THE INSTITUTE FOR SYSTEMS RESEARCH

ISR TECHNICAL REPORT 2015-01

Nonlinear Programming Methods for Distributed Optimization

Ion Matei
John Baras

The Institute for Systems Research

A. JAMES CLARK SCHOOL OF ENGINEERING

ISR develops, applies and teaches advanced methodologies of design and analysis to solve complex, hierarchical, heterogeneous and dynamic problems of engineering technology and systems for industry and government.

ISR is a permanent institute of the University of Maryland, within the A. James Clark School of Engineering. It is a graduated National Science Foundation Engineering Research Center.

www.isr.umd.edu

Nonlinear Programming Methods for Distributed Optimization

Ion Matei, John S. Baras

Abstract

In this paper we investigate how standard nonlinear programming algorithms can be used to solve constrained optimization problems in a distributed manner. The optimization setup consists of a set of agents interacting through a communication graph that have as common goal the minimization of a function expressed as a sum of (possibly non-convex) differentiable functions. Each function in the sum corresponds to an agent and each agent has associated an equality constraint. By re-casting the distributed optimization problem into an equivalent, augmented centralized problem, we show that distributed algorithms result naturally from applying standard nonlinear programming techniques. Due to the distributed formulation, the standard assumptions and convergence results no longer hold. We emphasize what changes are necessary for convergence to still be achieved for three algorithms: two algorithms based on Lagrangian methods, and an algorithm based the method of multipliers. The changes in the convergence results are necessary mainly due to the fact that the (local) minimizers of the lifted optimization problem are not regular, as a results of the distributed formulation. Unlike the standard algorithm based on the method of multipliers, for the distributed version we cannot show that the theoretical superlinear convergence rate can be achieved.

I. Introduction

Multi-agent, distributed optimization algorithms received a lot of attention in the recent years due to their applications in network resource allocation, collaborative control, estimation and identification problems. In these type of problems a group of agents has as common goal the optimization of a cost function under limited information and resources. The limited information may be induced by the fact that an agent can communicate with only a subset of the total set of agents, or/and by the fact that an agent is aware of only a part of the cost functions or constraint sets.

A distributed optimization algorithm was introduced in [20], where the convex optimization cost is expressed as a sum of functions and each function in the sum corresponds to an agent. In this formulation the agents interact with each other subject to a communication network, usually modeled as a undirected graph. The algorithm combines a standard (sub)gradient descent step with a consensus step; the latter being added to deal with the limited information about the cost function and about the actions of the agents. Extensions of this initial version followed in the literature. [17], [21] include communication noise and errors on subgradients, [12], [14] assume a random communication

Ion Matei is with the Palo Alto Research Center (PARC), Palo Alto, CA 94304 (ion.matei@parc.com) and John S. Baras is with the Institute for Systems Research at University of Maryland, College Park, MD 20742 (baras@umd.edu).

graph, [17], [24] study asynchronous versions of the algorithm, [13] considers state-dependent communication topologies, while [5] assumes directed communication graphs. Another modification of the algorithm described in [20] was introduced in [10], where the authors change the order in which the consensus-step and the subgradient descent step are executed. The algorithms discussed above became popular in the signal processing community as well, being used for solving distributed filtering and parameter identification problems [4], [22]. Consensus-based distributed optimization algorithms were further used to solve constrained convex optimization problems where all agents have the same constraint set [11], [17], [21] or where each agent has its own set of constraints [18], [24]. Other approaches for obtaining distributed algorithms use dual decomposition [25], augmented Lagragian [8], [9], or in particular, distributed versions of the Alternating Direction Method of Multipliers (ADMM) algorithm [3], [23], [27]. A summary of the relevant problem setups and approaches concerning distributed optimization is shown in Table I.

Cost type	**Constraints type**	**Communication graph**		**Approach**	**Authors**
convex	unconstrained	time invariant	undirected	combination of sub-gradient and consensus steps	[10], [20]
				augmented Lagrangian method with a consensus step included in the inner step	[8]
				randomized incremental subgradient method	[11]
				alternating direction method of multipliers	[3], [23], [27]
			directed	Laplacian-gradient dynamics	[5]
		time varying	random	combination of sub-gradient and consensus steps	[12], [14]
			state dependent	combination of sub-gradient and consensus steps	[13]
	global convex set	time invariant	undirected	combination of (projected) sub-gradient (with stochastic errors) and consensus steps	[17], [21]
			undirected bounded communication delays	dual decomposition	[25]
	local convex sets	time invariant, undirected		combination of (projected) sub-gradient and consensus steps	[19]
				augmented Lagrangian method with a gossip step included in the inner step	[8]
non-convex	global inequality constraints	time varying, periodic strong connectivity		approximate solution obtained using a Lagrangian duality method combined with consensus step	[28]
	local equality constraints	time invariant, undirected		(augmented)Lagrangian methods	[15], [16]

