MAXIMUM ENTROPY AND MINIMAL MUTUAL INFORMATION IN A NONLINEAR MODEL

Fabian J. Theis, Elmar W. Lang

Institute of Biophysics
University of Regensburg, D-93040 Regensburg, Germany
email: fabian.theis@mathematik.uni-regensburg.de

ABSTRACT

In blind source separation, two different separation techniques are mainly used: Minimal Mutual Information (MMI), where minimization of the mutual output information yields an independent random vector, and Maximum Entropy (ME), where the output entropy is maximized. However, it is yet unclear why ME should solve the separation problem, ie. result in an independent vector.

Amari has given a partial confirmation for ME in the linear case in [1], where he proves that under the assumption of vanishing expectancy of the sources ME does not change the solutions of MMI up to scaling and permutation.

In this paper, we generalize Amari's approach to nonlinear ICA problems, where random vectors have been mixed by output functions of layered neural networks. We show that certain solution points of MMI are kept fixed by ME if no scaling of the weight vectors is allowed. In general, ME however might leave those MMI solutions using diagonal weights in the first network layer. Therefore, we conclude this paper by suggesting that in nonlinear ME algorithms diagonal weights should be fixed in later epochs.

1. INTRODUCTION

Independent component analysis (ICA) describes methods to extract statistically independent components from a given random vector. One main application of ICA is to solve the blind source separation (BSS) problem which is, given only the mixtures of some underlying independent sources, to separate the mixed signals thus recovering the original sources. Here neither the sources nor the mixing process is known, hence the term *blind* source separation. In contrast to correlation-based transformations such as principal component analysis, ICA renders the output signals as statistically independent as possible by evaluating higher-order statistics. The idea of ICA was first expressed by Jutten and Herault [2] while the term ICA was first coined by Comon in [3]. However the field became popular only with the seminal paper by Bell and Sejnowski [4] who elaborated upon the Infomax-principle first advocated by Linsker [5] [6].

Classically, *linear* BSS has been treated most thoroughly [7] [8], where the mixing function of the source signals corresponds to a linear function (matrix). Two principles became popular: Minimal Mutual Information (MMI), where minimization of the mutual output information yields an independent random vector, and Maximum Entropy (ME), where the output entropy is maximized. Concerning MMI and ME, Comon proposed to use the mutual information of the output as a contrast function because minimizing the mutual information (MMI) induces statistical independence of the output. This has to be compared with Bell and Sejnowski's suggestion to maximize the entropy (ME) of the output. But ME does not always induce MMI ([4], section 4 and [9]) and therefore statistical independence. ME performs best when the non-linear demixing function in the ME algorithm matches with the cumulative distribution of the given source.

However, nowadays many algorithms are based on the ME contrast function. This raises the question how this large branch of ICA research compares to the MMI methods. For the linear case, Amari [1] provides a partial answer confirming the equivalence of MMI and ME under certain conditions. He shows that at solution points of ICA determined by MMI, ME algorithms will only change the demixing matrices in diagonal directions thus describing the same solution. This can be interpreted as a local justification for ME, showing that demixing matrices are stable fixpoints (up to scaling) of ME algorithms.

With the growing popularity of ICA, more and more nonlinear algorithms have been proposed (for example [10] [11], [12], [13]) and theoretic approaches have been taken [14]. Most of them are based on the ME contrast function. Yet, to the knowledge of the authors, in the more general nonlinear setting no justification for ME based algorithms has been given. In this paper, we want to discuss a nonlinear model described by the activation function of a multilayered neural network. We will show that at certain solution points of MMI, ME will have a local extremum thus only changing the solution in a well specified way (scaling in all weight matrices).

2. THE MODEL

In the quadratic case of blind source separation, a random vector $X : \Omega \to \mathbb{R}^N$ called **mixed vector** is given, where Ω is a fixed probability space; it comes from an independent random vector $S : \Omega \to \mathbb{R}^N$, which will be called **source vector**, by mixing with an invertible **mixing function** $\mu : \mathbb{R}^N \longrightarrow \mathbb{R}^N$, ie. $X = \mu \circ S = \mu(S)$. Only the mixed vector is known, and the task is to recover μ and therefore $S = \mu^{-1} \circ X$.

