Available online at www.sciencedirect.com
SCIENCE DIRECT®
Neurocomputing 68 (2005) 130–160
NEUROCOMPUTING
www.elsevier.com/locate/neucom

Unsupervised learning with stochastic gradient

Harold Szu, Ivica Kopriva*

Department of Electrical and Computer Engineering, George Washington University, 801, 22nd St. NW, Washington, DC 20052, USA

Received 24 February 2004; received in revised form 19 October 2004; accepted 23 November 2004
Available online 25 February 2005
Communicated by T. Heskes

Abstract

A stochastic gradient is formulated based on deterministic gradient augmented with Cauchy simulated annealing capable to reach a global minimum with a convergence speed significantly faster when simulated annealing is used alone. In order to solve space-time variant inverse problems known as blind source separation, a novel Helmholtz free energy contrast function, $H = E - T_0 S$, with imposed thermodynamics constraint at a constant temperature T_0 was introduced generalizing the Shannon maximum entropy S of the closed systems to the open systems having non-zero input–output energy exchange E. Here, only the input data vector was known while source vector and mixing matrix were unknown. A stochastic gradient was successfully applied to solve inverse space-variant imaging problems on a concurrent pixel-by-pixel basis with the unknown mixing matrix (imaging point spread function) varying from pixel to pixel.
Published by Elsevier B.V.

Keywords: Stochastic optimization; Cauchy annealing; Blind source separation; Helmholtz free energy

1. Introduction

Gradient optimization is generally incapable of reaching global minimum of the functional with multiple minimums [32]. A stochastic optimization known as simulated annealing [1,39,40], is guaranteed to reach global minimum but with the

*Corresponding author. Tel.: +202 994 5508; fax: +202 994 0227.
E-mail address: ikopriva@gwu.edu (I. Kopriva).

0925-2312/$ - see front matter Published by Elsevier B.V.
doi:10.1016/j.neucom.2004.11.010

very low speed of convergence. Geman and Geman proved in 1984 the convergence guaranteed to find the global minimum by means of classical Gaussian annealing to be an exceedingly slow admissible cooling schedule, $T_a(t) = T_0/\log t$. Their proof used the Metropolis annealing algorithm to generate random walks based on Gaussian distribution [17]. Szu in 1986 [39,40] had extended Geman–Geman convergence proof for the case of the Cauchy noise with unbounded variance combining naturally both Gaussian random walks with Levi random flights achieving the admissible cooling schedule $T_c(t) = T_0/t$. In this paper, we have augmented the classical gradient optimization with the fast fluctuating term using the rapid Cauchy annealing cooling schedule. We coined this approach the stochastic gradient optimization. One important application of the derived stochastic gradient optimization aimed to be unsupervised learning applied to the solution of the highly non-stationary linear inverse problems known as blind source separation (BSS) [11,6,4,3,10,9,12,13,23,24,26,33,46]. In this regard we have introduced a novel Helmholtz free energy contrast function, $H = E - T_0S$, with the imposed thermodynamics constraint at a constant temperature T_0 generalizing the Shannon maximum entropy S of the closed systems to the open systems having non-zero input–output energy exchange E. Following BSS terminology for linear data models, only the input data vector was known while the source vector and mixing matrix were unknown. In comparison with a number of the cost functions for BSS already proposed we have demonstrated a feature of the Helmholtz free energy cost function to have global minimum at the solution of the linear inverse problem. That enabled the applicability of the proposed cost function to solve the BSS problems when the unknown mixing matrix varied from measurement to measurement. In this paper, we have successfully applied a stochastic gradient optimization to solve inverse space-variant imaging problems on a concurrent pixel-by-pixel basis with the unknown mixing matrix (imaging point spread function) varying from pixel to pixel.

The organization of the paper is as follows. In Section 2, we have introduced the BSS problem as well as the Helmholtz free energy cost function with the classical gradient solution for the BSS problem. Section 3 gives the convergence proofs for both Cauchy and Gaussian annealing, along with their differences in free space and in gradient potential wells. The stochastic gradient is also introduced in Section 3. Performances of the 2-dimensional Cauchy and Gaussian annealing search algorithms as well as performances of stochastic gradient algorithm with Cauchy and Gaussian cooling schedule were compared with multiple minimums on the objective function. Section 4 gives more detailed description and illustration of the Helmholtz free energy $H = E - T_0S$ applied on the solution of both linear and nonlinear BSS problems. Comparison has been carried out with the adaptive independent component analysis (ICA) algorithms for linear [3,6,10,33] and post-nonlinear [46] mixtures. The conclusion is given in Section 5. For readers' convenience, Appendix A provides a derivation of the higher-dimensional Cauchy annealing algorithm based on the transformation of the higher-dimensional Cauchy pdf from Cartesian to hyper-spherical coordinates. Biological conjecture of the unsupervised learning based on the minimum of the Helmholtz free energy is given in Appendix B.

2. Helmholtz free energy cost function and blind source separation problem

The linear BSS problem is to find a solution to the linear inverse problem

$$\mathbf{x}(r) = \mathbf{A}(r)\mathbf{s}(r) \tag{1}$$

in terms of the mixing matrix $\mathbf{A}$ and source vector $\mathbf{s}$ given the data vector $\mathbf{x}$ only [11,6,4,3,10,9,12,13,23,24,26,33]. Here r represents the generalized coordinate emphasizing that both mixing matrix and source vector are r coordinate variant. Data model (1) can for example represent multispectral image [41,43] where $r(p,q) = 1 \ldots P \times Q$ $(p = 1 \ldots P,\ q = 1 \ldots Q)$ represents pixel location in the image with the size of P × Q pixels. Because we have focused our attention on imaging applications, the positivity constraints were imposed on the data vector, source vector, and mixing matrix. Without loss of generality we assume here $\mathbf{x}, \mathbf{s} \in R_0^{+N}$, $\mathbf{A} \in R_0^{+N\times N}$, where R_0^+ is a set of positive real numbers including zero and N represents a number of sensors and a number of sources. Generally, a solution is given in terms of the de-mixing matrix $\mathbf{W}$

$$\mathbf{u}(r) = \mathbf{W}(r)\mathbf{x}(r), \tag{2}$$

where $\mathbf{W} \cong \mathbf{A}^{-1}$. Hence, the algorithm initially presented in [41,43] and further elaborated here finds solution of the non-stationary BSS problem at the minimum of the to-be-defined Helmholtz free energy cost function by allowing both the mixing matrix and source vector to be r-variant. The independent component analysis (ICA) algorithms defined in [11,6,4,3,10,9,12,13,23,24,26,33] solve the problem on the statistical basis assuming the unknown mixing matrix to be r-invariant (Fig. 1).

In order to motivate our further work we first postulate with the Lyapunov convergence proof that a mammals' simultaneous (supervised and unsupervised) learning capability is achieved by minimizing the Helmholtz thermodynamic energy, $H = E - T_0 S$, at a constant cybernetic temperature T_0. This is the so-called homeostatic warm-blooded animal learning theory proposed by Szu [38].

Theorem 1. (Homeostatic learning). *It is assumed that neurodynamics are governed by the second law of thermodynamics in terms of the Helmholtz free energy of an open system in an isothermal dynamic balance*

$$H(s_1, \ldots, s_n; w_1, \ldots, w_n) = Energy(s_1, ..., s_n) - T_0\, Entropy(w_1, \ldots, w_n), \tag{3}$$

where the thermal reservoir temperature T_0 was assumed to be constant. It is also assumed that scalable local gradient dynamics are given with

$$\frac{\mathrm{d}u_i}{\mathrm{d}t} = -\frac{\partial\, Energy}{\partial v_i} \tag{4}$$

that corresponds with the minimum energy in the Hopfield–Grossberg–Kohonen sense and $v_i = \sigma(u_i)$ is standard artificial neural network (ANN) sigmoid output [16,18,20,21,34]. It is further assumed that learning of the de-mixing matrix is

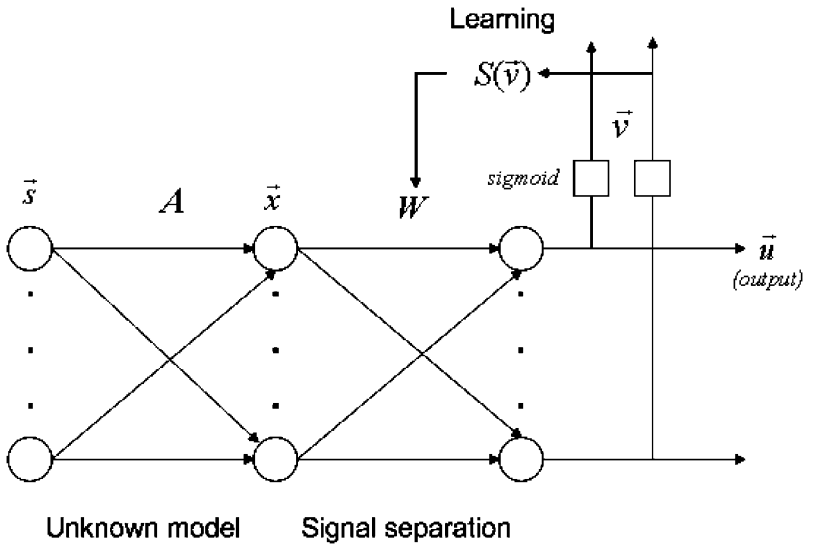

Fig. 1. Independent component analysis (ICA) artificial neural network (ANN) that performs the blind source separation task through maximization of the a posteriori entropy function that has been proven [6] to factorize the joint pdf of the ANN outputs. Consequently, a neighborhood data are required in order to make the joint pdf factorization possible, which indirectly assumed the space-invariant nature of the mixing.

governed by the natural gradient [3,4,10] *MaxEnt algorithms*

$$\mathrm{d}\mathbf{W}/\mathrm{d}t = \left\langle \frac{\mathrm{d}\,Entropy}{\mathrm{d}\mathbf{W}}\,\mathbf{W}^{\mathrm{T}}\mathbf{W} \right\rangle. \tag{5}$$

Then, both supervised and unsupervised learning occured concurrently at the Lyapunov equilibrium at the minimum of the thermodynamic Helmholtz free energy

$$\frac{\mathrm{d}H}{\mathrm{d}t} = \frac{\mathrm{d}\,Energy}{\mathrm{d}t} - T_0\,\frac{\mathrm{d}\,Entropy}{\mathrm{d}t} \leqslant 0. \tag{6}$$

Proof. By means of a monotonic logic $\mathrm{d}v_i/\mathrm{d}u_i > 0$ and by means of (4) we prove

$$\begin{aligned} \mathrm{d}E/\mathrm{d}t &= \textstyle\sum_i (\mathrm{d}\,Energy/\mathrm{d}v_i)(\mathrm{d}v_i/\mathrm{d}u_i)(\mathrm{d}u_i/\mathrm{d}t) \\ &\cong -\textstyle\sum_i (\mathrm{d}\,Energy/\mathrm{d}v_i)^2(\mathrm{d}v_i/\mathrm{d}u_i) \leqslant 0. \end{aligned} \tag{7}$$

Also by using (5) we prove

$$\mathrm{d}\,Entropy/\mathrm{d}t = \left\langle \frac{\mathrm{d}\,Entropy}{\mathrm{d}\mathbf{W}}\,\mathbf{W}^{\mathrm{T}}\,\mathbf{W} \right\rangle (\mathrm{d}\mathbf{W}/\mathrm{d}t) = \left(\left\langle \frac{\mathrm{d}\,Entropy}{\mathrm{d}\mathbf{W}}\,\mathbf{W}^{\mathrm{T}}\,\mathbf{W} \right\rangle \right)^2 \geqslant 0. \tag{8}$$

In (8) the Euclidean gradient $\mathrm{d}\,Entropy/\mathrm{d}\mathbf{W}$ was corrected by the metric tensor $\mathbf{W}^{\mathrm{T}}\mathbf{W}$ in order to obtain the natural or relative gradient [3,19]. Now, combining (7) and (8)

we prove Eq. (6) as

$$\frac{\mathrm{d}H}{\mathrm{d}t} = \frac{\mathrm{d}\,Energy}{\mathrm{d}t} - T_0\,\frac{\mathrm{d}\,Entropy}{\mathrm{d}t} \leqslant 0. \qquad \square$$

We have shown from the definition of the thermodynamic Helmholtz free energy that a constant cybernetic temperature T_0 allowed simultaneous minimization of the internal energy for supervised categorization and maximization of the entropy for unsupervised component analysis, such as simultaneous de-noising and associative recall in the cocktail party effect.

