ICASSP 2001, May 7-11, Salt Lake City, UT, USA, Vol. 5, pp. 2793–2796

ELEMENTARY COST FUNCTIONS FOR BLIND SEPARATION OF NON-STATIONARY SOURCE SIGNALS

Marcel Joho[1], *Russell H. Lambert*[2], *and Heinz Mathis*[1]

[1]Signal and Information Processing Laboratory,
Swiss Federal Institute of Technology Zurich, Switzerland
[2]Broadcom, 70 S. Lake Ave., Suite 900, Pasadena, CA 91101, USA

joho@isi.ee.ethz.ch, *russ@broadcom.com*, *mathis@isi.ee.ethz.ch*

ABSTRACT

Blind source separation (BSS) is a problem found in many applications related to acoustics or communications. This paper addresses the blind source separation problem for the case where the source signals are non-stationary and the sensors are noisy. To this end, we propose several useful elementary cost functions which can be combined to an overall cost function. The elementary cost functions might have different objectives, such as uncorrelated output signals or power normalization of the output signals. Additionally, the corresponding gradients with respect to the adjustable parameters are given. We discuss the design of an overall cost function and also give a simulation example.

1. INTRODUCTION

1.1. Problem description

The general $M \times M$ mixing process is shown in Fig. 1 and described as

$$\mathbf{x}_t = \mathbf{A}\mathbf{s}_t + \mathbf{n}_t \tag{1}$$

where $\mathbf{s}_t = (s_1, \ldots, s_M)_t^T$, $\mathbf{x}_t = (x_1, \ldots, x_M)_t^T$, and $\mathbf{n}_t$ contain the samples of the unknown source signals, the sensor signals, and the sensor noise at time t, respectively, and $\mathbf{A}^{M\times M}$ is the unknown mixing matrix. The blind source separation problem is defined as finding a separation matrix $\mathbf{W}^{M\times M}$ such that the output of the separation process

$$\mathbf{u}_t = \mathbf{W}\mathbf{x}_t = \mathbf{W}(\mathbf{A}\mathbf{s}_t + \mathbf{n}_t) = \mathbf{G}\mathbf{s}_t + \mathbf{W}\mathbf{n}_t \tag{2}$$

is a vector of waveform-preserving estimates of the unknown source signals by using only the time series of the measured sensor signals $\{\mathbf{x}_t\}$ for $t = 1, 2, \ldots$. $\mathbf{G}$ is the total transfer matrix of the global system.

In the following, our main objective is to find a so-called *zero-forcing* solution for $\mathbf{W}$ such that $\mathbf{G}$ becomes close to a scaled permutation matrix. This is equivalent to minimizing the *inter-channel interference* (ICI) at the output $\mathbf{u}$, regardless of a possible noise amplification by $\mathbf{W}$. In fact, there are different statistical criteria which can be exploited for blind signal separation, e.g., non-Gaussianity, non-whiteness, cyclo-stationarity, and non-stationarity of the source signals. In the following, we assume non-stationary source signals.

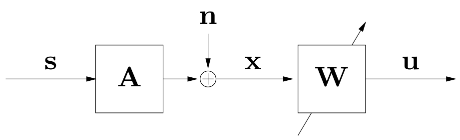

Fig. 1. Blind source separation setup with sensor noise.

1.2. Notation

The notation used throughout this paper is the following: Vectors are written in lower case, matrices in upper case. Matrix and vector transpose, complex conjugation and Hermitian transpose are denoted by $(.)^T$, $(.)^*$, and $(.)^H \triangleq ((.)^*)^T$, respectively. The sample index is denoted by t. The identity matrix is denoted by $\mathbf{I}$, a vector or a matrix containing only zeros by $\mathbf{0}$. $E\{.\}$ denotes the expectation operator. Vector or matrix dimensions are given in superscript. The Frobenius norm and the trace of a matrix are denoted by $\|\,.\,\|_F$ and $\mathrm{tr}\,(.)$, respectively. $\mathbf{a} = \mathrm{diag}(\,\mathbf{A}\,)$ is a vector whose elements are the diagonal elements of $\mathbf{A}$ and $\mathrm{diag}(\,\mathbf{a}\,)$ is a square diagonal matrix which contains the elements of $\mathbf{a}$. $\mathrm{ddiag}(\,\mathbf{A}\,)$ zeros the off-diagonal elements of $\mathbf{A}$ and

$$\mathrm{off}(\,\mathbf{A}\,) \triangleq \mathbf{A} - \mathrm{ddiag}(\,\mathbf{A}\,) \tag{3}$$

zeros the diagonal elements of $\mathbf{A}$. For a square matrix $\mathbf{A}$ we have $\mathrm{ddiag}(\,\mathbf{A}\,) = \mathrm{diag}(\,\mathrm{diag}(\,\mathbf{A}\,)\,)$.