Suppose, we are given a **transformation space** $\mathcal{T} \subset \{f : \mathbb{R}^N \longrightarrow \mathbb{R}^N \mid f \text{ measurable }\}$ and a **contrast function** $\kappa : \mathcal{T} \longrightarrow \mathbb{R}$. We want to study the space of all transformed random vectors $\tau \circ X$ where $\tau \in \mathcal{T}$ is a transformation as above. To be more precise, the goal of an ICA algorithm is to minimize κ. The reason for this is that if we take for example the **canonical contrast function** $\kappa(\tau) := I(\tau \circ X)$, $\tau \in \mathcal{T}$, where $I(Y)$ denotes the **mutual information** of the random vector Y, minimizing κ means minimizing the mutual information, so at a minimum τ_0, the demixed random vector $\tau_0 \circ X$ is as independent as possible. This method is then called **minimal mutual information** (MMI). Another often used contrast function is the negative entropy $\kappa(\tau) := -H(g\tau \circ X)$ (here g is a fixed vector of one-dimensional bounded functions in order to make the entropy finite). The induced minimization method, denoted by **maximum entropy** (ME), is usually easier to implement; we will show that it can lead to independence of components under mild assumptions, nonetheless.

If we furthermore assume that the inverse of the mixing function lies already in the transformation space ($\mu^{-1} \in \mathcal{T}$), then we know that the global minimum of the canonical contrast function has value 0, so indeed a global minimum will give us an independent random vector. Of course we cannot hope that μ^{-1} will be found because uniqueness in this general setting cannot be achieved, see for example [15] — in contrast to the linear case ($\mathcal{T} \subset \mathrm{Gl}(N)$) as shown in [3]. This will give us a further restriction of our method.

Obviously the setting above is far too general to give any computational results, because it is not possible to do minimization if it is not clear how to describe elements of $\mathcal{T}$. Therefore we need a means of describing $\mathcal{T}$ by a finite set of real-valued parameters. For this, we consider output functions of neural networks, see [16] and [17] . This has the advantage that in neural networks, being adaptive systems, we know for a given energy function how to algorithmically minimize this function for example using the standard accelerated gradient descent method. Moreover, more general functions can then be approximately learned using the fact that sufficiently complex neural networks are so called universal approximators, see for example [18].

Fixing notation, for any neural net Γ with N inputs and N outputs , $A(\Gamma) : \mathbb{R}^N \longrightarrow \mathbb{R}^N$ denotes its output function. For fixed Γ, $\Gamma(w)$ is the net Γ, where the weights have been replaced with the new weights $w \in \mathbb{R}^\omega$. Here ω denotes the total number of weights of the neural network. So we have $\mathcal{T} \subset \{A(\Gamma(w)) \mid w \in \mathbb{R}^\omega\}$ and $\overline{\kappa}(w) := \kappa(A(\Gamma(w))$ is to be minimized.

3. THE MAIN THEOREM

First note that MMI is obviously better suited than ME in terms of finding solutions; ME may terminate at points that do not represent demixing functions. Therefore even in the linear case ([1]) it can only be shown that solutions of MMI are also solutions of ME. For this, a uniqueness result by Comon ([3]) has to be used, where under the assumption that at most one source S_i is gaussian, it is shown that in the linear case all solutions of MMI are of the form PLA^{-1}, where A is the mixing matrix, P a permutation matrix and L a non-degenerate scaling matrix. Since no such uniqueness results have been found in the more general nonlinear setting, we will only be able to show that special demixing functions are solutions of ME. More precisely, let the mixing function satisfy (locally)

$$\mu = A^{(n)} \circ h^{(n)} \circ \ldots \circ A^{(1)} \circ h^{(1)},$$

where $A^{(\nu)} \in \mathrm{Mat}(N^{(\nu+1)} \times N^{(\nu)}; \mathbb{R})$ are matrices of full rank and $h^{(\nu)}$ are cartesian products of differentiable invertible functions (typically activation functions) for $\nu = 1, \ldots, n$. Hence, the mixing function is the output function of a layered neural network Γ with n layers.

Definition 3.1. *Let the network Γ be fully quadratic, ie. each layer has the same size. Then all activation functions $h^{(\nu)}$ and weight matrices $A^{(\nu)}$ are invertible for $\nu = 1, \ldots, n$, and we call functions of the type*

$$h^{(1)-1} \circ L^{(1)}P^{(1)}A^{(1)-1} \circ \ldots \circ h^{(n)-1} \circ L^{(n)}P^{(n)}A^{(n)-1}$$

scaled solutions of the BSS problem *if $L^{(i)}$ and $P^{(i)}$ are non-degenerate scaling and permutation matrices.*

Note that if the layers of the network have different sizes, additional nodes are to be added so that $A^{(i)-1}$ is well defined. This will be discussed in more detail in section 5. Since $h^{(1)}$ is a vector of activation functions of the network, it can be written as $h^{(1)} = h_1^{(1)} \times \ldots \times h_n^{(1)}$. Using the transformation theorem for entropies, we then calculate

$$\begin{aligned} I(h \circ X) &= \textstyle\sum_i H((h \circ X)_i) - H(h \circ X) \\ &= \textstyle\sum_i (H(X_i) + E_{X_i}(\log |h_i'|)) \\ &\quad -H(X) - E_X(\log |detDh|) \\ &= I(X) + \textstyle\sum_i E_{X_i}(\log |h_i'|) \\ &\quad -E_X(\log |\prod_i h_i'|) = I(X), \end{aligned}$$

where $Dh := \left(\frac{\partial h_i}{\partial x_j}\right)_{ij}$ denotes the derivative of $h : \mathbb{R}^n \to \mathbb{R}^n$. Therefore, scaled solutions with the same matrices