A natural question arises: what are mathematical principles of learning without a teacher? We can furthermore derive unsupervised neural networks if we assume that the first-order estimation *Energy* term is defined as

$$E \equiv \boldsymbol{\lambda}^{\mathrm{T}}(x - \mathbf{As}) = \boldsymbol{\mu}^{\mathrm{T}}(\mathbf{Wx} - \mathbf{s}), \tag{9}$$

where $\boldsymbol{\lambda} \in \mathrm{R}^N$ and $\boldsymbol{\mu} \in \mathrm{R}^N$ represent vectors of the Lagrange multipliers and superscript T denotes transpose operation. When the external information enters the system through a set of smart sensors, hundreds of millions of excitations are generated by sensory neurons. In theory, these neurons could take energy to sustain themselves and, thereby, make sense of incoming data and reduce an unwanted redundancy among brain waves in order to make room for further stimuli. Did in situ neural pathways take the advantage of inevitable decay to learn something quickly without supervision, based on data in memory? We conjectured the affirmative: systematic decay by gradient descent and stochastic decay by "adrenal" annealing. In fact, in this paper, we combined both into one stochastic gradient descent learning methodology.

We begin to build up a truly unsupervised form of learning by first considering that, without macroscopic constraints defined by data vector **x**, the Shannon entropy alone would at the equilibrium point produce a trivial solution in terms of the unknown source vector **s**. This is stated by the following physics theorem.

Theorem 2. (Closed system equal partition equilibrium law). *If the total number of components is N, then the maximum entropy solution of a closed system is given by* $\bar{s}_j = 1/N$, *for all j, where Shannon–Boltzmann entropy* [22] *is defined with the Lagrange constraint* $K_{\mathrm{B}}(\lambda_0 + 1)$ *to incorporate the unit normalization of the total of unknown source components as in* (10):

$$S = -K_{\mathrm{B}}|\mathbf{s}| \sum_{j=1}^{N} s_j' \ln s_j' + K_{\mathrm{B}}|\mathbf{s}|(\lambda_0 + 1)\left(\sum_{j=1}^{N} s_j' - 1\right). \tag{10a}$$

In (10a) K_B *represents Boltzmann's constant and* $|\boldsymbol{s}|$ *represents* L_1 *norm of the source vector* **s**, *i.e.*

$$|\mathbf{s}| = \sum_{j=1}^{N} s_j \quad s_j' = s_j/|\mathbf{s}| \tag{10b}$$

or in equivalent vector notation

$$\mathbf{s} = |\mathbf{s}|\mathbf{s}'. \tag{10c}$$

In the subsequent derivations we shall assume that $|\mathbf{s}|$ and s_j' are independent variables that will be estimated separately using optimization procedure to be defined.

Proof of Theorem 2. From (10a) we obtain

$$\frac{\partial S}{\partial s_j'} = -K_B|\mathbf{s}|(\ln s_j' + 1) + K_B|\mathbf{s}|(\lambda_0 + 1) = 0 \tag{11}$$

from which it follows that

$$\bar{s}_j' = \exp(\lambda_0). \tag{12}$$

The unit sum constraint $\sum_{j=1}^{N} s_j' = 1$ implies

$$\exp(\lambda_0) = \frac{1}{N} \tag{13}$$

and from (12) and (13) we get

$$\bar{s}_j' = \frac{1}{N}. \qquad \square$$

We have generalized the Shannon entropy S of a closed system to an open system in dynamic equilibrium with the external energy E in terms of the Lagrange constraint. Then the following optimization problem is formulated, for the multispectral imaging data, for example, on the pure space-variant basis

$$[\mathbf{A}^*, |\mathbf{s}|^*, \mathbf{s}'^*] = \arg\min \mathrm{H}(|\mathbf{s}|, \mathbf{s}', \mathbf{A}), \tag{14a}$$

where Helmholtz free energy cost function is defined as follows:

$$\begin{aligned} H(|\mathbf{s}|, \mathbf{s}', \mathbf{A}) &= E(|\mathbf{s}|, \mathbf{s}', \mathbf{A}) - T_0 S(|\mathbf{s}|, \mathbf{s}') \\ &= |\mathbf{s}| \sum_{i=1}^{N} \lambda_i \left(x_i - \sum_{j=1}^{N} a_{ij} s_j' \right) + K_B T_0 |\mathbf{s}| \sum_{j=1}^{N} s_j' \ln s_j' \\ &\quad - K_B T_0 |\mathbf{s}| (\lambda_0 + 1) \left(\sum_{j=1}^{N} s_j' - 1 \right), \end{aligned} \tag{14b}$$

where in (14a) superscript * denotes optimal values in the sense of the minimum of the Helmholtz free energy $H(|\mathbf{s}|, \mathbf{s}', \mathbf{A})$. In (14b) x_i are components of the multispectral image at one particular pixel location, i.e., we have dropped the r term from (1)–(2) in order to simplify the notation. Nonlinear constraints [32] could be also used in (14b) as will be considered later. As already pointed out, in the cost function (14b) we treat the L_1 norm $|\mathbf{s}|$ and components of the scaled sources vector s_j' as independent (separate) variables.

Theorem 3. (The open system sigmoid partition law). *The analytical solution for the class vector* $\mathbf{s}'$ *obtained at the minimum of the Helmholtz free energy has the classic ANN sigmoid logic* [41]

$$s'_j = \frac{1}{1 + \sum_{\substack{k=1 \\ k \neq j}}^{N} \exp\left(\frac{1}{K_B T_0} \sum_{i=1}^{N} \lambda_i (a_{ik} - a_{ij})\right)} = \sigma_j(\mathbf{A}, \boldsymbol{\lambda}). \tag{15}$$

Proof. Equating the derivation of the Helmholtz free energy (14a)/(14b) w.r.t. $\mathbf{s}'$ with zero [41]

$$\frac{\partial H}{\partial s'_j} = -\sum_{i=1}^{N} \lambda_i a_{ij} |\mathbf{s}| + |\mathbf{s}| K_B T_0 \ln(s'_j) + |\mathbf{s}| K_B T_0 \lambda_0 = 0 \tag{16}$$

yields

$$s'_j = \exp\left(\frac{1}{K_B T_0}\left(\sum_{i=1}^{N} a_{ij}\lambda_i - \lambda_0\right)\right), \tag{17}$$

where the minimum is guaranteed by the positive second derivative

$$\frac{\partial^2 H}{\partial s'^2_j} = \frac{1}{s'_j} > 0$$

because, from (17), it follows that s'_j is non-negative. Likewise, the total class probability can be normalized. Thus, we summed (17) over an arbitrary number N of total classes to determine λ_0

$$\exp\left(\frac{\lambda_0}{K_B T_0}\right) = \sum_{k=1}^{N} \exp\left(\frac{1}{K_B T_0} \sum_{i=1}^{N} a_{ik}\lambda_i\right), \tag{18}$$

where λ_0 is the Helmholtz free energy and the right-hand side of (18) is the (canonical ensemble) partition function in statistical mechanics [22]. Inserting (18) into (17), we obtain

$$s'_j = \frac{\exp(\sum_{i=1}^{N} a_{ij}\lambda_i)}{\sum_{k=1}^{N} \exp(\sum_{i=1}^{N} a_{ik}\lambda_i)}. \tag{19}$$

Dividing both numerator and denominator of (19) by the numerator yields the ANN sigmoid function (15) without assuming it. □

We could justify the name of Lagrange Constraint ANN if we further derive the unsupervised Hebbian learning rule using only input data vector and neglecting any desirable output as follows.

Given the associative recall data $\mathbf{x} = \mathbf{A}\mathbf{s}$ and assuming the message $\mathbf{s}$ has been found for the time being, we estimated the associative memory (AM) matrix $\mathbf{A}$ from the given pair through the vector outer product formula [41]

$$\mathbf{A} = \mathbf{x}\frac{\hat{\mathbf{s}}^{\mathrm{T}}}{|\hat{\mathbf{s}}|^2}, \tag{20}$$

which obviously recalls the solution $\mathbf{As} = \mathbf{x}$. Because the mixing matrix in (20) was rank-1, what was still missing was the fault tolerance (FT) for the inverse association $\mathbf{A}^{-1}\mathbf{x} = \mathbf{s}$. It had been shown in [43] that, for pattern recall in a case of adaptive learning matrix $\mathbf{A}$, the classifier distance is modified by the Riemannian learning metric tensor $\mathbf{G} = \mathbf{A}^{\mathrm{T}}\mathbf{A}$. It follows from the fact that, while the Euclidean distance measure d_{x} of data $\mathbf{x}$ had the identity metric $\mathbf{I}$, the recall message $\mathbf{s}$ has the equivalent distance $d_{\mathbf{s}}$ in the $\mathbf{s}$-space with the Riemannian metric $\mathbf{G}$

$$d_{\mathrm{x}} \equiv (\mathbf{x}, \mathbf{Ix}) = \mathbf{x}^{\mathrm{T}}\mathbf{x} = \mathbf{s}^{\mathrm{T}}\mathbf{A}^{\mathrm{T}}\mathbf{As} = (\mathbf{s}, \mathbf{A}^{\mathrm{T}}\mathbf{As}) \equiv (\mathbf{s}, \mathbf{Gs}) \equiv d_{\mathrm{s}}. \tag{21}$$

The implication of Eq. (21) is important for machine ATR [15], i.e., given the identical data set, a machine could be trained to classify better by altering the metric distance measure $d_{\mathrm{s}} = (\mathbf{s}, \mathbf{Gs})$. It might be quite profound that we have explicitly proven Amari's assertion of the brain learning geometry utilizing the Riemannian metric for pattern recognition [2,3]. In [43] we have applied Riemannian metric tensor $\mathbf{G} = \mathbf{A}^{\mathrm{T}}\mathbf{A}$ to embed the flat space Euclidean learning within the Riemannian learning hyper-sphere to produce the full rank AM Hebbian learning algorithm.

The full rank AM learning rule is computed based on the natural gradient of the minimum estimation error averaged over the source. Minimizing Helmholtz free energy H in (14a)/(14b) w.r.t. $\mathbf{A}$, we obtain

$$\frac{\partial \mathbf{A}}{\partial t} = -\left\langle \frac{\partial H(\mathbf{s}, \mathbf{A})}{\partial \mathbf{A}} \mathbf{A}^{\mathrm{T}}\mathbf{A} \right\rangle_{\text{sources}}, \tag{22}$$

where the Riemannian metric tensor for the sources' distance was derived from the Euclidean inner product distance of sources. The unsupervised learnt matrix $\mathbf{A}$ has to be full rank for the inverse of the associative learning to exist. From both (22) and (14a), we obtain

$$\begin{aligned} \mathbf{A}(k+1) &= \mathbf{A}(k) + \eta \partial \mathbf{A}/\partial t = \mathbf{A}(k) - \eta \langle \partial H(\mathbf{s}, \mathbf{A})/\partial \mathbf{A} \rangle_{\mathrm{s}} \mathbf{A}^{\mathrm{T}}\mathbf{A} \\ &= \mathbf{A}(k) + \eta \boldsymbol{\lambda} \mathbf{s}^{\mathrm{T}} \mathbf{A}(k)^{\mathrm{T}} \mathbf{A}(k) = (\mathbf{I} + \eta \boldsymbol{\lambda} \mathbf{x}^{\mathrm{T}}) \mathbf{A}(k), \end{aligned} \tag{23}$$

where in (23) η represents small learning gain. Here, we have applied the natural gradient learning rule [3] at the single pixel level [43]. The termination of iterations without a teacher occurs after the input data $\mathbf{x}$ becomes orthogonal to the Lagrange multipliers vector $\boldsymbol{\lambda}$: this is similar to Oja's hyper-spherical rotation. The subscript index $\mathbf{s}$ in (23) means that iterations are done over the source vector $\mathbf{s}$: that explains why we did not need to use the neighborhood data averaging. Update Eq. (23) gives the full rank mixing matrix $\mathbf{A}$. In (18), (19) and (23) $\boldsymbol{\lambda}$ is the vector of the Lagrange multipliers whose update $\Delta\boldsymbol{\lambda}$, according to [43], is obtained as a solution of the variation equation

$$\Delta x_i = \sum_{k=1}^{N} \frac{\partial \tilde{x}_i}{\partial \lambda_k} \Delta \lambda_k,$$

where $\tilde{x}_i = \sum_{j=1}^{N} a_{ij} s_j$ represents approximation of the data component x_i at some iteration l and $\Delta x_i = x_i - \tilde{x}_i$ represents approximation error. Solution of the above

variation equation is given with

$$\Delta\boldsymbol{\lambda} = (\tilde{\mathbf{x}}(l)\tilde{\mathbf{x}}^{\mathrm{T}}(l) - \mathbf{A}(l)\mathrm{diag}\{s_j(l)\}_{j=1}^{N}\mathbf{A}^{\mathrm{T}}(l))^{-1}(\mathbf{x} - \mathbf{A}(l)\mathbf{s}(l)). \tag{24}$$

We want to point out here that while the mixing matrix gradient descent learning rule (23) together with the Lagrange multipliers learning rule (24) and source probabilities equation (17) enables solution of the linear inverse problem (1) there is no guarantee that global minimum of the cost function (14a)/(14b) will be reached. That motivated our efforts to formulate stochastic gradient descent learning as a combination of the gradient descent and stochastic Cauchy annealing.

