A Unified Algorithmic Framework for Distributed Adaptive Signal and Feature Fusion Problems — Part II: Convergence Properties

Cem Ates Musluoglu, Charles Hovine, and Alexander Bertrand, *Senior Member, IEEE*

***Abstract*—This paper studies the convergence conditions and properties of the distributed adaptive signal fusion (DASF) algorithm, the framework itself having been introduced in a 'Part I' companion paper. The DASF algorithm can be used to solve linear signal and feature fusion optimization problems in a distributed fashion, and is in particular well-suited for solving spatial filtering optimization problems encountered in wireless sensor networks. The convergence conditions and results are provided along with rigorous proofs and analyses, as well as various example problems to which they apply. Additionally, we describe procedures that can be added to the DASF algorithm to ensure convergence in specific cases where some of the technical convergence conditions are not satisfied.**

***Index Terms*—Distributed optimization, distributed signal processing, spatial filtering, signal fusion, feature fusion, wireless sensor networks.**

I. Introduction

The Distributed Adaptive Signal Fusion (DASF) algorithm introduced in [1] can be used to solve a wide range of spatial filtering and signal fusion problems in a distributed fashion, e.g., within a wireless sensor network (WSN). Examples of such problems include minimum mean square error estimation, discriminant analysis based on generalized eigenvalue decomposition [3], canonical correlation analysis [4], [5], minimum variance beamforming [6], etc. The DASF algorithm is designed to cope with the typical bandwidth or energy limitations of WSNs.

A typical spatial filtering or signal fusion problem in a WSN involves optimizing a cost function depending on the sensor data collected by each node in the network. Contrarily to a centralized procedure requiring the sensor data of each node to be aggregated at a fusion center, the DASF algorithm requires the nodes to share only compressed data between each other. This data is then used to locally build a compressed version of the global optimization problem within a node at every iteration. As a result, any solver for the global (centralized) problem can be used to solve the local problems at each node within the DASF iterations.

In this paper, we provide a set of sufficient conditions for convergence and optimality of the DASF algorithm, based on which we can show that the DASF algorithm converges to the centralized solution of the problem despite the compression, as if all the raw sensor data were centrally available. The technical conditions required for convergence are akin to the well-known linear independence constraint qualifications in the optimization literature, which in the case of DASF lead to an upper bound on the number of constraints the global (centralized) problem is allowed to have. Furthermore, since the local problems in each node are compressed versions of the global problem, we assume that the local problems satisfy the same assumptions as the global problem as outlined in [1], which is generally the case as they are directly inherited. Finally, we impose a condition on the finiteness of the number of possible solutions that are achievable by the solver used to solve the local problems in each node. We will see that these conditions are often satisfied for spatial filtering and signal fusion problems in practical scenarios. Furthermore, we provide several examples and illustrations on how the convergence conditions can be checked either a priori or during operation of the algorithm. We will also show how the insights obtained from the convergence analysis can be used to design new strategies to enforce convergence in cases where a violation of the convergence conditions is detected.

The outline of the paper is as follows. After a short review of the DASF framework in Section II, we study the convergence and optimality guarantees of the DASF algorithm in Section III. In particular, we show that under some technical conditions, accumulation points of the sequence of points produced by the algorithm are also fixed points, and that fixed points are solutions of the centralized problem. Examples of typical spatial filtering and signal fusion problems, such as minimum mean square error or minimum variance beamforming, and how the convergence conditions apply to these cases are discussed in Section IV. Finally, in the contrived cases where some of the technical requirements are violated, we describe methods to still achieve convergence for the DASF algorithm in Section V. Conclusions are drawn in Section VI.

This project has received funding from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme (grant agreement No 802895). The authors also acknowledge the financial support of the FWO (Research Foundation Flanders) for project G081722N, and the Flemish Government (AI Research Program).

C.A. Musluoglu, C. Hovine and A. Bertrand are with KU Leuven, Department of Electrical Engineering (ESAT), Stadius Center for Dynamical Systems, Signal Processing and Data Analytics, Kasteelpark Arenberg 10, box 2446, 3001 Leuven, Belgium and with Leuven.AI - KU Leuven institute for AI. e-mail: cemates.musluoglu, charles.hovine, alexander.bertrand @esat.kuleuven.be

A companion paper submitted together with this paper is provided in [1]. This paper has supplementary downloadable material [2] available at http://ieeexplore.ieee.org, provided by the authors. The material includes the steps we follow to prove Lemma 3. This material is 397 KB in size.

Digital Object Identifier: 10.1109/TSP.2023.3275273

Notation: Uppercase letters are used to represent matrices and sets, the latter in calligraphic script, while scalars, scalar-valued functions and vectors are represented by lowercase letters, the latter in bold. We use the notation χ_q^i to refer to a certain mathematical object χ (such as a matrix, set, etc.) at node q and iteration i. The notation $\left(\chi^i\right)_{i\in\mathcal{I}}$ refers to a sequence of elements χ^i over every index i in the ordered index set $\mathcal{I}$. If it is clear from the context (often in the case where i is over all natural numbers), we omit the index set $\mathcal{I}$ and simply write $\left(\chi^i\right)_i$. A similar notation $\{\chi^i\}_{i\in\mathcal{I}}$ is used for non-ordered sets. Additionally, I_Q denotes the $Q\times Q$ identity matrix, $\mathbb{E}[\cdot]$ the expectation operator, $\text{tr}(\cdot)$ the trace operator, *BlkDiag* $(\cdot)$ the operator that creates a block-diagonal matrix from its arguments and $|\cdot|$ the cardinality of a set.

II. Review of the DASF Framework

In this section, we briefly restate the scope and operation of the DASF framework, which was extensively described in [1]. We do this for completeness and with the purpose to re-introduce key equations and introduce some new ones that will be needed in the convergence analysis. The reader who is not yet familiar with the DASF framework is encouraged to read [1] first, since the review in this section skips several useful insights and details.

A. Problem Description and Assumptions

We consider a WSN consisting of a set of K nodes $\mathcal{K}=\{1,\dots,K\}$ interconnected according to a network graph $\mathcal{G}$, which contains edges between nodes that are able to share data with each other. Each node k collects observations of an M_k-channel sensor signal $\mathbf{y}_k$. We define the network-wide sensor signal $\mathbf{y}\in\mathbb{R}^M$ as

$$\mathbf{y}(t)=[\mathbf{y}_1^T(t),\dots,\mathbf{y}_K^T(t)]^T, \tag{1}$$

where t denotes the time index and $M=\sum_{k\in\mathcal{K}}M_k$. $\mathbf{y}$ is assumed to be (short-term) stationary and ergodic, such that its statistical properties can be properly estimated given a sufficiently large number of samples at different time instances, e.g., $\{\mathbf{y}(\tau)\}_{\tau=0}^{N-1}$, where N denotes the number of time samples. Our objective is to find a linear filter $X\in\mathbb{R}^{M\times Q}$ that optimizes in some sense the output signal of the linear signal fusion $X^T\mathbf{y}$. More specifically, we consider problems of the following form:

$$\begin{aligned}\mathbb{P}:\underset{X\in\mathbb{R}^{M\times Q}}{\text{minimize}}\quad & \varphi\left(X^T\mathbf{y}(t),X^TB\right)\\ \text{subject to}\quad & \eta_j\left(X^T\mathbf{y}(t),X^TB\right)\leq 0\ \ \forall j\in\mathcal{J}_I,\\ & \eta_j\left(X^T\mathbf{y}(t),X^TB\right)=0\ \ \forall j\in\mathcal{J}_E,\end{aligned} \tag{2}$$

where φ and η_j's are differentiable real- and scalar-valued functions. We refer to such problems as *(distributed) signal fusion optimization* ((D)SFO) problems. In Problem (2), φ denotes the objective function and the η_j's denote the constraint functions, where the indices $j\in\mathcal{J}_I$ and $j\in\mathcal{J}_E$ correspond to inequality and equality constraints respectively. We denote the joint index set $\mathcal{J}_I\cup\mathcal{J}_E=\mathcal{J}$, such that the total number of constraints is $J=|\mathcal{J}|$.

The objective and constraint functions implicitly contain an operator that maps the stochastic variable $\mathbf{y}$ into a deterministic and time-independent quantity, such as an expectation operator, which in practice is typically replaced with an estimated quantity based on temporal averaging of N samples of $\mathbf{y}$.

B is an additional deterministic parameter of the problem. Similarly to (1), it can be partitioned as

$$B=[B_1^T,\dots,B_K^T]^T\in\mathbb{R}^{M\times L} \tag{3}$$

where L is some problem-dependent constant. It is assumed that each node k has a local access to its corresponding block B_k (i.e. they can be used in computations without being first requested from another node). In contrast to the signal $\mathbf{y}$, the matrix B is deterministic and independent of time. As noted in [1], Problem (2) can be generalized to more than one variable X, stochastic signal $\mathbf{y}$ and deterministic term B. Note that such a generalization covers deterministic quadratic terms in the form X^TAX, since two linear terms $X^TB^{(1)}$ and $X^TB^{(2)}$ can be combined into $(X^TB^{(1)})\cdot(X^TB^{(2)})^T$, with $A=B^{(1)}B^{(2)T}$.

As the actual optimization variable is X, and by removing the time-dependence of the problem by stationarity of the random signal $\mathbf{y}$, we define the functions f and h_j, $j\in\mathcal{J}$, which express the objective and constraints as a function of X only:

$$f(X)\triangleq\varphi\left(X^T\mathbf{y}(t),X^TB\right), \tag{4}$$

$$h_j(X)\triangleq\eta_j\left(X^T\mathbf{y}(t),X^TB\right),\ \forall j\in\mathcal{J}, \tag{5}$$

which are assumed to be continuously differentiable with respect to the variable X over their respective domains. This allows us to write (2) compactly as

$$\begin{aligned}\mathbb{P}:\underset{X\in\mathbb{R}^{M\times Q}}{\text{minimize}}\quad & f(X)\\ \text{subject to}\quad & h_j(X)\leq 0\ \ \forall j\in\mathcal{J}_I,\\ & h_j(X)=0\ \ \forall j\in\mathcal{J}_E.\end{aligned} \tag{6}$$

Furthermore, we denote the constraint set of (2) or (6) as $\mathcal{S}$, the complete solution set as $\mathcal{X}^*$ and a single solution as X^*, i.e., $X^*\in\mathcal{X}^*$.

In order to guarantee the theoretical convergence of the DASF algorithm, we restrict its application to problems satisfying the following three general assumptions[1] (in Section IV, we explain how these assumptions can be checked in practical examples):

Assumption 1. *The targeted instance of Problem (2) or (6) is well-posed, in the sense that the solution set is not empty and varies continuously with a change in the parameters of the problem.*

The notion of (generalized Hadamard) well-posedness we require is based on [7], [8]. The main difference is that we require the map from the space of inputs of the problem to the solution space to be continuous instead of upper semicontinuous, which is required for the convergence proof. We formally define the continuity of this map in Section III. Even though this condition might seem restrictive, it applies

[1] Throughout this text, if assumptions or conditions are labeled as "**Xa**", it implies that this assumption/condition can be replaced with a different assumption/condition "**Xb**". When we only mention the label "**X**", we refer to either of the two.

to many practical instances of the problems of interest (see Section IV).

Assumption 2. *The linear independence constraint qualifications (LICQ) [9] hold at the solutions of Problem (2) or (6), i.e., the solutions satisfy the Karush-Kuhn-Tucker (KKT) conditions.*

If X^* is a solution of Problem (2) or (6), Assumption 2 implies that, for $j \in \mathcal{J}^*$, the gradients $\nabla_X h_j(X^*)$ are linearly independent[2], where $\mathcal{J}^* \subseteq \mathcal{J}$ is the set of all indices j satisfying $h_j(X^*) = 0$. If the problem is unconstrained, we have $\nabla_X f(X^*) = 0$.

Assumption 3a. *f has compact sublevel sets in $\mathcal{S}$, i.e., for all $m \in \mathbb{R}$, $\{X \in \mathcal{S} \mid f(X) \leq m\}$ is compact, i.e., closed and bounded.*

It is noted here that this assumption can be further relaxed to an alternative Assumption 3b presented below. This relaxed version only requires that the DASF algorithm's initialization point is in a compact sublevel set, in which case not all sublevel sets of f in $\mathcal{S}$ should be compact.

Assumption 3b. *The sublevel set of f $\{X \in \mathcal{S} \mid f(X) \leq f(X^0)\}$ corresponding to X^0 is compact.*

As will be shown in Section III-A, Assumption 3a or 3b is needed to ensure that the elements of the sequence $(X^i)_i$ obtained from the DASF algorithm lie in a compact set, which is required to show convergence.

In the remaining of this paper, problems $\mathbb{P}$ which can be written as (2) or (6) will be satisfying the assumptions above, i.e., we will not repeat these assumptions in any of the convergence theorems. As discussed in [1], we also implicitly assume that there exists a centralized solver able to solve the targeted problem instance $\mathbb{P}$, which will be used by the DASF algorithm to solve local per-node compressed versions of the centralized problem $\mathbb{P}$.

B. The DASF Algorithm

In order to solve Problem (2) in a distributed setting, we consider a partitioning of the global variable X into local variables:

$$X = [X_1^T, \dots, X_K^T]^T, \tag{7}$$

where each local variable $X_k \in \mathbb{R}^{M_k \times Q}$ is assigned to a single node k. At any given iteration i, the nodes k all have an estimation X_k^i of their local variable X_k. At each iteration i, some node $q \in \mathcal{K}$ is selected to be the "updating node", and will solve a compressed version of Problem (2), which will be defined later in this section. We note that we choose a different updating node at each iteration. As the network graph $\mathcal{G}$ can contain loops, we prune the network into a tree, which we denote as $\mathcal{T}^i(\mathcal{G}, q)$, such that there is a unique path from any node $k \neq q$ to the updating node q (as explained in [1], the tree $\mathcal{T}^i(\mathcal{G}, q)$ should preserve all the neighbors of node q to maximize the degrees of freedom in the updating process).

