NEWTON METHOD FOR JOINT APPROXIMATE DIAGONALIZATION OF POSITIVE DEFINITE HERMITIAN MATRICES

MARCEL JOHO*

Abstract. In this paper we present a Newton method to jointly approximately diagonalize a set of positive definite Hermitian matrices. To this end, we derive the local gradient and Hessian of the underlying cost function in closed form. The algorithm is derived for the complex case and can also update a non-square diagonalization matrix. We analyze the cost function at the critical points and show its relation to a different cost function that is commonly studied.

Key words. Joint approximate diagonalization, Independent Component Analysis (ICA), Newton method, gradient and Hessian of a complex matrix-valued cost function.

AMS subject classifications. 15A23, 49M15, 49Q12, 65F30.

1. Problem formulation and cost function. The mathematical problem of jointly approximately diagonalizing a set of P covariance matrices, $\{\mathbf{R}_p\}_{p=1}^P$, is of interest in statistical methods such as common principal component analysis [7]. Recently the joint diagonalization problem has received more attention also in the community of independent component analysis and blind signal separation, as some problems that occur in those fields can be formulated as a joint-diagonalization optimization problem. As an example, we can describe one class of blind signal separation problems as follows:

Let $\mathbf{s}(t)$ be an M-dimensional vector containing the time series of M mutually independent *source signals* that are mixed by a *mixing matrix* $\mathbf{A}^{N \times M}$, such that $\mathbf{x}(t) = \mathbf{A}\,\mathbf{s}(t)$ is an N-dimensional vector containing the time-series of N *sensor signals*. In the blind signal separation problem the assumption is made that $\mathbf{s}(t)$ and $\mathbf{A}$ are unknown, only the signals at the sensors, $\mathbf{x}(t)$, are known. Furthermore, the assumption is made that the source signals are non-stationary and mutually independent. Hence, $\mathbf{R}_{\mathbf{ss}}(t) \triangleq E\{\mathbf{s}(t)\mathbf{s}^H(t)\}$ depends on t and has a diagonal structure. If we further assume that the source signals are non-stationary, then $\mathbf{R}_{\mathbf{ss}}(t)$ is time-varying, but always has a diagonal structure. The correlation matrix of the sensor signals $\mathbf{R}_{\mathbf{xx}}(t) \triangleq E\{\mathbf{x}(t)\mathbf{x}^H(t)\} = \mathbf{A}\,E\{\mathbf{s}(t)\mathbf{s}^H(t)\}\mathbf{A}^H = \mathbf{A}\,\mathbf{R}_{\mathbf{ss}}(t)\mathbf{A}^H$ becomes also time-varying, but has no diagonal structure in general. In order to recover the unknown source signals, we aim at finding a *separation matrix* $\mathbf{W}$ such that $\mathbf{u}(t) = \mathbf{W}\mathbf{x}(t)$ becomes an estimate of the original source signals $\mathbf{s}(t)$, aside from a possible scaling and permutation of the elements in $\mathbf{u}(t)$. In order for $\mathbf{u}(t)$ to become an estimate of $\mathbf{s}(t)$, the correlation matrix

$$\mathbf{R}_{\mathbf{uu}}(t) \triangleq E\{\mathbf{u}(t)\mathbf{u}^H(t)\} = \mathbf{W}\,\mathbf{R}_{\mathbf{xx}}(t)\mathbf{W}^H \tag{1}$$

has to be diagonal for all t. Assuming that $\mathbf{W}$ is time-invariant, $\mathbf{W}$ has to be of the following structure:

$$\mathbf{W} = \mathbf{DPA}^{-1} \tag{2}$$

---

*Bose Corp., Framingham, MA, USA (`joho@ieee.org`).

or, if $\mathbf{A}^{N \times M}$ is a tall matrix $(N > M)$,

$$\mathbf{W} = \mathbf{DPA}^{\#} \qquad (3)$$

where $\mathbf{D}$ is an arbitrary diagonal matrix, $\mathbf{P}$ is an arbitrary permutation matrix, and $\mathbf{A}^{\#}$ denotes the pseudo-inverse of $\mathbf{A}$. If we take P snapshots of $\mathbf{R_{xx}}(t)$, $\mathbf{R}_p \triangleq \mathbf{R_{xx}}(t_p)$, where t_p is the time instance of the pth snapshot, then we can formulate the task of finding a proper $\mathbf{W}$ as the following *joint-diagonalization problem*:

Given a set of P positive definite Hermitian matrices $\mathbf{R}_p$, find a single matrix $\mathbf{W}$ that approximately joint diagonalizes the whole set $\{\mathbf{R}_p\}$ such that $\mathbf{WR}_p\mathbf{W}^H$ is diagonal $\forall p$.

Perfect diagonalization is typically not possible for a set of random positive definite Hermitian matrixes $\{\mathbf{R}_p\}$, unless $\{\mathbf{R}_p\}$ is a set of commuting matrices [9]. However, the set $\{\mathbf{W}\,\mathbf{R}_p\mathbf{W}^H\}$ can still be approximately jointly diagonalized subject to a given cost function $\mathcal{J}(\mathbf{W}; \{\mathbf{R}_p\}_{p=1}^P)$ that measures the degree of joint diagonalization.

Several cost functions for the joint-diagonalization problem have been published in the last decade. In [5] Cardoso and Souloumiac have proposed a joint-diagonalization cost function and have given a very effective Jacobi-type algorithm that minimizes their cost function under the constraint that the diagonalization matrix $\mathbf{W}$ is unitary. The core idea of their algorithm has been used in in JADE [4] and SOBI [1], both very popular BSS algorithms.

In [17] Yeredor has published an algorithm, AC-DC, that uses a different cost function and is based on a subspace-fitting formulation. One advantage is that no orthogonality constraints are imposed on the diagonalization matrix $\mathbf{W}$. However, $\mathbf{W}$ needs to be square and real. In [19] Yeredor *et. al* have also derived an algorithm based on the *natural gradient* that works even with non-positive definite matrices $\mathbf{R}_p$, however, $\mathbf{W}$ needs to be real and square. Yeredor has also described in [18] how to compute a good initial value $\mathbf{W}_0$ for any iterative algorithm.

Another joint-diagonalization algorithm, FFDIAG, that seems to have a very fast convergence rate, was presented by Ziehe *et. al.* in [20]. FFDIAG also requires that $\mathbf{R}_p$ and $\mathbf{W}$ need to be a real square matrices. An extension of FFDIAG where $\mathbf{W}$ can be non-square or complex has not been published so far.

Recently, Vollgraf and Obermayer [16] have published an algorithm for the real case, QDIAG, that also works for a non-square $\mathbf{W}$. Their algorithm sequentially solves a quadratic sub-problem, which avoids the appearance of any higher-order terms in their cost function. QDIAG seems to have a similar convergence performance as FFDIAG and is very appealing from a computational point of view for joint-diagonalizing large sets of matrices.

In [14], Pham has presented an efficient joint-diagonalization algorithm that imposes no optimization constraints on $\mathbf{W}$, except that $\mathbf{W}$ needs to be square $(N = M)$. This algorithm is also a Jacobi-type algorithm. In contrast to most other joint-diagonalization algorithms, which minimize a constraint optimization problem, Pham's algorithm is formulated to minimize an unconstraint optimization problem. The underlying cost function of this algorithm is based on some preliminary work by Fluri and Gautschi [7,8], namely

$$\mathcal{J}(\mathbf{W}) \triangleq \sum_{p=1}^{P} \beta_p \left[\log\left(\det\left(\mathrm{diag}\left(\mathbf{W}\,\mathbf{R}_p\,\mathbf{W}^H\right)\right)\right) - \log\left(\det\left(\mathbf{W}\,\mathbf{R}_p\,\mathbf{W}^H\right)\right) \right] \qquad (4)$$

where the matrices $\mathbf{R}_p \in \mathbb{C}^{N \times N}$ are Hermitian and need to be positive definite. The weights β_p are positive scalars. Consequently, the matrix products $\mathbf{W}\,\mathbf{R}_p\,\mathbf{W}^H \in$

$\mathbb{C}^{M \times M}$ are also Hermitian and positive definite, assuming $\mathbf{W}$ has full rank. The cost function (4) is motivated by the Hadamard inequality

$$\det(\mathbf{Q}) \leq \det(\operatorname{diag}(\mathbf{Q})) \tag{5}$$

with equality if and only if $\mathbf{Q}$ is diagonal [9].

In the following, we will also use the cost function (4) as the basis of a Newton-type algorithm that we will derive. There are a few fundamental differences between the derived Newton algorithm and Pham's Jacobi-type algorithm. In Pham's algorithm, every iteration consists of a pairwise update of two rows at a time. One *sweep* of the algorithm consists of iterating once through all possible combinations of pairing two rows. Pham has shown that near a minimum, an iteration of his algorithm behaves similarly to a *quasi Newton-Raphson* iteration.

In contrast to Pham's algorithm, we will use a pure Newton algorithm that updates *all coefficients* in $\mathbf{W}$ in every iteration. The main difficulty in deriving a pure Newton algorithm, as opposed to a *quasi*-Newton algorithm, is that the Hessian needs to be known in closed form in every iteration. One major difference to Pham's algorithm will be that we drop the constraint that $\mathbf{W}$ needs to be square, and allow $\mathbf{W} \in \mathbb{C}^{M \times N}$ to be rectangular with $M \leq N$. Translated to the blind signal separation problem, our algorithm is also capable to work for the case where more sensor signals than source signals are present. This is particularly useful in case the number of source signals is not known beforehand.

In the following sections, we will derive the gradient and Hessian of the cost function (4) in close form. To this end, we will use the matrix-form representation of the second-order Taylor series expansion as described by Manton in [13]. This form allows to represent the gradient and the Hessian in a very compact form with the help of Kronecker products. As we will see, just as products of matrices reveal more structure than nested sums, the use of Kronecker products reveals the structure of the Hessian on an even higher level than by using matrices inside nested sums. Thorough treatments of Kronecker products and their properties are given in [3, 12].