TABLE I: Problem setups and approaches for distributed optimization

In this paper we study as well a distributed optimization problem whose goal is to minimize an objective function expressed as a sum of functions. Each function in the sum is associated to an agent that has assigned an equality constraint, as well. We propose three distributed algorithms: two first-order algorithms for solving the first order necessary optimality conditions and an algorithm inspired by the method of multipliers. The second

first-order algorithm uses an augmented Lagrangian idea to obtain weaker conditions for local convergence. The main message of this paper is that standard optimization techniques can be used to solve optimization problems in a distributed manner **as is**, provided appropriate changes in the convergence proofs are made to deal with the fact that the standard assumptions no longer hold as a result of lack of complete information. We make no convexity assumptions on the cost and constraint functions, but we assume they are continuously differentiable. Consequently, our convergence results are local. Distributed algorithms for solving constrained, non-convex optimization problems were also proposed in [15] and [28]. This paper is based on two conference papers [15], [16] where only preliminary results are shown and most of the proofs were omitted due to space restrictions.

The paper is organized as follows: in Section II we formulate the constrained optimization problem while in Section III we introduced three distributed optimization algorithms for solving optimization problems with equality constraints. Section IV presents the origins of the algorithms by demonstrating that our initial optimization problem is equivalent to a lifted optimization problem with equality constraints. Sections V and VI. Some proofs and supporting results are included in the Appendix.

Notations and definitions: For a matrix A, its (i,j) entry is denoted by $[A]_{ij}$ and its transpose is given by A'. If A is a symmetric matrix, $A \succ 0$ ($A \succeq 0$) means that A is positive (semi-positive) definite. The nullspace and range of A are denoted by Null(A) and Range(A), respectively. The symbol $\otimes$ is used to represent the Kronecker product between two matrices. The vector of all ones is denoted by $\mathbb{1}$. Let x and Q be a vector and a set of vectors, respectively. By $x+Q$ we understand the set of vectors produced by adding x to each element of Q, that is, $x+Q \triangleq \{x+y \mid y \in Q\}$. Let $\|\cdot\|$ be a vector norm. By $\|x-Q\|$ we denote the distance between the vector x and the set Q, that is, $\|x-Q\| \triangleq \inf_{y\in Q}\|x-y\|$. Let $f:\mathbb{R}^n \to \mathbb{R}$ be a function. We denote by $\nabla f(x)$ and by $\nabla^2 f(x)$ the gradient and the Hessian of f at x, respectively. Let $F:\mathbb{R}^n\times\mathbb{R}^m \to \mathbb{R}$ be a function of variables (x,y). The block descriptions of the gradient and of the Hessian of F at (x,y) are given by $\nabla F(x,y)' = \left(\nabla_x F(x,y)', \nabla_y F(x,y)'\right)$, and

$$\nabla^2 F(x,y) = \begin{pmatrix} \nabla^2_{xx}F(x,y) & \nabla^2_{xy}F(x,y) \\ \nabla^2_{xy}F(x,y) & \nabla^2_{yy}F(x,y) \end{pmatrix},$$

respectively. Let $\{A_i\}_{i=1}^N$ be a set of matrices. By diag(A_i, $i=1,\ldots,N$) we understand a block diagonal matrix, where the ith block matrix is given by A_i. We say that the set $\mathcal{X}$ is an *attractor* for the dynamics $x_{k+1} = f(x_k)$, if there exists $\epsilon > 0$, such that for any $x_0 \in S_\epsilon$, with $S_\epsilon = \{x \mid \|x-\mathcal{X}\| < \epsilon\}$, $\lim_{k\to\infty}\|x_k - \mathcal{X}\| = 0$.

II. Problem description

In this section we describe the setup of our problem. We present first the communication model followed by the optimization model.