At the beginning of each iteration, every node compresses N samples of its local signal $\mathbf{y}_k$ using its current estimate X_k^i of the local variable X_k to obtain the compressed $Q-$channel local signal:

$$\widehat{\mathbf{y}}_k^i(t) \triangleq X_k^{iT}\mathbf{y}_k(t) \in \mathbb{R}^Q, \tag{8}$$

while a similar operation is done to compress each B_k:

$$\widehat{B}_k^i \triangleq X_k^{iT} B_k \in \mathbb{R}^{Q \times L}. \tag{9}$$

A decentralized fuse-and-forward process is then started, in which the compressed data from all the nodes is summed on its way towards node q. This fuse-and-forward flow arises naturally when using a recursive computation rule, as explained in [1]. Indeed, the data that node k sends to its neighbor n (which is closest to the updating node q) is defined as

$$\widehat{\mathbf{y}}_{k\to n}^i(t) \triangleq X_k^{iT}\mathbf{y}_k(t) + \sum_{l \in \mathcal{N}_k \setminus \{n\}} \widehat{\mathbf{y}}_{l \to k}^i(t), \tag{10}$$

which can be computed as soon as node k receives $\widehat{\mathbf{y}}_{l\to k}$ from all its neighbors $l \in \mathcal{N}_k$, except node n. We see that this recursive expression starts at leaf nodes (nodes with only one neighbor) and extends naturally to node q, which eventually receives

$$\widehat{\mathbf{y}}_{n\to q}^i(t) = X_n^{iT}\mathbf{y}_n(t) + \sum_{k \in \mathcal{N}_n \setminus \{q\}} \widehat{\mathbf{y}}_{k\to n}^i(t) = \sum_{k \in \mathcal{B}_{nq}} \widehat{\mathbf{y}}_k^i(t) \tag{11}$$

from all its neighbors $n \in \mathcal{N}_q$. $\mathcal{B}_{nq}$ is the subgraph rooted at node q that contains $n \in \mathcal{N}_q$, i.e., the set of nodes which would be disconnected from the subgraph containing node q if we were to cut the connection between node n and q (see Figure 1). A similar fuse-and-forward process is performed for the terms (9), resulting in fused parameters $\widehat{B}_{n\to q}^i$ [1].

The data received by node q can then be structured in order to act as a local version of the global data. For this purpose, we define the $\widetilde{M}_q-$channel signal $\widetilde{\mathbf{y}}_q^i$ at node q as the local signal $\mathbf{y}_q$ stacked with the compressed signals it receives from its neighbors, given by

$$\widetilde{\mathbf{y}}_q^i(t) = [\mathbf{y}_q^T(t), \widehat{\mathbf{y}}_{n_1\to q}^{iT}(t), \dots, \widehat{\mathbf{y}}_{n_{|\mathcal{N}_q|}\to q}^{iT}(t)]^T. \tag{12}$$

The matrix $\widetilde{B}_q^i$ is defined similarly. Then, we define the local variable $\widetilde{X}_q$ at node q as

$$\widetilde{X}_q = [X_q^T, G_{n_1}^T, \dots, G_{n_{|\mathcal{N}_q|}}^T]^T \in \mathbb{R}^{\widetilde{M}_q \times Q}, \tag{13}$$

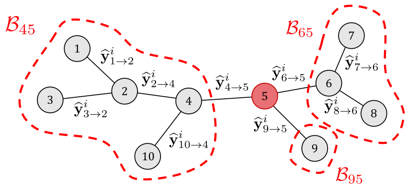
where $X_q \in \mathbb{R}^{M_q \times Q}$ and $G_n \in \mathbb{R}^{Q \times Q}$, $\forall n \in \mathcal{N}_q$. This variable acts as a spatial fusion filter on the locally available data at node q, analogous to the way X acts on $\mathbf{y}$ and B for the global problem. From (12) and (13), we can then write

$$\widetilde{X}_q^T \widetilde{\mathbf{y}}_q^i(t) = X_q^T \mathbf{y}_q(t) + \sum_{n \in \mathcal{N}_q} G_n^T \widehat{\mathbf{y}}_{n\to q}^i(t), \tag{14}$$

and replacing the signals $\widehat{\mathbf{y}}_{n\to q}^i$ by their definition in (11), we have

$$\begin{aligned} \widetilde{X}_q^T \widetilde{\mathbf{y}}_q^i(t) &= X_q^T \mathbf{y}_q(t) + \sum_{n \in \mathcal{N}_q} \sum_{k \in \mathcal{B}_{nq}} G_n^T \widehat{\mathbf{y}}_k^i(t), & (15) \\ &= X_q^T \mathbf{y}_q(t) + \sum_{n \in \mathcal{N}_q} \sum_{k \in \mathcal{B}_{nq}} (X_k^i G_n)^T \mathbf{y}_k(t). & (16) \end{aligned}$$

[2] A set of matrices $\{A_j\}_{j\in\mathcal{J}}$ is linearly independent when $\sum_{j\in\mathcal{J}} \alpha_j A_j = 0$ is satisfied if and only if $\alpha_j = 0$, $\forall j \in \mathcal{J}$, or equivalently, when $\{\text{vec}(A_j)\}_{j\in\mathcal{J}}$ is a set of linearly independent vectors, where $\text{vec}(\cdot)$ is the vectorization operator.


$\mathcal{B}_{45}$
$\mathcal{B}_{65}$
$\mathcal{B}_{95}$
1
2
3
4
5
6
7
8
9
10
$\widehat{\mathbf{y}}^i_{1\to 2}$
$\widehat{\mathbf{y}}^i_{2\to 4}$
$\widehat{\mathbf{y}}^i_{3\to 2}$
$\widehat{\mathbf{y}}^i_{4\to 5}$
$\widehat{\mathbf{y}}^i_{6\to 5}$
$\widehat{\mathbf{y}}^i_{7\to 6}$
$\widehat{\mathbf{y}}^i_{8\to 6}$
$\widehat{\mathbf{y}}^i_{9\to 5}$
$\widehat{\mathbf{y}}^i_{10\to 4}$

$$C_5^i = \begin{bmatrix} 0 & X_1^i & 0 & 0 \\ 0 & X_2^i & 0 & 0 \\ 0 & X_3^i & 0 & 0 \\ 0 & X_4^i & 0 & 0 \\ I_{M_5} & 0 & 0 & 0 \\ 0 & 0 & X_6^i & 0 \\ 0 & 0 & X_7^i & 0 \\ 0 & 0 & X_8^i & 0 \\ 0 & 0 & 0 & X_9^i \\ 0 & X_{10}^i & 0 & 0 \end{bmatrix}$$

Fig. 1. [10] Example of a tree network where the updating node is node 5. Each neighbor of node 5 creates its own cluster containing the nodes “hidden” from node 5 behind them, shown here as $\mathcal{B}_{45}$, $\mathcal{B}_{65}$, $\mathcal{B}_{95}$. The resulting transition matrix is given by C_5^i.

A similar expression can be derived for $\widetilde{X}_q^T\widetilde{B}_q^i$.

These local counterparts of the global expressions lead us to define the local problem at node q using the previous parameterization of X, which is a compressed version of the global problem (2):

$$\begin{aligned} \underset{\widetilde{X}_q\in\mathbb{R}^{\widetilde{M}_q\times Q}}{\text{minimize}} \quad & \varphi(\widetilde{X}_q^T\widetilde{\mathbf{y}}_q^i(t),\widetilde{X}_q^T\widetilde{B}_q^i) \\ \text{subject to} \quad & \eta_j(\widetilde{X}_q^T\widetilde{\mathbf{y}}_q^i(t),\widetilde{X}_q^T\widetilde{B}_q^i)\leq 0 \;\; \forall j\in\mathcal{J}_I, \\ & \eta_j(\widetilde{X}_q^T\widetilde{\mathbf{y}}_q^i(t),\widetilde{X}_q^T\widetilde{B}_q^i) = 0 \;\; \forall j\in\mathcal{J}_E. \end{aligned} \tag{17}$$

The fact that (2) and (17) have an equivalent structure implies that the same solver can be used for both.

From (16), we observe that the local inner product $\widetilde{X}_q^T\widetilde{\mathbf{y}}_q^i(t)$ at node q is related to the network-wide inner product $X^T\mathbf{y}(t)$ if X is defined as (details in [1])

$$X = C_q^i\widetilde{X}_q, \tag{18}$$

with

$$\begin{aligned} C_q(X) &= \left[\begin{array}{c|c} \begin{matrix} 0 \\ I_{M_q} \\ 0 \end{matrix} & \Theta_{-q}(X) \end{array}\right] \in \mathbb{R}^{M\times\widetilde{M}_q}, \\ C_q^i &\triangleq C_q(X^i), \end{aligned} \tag{19}$$

where I_{M_q} is placed in the $q-$th block-row, and $\Theta_{-q}(X)$ is a block matrix with K block-rows and $|\mathcal{N}_q|$ block-columns. An example of such a matrix C_q^i is provided in Figure 1, which can be formally defined as follows. Each block-column corresponds to one of the neighbors $n\in\mathcal{N}_q$ of q, which we re-index as $m_n\in\{1,\ldots,|\mathcal{N}_q|\}$. The block at the $k-$th block-row and m_n-th block-column is then defined as

$$\Big[\Theta_{-q}(X)\Big](k,m_n) = \begin{cases} X_k & \text{if } k\in\mathcal{B}_{nq} \\ 0 & \text{otherwise.} \end{cases} \tag{20}$$

This transition matrix allows us to relate the global data or global variables with their local counterparts:

$$\widetilde{\mathbf{y}}_q^i(t) = C_q^{iT}\mathbf{y}(t),\; \widetilde{B}_q^i = C_q^{iT}B,\; X = C_q^i\widetilde{X}_q \tag{21}$$

and also write the local problem (17) in a compact way:

$$\begin{aligned} \underset{\widetilde{X}_q\in\mathbb{R}^{\widetilde{M}_q\times Q}}{\text{minimize}} \quad & f(C_q^i\widetilde{X}_q) \\ \text{subject to} \quad & h_j(C_q^i\widetilde{X}_q)\leq 0 \;\; \forall j\in\mathcal{J}_I, \\ & h_j(C_q^i\widetilde{X}_q) = 0 \;\; \forall j\in\mathcal{J}_E. \end{aligned} \tag{22}$$

Moreover, denoting the constraint set of the global problem (2) or (6) as $\mathcal{S}$ and the constraint set of the local problem (17) or (22) as $\widetilde{\mathcal{S}}_q^i$, it can be shown that (see [1, Lemma 1])

$$\widetilde{X}_q\in\widetilde{\mathcal{S}}_q^i \iff C_q^i\widetilde{X}_q\in\mathcal{S}, \tag{23}$$

i.e., a point $\widetilde{X}_q$ in the constraint set of the local problem (17) leads to a corresponding point which by definition is also in the constraint set of the global problem (2). Using this notation, we define the solution of the local problem (17) or equivalently (22) as

$$\begin{aligned} \widetilde{X}_q^{i+1} &\triangleq \underset{\widetilde{X}_q\in\widetilde{\mathcal{S}}_q^i}{\text{argmin}}\; f\left(C_q^i\widetilde{X}_q\right), \\ &= \underset{\widetilde{X}_q\in\widetilde{\mathcal{S}}_q^i}{\text{argmin}}\; \varphi\left(\widetilde{X}_q^T\widetilde{\mathbf{y}}_q^i(t),\widetilde{X}_q^T\widetilde{B}_q^i\right) \end{aligned} \tag{24}$$

Considering instances where (17) or (22) would have multiple global minima, we choose $\widetilde{X}_q^{i+1}$ as the solution minimizing the distance $||\widetilde{X}_q^{i+1}-\widetilde{X}_q^i||_F$, where

$$\widetilde{X}_q^i \triangleq [X_q^{iT}, I_Q, \ldots, I_Q]^T. \tag{25}$$

Finally, the matrices $G_{n_1}^{i+1},\ldots,G_{n_{|\mathcal{N}_q|}}^{i+1}$ obtained from the partitioning (13) of $\widetilde{X}_q^{i+1}$ need to be disseminated in the network, so that every node can update their local estimator X_k as

$$X_k^{i+1} = \begin{cases} X_q^{i+1} & \text{if } k=q \\ X_k^iG_n^{i+1} & \text{if } k\neq q,\; k\in\mathcal{B}_{nq},\; n\in\mathcal{N}_q, \end{cases} \tag{26}$$

which follows from the parameterization (18) of X.

This process is then repeated at a different node at each iteration, as summarized in Algorithm 1.

III. Convergence and Optimality

In this section, we analyze the convergence of the sequence of points $(X^i)_i$ that are generated by the DASF algorithm. Note that X^i is formed by stacking all the local X_k^i's at all nodes $k\in\mathcal{K}$, or equivalently, by using (18) which shows that $X^i = C_q^i\widetilde{X}_q^i$. We show that under some mild technical conditions (which are generally satisfied in practice), the DASF algorithm converges to a stationary point of (2), and with some additional conditions even to the globally optimal signal fusion rule X^*, i.e., $(X^i)_i\to X^*\in\mathcal{X}^*$. As a result, the fused output signal $X^{iT}\mathbf{y}(t)$, which can be computed as $\widetilde{X}_k^{iT}\widetilde{\mathbf{y}}_k^i(t)$ at any node k, will become equal to $(X^*)^T\mathbf{y}(t)$ when the sample time t is large (note that the iteration index i is linked to the time index t). This convergence is without any of the nodes knowing the full signal $\mathbf{y}$ or matrix B, and despite the compression and fusion performed at each node.

For mathematical tractability, we make abstraction of estimation errors appearing from approximating the signal statistics using a finite N, which means the proofs only hold in the asymptotic case where $N\to+\infty$. In other words, we assume that all signal statistics are perfectly estimated, which is only an approximation of the practical case where the signal statistics are (re-)estimated based on blocks of N samples. In practice, a finite N has to be used, in which case the algorithm will hover around the optimal solution due to these aforementioned estimation errors in each iteration. We note that this is also the case for the centralized equivalent of the algorithm, in case the latter has to estimate the signal statistics over finite windows, e.g., in a tracking context.

