. Natural gradient learning for over- and under-complete bases in ICA. *Neural Computation*, 11(8):1875–1883, 1999.

S.-I. Amari, A. Cichocki, and H. H. Yang. A new learning algorithm for blind source separation. In D. S. Touretzky, M. C. Mozer, and M. E. Hasselmo, editors, *Neural Information Processing Systems 8*, pages 757–763. MIT Press, Cambridge, 1996.

M. Biehl. An exactly solvable model of unsupervised learning. *Europhysics Letters*, 25:391–396, 1994.

M. Biehl and E. Schlösser. The dynamics of on-line principle component analysis. *Journal of Physics A*, 31:L97–L103, 1998.

M. Biehl and H. Schwarze. Learning by on-line gradient descent. *Journal of Physics A*, 28:643–656, 1995.

J.-F. Cardoso and B. Laheld. Equivariant adaptive source separation. *IEEE Transactions on Signal Processing*, 44:3017–3030, 1996.

A. Edelman, T. A. Arias, and S. T. Smith. The geometry of algorithms with orthogonality constraints. *SIAM Journal on Matrix Analysis and Applications*, 20(2):303–353, 1999.

A. Engel and C. Van den Broeck. *Statistical Mechanics of Learning*. Cambridge University Press, 2001.

C. W. Gardiner. *Handbook of Stochastic Methods*. Springer-Verlag, New York, 1985.

A. Hyvärinen. Survey on independent component analysis. *Neural Computing Surveys*, 2:94–128, 1999.

A. Hyvärinen and E. Oja. A fast fixed-point algorithm for independent component analysis. *Neural Computation*, 9:1483–1492, 1997.

A. Hyvärinen and E. Oja. Independent component analysis by general non-linear Hebbian-like learning rules. *Signal Processing*, 64:301–313, 1998.

H. J. Kushner. Asymptotic global behaviour for stochastic approximation and diffusions with slowly decreasing noise effects: Global minimization via monte carlo. *SIAM Journal of Applied Mathematics*, 47:169–185, 1987.

H. J. Kushner and D. S. Clark. *Stochastic Approximation Methods for Constrained and Unconstrained Systems*. Springer-Verlag, New York, 1978.

T. K. Moon and J. H. Gunther. Contravariant adaptation on structured matrix spaces. *Signal Processing*, 82(10):1389–1410, 2002.

W. H. Press, S. A. Teukolsky, and W. T. Vetterling. *Numerical Recipes in C*. Cambridge University Press, 2nd edition, 1992.

M. Rattray. Stochastic trapping in a solvable model of independent component analysis. *Neural Computation*, 14:421–435, 2002.

M. Rattray and G. Basalyga. Scaling laws and local minima in Hebbian ICA. In T. G. Dietterich, S. Becker, and Z. Ghahramani, editors, *Advances in Neural Information Processing Systems 14*, pages 495–501. MIT Press, Cambridge, 2002.

M. Rattray and D. Saad. Analysis of natural gradient descent for multilayer neural networks. *Physical Review E*, 59:4523–4532, 1999.

M. Rattray, D. Saad, and S.-I. Amari. Natural gradient descent for on-line learning. *Physical Review Letters*, 81:5461–5464, 1998.

D. Saad, editor. *On-line learning in neural networks*. Cambridge University Press, 1998.

D. Saad and S. A. Solla. Exact solution for on-line learning in multilayer neural networks. *Physical Review Letters*, 74:4337–4340, 1995.

N. G. van Kampen. *Stochastic Processes in Physics and Chemistry*. Elsevier, Amsterdam, 1992.

H. White. Learning in artificial neural networks: A statistical perspective. *Neural Computation*, 1 (4):425–464, 1989.