$L^{(i)}$ and $P^{(i)}$ for $i > 1$ have the same mutual information, but with different scaling or permutation matrices, the information might be different. Hence, *not* all scaled solutions are solutions of the BSS problem, but only those with $L^{(i)} = P^{(i)} = I$ for $i > 2$, where I denotes the identity matrix; in the $n = 1$ case those are already all solutions. However, this it not true for $n > 1$.

We assume that the used neural networks as well as the density functions of the appearing random variables are all twice continously differentiable.

The main goal of the paper is the following theorem:

Theorem 3.2. *Assume the expectancy $E(S)$ of the sources vanishes. Then scaled solutions that are solutions of the BSS problem are local extrema of ME. To be more precise: The ME algorithm transforms scaled solutions that are solutions of BSS into scaled solutions.*

4. PROOF OF THE MAIN THEOREM

The case of a single layer has been treated already by Amari. We want to restrict ourselves to the two-layered case. The case of multiple layers is an easy generalization.

Let Γ be a two-layered neural network with N inputs and N outputs, and let $w \in \mathbb{R}^n$ be a weight vector of Γ. For the network $\Gamma(w)$ with weights w, we will use the following notations: Let N_h be the number of neurons in the hidden layer. For now, let us assume that $N_h = N$. The more general case $N_h > N$ will be discussed in section 5. $W^{(1)} \in \mathrm{Mat}(N_h \times N; \mathbb{R})$ denotes the linear mapping from the input layer to the hidden layer, $W^{(2)} \in \mathrm{Mat}(N \times N_h; \mathbb{R})$ is the mapping from the hidden layer to the output layer. $g_i^{(1)}$, $i = 1, \ldots, N_h$ denote the activation functions of the hidden layer, $g_j^{(2)}$, $j = 1, \ldots, N$ those of the output layer. All activation functions are at least $\mathcal{C}^2$-Diffeomorphisms of $\mathbb{R} \longrightarrow (a, b)$ with $a, b \in \mathbb{R}$ fixed. Since we are only interested in a local description, we sometimes also restrict the domains of the activation functions to some small open interval. Furthermore, we write $g^{(1)}$ respectively $g^{(2)}$ for the cartesian product of the activation functions. Moreover, define random vectors $Y^{(1)} := W^{(1)}X$, $Z^{(1)} := g^{(1)}Y^{(1)}$, $Y^{(2)} := W^{(2)}Z^{(1)}$, and $Z^{(2)} := g^{(2)}Y^{(2)}$. Then $Z^{(2)} = A(\Gamma(w)) \circ X$. Instead of indexing the above mappings by the weight vector $w \in \mathbb{R}^n$, we will use the pair $(W^{(1)}, W^{(2)})$ of the weight matrices. The transformed random vectors $Y^{(i)}$ and $Z^{(i)}$ however will not get indices for clarity of notation. The random propabilities will be denoted by $\rho_Y^{(\nu)}$ and $\rho_Z^{(\nu)}$ for $\nu = 1, 2$ respectively, and we write for the marginal densities $\rho_{Y,i}^{(\nu)} := \pi_i \rho_Y^{(\nu)}$ and $\rho_{Z,i}^{(\nu)} := \pi_i \rho_Z^{(\nu)}$, where $\pi_i : \mathbb{R}^N \longrightarrow \mathbb{R}$ is the projection onto the i-th coordinate, ie. $\pi_i(x_1, \ldots, x_N) = x_i$.

Since the activation functions have been chosen to be bounded, the entropy of $Z^{(2)}$ is bounded for all weights:

$$\begin{array}{rcl} H(Z^{(2)}) & = & H(g^{(2)} \circ Y^{(2)}) \\ & \leq & \sum_{i=1}^{N} H(g_i^{(2)} \circ \pi_i \circ Y^{(2)}) \\ & \leq & N \ln(b - a) \end{array}$$

We have to show that for $i \neq j$ $\frac{\partial}{\partial w_{i,j}^{(\nu)}} H(Z^{(2)}) = 0$ holds at scaled solutions, where $W^{(\nu)} = \left(w_{i,j}^{(\nu)}\right)_{i,j}$.