Algorithm 1: Distributed Adaptive Signal Fusion (DASF) Algorithm
Code available in [11]

```
output: $X^*$
Initialize $X^0$, $i \leftarrow 0$.
repeat
    Choose the updating node as $q \leftarrow (i \mod K) + 1$.
    1) The network $\mathcal{G}$ is pruned into a tree $\mathcal{T}^i(\mathcal{G}, q)$.
    2) Every node $k$ collects a new batch of $N$ samples
       of $\mathbf{y}_k$. All nodes compress these to $N$ samples of
       $\widehat{\mathbf{y}}_k^i$ as in (8). $\widehat{B}_k^i$ is computed using (9).
    3) The nodes sum-and-forward their compressed
       data towards node $q$ via the recursive rule (10)
       (and a similar rule for the $\widehat{B}_k^i$'s). Node $q$
       eventually receives $N$ samples of $\widehat{\mathbf{y}}_{n \to q}^i$ given in
       (11) along with $\widehat{B}_{n \to q}^i$ similarly defined, from all
       its neighbors $n \in \mathcal{N}_q$.
    at Node $q$ do
        4a) Compute $\widetilde{X}_q^{i+1}$ as the solution of (17)
            where $\widetilde{\mathbf{y}}_q^i$, $\widetilde{B}_q^i$ and $\widetilde{X}_q^i$ are defined in (12) and
            (13). If the solution of (17) is not unique,
            select the solution which minimizes
            $||\widetilde{X}_q^{i+1} - \widetilde{X}_q^i||_F$ with $\widetilde{X}_q^i$ defined as in (25).
        4b) Partition $\widetilde{X}_q^{i+1}$ as in (13).
        4c) Disseminate $G_n^{i+1}$ to all nodes in $\mathcal{B}_{nq}$,
            $\forall n \in \mathcal{N}_q$.
    end
    5) Every node updates $X_k^{i+1}$ according to (26).
    $i \leftarrow i + 1$
```

Note: The fused output signal $\widehat{\mathbf{y}}(t)$ for the current batch of N samples can be computed at node q as $\widetilde{X}_q^{(i+1)T} \widetilde{\mathbf{y}}_q^i(t)$.

A. Convergence of the Objective

The following result states the convergence of the objective function under Algorithm 1's update rule.

Lemma 1. *Let $(X^i)_i$ be any sequence of iterates satisfying Algorithm 1's update rule for an instance $\mathbb{P}$ of (2) or (6). Then, all $(X^i)_{i>0}$ belong to the constraint set $\mathcal{S}$ of (6) and $\big(f(X^i)\big)_{i>0}$ is a monotonically decreasing convergent sequence.*

Proof. Because $\widetilde{X}_q^{i+1}$ satisfies the constraint set $\widetilde{\mathcal{S}}_q^i$ (see (24)) for any update i, we conclude from (23) that each X^i, satisfies the constraint set $\mathcal{S}$ of the network-wide problem (6) for any $i \geq 1$ (as also shown in [1, Lemma 1]). Furthermore, since $\widetilde{X}_q^{i+1}$ is the solution of (24), we have $f(C_q^i \widetilde{X}_q^{i+1}) \leq f(C_q^i \widetilde{X}_q)$ for any $\widetilde{X}_q \in \widetilde{\mathcal{S}}_q^i$. In particular, this inequality is verified for $\widetilde{X}_q = \widetilde{X}_q^i$ as defined in (25) for which $C_q^i \widetilde{X}_q^i = X^i$. This is because $\widetilde{X}_q^i$ indeed also belongs to the constraint set $\widetilde{\mathcal{S}}_q^i$ in (24), because of (23) and the fact that X^i belongs to $\mathcal{S}$ if $i \geq 1$ (see the beginning of the proof). Then, we have $f(C_q^i \widetilde{X}_q^{i+1}) = f(X^{i+1}) \leq f(C_q^i \widetilde{X}_q^i) = f(X^i)$. Since the sequence is monotonically decreasing and since it has a lower bound (defined by the global minimum of $\mathbb{P}$), it must converge. □

Lemma 1 guarantees that the sequence of objective values $\big(f(X^i)\big)_i$ associated with the iterates generated by the DASF algorithm converges and that the iterates correspond to feasible points of Problem (2) or (6). However, this does not imply convergence of the sequence $(X^i)_i$ itself, which is typically much more challenging to guarantee, even for centralized optimization algorithms such as line-search or trust region methods [12]–[16]. Moreover, even if the convergence of $(X^i)_i$ can be proven for the DASF algorithm, we still need to show that it converges to an "interesting" point, i.e., a stationary point of Problem (6), and preferably a global minimum. In the next two subsections, we will introduce two technical conditions (on top of the assumptions 1-3 in Subsection II-A) which will be combined in Section III-D to state convergence and optimality results for $(X^i)_i$. We end this subsection with a corollary of Lemma 1 showing that the elements of $(X^i)_{i>0}$ lie in a compact set which is required to show convergence of the DASF algorithm (see Section III-C and Appendix C).

Corollary 1. *If either Assumption 3a or 3b is satisfied, then the elements of the sequence $(X^i)_{i>0}$ obtained from the DASF algorithm lie in a compact set.*

The proof of Corollary 1 comes directly from the monotonic decrease of the objective obtained in Lemma 1.

B. Technical Conditions for Stationarity of Fixed Points

The first condition comes in two versions, where either one of the two has to be satisfied in order to prove that the fixed points of the algorithm are stationary points. These conditions are akin to the linear independence constraint qualification (LICQ) in classical optimization theory [9], and can be seen as compressed versions of these. We define a fixed point $\overline{X}$ of the DASF algorithm as a point that is invariant under a DASF update step at any updating node q, i.e., $X^{i+1} = X^i = \overline{X}$ independently of the updating node q at iteration i.

Condition 1a. *For a fixed point $\overline{X}$ of Algorithm 1, the elements of the set $\{\overline{X}^T \nabla_X h_j(\overline{X})\}_{j \in \mathcal{J}}$ are linearly independent.*

Since the set $\{\overline{X}^T \nabla_X h_j(\overline{X})\}_{j \in \mathcal{J}}$ consists of J matrices of size $Q \times Q$, the number of constraints J is upper bounded as

$$J \leq Q^2, \tag{27}$$

in order to allow the set to be linearly independent. As we will see in the proof of Theorem 1, this condition ensures that the Lagrange multipliers of the local problem are unique at a fixed point, eventually leading to the global optimality conditions being satisfied. Note that Condition 1a is highly likely to be satisfied in practice if (27) is satisfied, as a linear dependency would be highly coincidental if there are less matrices in the set than entries in each matrix (the points where this condition is violated is then a discrete set of points within a continuum of points). It is noted that Condition 1a can sometimes be shown to be satisfied a priori, based on the structure of the constraint set in the DSFO problem to which it is applied, without knowing the fixed points of the algorithm. The following example demonstrates how this can be proven when the constraint set is the Stiefel manifold, i.e., $X^TX = I_Q$, which is the case in, e.g., principal component analysis.

Example 1. Let $\mathcal{S}$ be the Stiefel manifold, i.e., we have the constraints $X^TX = I_Q$. There are $J = \frac{Q(Q-1)}{2}$ distinct constraints, where each constraint function is written as $h_{ml}(X) = \mathbf{x}(m)^T\mathbf{x}(l) - \mathbb{1}\{m = l\}$, where $\mathbf{x}(m)$ is the $m-$th column of X, $m, l \in \{1, \dots, Q\}$, $l \leq m$ and $\mathbb{1}$ is the indicator function. The derivative with respect to X can be found to be

$$\nabla_X h_{ml}(X) = \mathbf{x}(m)\mathbf{e}_l^T + \mathbf{x}(l)\mathbf{e}_m^T, \tag{28}$$

where the $\mathbf{e}$'s are the standard basis vectors of $\mathbb{R}^Q$. Multiplying this expression by X^T from the left and applying the constraint $X^TX = I_Q$ (assuming $X \in \mathcal{S}$), we have

$$X^T\nabla_X h_{ml}(X) = \mathbf{e}_m\mathbf{e}_l^T + \mathbf{e}_l\mathbf{e}_m^T. \tag{29}$$

Note that the right-hand side of (29) is independent of X. Since the J elements of $\{\mathbf{e}_m\mathbf{e}_l^T + \mathbf{e}_l\mathbf{e}_m^T\}_{m,l}$, $l \leq m$, are linearly independent, Condition 1a is satisfied if $X \in \mathcal{S}$. From [1, Lemma 1], every X^i that is produced by Algorithm 1 belongs to $\mathcal{S}$, therefore Condition 1a is satisfied for $i > 0$.

Condition 1a allows us to state a first important result:

Theorem 1. *Let $\mathbb{P}$ be an instance of (2) or (6). Then, if Condition 1a is satisfied, any fixed point of Algorithm 1 must be a stationary point of $\mathbb{P}$ satisfying its KKT conditions.*

Proof. See Appendix A. □

By definition, convergence of an algorithm can only be towards a fixed point of the algorithm, hence Theorem 1 guarantees that *if* the DASF algorithm converges, it converges to a stationary point of the global problem. For the special case of problems that are unconstrained, we do not require Condition 1a to hold (as it would lead to an empty set), in which case Theorem 1 can be simplified as follows.

Corollary 2. *Let $\mathbb{P}$ be an instance of (2) or (6) which is unconstrained. Then, any fixed point of Algorithm 1 must be a stationary point of $\mathbb{P}$ satisfying its KKT conditions.*

Alternatively, we propose a less restrictive – albeit more complicated – condition for cases with more constraints than Q^2 (see (27)), which is especially of interest for problems where $Q = 1$, for which Condition 1a would only allow a single constraint.

Condition 1b. *For a fixed point $\overline{X}$ of Algorithm 1, the elements of the set $\{D_{j,q}(\overline{X})\}_{j\in\mathcal{J}}$ are linearly independent for any q, where*

$$D_{j,q}(\overline{X}) \triangleq \begin{bmatrix} \overline{X}_q^T\nabla_{X_q}h_j(\overline{X}) \\ \sum\limits_{k\in\mathcal{B}_{n_1q}} \overline{X}_k^T\nabla_{X_k}h_j(\overline{X}) \\ \vdots \\ \sum\limits_{k\in\mathcal{B}_{n_{|\mathcal{N}_q|}q}} \overline{X}_k^T\nabla_{X_k}h_j(\overline{X}) \end{bmatrix}, \tag{30}$$

which is a block-matrix containing $(1+|\mathcal{N}_q|)$ blocks of $Q \times Q$ matrices.

For a given node q, the elements of the set $\{D_{j,q}(\overline{X})\}_{j\in\mathcal{J}}$ are now $(1 + |\mathcal{N}_q|)Q \times Q$ matrices and therefore their size depends on the nodes and the topology of the network. This means that we require the number of constraints J to satisfy:

$$J \leq (1 + \min_{k\in\mathcal{K}} |\mathcal{N}_k|)Q^2. \tag{31}$$

This condition assumes that the pruning of the network $\mathcal{T}^i(\mathcal{G}, q)$ preserves all the links with the neighbors of the updating node q. Furthermore, the proof in Appendix B will reveal that the number of constraints should also satisfy a second bound, which is also necessary for Condition 1b to hold:

$$J \leq \frac{Q^2}{K-1}\sum_{k\in\mathcal{K}} |\mathcal{N}_k|. \tag{32}$$

The reason is less obvious, but is related to specific interdependencies between the $D_{j,q}$'s across different nodes q (see Appendix B). It is noted that both bounds (31)-(32) are necessary, i.e., satisfying the first does not necessarily imply that the second one is satisfied and vice versa.

Similarly to the previous condition, Condition 1b is typically satisfied in practice when J satisfies both bounds in (31)-(32), i.e.,

$$J \leq \min\left(\frac{Q^2}{K-1}\sum_{k\in\mathcal{K}} |\mathcal{N}_k|,\ (1 + \min_{k\in\mathcal{K}} |\mathcal{N}_k|)Q^2\right). \tag{33}$$

Nevertheless, it is still possible that there exists a fixed point that is "close" to violating this condition, in which case the convergence of the DASF algorithm might become very slow if it reaches a neighborhood of such a fixed point. We refer to Section V on how to deal with these rare situations.

The following example illustrates why Condition 1b is less restrictive than Condition 1a.

Example 2. Suppose that $\mathcal{S} = \{X \in \mathbb{R}^{M\times Q} \mid X^TB = A,\ A \in \mathbb{R}^{Q\times L}\}$, which is a typical constraint used in linearly constrained minimum variance (LCMV) beamforming [6]. We have $J = QL$ constraints and each constraint function is given by $h_{ml}(X) = \mathbf{x}(m)^T\mathbf{b}(l) - A_{ml}$, where $\mathbf{x}(m)$ is the $m-$th column of X, $\mathbf{b}(l)$ is the $l-$th column of B and A_{ml} is the entry (m, l) of the matrix A. It is straightforward to show that requiring the elements $X^T\nabla_X h_{ml}(X) = X^T\mathbf{b}(l)\mathbf{e}_m^T$ to be linearly independent for every m and l, i.e., satisfying Condition 1a, is equivalent to requiring $\text{rank}(A) = L$. If $Q < L$, this condition cannot be satisfied. Looking now at Condition 1b, we have $X_k^T\nabla_{X_k}h_{ml}(X) = X_k^T\mathbf{b}_k(l)\mathbf{e}_m^T$ for every $k \in \mathcal{K}$, where $\mathbf{b}_k(l)$ is the block of $\mathbf{b}(l)$ corresponding to node k. Then, we can show that satisfying Condition 1b is equivalent to requiring that the matrix

$$[B_q^TX_q^i, \sum_{k\in\mathcal{B}_{n_1q}} B_k^TX_k^i, \dots, \sum_{k\in\mathcal{B}_{n_{|\mathcal{N}_q|}q}} B_k^TX_k^i]^T \tag{34}$$

has rank L at at a fixed point X^i, where the B_k's are obtained from the partitioning of B as in (3). This is possible even when $Q < L$, i.e., if node q has sufficient neighbors such that $(1 + |\mathcal{N}_q|)Q \geq L$.

Theorem 2. *Let $\mathbb{P}$ be an instance of (2) or (6). Then, if Condition 1b is satisfied, a fixed point of Algorithm 1 must be a stationary point of $\mathbb{P}$ satisfying its KKT conditions.*

Proof. See Appendix B. □

Finally, we note that these conditions are complementary in the sense that Condition 1a is not necessary for Condition 1b to hold, and vice versa.