1.3. Assumptions

In addition to the problem proposed above, we make the following assumptions:

- $\mathcal{A}1$ Time-invariant mixing matrix $\mathbf{A}$.
- $\mathcal{A}2$ $\mathbf{A}$ has full rank M.
- $\mathcal{A}3$ Source signals s_m, $m = 1, \ldots, M$, are mutually independent.
- $\mathcal{A}4$ All source signals s_m but possibly one are non-stationary.
- $\mathcal{A}5$ All source signals are unknown
- $\mathcal{A}6$ The sensor noise signals are stationary additive white Gaussian processes and mutually independent.
- $\mathcal{A}7$ The source signals and the sensor noise are mutually independent.

As a consequence, $\mathcal{A}3$ and $\mathcal{A}6$ imply

$$\mathbf{R}_{\mathbf{ss}_t} \triangleq E\{\mathbf{s}_t\mathbf{s}_t^H\} = \mathrm{diag}(\,\sigma_{\mathrm{s}1,t}^2, \ldots, \sigma_{\mathrm{s}M,t}^2\,) \tag{4}$$

$$\mathbf{R}_{\mathbf{nn}} \triangleq E\{\mathbf{n}_t\mathbf{n}_t^H\} = \mathrm{diag}(\,\sigma_{\mathrm{n}1}^2, \ldots, \sigma_{\mathrm{n}M}^2\,)\,. \tag{5}$$

and from (1) and $\mathcal{A}7$ we have

$$\mathbf{R}_{\mathbf{xx}_t} \triangleq E\{\mathbf{x}_t \mathbf{x}_t^H\} = \mathbf{A}\mathbf{R}_{\mathbf{ss}_t}\mathbf{A}^H + \mathbf{R}_{\mathbf{nn}} \,. \tag{6}$$

2. OVERALL COST FUNCTION

In order to successfully separate the unknown source signals, we define an *overall cost function* $\mathcal{J}$ which consists of a weighted sum of *elementary cost functions*.

$$\mathcal{J}(\mathbf{W}, \mathbf{R}_{\mathbf{nn}}) = \sum_i \alpha_i \mathcal{J}_i(\mathbf{W}, \mathbf{R}_{\mathbf{nn}}) \,. \tag{7}$$

The elementary cost functions $\mathcal{J}_i$ can have different objectives, e.g., decorrelation of the output signals, normalization of the separation matrix $\mathbf{W}$, output power control, etc.

After choosing a suitable cost function $\mathcal{J}$, we use a stochastic gradient algorithm to find the unknown parameters which minimize $\mathcal{J}$

$$\mathbf{W}_{k+1} = \mathbf{W}_k + \Delta\mathbf{W}_k \tag{8}$$

$$\Delta\mathbf{W}_k = -\mu \, \nabla_{\mathbf{W}} \left\{ \mathcal{J}(\mathbf{W}_k, \mathbf{R}_{\mathbf{nn}}) \right\} \tag{9}$$

where $\Delta\mathbf{W}_k$ is the incremental update of $\mathbf{W}_k$, and

$$\nabla_{\mathbf{W}} \left\{ \mathcal{J}(\mathbf{W}_k, \mathbf{R}_{\mathbf{nn}}) \right\} = \sum_i \alpha_i \nabla_{\mathbf{W}} \left\{ \mathcal{J}_i(\mathbf{W}_k, \mathbf{R}_{\mathbf{nn}}) \right\} \tag{10}$$

is the gradient of the overall cost function $\mathcal{J}$ with respect to $\mathbf{W}$.

In the case where $\mathbf{R}_{\mathbf{nn}}$ is unknown, we can also use a stochastic gradient algorithm to find an estimate $\hat{\mathbf{R}}_{\mathbf{nn}}$

$$\hat{\mathbf{R}}_{\mathbf{nn}_{k+1}} = \hat{\mathbf{R}}_{\mathbf{nn}_k} + \Delta\hat{\mathbf{R}}_{\mathbf{nn}_k} \tag{11}$$

$$\Delta\hat{\mathbf{R}}_{\mathbf{nn}_k} = -\eta \, \nabla_{\mathbf{R}_{\mathbf{nn}}} \left\{ \mathcal{J}(\mathbf{W}_k, \hat{\mathbf{R}}_{\mathbf{nn}_k}) \right\} \tag{12}$$

where

$$\nabla_{\mathbf{R}_{\mathbf{nn}}} \left\{ \mathcal{J}(\mathbf{W}_k, \hat{\mathbf{R}}_{\mathbf{nn}_k}) \right\} = \sum_i \alpha_i \nabla_{\mathbf{R}_{\mathbf{nn}}} \left\{ \mathcal{J}_i(\mathbf{W}_k, \hat{\mathbf{R}}_{\mathbf{nn}_k}) \right\} \,. \tag{13}$$