By definition of the mutual information I, we have

$$H(Z^{(2)}) = -I(Z^{(2)}) + \sum_{i=1}^{N} H(Z_i^{(2)}).$$

The first summand vanishes at scaled solutions that are minima of MMI; we decompose the second term using the source densities

$$-\sum_{i=1}^{N} H(Z_i^{(2)}) = D(\tilde{\rho}_Z^{(2)} \| \rho_S) + C(W^{(1)}, W^{(2)})$$

where $C(W^{(1)}, W^{(2)}) := \sum_i \int_{\mathbb{R}} \log(\rho_{S_i}(z_i))\rho_{Z,i}(z_i)dz_i$ and $\tilde{\rho}_Z^{(2)} := \prod_{i=1}^{N} \rho_{Z,i}^{(2)}$, and $D(\tilde{\rho}_Z^{(2)} \| \rho_S) = \int_{\mathbb{R}^N} \tilde{\rho}_Z^{(2)} \log \frac{\tilde{\rho}_Z^{(2)}}{\rho_S}$ denotes the Kullback-Leibler divergence of $\tilde{\rho}_Z^{(2)}$ and ρ_S. Since the source densities ρ_S are independent, the random densities factorize and the decomposition follows.

At the very special scaled solution $A(\Gamma(w)) = \mu^{-1}$, we have $S = Z^{(2)}$, so that the first two summands $I(Z^{(2)})$ and $D(\tilde{\rho}_Z \| \rho_S)$ in the decomposition of $-H(Z^{(2)})$ at $W^{(\nu)} = A^{(\nu)-1}$ vanish and are therefore minimal. This is the case at all scaled solutions that are global minima of the mutual information. Hence, we only have to show that $C(W)$ has a local extremum at those points.

We restrict ourselves to the point $W^{(\nu)} = A^{(\nu)-1}$. As the invertible matrices $\mathrm{Gl}(N)$ are open in $\mathrm{Mat}(N \times N; \mathbb{R})$, there exists an open neighbourhood $U \subset \mathrm{Mat}(N \times N; \mathbb{R})$ of the identity I with $U \subset \mathrm{Gl}(N) \cap \{W | \ \|W\| < 1\}$ and $B^{(\nu)} := W^{(\nu)}A^{(\nu)} - I \in U$ for $W^{(\nu)}$ from a neighbourhood of $A^{(\nu)-1}$ and $\nu = 1, 2$. Hence locally, $W^{(\nu)}$ can be written as $W^{(\nu)} = (I + B^{(\nu)})A^{(\nu)-1}$. Using this decomposition, we can prove the following lemma:

Lemma 4.1. *Assume $E(S) = 0$. Then at the point $B^{(\nu)} = 0$, $\nu = 1, 2$ that is at $W^{(\nu)} = A^{(\nu)-1}$, the partial derivatatives of $C((I + B^{(1)})A^{(1)-1}, (I + B^{(2)})A^{(2)-1})$ satisfy*

$$\frac{\partial C}{\partial B_{ij}^{(1)}} = \frac{\partial C}{\partial B_{ij}^{(2)}} = -\delta_{ij}$$

The proof is given in section 7. The lemma shows that $C(W)$ is constant in non-diagonal directions; hence ME does not change solutions up to scaling, that is it maps into scaled solutions. The claim at other scaled solutions that are solutions of the BSS problem is shown similiarily by parametrizing $W^{(\nu)}$ by

$$W^{(\nu)} = (I + B^{(\nu)})P^{(\nu)}L^{(\nu)}A^{(\nu)-1}.$$

This shows the theorem for fully quadratic two-layered networks.

5. GENERALIZATIONS

For not fully quadratic neural nets, that is for nets with $N_h > N$ (this is the normal case, but $N_h < N$ can be shown similarily), we propose the following strategy to reduce this case to the fully quadratic case: Add $N_h - N$ in- and output neurons and introduce uniform independent input signals (enlarge X to a random vector on $\mathbb{R}^{N_h}$ such that the $N_h - N$ last components are independent). Then scaled solutions in the normal case induce scaled solutions in the fully quadratic case; using the above, we deduce that those are local extrema of ME in the quadratic case and hence extrema in the normal case, because the additional sources are uniformly distributed.

The generalization to multilayered neural networks can be done similarily using the above iteratively. This indicates how to prove the main theorem for those more general cases.

6. CONCLUSION

The ME algorithm has been studied, and we have shown that special MMI solutions are indeed local extrema of the maximum entropy method, up to scaling in the weight matrices. Therefore, we suggest ME algorithms which try to keep constant the diagonal weight entries in later epochs in order to stay at MMI minima.

We propose two further studies: It would be nice to know that those extrema are indeed stable fixpoints so it should be shown that scaled solutions are not only extrema but also maxima up to scaling and permutation. Furthermore, the theorem should be extended to general solutions of MMI which would solve the uniqueness problem in the layered neural network setting.

7. APPENDIX: PROOF OF LEMMA 4.1

Note that for calculating the derivative of a function of one variable it suffices to expand this function in this variable and discard terms of order greater than one.