C. Technical Conditions for Convergence

Conditions 1a and 1b are sufficient to show that fixed points of the DASF algorithm are stationary points. The next step is to show that accumulation points[3] of the algorithm are fixed points (and therefore stationary points of (2) or (6)), for which we require a second condition.

Condition 2. *The local problems (17) or (22) satisfy Assumptions 1-3.*

It is important to note here that this condition is usually satisfied in practice because the local problems have the same structure as the global problem (2), which was already assumed to satisfy Assumptions 1-3. It is therefore reasonable to assume that these local problems also inherit these same properties. In Section IV, we will give several examples to illustrate how Condition 2 can be checked in various problems. We will also present some examples of rare cases where this condition is not satisfied and provide fixes for it.

The well-posedness of the problem as required in Assumption 1 requires a continuity assumption on the point-to-set mapping from the space of inputs of the problem to its solution space. Formally, let $\widetilde{\mathcal{F}}_q : \mathbb{R}^{M\times Q} \rightrightarrows \mathbb{R}^{\widetilde{M}_q\times Q}$ be the following point-to-set mapping:

$$\widetilde{\mathcal{F}}_q(X) \triangleq \underset{\widetilde{W}_q : C_q(X)\widetilde{W}_q\in\mathcal{S}_q(X)}{\operatorname{argmin}} \quad f(C_q(X)\widetilde{W}_q), \tag{35}$$

where $C_q(X)$ is defined in (19) and $\mathcal{S}_q$ is the point-to-set mapping corresponding to the local parameterized constraint set for X when node q is the updating node:

$$\mathcal{S}_q(X) = \{W \in \mathcal{S} \mid W_k \in \mathcal{C}(X_k) \quad \forall k \neq q\} \tag{36}$$

with the subscript k referring to the per-node partitioning of X and W as in (7) and $\mathcal{C}(X_k)$ the set of all matrices with the same size and the same column space as X_k. To appreciate how the set (36) relates to the local constraint set, note that $\mathcal{S}_q(X^i) = \{X = C_q^i\widetilde{X}_q^i \mid \widetilde{X}_q^i \in \widetilde{\mathcal{S}}_q^i\}$. We require from our well-posedness property in Assumption 1 and Condition 2 that $\widetilde{\mathcal{F}}_q$ should be a continuous mapping (see [17, Definition 17.2] for a formal definition). Intuitively, this means that we expect that an arbitrarily small change in the inputs results in the addition or removal of points arbitrarily close to other points in the output set. In Section IV, we will illustrate on a few selected examples how the continuity of such a mapping can be checked.

Let the point-to-set mapping $\widetilde{\mathcal{M}}_q : \mathbb{R}^{M\times Q} \rightrightarrows \mathbb{R}^{\widetilde{M}_q\times Q}$ be defined as

$$\widetilde{\mathcal{M}}_q(X) \triangleq \underset{\widetilde{W}_q\in\widetilde{\mathcal{F}}_q(X)}{\operatorname{argmin}} \; ||W_q - X_q||_F^2 + \sum_{k\neq q} ||W_k - I_Q||_F^2, \tag{37}$$

where $\widetilde{W}_q = [W_q^T, W_1^T, \ldots, W_{q-1}^T, W_{q+1}^T, \ldots, W_K^T]^T$, i.e., $\widetilde{\mathcal{M}}_q$ selects the point in the set $\widetilde{\mathcal{F}}_q(X)$ that is closest to $[X_q^T, I_Q, \ldots, I_Q]^T$. We then define the point-to-set mapping $\mathcal{M}_q : \mathbb{R}^{M\times Q} \rightrightarrows \mathbb{R}^{M\times Q}$ as

$$\mathcal{M}_q(X) \triangleq C_q(X)\widetilde{\mathcal{M}}_q(X). \tag{38}$$

A single iteration of Algorithm 1 can then be summarized as

$$X^{i+1} \in \mathcal{M}_q(X^i). \tag{39}$$

In very contrived cases, it could happen that $\mathcal{M}_q(X^i)$ is not a singleton, i.e., there exists more than one solution at a certain iteration i which are equidistant to the previous estimate X^i. In that case, selecting by any means one particular solution is sufficient to resolve this ambiguity.

Theorem 3. *Suppose that for an instance $\mathbb{P}$ of (2) or (6), under the updates of Algorithm 1, Condition 2 is satisfied. Then:*

1) *Any accumulation point $\overline{X}$ of $(X^i)_i$ is a fixed point of the map $\mathcal{M}_q : \mathbb{R}^{M\times Q} \rightrightarrows \mathbb{R}^{M\times Q}$ for any q.*
2) $\lim_{i\to+\infty} ||X^{i+1} - X^i||_F = 0$.

Proof. See Appendix C. □

An important corollary is that any accumulation point is a fixed point of the full DASF algorithm as it is a fixed point for an update at any node q. However, note that Theorem 3 still does not guarantee convergence to a single point. The latter can be established if we assume the following condition:

Condition 3a. *The number of stationary points of the global problem $\mathbb{P}$ is finite.*

Theorem 4. *If Conditions 2 and 3a are satisfied, then $(X^i)_i$ converges to a single point.*

Proof. See Appendix C. □

The condition on the finiteness of the number of stationary points can be relaxed to the following condition:

Condition 3b. *The number of solutions of each local problem (22) is finite or the solver of the local problems (22) can only obtain a finite subset of the solutions of (22).*

In other words, we only require the finiteness of the number of solutions *obtainable* through the solver used for solving the local problems. For example, when maximizing tr(X^TAX) over $X^TBX = I$, there are infinitely many options for a solution X^* represented as $X^* = V^*U$ where U is an orthogonal matrix and V^* contains the principal generalized eigenvectors of the matrix pencil (A, B). However, a solver using a generalized eigenvalue decomposition to solve this problem can only output V^* itself up to a sign change of its columns, hence the solver is only able to select solutions from a finite subset of the complete solution set.

Theorem 5. *If Conditions 2 and 3b are satisfied, then $(X^i)_i$ converges to a single point.*

The proof of Theorem 4 can be straightforwardly applied to Theorem 5 as well.

[3] We define an accumulation point of a sequence $(X^i)_{i\in\mathbb{N}}$ as the limit of a converging subsequence $(X^i)_{i\in\mathcal{I}}$ of $(X^i)_{i\in\mathbb{N}}$, with $\mathcal{I} \subseteq \mathbb{N}$.

D. Convergence to Stationary Points and Global Minima

We can now combine all of the previous results to eventually obtain complete convergence results of the DASF algorithm to stationary points of Problem (6).

Theorem 6. *Suppose that for an instance $\mathbb{P}$ of (2) or (6), under the updates of Algorithm 1, Conditions 1, 2, 3 (for 1 and 3 either the form a or b) are satisfied, then $(X^i)_i$ converges and $\lim_{i\to+\infty} X^i = \overline{X}$, where $\overline{X}$ is a stationary point of $\mathbb{P}$ satisfying its KKT conditions.*

Proof. From Theorems 4 and 5, $(X^i)_i$ converges to a single point $\overline{X}$. Therefore $\overline{X}$ is a fixed point of Algorithm 1 (Theorem 3). From Theorems 1 and 2, fixed points of the DASF algorithm are stationary points of $\mathbb{P}$ satisfying its KKT conditions, proving the theorem. □

Theorem 6 can lead to stronger convergence guarantees if all minima of the problem $\mathbb{P}$ are global minima (i.e., the value of the objective function is the same in all minima), which will be explained next. It is noted that many of the common spatial filtering design criteria satisfy this assumption, including PCA, canonical correlation analysis, minimum variance beamformers, generalized eigenvalue decomposition and trace ratio optimization.

Theorem 7. *Under the same settings of Theorem 6, if all minima of $\mathbb{P}$ are global minima, the only stable fixed points of Algorithm 1 are in $\mathcal{X}^*$.*

Proof. For any fixed point $\overline{X} \notin \mathcal{X}^*$, there exists a descent direction in $\mathcal{S}$ (as $\overline{X}$ cannot be a minimum), and therefore there exists a perturbation ΔX, with $X + \Delta X \in \mathcal{S}$, such that $f(X + \Delta X) \leq f(X)$. Due to the monotonic decrease of $(f(X^i))_i$ (see Lemma 1), the sequence $(X^i)_i$ is kicked out of equilibrium and cannot return to it, hence the equilibrium is unstable. Therefore, in the absence of local minima, the only stable fixed points of Algorithm 1 must be in $\mathcal{X}^*$. □

Corollary 3. *If all minima of f over $\mathcal{S}$ are global minima and all conditions of Theorem 6 are satisfied, $(X^i)_i$ converges to the global minimum of (2) or (6) with high probability*[4]*.*

This corollary follows immediately from Theorems 6 and 7. Indeed, from Theorem 6, we know that $(X^i)_i$ converges to a stationary point. Since all fixed points that are not in $\mathcal{X}^*$ are unstable (Theorem 7), the algorithm will eventually escape from the neighborhoods of such unstable equilibria and will end up in a point of $\mathcal{X}^*$.

Corollary 4. *If the global problem (2) or (6) is a convex problem with a strongly convex objective f and all conditions of Theorem 6 are satisfied, $(X^i)_i$ converges to the unique global minimum X^*.*

Proof. Theorem 6 guarantees convergence to a stationary point of Problem (6). The result comes from the fact that for a convex problem with a strongly convex objective, the unique stationary point is the global minimum. □

The previous corollary can be further relaxed by removing the requirements of Conditions 1a and 1b for a problem which is unconstrained (see Corollary 2):

Corollary 5. *If the global problem (2) or (6) is unconstrained with a strongly convex objective f and Condition 2 is satisfied, $(X^i)_i$ converges to the unique global minimum X^*.*

IV. Selected Examples

In this section, we illustrate how the different convergence results and conditions translate to some commonly encountered spatial filtering problems, and how the problems can be manipulated in order to satisfy the conditions in case they are violated. In the following, $\mathbf{y}$ and $\mathbf{d}$ are stochastic signals, and we denote their covariance and cross-covariance matrices $R_{\mathbf{yy}} = \mathbb{E}[\mathbf{y}(t)\mathbf{y}^T(t)]$ and $R_{\mathbf{yd}} = \mathbb{E}[\mathbf{y}(t)\mathbf{d}^T(t)]$. We also define the corresponding compressed matrices $R^i_{\widetilde{\mathbf{y}}_q\widetilde{\mathbf{y}}_q} = \mathbb{E}[\widetilde{\mathbf{y}}^i_q(t)\widetilde{\mathbf{y}}^{iT}_q(t)] = C^{iT}_q R_{\mathbf{yy}} C^i_q$ and $R^i_{\widetilde{\mathbf{y}}_q\mathbf{d}} = \mathbb{E}[\widetilde{\mathbf{y}}^i_q(t)\mathbf{d}^T(t)] = C^{iT}_q R_{\mathbf{yd}}$ for the local problems.

As will be shown, we typically require $R_{\mathbf{yy}}$ to be non-singular for Condition 2 to be satisfied, in particular for the local problem to be well-posed (Assumption 1), as Assumptions 2 and 3a (or 3b) are satisfied automatically if the global problem satisfies these. We will see that a question that will arise is whether the local $R^i_{\widetilde{\mathbf{y}}_q\widetilde{\mathbf{y}}_q}$ is non-singular. If this is not satisfied, we cannot ignore situations where a small change in the local problem parameters leads to discontinuous changes in their solution sets, making Condition 2 (in particular Assumption 1) invalid. The following result relates the (non-)singularity of $C^{iT}_q R C^i_q$ to the (non-)singularity of R.

Lemma 2. *Suppose that a matrix $R \in \mathbb{R}^{M\times M}$ is non-singular. Then, given $C^i_q \in \mathbb{R}^{M\times\widetilde{M}_q}$ as defined in (19), if C^i_q has full rank, the matrix $C^{iT}_q R C^i_q$ is also non-singular.*

The proof is omitted as this is a well-known property of the rank of a matrix. Lemma 2 will be used in many of the examples that are discussed in this section.

Remark 1. Note that — by construction — the matrix C^i_q has full rank unless one of the matrices X_k has linearly dependent columns. The latter is a contrived case and is to be avoided anyway, since it would imply that redundant data is transmitted via $\widehat{\mathbf{y}}_k$ (in which case the algorithm should only transmit the non-redundant part). The result of Lemma 2 can therefore be interpreted in such a way that if the global problem is well-posed, then so is the local one because from this result, we have a guarantee that the local problems' covariance matrix will also be non-singular.

A. Least Squares / Minimum Mean Square Error and Ridge Regression

The least squares (LS) / minimum mean square error (MMSE) problem can be written as

$$\underset{X\in\mathbb{R}^{M\times Q}}{\text{minimize}}\ \mathbb{E}[||X^T\mathbf{y}(t) - \mathbf{d}(t)||^2], \tag{40}$$

which is an unconstrained problem with a convex quadratic objective, since $R_{\mathbf{yy}}$ is positive semi-definite by definition. A solution X^* of (40) needs to satisfy the normal equations given

[4] The phrasing "with high probability" here refers to the fact that it is expected that the algorithm cannot end up in an unstable equilibrium, as it would always escape from it due to numerical or estimation noise.

as $R_{\mathbf{yy}}X^* = R_{\mathbf{yd}}$. If additionally $R_{\mathbf{yy}}$ is positive definite, then it is invertible, and the objective is strongly convex leading to the unique solution $X^* = R_{\mathbf{yy}}^{-1}R_{\mathbf{yd}}$ of the global problem. In this case, (40) is well-posed and satisfies **Assumptions 1 and 3**, while **Assumption 2** is automatically satisfied as there are no constraints. We write the corresponding local problem (17) as

$$\underset{\widetilde{X}_q\in\mathbb{R}^{\widetilde{M}_q\times Q}}{\text{minimize}}\ \mathbb{E}[||\widetilde{X}_q^T\widetilde{\mathbf{y}}_q^i(t)-\mathbf{d}(t)||^2]. \tag{41}$$

Since (41) does not have any constraints, **Condition 1a** is trivially satisfied (see also Corollary 2). Similarly to the centralized case, $\widetilde{X}_q$ is a solution of (41) if and only if it satisfies the normal equations

$$R^i_{\widetilde{\mathbf{y}}_q\widetilde{\mathbf{y}}_q}\widetilde{X}_q = R^i_{\widetilde{\mathbf{y}}_q\mathbf{d}}. \tag{42}$$

In general cases, $R^i_{\widetilde{\mathbf{y}}_q\widetilde{\mathbf{y}}_q} = C_q^{iT}R_{\mathbf{yy}}C_q^i$ is invertible from Lemma 2, and the solution of the local problem at iteration i and node q is unique (satisfying **Condition 3a**) and equal to

$$\widetilde{X}_q^{i+1} = (R^i_{\widetilde{\mathbf{y}}_q\widetilde{\mathbf{y}}_q})^{-1}R^i_{\widetilde{\mathbf{y}}_q\mathbf{d}}. \tag{43}$$

We then have $\widetilde{\mathcal{F}}_q(X^i) = \{\widetilde{X}_q^{i+1}\}$, which is continuous since matrix inversion is continuous (see Cramer's rule). Hence **Condition 2** is satisfied if $R^i_{\widetilde{\mathbf{y}}_q\widetilde{\mathbf{y}}_q}$ is non-singular. Since (40) is unconstrained, we satisfy the conditions of **Corollary 5**, and we conclude that the DASF algorithm applied to the LS / MMSE problem converges to the optimal solution X^*.