A. Communication model

A set of N agents interact with each other through a communication topology modeled as an undirected communication graph $\mathcal{G} = (\mathcal{V},\mathcal{E})$, where $\mathcal{V} = \{1,2,\ldots,N\}$ is the set of nodes and $\mathcal{E} = \{e_{ij}\}$ is the set of edges. An edge between two nodes i and j means that agents i and j can exchange information (or can cooperate). We

assume that at each time instant k the agents can synchronously exchange information with their neighbors. We denote by $\mathcal{N}_i \triangleq \{j \mid e_{ij} \in \mathcal{E}\}$ the set of neighbors of agent i. Consider the set of pairs $\{(i,j), j \in \mathcal{N}_i,\ i = 1,\ldots,N\}$ and let $\bar{N} = \sum_{i=1}^{N} |\mathcal{N}_i|$, where $|\cdot|$ denotes the cardinality of a set. We denote by $S \in \mathbb{R}^{\bar{N}\times N}$ the (weighted) edge-node incidence matrix of graph $\mathcal{G}$, for which each row number corresponds to a unique pair (i,j) from the previously defined set. The matrix S is defined as

$$[S]_{(ij),l} = \begin{cases} s_{ij} & i = l, \\ -s_{ij} & j = l, \\ 0 & \text{otherwise}, \end{cases} \tag{1}$$

where s_{ij} are given positive scalars.

Remark 2.1: It is not difficult to observe that the matrix $L = S'S = (l_{ij})$ is a (weighted) Laplacian matrix corresponding to the graph $\mathcal{G}$ and $\text{Null}(L) = \text{Null}(S)$. Moreover, for any $i \neq j$ we have that $l_{ij} = -(s_{ij}^2 + s_{ji}^2)$. □

In the next sections we are going to make use of a set of properties of the matrices S and L; properties that are grouped in what follows.

Proposition 2.1: The matrix S and L defined with respect to a connected graph $\mathcal{G}$ satisfies the following properties:

(a) The nullspaces of S and L are given by $\text{Null}(S) = \text{Null}(L) = \{\gamma\mathbb{1} \mid \gamma \in \mathbb{R}\}$;

(b) Let $\mathbf{S} = S \otimes I$ and $\mathbf{L} = L \otimes I$, where I is the n-dimensional identity matrix. Then the nullspaces of $\mathbf{S}$ and $\mathbf{L}$ are given by $\text{Null}(\mathbf{S}) = \text{Null}(\mathbf{L}) = \{\mathbb{1} \otimes x \mid x \in \mathbb{R}^n\}$.

(c) Let $\boldsymbol{\lambda}$ be a vector in $\mathbb{R}^{n\bar{N}}$ and let $\{u_1,...,u_m\}$ be an orthogonal basis spanning $\text{Null}(S')$. Then the orthogonal projection of $\boldsymbol{\lambda}$ on $\text{Null}(\mathbf{S}')$ is given by $\boldsymbol{\lambda}_\perp = \mathbf{J}\boldsymbol{\lambda}$, where $\mathbf{J}$ is the orthogonal projection matrix (operator) defined as

$$\mathbf{J} \triangleq J \otimes I,$$

where $J = UU'$ with $U = [u_1,...,u_m]$ is the orthogonal projection operator on $\text{Null}(S')$.

B. Optimization model

We consider a function $f : \mathbb{R}^n \to \mathbb{R}$ expressed as a sum of N functions $f(x) = \sum_{i=1}^{N} f_i(x)$, and a vector-valued function $h : \mathbb{R}^n \to \mathbb{R}^N$ where $h \triangleq (h_1, h_2, \ldots, h_N)'$, with $h_i : \mathbb{R}^n \to \mathbb{R}$ and $N \leq n$.

We make the following assumptions on the functions f and h and on the communication model.

Assumption 2.1: (a) The functions $f_i(x)$ and $h_i(x)$, $i = 1,\ldots,N$ are twice continuously differentiable;

(b) Agent i has knowledge of only functions $f_i(x)$ and $h_i(x)$, and scalars s_{ij}, for $j \in \mathcal{N}_i$;

(c) Agent i can exchange information only with agents in the set of neighbors defined by $\mathcal{N}_i$;

(d) The communication graph $\mathcal{G}$ is connected.

The common goal of the agents is to minimize the following optimization problem with equality constraints

$$\begin{aligned} (\mathrm{P}_1) \qquad & \min_{x\in\mathbb{R}^n} \quad f(x), \\ & \text{subject to:} \quad h(x) = 0, \end{aligned}$$

under Assumptions 2.1. Throughout the rest of the paper we assume that problem ($\mathtt{P}_1$) has at least one local minimizer.

Let x^* be a local minimizer of ($\mathtt{P}_1$) and let $\nabla h(x^*) \triangleq [\nabla h_1(x^*), \nabla h_2(x^*), \ldots, \nabla h_N(x^*)]$ be a matrix whose columns are the gradients of the functions $h_i(x)$ computed at x^*. The following assumption is used to guarantee the uniqueness of the Lagrange multiplier vector ψ^* appearing in the first order necessary conditions of ($\mathtt{P}_1$), namely $\nabla f(x^*) + \nabla h(x^*)\psi^* = 0$.