Let $B^{(\nu)} \in U^{(\nu)}$. Then $\|B^{(\nu)}\| < \epsilon < 1$, so the matrices $I+B^{(\nu)}$ are invertible and $(I+B^{(\nu)})^{-1} = I-B^{(\nu)}+O(\epsilon^2)$ and $\det(I+B^{(\nu)})^{-1} = 1-\operatorname{tr}(B^{(\nu)})+O(\epsilon^2)$ holds, where we write $O(\epsilon^2)$ for $O(\min(\|B^{(1)}\|, \|B^{(2)}\|)^2)$. Expand the activation function linearly $g^{(\nu)}(x+h) = g^{(\nu)}(x)+Dg^{(\nu)}h+O(\|h\|^2)$ and approximate $Z^{(2)}$:

$$\begin{aligned}
&Z^{(2)} = g^{(2)} \circ W^{(2)} \circ g^{(1)} \circ (I + B^{(1)}) \circ g^{(1)-1} \\
&\circ A^{(2)} \circ g^{(2)-1} \circ S \\
&= g^{(2)} \circ W^{(2)} \big(g^{(1)}(g^{(1)-1} \circ A^{(2)}) \\
&+ Dg^{(1)}(g^{(1)-1} \circ A^{(2)} \circ g^{(2)-1} \circ S) \circ B^{(1)} \circ g^{(1)-1} \\
&\circ A^{(2)} + O(\epsilon^2)\big) \circ g^{(2)-1} \circ S \\
&= g^{(2)} \circ \big(W^{(2)}A^{(2)} + W^{(2)} \circ D^{(1)} \circ B^{(1)} \circ g^{(1)-1} \\
&\circ A^{(2)} + O(\epsilon^2)\big) \circ g^{(2)-1} \circ S
\end{aligned}$$

Here, $D^{(1)} := Dg^{(1)}(g^{(1)-1} \circ A^{(2)} \circ g^{(2)-1} \circ S)$ is a $N \times N$-matrix in diagonal form, which does not depend on $W^{(\nu)}$. The inverse of $D^{(1)}$ can then be calculated as follows:

$$D^{(1)-1} = Dg^{(1)-1}(A^{(2)} \circ g^{(2)-1} \circ S)$$

Furthermore

$$\begin{aligned}
&Z^{(2)} = g^{(2)} \circ \big(I + B^{(2)} + (I + B^{(2)}) \circ A^{(2)-1} \\
&\circ D^{(1)} \circ B^{(1)} \circ g^{(1)-1} \circ A^{(2)} + O(\epsilon^2)\big) \circ g^{(2)-1} \circ S \\
&= g^{(2)} \circ \big(I + B^{(2)} + A^{(2)-1} \circ D^{(1)} \circ B^{(1)} \circ g^{(1)-1} \\
&\circ A^{(2)} + O(\epsilon^2)\big) \circ g^{(2)-1} \circ S
\end{aligned}$$

Hence we are left with the interior term $g^{(1)-1} \circ A^{(2)}$. For this use $f(x) = Df(x)x + O(x^2)$ in a neighbourhood of 0. In our case $g^{(1)-1} \circ A^{(2)} \circ g^{(2)-1} \circ S = D^{(1)-1} \circ A^{(2)} \circ g^{(2)-1} \circ S + O(\epsilon^2)$ and therefore

$$\begin{aligned}
Z^{(2)} &= g^{(2)} \circ \big(I + B^{(2)} + Q \circ B^{(1)} \circ Q^{-1} \\
&\quad + O(\epsilon^2)\big) \circ g^{(2)-1} \circ S,
\end{aligned}$$

where $Q := (q_{ij}) := A^{(2)-1} \circ D^{(1)} \in \mathrm{Gl}(N)$. Denote the components of the inverse of Q by $(q'_{i,j}) := Q^{-1}$.

Now note that we can assume the activation functions at the network output to be the identity, because only independence of S and entropy maximization of $Z^{(2)}$ will be used, which both are invariant under *componentwise orientation-preserving* transformations of the $g_i^{(2)}$. So let $Z^{(2)} = Y^{(2)}$, and we have shown $Z^{(2)} = J \circ S$ with $J := I + B^{(2)} + Q \circ B^{(1)} \circ Q^{-1} + O(\epsilon^2)$. The inverse of J is given to first approximation by $J^{-1} = I - B^{(2)} - Q \circ B^{(1)} \circ Q^{-1} + O(\epsilon^2)$ and the determinant satisfies $\det J^{-1} = 1 - \operatorname{tr} B^{(2)} - \operatorname{tr} B^{(1)} + O(\epsilon^2)$. Applying the transformation theorem for densities to