3. ELEMENTARY COST FUNCTIONS

In this section, we present some elementary cost functions which are effective for blind signal separation of non-stationary source signals. A summary of elementary cost functions and their corresponding gradients is given in Table 1. Appendix A lists all equations used for the derivation of the gradients.

3.1. Decorrelation of output signals

An elementary cost function for blind signal separation of non-stationary source signals can be a cost function which penalizes uncorrelated output signals. While decorrelation of the output signals is a necessary but not sufficient criterion for the separation of stationary source signals, the decrorrelation criterion can be sufficient for non-stationary source signals under some weak conditions. We define an elementary cost function

$$\mathcal{J}_1 \triangleq \left\| \operatorname{off}\Big(\mathbf{W} \left(\mathbf{R}_{\mathbf{xx}} - \mathbf{R}_{\mathbf{nn}} \right) \mathbf{W}^H \Big) \right\|_F^2 \tag{14}$$

which measures the deviation from having uncorrelated output signals in the noise-free case. Since we want to adapt $\mathbf{W}$ by a stochastic-gradient learning algorithm, we need the gradient of the cost function with respect to $\mathbf{W}$

$$\nabla_{\mathbf{W}} \mathcal{J}_1 = 4 \operatorname{off}\Big(\mathbf{W} \left(\mathbf{R}_{\mathbf{xx}} - \mathbf{R}_{\mathbf{nn}} \right) \mathbf{W}^H \Big) \mathbf{W} \left(\mathbf{R}_{\mathbf{xx}} - \mathbf{R}_{\mathbf{nn}} \right) \tag{15}$$

which we can use in (9) to update $\mathbf{W}$. With (14) we obtain a bias-free separation matrix $\mathbf{W}$ after convergence. This goes also in line with the bias-removal technique proposed by Douglas *et al.* in [1]. However, if $\mathbf{R}_{\mathbf{nn}}$ is unknown, by simply setting $\mathbf{R}_{\mathbf{nn}} = \mathbf{0}$ we usually obtain a biased separation matrix in the noisy case, except for some special cases, e.g., if $\mathbf{R}_{\mathbf{nn}}$ is a scalar matrix and $\mathbf{A}$ is a unitary matrix. For this reason, if $\mathbf{R}_{\mathbf{nn}}$ is unknown we also use a stochastic gradient method to estimate $\hat{\mathbf{R}}_{\mathbf{nn}}$

$$\hat{\mathbf{R}}_{\mathbf{nn}_{k+1}} = \hat{\mathbf{R}}_{\mathbf{nn}_k} + \Delta\hat{\mathbf{R}}_{\mathbf{nn}_k} \tag{16}$$

$$\Delta\hat{\mathbf{R}}_{\mathbf{nn}_k} = -\eta \, \nabla_{\mathbf{R}_{\mathbf{nn}}} \left\{ \mathcal{J}_1(\mathbf{W}_k, \hat{\mathbf{R}}_{\mathbf{nn}_k}) \right\} \,. \tag{17}$$

We restrict ourself to adapt a diagonal matrix $\hat{\mathbf{R}}_{\mathbf{nn}}$, since we assume from $\mathcal{A}6$ that the sensor noise is mutually uncorrelated. Therefore,

$$\nabla_{\mathbf{R}_{\mathbf{nn}}} \mathcal{J}_1 = -4 \operatorname{ddiag}\Big(\mathbf{W}^H \operatorname{off}\Big(\mathbf{W} \left(\mathbf{R}_{\mathbf{xx}} - \hat{\mathbf{R}}_{\mathbf{nn}} \right) \mathbf{W}^H \Big) \mathbf{W} \Big) \tag{18}$$

where we choose the initial value $\hat{\mathbf{R}}_{\mathbf{nn}_0}$ to be a diagonal matrix. Moreover, if we know that each sensor has the same noise characteristics, i.e. $\sigma_{\mathrm{n}m}^2 = \sigma_{\mathrm{n}}^2$, we only have to adapt