3. Fast Cauchy annealing

The constraint optimization could be generalized for both supervised and unsupervised ANN depending on the construction of the energy landscape. Our empirical data indicated that the supervised ANN was associated with a gentle, Piedmont-like landscape with multiple lakes of different depths; in contrast, the unsupervised ANN, having unlabelled training data, was associated with much narrower attractor basins, including the singularity-like golf-course landscape. This type of singular landscape was obtained when deterministic blind inversion of the space-variant imaging problem was solved by minimizing the Helmholtz free energy [27,28,44]. When the cost function $C(\mathbf{x})$ had a single minimum, a gradient descent gave a unique ground state, and any reasonable method could approach it. However, when the cost function had multiple extremes, more powerful techniques were required for escaping from local extremes. Simulated annealing was just such a stochastic strategy for searching the global ground state. Geman and Geman had proved the cooling schedule of simulated annealing theorem [17], using the Metropolis annealing algorithm to generate random walks with reduced variances. As a result, the algorithm reached the global minimum at an exceedingly slow inversely logarithmic cooling schedule [17]. Fast Cauchy simulated annealing (CSA) combining Gaussian random walks with occasional Levy random flights could achieve the global optimization with a much faster cooling schedule: inversely linear in the state-transition time steps [39,40]. The cooling schedule of the Cauchy annealing algorithm was proved to be inversely linear in time—which was fast compared to Geman–Geman Gaussian simulated annealing (GSA), which is strictly a local search requiring that the cooling schedule be inversely proportional to the logarithmic function of time [17].

We recapitulated the proof of the convergence of the cooling schedules for both fast CSA [40] and GSA [1,17].

Theorem 4. (Fast Cauchy annealing). *The Cauchy annealing with the cooling schedule* $T_{\mathrm{c}}(t) = T_0/t$ *for some* $T_0 > 0$ *could visit the local neighborhood with the probability* g_t *at any time t. Similarly, the Gaussian annealing with the cooling schedule* $T_{\mathrm{a}}(t) = T_0/\log t$ *for some* $T_0 > 0$ *could visit the local neighborhood with the probability* g_t *at any time t.*

Proof. Let the state-generating probability at the cooling temperature $T_c(t)$ at the time t and within a neighborhood be (bounded below) $\geqslant g_t$. Then the probability of not generating a state in the neighborhood is obviously (bounded above by) $\leqslant(1-g_t)$. To prove that a specific cooling schedule maintains the state-generation infinite often in time it is easier to prove the negation of the converse, i.e., the impossibility of never generating a state in the neighborhood after an arbitrary time t_0. In other words, the negation probability vanishes

$$\prod_{t=t_0}^{\infty}(1-g_t)=0. \tag{25}$$

Taking the logarithm of (25) and expanding in Taylor series (noting that $\log 0=-\infty, \quad \log(1-g_t)\approx -g_t$), shows that proving (25) is equivalent to proving (26)

$$\sum_{t=t_0}^{\infty} g_t=\infty. \tag{26}$$

We can now verify cooling schedules satisfying (26) in the D-dimensional neighborhood for an arbitrary size $|\Delta\mathbf{x}|=|\mathbf{x}-\mathbf{x_0}|$ and t_0, where $\mathbf{x_0}$ is the previously chosen point to test. For the bounded variance-type GSA, there exists an initial T_0. For $t>0$,

$$T_a(t)=T_0/\log t, \tag{27}$$

$$g_t(\mathbf{x})=(2\pi T_a(t))^{-D/2}\exp[-|\Delta\mathbf{x}|^2/(2T_a(t))], \tag{28}$$

$$\sum_{t=t_0}^{\infty} g_t\geqslant\sum_{t=t_0}^{\infty}\exp(-\ln t)=\sum_{t=t_0}^{\infty}1/t=\infty. \tag{29}$$

For the unbounded variance-type Cauchy annealing for arbitrary $T_0>0$,

$$T_c(t)=T_0/t, \tag{30}$$

$$g_t(\mathbf{x})=\frac{T_c(t)}{[T_c^2(t)+|\Delta\mathbf{x}|^2]^{(D+1)/2}}\approx\frac{T_0}{t|\Delta\mathbf{x}|^{D+1}}, \tag{31}$$

$$\sum_{t=t_0}^{\infty} g_t\approx\frac{T_0}{|\Delta\mathbf{x}|^{D+1}}\sum_{t=t_0}^{\infty}\frac{1}{t}=\infty. \tag{32}$$

So, the local neighborhood is visited infinite number of times at each time t, and the cooling schedule algorithm is admissible. □

To apply CSA to D-dimensional problems, we had to introduce the transformation from Cartesian to hyper-spherical coordinates in order to simplify generation of D-dimensional Cauchy pdf (in Cartesian coordinates) as a product of D one-dimensional pdf's (in the hyper-spherical coordinates). The full derivation of the annealing algorithm based on this transformation is given in Appendix A. The CSA

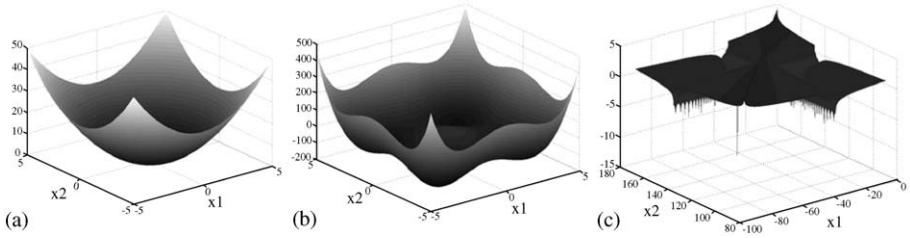

Fig. 2. From left to right are: (a) 2*D* objective functions with the 'ocean'-like type of landscape; (b) multiple 'lakes'-like type of landscape; (c) 'golf hole'-like type of landscape.

algorithm could then be applied to search both the supervised learning multiple lake-like landscape and the unsupervised learning golf-course-like landscape. Fig. 2 illustrates three possible landscapes (a) left: ocean landscape, (b) center: multiple minima in a lake landscape, and (c) right: golf course landscape. Convergence of fast Cauchy annealing had been demonstrated in [39] for a 1-*D* double-well potential function, the 2-*D* equivalent of which is shown in Fig. 2b. The Helmholtz free energy-based objective function used in blind space-variant imaging problems [27,28,44] had a very difficult golf-course landscape: note the two spikes in Fig. 2c. Here, we demonstrated the application of the higher-dimensional fast Cauchy annealing on the global minimization of a golf-course landscape using techniques such as blind de-mixing of a space-variant mixture of images. Other applications for such de-mixing include telescope images in astronomy or remotely sensed images where pixel values are represented as positive intensities [41,42].

In Fig. 3, we illustrated the fundamental difference between Gaussian annealing [1] and Cauchy annealing [39,40]. The top portion of Fig. 3 shows free space random walks for an independent variable with a Gaussian distribution where variance of the distribution was equivalent to the temperature, the change of which is inversely proportional with logarithm of time, i.e. $T_a(t)/T_0 = 1/\log(1+t)$. The bottom portion of Fig. 3 shows the free space random walk for the same independent variable generated with a Cauchy or Lorentz distribution. Here, the temperature equivalent parameter is inversely proportional to time:

$$c(t) = T_c(t)/T_0 = 1/(1+t) \tag{33}$$

A faster cooling rate combined with occasional long jumps is what enables the Cauchy annealing to converge faster than Gaussian annealing. To apply Cauchy annealing theory on the *D*-dimensional non-convex optimization problems, we needed to generate the *D*-dimensional Cauchy distribution given by [39]

$$p(\mathbf{x}) = \frac{c}{[c^3 + |\mathbf{x}|^2]^{(D+1)/2}} \tag{34}$$

which we showed was easier to generate if the parameter vector $\mathbf{x}$ was transformed from the Cartesian to hyper-spherical coordinates. We give a comprehensive mathematical treatment of this transformation in Appendix A. Here, we pointed out that *D*-dimensional distribution $p(\mathbf{x})$ can be generated as the product of *D*

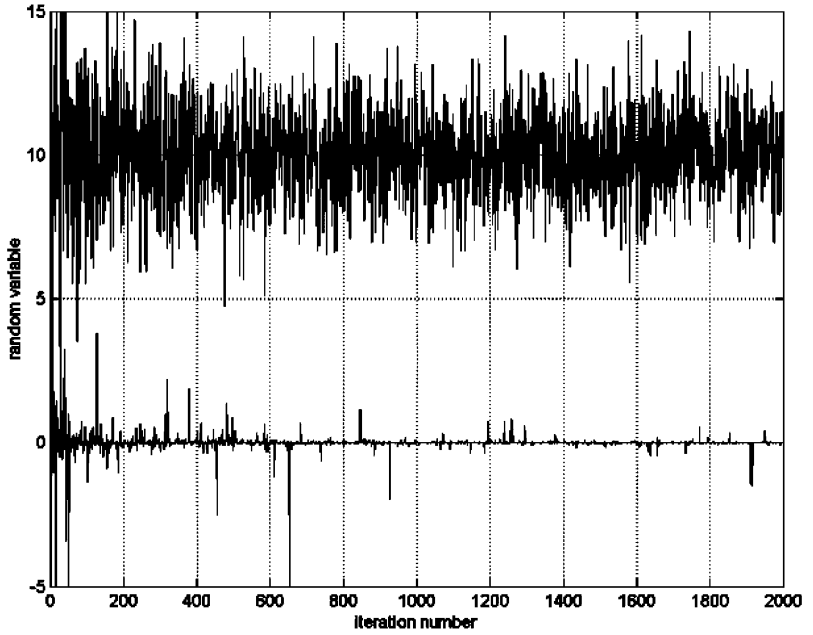

Fig. 3. Comparison of the free-space random walk for Gaussian distribution with cooling rate inversely proportional with the logarithm of time (top) and for Cauchy distribution with the cooling rate inversely proportional with time (down).

one-dimensional distributions $p(\theta_1)p(\theta_2)....p(\theta_{D-1})p(r)$, where $\theta_i\ i = 1, 2, \ldots, D-1$ are angles in the hyper-spherical coordinate system, and r is the magnitude of the D-dimensional parameter vector $\mathbf{x}$. The one-dimensional distributions are derived from the identity

$$\int\int\ldots\int p(x_1, x_2, \ldots, x_D)\,\mathrm{d}x_1\,\mathrm{d}x_2\ldots\mathrm{d}x_D = \int\int\ldots\int p(\theta_1)p(\theta_2)\ldots p(\theta_{D-1})p(r)\,\mathrm{d}\theta_1\,\mathrm{d}\theta_2\ldots\mathrm{d}\theta_{D-1}\,\mathrm{d}r = 1. \tag{35}$$

We illustrated performance of two-dimensional Cauchy annealing on an example: three objective functions with the three different landscapes shown in Fig. 2. The objective function for Fig. 2a is $C_1(x_1, x_2) = x_1^2 + x_2^2$. It has a single minimum that can be reached by standard gradient descent methods. Fig. 2b has this objective function

$$C_2(x_1, x_2) = x_1^4 - 16x_1^2 + 5x_1 + x_2^4 - 16x_2^2 + 5x_2. \tag{36}$$

For (36), the interval $x_1, x_2 \in [-5, 5]$ has four minimums and one maximum. The global minimum of the objective function $C_2(x_1, x_2)$ is located at the point $x_1 = x_2 = -2.9$ when $C_2(-2.9; -2.9) = -156.66$. The third objective function, shown in