$\rho_Z^{(2)}$, we calculate at $z \in \mathbb{R}^N$ as approximation of $B^{(\nu)}$:

$$\begin{aligned}
&\rho_Z^{(2)}(z) = (\det J^{-1})\,\rho_S(J^{-1}z) \\
&= (\det J^{-1})\,\rho_S(z - B^{(2)}z - Q \circ B^{(1)} \circ Q^{-1}z + O(\epsilon^2)) \\
&= (\det J^{-1}) \prod_{i=1}^N \rho_{S,i}\,(z_i - \\
&\qquad \Big(\textstyle\sum_j B_{i,j}^{(2)} z_j + \sum_{mjn} q_{i,m} B_{m,n}^{(1)} q'_{n,j} z_j\Big)\Big) + O(\epsilon^2) \\
&= (1 - \operatorname{tr} B^{(1)} - \operatorname{tr} B^{(2)})\,(\rho_S(z) - \\
&\qquad \textstyle\sum_{i=1}^N \rho'_{S,i}(z_i) \prod_{k\neq i} \rho_{S,k}(z_k) \\
&\qquad \Big(\textstyle\sum_j B_{i,j}^{(2)} z_j + \sum_{mjn} q_{i,m} B_{m,n}^{(1)} q'_{n,j} z_j\Big)\Big) + O(\epsilon^2) \\
&= (1 - \operatorname{tr} B^{(1)} - \operatorname{tr} B^{(2)}) \Big(1 - \textstyle\sum_{i=1}^N l_i(z_i) \\
&\qquad \Big(\textstyle\sum_j B_{i,j}^{(2)} z_j + \sum_{mjn} q_{i,m} B_{m,n}^{(1)} q'_{n,j} z_j\Big)\Big)\,\rho_S(z) \\
&\qquad + O(\epsilon^2) \\
&= \rho_S(z) - \Big(\operatorname{tr} B^{(1)} + \operatorname{tr} B^{(2)} + \textstyle\sum_{i,j=1}^N l_i(z_i) B_{i,j}^{(2)} z_j \\
&\qquad + \textstyle\sum_{i,j,m,n=1}^N l_i(z_i) q_{i,m} B_{m,n}^{(1)} q'_{n,j} z_j\Big)\,\rho_S(z) + O(\epsilon^2)
\end{aligned}$$

Here, we have defined $l_i(z_i) := \frac{\rho'_{S,i}(z_i)}{\rho_{S,i}(z_i)}$. In the second last line, the $\mathcal{C}^2$-function $\rho_S = \prod_i \rho_{S,i}$ has been expanded linearly at z using Taylor. We get a formula for $C(W^{(1)}, W^{(2)})$ at $B^{(\nu)} = 0$, which is linear in $B^{(\nu)}$:

$$\begin{aligned}
C = &\textstyle\sum_{k=1}^N \int_{\mathbb{R}^N} \big(1 - \operatorname{tr} B^{(1)} - \operatorname{tr} B^{(2)} \\
&- \textstyle\sum_{i,j=1}^N l_i(z_i) B_{i,j}^{(2)} z_j - \sum_{m,n,i,j=1}^N l_m(z_m) q_{m,i} \\
&B_{i,j}^{(1)} q'_{j,n} z_n\big)\ \ \rho_S(z) \log \rho_{S,k}(z_k) dz + O(\epsilon^2)
\end{aligned}$$

Using this, we calculate the derivative of C at $B^{(\nu)} = 0$: The case $\nu = 2$ is done similiar to the linear case: Let $i \neq j$. Then

$$\begin{aligned}
&\tfrac{\partial C}{\partial B_{ij}^{(2)}} = -\textstyle\sum_{k=1}^N \int_{\mathbb{R}^N} l_i(z_i) z_j \rho_S(z) \log \rho_{S,k}(z_k) dz \\
&= -\textstyle\sum_{k\neq i} \int_{\mathbb{R}^N} l_i(z_i) z_j \rho_S(z) \log \rho_{S,k}(z_k) dz \\
&\qquad - \textstyle\int_{\mathbb{R}^N} l_i(z_i) z_j \rho_S(z) \log \rho_{S,i}(z_i) dz \\
&= -\textstyle\sum_{k\neq i} \big(\int_{\mathbb{R}} \rho'_{S,i}(z_i) dz_i\big) \big(\int_{\mathbb{R}^{N-1}} z_j \rho_{S,j}(z_j) \\
&\qquad \log \rho_{S,k}(z_k) dz_1 .. dz_{i-1} dz_{i+1} .. dz_N\big) \\
&\qquad - \big(\textstyle\int_{\mathbb{R}} z_j \rho_{S,j}(z_j) dz_j\big) \big(\int_{\mathbb{R}} \rho'_{S,i}(z_i) \log \rho_{S,i}(z_i) dz_i\big) \\
&= -\big(\textstyle\int_{\mathbb{R}} z_j \rho_{S,j}(z_j) dz_j\big) \big(\int_{\mathbb{R}} \rho'_{S,i}(z_i) \log \rho_{S,i}(z_i) dz_i\big)
\end{aligned}$$