$$\hat{\sigma}_{\mathrm{n}k+1}^2 = \hat{\sigma}_{\mathrm{n}k}^2 + \Delta\hat{\sigma}_{\mathrm{n}k}^2 \tag{19}$$

$$\Delta\hat{\sigma}_{\mathrm{n}k}^2 = -\eta \, \nabla_{\sigma_{\mathrm{n}k}^2} \left\{ \mathcal{J}_1(\mathbf{W}_k, \hat{\sigma}_{\mathrm{n}k}^2) \right\} \tag{20}$$

with

$$\nabla_{\sigma_{\mathrm{n}}^2} \mathcal{J}_1 = -\frac{4}{M} \operatorname{tr}\Big(\mathbf{W}^H \operatorname{off}\Big(\mathbf{W} \left(\mathbf{R}_{\mathbf{xx}} - \hat{\sigma}_{\mathrm{n}k}^2 \mathbf{I} \right) \mathbf{W}^H \Big) \mathbf{W} \Big) \tag{21}$$

where $\hat{\mathbf{R}}_{\mathbf{nn}} = \hat{\sigma}_{\mathrm{n}k}^2 \mathbf{I}$ is the current estimate of $\mathbf{R}_{\mathbf{nn}}$.

3.2. Constraints on the separation matrix W

Since $\mathbf{W} = \mathbf{0}$ also minimizes the elementary cost function in (14), we need an additional constraint which prevents this trivial solution. The elementary cost function

$$\mathcal{J}_3 \triangleq \left\| \operatorname{ddiag}\Big(\mathbf{W}\mathbf{W}^H - \mathbf{I} \Big) \right\|_F^2 \tag{22}$$

has its minimum when the row vectors of $\mathbf{W}$ are normalized to have length one. An alternative elementary cost function is

$$\mathcal{J}_6 \triangleq \| \operatorname{ddiag}(\mathbf{W} - \mathbf{I}) \|_F^2 \tag{23}$$

which, if included in the overall cost function $\mathcal{J}$, steers the diagonal elements of $\mathbf{W}$ towards $+1$.

3.3. Output power normalization

An alternative to directly constrain $\mathbf{W}$, is to use a constraint on the average output power. The elementary cost function

$$\mathcal{J}_2 \triangleq \left\| \operatorname{ddiag}\Big(\mathbf{W}\mathbf{R}_{\mathbf{xx}}\mathbf{W}^H - \mathbf{I} \Big) \right\|_F^2 \tag{24}$$

steers the average long-term output power of the output signals $\mathbf{u}$ to become one.

3.4. Relationship to other algorithms

An adaptive algorithm using the cost function $\mathcal{J} = \mathcal{J}_1 + \alpha_2 \mathcal{J}_2$ with $\mathbf{R_{nn}} = \mathbf{0}$ was recently published by Jones [2]. Another algorithm was proposed by Parra and Spence [3] which uses a cost function similar to $\mathcal{J}_1$, except that the assumption having stationary noise signals is dropped and the powers of the output signals u_m are estimated too. In addition, an extension is given for the case where the source signals are convolutively mixed. Related work was also done by Matsuoka *et al.* in [4].

4. SOURCE SEPARATION VIA EIGENVALUE DECOMPOSITION

We now present an alternative method, which is inspired from a Linear Algebra problem, namely the simultaneous diagonalization of two matrices [5], which is only possible if the two matrices commute. In our case, the two matrices we want to diagonalize are $(\mathbf{R_{xx}}_{t_1} - \mathbf{R_{nn}})$ and $(\mathbf{R_{xx}}_{t_2} - \mathbf{R_{nn}})$, where t_1 and t_2 are two different time instants. First, let us assume that $\mathbf{R_{xx}}_{t_1}$, $\mathbf{R_{xx}}_{t_2}$, and $\mathbf{R_{nn}}$ are known. Then we can define a matrix $\mathbf{Q}$ as

$$\mathbf{Q} \triangleq \left(\mathbf{R_{xx}}_{t_1} - \mathbf{R_{nn}}\right)^{-1} \left(\mathbf{R_{xx}}_{t_2} - \mathbf{R_{nn}}\right) \tag{25}$$

$$= [\mathbf{A}\mathbf{R_{ss}}_{t_1}\mathbf{A}^H]^{-1}\mathbf{A}\mathbf{R_{ss}}_{t_2}\mathbf{A}^H \tag{26}$$

$$= \mathbf{A}^{-H}\mathbf{R}_{\mathbf{ss}_{t_1}}^{-1}\mathbf{R_{ss}}_{t_2}\mathbf{A}^H . \tag{27}$$