1.1. 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 $(.)^H$, respectively. The $M \times M$ identity matrix is denoted by $\mathbf{I}_{M \times M}$. The Frobenius norm and the trace of a matrix are denoted by $\| . \|_F$ and $\operatorname{tr}(.)$, respectively. The spectral radius of a matrix $\mathbf{Q}$ is the nonnegative real number $\rho(\mathbf{Q}) = \max\{|\lambda_{\max}(\mathbf{Q})|\}$, see [9]. Matrix dimensions are given in superscript, e.g., $\mathbf{W}^{M \times N}$. The operator $\operatorname{vec}(\mathbf{W})$ forms a column vector by stacking the columns of $\mathbf{W}$, and $\mathbf{W}^{M \times N} = \operatorname{mat}_{M \times N}(\mathbf{w})$ is the inverse operation of $\mathbf{w} = \operatorname{vec}(\mathbf{W}^{M \times N})$. The Kronecker product [3] is denoted by $\otimes$. The $MN \times MN$ dimensional permutation matrix $\mathbf{P}_{M \times N}$, where the subscript $M \times N$ is the argument of $\mathbf{P}_{M \times N}$, is uniquely defined with

$$\operatorname{vec}(\mathbf{W}^T) \equiv \mathbf{P}_{M \times N} \operatorname{vec}(\mathbf{W}^{M \times N}) \tag{6}$$

as $\operatorname{vec}(\mathbf{W})$ and $\operatorname{vec}(\mathbf{W}^T)$ contain the same elements, just arranged in a different order. With $\bar{\mathbf{Q}} = \operatorname{diag}(\mathbf{q})$ we get a square diagonal matrix that contains the elements of the vector $\mathbf{q}$ in its diagonal. The matrix $\bar{\mathbf{Q}} = \operatorname{diag}(\mathbf{Q})$ is a diagonal matrix where its diagonal elements are the same as the diagonal elements of $\mathbf{Q}$, and

$$\operatorname{off}(\mathbf{Q}) \triangleq \mathbf{Q} - \operatorname{diag}(\mathbf{Q}) \tag{7}$$

keeps all off-diagonal elements of $\mathbf{Q}$ and sets all diagonal elements of $\mathbf{Q}$ to zero. Furthermore, we define the two following $M^2 \times M^2$ diagonal projection matrices

$$\mathbf{P}_{\text{diag}} \triangleq \text{diag}(\,\text{vec}(\,\mathbf{I}_{M \times M}\,)\,) \tag{8}$$

$$\mathbf{P}_{\text{off}} \triangleq \mathbf{I}_{M^2 \times M^2} - \mathbf{P}_{\text{diag}} \tag{9}$$

which appear in the following two relations:

$$\text{vec}(\,\text{diag}(\,\mathbf{Z}\,)\,) = \mathbf{P}_{\text{diag}}\,\text{vec}(\,\mathbf{Z}\,) \tag{10}$$

$$\text{vec}(\,\text{off}(\,\mathbf{Z}\,)\,) = \mathbf{P}_{\text{off}}\,\text{vec}(\,\mathbf{Z}\,). \tag{11}$$

2. Second order Taylor series of the cost function $\mathcal{J}(\mathbf{W})$.

2.1. Matrix form of second order Taylor approximation. In the following we will derive the gradient and Hessian of the cost function (4) with respect to the free parameters, i.e., the elements of $\mathbf{W}$. First we need to define how the gradient and Hessian are represented. Since the free parameters in the cost function (4) are arranged in the matrix $\mathbf{W}$, we decide to use the matrix form of the second order Taylor series as given by Manton in [13]:

Let $\mathcal{J} : \mathbb{C}^{M \times N} \to \mathbb{R}$ be a cost function. Then we can describe the Taylor series expansion of $\mathcal{J}$ at $\mathbf{W}$ as

$$\begin{aligned}
\mathcal{J}(\mathbf{W} + \delta \mathbf{Z}) = \mathcal{J}(\mathbf{W}) &+ \delta\,\mathfrak{Re}\{\text{tr}\,(\mathbf{Z}^H \mathbf{D_W})\} \\
&+ \frac{\delta^2}{2}\,\text{vec}(\,\mathbf{Z}\,)^H \mathbf{H_W}\,\text{vec}(\,\mathbf{Z}\,) \\
&+ \frac{\delta^2}{2}\mathfrak{Re}\{\text{vec}(\,\mathbf{Z}\,)^T \mathbf{C_W}\,\text{vec}(\,\mathbf{Z}\,)\} + O(\delta^3)
\end{aligned} \tag{12}$$

where $\mathbf{W}, \mathbf{Z} \in \mathbb{C}^{M \times N}$, $\mathbf{D_W} \in \mathbb{C}^{M \times N}$ is the gradient of $\mathcal{J}$ evaluated at $\mathbf{W}$, and $\mathbf{H_W}, \mathbf{C_W} \in \mathbb{C}^{MN \times MN}$ are the Hessian of $\mathcal{J}$ evaluated at $\mathbf{W}$. The scalar δ is a small real number. Uniqueness can be achieved by requiring $\mathbf{H_W}^H = \mathbf{H_W}$ and $\mathbf{C_W}^T = \mathbf{C_W}$.

In contrast to the commonly known vector form of the Taylor series expansion

$$\mathcal{J}(\mathbf{w} + \delta \mathbf{z}) = \mathcal{J}(\mathbf{w}) + \delta\,\mathbf{z}^T \mathbf{d} + \frac{\delta^2}{2}\mathbf{z}^T \mathbf{H}\,\mathbf{z} + O(\delta^3) \tag{13}$$

where the coefficients of $\mathbf{W}$ are rearranged in the real vector

$$\mathbf{w} \triangleq \begin{pmatrix} \mathbf{w}^{\text{re}} \\ \mathbf{w}^{\text{im}} \end{pmatrix} \triangleq \begin{pmatrix} \mathfrak{Re}\{\text{vec}(\mathbf{W})\} \\ \mathfrak{Im}\{\text{vec}(\mathbf{W})\} \end{pmatrix}, \tag{14}$$

the matrix form often reveals the structure of the gradient and the Hessian in a much more transparent form via matrix and Kronecker products. The gradients and the Hessians of the two Taylor expansion forms (12) and (13) can be transformed into each other, as described in [10].

2.2. Derivation of the gradient and Hessian. In order to simplify the derivation of the gradient and Hessian of $\mathcal{J}(.)$, we rewrite the cost function (4) as

$$\mathcal{J}(\mathbf{W}\,;\{\beta_p\},\{\mathbf{R}_p\}) \triangleq \sum_{p=1}^{P} \beta_p\,\widetilde{\mathcal{J}}(\mathbf{W};\mathbf{R}_p) \tag{15}$$

with

$$\tilde{\mathcal{J}}\left(\mathbf{W};\mathbf{R}_p\right) \triangleq \mathcal{J}^{(1)}\left(\mathbf{W};\mathbf{R}_p\right) - \mathcal{J}^{(2)}\left(\mathbf{W};\mathbf{R}_p\right) \tag{16}$$

$$\mathcal{J}^{(1)}\left(\mathbf{W};\mathbf{R}\right) \triangleq \log\left(\det\left(\operatorname{diag}\left(\mathbf{W}\,\mathbf{R}\,\mathbf{W}^H\right)\right)\right) \tag{17}$$

$$\mathcal{J}^{(2)}\left(\mathbf{W};\mathbf{R}\right) \triangleq \log\left(\det\left(\mathbf{W}\,\mathbf{R}\,\mathbf{W}^H\right)\right). \tag{18}$$

2.3. Gradient and Hessian of $\mathcal{J}^{(1)}(.)$. In order to derive the gradient and Hessian of $\mathcal{J}^{(1)}$, defined in (17), we do the following expansion

$$\mathcal{J}^{(1)}\left(\mathbf{W}+\delta\,\mathbf{Z}\right) \triangleq \log\left(\det\left(\operatorname{diag}\left((\mathbf{W}+\delta\,\mathbf{Z})\,\mathbf{R}\,(\mathbf{W}+\delta\,\mathbf{Z})^H\right)\right)\right) \tag{19}$$

$$= \log\left(\det\left(\operatorname{diag}\left((\mathbf{W}\,\mathbf{R}\,\mathbf{W}^H + \delta\left(\mathbf{W}\,\mathbf{R}\,\mathbf{Z}^H + \mathbf{Z}\,\mathbf{R}\,\mathbf{W}^H\right) + \delta^2\mathbf{Z}\,\mathbf{R}\,\mathbf{Z}^H\right)\right)\right)\right). \tag{20}$$

By substituting

$$\bar{\mathbf{Q}} \triangleq \operatorname{diag}\left(\mathbf{W}\,\mathbf{R}\,\mathbf{W}^H\right) \tag{21}$$

we can formulate (20) as

$$\mathcal{J}^{(1)}\left(\mathbf{W}+\delta\,\mathbf{Z}\right) = \log\Big(\det\big(\bar{\mathbf{Q}}\,\big(\mathbf{I}+\delta\,\bar{\mathbf{Q}}^{-1}\operatorname{diag}\left(\mathbf{W}\,\mathbf{R}\,\mathbf{Z}^H + \mathbf{Z}\,\mathbf{R}\,\mathbf{W}^H\right)$$
$$+ \delta^2\,\bar{\mathbf{Q}}^{-1}\operatorname{diag}(\mathbf{Z}\,\mathbf{R}\,\mathbf{Z}^H)\big)\big)\Big) \tag{22}$$

$$= \mathcal{J}^{(1)}\left(\mathbf{W}\right) + \log\Big(\det\big(\mathbf{I}+\delta\,\bar{\mathbf{Q}}^{-1}\operatorname{diag}\left(\mathbf{W}\,\mathbf{R}\,\mathbf{Z}^H + \mathbf{Z}\,\mathbf{R}\,\mathbf{W}^H\right)$$
$$+ \delta^2\,\bar{\mathbf{Q}}^{-1}\operatorname{diag}\left(\mathbf{Z}\,\mathbf{R}\,\mathbf{Z}^H\right)\big)\Big) \tag{23}$$

$$= \mathcal{J}^{(1)}\left(\mathbf{W}\right) + \log\left(\det\left(\mathbf{I}+\delta\,\mathbf{A}+\delta^2\mathbf{B}\right)\right) \tag{24}$$

with

$$\mathbf{A} \triangleq \bar{\mathbf{Q}}^{-1}\operatorname{diag}\left(\mathbf{W}\,\mathbf{R}\,\mathbf{Z}^H + \mathbf{Z}\,\mathbf{R}\,\mathbf{W}^H\right) \tag{25}$$

$$\mathbf{B} \triangleq \bar{\mathbf{Q}}^{-1}\operatorname{diag}\left(\mathbf{Z}\,\mathbf{R}\,\mathbf{Z}^H\right). \tag{26}$$

Here we made use of $\det\mathbf{XY} = \det\mathbf{X}\det\mathbf{Y}$ which is valid for square matrices $\mathbf{X}$, $\mathbf{Y}$. Before we continue to simplify (24), we introduce the following proposition:

PROPOSITION 2.1. *Let $\mathbf{A}, \mathbf{B} \in \mathbb{C}^{N \times N}$. Then, for $\delta \to 0$,*

$$\log\left(\det\left(\mathbf{I}+\delta\,\mathbf{A}+\delta^2\mathbf{B}\right)\right) = \delta\operatorname{tr}\mathbf{A} + \delta^2\operatorname{tr}\mathbf{B} - \frac{1}{2}\delta^2\operatorname{tr}\mathbf{A}^2 + O(\delta^3). \tag{27}$$

Proof. Let $\mathbf{S} \in \mathbb{C}^{N \times N}$ with *spectral radius* $\rho(\mathbf{S}) < 1$. We know that for any non-singular $\mathbf{G} \in \mathbb{C}^{N \times N}$ we have $\log\left(\det\left(\mathbf{G}\right)\right) = \operatorname{tr}\left(\log\left(\mathbf{G}\right)\right)$. We also know that $\log\left(\mathbf{I}+\mathbf{G}\right) = \sum_k \frac{(-1)^{k+1}}{k}\mathbf{G}^k$ for any $\mathbf{G} \in \mathbb{C}^{N \times N}$ with $\rho(\mathbf{G}) < 1$. Note that the assumption $\rho(\mathbf{S}) < 1$ implies that $\mathbf{I}+\mathbf{S}$ is non-singular. Hence, combining the above results we have

$$\log(\det(\mathbf{I}+\mathbf{S})) = \sum_k \frac{(-1)^{k+1}}{k}\mathbf{S}^k. \tag{28}$$

Now, if we insert $\mathbf{S} = \delta\,\mathbf{A}+\delta^2\,\mathbf{B}$ into (28) and let $\delta \to 0$ we obtain (27). $\square$

Note that the logarithm for complex arguments is not uniquely defined. However, we will only apply (28) for Hermitian matrices $\mathbf{S}$ with $\rho(\mathbf{S}) < 1$. Consequently all eigenvalues of $\mathbf{I}+\mathbf{S}$, and also $\det(\mathbf{I}+\mathbf{S})$, will be real and positive.

If we apply (27) in (24) we obtain

$$\mathcal{J}^{(1)}\left(\mathbf{W}+\delta\,\mathbf{Z}\right)=\mathcal{J}^{(1)}\left(\mathbf{W}\right)+\delta\operatorname{tr}\mathbf{A}+\delta^{2}\operatorname{tr}\mathbf{B}-\frac{1}{2}\delta^{2}\operatorname{tr}\mathbf{A}^{2}+O(\delta^{3}) \qquad (29)$$

where $\mathbf{A}$ and $\mathbf{B}$ are defined in (25) and (26), respectively. We now analyze each of the three δ terms in (29) separately. We start with

$$\operatorname{tr}\mathbf{A}=\operatorname{tr}\left(\bar{\mathbf{Q}}^{-1}\mathbf{W}\,\mathbf{R}\,\mathbf{Z}^{H}+\bar{\mathbf{Q}}^{-1}\mathbf{Z}\,\mathbf{R}\,\mathbf{W}^{H}\right) \qquad (30)$$
$$=\operatorname{tr}\left(\mathbf{Z}^{H}\,\bar{\mathbf{Q}}^{-1}\mathbf{W}\,\mathbf{R}+\mathbf{R}\,\mathbf{W}^{H}\,\bar{\mathbf{Q}}^{-1}\mathbf{Z}\right) \qquad (31)$$
$$=2\,\mathfrak{Re}\{\operatorname{tr}\left(\mathbf{Z}^{H}\,\bar{\mathbf{Q}}^{-1}\mathbf{W}\,\mathbf{R}\right)\}. \qquad (32)$$

Here we made use of $\operatorname{tr}\left(\operatorname{diag}(\mathbf{X})\operatorname{diag}(\mathbf{Y})\right)=\operatorname{tr}\left(\mathbf{X}\operatorname{diag}(\mathbf{Y})\right)=\operatorname{tr}\left(\operatorname{diag}(\mathbf{X})\mathbf{Y}\right)$. The third term in (29), where $\mathbf{A}$ is defined in (25), can be modified as

$$\operatorname{tr}\mathbf{A}^{2}=\operatorname{tr}\left(\bar{\mathbf{Q}}^{-1}\left(\mathbf{W}\,\mathbf{R}\,\mathbf{Z}^{H}+\mathbf{Z}\,\mathbf{R}\,\mathbf{W}^{H}\right)\bar{\mathbf{Q}}^{-1}\operatorname{diag}\left(\mathbf{W}\,\mathbf{R}\,\mathbf{Z}^{H}+\mathbf{Z}\,\mathbf{R}\,\mathbf{W}^{H}\right)\right) \qquad (33)$$
$$\begin{aligned}
=\ &\operatorname{tr}\left(\bar{\mathbf{Q}}^{-1}\mathbf{W}\,\mathbf{R}\,\mathbf{Z}^{H}\bar{\mathbf{Q}}^{-1}\operatorname{diag}\left(\mathbf{W}\,\mathbf{R}\,\mathbf{Z}^{H}\right)\right)\\
&+\operatorname{tr}\left(\bar{\mathbf{Q}}^{-1}\mathbf{W}\,\mathbf{R}\,\mathbf{Z}^{H}\bar{\mathbf{Q}}^{-1}\operatorname{diag}\left(\mathbf{Z}\,\mathbf{R}\,\mathbf{W}^{H}\right)\right)\\
&+\operatorname{tr}\left(\bar{\mathbf{Q}}^{-1}\mathbf{Z}\,\mathbf{R}\,\mathbf{W}^{H}\bar{\mathbf{Q}}^{-1}\operatorname{diag}\left(\mathbf{W}\,\mathbf{R}\,\mathbf{Z}^{H}\right)\right)\\
&+\operatorname{tr}\left(\bar{\mathbf{Q}}^{-1}\mathbf{Z}\,\mathbf{R}\,\mathbf{W}^{H}\bar{\mathbf{Q}}^{-1}\operatorname{diag}\left(\mathbf{Z}\,\mathbf{R}\,\mathbf{W}^{H}\right)\right)
\end{aligned} \qquad (34)$$
$$\begin{aligned}
=\ &2\operatorname{tr}\left(\mathbf{Z}^{H}\,\bar{\mathbf{Q}}^{-1}\operatorname{diag}\left(\mathbf{Z}\,\mathbf{R}\,\mathbf{W}^{H}\right)\bar{\mathbf{Q}}^{-1}\mathbf{W}\,\mathbf{R}\right)\\
&+\operatorname{tr}\left(\mathbf{Z}\,\mathbf{R}\,\mathbf{W}^{H}\,\bar{\mathbf{Q}}^{-1}\operatorname{diag}\left(\mathbf{Z}\,\mathbf{R}\,\mathbf{W}^{H}\right)\bar{\mathbf{Q}}^{-1}\right)\\
&+\operatorname{tr}\left(\mathbf{Z}^{H}\,\bar{\mathbf{Q}}^{-1}\operatorname{diag}\left(\mathbf{W}\,\mathbf{R}\,\mathbf{Z}^{H}\right)\bar{\mathbf{Q}}^{-1}\mathbf{W}\,\mathbf{R}\right)
\end{aligned} \qquad (35)$$

where we used $\operatorname{tr}\left(\mathbf{X}\,\mathbf{Y}\right)=\operatorname{tr}\left(\mathbf{Y}\,\mathbf{X}\right)$ and some elementary properties of $\operatorname{tr}(.)$ with diagonal matrices. With the help of (94) and (95) we obtain

$$\begin{aligned}
\operatorname{tr}\mathbf{A}^{2}=\ &2\operatorname{vec}(\mathbf{Z})^{H}\left[\left(\bar{\mathbf{Q}}^{-1}\mathbf{W}\,\mathbf{R}\right)^{T}\otimes\bar{\mathbf{Q}}^{-1}\right]\operatorname{vec}\left(\operatorname{diag}\left(\mathbf{Z}\,\mathbf{R}\,\mathbf{W}^{H}\right)\right)\\
&+\operatorname{vec}(\mathbf{Z})^{T}\mathbf{P}_{M\times N}^{T}\left[\bar{\mathbf{Q}}^{-T}\otimes\mathbf{R}\,\mathbf{W}^{H}\,\bar{\mathbf{Q}}^{-1}\right]\operatorname{vec}\left(\operatorname{diag}\left(\mathbf{Z}\,\mathbf{R}\,\mathbf{W}^{H}\right)\right)\\
&+\operatorname{vec}(\mathbf{Z})^{H}\left[\left(\bar{\mathbf{Q}}^{-1}\mathbf{W}\,\mathbf{R}\right)^{T}\otimes\bar{\mathbf{Q}}^{-1}\right]\operatorname{vec}\left(\operatorname{diag}\left(\mathbf{W}\,\mathbf{R}\,\mathbf{Z}^{H}\right)\right).
\end{aligned} \qquad (36)$$

Next we make use of (10). With further help of (93) and (84) we get

$$\begin{aligned}
\operatorname{tr}\mathbf{A}^{2}=\ &2\operatorname{vec}(\mathbf{Z})^{H}\left[\left(\bar{\mathbf{Q}}^{-1}\mathbf{W}\,\mathbf{R}\right)^{T}\otimes\bar{\mathbf{Q}}^{-1}\right]\mathbf{P}_{\mathrm{diag}}\left[\left(\mathbf{R}\,\mathbf{W}^{H}\right)^{T}\otimes\mathbf{I}_{M}\right]\operatorname{vec}(\mathbf{Z})\\
&+\operatorname{vec}(\mathbf{Z})^{T}\mathbf{P}_{M\times N}^{T}\left[\bar{\mathbf{Q}}^{-T}\otimes\mathbf{R}\,\mathbf{W}^{H}\,\bar{\mathbf{Q}}^{-1}\right]\mathbf{P}_{\mathrm{diag}}\left[\left(\mathbf{R}\,\mathbf{W}^{H}\right)^{T}\otimes\mathbf{I}_{M}\right]\operatorname{vec}(\mathbf{Z})\\
&+\operatorname{vec}(\mathbf{Z})^{H}\left[\left(\bar{\mathbf{Q}}^{-1}\mathbf{W}\,\mathbf{R}\right)^{T}\otimes\bar{\mathbf{Q}}^{-1}\right]\mathbf{P}_{\mathrm{diag}}\left[\mathbf{I}_{M}\otimes\mathbf{W}\,\mathbf{R}\right]\mathbf{P}_{M\times N}\operatorname{vec}(\mathbf{Z})^{*}.
\end{aligned} \qquad (37)$$