The last line follows, because due to $\int_{\mathbb{R}} \rho_{S,i}(z_i) = 1$ the integral over its derivative $\int_{\mathbb{R}} \rho'_{S,i}(z_i) = 0$ vanishes since $\lim_{z_i \to \infty} \int_{-\infty}^{z_i} \rho'_{S,i}(t) dt = \lim_{z_i \to \infty} \rho_{S,i}(z_i) = 0$. For $i = j$, we calculate

$$\begin{aligned}
&\tfrac{\partial C}{\partial B_{ii}^{(2)}} = -\textstyle\sum_{k=1}^N \int_{\mathbb{R}^N} (1 + l_i(z_i) z_j) \rho_S(z) \log \rho_{S,k}(z_k) dz \\
&= \textstyle\sum_{k\neq i} \big(H(S_k) + \int_{\mathbb{R}} \rho'_{S,i}(z_i) z_i dz_i H(S_k)\big) + H(S_i) \\
&\qquad - \textstyle\int_{\mathbb{R}} \rho'_{S,i}(z_i) z_i \log \rho_{S,i}(z_i) dz_i \\
&= 0 + H(S_i) + \textstyle\int_{\mathbb{R}} \rho_{S,i}(z_i) \Big(\log \rho_{S,i}(z_i) + z_i \frac{\rho'_{S,i}(z_i)}{\rho_{S,i}(z_i)}\Big) dz_i \\
&= H(S_i) - H(S_i) + \textstyle\int_{\mathbb{R}} z_i \rho'_{S,i}(z_i) dz_i \\
&= -\textstyle\int_{\mathbb{R}} \rho_{S,i}(z_i) dz_i = -1.
\end{aligned}$$

For $B^{(1)}$, a little more sophisticated calculation is needed.We have

$$\begin{aligned}
&\tfrac{\partial C}{\partial B_{ij}^{(1)}} = \delta_{ij} \textstyle\sum_k H(S_k) - \sum_{k,m,n=1}^N \int_{\mathbb{R}^N} l_m(z_m) q_{m,i} q'_{j,n} \\
&\qquad z_n \rho_S(z) \log \rho_{S,k}(z_k) dz \\
&= \delta_{ij} \textstyle\sum_k H(S_k) + \sum_{k,m,n=1}^N T_{k,m,n}
\end{aligned}$$

where we define

$$T_{k,m,n} := -\int_{\mathbb{R}^N} l_m(z_m) q_{m,i} q'_{j,n} z_n \rho_S(z) \log \rho_{S,k}(z_k) dz.$$

The sum then is decomposed into the following five terms:

$$\begin{aligned}
\tfrac{\partial C}{\partial B_{ij}^{(1)}} = &\textstyle\sum_k \sum_{\substack{m \\ m\neq k}} \sum_{\substack{n \\ n\neq k, n\neq m}} T_{k,m,n} \\
&+ \textstyle\sum_k \sum_{\substack{m \\ m\neq k}} T_{k,m,k} + \sum_k \sum_{\substack{m \\ m\neq k}} T_{k,m,m} \\
&+ \textstyle\sum_k \sum_{\substack{n \\ n\neq k}} T_{k,k,n} + \sum_k T_{k,k,k}
\end{aligned}$$

The first two summands both vanish, because one can factor out $\int \rho'_{S,m}(z_m) dz_m$, and this term is zero as we have seen above. The fourth summand has a factor $E(S_n)$:

$$\begin{aligned}
\textstyle\sum_k \sum_{\substack{n \\ n\neq k}} T_{k,k,n} \quad = \quad &\textstyle\sum_k \sum_{\substack{n \\ n\neq k}} E(S_n) q_{k,i} q'_{j,n} H(S_k) \\
&\textstyle\int_{\mathbb{R}^{N-1}} l_m(z_m) \rho_S(z_m) dz_m
\end{aligned}$$

According to the assumptions $E(S) = 0$, so this summand vanishes as well. Using partial integration, the third summand of C can be simplified as follows

$$\begin{aligned}
&\textstyle\sum_k \sum_{\substack{m \\ m\neq k}} T_{k,m,m} = \\
&= -\textstyle\sum_{\substack{k,m \\ m\neq k}} \int_{\mathbb{R}^N} l_m(z_m) q_{m,i} q'_{j,m} z_m \rho_S(z) \log \rho_{S,k}(z_k) dz \\
&= \textstyle\sum_{\substack{k,m \\ m\neq k}} q_{m,i} q'_{j,m} \big(\int_{\mathbb{R}} z_m \rho'_{S,m}(z_m) dz_m\big)\, H(S_k) \\
&= -\textstyle\sum_k \Big(\sum_{\substack{m \\ m\neq k}} q_{m,i} q'_{j,m}\Big)\, H(S_k)
\end{aligned}$$

and for the last summand we calculate with the same integration as in the $\nu = 2$ case:

$$\begin{aligned}
\textstyle\sum_k T_{k,k,k} \quad &= \quad -\textstyle\sum_k q_{k,i} q'_{j,k} \int_{\mathbb{R}} \rho'_{S,k}(z_k) z_k \rho_{S,k}(z_k) dz_k \\
&= \quad -\textstyle\sum_k q_{k,i} q'_{j,k} (H(S_k) + 1)
\end{aligned}$$