Fig. 2c is of the form

$$C_3(x_1,x_2)=\left(\frac{\cos x_1}{\sin(x_2-x_1)}\,4.3132+\frac{\sin x_1}{\sin(x_2-x_1)}\,6.6153-5\right)^2 +\left(\frac{\cos x_2}{\sin(x_2-x_1)}\,4.3132+\frac{\sin x_2}{\sin(x_2-x_1)}\,6.6153-3\right)^2, \tag{37}$$

where $x_1\in[\frac{\pi}{2},\pi]$ and $x_2\in[-\frac{\pi}{2},0]$. The objective function (37) is minimized by solving the two-dimensional linear space-variant inverse imaging problem [44], where independent variables x_1 and x_2 are angles of the parameterized de-mixing matrix. From the optimization point of view, it was interesting that the global minimum of Fig. 2c's objective function, $C_3(x_1,x_2)$, which had the so-called singularity golf course landscape, again was practically impossible to find by means of gradient descent algorithms. In Section 4, we demonstrated performance of the Cauchy annealing algorithm applied to the five-dimensional objective function with the same type landscape as $C_3(x_1,x_2)$. For the 2-*D* objective function given in (36), we computed gradients as

$$\frac{\partial C(x_1,x_2)}{\partial x_1}=4x_1^3-32x_1+5,$$

$$\frac{\partial C(x_1,x_2)}{\partial x_2}=4x_2^3-32x_2+5. \tag{38}$$

Now, the stochastic gradient learning algorithm for x_1 and x_2 was given by

$$\frac{\partial x_i}{\partial t}=F_i(t)=\overline{F}_i(t)+\tilde{F}_i(t)=-\frac{\partial C}{\partial x_i}+\tilde{F}_i(t), \tag{39}$$

where according to Uhlenbeck and Lawson [29], we have decomposed a macroscopic dynamic variable $F_i(t)$ into two terms: (i) a systematic behavior denoted by a time-averaging superscript bar or an ensemble expectation bracket in a slow time scale—$\bar{F}_i(t)$, and (ii) the fluctuation difference denoted by the superscript tilde in a fast time scale—$\tilde{F}_i(t)$. Evaluating both sides of (39) at transition time step $k+1$, we obtained

$$x_i(k+1)=\bar{x}_i(k)-\eta\,\frac{\partial C(x_1(k+1),x_2(k+1))}{\partial x_i}+\gamma_i\tilde{x}_i(k+1);\quad \gamma_i\equiv\partial\tilde{F}_i/\partial\tilde{x}_i, \tag{40}$$

where $\bar{x}_i$ represented the last value accepted by following either criterion in (41)

$$C(x_1(k+1),x_2(k+1))<C(\bar{x}_1,\bar{x}_2) \tag{41}$$

or the Metropolis criterion [30], in (42)

$$p_k\leqslant\frac{1}{(1+\mathrm{e}^{\Delta C_k/T})}, \tag{42}$$

where p_k is uniformly generated probability and $\Delta C_k=C(x_1(k+1),x_2(k+1))-C(\bar{x}_1,\bar{x}_2)>0$ is the increase in the error energy at iteration k. The Metropolis criterion (42) was derived from the normalization of two transition states in the

canonical probability $\exp(-E/K_B T) = \exp(-C/T)$, which accepts the new state even if the current 'guess' is not in the decreasing energy direction of the objective function. Either of the Cauchy or Gaussian pdf can generate the perturbation $\tilde{x}_i$ in (40). The Cauchy pdf exceeded the Gaussian in the sense that it occasionally generated random Levy flights or long jumps and local random walks otherwise. This was helpful if the last accepted solution $\bar{\mathbf{x}}$ was already in the proximity of the global minimum, because the gradient part $\partial C(x_1, x_2)/\partial x_i$ of the learning rule (40) would continue toward global minimum's direction even when contribution from the thermal noise part $\tilde{x}_i$ was negligible. Once algorithm (40) approached the proximity of global minimum, contributions from the gradient part $\partial C(x_1, x_2)/\partial x_i$ would approach zero. If the system was cooled enough, the noise part $\tilde{x}_i$ of the learning rule (40) would not be able to move the solution away from stable global minimum because both criteria (41) and (42) could not be satisfied anymore. If, however, the algorithm sits in a local minimum, the gradient part of the learning rule $\partial C(x_1, x_2)/\partial x_i$ will approach zero. But contribution from the thermal noise $\tilde{x}_i$ would be able to move the system from the local minimum because a new "guess" could satisfy either criterion (41), or if the system was not completely cooled, criterion (42). Fig. 4 showed convergence, in terms of the mean-square error, of the stochastic gradient algorithm with the Cauchy annealing (solid line) and Gaussian

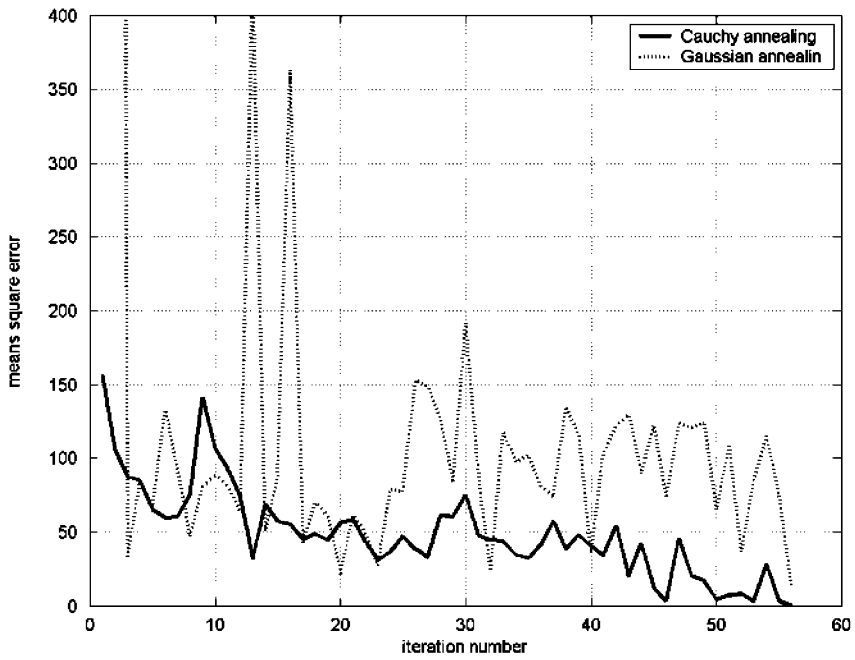

Fig. 4. Mean-square error of the multiple well cost function $C_2(x_1, x_2)$ for stochastic gradient algorithms with the Cauchy annealing (solid line) and Gaussian annealing (dashed line).

annealing (dashed line) applied to the stochastic optimization of the multiple-well objective function (40), where two-dimensional ground state was sampled in polar coordinates using two one-dimensional distributions: $p(\theta)$ and $p(r)$. Fig. 5 shows convergence of the two-dimensional stochastic gradient algorithms with Cauchy annealing (dashed line with '$*$' marks) and standard deterministic gradient descent ('$+$' marks) algorithm in the plane spanned by the independent variables x_1 and x_2. Both algorithms started from the initial point (2.3,3.4), ('$\square$' mark) that was closed to local minimum (2.9;2.9), ('∇' mark). Global minimum was located at (-2.9,-2.9), ('O' mark). As seen in Fig. 5, standard gradient descent becomes trapped in the local minimum while the stochastic gradient with Cauchy annealing manages to converge toward the global minimum. Fig. 6 shows the cumulative value of the average number of iterations necessary to reach the global minimum under given criterion

$$\frac{|C(x_1, x_2) - C(x_1^*, x_2^*)|}{C(x_1^*, x_2^*)} \leqslant 0.01. \tag{43}$$

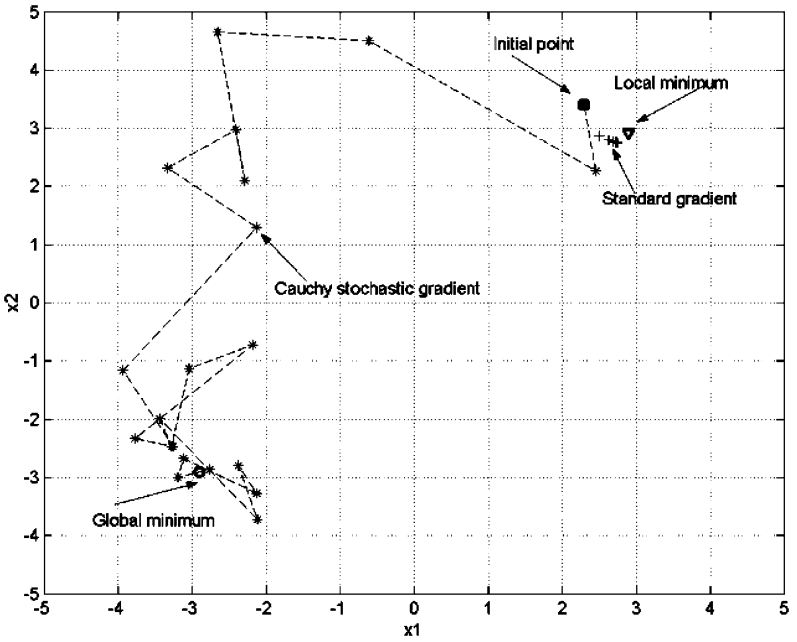

Fig. 5. Convergence of the two-dimensional stochastic gradient algorithms with Cauchy annealing (dashed line with '$*$' marks) and standard deterministic gradient descent ('$+$' marks) algorithm in the plane spanned by independent variables x_1 and x_2. Both algorithms started from the initial point (2.3,3.4), ('$\square$' mark) that was closed to local minimum (2.9,2.9), ('∇' mark). Global minimum was located at (-2.9,-2.9), ('O' mark). As could be seen, standard gradient descent was trapped in the local minimum while stochastic gradient with Cauchy annealing managed to converge toward global minimum.

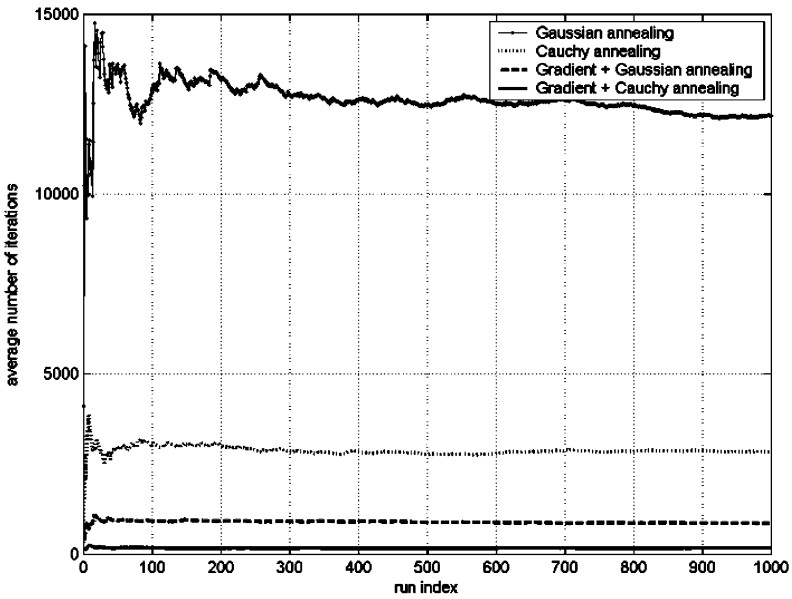

Fig. 6. Cumulative value of the average number of iterations necessary to reach the global minima under given criteria $|C_2(x_1, x_2) - C_2(x_1^*, x_2^*)|/C_2(x_1^*, x_2^*) \leqslant 0.01$. Average number of iterations was evaluated as a function of the run index over 1000 runs. From top to down are: (a) Gaussian annealing algorithm (dot-dashed line); (b) Cauchy annealing algorithm (dotted line); (c) Stochastic gradient algorithm with Gaussian annealing (dashed line); (d) Stochastic gradient algorithm with Cauchy annealing (solid line). In average 12,175 iterations were necessary for Gaussian annealing algorithm, 2833 iteration for Cauchy annealing algorithm, 844 iterations for stochastic gradient algorithm with Gaussian annealing and only 155 iterations for stochastic gradient with Cauchy annealing. For comparison 10,000 iterations are necessary for exhaustive search in order to find the independent variables with precision on the first decimal place. This gives speed-up factor of ≈ 64.5 for the stochastic gradient algorithm with Cauchy annealing.