Assumption 2.2: Let x^* be a local minimizer of ($\mathtt{P}_1$). The matrix $\nabla h(x^*)$ is full rank, or equivalently, the vectors $\{\nabla h_i(x^*)\}_{i=1}^N$ are linearly independent.

Together with some additional assumptions on $f(x)$ and $h(x)$, Assumption 2.2 is typically used to prove local convergence for the "original" method of multipliers applied to Problem ($\mathtt{P}_1$) (see for example Section 2.2, page 104 of [2]). As we will see in the next sections, the same assumption will be used to prove local convergence for a distributed version of the method of multipliers used to solved a 'lifted' optimization problem with equality constrains.

Remark 2.2: We assumed that each agent has an equality constraint of the type $h_i(x) = 0$. All the results presented in what follows can be easily adapted for the case where only $m \leq N$ agents have equality constraints.

III. Main results

In this section we present the main results of the paper, namely three distributed algorithms for solving ($\mathtt{P}_1$). As seen later in the paper, these algorithm are a result of applying Lagrangian methods and the method of multipliers to an augmented but equivalent version of ($\mathtt{P}_1$). The algorithms are used to solve the equivalent problem in a centralized manner, however due to the nature of the cost function and the constraints, they can be naturally implemented in a distributed manner.

A. Distributed Algorithms

Let x^* be a local minimizer of ($\mathtt{P}_1$) and let $x_{i,k}$ denote agent i's *estimate* of x^*, at time-slot k. In addition, let us denote by $\mathbf{x}_k \in \mathbb{R}^{nN}$, $\boldsymbol{\mu}_k \in \mathbb{R}^N$ and $\boldsymbol{\lambda}_k \in \mathbb{R}^{n\bar{N}}$ the vectors $\mathbf{x}_k = (x_{i,k})$, $\boldsymbol{\mu}_k = (\mu_{i,k})$ and $\boldsymbol{\lambda}_k = (\lambda_{i,k})$, with $\lambda_{i,k} = (\lambda_{ij,k})$ for all $j \in \mathcal{N}_i$. The first algorithm based on Lagrangian methods for solving ($\mathtt{P}_1$), denoted by Algorithm ($\mathtt{A}_1$) is given by the following iterations:

$$
\begin{aligned}
x_{i,k+1} &= x_{i,k} - \alpha\nabla f_i(x_{i,k}) - \alpha\mu_{i,k}\nabla h_i(x_{i,k}) - && (2)\\
&- \alpha\sum_{j\in\mathcal{N}_i}\Big(s_{ij}\lambda_{ij,k} - s_{ji}\lambda_{ji,k}\Big),\ x_{i,0} = x_i^0,\\
\mu_{i,k+1} &= \mu_{i,k} + \alpha h_i(x_{i,k}),\ \mu_{i,0} = \mu_i^0, && (3)\\
\lambda_{ij,k+1} &= \lambda_{ij,k} + \alpha\Big(s_{ij}x_{i,k} - s_{ji}x_{j,k}\Big),\ \lambda_{ij,0} = \lambda_{ij}^0,\ j\in\mathcal{N}_i, && (4)
\end{aligned}
$$

where $\alpha > 0$ is the step-size of the algorithm, $\nabla f_i(x_{i,k})$ and $\nabla h_i(x_{i,k})$ denote the gradients of functions $f_i(x)$ and $h_i(x)$, respectively, computed at $x_{i,k}$, and x_i^0, μ_i^0 and λ_{ij}^0 are given scalars. In addition, the positive scalars s_{ij} are the

entries of the incidence matrix S of the graph $\mathcal{G}$ defined in (1).

In the second algorithm referred to as Algorithm ($\mathtt{A}_2$), the iteration for updating agent i^{th} estimate $x_{i,k}$ has two additional terms that will allow for weaker assumptions for proving convergence of the algorithm, compared to the previous algorithm:

$$\begin{aligned} x_{i,k+1} &= x_{i,k} - \alpha\nabla f_i(x_{i,k}) - \alpha\mu_{i,k}\nabla h_i(x_{i,k}) - \\ &- \alpha\sum_{j\in\mathcal{N}_i}\Big(s_{ij}\lambda_{ij,k} - s_{ji}\lambda_{ji,k}\Big) - \alpha c h_i(x_{i,k})\nabla h_i(x_{i,k}) \\ &- \alpha c\sum_{j\in\mathcal{N}_i} l_{ij}(x_{i,k} - x_{j,k}),\ x_{i,0} = x_i^0, \end{aligned} \tag{5}$$

$$\mu_{i,k+1} = \mu_{i,k} + \alpha h_i(x_{i,k}),\ \mu_{i,0} = \mu_i^0, \tag{6}$$

$$\lambda_{ij,k+1} = \lambda_{ij,k} + \alpha\Big(s_{ij}x_{i,k} - s_{ji}x_{j,k}\Big),\ \lambda_{ij,0} = \lambda_{ij}^0,\ j\in\mathcal{N}_i, \tag{7}$$

where in addition to the parameters of Algorithm ($\mathtt{A}_1$), we have a new positive parameter c, and $l_{ij} = s_{ij}^2 + s_{ji}^2$.