In cases where $R^i_{\widetilde{\mathbf{y}}_q\widetilde{\mathbf{y}}_q}$ is singular, the normal equations have infinitely many solutions and (41) therefore admits infinitely many solutions as well. However, a small change in inputs can lead to $R^i_{\widetilde{\mathbf{y}}_q\widetilde{\mathbf{y}}_q}$ suddenly becoming non-singular, reducing the number of solutions from infinitely many to a single one, which would not satisfy **Condition 2**. We can resolve this problem by additionally requiring the column space of $\widetilde{X}_q$ to be orthogonal to the null space of $R^i_{\widetilde{\mathbf{y}}_q\widetilde{\mathbf{y}}_q}$, which corresponds to the solution with minimum norm [18, Chapter 3, Section 2]. The solution of this surrogate problem is always uniquely defined and varies continuously with $R^i_{\widetilde{\mathbf{y}}_q\widetilde{\mathbf{y}}_q}$ and $R^i_{\widetilde{\mathbf{y}}_q\mathbf{d}}$, even at points where $R^i_{\widetilde{\mathbf{y}}_q\widetilde{\mathbf{y}}_q}$ becomes singular. In practice though, $R^i_{\widetilde{\mathbf{y}}_q\widetilde{\mathbf{y}}_q}$ will never be singular (see also Remark 1), and the additional constraint is always trivially satisfied, making the surrogate problem equivalent to the original one.

If $R_{\mathbf{yy}}$ is singular (therefore implying that **Assumptions 1 and 3** do not hold in the general case for (40)), one can consider adding an ℓ_2-norm constraint or penalty to (40), leading to the ridge regression (RR) problem. As this can be rewritten as a least squares problem where $R_{\mathbf{yy}}$ is replaced by $R_{\mathbf{yy}}+\alpha I_M$, the resulting matrix becomes non-singular. In this case, both the local and global problems will satisfy the well-posedness assumption, and therefore **Corollary 5** straightforwardly applies, such that convergence to the global solution is guaranteed.

B. Linearly Constrained Minimum Variance

The linearly constrained minimum variance (LCMV) problem is a convex problem often used in beamforming applications [6], and written as

$$\begin{aligned}\underset{X\in\mathbb{R}^{M\times Q}}{\text{minimize}}\quad & \mathbb{E}[||X^T\mathbf{y}(t)||^2] = \text{tr}(X^TR_{\mathbf{yy}}X)\\ \text{subject to}\quad & X^TB = A,\end{aligned} \tag{44}$$

with the linear term $B\in\mathbb{R}^{M\times L}$, $M > L$. If $R_{\mathbf{yy}}$ is positive definite and $\text{rank}(B) = L$, Problem (44) has a strongly convex objective and the unique solution is given by $X^* = R_{\mathbf{yy}}^{-1}B(B^TR_{\mathbf{yy}}^{-1}B)^{-1}A^T$. Problem (44) is then well-posed and satisfies **Assumption 1**. As shown in Example 2, the gradient of each constraint function h_{ml} of (44) is given by $\nabla_X h_{ml}(X) = \mathbf{b}(l)\mathbf{e}_m$, where $\mathbf{b}(l)$ corresponds to the $l-$th column of B. Therefore, **Assumption 2** is satisfied when B is full (column) rank. Finally, since the objective is continuous and the objective strongly convex, the sublevel sets of the objective are compact. Adding the constraints $X^TB = A$ preserves compactness as the intersection of a closed set with a compact one is compact, satisfying **Assumption 3**. The local LCMV problem is written as

$$\begin{aligned}\underset{\widetilde{X}_q\in\mathbb{R}^{\widetilde{M}_q\times Q}}{\text{minimize}}\quad & \text{tr}(\widetilde{X}_q^TR^i_{\widetilde{\mathbf{y}}_q\widetilde{\mathbf{y}}_q}\widetilde{X}_q)\\ \text{subject to}\quad & \widetilde{X}_q^T\widetilde{B}_q^i = A,\end{aligned} \tag{45}$$

at iteration i and node q. As discussed in Example 2, **Condition 1a** is satisfied if A has rank L, or alternatively **Condition 1b** is satisfied if the matrix given in (34) has rank L. Excluding the rare cases where C_q^i is rank deficient (which can be dealt with, see Remark 1), $R^i_{\widetilde{\mathbf{y}}_q\widetilde{\mathbf{y}}_q}$ is invertible from Lemma 2, and it can be shown from the property that B is full rank that $\widetilde{B}_q^{iT}(R^i_{\widetilde{\mathbf{y}}_q\widetilde{\mathbf{y}}_q})^{-1}\widetilde{B}_q^i$ is also invertible. Therefore, the unique solution of the local problem (implying **Condition 3a** is satisfied) is given by an analogous expression to the one of the global LCMV problem:

$$\widetilde{X}_q^{i+1} = (R^i_{\widetilde{\mathbf{y}}_q\widetilde{\mathbf{y}}_q})^{-1}\widetilde{B}_q^i[\widetilde{B}_q^{iT}(R^i_{\widetilde{\mathbf{y}}_q\widetilde{\mathbf{y}}_q})^{-1}\widetilde{B}_q^i]^{-1}A^T. \tag{46}$$

From the continuity of matrix inversion, **Condition 2** is satisfied. This implies that we can apply **Corollary 4** to conclude that the DASF algorithm will converge to the globally optimal LCMV solution. We note that **Condition 1b** is satisfied in practice with high probability when J satisfies the upper bound (33) as illustrated in Figure 2. However, there still exist situations where slow convergence is observed in cases where **Condition 1b** is "close" to being violated. We propose a fix for those situations in Section V-A. Note that Condition 1b is sufficient for showing convergence to the optimal point but it is not a necessary condition, as in Figure 2, we can still see (slow) convergence for some cases when J does not satisfy (33).

C. Generalized Eigenvalue Decomposition and Principal Component Analysis

Let us consider the problem:

$$\begin{aligned}\underset{X\in\mathbb{R}^{M\times Q}}{\text{minimize}}\quad & -\mathbb{E}\Big[||X^T\mathbf{y}(t)||^2\Big] = -\text{tr}(X^TR_{\mathbf{yy}}X)\\ \text{subject to}\quad & \mathbb{E}[X^T\mathbf{v}(t)\mathbf{v}^T(t)X] = X^TR_{\mathbf{vv}}X = I_Q,\end{aligned} \tag{47}$$

where $\mathbf{y}$ and $\mathbf{v}$ are $M-$dimensional time signals. Note that (47) can be transformed into (2) by minimizing the negative or the reciprocal of the objective. A global solution of this

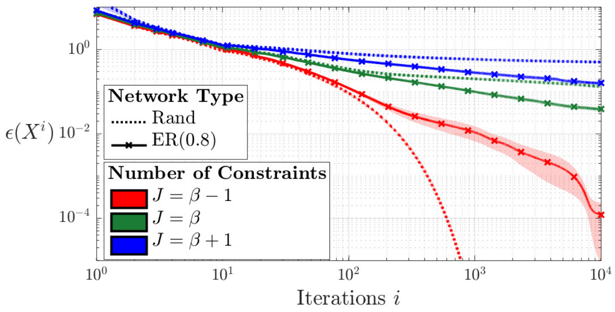


Fig. 2. Convergence comparison of the DASF algorithm solving the LCMV problem (44) for Erdős-Rényi random graphs with connection probability 0.8 (*ER(*0.8*)*) and randomly generated trees (*Rand*). The bold lines represent the mean values across 100 Monte Carlo runs, while the shaded areas delimit the standard error of the mean around them. β represents the upper bound given in (33) on the number of constraints J.

problem is obtained by computing the Q principal generalized eigenvectors when computing the generalized eigenvalue decomposition (GEVD) of the matrix pencil $(R_{\mathbf{yy}}, R_{\mathbf{vv}})$. This GEVD problem is often encountered in discriminant analysis or max-SNR filtering [3], [6], [19]. It also contains the standard eigenvalue decomposition (EVD) problem or principal component analysis (PCA) as a special case, which is obtained when the constraint set of (47) is replaced with $X^TX = I_Q$, i.e., the Stiefel manifold (or equivalently, when $\mathbf{v}$ is a white noise process). Therefore, the discussions below apply to the EVD and PCA problems as well.

If both $R_{\mathbf{yy}}$ and $R_{\mathbf{vv}}$ are positive definite and the $Q+1$ largest generalized eigenvalues (GEVLs) of $(R_{\mathbf{yy}}, R_{\mathbf{vv}})$ are all distinct, Problem (47), is well-posed [20], therefore satisfying **Assumption 1**. A solution of (47) is given by $X^* = \text{GEVD}_Q(R_{\mathbf{yy}}, R_{\mathbf{vv}})$, where we consider that, given a matrix pencil (A, B), $\text{GEVD}_Q(A, B)$ is a matrix containing the Q generalized eigenvectors of the pencil in its columns, corresponding to its largest GEVLs. We note that the solution of this problem is not unique, and applying any orthogonal transformation on $X^* = \text{GEVD}_Q(R_{\mathbf{yy}}, R_{\mathbf{vv}})$ is also a valid solution. Similarly to Example 1, we have $\nabla_X h_{ml}(X^*) = R_{\mathbf{vv}}(\mathbf{x}^*(m)\mathbf{e}_l^T + \mathbf{x}^*(l)\mathbf{e}_m^T)$. It can be shown that the linear independence of the set $\{\nabla_X h_{ml}(X^*)\}_{m,l}$ is equivalent to the linear independence of the columns of $R_{\mathbf{vv}}X^*$. Under Assumption 1, $R_{\mathbf{vv}}$ is positive definite, hence invertible and since X^* contains generalized eigenvectors of $(R_{\mathbf{yy}}, R_{\mathbf{vv}})$ in its columns, it is by definition full column rank. Therefore the solutions of (47) satisfy the LICQ conditions hence **Assumption 2** is satisfied. Additionally, the sublevel sets of the objective of (47) are closed, while the constraint of (47) defines a compact set. From the compactness of their intersection, we satisfy **Assumption 3**. From (47), we observe that the corresponding local problem (17) is given by

$$\begin{aligned} \underset{\widetilde{X}_q\in\mathbb{R}^{\widetilde{M}_q\times Q}}{\text{minimize}} \quad & -\text{tr}(\widetilde{X}_q^T R^i_{\widetilde{\mathbf{y}}_q\widetilde{\mathbf{y}}_q}\widetilde{X}_q) \\ \text{subject to} \quad & \widetilde{X}_q^T R^i_{\widetilde{\mathbf{v}}_q\widetilde{\mathbf{v}}_q}\widetilde{X}_q = I_Q, \end{aligned} \tag{48}$$

and for any fixed iteration i and node q, a solution of the local problem is

$$\widetilde{X}_q^{i+1} = \text{GEVD}_Q(R^i_{\widetilde{\mathbf{y}}_q\widetilde{\mathbf{y}}_q},\ R^i_{\widetilde{\mathbf{v}}_q\widetilde{\mathbf{v}}_q}). \tag{49}$$

Similarly to Example 1, we can show that $\nabla_X h_{ml}(X) = R_{\mathbf{vv}}(\mathbf{x}(m)\mathbf{e}_l^T + \mathbf{x}(l)\mathbf{e}_m^T)$. Therefore, for any X satisfying the constraints of (47), we have $X^T\nabla_X h_{ml}(X) = \mathbf{e}_m\mathbf{e}_l^T + \mathbf{e}_l\mathbf{e}_m^T$, which, for every (m, l), form a linearly independent set hence **Condition 1a** is satisfied at any iteration.

On the other hand, we do not have a guarantee that the algorithm does not converge to a local problem where the $Q+1$ largest generalized eigenvalues of the local matrix pencil are all distinct. This would lead to a violation of the continuity of the problem as required in **Condition 2**, which is otherwise satisfied. This event, however improbable, can be monitored and a particular fix is described in Subsection V-B.

As noted previously, there exists infinitely many solutions of Problem (47) therefore **Condition 3a** cannot be satisfied. However, suppose the solver we choose to solve the local problems (48) computes the generalized eigenvalue decomposition of the matrix pencil $(R^i_{\widetilde{\mathbf{y}}_q\widetilde{\mathbf{y}}_q},\ R^i_{\widetilde{\mathbf{v}}_q\widetilde{\mathbf{v}}_q})$ as in (49). Then, the solver can only output one of the 2^Q possible solutions of the local problems at each iteration, namely one of the matrices containing the Q most significant generalized eigenvectors of the matrix pencil $(R^i_{\widetilde{\mathbf{y}}_q\widetilde{\mathbf{y}}_q},\ R^i_{\widetilde{\mathbf{v}}_q\widetilde{\mathbf{v}}_q})$, which are equal up to a sign change of the columns. This allows to eliminate all other solutions of (48) from the set of candidate solutions, making the solution set obtainable by the solver finite and leading to **Condition 3b** being satisfied.