In this case $\mathbf{P}_{\mathrm{diag}}$ is an $M^{2}\times M^{2}$ dimensional matrix. By making use of (92) we obtain

$$\begin{aligned}
\operatorname{tr}\mathbf{A}^{2}=\ &2\operatorname{vec}(\mathbf{Z})^{H}\left[\mathbf{R}^{T}\,\mathbf{W}^{T}\,\bar{\mathbf{Q}}^{-T}\otimes\bar{\mathbf{Q}}^{-1}\right]\mathbf{P}_{\mathrm{diag}}\left[\mathbf{W}^{*}\,\mathbf{R}^{T}\otimes\mathbf{I}_{M}\right]\operatorname{vec}(\mathbf{Z})\\
&+\operatorname{vec}(\mathbf{Z})^{T}\left[\mathbf{R}\,\mathbf{W}^{H}\,\bar{\mathbf{Q}}^{-1}\otimes\bar{\mathbf{Q}}^{-T}\right]\mathbf{P}_{M\times M}\,\mathbf{P}_{\mathrm{diag}}\left[\mathbf{W}^{*}\mathbf{R}^{T}\otimes\mathbf{I}_{M}\right]\operatorname{vec}(\mathbf{Z})\\
&+\operatorname{vec}(\mathbf{Z})^{H}\left[\mathbf{R}^{T}\,\mathbf{W}^{T}\bar{\mathbf{Q}}^{-T}\otimes\bar{\mathbf{Q}}^{-1}\right]\mathbf{P}_{\mathrm{diag}}\,\mathbf{P}_{M\times M}\left[\mathbf{W}\,\mathbf{R}\otimes\mathbf{I}_{M}\right]\operatorname{vec}(\mathbf{Z})^{*}.
\end{aligned} \qquad (38)$$

Since $\mathbf{P}_{N\times M} = \mathbf{P}_{M\times N}^{-1} = \mathbf{P}_{M\times N}^{T}$ and $\mathbf{P}_{M\times M}\,\mathbf{P}_{\mathrm{diag}} = \mathbf{P}_{\mathrm{diag}}\,\mathbf{P}_{M\times M}$, we get

$$
\begin{aligned}
\operatorname{tr}\mathbf{A}^2 &= 2\operatorname{vec}(\mathbf{Z})^H\left[\mathbf{R}^T\mathbf{W}^T\bar{\mathbf{Q}}^{-T}\otimes\bar{\mathbf{Q}}^{-1}\right]\mathbf{P}_{\mathrm{diag}}\left[\mathbf{W}^*\mathbf{R}^T\otimes\mathbf{I}_M\right]\operatorname{vec}(\mathbf{Z}) \\
&\quad + \operatorname{vec}(\mathbf{Z})^T\left[\mathbf{R}\,\mathbf{W}^H\bar{\mathbf{Q}}^{-1}\otimes\bar{\mathbf{Q}}^{-T}\right]\mathbf{P}_{\mathrm{diag}}\,\mathbf{P}_{M\times M}\left[\mathbf{W}^*\mathbf{R}^T\otimes\mathbf{I}_M\right]\operatorname{vec}(\mathbf{Z}) \\
&\quad + \operatorname{vec}(\mathbf{Z})^H\left[\mathbf{R}^T\mathbf{W}^T\bar{\mathbf{Q}}^{-T}\otimes\bar{\mathbf{Q}}^{-1}\right]\mathbf{P}_{\mathrm{diag}}\,\mathbf{P}_{M\times M}\left[\mathbf{W}\,\mathbf{R}\otimes\mathbf{I}_M\right]\operatorname{vec}(\mathbf{Z})^* \quad (39) \\
&= 2\operatorname{vec}(\mathbf{Z})^H\left[\mathbf{R}^T\mathbf{W}^T\bar{\mathbf{Q}}^{-T}\otimes\bar{\mathbf{Q}}^{-1}\right]\mathbf{P}_{\mathrm{diag}}\left[\mathbf{W}^*\mathbf{R}^T\otimes\mathbf{I}_M\right]\operatorname{vec}(\mathbf{Z}) \\
&\quad + 2\,\mathfrak{Re}\{\operatorname{vec}(\mathbf{Z})^T\left[\mathbf{R}\,\mathbf{W}^H\bar{\mathbf{Q}}^{-1}\otimes\bar{\mathbf{Q}}^{-T}\right]\mathbf{P}_{\mathrm{diag}}\,\mathbf{P}_{M\times M}\left[\mathbf{W}^*\mathbf{R}^T\otimes\mathbf{I}_M\right]\operatorname{vec}(\mathbf{Z})\}.
\end{aligned}
$$
$$(40)$$

In the last step we used the fact that $\mathbf{R}^* = \mathbf{R}^T$ and $\bar{\mathbf{Q}}^* = \bar{\mathbf{Q}}^T$, as $\mathbf{R}$ and $\bar{\mathbf{Q}}$ are both Hermitian. Finally, the second term in (29) can be modified with (95) as

$$
\begin{aligned}
\operatorname{tr}\mathbf{B} = \operatorname{tr}\left(\bar{\mathbf{Q}}^{-1}\operatorname{diag}(\mathbf{Z}\,\mathbf{R}\,\mathbf{Z}^H)\right) &= \operatorname{tr}\left(\bar{\mathbf{Q}}^{-1}\mathbf{Z}\,\mathbf{R}\,\mathbf{Z}^H\right) && (41) \\
&= \operatorname{tr}\left(\mathbf{Z}^H\bar{\mathbf{Q}}^{-1}\mathbf{Z}\,\mathbf{R}\right) && (42) \\
&= \operatorname{vec}(\mathbf{Z})^H\left[\mathbf{R}^T\otimes\bar{\mathbf{Q}}^{-1}\right]\operatorname{vec}(\mathbf{Z}). && (43)
\end{aligned}
$$

Inserting (32), (40), and (43) into (29) yields

$$
\begin{aligned}
\mathcal{J}^{(1)}&(\mathbf{W}+\delta\,\mathbf{Z}) \\
&= \mathcal{J}^{(1)}(\mathbf{W}) + 2\,\delta\,\mathfrak{Re}\{\operatorname{tr}\left(\mathbf{Z}^H\bar{\mathbf{Q}}^{-1}\mathbf{W}\,\mathbf{R}\right)\} + \delta^2\operatorname{vec}(\mathbf{Z})^H\cdot \\
&\quad\cdot\left(\left[\mathbf{R}^T\otimes\bar{\mathbf{Q}}^{-1}\right]-\left[\mathbf{R}^T\mathbf{W}^T\bar{\mathbf{Q}}^{-T}\otimes\bar{\mathbf{Q}}^{-1}\right]\mathbf{P}_{\mathrm{diag}}\left[\mathbf{W}^*\mathbf{R}^T\otimes\mathbf{I}_M\right]\right)\operatorname{vec}(\mathbf{Z}) \\
&\quad - \delta^2\,\mathfrak{Re}\{\operatorname{vec}(\mathbf{Z})^T\left[\mathbf{R}\,\mathbf{W}^H\bar{\mathbf{Q}}^{-1}\otimes\bar{\mathbf{Q}}^{-T}\right]\mathbf{P}_{\mathrm{diag}}\,\mathbf{P}_{M\times M}\left[\mathbf{W}^*\mathbf{R}^T\otimes\mathbf{I}_M\right]\operatorname{vec}(\mathbf{Z})\} \\
&\quad + O(\delta^3).
\end{aligned}
$$
$$(44)$$

By coefficient comparison between (44) and the matrix form of the second-order Taylor series (12), and using $\mathbf{R}^T = \mathbf{R}^*$ and (91), we finally obtain

$$
\mathbf{D}_{\mathbf{W}}^{(1)} = 2\,\bar{\mathbf{Q}}^{-1}\mathbf{W}\,\mathbf{R} \tag{45}
$$

$$
\mathbf{H}_{\mathbf{W}}^{(1)} = 2\left[\mathbf{R}^T\otimes\bar{\mathbf{Q}}^{-1}\right] - 2\left[\mathbf{R}^T\mathbf{W}^T\otimes\mathbf{I}_M\right]\left[\bar{\mathbf{Q}}^{-T}\otimes\bar{\mathbf{Q}}^{-1}\right]\mathbf{P}_{\mathrm{diag}}\left[\mathbf{W}^*\mathbf{R}^*\otimes\mathbf{I}_M\right] \tag{46}
$$

$$
\mathbf{C}_{\mathbf{W}}^{(1)} = -2\left[\mathbf{R}\,\mathbf{W}^H\otimes\mathbf{I}_M\right]\left[\bar{\mathbf{Q}}^{-1}\otimes\bar{\mathbf{Q}}^{-T}\right]\mathbf{P}_{\mathrm{diag}}\,\mathbf{P}_{M\times M}\left[\mathbf{W}^*\mathbf{R}^*\otimes\mathbf{I}_M\right] \tag{47}
$$

where $\bar{\mathbf{Q}}$ is defined in (21).

2.4. Gradient and Hessian of $\mathcal{J}^{(2)}(.)$. Deriving the gradient and Hessian of $\mathcal{J}^{(2)}(\mathbf{W}) \triangleq \log\left(\det\left(\mathbf{W}\,\mathbf{R}\,\mathbf{W}^H\right)\right)$, as defined in (18), can be done in a similarly way as it was done for $\mathcal{J}^{(1)}$ in the previous section. To this end, we expand $\mathcal{J}^{(2)}$ as

$$
\begin{aligned}
\mathcal{J}^{(2)}(\mathbf{W}+\delta\,\mathbf{Z}) &= \log\left(\det\left((\mathbf{W}+\delta\,\mathbf{Z})\,\mathbf{R}\,(\mathbf{W}+\delta\,\mathbf{Z})^H\right)\right) && (48) \\
&= \log\left(\det\left(\mathbf{W}\,\mathbf{R}\,\mathbf{W}^H + \delta\left(\mathbf{W}\,\mathbf{R}\,\mathbf{Z}^H + \mathbf{Z}\,\mathbf{R}\,\mathbf{W}^H\right) + \delta^2\mathbf{Z}\,\mathbf{R}\,\mathbf{Z}^H\right)\right).
\end{aligned}
$$
$$(49)$$

By substituting

$$
\mathbf{Q} \triangleq \mathbf{W}\,\mathbf{R}\,\mathbf{W}^H \tag{50}
$$

we can rewrite (49) as

$$\mathcal{J}^{(2)}\left(\mathbf{W}+\delta\,\mathbf{Z}\right)$$

$$= \log\left(\det\left(\mathbf{Q}\left(\mathbf{I}+\delta\,\mathbf{Q}^{-1}\left(\mathbf{W}\,\mathbf{R}\,\mathbf{Z}^H+\mathbf{Z}\,\mathbf{R}\,\mathbf{W}^H\right)+\delta^2\,\mathbf{Q}^{-1}\mathbf{Z}\,\mathbf{R}\,\mathbf{Z}^H\right)\right)\right) \quad (51)$$

$$= \mathcal{J}^{(2)}\left(\mathbf{W}\right)+\log\left(\det\left(\mathbf{I}+\delta\,\mathbf{Q}^{-1}\left(\mathbf{W}\,\mathbf{R}\,\mathbf{Z}^H+\mathbf{Z}\,\mathbf{R}\,\mathbf{W}^H\right)+\delta^2\,\mathbf{Q}^{-1}\mathbf{Z}\,\mathbf{R}\,\mathbf{Z}^H\right)\right). \quad (52)$$