Finally the third distributed algorithm denoted by Algorithm ($\mathtt{A}_3$) has at its origin the method of multipliers applied to the augmented and equivalent version of ($\mathtt{P}_1$). The iterations of the algorithm are given by

$$\begin{aligned} \mathbf{x}_k &= \arg\min_{\mathbf{x}}\sum_i f_i(x_i) + \mu_{i,k}h_i(x_i) + \\ &+ \sum_{j\in\mathcal{N}_i}\lambda_{ij,k}s_{ij}(x_i - x_j) + \frac{c_k}{2}h_i(x_i)^2 + \\ &+ \frac{c_k}{2}\sum_{j\in\mathcal{N}_i}(s_{ij}^2 + s_{ji}^2)x_i(x_i - x_j),\ \mathbf{x}_0 = \mathbf{x}^0, \end{aligned} \tag{8}$$

$$\mu_{i,k+1} = \mu_{i,k} + c_k h_i(x_{i,k}),\ \mu_{i,0} = \mu_i^0,\ i = 1,\ldots,N, \tag{9}$$

$$\lambda_{ij,k+1} = \lambda_{ij,k} + c_k s_{ij}(x_{i,k} - x_{j,k}),\ \lambda_{ij,0} = \lambda_{ij}^0,\ j\in\mathcal{N}_i, \tag{10}$$

where $\{c_k\}$ is an non-decreasing sequence of positive numbers known by all agents. Note that at each time instant k we need to solve the unconstrained optimization problem (8). For Algorithm ($\mathtt{A}_3$) to be distributed we need to provide a distributed algorithm that solves (8). Due to the structure of the cost function in (8), such an algorithm results from using a gradient-descent method, namely

$$\begin{aligned} x_{i,\tau+1} &= x_{i,\tau} - \alpha_\tau\Big[\nabla f_i(x_{i,\tau}) + \mu_{i,k}\nabla h_i(x_{i,\tau}) + \\ &+ \sum_{j\in\mathcal{N}_i}(s_{ij}\lambda_{ij,k} - s_{ji}\lambda_{ji,k}) + c_k\nabla h_i(x_{i,\tau})h_i(x_{i,\tau}) \\ &+ c_k\sum_{j\in\mathcal{N}_i}(s_{ij}^2 + s_{ji}^2)(x_{i,\tau} - x_{j,\tau})],\ i = 1,\ldots,N \end{aligned} \tag{11}$$

with $x_{i,0} = (\mathbf{x}_k)_i$, and where $\{\alpha_\tau\}$ is a globally known sequence of step-sizes for the iteration (11). Note that we denote by τ the iteration index for the algorithm used to solve (8).

B. Considerations on the distributed algorithm

Algorithms ($\mathtt{A}_1$), ($\mathtt{A}_2$) are part of the general class of methods, called Lagrangian methods (see for example Section 4.4.1, page 386, [1]). They are based on a first order method, and therefore they achieve a linear rate of

convergence, while Algorithm ($\mathtt{A}_3$) is based on the method of multipliers that theoretically can reach super-linear rate of convergence. In the case of all three algorithms we assume that the stepsize α and the sequences of stepsizes $\{c_k\}_{k\geq 0}$ and $\{\alpha_\tau\}_{\tau\geq 0}$ are globally known by all agents. It can be observed that the algorithms are indeed distributed since for updating their local variables $x_{i,k}$, $\mu_{i,k}$ and $\lambda_{ij,k}$ they use only local information ($\nabla f_i(x_{i,k})$ and $\nabla h_i(x_{i,k})$) and information from their neighbors ($x_{j,k}$, $\lambda_{ji,k}$, and s_{ji} for $j \in \mathcal{N}_i$). In the case of Algorithm ($\mathtt{A}_1$), equation (8) describing the minimizer estimate update is comprised of a standard gradient descent step and two additional terms used to cope with the local equality constraint and the lack of complete information. Intuitively, $\mu_{i,k}$ can be seen as the price paid by agent i for satisfying the local equality constraint, while $\lambda_{i,k}$ is the price paid by the same agent for having its estimate $x_{i,k}$ far away from the estimates of its neighbors. In the case of Algorithms ($\mathtt{A}_2$), ($\mathtt{A}_3$), the iteration for computing the minimizer has two additional terms. These terms have their origin in the use of an augmented Lagrangian and ensure the local convergence to a local minimizer under weaker conditions. In the next sections we will focus on proving converge results corresponding to the three algorithms. Although these three algorithms result from applying standard methods, their convergence results are no longer standard since the local minimizers corresponding to ($\mathtt{P}_1$) are no longer regular due to the distributed setup. We would like to emphasize that unless the communication topology is assumed *undirected*, the updates of the minimizer estimates can no longer be implemented in a distributed manner since the agents would require information from agents not in their neighborhoods.