Under these conditions, we conclude from **Theorem 6** that the DASF algorithm converges to a stationary point of the GEVD problem. As all minima/maxima of (47) are global minima/maxima, we obtain convergence to the global solution according to **Corollary 3**.

D. Trace Ratio Optimization

The trace ratio optimization (TRO) problem [21], [22] is defined as

$$\begin{aligned} \underset{X\in\mathbb{R}^{M\times Q}}{\text{minimize}} \quad & -\frac{\mathbb{E}\Big[||X^T\mathbf{y}(t)||^2\Big]}{\mathbb{E}\Big[||X^T\mathbf{v}(t)||^2\Big]} = -\frac{\text{tr}(X^TR_{\mathbf{yy}}X)}{\text{tr}(X^TR_{\mathbf{vv}}X)} \\ \text{subject to} \quad & X^TX = I_Q. \end{aligned} \tag{50}$$

Considering that $R_{\mathbf{yy}}$ and $R_{\mathbf{vv}}$ are positive definite, there exists a scalar ρ such that a solution of this problem is given by $X^* = \text{EVD}_Q(R_{\mathbf{yy}} - \rho R_{\mathbf{vv}})$ [21], where $\text{EVD}_Q(A)$ is a matrix containing the Q eigenvectors corresponding to the Q largest eigenvalues of A in its columns. This solution returned by the TRO solver defined in [21] is unique up to a sign change of its columns if the $Q+1$ eigenvalues of $R_{\mathbf{yy}} - \rho R_{\mathbf{vv}}$ are distinct, in which case the problem is well-posed, satisfying **Assumption 1**. It can be shown in a similar fashion as to IV-C that (50) satisfies **Assumptions 2 and 3**. Additionally, From Example 1, we know that the Stiefel manifold satisfies **Condition 1a**. The local problem that node q solves at iteration i is written as

$$\begin{aligned} \underset{\widetilde{X}_q\in\mathbb{R}^{\widetilde{M}_q\times Q}}{\text{minimize}} \quad & -\frac{\text{tr}(\widetilde{X}_q^T R^i_{\widetilde{\mathbf{y}}_q\widetilde{\mathbf{y}}_q}\widetilde{X}_q)}{\text{tr}(\widetilde{X}_q^T R^i_{\widetilde{\mathbf{v}}_q\widetilde{\mathbf{v}}_q}\widetilde{X}_q)} \\ \text{subject to} \quad & \widetilde{X}_q^T\widetilde{\Gamma}_q^i\widetilde{X}_q = I_Q, \end{aligned} \tag{51}$$

with[5] $\widetilde{\Gamma}_q^i = C_q^{iT} C_q^i$. There exists a scalar ρ_q^i such that the solution of the local problem is given by

$$\widetilde{X}_q^{i+1} = \text{GEVD}_Q\left(R^i_{\widetilde{\mathbf{y}}_q\widetilde{\mathbf{y}}_q} - \rho_q^i R^i_{\widetilde{\mathbf{v}}_q\widetilde{\mathbf{v}}_q},\ \widetilde{\Gamma}_q^i\right), \tag{52}$$

if both matrices of the pencil are positive definite (this holds if C_q^i is full rank from Lemma 2) and the $Q+1$ largest GEVLs are distinct (in Subsection V-B, we will cover the case where this is not satisfied). Therefore, the local problems are generally well-posed, satisfying **Condition 2**. Similarly to the GEVD example, a solver using (52) to solve (51) satisfies **Condition 3b**. As there are no local minima, in practical cases we obtain convergence to a global minimum from **Corollary 3**.

E. Canonical Correlation Analysis

The goal of canonical correlation analysis (CCA) is to find two spatial filters for two different multi-channel signals such that their outputs are maximally correlated [4], [5]. The CCA problem can be (re)written as [5]

$$\begin{aligned} \underset{X,W\in\mathbb{R}^{M\times Q}}{\text{minimize}} \quad & -\mathbb{E}\Big[\text{tr}\left(X^T\mathbf{y}(t)\mathbf{v}^T(t)W\right)\Big] = -\text{tr}(X^T R_{\mathbf{yv}} W) \\ \text{subject to} \quad & \mathbb{E}[X^T\mathbf{y}(t)\mathbf{y}^T(t)X] = X^T R_{\mathbf{yy}} X = I_Q, \\ & \mathbb{E}[W^T\mathbf{v}(t)\mathbf{v}^T(t)W] = W^T R_{\mathbf{vv}} W = I_Q. \end{aligned} \tag{53}$$

Assuming that $R_{\mathbf{yy}}$ and $R_{\mathbf{vv}}$ are positive definite the solution of Problem (53) is given by $X^* = \text{GEVD}_Q(R_{\mathbf{yv}}R_{\mathbf{vv}}^{-1}R_{\mathbf{vy}},\ R_{\mathbf{yy}})$ and $W^* = R_{\mathbf{vv}}^{-1}R_{\mathbf{vy}}X^*\Lambda^{-1/2}$, where Λ is a $Q\times Q$ diagonal matrix containing the Q largest GEVLs of the pencil $(R_{\mathbf{yv}}R_{\mathbf{vv}}^{-1}R_{\mathbf{vy}},\ R_{\mathbf{yy}})$. If the $Q+1$ largest GEVLs of this pencil are all distinct, the CCA problem is well-posed, satisfying **Assumption 1**. Similarly to IV-C, it can be shown that (53) satisfies **Assumptions 2 and 3**. As shown previously, **Condition 1a** is satisfied at any iteration i. We can write the local problems as

$$\begin{aligned} \underset{\widetilde{X}_q,\widetilde{W}_q\in\mathbb{R}^{\widetilde{M}_q\times Q}}{\text{minimize}} \quad & -\text{tr}(\widetilde{X}_q^T R^i_{\widetilde{\mathbf{y}}_q\widetilde{\mathbf{v}}_q}\widetilde{W}_q) \\ \text{subject to} \quad & \widetilde{X}_q^T R^i_{\widetilde{\mathbf{y}}_q\widetilde{\mathbf{y}}_q}\widetilde{X}_q = I_Q, \\ & \widetilde{W}_q^T R^i_{\widetilde{\mathbf{v}}_q\widetilde{\mathbf{v}}_q}\widetilde{W}_q = I_Q. \end{aligned} \tag{54}$$

From Lemma 2 and Remark 1, we conclude that both $R^i_{\widetilde{\mathbf{y}}_q\widetilde{\mathbf{y}}_q}$ and $R^i_{\widetilde{\mathbf{v}}_q\widetilde{\mathbf{v}}_q}$ are non-singular. The solution of the local problems is then

$$\widetilde{X}_q^{i+1} = \text{GEVD}_Q\left(R^i_{\widetilde{\mathbf{y}}_q\widetilde{\mathbf{v}}_q}(R^i_{\widetilde{\mathbf{v}}_q\widetilde{\mathbf{v}}_q})^{-1}R^i_{\widetilde{\mathbf{v}}_q\widetilde{\mathbf{y}}_q},\ R^i_{\widetilde{\mathbf{y}}_q\widetilde{\mathbf{y}}_q}\right) \tag{55}$$

$$\widetilde{W}_q^{i+1} = (R^i_{\widetilde{\mathbf{v}}_q\widetilde{\mathbf{v}}_q})^{-1}R^i_{\widetilde{\mathbf{v}}_q\widetilde{\mathbf{y}}_q}\widetilde{X}_q^{i+1}(\Lambda_q^i)^{-1/2}, \tag{56}$$

where Λ_q^i is the $Q\times Q$ diagonal matrix containing the Q largest GEVLs of the pencil given in (55). If the $Q+1$ GEVLs are all distinct, the local problem (54) is again well-posed and **Condition 2** is satisfied. If this is not the case, the same fixes we discussed in the examples of the GEVD and TRO problems can be applied. As in these latter examples, **Condition 3b** is also satisfied when choosing a solver which computes the solutions of (54) using (55)-(56). Therefore, the corresponding DASF algorithm will converge to the centralized CCA solution from **Corollary 3**.

V. Fixes in Case of Violations of Conditions

A. Avoiding Violations of Condition 1b

In certain cases, the linear independence requirements of Condition 1a or 1b cannot be a priori guaranteed. In such cases, even though they are expected to hold (since they are only violated in a discrete set of points within a continuum of possibilities), it is possible that the DASF algorithm passes a neighborhood of a fixed point $\overline{X}$ where the conditions are not satisfied, especially when J is close to the upper bound (33) or even larger (see Figure 2). When an updating node q observes that the elements of $\{D_{j,q}(X^i)\}_j$ are close to being linearly dependent, the algorithm is potentially converging to a fixed point that does not satisfy the linear independence requirements. This can be checked at the node itself, e.g., when the $J-$th singular value of $[\text{vec}(D_{1,q}(X^i)),\dots,\text{vec}(D_{J,q}(X^i))]$ is close to zero[6]. In that case, we propose that the neighbors k of node q split their local X_k^i into a sum of two matrices $X_{k,a}^i$ and $X_{k,b}^i$ (which both have linearly independent columns) such that $X_k^i = X_{k,a}^i + X_{k,b}^i$, and start to temporarily communicate two sets of compressed signals $\widehat{\mathbf{y}}_{k,a}^i = X_{k,a}^{iT}\mathbf{y}_k$ and $\widehat{\mathbf{y}}_{k,b}^i = X_{k,b}^{iT}\mathbf{y}_k$. This implies that the elements of $\{D_{j,q}(X^i)\}_j$ have grown in dimension thereby making the linear independence requirement of the set $\{D_{j,q}(X^i)\}_j$ more likely to be met. As soon as the algorithm escapes the suboptimal point, node k can again merge $X_{k,a}$ and $X_{k,b}$ for future iterations, returning to the minimal communication bandwidth setting.

B. Avoiding Violations of Condition 2

In the very contrived cases where the algorithm would produce a subsequence converging to a stationary point at which Condition 2 does not hold, convergence of the overall sequence cannot be guaranteed anymore and the algorithm will possibly oscillate between points which are solutions of the local problems, but not necessarily corresponding to global solutions. Indeed, without Condition 2, we cannot guarantee that $\lim_{X\to\overline{X}}\mathcal{M}_q(X)$ is well-defined (i.e., is unique and a singleton). A practical and pragmatic fix for such scenarios is to monitor both the potential oscillatory behavior and the continuity of $\widetilde{\mathcal{F}}_q(X^i)$ and skip the update at the node where a problem occurs. To detect such an oscillation, select an arbitrarily small $\varepsilon > 0$, and monitor whether $|f(X^{i+1}) - f(X^i)|\cdot||X^{i+1} - X^i||_F^{-1} > \varepsilon$ which can be interpreted as a sufficient decrease condition. If this condition is violated, a potential oscillation is flagged. In addition, one could monitor whether a particular near-discontinuity condition is met, which is problem specific. This condition can be interpreted as a sufficient decrease condition. In the case of GEVD and TRO problems discussed above, monitoring the difference between

[5] The expression X^TX can be rewritten in the form required by Z_B in (2) if B is set to $B = I_M$, which results in $(X^TB)\cdot(X^TB)^T$. It is then observed from (21) that $\widetilde{B}_q^i = C_q^i$.

[6] Note that, depending on the problem, constructing $D_{j,q}$'s at node q might require some additional data exchange, namely each node k should compute and transmit $X_k^{iT}\nabla_{X_k}h_j(X^i)$'s towards node q. However, this would not add a significant burden to the communication cost as these matrices are $Q\times Q$, while there are also many cases where the compressed gradients are already required to be communicated due to the problem's structure.

the $Q-$th and $(Q+1)-$th largest eigenvalue is sufficient to detect such a discontinuity. If both such a near-discontinuity and an insufficient decrease are flagged, the update at that node should be skipped.

VI. Conclusion

In this paper, we have analyzed the technical convergence properties and conditions of the DASF algorithm and provided formal proofs of convergence. The conditions required for convergence were shown to be satisfied in many practical problems, assuming some bounds on the dimension of the problem are satisfied, which depend on the number of constraints and the network topology. We have provided some illustrative examples to demonstrate how the — sometimes rather technical — conditions can be validated in practice. These examples also showed in which contrived cases a problem could occur, e.g., in case of singularities or eigenvalue collisions in sensor signal covariance matrices, which are expected to be rare in practice. Nevertheless, we have discussed various methods to fix these convergence problems for the rare cases where these should occur.