By inspection, we see that (52) has the same structure as (23) and therefore can also be written in the form

$$\mathcal{J}^{(2)}\left(\mathbf{W}+\delta\,\mathbf{Z}\right) = \mathcal{J}^{(2)}\left(\mathbf{W}\right)+\log\left(\det\left(\mathbf{I}+\delta\,\mathbf{A}+\delta^2\mathbf{B}\right)\right), \quad (53)$$

this time with

$$\mathbf{A} \triangleq \mathbf{Q}^{-1}\left(\mathbf{W}\,\mathbf{R}\,\mathbf{Z}^H+\mathbf{Z}\,\mathbf{R}\,\mathbf{W}^H\right) \quad (54)$$

$$\mathbf{B} \triangleq \mathbf{Q}^{-1}\mathbf{Z}\,\mathbf{R}\,\mathbf{Z}^H. \quad (55)$$

One difference, though, is that the matrices $\mathbf{Q}$, $\mathbf{A}$, and $\mathbf{B}$ are this time no longer diagonal, in general. However, since in the previous calculations we only made the assumption that these matrices needed to be Hermitian and not diagonal, we can use (27) again and carry out the further steps in exactly the same manner as it was done for $\mathcal{J}^{(1)}$. By going through the same derivation steps, we finally obtain

$$\mathbf{D}_{\mathbf{W}}^{(2)} = 2\,\mathbf{Q}^{-1}\,\mathbf{W}\,\mathbf{R} \quad (56)$$

$$\mathbf{H}_{\mathbf{W}}^{(2)} = 2\left[\mathbf{R}^T\otimes\mathbf{Q}^{-1}\right]-2\left[\mathbf{R}^T\,\mathbf{W}^T\otimes\mathbf{I}_M\right]\left[\mathbf{Q}^{-T}\otimes\mathbf{Q}^{-1}\right]\left[\mathbf{W}^*\,\mathbf{R}^*\otimes\mathbf{I}_M\right] \quad (57)$$

$$\mathbf{C}_{\mathbf{W}}^{(2)} = -2\left[\mathbf{R}\,\mathbf{W}^H\otimes\mathbf{I}_M\right]\left[\mathbf{Q}^{-1}\otimes\mathbf{Q}^{-T}\right]\mathbf{P}_{M\times M}\left[\mathbf{W}^*\mathbf{R}^*\otimes\mathbf{I}_M\right]. \quad (58)$$

By comparing (56) to (58) with (45) to (47), respectively, basically $\bar{\mathbf{Q}}$ is replaced by its non-diagonal version $\mathbf{Q}$, and $\mathbf{P}_{\mathrm{diag}}$ is replaced by $\mathbf{I}_M$.

2.5. Gradient and Hessian of $\mathcal{J}\left(.\right)$. Because of the definitions (16), (17), and (18) we can write the gradient and the Hessian of $\widetilde{\mathcal{J}}(.)$ as

$$\widetilde{\mathbf{D}}_{\mathbf{W}}\left(\mathbf{W};\mathbf{R}_p\right) = \mathbf{D}_{\mathbf{W}}^{(1)}\left(\mathbf{W};\mathbf{R}_p\right)-\mathbf{D}_{\mathbf{W}}^{(2)}\left(\mathbf{W};\mathbf{R}_p\right) \quad (59)$$

$$\widetilde{\mathbf{H}}_{\mathbf{W}}\left(\mathbf{W};\mathbf{R}_p\right) = \mathbf{H}_{\mathbf{W}}^{(1)}\left(\mathbf{W};\mathbf{R}_p\right)-\mathbf{H}_{\mathbf{W}}^{(2)}\left(\mathbf{W};\mathbf{R}_p\right) \quad (60)$$

$$\widetilde{\mathbf{C}}_{\mathbf{W}}\left(\mathbf{W};\mathbf{R}_p\right) = \mathbf{C}_{\mathbf{W}}^{(1)}\left(\mathbf{W};\mathbf{R}_p\right)-\mathbf{C}_{\mathbf{W}}^{(2)}\left(\mathbf{W};\mathbf{R}_p\right) \quad (61)$$

where $\mathbf{D}_{\mathbf{W}}^{(1)}$, $\mathbf{H}_{\mathbf{W}}^{(1)}$, and $\mathbf{C}_{\mathbf{W}}^{(1)}$ are defined in (45), (46), and (47), respectively, and $\mathbf{D}_{\mathbf{W}}^{(2)}$, $\mathbf{H}_{\mathbf{W}}^{(2)}$, and $\mathbf{C}_{\mathbf{W}}^{(2)}$ are defined in (56), (57), and (58), respectively. By inserting these terms into (59), (60), and (61), and after some rearranging we finally obtain $\forall p$

$$\widetilde{\mathbf{D}}_{\mathbf{W}}\left(\mathbf{W};\mathbf{R}\right) = 2\left(\bar{\mathbf{Q}}^{-1}-\mathbf{Q}^{-1}\right)\mathbf{W}\,\mathbf{R} \quad (62)$$

$$\widetilde{\mathbf{H}}_{\mathbf{W}}\left(\mathbf{W};\mathbf{R}\right) = 2\left[\mathbf{R}^T\otimes\left(\bar{\mathbf{Q}}^{-1}-\mathbf{Q}^{-1}\right)\right]$$
$$+2\left[\mathbf{R}^T\,\mathbf{W}^T\otimes\mathbf{I}_M\right]\left(\left[\mathbf{Q}^{-T}\otimes\mathbf{Q}^{-1}\right]-\left[\bar{\mathbf{Q}}^{-T}\otimes\bar{\mathbf{Q}}^{-1}\right]\mathbf{P}_{\mathrm{diag}}\right)$$
$$\cdot\left[\mathbf{W}^*\,\mathbf{R}^*\otimes\mathbf{I}_M\right] \quad (63)$$

$$\widetilde{\mathbf{C}}_{\mathbf{W}}\left(\mathbf{W};\mathbf{R}\right) = 2\left[\mathbf{R}\,\mathbf{W}^H\otimes\mathbf{I}_M\right]\left(\left[\mathbf{Q}^{-1}\otimes\mathbf{Q}^{-T}\right]-\left[\bar{\mathbf{Q}}^{-1}\otimes\bar{\mathbf{Q}}^{-T}\right]\mathbf{P}_{\mathrm{diag}}\right)$$
$$\cdot\mathbf{P}_{M\times M}\left[\mathbf{W}^*\mathbf{R}^*\otimes\mathbf{I}_M\right]. \quad (64)$$

Finally, because of (15), we can write the gradient and the Hessian of the cost function $\mathcal{J}$ as

$$\mathbf{D_W} = \sum_{p=1}^{P} \beta_p \, \widetilde{\mathbf{D}}_{\mathbf{W}} \left(\mathbf{W}; \mathbf{R}_p \right) \tag{65}$$

$$\mathbf{H_W} = \sum_{p=1}^{P} \beta_p \, \widetilde{\mathbf{H}}_{\mathbf{W}} \left(\mathbf{W}; \mathbf{R}_p \right) \tag{66}$$

$$\mathbf{C_W} = \sum_{p=1}^{P} \beta_p \, \widetilde{\mathbf{C}}_{\mathbf{W}} \left(\mathbf{W}; \mathbf{R}_p \right) \tag{67}$$

The equations (62) to (67) are the gradient and the Hessian of the cost function (4).

2.6. Comparison between gradient and Hessian of $\mathcal{J}$ and $\mathcal{J}_{\mathrm{off}}$. A well know cost function used for the joint-diagonalization problem is [5]

$$\mathcal{J}_{\mathrm{off}} \left(\mathbf{W} \, ; \{\beta_p\}, \{\mathbf{R}_p\} \right) \triangleq \sum_{p=1}^{P} \beta_p \, \left\| \, \mathrm{off}(\, \mathbf{W} \, \mathbf{R}_p \, \mathbf{W}^H \,) \, \right\|_F^2 \tag{68}$$

subject to a constraint that prevents $\mathbf{W}$ to become zero. We now like to make a comparison between the gradient and Hessian of $\mathcal{J}$ and $\mathcal{J}_{\mathrm{off}}$. The gradient and Hessian of the term $\hat{\mathcal{J}} \triangleq \| \, \mathrm{off}(\, \mathbf{W} \, \mathbf{R}_p \, \mathbf{W}^H \,) \|_F^2$, which appears in (68), are [10]

$$\widehat{\mathbf{D}}_{\mathbf{W}} \left(\mathbf{W}; \mathbf{R} \right) = 4 \, \mathrm{off}(\, \mathbf{Q} \,) \, \mathbf{W} \, \mathbf{R} \tag{69}$$

$$\widehat{\mathbf{H}}_{\mathbf{W}} \left(\mathbf{W}; \mathbf{R} \right) = 4 \left[\, \mathbf{R}^T \otimes \mathrm{off}(\, \mathbf{Q} \,) \, \right] + 4 \left[\, \mathbf{R}^T \, \mathbf{W}^T \otimes \mathbf{I}_M \, \right] \mathbf{P}_{\mathrm{off}} \left[\, \mathbf{W}^* \, \mathbf{R}^* \otimes \mathbf{I}_M \right] \tag{70}$$

$$\widehat{\mathbf{C}}_{\mathbf{W}} \left(\mathbf{W}; \mathbf{R} \right) = 4 \left[\, \mathbf{R} \, \mathbf{W}^H \otimes \mathbf{I}_M \, \right] \mathbf{P}_{\mathrm{off}} \, \mathbf{P}_{M \times M} \left[\, \mathbf{W}^* \mathbf{R}^* \otimes \mathbf{I}_M \right] . \tag{71}$$