The average number of iterations was evaluated as a function of the run index over 1000 runs according to

$$\bar{M}(k) = \frac{1}{k}[M(k)] = \frac{1}{k}\sum_{m=1}^{k} M(m), \tag{44}$$

where $M(k)$ represents cumulative number of iterations necessary to reach the global minimum when algorithm is run k times. In our performance evaluation experiment k was run from 1 to 1000. From top-down, Fig. 6 has four parts: (a) the Gaussian annealing algorithm (dot-dashed line), (b) the Cauchy annealing algorithm (dotted line), (c) the stochastic gradient algorithm with Gaussian annealing (dashed line), and (d) the stochastic gradient algorithm with Cauchy annealing (solid line). On the average, 12,175 iterations were necessary for Gaussian annealing algorithm, 2833 iteration for Cauchy annealing algorithm, 844 iterations for stochastic gradient

algorithm with Gaussian annealing, and only 155 iterations for stochastic gradient with Cauchy annealing. For comparison, 10,000 iterations are necessary for exhaustive search in order to find the independent variables x_1 and x_2 with precision to the first decimal place. This demonstrates an increase in speed of ≈ 64.5 for the stochastic gradient algorithm with Cauchy annealing.

4. Nonlinear blind space-variant imaging

In this section, we applied the Cauchy annealing algorithm described in Section 3 to the minimization of the Helmholtz free energy cost function in blind space-variant linear and nonlinear imaging problems [27,44]. The cost function for one particular set of parameters was given by (37) and, as illustrated in Fig. 2c, had a singularity landscape that makes the gradient term vanish in the vicinity of the global minimum. Therefore, only stochastic search based on the Cauchy pdf has been used to find the global minimum of the cost function (37); that for more general nonlinear case is given in (49). Based on [27], we formulated the nonlinear space-variant imaging problem as

$$\mathbf{x}(r) = g(\mathbf{A}(r)\mathbf{s}(r)), \tag{45}$$

where r in (45) means that both unknown mixing matrix $\mathbf{A}$ and unknown source vector $\mathbf{s}$ are pixel dependent, i.e., space-variant. Unlike algorithms described in [35,46], that can solve the post-nonlinear (PNL) BSS problem with unknown nonlinearity, the algorithm described here is not capable to do so. Like in [8] we did assume that functional form of the sensor nonlinearity $g()$ was known depending on the unknown parameters. The main reason why the algorithm described here is not able to estimate inverse of the unknown nonlinearity is the space variant nature of model (45). Statistical approaches [35,46], used to estimate the inverse nonlinearity, cannot be used in the space-variant mixture. In the simulation example shown in Figs. 11 and 12 we had used a typical saturation type of infrared sensor nonlinearity

$$x_i^n = g(x_i) = 2^B(1 - \mathrm{e}^{-\alpha_i x_i}), \tag{46}$$

where $B = 8$ represented number of bits and α_i was unknown slope parameter to be determined with the Cauchy annealing stochastic optimization algorithm. The value of the slope parameter used in the two-dimensional simulation experiments reported in Figs. 11 and 12 was $\alpha_i = 0.01$ for both sensors. In nonlinear blind de-mixing process, the linearized part of the data vector (46) had been modeled as

$$\begin{bmatrix} \tilde{x}_1 \\ \tilde{x}_2 \end{bmatrix} = |\mathbf{s}| \begin{bmatrix} \cos\theta & \cos\varphi \\ \sin\theta & \sin\varphi \end{bmatrix} \begin{bmatrix} s_1' \\ s_2' \end{bmatrix}, \tag{47}$$

where $\tilde{x}_i = g^{-1}(x_i)$. Although this representation might seem to be quite specific it describes well the conservative propagation medium and is physically relevant. It also incorporates positivity constraints necessary to deal with the imaging problems. The model can be generalized to higher dimensions [28,44], in which case the $N-1$

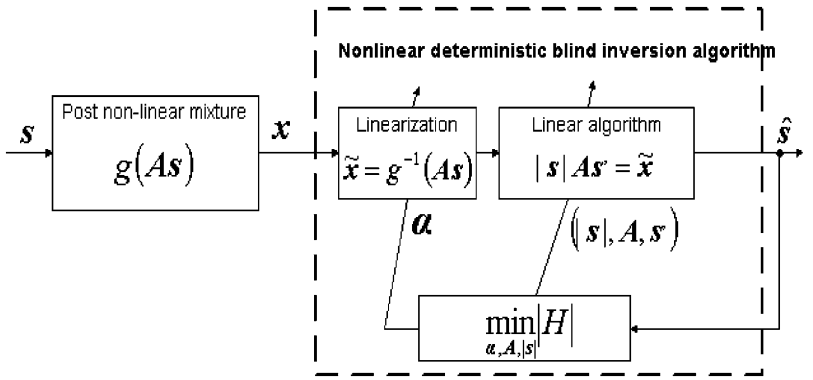

Fig. 7. Illustration of the blind inversion of the nonlinear space-variant imaging problem.

angle per column vector of the mixing matrix is necessary to describe a position of the unit norm column vector in the N-dimensional space. Fig. 7 shows that the blind inversion algorithm for the nonlinear space-variant imaging problem consisted of two steps: (i) linearization of the nonlinear data vector $\mathbf{x}$ and (ii) blind inversion of the linearized space-variant problem. However, both linearization and solution of the linear BSS problem are obtained simultaneously. The nonlinear blind inversion approach illustrated in Fig. 7 is in principle equivalent to the post-nonlinear mixture model studied by Taleb and Jutten [46]. In implementing the Taleb–Jutten algorithm, demonstrated in Figs. 11 and 12, we took advantage of the fact that type of the nonlinearity was known and given in (46). That way, estimation of the inverse nonlinearity was avoided, and the Taleb–Jutten algorithm was in a fair position in comparison with our algorithm. As already pointed out for the algorithm presented here the type of the nonlinearity must be known with the unknown parameters. For the simulation experiments reported in Figs. 9–12, space-(in)variant mixtures were generated by using data model (47): their mixing matrices were parameterized in terms of two mixing angles θ and φ. For the space-invariant cases, mixing angles were constant for all the pixels with values $\theta = 64^\circ$ and $\varphi = 45^\circ$. For the space-variant cases, mixing angles were changed row-wise as illustrated in Fig. 8. The image size used in the simulation experiments was 72×88 pixels. The φ angle was changing row-wise 1° per row, ranging from 1° to 72°. The θ angle was always 4° greater than the φ angle. Based on (46), an inverse nonlinearity has been obtained

$$g^{-1}(x_i) = \frac{1}{\alpha_i} \ln \frac{1}{1 - (x_i/2^B)}. \tag{48}$$

To solve the two-dimensional nonlinear space-variant imaging problem, assuming that each sensor's nonlinearity could be modeled by using one parameter only, it was necessary to minimize the five-dimensional objective function with the golf course

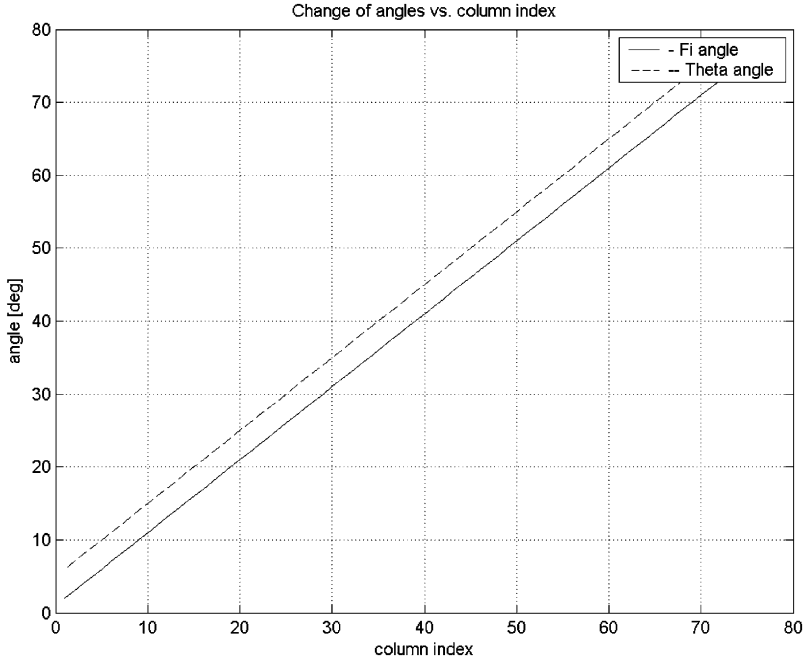

Fig. 8. A columnwise change of the mixing angles φ and θ in accordance with the linear mixing model (49). Solid line—φ angle; dashed line—θ angle.

type of landscape such as (37)

$$(\boldsymbol{\alpha}^*, \mathbf{A}^*, |\mathbf{s}|^*) = \arg\min |H|^2 \cong \arg\min |E|^2 \tag{49a}$$

with

$$|E|^2 = (g^{-1}(\mathbf{x}) - \mathbf{A}|\mathbf{s}|\mathbf{s}')^{\mathrm{T}}(g^{-1}(\mathbf{x}) - \mathbf{A}|\mathbf{s}|\mathbf{s}'), \tag{49b}$$

where $\boldsymbol{\alpha}$ was the vector of the unknown nonlinearity parameters; $g^{-1}(\circ)$ was the inverse nonlinearity (48) such that $\tilde{\mathbf{x}} = g^{-1}(\mathbf{x})$; $\mathbf{x}$ was the given data vector; $\mathbf{A}$ was the unknown mixing matrix parameterized in terms of the two angles θ and φ; and $|\mathbf{s}|$and $\mathbf{s}'$ were, respectively, the magnitude and scaled components of the unknown source vector $\mathbf{s} = |\mathbf{s}|\mathbf{s}'$. Because cost function (49b) is non-negative, it is bounded from below. In (49b) $\mathbf{s}'$ was found at the equilibrium of the Helmholtz free energy, defined as

$$H(\boldsymbol{\alpha}, \mathbf{A}, |\mathbf{s}|, \mathbf{s}') = E - T_0 S = \boldsymbol{\lambda}^{\mathrm{T}}[g^{-1}(\mathbf{x}) - \mathbf{A}|\mathbf{s}|\mathbf{s}'] + |\mathbf{s}|K_{\mathrm{B}}T_0 \sum_{i=1}^{N} s_i' \ln s_i'$$
$$- |\mathbf{s}|K_{\mathrm{B}}T_0(\lambda_0 + 1)\left(\sum_{i=1}^{N} s_i' - 1\right). \tag{50}$$

We want to comment here that formulation of the Helmholtz free energy given by (14b) is a special case, obtained just for the linear data model, of the more general formulation (50). The reason that minimizing the objective function (49b) was equivalent to minimizing the Helmholtz free energy defined by (50) was that Shannon entropy S used in (50) was invariant w.r.t. triplet $(\boldsymbol{\alpha}, \mathbf{A}, |\mathbf{s}|)$. Here, the vector of nonlinearity parameters $\boldsymbol{\alpha}$ and unknown de-mixing matrix were found by using Cauchy annealing. This is how the algorithm works. At some iteration, l, triplet $(\boldsymbol{\alpha}^{(l)}, \mathbf{A}^{(l)}, |\mathbf{s}|^{(l)})$ is generated as an output of Cauchy annealing in an attempt to reach possibly global minimum of the estimation error energy (49b), i.e.

$$(\boldsymbol{\alpha}^{(l)}, \mathbf{A}^{(l)}, |\mathbf{s}|^{(l)}) = \arg\min |E|.$$

For a given triplet $(\boldsymbol{\alpha}^{(l)}, \mathbf{A}^{(l)}, |\mathbf{s}|^{(l)})$, the MaxEnt-like algorithm in [44] computed the most probable solution for the vector of source probabilities, $\mathbf{s}'^{(l)}$ based on the (17)

$$s_j'^{(l,k)} = \exp\left(\frac{1}{K_B T_0}\left(\sum_{i=1}^{N} a_{ij}^{(l)} \lambda_i^{(k)} - \lambda_0\right)\right)$$

and Lagrange multipliers learning equation derived in a similar way as (24)

$$\begin{aligned}\Delta\lambda_j &= \sum_{i=1}^{N} \frac{\partial\lambda_j}{\partial s_i'} \Delta s_i' \\ &= \sum_{i=1}^{N}\left(\sum_{m=1}^{N} w_{mj}\left(\frac{K_B T_0}{s_i'^{(k)}}\,\delta_{im} + \sum_{n=1}^{N} \lambda_n^{(k)} a_{ni}^{(l)}\right)\right)\left(\sum_{n=1}^{N} w_{in}^{(l)} \frac{x_n}{|\mathbf{s}|} - s_i'^{(k)}\right),\end{aligned}$$

where k represents another iteration index associated with learning of the scaled source vector components s_j' for a given triplet $(\boldsymbol{\alpha}^{(l)}, \mathbf{A}^{(l)}, |\mathbf{s}|^{(l)})$ and w_{mj} and w_{in} are components of the de-mixing matrix $\mathbf{W}$. After each iteration l is completed, we get a quadruple, $\left(\boldsymbol{\alpha}^{(l)}, \mathbf{A}^{(l)}, |\mathbf{s}|^{(l)}, \mathbf{s}'^{(l)}\right)$. The algorithm accepts as a final solution the quadruple $(\boldsymbol{\alpha}^*, \mathbf{A}^*, |\mathbf{s}|^*, \mathbf{s}'^*)$ for which the estimation error energy (49b) reaches a possibly global minimum. Because the nonlinear function (46) has unique inverse, the quadruple $(\boldsymbol{\alpha}^*, \mathbf{A}^*, |\mathbf{s}|^*, \mathbf{s}'^*)$, corresponding with given data model (45)–(47), gives a global minimum of the error energy function (49b). The described algorithm is summarized in the pseudocode given in Table 1.