Appendix A

Proof of Theorem 1. Let us write the KKT conditions of Problem (6), as mentioned in Assumption 2 [23]:

$$\nabla_X \mathcal{L}(X, \boldsymbol{\lambda}) = 0, \tag{57}$$

$$h_j(X) \leq 0 \ \ \forall j \in \mathcal{J}_I, \ h_j(X) = 0 \ \ \forall j \in \mathcal{J}_E, \tag{58}$$

$$\lambda_j \geq 0 \ \ \forall j \in \mathcal{J}_I, \tag{59}$$

$$\lambda_j h_j(X) = 0 \ \ \forall j \in \mathcal{J}_I, \tag{60}$$

where

$$\mathcal{L}(X, \boldsymbol{\lambda}) \triangleq f(X) + \sum_{j\in\mathcal{J}} \lambda_j h_j(X) \tag{61}$$

is the Lagrangian with $\lambda_j \in \mathbb{R}$ the Lagrange multiplier corresponding to the constraint h_j and $\boldsymbol{\lambda}$ in bold is used as a shorthand notation for the set of all Lagrange multipliers. These KKT conditions can also be formalized for the local optimization problem (22) defined at the updating node q. Since $\widetilde{X}_q^{i+1}$ solves the local problem (22), it must satisfy the KKT conditions of the local problem, and therefore based on the parameterization in (22), the stationarity condition can be written as

$$\nabla_{\widetilde{X}_q} \mathcal{L}(C_q^i \widetilde{X}_q^{i+1}, \widetilde{\boldsymbol{\lambda}}(q)) = 0, \tag{62}$$

where $\widetilde{\boldsymbol{\lambda}}(q)$ represents the set of Lagrange multipliers corresponding to the local problem (22) at node q and iteration i. Note that (62) contains the Lagrangian of the centralized problem (6), yet it is parameterized based on (22) to transform it into the local problem at node q. From the chain rule with $X = C_q^i \widetilde{X}_q$, we have

$$C_q^{iT} \nabla_X \mathcal{L}(C_q^i \widetilde{X}_q^{i+1}, \widetilde{\boldsymbol{\lambda}}(q)) = 0. \tag{63}$$

The KKT conditions for optimality of the local problem (22) can then be written as

$$C_q^{iT} \nabla_X \Big[f\left(C_q^i \widetilde{X}_q^{i+1}\right) + \sum_{j\in\mathcal{J}} \lambda_j(q) h_j\left(C_q^i \widetilde{X}_q^{i+1}\right) \Big] = 0, \tag{64}$$

$$h_j\left(C_q^i \widetilde{X}_q^{i+1}\right) \leq 0 \ \ \forall j \in \mathcal{J}_I, \ h_j\left(C_q^i \widetilde{X}_q^{i+1}\right) = 0 \ \ \forall j \in \mathcal{J}_E, \tag{65}$$

$$\lambda_j(q) \geq 0 \ \ \forall j \in \mathcal{J}_I, \tag{66}$$

$$\lambda_j(q) h_j\left(C_q^i \widetilde{X}_q^{i+1}\right) = 0 \ \ \forall j \in \mathcal{J}_I, \tag{67}$$

where the $\lambda_j(q)$'s are the Lagrange multipliers at updating node q and iteration i. Since (65) is exactly the same as (58), we conclude that the local primal feasibility condition is also satisfied globally (which we already knew from (23) and [1, Lemma 1]). Let us now look at the three other equations and assume that the algorithm has reached a fixed point, i.e., $X^{i+1} = X^i = \overline{X}$. From this fixed point assumption, we can replace $C_q^i \widetilde{X}_q^{i+1} = X^{i+1}$ with $C_q^i \widetilde{X}_q^i = X^i = \overline{X}$ in (64) such that the local stationarity condition can be rewritten as

$$C_q^{iT} \nabla_X \Big[f(\overline{X}) + \sum_{j\in\mathcal{J}} \lambda_j(q) h_j(\overline{X}) \Big] = 0. \tag{68}$$

Selecting the first M_q rows of (68), we have (note that from (19), the matrix C_q^{iT} selects the $q-$th block-row from X as the first M_q rows)

$$\nabla_{X_q} f(\overline{X}) = -\sum_{j\in\mathcal{J}} \lambda_j(q) \nabla_{X_q} h_j(\overline{X}). \tag{69}$$

Since this result is valid for any node q due to the fixed point assumption, we may stack the variations of equation (69) for every node $q \in \mathcal{K}$:

$$\begin{bmatrix} \nabla_{X_1} f(\overline{X}) \\ \vdots \\ \nabla_{X_K} f(\overline{X}) \end{bmatrix} = \nabla_X f(\overline{X}) = -\begin{bmatrix} \sum_{j\in\mathcal{J}} \lambda_j(1) \nabla_{X_1} h_j(\overline{X}) \\ \vdots \\ \sum_{j\in\mathcal{J}} \lambda_j(K) \nabla_{X_K} h_j(\overline{X}) \end{bmatrix}. \tag{70}$$

Multiplying (68) by $\widetilde{X}_q^i$ defined in (25) and using the fact that $C_q^i \widetilde{X}_q^i = X^i = \overline{X}$ (this follows from (25) and the definition of C_q^i in (19)-(20)), we obtain

$$\overline{X}^T \nabla_X f(\overline{X}) = -\sum_{j\in\mathcal{J}} \lambda_j(q) \overline{X}^T \nabla_X h_j(\overline{X}). \tag{71}$$

From Condition 1a, the set $\{\overline{X}^T \nabla_X h_j(\overline{X})\}_j$ is linearly independent and therefore the Lagrange multipliers $\{\lambda_j(q)\}_j$ that satisfy (71) are unique. Moreover, since the left-hand side of (71) does not depend on the node q, we have that $\lambda_j(q) = \lambda_j$ for any node q. Therefore, (70) becomes

$$\nabla_X f(\overline{X}) = -\sum_{j\in\mathcal{J}} \lambda_j \nabla_X h_j(\overline{X}), \tag{72}$$

which implies that $(\overline{X}, \{\lambda_j\}_j)$ satisfy the global stationarity conditions (57). Since $(\overline{X}, \{\lambda_j\}_j)$ satisfies the local dual feasibility and the local complementary slackness conditions (66) and (67) (with $C_q^i \widetilde{X}_q^{i+1} = X^{i+1}$ replaced by $\overline{X}$ due to it being a fixed point), it also satisfies their global counterparts (59) and (60). Hence, $(\overline{X}, \{\lambda_j\}_j)$ satisfies all the global KKT optimality conditions. This proves that any fixed point of Algorithm 1 is a stationary point of the global problem (6).

□

Appendix B

Proof of Theorem 2. The arguments in this proof are very similar to the ones in the proof of Theorem 1, where the main difference is to show the uniqueness of the Lagrange multipliers when $\{D_{j,q}(\overline{X})\}_j$ is a linearly independent set for any $q\in\mathcal{K}$ at fixed points $\overline{X}=X^{i+1}=X^i$. Therefore we only make changes to that part of the proof.

From the definition of C_q^i in (19)-(20), left-multiplying each block-row of (68) by the corresponding block-column of $\widetilde{X}_q^{(i+1)T}$ as structured in (13), we have

$$X_q^{(i+1)T}\nabla_{X_q}\Big[f(X^i)+\sum_{j\in\mathcal{J}}\lambda_j(q)h_j(X^i)\Big]=0, \tag{73}$$

$$G_n^{(i+1)T}\sum_{k\in\mathcal{B}_{nq}}X_k^{iT}\nabla_{X_k}\Big[f(X^i)+\sum_{j\in\mathcal{J}}\lambda_j(q)h_j(X^i)\Big]=0, \tag{74}$$

$\forall n\in\mathcal{N}_q$. Note that we here again assume that the algorithm has reached a fixed point for which $X^{i+1}=X^i=\overline{X}$ (as this was also assumed to derive (68)), which implies that we can make the substitutions $X_q^{i+1}=X_q^i=\overline{X}_q$ and $G_n^{i+1}=I_Q$ for all $n\in\mathcal{N}_q$ (see (26)) within (73)-(74). By doing so and, from the definition (30) of $D_{j,q}(\overline{X})$, (73) and (74) become

$$C_{X_q}^{iT}\nabla_X f(\overline{X})+\sum_{j\in\mathcal{J}}\lambda_j(q)D_{j,q}(\overline{X})=0, \tag{75}$$

where $C_{X_q}^i$ is the matrix C_q^i but the identity matrix of the first block-column in (19) has been replaced by $X_q^i=\overline{X}_q$. From (75) and the linear independence assumption over the set $\{D_{j,q}(\overline{X})\}_j$ (see Condition 1b), the Lagrange multipliers $\lambda_j(q)$ for all $j\in\mathcal{J}$ that satisfy (75) must be unique. We can repeat the same argument for any updating node, which implies that (73)-(75) holds for any node q, each time with its own unique set of Lagrange multipliers. We will now prove that this unique set of Lagrange multipliers is the same for any updating node q.

Slightly rewriting (73)-(74) (with $X^{i+1}=X^i=\overline{X}$ and $G_n^{i+1}=I_Q$) gives

$$\overline{X}_q^T\nabla_{X_q}f(\overline{X})=-\sum_{j\in\mathcal{J}}\lambda_j(q)\overline{X}_q^T\nabla_{X_q}h_j(\overline{X}), \tag{76}$$

$$\sum_{k\in\mathcal{B}_{nq}}\overline{X}_k^T\nabla_{X_k}f(\overline{X})=-\sum_{j\in\mathcal{J}}\sum_{k\in\mathcal{B}_{nq}}\lambda_j(q)\overline{X}_k^T\nabla_{X_k}h_j(\overline{X}), \tag{77}$$

where (76)-(77) holds for every $q\in\mathcal{K}$ and for all $n\in\mathcal{N}_q$. Substituting (76) into (77), we obtain

$$\begin{aligned}&\sum_{j\in\mathcal{J}}\sum_{k\in\mathcal{B}_{nq}}\lambda_j(k)\overline{X}_k^T\nabla_{X_k}h_j(\overline{X})\\ &=\sum_{j\in\mathcal{J}}\sum_{k\in\mathcal{B}_{nq}}\lambda_j(q)\overline{X}_k^T\nabla_{X_k}h_j(\overline{X}).\end{aligned} \tag{78}$$

Vectorizing the matrices in (78) such that $\mathbf{h}_{j,k}=\mathrm{vec}\left(\overline{X}_k^T\nabla_{X_k}h_j(\overline{X})\right)\in\mathbb{R}^{Q^2}$ and defining $H_k=[\mathbf{h}_{1,k},\dots,\mathbf{h}_{J,k}]\in\mathbb{R}^{Q^2\times J}$ $\forall k\in\mathcal{K}$, we obtain

$$\sum_{k\in\mathcal{B}_{nq}}H_k\boldsymbol{\lambda}(k)=\left(\sum_{k\in\mathcal{B}_{nq}}H_k\right)\boldsymbol{\lambda}(q), \tag{79}$$

where $\boldsymbol{\lambda}(k)=[\lambda_1(k),\dots,\lambda_J(k)]^T$. At node q and for its corresponding neighbor $n\in\mathcal{N}_q$, we can then write the following linear system of equation:

$$\mathbf{H}_{nq}\cdot\boldsymbol{\lambda}_{\mathcal{K}}=0, \tag{80}$$

where $\boldsymbol{\lambda}_{\mathcal{K}}=[\boldsymbol{\lambda}^T(1),\dots,\boldsymbol{\lambda}^T(K)]^T\in\mathbb{R}^{KJ}$ and $\mathbf{H}_{nq}\in\mathbb{R}^{Q^2\times KJ}$ is a block-column matrix where the $l-$th block of size $Q^2\times J$ is given by

$$\mathbf{H}_{nq}(l)=\begin{cases}-\sum_{k\in\mathcal{B}_{nq}}H_k & \text{if } l=q\\ H_l & \text{if } l\in\mathcal{B}_{nq}\\ 0 & \text{otherwise.}\end{cases} \tag{81}$$

Stacking vertically the matrices $\mathbf{H}_{nq}$, for every neighbor $n\in\mathcal{N}_q$ and every node $q\in\mathcal{K}$, results in $\mathbf{H}\in\mathbb{R}^{Q^2\sum_k|\mathcal{N}_k|\times KJ}$ and we write

$$\mathbf{H}\cdot\boldsymbol{\lambda}_{\mathcal{K}}=0. \tag{82}$$

Note that from (81), the sum of all $Q^2\times J$ block-columns of $\mathbf{H}_{nq}(l)$ must be equal to the zero matrix. Therefore, every $\boldsymbol{\lambda}_{\mathcal{K}}$ such that $\boldsymbol{\lambda}(1)=\cdots=\boldsymbol{\lambda}(K)$ is in the null space of $\mathbf{H}$ and would satisfy (82). The dimension of the set $\{\boldsymbol{\lambda}_{\mathcal{K}}\in\mathbb{R}^{KJ}\mid\boldsymbol{\lambda}(1)=\cdots=\boldsymbol{\lambda}(K)\}$ is equal to J and therefore $\mathrm{rank}(\mathbf{H})\le KJ-J$. To ensure that these are the only solutions of (82), we require $\mathrm{rank}(\mathbf{H})=KJ-J$. Note that a necessary condition to satisfy this is that $KJ-J\le Q^2\sum_k|\mathcal{N}_k|$, i.e., $KJ-J$ is less than the number of rows of $\mathbf{H}$, or equivalently $J\le\frac{Q^2}{K-1}\sum_k|\mathcal{N}_k|$, leading to the upper bound given in (32).

Lemma 3. *If Condition 1b holds, then rank*$(\mathbf{H})=KJ-J$.

Proof. The proof of this lemma is provided as supplementary material to the paper as it is too elaborate to fit in this main text. □

Since $\mathrm{rank}(\mathbf{H})=KJ-J$, the set $\{\boldsymbol{\lambda}_{\mathcal{K}}\in\mathbb{R}^{KJ}\mid\boldsymbol{\lambda}(1)=\cdots=\boldsymbol{\lambda}(K)\}$ contains the full null space of $\mathbf{H}$ and hence all the solutions of (82). Because we now have established that all the Lagrange multipliers are the same, we can conclude that (70) holds in this case as well, allowing to obtain the same result as in (72). The remaining arguments of the proof of Theorem 1 can be applied here to conclude that each fixed point is a point satisfying the KKT conditions (57)-(60) of the original problem (6). □

Appendix C

Proof of Theorem 3. From Corollary 1, all points in $(X^i)_i$ remain in a compact set. Since each compact set has at least one accumulation point $\overline{X}$, there exists an infinite subsequence of $(X^i)_i$ which converges to $\overline{X}$. Because the number of nodes is finite, there exists a node $k\in\mathcal{K}$ that acts as an updating node in an infinite number of iterations that are sampled in this subsequence. In other words, we can find some node k such that there exists a set of iteration indices $\mathcal{I}_k\subseteq\mathbb{N}$ such that $(X^i)_{i\in\mathcal{I}_k}$ converges to $\overline{X}$ and $(i \mod K)_{i\in\mathcal{I}_k}=(k)_{i\in\mathcal{I}_k}$, i.e., the iteration indices $\mathcal{I}_k$ correspond only to iterations where node k is the updating node in Algorithm 1. From Berge's Maximum Theorem [24], the continuity of $\widetilde{\mathcal{F}}_k$ and $g(W,V) \triangleq ||W-V||_F$ implies that $\widetilde{\mathcal{M}}_k$ and thus $\mathcal{M}_k = C_k\widetilde{\mathcal{M}}_k$ as in (38) are upper semicontinuous. This implies that any accumulation point $\overline{X}^{(+1)}$ of the set $(X^{i+1})_{i\in\mathcal{I}_k}$ must satisfy

$$\overline{X}^{(+1)} \in \mathcal{M}_k(\overline{X}). \tag{83}$$

As from Lemma 4 (see end of this appendix), we have $\mathcal{M}_k(\overline{X}) = \{\overline{X}\}$ (i.e., $\overline{X}$ is a fixed point of $\mathcal{M}_k$), (83) implies that

$$\overline{X}^{(+1)} = \overline{X}. \tag{84}$$

Inductively applying the above argument for the new subsequence $(X^{i+1})_{i\in\mathcal{I}_k}$ and accumulation point $\overline{X}^{(+1)}$ and for node $k+1 \mod K$ yields

$$\overline{X}^{(+l)} = \overline{X} \quad \forall l \geq 0. \tag{85}$$

As $\overline{X}^{(+l)}$ is *any* accumulation point $(X^{i+l})_{i\in\mathcal{I}_k}$, all the accumulation points of $(X^{i+l})_{i\in\mathcal{I}_k}$ are equal to $\overline{X}$ and all the sequences

$$(X^i)_{i\in\mathcal{I}_{(k+l \mod K)}} \triangleq (X^{i+l})_{i\in\mathcal{I}_k} \tag{86}$$

converge to the same point $\overline{X}$. From Lemma 4 (here applied to the node $k+l \mod K$ instead of k), $\overline{X}^{(+l)}$ is a fixed point of $\mathcal{M}_{(q+l \mod K)}$ and $\overline{X}$ is therefore a fixed point of $\mathcal{M}_k$ for any k, proving the first part of the theorem.