We wish to bring the gradient and Hessian of $\mathcal{J}$ into a similar form. To this end, we reformulate the term $\bar{\mathbf{Q}}^{-1} - \mathbf{Q}^{-1}$ as

$$\bar{\mathbf{Q}}^{-1} - \mathbf{Q}^{-1} = \bar{\mathbf{Q}}^{-1} \, (\, \mathbf{Q} - \bar{\mathbf{Q}} \,) \mathbf{Q}^{-1} = \bar{\mathbf{Q}}^{-1} \, \mathrm{off}(\, \mathbf{Q} \,) \, \mathbf{Q}^{-1} \tag{72}$$

where $\mathrm{off}(\, . \,)$ is defined in (7) and $\bar{\mathbf{Q}} \triangleq \mathrm{diag}(\, \mathbf{Q} \,)$, see (21) and (50). Furthermore, since $\mathbf{P}_{\mathrm{diag}} + \mathbf{P}_{\mathrm{off}} = \mathbf{I}$ we have

$$
\begin{aligned}
& \left[\, \mathbf{Q}^{-T} \otimes \mathbf{Q}^{-1} \, \right] - \left[\, \bar{\mathbf{Q}}^{-T} \otimes \bar{\mathbf{Q}}^{-1} \, \right] \mathbf{P}_{\mathrm{diag}} \\
& \quad = \left(\left[\, \mathbf{Q}^{-T} \otimes \mathbf{Q}^{-1} \, \right] - \left[\, \bar{\mathbf{Q}}^{-T} \otimes \bar{\mathbf{Q}}^{-1} \, \right] \right) \mathbf{P}_{\mathrm{diag}} + \left[\, \mathbf{Q}^{-T} \otimes \mathbf{Q}^{-1} \, \right] \mathbf{P}_{\mathrm{off}} .
\end{aligned} \tag{73}
$$

By inserting (72) into (62) and (63), and (73) into (63) and (64) we can write the gradient and Hessian of $\mathcal{J}$ as

$$\widetilde{\mathbf{D}}_{\mathbf{W}} \left(\mathbf{W}; \mathbf{R} \right) = 2 \, \bar{\mathbf{Q}}^{-1} \, \mathrm{off}(\, \mathbf{Q} \,) \, \mathbf{Q}^{-1} \, \mathbf{W} \, \mathbf{R} \tag{74}$$

$$
\begin{aligned}
\widetilde{\mathbf{H}}_{\mathbf{W}} \left(\mathbf{W}; \mathbf{R} \right) = {} & 2 \left[\, \mathbf{R}^T \otimes \bar{\mathbf{Q}}^{-1} \, \mathrm{off}(\, \mathbf{Q} \,) \, \mathbf{Q}^{-1} \, \right] \\
& + 2 \left[\, \mathbf{R}^T \, \mathbf{W}^T \otimes \mathbf{I}_M \, \right] \left(\left[\, \mathbf{Q}^{-T} \otimes \mathbf{Q}^{-1} \, \right] - \left[\, \bar{\mathbf{Q}}^{-T} \otimes \bar{\mathbf{Q}}^{-1} \, \right] \right) \mathbf{P}_{\mathrm{diag}} \left[\, \mathbf{W}^* \, \mathbf{R}^* \otimes \mathbf{I}_M \, \right] \\
& + 2 \left[\, \mathbf{R}^T \, \mathbf{W}^T \otimes \mathbf{I}_M \, \right] \left[\, \mathbf{Q}^{-T} \otimes \mathbf{Q}^{-1} \, \right] \mathbf{P}_{\mathrm{off}} \left[\, \mathbf{W}^* \, \mathbf{R}^* \otimes \mathbf{I}_M \, \right]
\end{aligned} \tag{75}
$$

$$
\begin{aligned}
\widetilde{\mathbf{C}}_{\mathbf{W}} \left(\mathbf{W}; \mathbf{R} \right) = {} & \\
& 2 \left[\, \mathbf{R} \, \mathbf{W}^H \otimes \mathbf{I}_M \, \right] \left(\left[\, \mathbf{Q}^{-1} \otimes \mathbf{Q}^{-T} \, \right] - \left[\, \bar{\mathbf{Q}}^{-1} \otimes \bar{\mathbf{Q}}^{-T} \, \right] \right) \mathbf{P}_{\mathrm{diag}} \, \mathbf{P}_{M \times M} \left[\, \mathbf{W}^* \, \mathbf{R}^* \otimes \mathbf{I}_M \, \right] \\
& + 2 \left[\, \mathbf{R} \, \mathbf{W}^H \otimes \mathbf{I}_M \, \right] \left[\, \mathbf{Q}^{-1} \otimes \mathbf{Q}^{-T} \, \right] \mathbf{P}_{\mathrm{off}} \, \mathbf{P}_{M \times M} \left[\, \mathbf{W}^* \, \mathbf{R}^* \otimes \mathbf{I}_M \, \right] .
\end{aligned} \tag{76}
$$

This form allows us to see clear similarities between the gradient and Hessian terms of the two cost functions $\mathcal{J}_{\text{off}}$ and $\mathcal{J}$ when comparing (69) to (71) with (74) to (76), respectively: In the gradient and the first term of the Hessian, off($\mathbf{Q}$) is replaced by $\bar{\mathbf{Q}}^{-1}$ off($\mathbf{Q}$) $\mathbf{Q}^{-1}$. Note that $\mathbf{Q}$ and $\bar{\mathbf{Q}}$ are both positive definite matrices. The last terms of the Hessian terms also show a very similar structure. The only difference is the additional term $\mathbf{Q}^{-T} \otimes \mathbf{Q}^{-1}$ in the Hessian of $\mathcal{J}$, which is also positive definite. The second term of $\widetilde{\mathbf{H}}_{\mathbf{W}}$ and the first term of $\widetilde{\mathbf{C}}_{\mathbf{W}}$ have no corresponding terms in $\widehat{\mathbf{H}}_{\mathbf{W}}$ and $\widehat{\mathbf{C}}_{\mathbf{W}}$.

2.7. Discussion of critical points. The critical points are defined where the gradient of $\mathcal{J}$ becomes zero. From (74) we see that $\widehat{\mathbf{D}}_{\mathbf{W}}$ becomes zero if off($\mathbf{Q}$) $\equiv \mathbf{0}$ and consequently $\mathbf{Q} \equiv \bar{\mathbf{Q}}$. Since we can reformulate, similar to (72),

$$\left[\mathbf{Q}^{-T} \otimes \mathbf{Q}^{-1} \right] - \left[\bar{\mathbf{Q}}^{-T} \otimes \bar{\mathbf{Q}}^{-1} \right] = - \left[\mathbf{Q}^{-T} \otimes \mathbf{Q}^{-1} \right] \text{off}(\mathbf{Q}^T \otimes \mathbf{Q}) \left[\bar{\mathbf{Q}}^{-T} \otimes \bar{\mathbf{Q}}^{-1} \right] \tag{77}$$

and the Kronecker product of two diagonal matrices is a diagonal matrix, the term off($\bar{\mathbf{Q}}^T \otimes \bar{\mathbf{Q}}$) becomes zero. Hence, the Hessian terms of $\mathcal{J}$ at the critical points are

$$\widetilde{\mathbf{H}}_{\mathbf{W}}(\mathbf{W}; \mathbf{R}) = 2 \left[\mathbf{R}^T \mathbf{W}^T \otimes \mathbf{I}_M \right] \left[\bar{\mathbf{Q}}^{-T} \otimes \bar{\mathbf{Q}}^{-1} \right] \mathbf{P}_{\text{off}} \left[\mathbf{W}^* \mathbf{R}^* \otimes \mathbf{I}_M \right] \tag{78}$$

$$\widetilde{\mathbf{C}}_{\mathbf{W}}(\mathbf{W}; \mathbf{R}) = 2 \left[\mathbf{R} \, \mathbf{W}^H \otimes \mathbf{I}_M \right] \left[\bar{\mathbf{Q}}^{-1} \otimes \bar{\mathbf{Q}}^{-T} \right] \mathbf{P}_{\text{off}} \, \mathbf{P}_{M \times M} \left[\mathbf{W}^* \mathbf{R}^* \otimes \mathbf{I}_M \right]. \tag{79}$$

When comparing (78) and (79) with (70) and (71), respectively, we make the interesting discovery that, besides the diagonal matrix $\bar{\mathbf{Q}}^{-T} \otimes \bar{\mathbf{Q}}^{-1}$ and a scaling factor, the Hessian of $\mathcal{J}$ and $\mathcal{J}_{\text{off}}$ at the critical points have an identical structure.
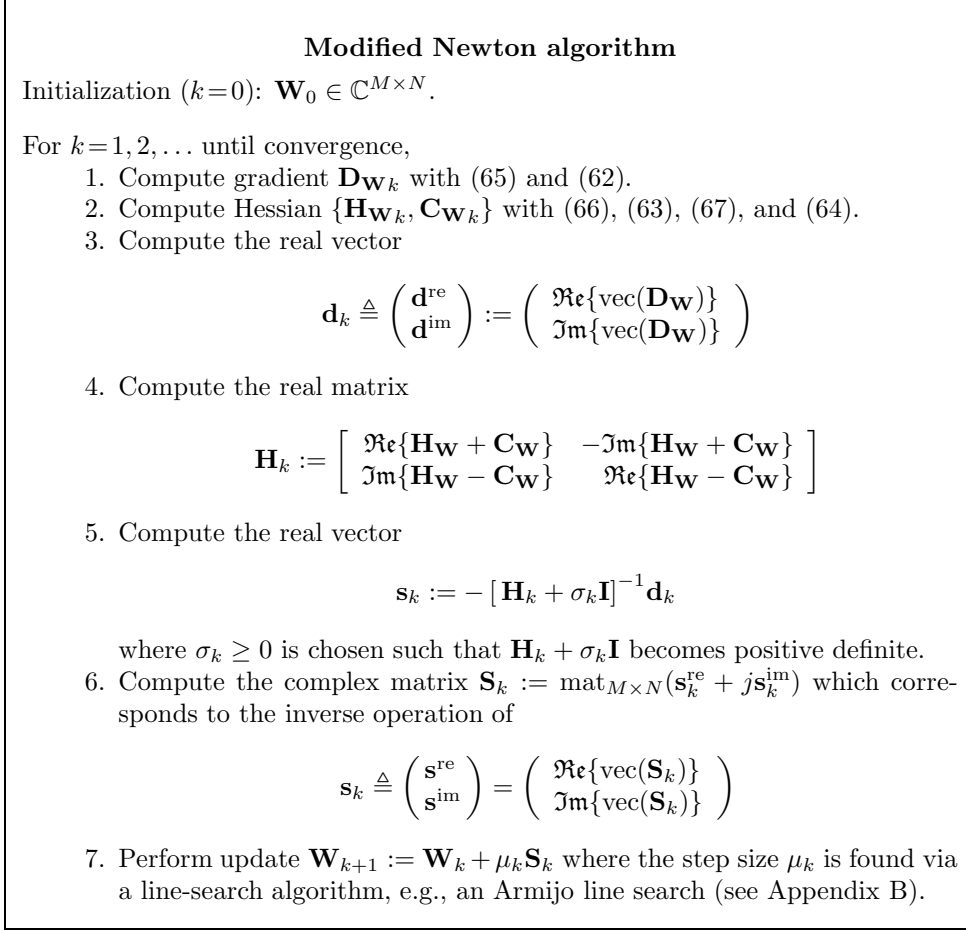
3. The Newton Algorithm. Once we have derived the gradient and Hessian of our cost function, we can now formulate the Newton algorithm. The Newton update at iteration k can be written as

$$\mathbf{W}_{k+1} = \mathbf{W}_k + \mu_k \mathbf{S}_k \tag{80}$$

where $\mathbf{S}_k$ is the *search direction* and μ_k is the *step size* of the kth update. The individual steps of the Newton algorithm are given in Fig. 3.1. Since our cost function is non-quadratic, we use a modified Newton step that incorporates an Armijo line search. In the vicinity of the minimum the update will approach the pure Newton step.