IV. An equivalent optimization problem with equality constraints

In this section we define an augmented optimization problem, from whose solution we can extract the solution of problem ($\mathtt{P}_1$). As made clear in what follows, the distributed algorithms proposed in this paper, follow from applying standard techniques for solving optimization problem with equality constraints to this particular problem.

Let us define the function $\mathbf{F}:\mathbb{R}^{nN}\to\mathbb{R}$ given by $\mathbf{F}(\mathbf{x})=\sum_{i=1}^N f_i(x_i)$, where $\mathbf{x}'=(x_1',x_2',\ldots,x_N')$, with $x_i\in\mathbb{R}^n$. In addition we introduce the vector-valued functions $\mathbf{h}:\mathbb{R}^{nN}\to\mathbb{R}^N$ and $\mathbf{g}:\mathbb{R}^{nN}\to\mathbb{R}^{nN}$, where $\mathbf{h}(\mathbf{x})=(\mathbf{h}_1(\mathbf{x}),\mathbf{h}_2(\mathbf{x}),\ldots,\mathbf{h}_N(\mathbf{x}))'$, with $\mathbf{h}_i:\mathbb{R}^{nN}\to\mathbb{R}$ given by $\mathbf{h}_i(\mathbf{x})=h_i(x_i)$, and $\mathbf{g}(\mathbf{x})'=(g_1(\mathbf{x})',g_2(\mathbf{x})',\ldots,g_N(\mathbf{x})')$, with $g_i:\mathbb{R}^{nN}\to\mathbb{R}^{|\mathcal{N}_i|n}$ given by $g_i(\mathbf{x})=(g_{ij}(\mathbf{x}))$, where $g_{ij}(\mathbf{x})=s_{ij}(x_i-x_j)$, with s_{ij} positive scalars. The vector valued function $\mathbf{g}(\mathbf{x})$ can be compactly expressed as $g(\mathbf{x})=\mathbf{S}\mathbf{x}$, where $\mathbf{S}=S\otimes I$, with I the n-dimensional identity matrix and S defined in (1). We introduce the optimization problem

$$(\mathtt{P}_2)\qquad \min_{\mathbf{x}\in\mathbb{R}^{nN}}\quad \mathbf{F}(\mathbf{x}), \tag{12}$$

$$\text{subject to:}\quad \mathbf{h}(\mathbf{x})=0, \tag{13}$$

$$\mathbf{g}(\mathbf{x})=\mathbf{S}\mathbf{x}=0. \tag{14}$$

The Lagrangian function of `Problem` ($\mathtt{P}_2$) is a function $\mathcal{L}:\mathbb{R}^{nN}\times\mathbb{R}^N\times\mathbb{R}^{n\bar{N}}\to\mathbb{R}$, defined as

$$\mathcal{L}(\mathbf{x},\boldsymbol{\mu},\boldsymbol{\lambda})\triangleq\mathbf{F}(\mathbf{x})+\boldsymbol{\mu}'\mathbf{h}(\mathbf{x})+\boldsymbol{\lambda}'\mathbf{S}\mathbf{x}. \tag{15}$$

We define also the following augmented Lagrangian for problem (P_2):

$$\mathcal{L}_c(\mathbf{x},\boldsymbol{\mu},\boldsymbol{\lambda}) = \mathbf{F}(\mathbf{x}) + \boldsymbol{\mu}'\mathbf{h}(\mathbf{x}) + \boldsymbol{\lambda}'\mathbf{S}\mathbf{x} + \frac{c}{2}\|\mathbf{h}(\mathbf{x})\|^2 + \frac{c}{2}\mathbf{x}'\mathbf{L}\mathbf{x}, \tag{16}$$

where $\mathbf{L} = \mathbf{S}'\mathbf{S}$ is a Laplacian type of matrix and c is a positive scalar. The gradient and the Hessian of $\mathcal{L}_c(\mathbf{x},\boldsymbol{\mu},\boldsymbol{\lambda})$ are given by