We now prove that $\lim_{i\to+\infty} ||X^{i+1} - X^i||_F = 0$ by contradiction. Let us assume that the above statement is not true. We first note that $X, W \to ||X-W||_F$ is a continuous mapping and X^i and X^{i+1} both live in a compact set (see beginning of the proof). Since the continuous image of a compact set is itself a compact set, $(||X^{i+1} - X^i||_F)_i$ has at least one accumulation point. There must therefore be some index set $\mathcal{I}$ such that

$$\lim_{i\in\mathcal{I}\to\infty} ||X^{i+1} - X^i||_F > 0, \tag{87}$$

that is, there is one convergent subsequence converging to a point different from zero. Indeed, if zero was the only accumulation point, the sequence would be convergent (see Lemma 5). Furthermore, based on the same reasoning as the beginning of this proof, there is some $\mathcal{I}'_k \subseteq \mathcal{I}$ such that $(X^i)_{i\in\mathcal{I}'_k}$ is a convergent sequence such that $(i \mod K)_{i\in\mathcal{I}'_k} = (k)_{i\in\mathcal{I}'_k}$ for some k. We have shown above that the convergence of such a sequence $(X^i)_{i\in\mathcal{I}'_k}$ implied that

$$\lim_{i\in\mathcal{I}'_k\to\infty} X^i = \lim_{i\in\mathcal{I}'_k\to\infty} X^{i+1}. \tag{88}$$

Therefore, by continuity of the Frobenius norm, it must be that

$$\lim_{i\in\mathcal{I}'_k\to\infty} ||X^{i+1} - X^i||_F = 0. \tag{89}$$

As (89) contradicts (87), every convergent subsequence of $(||X^{i+1} - X^i||_F)_{i\in\mathbb{N}}$ converges to 0 and $(||X^{i+1} - X^i||_F)_{i\in\mathbb{N}}$ is therefore a convergent sequence.

□

Proof of Theorem 4. Let us assume that the mapping (39) corresponding to the DASF algorithm has a finite set of fixed points denoted Φ. As the set of fixed points is finite, it must be that there exists some $\delta > 0$ such that for any pair of fixed points $\overline{X}, \overline{W}$ $||\overline{X} - \overline{W}||_F > \delta$. From Lemma 5, as the sublevel sets of f are compact, $(X^i)_i$ converges to the set of its accumulation points $\mathcal{A}$. From Theorem 3, this set $\mathcal{A}$ must be a subset of Φ, and therefore finite.

We will now show that $\mathcal{A}$ must be a singleton. From Lemma 5, we have

$$\forall\varepsilon, \exists i_\varepsilon > 0 : \inf_{W\in\mathcal{A}} ||W - X^i||_F < \varepsilon, \quad \forall i > i_\varepsilon. \tag{90}$$

From Theorem 3, we have

$$\forall\varepsilon, \exists i_\varepsilon > 0 : ||X^{i+1} - X^i||_F < \varepsilon, \quad \forall i > i_\varepsilon. \tag{91}$$

From (90) and (91), there exists an $i_\varepsilon > 0$ such that

$$\begin{aligned} \forall i > i_\varepsilon\ \exists \overline{X}, \overline{X}^{(+1)} \in \mathcal{A} : &||\overline{X} - X^i||_F < \delta/3, \\ &||\overline{X}^{(+1)} - X^{i+1}||_F < \delta/3, \\ &||X^{i+1} - X^i||_F < \delta/3. \end{aligned} \tag{92}$$

We then have from the triangle inequality

$$\begin{aligned} ||\overline{X}^{(+1)} - \overline{X}||_F &\leq ||\overline{X}^{(+1)} - X^i||_F \\ &+ \left\|X^{i+1} - X^i\right\|_F + ||X^{i+1} - \overline{X}^{(+1)}||_F < \delta. \end{aligned} \tag{93}$$

If $\overline{X} \neq \overline{X}^{(+1)}$ then this would imply that $||\overline{X}^{(+1)} - \overline{X}||_F > \delta$ (by definition of δ). However, this would be a contradiction with (93). Therefore, $\overline{X}$ and $\overline{X}^{(+1)}$ must be equal. Since (92) holds for any i, we find by induction that $\mathcal{A}$ is a singleton. From Lemma 5, this results in the convergence of $(X^i)_i$. □

Lemma 4. *Let $\mathcal{I}_k \subseteq \mathbb{N}$ be such that $(X^i)_{i\in\mathcal{I}_k}$ converges to $\overline{X}$ and $(i \mod K)_{i\in\mathcal{I}_k} = (k)_{i\in\mathcal{I}_k}$, i.e., we only consider iterates related to some node k. Then if $\widetilde{\mathcal{F}}_k : \mathbb{R}^{M\times Q} \rightrightarrows \mathbb{R}^{\widetilde{M}_Q\times Q}$ is a continuous mapping*[7]*, $\overline{X}$ is a fixed point of the map $\mathcal{M}_k$, i.e., $\mathcal{M}_k(\overline{X}) = \{\overline{X}\}$.*

Proof. From Corollary 1, all points in $(X^i)_i$ remain in a compact set therefore $(X^{i+1})_{i\in\mathcal{I}_k}$ has an accumulation point $\overline{X}^{(+1)}$. The continuity of $\widetilde{\mathcal{F}}_k$, and thus of $C_k(X)\widetilde{\mathcal{F}}_k$, implies that it is also upper semicontinuous. Therefore we have (by definition, see [17])

$$\overline{X}^{(+1)} \in C_k(\overline{X})\widetilde{\mathcal{F}}_k(\overline{X}). \tag{94}$$

We can now prove that

$$\min_{W\in\mathcal{S}_k(\overline{X})} f(W) = f(\overline{X}^{(+1)}) = f(\overline{X}). \tag{95}$$

The first equality directly follows from (94) and the definition (35) of $\widetilde{\mathcal{F}}_k$, that is

$$\min_{W\in\mathcal{S}_k(\overline{X})} f(W) = \min_{\widetilde{W}_q : C_q(\overline{X})\widetilde{W}_q\in\mathcal{S}_k(\overline{X})} f(C_q(\overline{X})\widetilde{W}_q) = f(X) \tag{96}$$

$\forall X \in C_k(\overline{X})\widetilde{\mathcal{F}}_k(\overline{X})$. The second equality in (95) follows from the fact that $\overline{X}$ is an accumulation point and that f is continuous together with the fact that $(f(X^i))_i$ is monotonically decreasing (Lemma 1) (i.e., all accumulation points have the same objective value). Because of (95), and since $\overline{X}$ is by definition in $\mathcal{S}_k(\overline{X})$, it must be that $\overline{X} \in C_k(\overline{X})\widetilde{\mathcal{F}}_k(\overline{X})$ and

[7] As seen in the proof of Lemma 4, this can be relaxed to upper semicontinuity, but we keep the continuity condition for consistency with previous results.

thus $[\overline{X}_k^T, I_Q, \dots, I_Q]^T \in \widetilde{\mathcal{F}}_k(\overline{X})$. Using this in (37) with X replaced by $\overline{X}$ results in $\{\overline{X}\} = \mathcal{M}_k(\overline{X})$. □

Lemma 5. *In a compact metric space, a sequence converges to the set of its accumulation points. Additionally, the sequence converges if and only if it has a single accumulation point.*

Proof. Let $(X^i)_i$ be some sequence in a compact metric space. Let $\mathcal{A}$ denote the set of accumulation points of $(X^i)_i$. We have $\forall X^* \in \mathcal{A}, \exists \mathcal{I} \subseteq \mathbb{N} : \forall \varepsilon > 0, \exists i_\varepsilon > 0 : ||X^* - X^i||_F < \varepsilon, \quad \forall i \in \mathcal{I} > i_\varepsilon$. We wish to prove that $\forall \varepsilon > 0, \exists i_\varepsilon > 0 : \inf_{X\in\mathcal{A}} ||X - X^i||_F < \varepsilon, \quad \forall i > i_\varepsilon$. Let us assume that our claim is not true. Then

$$\exists \varepsilon > 0, \forall i_\varepsilon > 0, \exists i > i_\varepsilon : \inf_{X\in\mathcal{A}} ||X - X^i||_F \geq \varepsilon. \tag{97}$$

which implies that there exists some infinite set $\mathcal{I} \subseteq \mathbb{N}$ such that

$$\exists i_\varepsilon > 0 : \inf_{X\in\mathcal{A}} ||X - X^i||_F \geq \varepsilon, \quad \forall i \in \mathcal{I} > i_\varepsilon. \tag{98}$$

Since the space is compact, the subsequence $(X^i)_{i\in\mathcal{I}}$ has itself a convergent subsequence converging to a point in $\mathcal{A}$, contradicting (98).

The convergence in the case of a single accumulation point follows directly from the previous result and the converse is a well-known result. □

References

[1] C. A. Musluoglu and A. Bertrand, "A unified algorithmic framework for distributed adaptive signal and feature fusion problems — Part I: Algorithm derivation," 2022.

[2] C. A. Musluoglu, C. Hovine, and A. Bertrand, "A unified algorithmic framework for distributed adaptive signal and feature fusion problems — Part II: Convergence properties: Supplementary material," 2022.

[3] K. Liu, Y.-Q. Cheng, and J.-Y. Yang, "A generalized optimal set of discriminant vectors," *Pattern Recognition*, vol. 25, no. 7, pp. 731–739, 1992.

[4] M. Borga, "Learning multidimensional signal processing," Ph.D. dissertation, Linköping University Electronic Press, 1998.

[5] D. R. Hardoon, S. Szedmak, and J. Shawe-Taylor, "Canonical correlation analysis: An overview with application to learning methods," *Neural computation*, vol. 16, no. 12, pp. 2639–2664, 2004.

[6] B. D. Van Veen and K. M. Buckley, "Beamforming: A versatile approach to spatial filtering," *IEEE ASSP magazine*, vol. 5, no. 2, pp. 4–24, 1988.

[7] J. Hadamard, "Sur les problèmes aux dérivées partielles et leur signification physique," *Princeton university bulletin*, pp. 49–52, 1902.

[8] Y.-h. Zhou, J. Yu, H. Yang, and S.-w. Xiang, "Hadamard types of well-posedness of non-self set-valued mappings for coincide points," *Nonlinear Analysis: Theory, Methods & Applications*, vol. 63, no. 5-7, pp. e2427–e2436, 2005.

[9] D. W. Peterson, "A review of constraint qualifications in finite-dimensional spaces," *Siam Review*, vol. 15, no. 3, pp. 639–654, 1973.

[10] C. A. Musluoglu and A. Bertrand, "Distributed adaptive trace ratio optimization in wireless sensor networks," *IEEE Transactions on Signal Processing*, vol. 69, pp. 3653–3670, 2021.

[11] ——, "DASF Toolbox," 2022. [Online]. Available: https://github.com/AlexanderBertrandLab/DASF_toolbox

[12] P.-A. Absil, R. Mahony, and B. Andrews, "Convergence of the iterates of descent methods for analytic cost functions," *SIAM Journal on Optimization*, vol. 16, no. 2, pp. 531–547, 2005.

[13] A. A. Goldstein, "On steepest descent," *Journal of the Society for Industrial and Applied Mathematics, Series A: Control*, vol. 3, no. 1, pp. 147–151, 1965.

[14] R. Fletcher, *Practical methods of optimization*. John Wiley & Sons, 2013.

[15] J. Nocedal and S. Wright, *Numerical optimization*. Springer Science & Business Media, 2006.

[16] A. R. Conn, N. I. Gould, and P. L. Toint, *Trust region methods*. SIAM, 2000.

[17] D. Charalambos and B. Aliprantis, *Infinite Dimensional Analysis: A Hitchhiker's Guide*. Springer-Verlag Berlin and Heidelberg GmbH & Company KG, 2013.

[18] A. Ben-Israel and T. N. Greville, *Generalized inverses: theory and applications*. Springer Science & Business Media, 2003, vol. 15.

[19] K. Fukunaga, *Introduction to statistical pattern recognition*. Elsevier, 2013.

[20] T. Kato, *Perturbation theory for linear operators*. Springer Science & Business Media, 2013, vol. 132.

[21] H. Wang, S. Yan, D. Xu, X. Tang, and T. Huang, "Trace ratio vs. ratio trace for dimensionality reduction," in *2007 IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, 2007, pp. 1–8.

[22] T. T. Ngo, M. Bellalij, and Y. Saad, "The trace ratio optimization problem," *SIAM review*, vol. 54, no. 3, pp. 545–569, 2012.

[23] D. P. Bertsekas, *Nonlinear Programming: 2nd Edition*. Athena Scientific, 1999.

[24] C. Berge, *Topological Spaces: including a treatment of multi-valued functions, vector spaces, and convexity*. Courier Corporation, 1997.