Fig. 9 shows results of blind de-mixing for the linear space-invariant imaging problem. From left to right: (a) source images; (b) space-invariant, noise-free linear mixtures; (c) recovery of the source images using Helmholtz free energy and linear version of the Helmholtz free energy blind inversion algorithm in (43)–(46) with the Cauchy annealing-based algorithm described in Section 3, and (d) recovery of the source images using Infomax ICA algorithm [6]. In Fig. 9d, the source images were recovered using the maximum likelihood/Infomax ICA algorithm [6,33] in combination with Amari's natural [3] or Cardoso's relative [10] gradient. Thus, an adaptive ICA algorithm for estimation of the de-mixing matrix is obtained

$$\mathbf{W}(k+1) = \mathbf{W}(k) + \eta\left(\mathbf{I} - \varphi(\mathbf{y})\mathbf{y}(k)^{\mathrm{T}}\right)\mathbf{W}(k), \tag{51}$$

Table 1
A pseudocode of the Helmholtz free energy based algorithm for solving space-variant post-nonlinear BSS problem

1. *START with some* $\boldsymbol{\alpha}^{(0)}$, $\mathbf{A}^{(0)}$, $|\mathbf{s}|^{(0)}$, $l=0$;
2. $s_j^{(l,0)}=1/N$; $\lambda_j^{(0)}=0$, $k=0$;
   DO
   $$\Delta\lambda_j=\sum_{i=1}^{N}\left(\sum_{m=1}^{N} w_{mj}\left(\frac{K_B T_0}{s_i^{(l,k)}}\delta_{im}+\sum_{n=1}^{N}\lambda_n^{(k)}a_{ni}^{(l)}\right)\right)\left(\sum_{n=1}^{N} w_{in}^{(l)}\frac{x_n}{|\mathbf{s}|^{(l)}}-s_i^{(l,k)}\right)$$
   $$\lambda_j^{(k+1)}=\lambda_j^{(k)}+\Delta\lambda_j$$
   $$s_j^{(l,k+1)}=\frac{1}{1+\sum_{\substack{m=1\\ m\neq j}}^{N}\exp\left(\frac{1}{K_B T_0}\sum_{i=1}^{N}\lambda_i(a_{im}-a_{ij})\right)}$$
   $k=k+1$;
   UNTIL$(|s^{(l,k)}-s^{(l,k-1)}|<\varepsilon)$
3. $l=l+1$;
   $$(\boldsymbol{\alpha}^{(l)},\mathbf{A}^{(l)},|\mathbf{s}|^{(l)})=\arg\min|E|^2$$
   $$=\arg\min(g^{-1(l)}(\mathbf{x})-\mathbf{A}^{(l)}|\mathbf{s}|^{(l)}\mathbf{s}^{\cdot(l)})^{\mathrm{T}}(g^{-1(l)}(\mathbf{x})-\mathbf{A}^{(l)}|\mathbf{s}|^{(l)}\mathbf{s}^{\cdot(l)})$$
   IF $(|E|^2<\varepsilon_E)$ *THEN*
   Solution is given with $(\boldsymbol{\alpha}^*,\mathbf{A}^*,|\mathbf{s}|^*,\mathbf{s}^{\cdot*})=(\boldsymbol{\alpha}^{(l)},\mathbf{A}^{(l)},|\mathbf{s}|^{(l)},\mathbf{s}^{\cdot(l)})$
   ELSE
   go to 2
   END

Fig. 9. Results of blind de-mixing for *linear space-invariant* imaging problem. From left to right are: (a) source images; (b) *space-invariant* noise free linear mixture; (c) recovery of the source images using Helmholtz free energy (50)/(51) and Cauchy annealing based linear blind inversion algorithm described in Section 3; (d) recovery of the source images using Infomax ICA algorithm [6]. Due to the space-invariant nature of the mixing only one pixel had to be solved by Helmholtz free energy and Cauchy annealing algorithm.

where η is a small learning gain, $\mathbf{y}$ is the ANN's output vector of separated signals, and φ is the nonlinearity in the optimal case given with $\varphi(y_j)=-(\mathrm{d}p(y_j)/\mathrm{d}y_j)/p(y_j)$. Due to both the space-invariant and the linear nature of the mixture, the ICA algorithm (51) gave good result. However, we emphasize that, in the space-invariant

case only, one pixel has to be solved with the Helmholtz free energy and Cauchy annealing algorithm first; then, the mixing matrix that is found can be applied to all the pixels. This makes the computation of the Helmholtz free energy and Cauchy annealing algorithm invariant w.r.t. image size. Fig. 10 shows results of the blind de-mixing for the linear space-variant imaging problem using a space-variant mixing matrix based on data model (47) and Fig. 8. From left to right: (a) source images; (b) space-variant, noise-free linear mixtures; (c) recovery of the source images using the linear version of the Helmholtz free energy blind inversion algorithm (45)–(46) with Cauchy annealing; and (d) recovery of the source images using ICA algorithm (51) [6,9,46]. Due to the space-variant nature of the mixture that changes from pixel to pixel, the stochastic ICA algorithm (51) gave poor result. Why? The changes occurred so quickly that the adaptive ICA algorithm (51) could not converge. This happened because, as already explained and illustrated by Fig. 8, the mixing matrix was changing row-wise, but conversion from a $2-D$ image to a $1-D$ signal to form the data model (45)/(47) had been done columnwise. The result: the mixing matrix was effectively changing from pixel to pixel, so the adaptive ICA algorithm (51) was not able to converge. Fig. 11 shows results of blind de-mixing for the nonlinear space-invariant imaging problem. From left to right: (a) source images; (b) space-invariant, noise-free nonlinear mixtures; (c) recovery of the source images using the Helmholtz free energy-based nonlinear blind inversion algorithm (49)–(50) with Cauchy annealing; and (d) recovery of the source images using Taleb–Jutten BSS algorithm [46] derived for post-nonlinear mixtures. This took advantage of the fact that the type of nonlinearity was known and given with (50). As already discussed, this inverse nonlinearity was formulated in (48) avoiding estimation of the inverse nonlinearity. Thus, Taleb–Jutten algorithm was put in a fair position relative to our method. Also, we took advantage of the fact that image data used in the simulation

Fig. 10. Results of blind de-mixing for *linear space-variant* imaging problem. From left to right are: (a) source images; (b) *space-variant* noise free linear mixture; (c) recovery of the source images using Cauchy annealing based linear blind inversion algorithm described in Section 3; (d) recovery of the source images using Infomax ICA algorithm [6]. Due to the space-variant nature of the mixture that changes from pixel to pixel, stochastic ICA algorithm gave poor result.

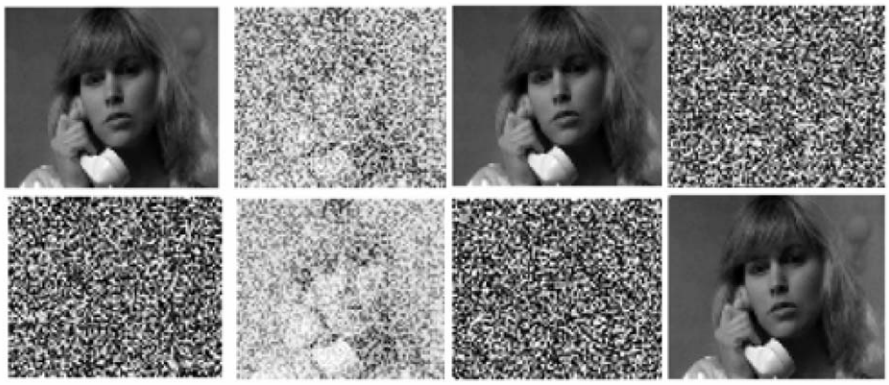

Fig. 11. Results of blind de-mixing for *nonlinear space-invariant* imaging problem. From left to right are: (a) source images; (b) *space-invariant* noise free *nonlinear* mixture; (c) recovery of the source images using Cauchy annealing based nonlinear blind inversion algorithm described in Section 3; (d) recovery of the source images using BSS algorithm derived for post-nonlinear mixture [46]. Due to the space-invariant nature of the mixture linear BSS algorithm [46] gave good results.

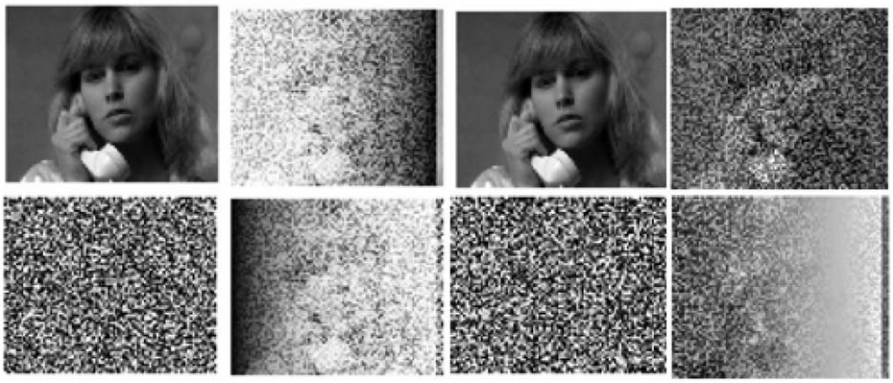

Fig. 12. Results of blind de-mixing for *nonlinear space-variant* imaging problem. From left to right are: (a) source images; (b) *space-variant* noise free *nonlinear* mixture; (c) recovery of the source images using Cauchy annealing based nonlinear blind inversion algorithm described in Section 3; (d) recovery of the source images using BSS algorithm derived for post-nonlinear mixture [46]. Due to the space-variant nature of the mixture stochastic nonlinear BSS algorithms gave poor result.

are sub-Gaussian, so that the nonlinearity φ used in (51) could be fixed. Consequently, there was no necessity to estimate score functions in the Taleb–Jutten algorithm. We have used nonlinear function $\varphi(y_j) = \text{sign}(y_j)y_j^2$, which is known to be good for sub-Gaussian data. As can be seen from Fig. 11, due to the space-invariant nature of the mixture, the post-nonlinear BSS algorithm [46] performed well. Finally, Fig. 12 shows results of blind de-mixing for the nonlinear space-variant imaging problem where nonlinearity (46) has been included and the mixing matrix has been

changed row-wise, as already described and illustrated in Fig. 8. From left to right: (a) source images; (b) space-variant, noise-free nonlinear mixture; (c) recovery of the source images using the Helmholtz free energy-based nonlinear blind inversion algorithm (49)–(50) with Cauchy annealing; and (d) recovery of the source images using Taleb–Jutten post-nonlinear mixture BSS algorithm [46]. Due to the space-variant nature of the mixture that effectively changed from pixel to pixel, the stochastic BSS algorithm [46] performed poorly.