$$\nabla_{\mathbf{x}}\mathcal{L}_c(\mathbf{x},\boldsymbol{\mu},\boldsymbol{\lambda}) = \nabla\mathbf{F}(\mathbf{x}) + \nabla\mathbf{h}(\mathbf{x})\boldsymbol{\mu} + \mathbf{S}'\boldsymbol{\lambda} + c\nabla\mathbf{h}(\mathbf{x})\mathbf{h}(\mathbf{x}) + c\mathbf{L}\mathbf{x}, \tag{17}$$

and

$$\nabla^2_{\mathbf{x}\mathbf{x}}\mathcal{L}_c(\mathbf{x},\boldsymbol{\mu},\boldsymbol{\lambda}) = \nabla^2\mathbf{F}(\mathbf{x}) + \sum_{i=1}^{N}\boldsymbol{\mu}_i\nabla^2\mathbf{h}_i(\mathbf{x}) + c\mathbf{L} +$$

$$+c\sum_{i=1}^{N}\Big(\mathbf{h}_i(\mathbf{x})\nabla^2\mathbf{h}_i(\mathbf{x}) + \nabla\mathbf{h}_i(\mathbf{x})\nabla\mathbf{h}_i(\mathbf{x})'\Big), \tag{18}$$

respectively.

The following proposition states that by solving (P_2) we solve in fact (P_1) as well, and vice-versa.

Proposition 4.1: Let Assumptions 2.1 hold. The vector x^* is a local minimizer of (P_1) if and only if $\mathbf{x}^* = \mathbb{1}\otimes x^*$ is a local minimizer of (P_2).

Proof: Since the Laplacian L corresponds to a connected graph, according to Proposition 2.1-(c), the nullspace of $\mathbf{S}$ is given by $\mathrm{Null}(\mathbf{S}) = \{\mathbb{1}\otimes x \mid x\in\mathbb{R}^n\}$. From the equality constraint (14), we get that any local minimizer $\mathbf{x}^*$ of (P_2) must be of the form $\mathbf{x}^* = \mathbb{1}\otimes x^*$, for some $x^*\in\mathbb{R}^n$. Therefore, the solution of (P_2) must be searched in the set of vectors with structure given by $\mathbf{x} = \mathbb{1}\otimes x$. Applying this constraint, the cost function (12) becomes

$$\mathbf{F}(\mathbf{x}) = \sum_{i=1}^{N} f_i(x) = f(x),$$

and the equality constraint (13) becomes

$$\mathbf{h}(\mathbf{x}) = h(x) = 0,$$

which shows that we have recovered the optimization problem (P_1). ■

Remark 4.1: We note from the above proposition the importance of having a connected communication topology. Indeed, if $\mathcal{G}$ is not connected, then the nullspace of $\mathbf{S}$ is much richer than $\{\mathbb{1}\otimes x \mid x\in\mathbb{R}^n\}$, and therefore the solution of (P_2) may not necessarily be of the form $\mathbf{x}^* = \mathbb{1}\otimes x^*$. However, the fact that we search a solution of (P_2) of this particular structure is *fundamental* for showing the equivalence between the two optimization problems.

V. Convergence analysis of the distributed algorithms based on Lagrangian methods

In this section we study the convergence properties of the two distributed algorithms for solving problem (P_1) that are based on Lagrangian methods. In particular, they are obtained by applying a first order method for solving the first-order necessary optimality conditions, where in the case of the second algorithm, the first order necessary conditions are derived in terms of the augmented Lagrangian. As we will see next, this will require weaker conditions for convergence.

A. Supporting results for the convergence analysis

This section introduces a set of results used for the convergence analysis of the algorithms. The proofs of these results together with their auxiliary results can be found in the Appendix section.

The following result characterize the tangent cone at a local minimizer of (P_2) in terms of the tangent cone at a local minimizer of (P_1).

Proposition 5.1: Let Assumptions 2.1-(a) and 2.2 hold, let $\mathbf{x}^* = \mathbb{1} \otimes x^*$ be a local minimizer of (P_2) and let $\mathbf{\Omega}$ denote the constraint set, that is, $\mathbf{\Omega} = \{\mathbf{x} \mid \mathbf{h}(\mathbf{x}) = 0, \mathbf{S}\mathbf{x} = 0\}$. Then the tangent cone to $\mathbf{\Omega}$ at $\mathbf{x}^*$ is given by $\mathrm{TC}(\mathbf{x}^*, \mathbf{\Omega}) = \mathrm{Null}\left([\nabla \mathbf{h}(\mathbf{x}^*), \mathbf{S}']'\right) = \{\mathbb{1} \otimes h \mid h \in \mathrm{Null}\left(\nabla h(x^*)'\right) = \mathrm{TC}(x^*, \Omega)\}$.