The described Newton algorithm is built upon a vector-form Newton algorithm, where the complex coefficients of $\mathbf{W}_k$ are arranged in a length $2MN$ real vector $\mathbf{w}_k$ as defined in (14). Hence, the vector $\mathbf{d}_k$ and the matrix $\mathbf{H}_k$ are the gradient and Hessian of the vector form of the Taylor series expansion of $\mathcal{J}(\mathbf{w}_k)$, see (13).

Since our cost function is not quadratic, the Hessian $\mathbf{H}_k$ can have negative eigenvalues. By choosing σ_k such that $\mathbf{H}_k + \sigma_k \mathbf{I}$ becomes positive definite, the inverse $\left[\mathbf{H}_k + \sigma_k \mathbf{I} \right]^{-1}$ will be positive definite as well. This will ensure that $\mathbf{s}_k$ and $\mathbf{S}_k$ will point to a descent direction, just like the negative gradient $-\mathbf{d}_k$ does. Hence, σ_k must be chosen larger than $-\lambda_{min}(\mathbf{H}_k)$ if $\mathbf{H}_k$ has non-positive eigenvalues, where λ_{min} is the smallest eigenvalue of $\mathbf{H}_k$. In the vicinity of a local minimum, σ_k will become zero and the update will approach the pure Newton update for $\mu_k = 1$. On the other hand, if σ_k is chosen very large, the search direction $\mathbf{s}_k$ will become close to the direction of the negative gradient. For efficiency reasons, $\mathbf{H}_k$ is often regularized via a modified Cholesky factorization method [2, Sec. 1.4]. Therefore the described modified Newton algorithm should be understood more as a prototype algorithm. The step size μ_k is

---

Modified Newton algorithm

Initialization ($k\!=\!0$): $\mathbf{W}_0 \in \mathbb{C}^{M \times N}$.

For $k\!=\!1, 2, \ldots$ until convergence,

    1. Compute gradient $\mathbf{D}_{\mathbf{W}k}$ with (65) and (62).

    2. Compute Hessian $\{\mathbf{H}_{\mathbf{W}k}, \mathbf{C}_{\mathbf{W}k}\}$ with (66), (63), (67), and (64).

    3. Compute the real vector

$$\mathbf{d}_k \triangleq \begin{pmatrix} \mathbf{d}^{\mathrm{re}} \\ \mathbf{d}^{\mathrm{im}} \end{pmatrix} := \begin{pmatrix} \mathfrak{Re}\{\mathrm{vec}(\mathbf{D_W})\} \\ \mathfrak{Im}\{\mathrm{vec}(\mathbf{D_W})\} \end{pmatrix}$$

    4. Compute the real matrix

$$\mathbf{H}_k := \begin{bmatrix} \mathfrak{Re}\{\mathbf{H_W} + \mathbf{C_W}\} & -\mathfrak{Im}\{\mathbf{H_W} + \mathbf{C_W}\} \\ \mathfrak{Im}\{\mathbf{H_W} - \mathbf{C_W}\} & \mathfrak{Re}\{\mathbf{H_W} - \mathbf{C_W}\} \end{bmatrix}$$

    5. Compute the real vector

$$\mathbf{s}_k := -\left[\mathbf{H}_k + \sigma_k \mathbf{I}\right]^{-1} \mathbf{d}_k$$

    where $\sigma_k \geq 0$ is chosen such that $\mathbf{H}_k + \sigma_k \mathbf{I}$ becomes positive definite.

    6. Compute the complex matrix $\mathbf{S}_k := \mathrm{mat}_{M \times N}(\mathbf{s}_k^{\mathrm{re}} + j\mathbf{s}_k^{\mathrm{im}})$ which corresponds to the inverse operation of

$$\mathbf{s}_k \triangleq \begin{pmatrix} \mathbf{s}^{\mathrm{re}} \\ \mathbf{s}^{\mathrm{im}} \end{pmatrix} = \begin{pmatrix} \mathfrak{Re}\{\mathrm{vec}(\mathbf{S}_k)\} \\ \mathfrak{Im}\{\mathrm{vec}(\mathbf{S}_k)\} \end{pmatrix}$$

    7. Perform update $\mathbf{W}_{k+1} := \mathbf{W}_k + \mu_k \mathbf{S}_k$ where the step size μ_k is found via a line-search algorithm, e.g., an Armijo line search (see Appendix B).

Fig. 3.1. *Modified Newton algorithm.*

obtained from a line-search step, e.g., an *Armijo line-search method* [2,6,11,15], which guarantees that $\mathcal{J}(\mathbf{W}_{k+1}) \leq \mathcal{J}(\mathbf{W}_k)$ and μ_k is chosen *not-to-small*. The reason to include a variable step size μ_k into the Newton update is motivated by the fact that in our case $\mathcal{J}$ is a non-quadratic cost function. In this case the Newton step, $\mu_k\!=\!1$, often overshoots the local minimum, even if the Hessian $\mathbf{H}_k$ is positive definite. Close to a minimum, where the cost function can be approximated by a quadratic curvature, the Armijo line-search method will often set μ_k automatically to one.

4. Simulation example. We give now a simulation example where the performance between the Newton algorithm and a steepest-descent algorithm are compared. This simulation is set up such that a perfect joint diagonalization is possible. The chosen parameters are $M = 3$ and $N = 5$. For each simulation trial we generate a random complex matrix $\mathbf{A} \in \mathbb{C}^{5 \times 3}$, such that $\mathbf{A}^H \mathbf{A} = \mathbf{I}_3$. Then we generate for each trial a set of $P = 15$ correlation matrices $\{\mathbf{R}_p\}_{p=1}^{15} = \{\mathbf{A}\, \mathbf{\Lambda}_p\, \mathbf{A}^H\}_{p=1}^{15}$ where each $\mathbf{\Lambda}_p \in \mathbb{R}^{3 \times 3}$ is a diagonal matrix whose elements are randomly chosen from a uniform distribution between 0.1 and 1. Hence, each $\mathbf{\Lambda}_p$ is positive definite, and each $\mathbf{R}_p$ is positive semi-definite and has rank 3. Fig. 4.1 compares the performance between
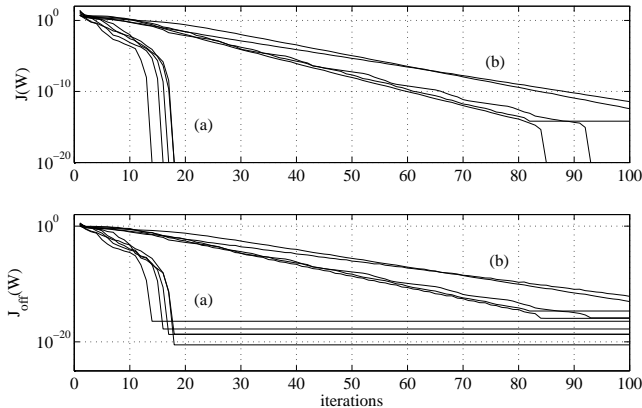
FIG. 4.1. *Learning curves of $\mathcal{J}(\mathbf{W}_k)$ (top) and $\mathcal{J}_{off}(\mathbf{W}_k)$ (bottom) for five independent simulations using (a) modified Newton method and (b) steepest-descent algorithm.*

the Newton algorithm, using Armijo line searches, and a gradient type update. The top curve shows the performance of the cost function $\mathcal{J}$, defined in (4), the bottom curve shows how the cost function $\mathcal{J}_{\text{off}}$, given in (68), behaves. For this simulation we pre-multiply $\mathbf{W}_k$ after every iteration with a diagonal matrix, such that the rows of $\mathbf{W}_k$ have unit norm, i.e., $\text{diag}(\mathbf{W}_k \mathbf{W}_k^H) = \mathbf{I}_3$. This normalization step does not affect the value of the cost function $\mathcal{J}$, as $\mathcal{J}$ is scale invariant to such an operation, however, it is important for a meaningful interpretation of $\mathcal{J}_{\text{off}}$. Without this normalization a small value of $\mathcal{J}_{\text{off}}$ could also be caused by $\| \mathbf{W}_k \|_F \ll 1$, which would lead to a misleading performance interpretation. From the simulation curves it is clearly seen that once the Newton algorithm approaches the vicinity of a minimum, it reveals a *superlinear convergence* and attains the minimum within a few steps.

5. Conclusion. The problem of joint-approximate diagonalization of a set of positive definite matrices has become of great interest in blind signal separation applications. Most algorithms known for the joint-diagonalization task impose some constraints on the diagonalization matrix $\mathbf{W}$, namely that $\mathbf{W}$ needs to be (i) real, (ii) unitary, or (iii) square. We have derived a Newton algorithm for this problem which has none of these restrictions. We allow the diagonalization matrix $\mathbf{W}$ to be complex, non-unitary, and even rectangular.

The most general case where the diagonalization matrix $\mathbf{W}$ can be rectangular, instead of being square, is of particular interest in blind signal separation. This scenario occurs when access to more sensor signals than source signals is available. In this case the correlation matrices $\mathbf{R}_p$ are no longer positive definite, they only will be positive semi-definite. Algorithms that use the same cost function as given in (4), but constraint $\mathbf{W}$ to be square, require that $\mathbf{R}_p$ are positive definite, otherwise $\det(\mathbf{W} \mathbf{R}_p \mathbf{W}^H)$ becomes zero. Since our algorithm can also update a rectangular $M \times N$ matrix $\mathbf{W}$, where $M \leq N$, we impose a much weaker constraint, namely that the product $\mathbf{W} \mathbf{R}_p \mathbf{W}^H$ needs to be positive definite and not $\mathbf{R}_p$. For a given set of $\{\mathbf{R}_p\}$ we can simply achive this by reducing M, the number of rows of $\mathbf{W}$, until $\{\mathbf{W} \mathbf{R}_p \mathbf{W}^H\}$ has full rank for all p.