5. Conclusion

The stochastic gradient is formulated based on deterministic gradient augmented with Cauchy simulated annealing capable of avoiding local minimums with a convergence speed significantly faster in relation to when simulated annealing is used alone and still being capable of reaching global minimum. In order to solve highly non-stationary linear inverse problems known as blind source separation a novel contrast function known as the Helmholtz free energy, $H = E - T_0 S$, with imposed thermodynamics constraint at a constant temperature T_0 was introduced generalizing the Shannon maximum entropy S of the closed systems to the open systems having non-zero input–output energy exchange E. Here, only the input data vector was known while source vector and mixing matrix were unknown. A stochastic gradient was successfully applied to solve inverse space-variant imaging problems on a concurrent pixel-by-pixel basis with the unknown mixing matrix (imaging point spread function) varying from pixel to pixel.

Appendix A. Cauchy PDF in hyper-spherical coordinates

To apply Cauchy annealing theory on the D-dimensional non-convex optimization problems, we need to generate the D-dimensional Cauchy distribution given by [43]

$$p(\mathbf{x}) = \frac{c}{[c^3 + |\mathbf{x}|^2]^{(D+1)/2}}. \tag{A.1}$$

If the parameter vector $\mathbf{x}$ is transformed from the Cartesian to hyper-spherical coordinates, then the problem of generating one D-dimensional distribution $p(\mathbf{x})$ is transformed into the problem of generating D one-dimensional pdfs, $p(\theta_1)p(\theta_2)....p(\theta_{D-1})p(r)$ where θ_i $i = 1, 2, \ldots, D-1$ angles are in the hyper-spherical coordinate system and r is the magnitude of the D-dimensional parameter vector $\mathbf{x}$. The one-dimensional distribution can then be derived from the identity

$$\begin{aligned} &\int\int\ldots\int p(x_1, x_2, \ldots, x_D)\,\mathrm{d}x_1\,\mathrm{d}x_2\ldots\mathrm{d}x_D \\ &\quad = \int\int\ldots\int p(\theta_1)p(\theta_2)...p(\theta_{D-1})p(r)\,\mathrm{d}\theta_1\,\mathrm{d}\theta_2\ldots\mathrm{d}\theta_{D-1}\,\mathrm{d}r = 1. \end{aligned} \tag{A.2}$$

In order to derive the right-hand side of Eq. (A.2), a determinant of the Jacobin $|\mathbf{J}|$ of the coordinate transformation must be derived due to the equality:

$$\mathrm{d}x_1\,\mathrm{d}x_2\ldots\mathrm{d}x_D = |\mathbf{J}|\,\mathrm{d}\theta_1\,\mathrm{d}\theta_2\ldots\mathrm{d}\theta_{D-1}\,\mathrm{d}r, \tag{A.3}$$

where the Jacobin matrix of the transformation is given by

$$\mathbf{J} = \begin{bmatrix} \frac{\partial x_1}{\partial\theta_1} & \frac{\partial x_1}{\partial\theta_2} & \ldots. & \frac{\partial x_1}{\partial r} \\ \frac{\partial x_2}{\partial\theta_1} & \frac{\partial x_2}{\partial\theta_2} & \ldots. & \frac{\partial x_2}{\partial r} \\ \ldots. & \ldots. & \ldots. & \ldots. \\ \frac{\partial x_D}{\partial\theta_1} & \frac{\partial x_D}{\partial\theta_2} & \ldots. & \frac{\partial x_D}{\partial r} \end{bmatrix} \tag{A.4}$$

and relations between Cartesian and hyper-spherical coordinates are given by

$$\begin{aligned} x_1 &= r\cos\theta_1, \\ x_k &= r\left(\prod_{i=1}^{k}\sin\theta_i\right)\cos\theta_{k+1}, \\ x_D &= r\prod_{i=1}^{D}\sin\theta_i. \end{aligned} \tag{A.5}$$

For the general D-dimensional case determinant of the Jacobian could be written as

$$|\mathbf{J}| = f_r(r)f_1(\theta_1)f_2(\theta_2)\ldots f_{D-1}(\theta_{D-1}). \tag{A.6}$$

For a polar coordinate system, the determinant of the Jacobian is given by $|\mathbf{J}| = r$, so, consequently, $f_r(r) = r$, $f_1(\theta_1) = 1$. For a spherical coordinate system, the determinant of the Jacobian is given by $|\mathbf{J}| = r^2\sin\theta_1$; consequently, $f_r(r) = r^2$, $f_1(\theta_1) = \sin\theta_1$, $f_2(\theta_2) = 1$. Now, taking into account that $|\mathbf{x}|^2 = r^2$ and taking into account (A.6), Eq. (A.2) can be written as

$$\begin{aligned} &\iiint p(x_1, x_2, \ldots, x_D)\,\mathrm{d}x_1\,\mathrm{d}x_2\ldots\mathrm{d}x_D \\ &\quad= \iiint \frac{c}{(c^2+r^2)^{(D+1)/2}} f_r(r)f_1(\theta_1)f_2(\theta_2)\ldots \\ &\qquad\times f_{D-1}(\theta_{D-1})\,\mathrm{d}r\,\mathrm{d}\theta_1\,\mathrm{d}\theta_2\ldots\mathrm{d}\theta_{D-1} \\ &\quad= \int \frac{c}{(c^2+r^2)^{(D+1)/2}} f_r(r)\,\mathrm{d}r \int f_1(\theta_1)\,\mathrm{d}\theta_1\ldots\int f_{D-1}(\theta_{D-1})\,\mathrm{d}\theta_{D-1} \\ &\quad= 1. \end{aligned} \tag{A.7}$$

Eq. (A.7) shows that, after transformation from Cartesian to polar coordinates, the problem of generating one D-dimensional Cauchy pdf is replaced by the problem of generating D one-dimensional pdfs. Related one-dimensional pdfs can be derived

from (A.7)

$$p(r) = K_r \frac{c}{(c^2 + r^2)^{(D+1)/2}} f_r(r),$$
$$p(\theta_k) = K_k f_k(\theta_k) \quad k = 1, 2, \ldots, D-1, \tag{A.8}$$

where constants K_r and K_k introduced in (A.8) are from normalization:

$$\int_{r_{\min}}^{r_{\max}} p(r)\,\mathrm{d}r = 1$$
$$\int_{\theta_{k\min}}^{\theta_{k\max}} p(\theta_k) = 1 \quad k = 1, 2, \ldots, D-1. \tag{A.9}$$

Upon transformation from Cartesian to polar coordinates, we get

$$p(\theta_1) = \frac{1}{\theta_{1\max} - \theta_{1\min}},$$

$$p(r) = K_r \frac{cr}{(r^2 + c^2)^{3/2}},$$

$$K_r = \frac{1}{c}\left(\frac{1}{\sqrt{r_{\min}^2 + c^2}} - \frac{1}{\sqrt{r_{\max}^2 + c^2}}\right)^{-1}. \tag{A.10}$$

In order to generate distributions for (A.10) w.r.t. uniform distribution on the interval [0,1], the use of the following identity is made:

$$|p(y)\,\mathrm{d}y| = |p(x)\,\mathrm{d}x|, \tag{A.11}$$

where y was distributed according to some general distribution and x was distributed uniformly on the interval [0,1]. For the two-dimensional case given by (A.10), from condition (A.11), we get

$$\theta_1 = (\theta_{1\max} - \theta_{1\min})x + \theta_{1\min},$$

$$r = K_r \frac{c}{\tilde{x}}\sqrt{1 - K_r^2 \tilde{x}^2},$$

$$\tilde{x} = (\tilde{x}_{\max} - \tilde{x}_{\min})x + \tilde{x}_{\min},$$

$$\tilde{x}_{\min} = \frac{K_r}{\sqrt{1 + r_{\min}^2}} \quad \tilde{x}_{\max} = \frac{K_r}{\sqrt{1 + r_{\max}^2}}, \tag{A.12}$$

where in (A.12), $\theta_{1\min}, \theta_{1\max}, r_{\min}$, and $r_{\max}$ are integration boundaries.

Appendix B. Biological conjecture of the Helmholtz free energy-based unsupervised learning

We conjectured that if Hebb synaptic weights were on the order of milli-Volts, then, from physical dimensionality analysis of the energy power viewpoint, Lagrange multipliers could be on the order of the pico-Ampere currents from the single dendrite ion channels. These currents were demonstrated by 1991 Nobel Laureates Erwin Neher and Bert Sakmann from the Max Planck Institute [31]. As matter of fact, it is almost a miraculous coincidence that ten of thousands of neurons forming the associative memory in fully connected neural networks in the hippocampus are energetically equivalent to the Poisson fluctuations of neurotransmittals. For these neurotransmittals, the mean by definition equals the variance, as is meaningfully defined to be the brain cybernetic temperature 37 °C (with the help of physics conversion factor $310\,K_B T$ at brain temperature $= (37/27) \times (1/40)$ eV). Furthermore, with brain anatomy revealing billions of neurons (defined by the binding between pair firing rates for the Hebbian synaptic memory) and an equal amount of housekeeping glial cells, this new Lagrange dynamic variable may represent the housekeeping glial cells' supporting dendrite tree signal pre-conditioning in the unsupervised learning methodology. Active roles of these glial cells in information processing are suspected but not yet explicitly verified. (While there are hundreds of types of neurons, only three classes of glial cells exist to keep the thermodynamic balance. The first type is (i) astrocytes, or glial cells that provide the glue, so-to-speak glue, of blood vessels to the neurons. The second type is the oligodendroglia/cytes, that provide the myelin sheath wrapped around the central nerve system like a link of sausages forming an express way. The third and last type is the Schwann cells, which wrap around the peripheral nervous system other than the brain and the spinal cord.) We expect that the role of glial cells is above and beyond the housekeeping of brain activity and that they are also necessary for unsupervised learning via the equilibrium heat reservoir. Thus, both actions by neurons and the reactions of housekeeping glial cells are simultaneously present in our ANN model. Any direct or indirect evidence in biological and neuro-psychological experiments would be welcomed. Statistical mechanics, which allow the macroscopic world to be understood in terms of microscopic properties of chemicals, provides some basic tools for investigating human natural intelligence. However, the results of applying statistical mechanics to learning in the human brain need to be supported and confirmed by other methods. These include (i) advanced instrumentation (PET, F_NMR, SQUIB, EO/IR Brain Imaging, etc.); (ii) interdisciplinary investigations into biological relevance mentioned in this conclusion (there are 10 orders of magnitude spanning the scales from CNS to DNA); and (iii) understanding/analysis of vector time series of input pairs used in unsupervised learning.

This study into truly unsupervised learning began with a question: why do mammalian brains employ pairs of inputs to feed a single output? One obvious answer would be to accommodate pairs of sensors (e.g., eyes, ears); another would be that the redundancy supports wet-ware fault tolerance. There was, however, another possibility: an unsupervised learning strategy that involved "squeezing out the

garbage.'' Consider this: if a sensor pair could reject misinformation (garbage) by simple in situ comparison, what remained was the unknown but wanted signal. Not unlike common mode rejection in an electrical circuit, this process could provide instantaneous correlation for selective feature amplification, a great binaural hearing aid. Thus, the computing adage ''garbage in, garbage out'' can be modified to suit this strategy for a pair of smart receivers: ''raw pair of energies in, garbage entropy out.'' This learning strategy might be discussed in terms of a sensor-pair time-series vector representation (as opposed to a single-sensor time-series scalar), which takes advantage of brainwave diffusion to reduce the redundancy to conserve energy and make room for upcoming new excitations. With this strategy (and vector representation), a synapse might filter out garbage and maintain an accurate external world representation without the need for a teacher.

There are vast differences between natural intelligence, as was encountered in human biology [5,7,36], and artificial intelligence, as encountered in artificial neural networks (ANN) [14,25], although some similarity exists. It is interesting to examine how each form of intelligence handles unsupervised learning. An unsupervised artificial neural network evolves—it essentially teaches itself to extract features or regularities present in input vectors it receives [19]. In biological naturally intelligent system, this functionality is realized with greater underlying complexity than more manipulation of network weights (e.g., temporally driven consolidation of short-term memory into long-term memory in humans [38]) and far more neurons than is possible to implement in an artificial network.