Since the source signals s_m are mutually independent by assumption, and therefore mutually uncorrelated, $\mathbf{R_{ss}}_{t_1}$ and $\mathbf{R_{ss}}_{t_2}$ are diagonal matrices, and so is $\mathbf{R}_{\mathbf{ss}_{t_1}}^{-1}\mathbf{R_{ss}}_{t_2}$. Hence, the similarity transform in (27) is just an eigenvalue decomposition (EVD) of $\mathbf{Q}$. Since an EVD is not unique, we can decompose $\mathbf{Q}$ as

$$\mathbf{Q} = \mathbf{T}\boldsymbol{\Lambda}\mathbf{T}^{-1} \tag{28}$$

where the column vectors $\mathbf{t}_m$ of $\mathbf{T}$ are the eigenvectors of $\mathbf{Q}$ and have unity length, i.e. $\|\mathbf{t}_m\|_2 = 1$. Furthermore,

$$\boldsymbol{\Lambda} = \operatorname{diag}(\lambda_1, \ldots, \lambda_M) \triangleq \mathbf{R}_{\mathbf{ss}_{t_1}}^{-1}\mathbf{R_{ss}}_{t_2} \tag{29}$$

contains the eigenvalues $\lambda_m = \sigma^2_{m,t_2}/\sigma^2_{m,t_1}$ of $\mathbf{Q}$. λ_m is the power ratio of the source signal s_m between time instant t_2 and t_1. If all λ_m are distinct, $\mathbf{W} = \mathbf{T}^H$ is a separation matrix such that $\mathbf{G}$ becomes a scaled permutation matrix. In that case, $\mathbf{W}(\mathbf{R_{xx}}_t - \mathbf{R_{nn}})\mathbf{W}^H$ is diagonal for all t. Hence, under these conditions, the blind signal separation problem for non-stationary source signals can be tracked down to solving a single eigenvalue decomposition task. Problems with this method arise if an eigenvalue λ_m appears with multiplicity greater than one, because then $\mathbf{Q}$ has an eigenspace which does not uniquely define the column vectors of $\mathbf{T}$, which are the row vectors of $\mathbf{W}$. This problem can be greatly reduced by simultaneously diagonalizing a set of K correlation matrices $\{\mathbf{R_{xx}}_{t_k} - \mathbf{R_{nn}}\}_{k=1}^{K}$.

If $\mathbf{R_{nn}}$ is unknown, it can be replaced by an estimate $\hat{\mathbf{R}}_{\mathbf{nn}}$. However, the drawback of this method is that the eigenvectors $\mathbf{t}_m$ are quite sensitive to bad estimates of $\mathbf{R_{xx}}$ and $\mathbf{R_{nn}}$. It can even cause $\mathbf{Q}$ to have negative eigenvalues λ_m, which have no physical meaning anymore. The solution of the EVD method can also be helpful to obtain a good initial value $\mathbf{W}_0$ in (8).

A similar EVD method was used by Molgedey and Schuster in [6] for the separation of stationary, but temporally correlated source signals. The simultaneous diagonalization of two correlation matrices can also be solved by a generalized eigenvalue decomposition, as pointed out by Tsatsanis and Kweon in [7]. Recently, Choi and Cichocki [8, 9] have proposed a method for blind separation of non-stationary and non-white source signals using simultaneous diagonalization of time-delayed correlation matrices $\mathbf{R_{xx}}_t(\tau) \triangleq E\{\mathbf{x}_t\mathbf{x}_{t-\tau}^H\}$ for $\tau \neq 0$.

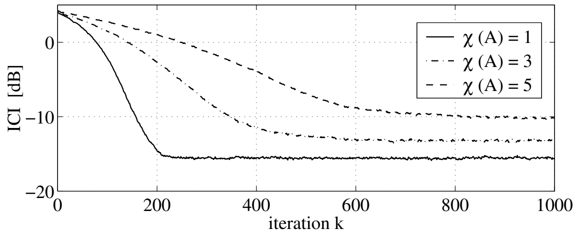

Fig. 2. Learning curve measuring the interchannel interference.