Putting together those last two equations, we therefore get for the derivative of C:

$$\begin{aligned}
\tfrac{\partial C}{\partial B_{ij}^{(1)}} \quad &= \quad \delta_{ij} \textstyle\sum_k H(S_k) \\
&\qquad - \textstyle\sum_k \Big(\sum_{\substack{m \\ m\neq k}} q_{m,i} q'_{j,m}\Big)\, H(S_k) \\
&\qquad - \textstyle\sum_k q_{k,i} q'_{j,k} (H(S_k) + 1) \\
&= \quad \delta_{ij} \textstyle\sum_k H(S_k) - \sum_k \big(\sum_m q'_{j,m} q_{m,i}\big)\, H(S_k) \\
&\qquad - \textstyle\sum_k q_{k,i} q'_{j,k} \\
&= \quad \delta_{ij} \textstyle\sum_k H(S_k) - \sum_k \delta_{ij} H(S_k) - \delta_{ij}
\end{aligned}$$

In the last line we used $Q^{-1}Q = I$ which means for $i \neq j$:

$$\sum_m q_{m,i} q'_{j,m} = \sum_m q'_{j,m} q_{m,i} = (Q^{-1}Q)_{j,i} = I_{j,i} = 0$$

This proves the lemma.

8. ACKNOWLEDGEMENTS

We want to thank Christian Jutten for his helpful comments.

9. REFERENCES

[1] H. H. Yang and S. Amari, "Adaptive on-line learning algorithms for blind separation – maximum entropy and minimum mutual information," *Neural Computation*, vol. 9, pp. 1457–1482, 1997.

[2] C. Jutten and J. Herault, "Blind separation of sources, part i: An adaptive algorithm based on neuromimetic architecture," *Signal Processing*, vol. 24, pp. 1–10, 1991.

[3] P. Comon, "Independent component analysis - a new concept?," *Signal Processing*, vol. 36, pp. 287–314, 1994.

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

[5] R. Linsker, "An application of the principle of maximum information preservation to linear systems," *Advances in Neural Information Processing Systems*, vol. 1, 1989.

[6] R. Linsker, "Local synaptic learning rules suffice to maximize mutual information in a linear network," *Neural Computation*, vol. 4, pp. 691–702, 1992.

[7] T.-W. Lee, "Independent component analysis, theory and applications," *Kluwer, Academic Publishers*, 1998.

[8] A. Hyvärinen, J. Karhunen, and E. Oja, "Independent component analysis," *John Wiley & Sons*, 2001.

[9] J.-P. Nadal and N. Parga, "Nonlinear neurons in the low noise limit: a factorial code maximizes information transfer," *Network*, vol. 5, pp. 561–581, 1994.

[10] A. Taleb and C. Jutten, "Source separation in post non linear mixtures," *IEEE Trans. on Signal Processing*, vol. 47, pp. 2807–2820, 1999.

[11] T. Lee, "Nonlinear approaches to independent component analysis," *Proceedings of the American Institute of Physics*, 1999.

[12] T. Lee, B. Koehler, and R. Orglmeister, "Blind separation of nonlinear mixing models," *IEEE NNSP*, pp. 406–415, 1997.

[13] G. C. Marques and L. B. Almeida, "Separation of nonlinear mixtures using pattern repulsion," *J. F. Cardoso, Ch. Jutten, Th. Loubaton EDS. ICA '99*, pp. 277–283.

[14] H. H. Yang, S. Amari, and A. Cichocki, "Information-theoretic approach to blind separation of sources in non-linear mixture," *Signal Processing*, vol. 64, pp. 291–300, 1998.

[15] A. Hyvärinen and P. Pajunen, "On existence and uniqueness of solutions in nonlinear independent component analysis," *Proceedings of the 1998 IEEE International Joint Conference on Neural Networks (IJCNN'98)*, vol. 2, pp. 1350–1355, 1998.

[16] M. Anthony and P. L. Bartlett, "Neural network learning: Theoretical foundations," *Cambridge University Press*.

[17] S. Haykin, "Neural networks," *Macmillan College Publishing Company*, 1994.

[18] K. Hornik, M. Stinchcombe, and H. White, "Multilayer feedforward networks are universal approximators," *Neural Networks*, vol. 2, pp. 359–366, 1989.