So far, we have briefly reviewed (i) global equilibrium body temperature regulation and (ii) housekeeping glial cells. We used these to prove (iii) the convergence of our statistical mechanics model of unsupervised learning. We also demonstrated (iv) a quenching of local r.m.s fluctuations in annealing to satisfy constraints. One important application of the discussed unsupervised learning methodology might be the solution of inverse problems. Can one solve the linear system of equations $\mathbf{x} = \mathbf{As}$ without knowing the mixing matrix $\mathbf{A}$? Several groups [11,6,4,3,10,9,12,13,23,24,26,33,41,44,46] offer ANN solutions in terms of two different approaches:

(i) The statistical approach is based on the ensemble average [11,6,4,3,10,9,12,13, 23,24,26,33,46], where one assumes for all pixels an unknown mixing matrix $\mathbf{A}$ which was valid in the space-invariant imaging $\mathbf{A}$. The ANN algorithm was able to solve for the inverse matrix $\mathbf{W} \cong \mathbf{A}^{-1}$ by exploiting the principle of statistical independence. The missing information could be derived from neighborhood pixels statistics by assuming space-invariant problem, in which case one gains additional information by taking into account those neighborhood pixel measurements without increasing the number of unknown $\mathbf{A}$'s. For that case, a MaxEnt natural gradient neural network algorithm has been derived see Fig. 1 as well as [3,4,6,10,33].

$$\frac{\partial \mathbf{W}}{\partial t} = \left\langle \frac{\partial S(\mathbf{y}(\mathbf{Wx}))}{\partial \mathbf{W}} \mathbf{W}^{\mathrm{T}}\mathbf{W} \right\rangle_{\text{pixels}}. \tag{B.1}$$

We would augment the following stochastic gradient learning with fast Cauchy annealing:

$$\frac{\partial \mathbf{W}}{\partial t} = F(t) = \left\langle \frac{\partial S}{\partial \mathbf{W}} \mathbf{W}^{\mathrm{T}}\mathbf{W} \right\rangle_{\text{pixels}} + \tilde{F}(t). \tag{B.2}$$

(ii) Space variant problems [37,41,43,44], where each pixel has a different mixing matrix **A**, can be solved by the Lagrange constrained neural network (LCNN). Giving the data model $\mathbf{x} = \mathbf{As}$ as a linear photon source de-mixing problem, the LCNN found unknown **s** and **A** by adopting the Lagrange constraint methodology with demonstrated real word applications in remote sensing [41] and infrared breast cancer detection [45].

References

[1] D.H. Ackley, G.E. Hinton, T.J. Sejnowski, A learning algorithm for Boltzmann Machines, Cognitive Sci. 9 (1985) 147–169.
[2] S. Amari, Information geometry, in: H. Nencka, J.-P. Bourguignon (Eds.), Geometry and Nature: Contemporary Mathematics, vol. 203, 1997, pp. 81–95.
[3] S. Amari, Natural gradient works efficiently in learning, Neural Comput. 10 (1998) 251–276.
[4] S. Amari, A. Cihocki, H.H. Yang, A new learning algorithm for blind signal separation, Advances in Neural Information Processing Systems, vol. 8, MIT Press, Cambridge, MA, 1996, pp. 757–763.
[5] E. Antebi, D. Fishlock, Biotechnology Strategies for life, MIT Press, Cambrdge, MA, 1986.
[6] A.J. Bell, T.J. Sejnowski, An information-maximization approach to blind separation and blind deconvolution, Neural Comput. 7 (1995) 1129–1159.
[7] E. Bruce Goldstein, Sensation & Perception, 4th ed., Brooks/Cole Publishing Co., Pacific Grove, CA, 1996.
[8] G. Bureal, Blind separation of sources: a nonlinear neural algorithm, Neural Networks 5 (1992) 937–947.
[9] J.F. Cardoso, Infomax and maximum likelihood for blind source separation, IEEE Signal Process. Lett. 4 (1997) 112–114.
[10] J.F. Cardoso, B. Laheld, Equivariant adaptive source separation, IEEE Trans. Signal Process. 44 (12) (1996) 3017–3030.
[11] J.F. Cardoso, A. Souloumiac, Blind beamforming for non-Gaussian signals, Proc. IEE F 140 (1993) 362–370.
[12] A. Cichocki, S. Amari, Adaptive Blind Signal and Image Processing, Wiley, New York, 2002.
[13] P. Comon, Independent component analysis, A new concept?, Signal Process. 36 (1994) 287–314.
[14] H. Curtis, N.S. Barnes, Invitation to Biology, 4th ed., Worth Publishers, Inc., New York, 1985.
[15] R.O. Duda, P.E. Hart, D.G. Stork, Pattern Classification, 2nd ed., Wiley, New York, 2001.
[16] R. Durbin, D. Wilshaw, An analogue approach to the traveling salesman problem using an elastic net method, Nature 326 (1987) 689–691.
[17] S. Geman, D. Geman, Stochastic relaxation, Gibbs distribution and the Bayesian restoration of images, IEEE Trans. Pattern Anal. Mach. Intel. PAMI-6 (6) (1984) 721–741.
[18] S. Grossberg, Adaptive pattern classification and universal recording: I. parallel development and coding of neural feature detectors, Biol. Cyber. 23 (1976) 121–134.
[19] A. Guyton, Textbook of Medical Physiology, 8th ed., W. B. Saunders Company (Harcourt Brace Jovanovich, Inc.), London, 1991.
[20] J.J. Hopfield, Neural network and physical systems with emergent collective computational abilities, Proc. Natl. Acad. Sci. 79 (1982) 2554–2558.

[21] J.J. Hopfield, D.W. Tank, Neural computation of decisions optimization problem, Biol. Cyber. 52 (1985) 141–152.
[22] K. Huang, Statistical Mechanics, Wiley, New York, 1963.
[23] A. Hyvärinen, J. Karhunen, E. Oja, Independent Component Analysis, Wiley, New York, 2001.
[24] A. Hyvärinen, E. Oja, A fast fixed-point algorithm for independent component analysis, Neural Comput. 9 (1997) 1483–1492.
[25] J.-S.R. Jang, C.-T. Sun, E. Mizutani, Neuro-fuzzy and Soft Computing: a Computational Approach to Learning and Machine Intelligence, Prentice-Hall, Englewood Cliffs, NJ, 1997.
[26] Ch. Jutten, J. Herault, Blind separation of sources, Part I: An adaptive algorithm based on neuromimetic architecture, Signal Process. 24 (1991) 1–10.
[27] I. Kopriva, H. Szu, Blind inversion in nonlinear space-variant imaging by using Cauchy machine, Proceedings of the SPIE, vol. 5102, Independent Component Analysis, Wavelets and Neural Networks, Orlando, FL, April 22–25, 2003, pp. 5–16.
[28] I. Kopriva, H. Szu, Space-time variant blind sources separation with additive noise, in: C.G. Puntonet, A. Prieto (Eds.), Proceedings of the Fifth International Conference on Independent Component Analysis and Blind Signal Separation, Granada, Spain, Lecture Notes in Computer Science, vol. 3195, Springer, Berlin, 2004, pp. 240–247.
[29] J.L. Lawson, G.E. Uhlenbeck, Threshold Signals, vol. 24, MIT Radiation Laboratory Series, 1950.
[30] N. Metropolis, A.W. Rosenbluth, M.N. Rosenbluth, A.H. Teller, E. Teller, Equation of state calculations by fast computing machines, J. Chem. Phys. 21 (1953) 1087–1092.
[31] E. Neher, B. Sakmann, Single-channel currents recorded from membranes of denervated from muscle fibers, Nature 260 (1976) 799–802.
[32] J. Nocedal, S.J. Wright, Numerical Optimization, Springer, Berlin, 1999, pp. 490–526.
[33] D.T. Pham, P. Garat, Blind separation of mixtures of independent sources through a quasimaximum likelihood approach, IEEE Trans. Signal Process. 45 (7) (1997) 1712–1725.
[34] D.E. Rumelhart, G.E. Hinton, R.J. Williams, Learning internal representations by error propagation, Parallel Distributed Processing, MIT Press, Cambridge, MA, 1986.
[35] J. Solé-Casals, M. Babaie-Zadeh, Ch. Juten, D.T. Pham, Improving algorithm speed in PNL mixture separation and Wiener system inversion, in: S. Amari, A. Cichocki, S. Makino, N. Murata (Eds.), Proceedings of the Fourth International Symposium on Independent Component Analysis and Blind Signal Separation, Nara, Japan, 2003, pp. 639–644.
[36] E.P. Solomon, R.R. Schmidt, P.J. Adragna, Human Antanomy & Physiology, 2nd ed., Sunders College Publishing, Harcourt Brae Publishers, 1990.
[37] H. Szu, Progresses in unsupervised artificial neural networks of blind image demixing, IEEE Ind. Elec. Soc. Newsletter, June, 1999, pp. 7–12.
[38] H. Szu, Thermodynamics energy for both supervised and unsupervised learning neural nets at a constant temperature, Int. J. Neural Systems 9 (1999) 175–186.
[39] H. Szu, R. Hartley, Fast simulated annealing, Phys. Lett. A 122 (3) (1987) 157–162.
[40] H.H. Szu, R.L. Hartley, Nonconvex optimization by simulated annealing, Proc. IEEE 75 (11) (1987) 1538–1540.
[41] H.H. Szu, C. Hsu, Landsat spectral Unmixing à la superresolution of blind matrix inversion by constraint MaxEnt neural nets, Proc. SPIE 3078 (1997) 147–160.
[42] H. Szu, I. Kopriva, Artificial neural networks for noisy image super-resolution, Opt. Commun. 198 (1–3) (2001) 71–81.
[43] H. Szu, I. Kopriva, Comparison of the Lagrange constrained neural network with traditional ICA methods, Proceedings of the IEEE 2002 World Congress on Computational Intelligence-International Joint Conference on Neural Networks, Hawaii, USA, May 17–22, 2002, pp. 466–471.
[44] H. Szu, I. Kopriva, Deterministic blind source separation for space variant imaging, in: S. Amari, A. Cichocki, S. Makino, N. Murata (Eds.), Proceedings of the Fourth International Symposium on Independent Component Analysis and Blind Signal Separation, Nara, Japan, 2003, pp. 669–674.
[45] H. Szu, I. Kopriva, P. Hoekstra, N. Diakides, M. Diakides, J. Buss, J. Lupo, Early tumor detection by multiple infrared unsupervised neural nets fusion, 25th Annual International Conference of the IEEE Engineering in Medicine and Biology Society, Cancun, Mexico, September 17–21, 2003, pp. 1133–1136.

[46] A. Taleb, Ch. Jutten, Source Separation in Post-Nonlinear Mixtures, IEEE Trans. Signal Process. 47 (10) (1999) 2807–2820.

Harold Szu received a Ph.D. in Statistical Mechanics from Rockefeller University in 1971 and worked at NRL in Plasma Physics, Optics, and EW over 15 years (1977–1990), and was Info Sci Group Leader of NSWC at White Oak (1990–1996) and Lamson endowed Chair Professor of U. of Louisiana at Lafayette (1996–1998). He returned to the Navy to build "WaveNet" for achieving a live video 1000:1-compression through SINCGARS radio and received an Army Medal, and then headed ONR all Digital Array Radar test bed. He is the director of Digital Media RF lab and Research Professor of GWU, Washington DC.

He has been one of the early neural network researchers and is responsible for one of the recent breakthrough's of unsupervised learning methodology of pair sensors with the motto:—as opposed to dumb PC "garbage in—garbage out"—smart bio-sensors—"raw sensory pairs inputs, squeeze out the known garbage" without the need of a teacher and what is retained in the synaptic weight matrix is known as the information factor code or independent-density component analyses (ICA), as opposed to classical PCA based on the covariance statistics. He is known for proving the fast cooling schedule for Cauchy simulated annealing algorithm for global optimum search: the completeness theorem of adaptive wavelet transform representing speech phonemes and applying ICA to slur word for de-hyphenation. He contributed the image wavelet textures and ICA remote sensing achieving systematically multiple class-labels per pixel for the surveillance of Amazon deforestation.

Dr. Szu is one of the founders, former president, and a governor of INNS for SIG/Chapter development, a Champion of brain-style computing. He has educated a dozen Ph.D. students, published three hundreds of papers, ten patents and several books and journals. He is Fellow of SPIE (1995) for neural nets, OSA(1996) for adaptive wavelets, IEEE (1997) for bio-sensors, and Academician of Russian Academy of Nonlinear Sciences (1999) for the convergence proof of homeostatic learning theorem postulating the Lyapunov-like Helmholtz energy based on a cybernetic temperature.

Ivica Kopriva received the B.S. degree in electrical engineering from Military Technical Faculty, Zagreb, Croatia in 1987, and M.S. and Ph.D. degrees in electrical engineering from the Faculty of Electrical Engineering and Computing, Zagreb, Croatia in 1990 and 1998, respectively. Currently, he is senior research scientist at George Washington University, Department of Electrical and Computer Engineering. His current research activities are related to the unsupervised learning theory with application in solving blind imaging problems and higher order statistics based array signal processing with the application on the direction finding problems.