A major contribution of this paper is the derivation of the Hessian in closed form for every $\mathbf{W}$ and not only at the critical points. It turned out that the matrix form of the Taylor-series expansion (12), as given by Manton in [13], has provided

the fundament in this derivation. This form preserves the matrix structure of the underlying problem which allows a compact-form representation of the gradient and Hessian through matrix- and Kronecker products. Finally, we have shown that there exists a close similarity between the gradient and Hessian of two commonly used cost functions for the joint-diagonalization problem.

Appendix A. Useful relations for deriving the gradient and Hessian of a matrix-valued cost function.

The following equalities were very useful for the derivation of the gradient and Hessian. Some basic relations are

$$\| \mathbf{A} \|_F^2 = \mathrm{tr}(\mathbf{A}\mathbf{A}^H) \tag{81}$$

$$\mathrm{tr}\,(\mathbf{A}\mathbf{B}) = \mathrm{tr}\,(\mathbf{B}\mathbf{A}) \tag{82}$$

$$\mathrm{tr}\left(\mathbf{A}^H\right) = \mathrm{tr}\left(\mathbf{A}^*\right) = \left(\mathrm{tr}\left(\mathbf{A}\right)\right)^* . \tag{83}$$

Furthermore, we have some useful equalities with the vec(.) operation and Kronecker product [3] with $\mathbf{Z} \in \mathbb{C}^{M \times N}$:

$$\mathrm{vec}(\,\mathbf{Z}^T\,) = \mathbf{P}_{M \times N}\,\mathrm{vec}(\,\mathbf{Z}\,) \tag{84}$$

$$\mathrm{tr}(\mathbf{Z}^H\mathbf{A}) = \mathrm{vec}(\,\mathbf{Z}\,)^H\,\mathrm{vec}(\,\mathbf{A}\,) \tag{85}$$

$$\mathrm{tr}\,(\mathbf{Z}\mathbf{A}) = \mathrm{vec}(\,\mathbf{Z}^T\,)^T\,\mathrm{vec}(\,\mathbf{A}\,) \tag{86}$$

$$= \mathrm{vec}(\,\mathbf{Z}\,)^T\,\mathbf{P}_{M \times N}^T\,\mathrm{vec}(\,\mathbf{A}\,) \tag{87}$$

$$(\mathbf{A} \otimes \mathbf{B})^T = \mathbf{A}^T \otimes \mathbf{B}^T \tag{88}$$

$$(\mathbf{A} \otimes \mathbf{B})^H = \mathbf{A}^H \otimes \mathbf{B}^H \tag{89}$$

$$(\mathbf{A} \otimes \mathbf{B})^{-1} = \mathbf{A}^{-1} \otimes \mathbf{B}^{-1} \tag{90}$$

$$(\,\mathbf{A}\,\mathbf{B} \otimes \mathbf{C}\,\mathbf{D}\,) = (\,\mathbf{A} \otimes \mathbf{C}\,)(\,\mathbf{B} \otimes \mathbf{D}\,) \tag{91}$$

$$\mathbf{A}^{P \times Q} \otimes \mathbf{B}^{R \times S} = \mathbf{P}_{P \times R}\,(\,\mathbf{B} \otimes \mathbf{A}\,)\,\mathbf{P}_{S \times Q} \tag{92}$$

$$\mathrm{vec}(\,\mathbf{A}\,\mathbf{Z}\,\mathbf{B}\,) = (\,\mathbf{B}^T \otimes \mathbf{A}\,)\,\mathrm{vec}(\,\mathbf{Z}\,) \tag{93}$$

where the permutation matrix $\mathbf{P}_{M \times N}$ is uniquely defined with (84). Sometimes $\mathbf{P}_{M \times N}$ is called the *commutation matrix* [12]. Since $\mathbf{P}_{M \times N}$ is a permutation matrix, $\mathbf{P}_{M \times N} = \mathbf{P}_{N \times M}^T = \mathbf{P}_{N \times M}^{-1}$. For the special case where $M = N$, the commutation matrix is *involutary*, $\mathbf{P}_{M \times M}^2 = \mathbf{I}$, as $\mathbf{P}_{M \times M}^T = \mathbf{P}_{M \times M}$ is symmetric. See [3, 12] for a thorough list of properties of Kronecker products.

The following relations, where $\mathbf{Z}_1, \mathbf{Z}_2 \in \mathbb{C}^{M \times N}$ and the argument of $\mathrm{tr}\,(.)$ is a square matrix, have been proven to be very useful in the derivation of the Hessian:

$$\mathrm{tr}\,(\mathbf{Z}_1\,\mathbf{A}\,\mathbf{Z}_2\,\mathbf{B}) = \mathrm{vec}(\,\mathbf{Z}_1\,)^T\,\mathbf{P}_{M \times N}^T\,(\,\mathbf{B}^T \otimes \mathbf{A}\,)\,\mathrm{vec}(\,\mathbf{Z}_2\,) \tag{94}$$

$$\mathrm{tr}\left(\mathbf{Z}_1^H\,\mathbf{A}\,\mathbf{Z}_2\,\mathbf{B}\right) = \mathrm{vec}(\,\mathbf{Z}_1\,)^H\,(\,\mathbf{B}^T \otimes \mathbf{A}\,)\,\mathrm{vec}(\,\mathbf{Z}_2\,) \tag{95}$$

$$\mathrm{tr}\left(\mathbf{Z}_1^H\,\mathbf{A}\,\mathbf{Z}_2^H\,\mathbf{B}\right) = \mathrm{vec}(\,\mathbf{Z}_1\,)^H\,(\,\mathbf{B}^T \otimes \mathbf{A}\,)\,\mathbf{P}_{M \times N}\,\mathrm{vec}(\,\mathbf{Z}_2\,)^* . \tag{96}$$

Eq. (95) can be derived with (85) and (93). Eq. (94) and (96) can be derived with (85), (93), (87), and (83).

Appendix B. Armijo rule for matrix form.

The Armijo rule for choosing a step size μ_k at the kth iteration is defined as $\mu_k = \mu_0 \, \gamma^m$ where m is the first nonnegative integer that fulfills

$$\mathcal{J}(\mathbf{W}_k) - \mathcal{J}(\mathbf{W}_k + \mu_0 \, \gamma^m \, \mathbf{S}_k) \geq -\eta \, \mu_0 \, \gamma^m \, \mathfrak{Re}\{\langle \mathbf{S}_k, \mathbf{D}_{\mathbf{W}\,k} \rangle\} \tag{97}$$

The search direction and the gradient of $\mathcal{J}$ at $\mathbf{W}_k$ are denoted as $\mathbf{S}_k$ and $\mathbf{D}_{\mathbf{W}\,k}$, respectively, and $\langle \mathbf{S}_k, \mathbf{D}_{\mathbf{W}\,k} \rangle \triangleq \mathrm{tr}\left(\mathbf{S}_k^H \mathbf{D}_{\mathbf{W}\,k}\right)$ defines an inner product between $\mathbf{S}_k$ and $\mathbf{D}_{\mathbf{W}\,k}$.

REFERENCES

[1] A. Belouchrani, K. Abed-Meraim, J.-F. Cardoso, and E. Moulines, *A blind source separation technique using second-order statistics*, IEEE Transactions on Signal Processing, 45 (1997), pp. 434–444.

[2] D. P. Bertsekas, *Nonlinear Programming*, Athena Scientific, 2nd ed., 1999.

[3] J. W. Brewer, *Kronecker products and matrix calculus in system theory*, IEEE Transactions on Circuits and Systems, 25 (1978), pp. 772–781.

[4] J.-F. Cardoso and A. Souloumiac, *Blind beamforming for non Gaussian signals*, IEE Proceedings-F, 140 (1993), pp. 362–370.

[5] J. F. Cardoso and A. Souloumiac, *Jacobi angles for simultaneous diagonalization*, SIAM Journal on Matrix Analysis and Applications, 17 (1996), pp. 161–164.

[6] R. Fletcher, *Practical Methods of Optimization*, John Wiley & Sons, 2nd ed., 1987.

[7] B. Flury, *Common principal components in k groups*, J. Amer. Statist. Assoc., 79 (1984), pp. 892–897.

[8] B. Flury and W. Gautschi, *An algorithm for simultaneous orthogonal transformation of several positive definite symmetric matrices to nearly diagonal form*, SIAM Journal on Scientific and Statistical Computing, 7 (1986), pp. 169–184.

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

[10] M. Joho and K. Rahbar, *Joint diagonalization of correlation matrices by using Newton methods with application to blind signal separation*, in IEEE Sensor Array and Multichannel Signal Processing Workshop, Rosslyn, VA, August 4–6, 2002, pp. 403–407.

[11] D. G. Luenberger, *Linear and Nonlinear Programming*, Addison-Wesley, 2nd ed., 1989.

[12] J. R. Magnus and H. Neudecker, *Matrix Differential Calculus with Applications in Statistics and Econometrics*, John Wiley & Sons, 2nd ed., 1999.

[13] J. H. Manton, *Optimisation algorithms exploiting unitary constraints*, IEEE Transactions on Signal Processing, 50 (2002), pp. 635–650.

[14] D. T. Pham, *Joint approximate diagonalization of positive definite hermitian matrices*, SIAM Journal on Matrix Analysis and Applications, 22 (2001), pp. 1136–1152.

[15] E. Polak, *Optimization: Algorithms and Consistent Approximations*, Springer Verlag, 1997.

[16] R. Vollgraf and K. Obermayer, *Quadratic optimization for simultaneous matrix diagonalization*, IEEE Transactions on Signal Processing, 54 (2006), pp. 3270–3278.

[17] A. Yeredor, *Non-orthogonal joint diagonalization in the least-squares sense with application in blind source separation*, IEEE Transactions on Signal Processing, 50 (2002), pp. 1545–1553.

[18] ——, *On using exact joint diagonalization for noniterative approximate joint diagonalization*, IEEE Signal Processing Letters, 12 (2005), pp. 645–648.

[19] A. Yeredor, A. Ziehe, and K. R. Müller, *Approximate joint diagonalization using a natural-gradient approach*, in Proc. International Conference on Independent Component Analysis and Blind Signal Separation (ICA), Granada, Spain, September 22–24, 2004, pp. 89–96.

[20] A. Ziehe, P. Laskov, and K.-R. Müller, *A fast algorithm for joint diagonalization with non-orthogonal transformations and its application to blind source separation*, Journal of Machine Learning Research, (2004), pp. 777–800.