5. SIMULATION

In the following, we give a simulation example to analyze the behavior of the proposed algorithm. We measure the average channel-wise *interchannel interference* (ICI) as our performance criterion of interest. The overall cost function is a combination of $\mathcal{J}_1$ and $\mathcal{J}_3$. Hence, we adapt $\mathbf{W}_k$ with (8) and

$$\Delta\mathbf{W}_k = -\frac{0.4}{0.05 + \|\nabla_{\mathbf{W}}\mathcal{J}_1\|_F}\nabla_{\mathbf{W}}\mathcal{J}_1 - 0.1\,\nabla_{\mathbf{W}}\mathcal{J}_3 \tag{30}$$

which includes a step-size normalization. $\hat{\sigma_{\mathrm{n}}}_k^2$ is adapted with (19) to (21) and $\eta = 0.04$. The randomly chosen complex mixing matrices $\mathbf{A}$ are normalized such that their largest singular value is always 1, the condition numbers are $\chi(\mathbf{A}) = 1$, 3, and 5, and the singular values of $\mathbf{A}$ are logarithmically distributed. We have $M = 10$ stationary Gaussian-distributed source signals s', each being complex and with power $\sigma_s^2 = 1$. The non-stationarity of the source signals is introduced by a block-wise randomly chosen complex gain $K_{m,k} \in [0.1, 1]$ for every source signal, e.g. $s_{m,t} = K_{m,k} \cdot s'_{m,t}$. Hence, $\sigma^2_{s_{m,k}} \in [0.01, 1]$. Furthermore, we have $\sigma_n^2 = 0.1$. The correlation matrices $\mathbf{R_{xx}}_k$ are estimated over blocks of $L = 100$ samples. Fig. 2 shows performance curves averaged over 30 runs for different $\chi(\mathbf{A})$. We see that convergence is reached quite fast, despite the high noise level. However, the performance depends strongly on the condition number of $\mathbf{A}$.

6. SUMMARY

We have presented several elementary cost functions, which can be combined to an overall cost function for blindly separating a noisy mixture of non-stationary source signals. In addition, the gradients of the elementary cost functions are given, which can be used for an online stochastic-gradient learning algorithm for adjusting the parameters of interest. Finally, a simulation example is given.

Table 1. Elementary cost functions with corresponding gradients

objective	elementary cost function $\mathcal{J}_i\left(\mathbf{W}, \mathbf{R_{nn}}\right)$	gradient $\nabla \mathcal{J}_i\left(\mathbf{W}, \mathbf{R_{nn}}\right)$
uncorrelated output signals	$\mathcal{J}_1 \triangleq \left\| \operatorname{off}\left(\mathbf{W}\left(\mathbf{R_{xx}} - \mathbf{R_{nn}}\right)\mathbf{W}^H \right) \right\|_F^2$	$\nabla_{\mathbf{W}} \mathcal{J}_1 = 4 \operatorname{off}\left(\mathbf{W}\left(\mathbf{R_{xx}} - \mathbf{R_{nn}}\right)\mathbf{W}^H \right) \mathbf{W}\left(\mathbf{R_{xx}} - \mathbf{R_{nn}}\right)$
if $\mathbf{R_{nn}}$ is diagonal		$\nabla_{\mathbf{R_{nn}}} \mathcal{J}_1 = -4 \operatorname{ddiag}\left(\mathbf{W}^H \operatorname{off}\left(\mathbf{W}\left(\mathbf{R_{xx}} - \mathbf{R_{nn}}\right)\mathbf{W}^H \right) \mathbf{W} \right)$
if $\mathbf{R_{nn}} = \sigma_n^2 \mathbf{I}$		$\nabla_{\sigma_n^2} \mathcal{J}_1 = -\frac{4}{M} \operatorname{tr}\left(\mathbf{W}^H \operatorname{off}\left(\mathbf{W}\left(\mathbf{R_{xx}} - \sigma_n^2 \mathbf{I}\right)\mathbf{W}^H \right) \mathbf{W} \right)$
normalized output power	$\mathcal{J}_2 \triangleq \left\| \operatorname{ddiag}\left(\mathbf{W}\left(\mathbf{R_{xx}} - \mathbf{R_{nn}}\right)\mathbf{W}^H \right) - \mathbf{I} \right\|_F^2$	$\nabla_{\mathbf{W}} \mathcal{J}_2 = 4 \operatorname{ddiag}\left(\mathbf{W}\left(\mathbf{R_{xx}} - \mathbf{R_{nn}}\right)\mathbf{W}^H - \mathbf{I} \right) \mathbf{W}\left(\mathbf{R_{xx}} - \mathbf{R_{nn}}\right)$
if $\mathbf{R_{nn}}$ is diagonal		$\nabla_{\mathbf{R_{nn}}} \mathcal{J}_2 = -4 \operatorname{ddiag}\left(\mathbf{W}^H \operatorname{ddiag}\left(\mathbf{W}\left(\mathbf{R_{xx}} - \mathbf{R_{nn}}\right)\mathbf{W}^H - \mathbf{I} \right) \mathbf{W} \right)$
if $\mathbf{R_{nn}} = \sigma_n^2 \mathbf{I}$		$\nabla_{\sigma_n^2} \mathcal{J}_2 = -\frac{4}{M} \operatorname{tr}\left(\mathbf{W}^H \operatorname{ddiag}\left(\mathbf{W}\left(\mathbf{R_{xx}} - \mathbf{R_{nn}}\right)\mathbf{W}^H - \mathbf{I} \right) \mathbf{W} \right)$
row-normalized $\mathbf{W}$	$\mathcal{J}_3 \triangleq \left\| \operatorname{ddiag}\left(\mathbf{W}\mathbf{W}^H - \mathbf{I} \right) \right\|_F^2$	$\nabla_{\mathbf{W}} \mathcal{J}_3 = 4 \operatorname{ddiag}\left(\mathbf{W}\mathbf{W}^H - \mathbf{I} \right) \mathbf{W}$
column-normalized $\mathbf{W}$	$\mathcal{J}_4 \triangleq \left\| \operatorname{ddiag}\left(\mathbf{W}^H\mathbf{W} - \mathbf{I} \right) \right\|_F^2$	$\nabla_{\mathbf{W}} \mathcal{J}_4 = 4\, \mathbf{W} \operatorname{ddiag}\left(\mathbf{W}^H\mathbf{W} - \mathbf{I} \right)$
norm constraint	$\mathcal{J}_5 \triangleq \left(\| \mathbf{W} \|_F^2 - M \right)^2$	$\nabla_{\mathbf{W}} \mathcal{J}_5 = 4 \left(\| \mathbf{W} \|_F^2 - M \right) \mathbf{W}$
$\operatorname{diag}(\mathbf{W}) = \mathbf{1}$	$\mathcal{J}_6 \triangleq \| \operatorname{ddiag}(\mathbf{W} - \mathbf{I}) \|_F^2$	$\nabla_{\mathbf{W}} \mathcal{J}_6 = 2 \operatorname{ddiag}(\mathbf{W} - \mathbf{I})$
real diag. elem. of $\mathbf{W}$	$\mathcal{J}_7 \triangleq \left\| \operatorname{ddiag}\left(\mathbf{W} - \mathbf{W}^H \right) \right\|_F^2$	$\nabla_{\mathbf{W}} \mathcal{J}_7 = 4 \operatorname{ddiag}\left(\mathbf{W} - \mathbf{W}^H \right)$
row-scaled unitary $\mathbf{W}$	$\mathcal{J}_8 \triangleq \left\| \operatorname{off}\left(\mathbf{W}\mathbf{W}^H \right) \right\|_F^2$	$\nabla_{\mathbf{W}} \mathcal{J}_8 = 4 \operatorname{off}\left(\mathbf{W}\mathbf{W}^H \right) \mathbf{W}$
column-scaled unitary $\mathbf{W}$	$\mathcal{J}_9 \triangleq \left\| \operatorname{off}\left(\mathbf{W}^H\mathbf{W} \right) \right\|_F^2$	$\nabla_{\mathbf{W}} \mathcal{J}_9 = 4\, \mathbf{W} \operatorname{off}\left(\mathbf{W}^H\mathbf{W} \right)$
unitary $\mathbf{W}$	$\mathcal{J}_{10} \triangleq \left\| \mathbf{W}\mathbf{W}^H - \mathbf{I} \right\|_F^2$	$\nabla_{\mathbf{W}} \mathcal{J}_{10} = 4 \left(\mathbf{W}\mathbf{W}^H - \mathbf{I} \right) \mathbf{W} = 4\, \mathbf{W} \left(\mathbf{W}^H\mathbf{W} - \mathbf{I} \right)$
$\lvert \det(\mathbf{W}) \rvert = 1$	$\mathcal{J}_{11} \triangleq \lvert \log(\lvert \det(\mathbf{W}) \rvert) \rvert^2$	$\nabla_{\mathbf{W}} \mathcal{J}_{11} = 2 \log(\lvert \det(\mathbf{W}) \rvert) \cdot \mathbf{W}^{-H}$

A. COMPUTATION OF THE GRADIENT OF A COST FUNCTION WITH RESPECT TO A MATRIX

The following equalities are useful for the computation of Frobenius norms and trace functions [10]

$$\| \mathbf{A} \|_F^2 = \operatorname{tr}(\mathbf{A}^H \mathbf{A}) \tag{31}$$

$$\| \operatorname{off}(\mathbf{A}) \|_F^2 = \| \mathbf{A} \|_F^2 - \| \operatorname{ddiag}(\mathbf{A}) \|_F^2 \tag{32}$$

$$\operatorname{tr}(\mathbf{AB}) = \operatorname{tr}(\mathbf{BA}) \tag{33}$$

$$\operatorname{tr}(\mathbf{A} \operatorname{ddiag}(\mathbf{B})) = \operatorname{tr}(\operatorname{ddiag}(\mathbf{A})\, \mathbf{B}) \tag{34}$$

$$= \operatorname{tr}(\operatorname{ddiag}(\mathbf{A}) \operatorname{ddiag}(\mathbf{B})) \tag{35}$$

$$\operatorname{tr}(\operatorname{off}(\mathbf{A}) \operatorname{ddiag}(\mathbf{B})) = 0\,. \tag{36}$$

For the computation of the gradient of a cost function based on the Frobenius norm, we have to differentiate a trace function with respect to a complex matrix $\mathbf{W} \triangleq \left[w_{mn}^{\mathrm{re}} + j w_{mn}^{\mathrm{im}} \right]$. After Haykin [11], the complex gradient can be defined as

$$\nabla_{\mathbf{W}} = 2 \frac{\partial}{\partial \mathbf{W}^*} = \left[\frac{\partial}{\partial w_{mn}^{\mathrm{re}}} + j \frac{\partial}{\partial w_{mn}^{\mathrm{im}}} \right]. \tag{37}$$

The following equalities are useful for the differentiation of a trace function with respect to a complex matrix

$$\frac{\partial}{\partial \mathbf{W}^*} \operatorname{tr}(\mathbf{WA}) = \mathbf{0} \tag{38}$$

$$\frac{\partial}{\partial \mathbf{W}^*} \operatorname{tr}(\mathbf{W}^H \mathbf{A}) = \mathbf{A} \tag{39}$$

$$\frac{\partial}{\partial \mathbf{W}^*} \operatorname{tr}(\mathbf{W}^H \mathbf{A} \mathbf{W} \mathbf{B}) = \mathbf{AWB} \tag{40}$$

$$\frac{\partial}{\partial \mathbf{W}^*} \operatorname{tr}(\mathbf{W}^H \mathbf{A} \mathbf{W}^H \mathbf{B}) = \mathbf{A}\mathbf{W}^H\mathbf{B} + \mathbf{B}\mathbf{W}^H\mathbf{A}\,. \tag{41}$$

B. REFERENCES

[1] S. C. Douglas, A. Cichocki, and S. Amari, “Bias removal technique for blind source separation with noisy measurements,” *Electron. Lett.*, vol. 34, no. 14, pp. 1379–1380, July 1998.

[2] D. L. Jones, “A new method for blind source separation of nonstationary signals,” in *Proc. ICASSP*, Phoenix, AZ, May 15–19, 1999, vol. 5, pp. 2893–2896.

[3] L. Parra and C. Spence, “Convolutive blind separation of non-stationary sources,” *IEEE Trans. Speech and Audio Processing*, vol. 8, no. 3, pp. 320–327, 2000.

[4] K. Matsuoka, M. Ohya, and M. Kawamoto, “A neural net for blind separation of nonstationary signals,” *Neural Networks*, vol. 8, no. 3, pp. 411–419, Mar. 1995.

[5] R. A. Horn and C. R. Johnson, *Matrix Analysis*, Cambridge University Press, 1990.

[6] L. Molgedey and H. Schuster, “Separation of independent signals using time-delayed correlations,” *Phys. Rev. Lett.*, vol. 72, no. 23, pp. 3634–3637, Oct. 1994.

[7] M. K. Tsatsanis and C. Kweon, “Blind source separation of non-stationary sources using second-order statistics,” in *Proc. Asilomar Conf. Signals, Syst., Comput.*, Pacific Grove, CA, Nov. 1998, vol. II, pp. 1574–1578.

[8] S. Choi and A. Cichocki, “Blind separation of nonstationary sources in noisy mixtures,” *Electron. Lett.*, vol. 36, no. 9, pp. 848–849, 2000.

[9] S. Choi and A. Cichocki, “Blind separation of nonstationary and temporally correlated sources from noisy mixtures,” in *Proc. NNSP*, Sydney, Australia, Dec. 2000, vol. 2, pp. 405–414.

[10] M. Joho, *A Systematic Approach to Adaptive Algorithms for Multichannel System Identification, Inverse Modeling, and Blind Identification*, Ph.D. thesis, ETH Zürich, Dec. 2000.

[11] S. Haykin, *Adaptive Filter Theory*, Prentice Hall, 3rd edition, 1996.