Under the assumption that the matrix $\nabla h(x^*)$ is full rank, the first order necessary conditions of (P_1) are given by $\nabla f(x^*) + \nabla h(x^*)\psi^* = 0$, $h(x^*) = 0$, where the vector ψ^* is unique (see for example Proposition 3.3.1, page 255, [1]). An interesting question is whether or not there is a connection between ψ^* and $\boldsymbol{\mu}^*$ shown in the first order necessary conditions of (P_2). As shown in the following, the two vectors are in fact equal.

Proposition 5.2: Let Assumptions 2.1 and 2.2 hold, let $\mathbf{x}^* = \mathbb{1} \otimes x^*$ be a local minimizer of (P_2) and let ψ^* and $\boldsymbol{\mu}^*$ be the unique Lagrange multiplier vectors corresponding to the first order necessary conditions of (P_1) and (P_2), respectively. Then $\psi^* = \boldsymbol{\mu}^*$.

Using augmented Lagrangian for improving convergence was first study in [6]. The basic idea is that under some assumptions on the standard Lagrangian function, the augmented Lagrangian can be made positive definite positive definite and therefore invertible when choosing a scalar c large enough. The next proposition states that this property holds in our setup as well.

Proposition 5.3: Let $(\mathbf{x}^*, \boldsymbol{\mu}^*, \boldsymbol{\lambda}^*)$ be a local minimizer-Lagrange multipliers pair of (P_2) and assume that $\mathbf{z}'\nabla^2_{\mathbf{xx}}\mathcal{L}_0(\mathbf{x}^*, \boldsymbol{\mu}^*, \boldsymbol{\lambda}^*)\mathbf{z} > 0$ for all $\mathbf{z} \in \mathrm{TC}(\mathbf{x}^*, \mathbf{\Omega})$. Then there exists a positive scalar $\bar{c}$, such that $\nabla^2_{\mathbf{xx}}\mathcal{L}_c(\mathbf{x}^*, \boldsymbol{\mu}^*, \boldsymbol{\lambda}^*) \succ 0$ for all $c \geq \bar{c}$.

Proof: We recall that the the Hessian $\nabla^2_{\mathbf{xx}}\mathcal{L}_c(\mathbf{x}^*, \boldsymbol{\mu}^*, \boldsymbol{\lambda}^*)$ is given by

$$\nabla^2_{\mathbf{xx}}\mathcal{L}_c(\mathbf{x}^*, \boldsymbol{\mu}^*, \boldsymbol{\lambda}^*) = \nabla^2_{\mathbf{xx}}\mathcal{L}_0(\mathbf{x}^*, \boldsymbol{\mu}^*, \boldsymbol{\lambda}^*) + c\nabla \mathbf{h}(\mathbf{x}^*)\nabla \mathbf{h}(\mathbf{x}^*)' + c\mathbf{S}'\mathbf{S}.$$

We have that $\mathbf{z}'\left[\nabla \mathbf{h}(\mathbf{x}^*)\nabla \mathbf{h}(\mathbf{x}^*)' + \mathbf{S}'\mathbf{S}\right]\mathbf{z} = 0$ if and only if $z \in \mathrm{Null}\left([\nabla \mathbf{h}(\mathbf{x}^*), \mathbf{S}']'\right)$. By Proposition 5.1 we have $\mathrm{Null}\left([\nabla \mathbf{h}(\mathbf{x}^*), \mathbf{S}']'\right) = \mathrm{TC}\left(\mathbf{x}^*, \mathbf{\Omega}\right)$. Finally, using Lemma 1.25, page 68 of [2], there exists $\bar{c} > 0$ so that $\nabla^2_{\mathbf{xx}}\mathcal{L}_c(\mathbf{x}^*, \boldsymbol{\mu}^*, \boldsymbol{\lambda}^*) \succ 0$, for all $c \geq \bar{c}$. ■

B. Lagrangian methods - Algorithm (A_1)

To find a solution of problem (P_2) the first thing we can think about is solving the set of necessary conditions:

$$\nabla \mathbf{F}(\mathbf{x}) + \mathbf{S}'\boldsymbol{\lambda} + \nabla \mathbf{h}(\mathbf{x})\boldsymbol{\mu} = 0, \tag{19}$$

$$\mathbf{h}(\mathbf{x}) = 0, \tag{20}$$

$$\mathbf{S}\mathbf{x} = 0. \tag{21}$$

Solving (19)-(21) does not guarantee finding a local minimizer, but at least the local minimizers are among the solutions of the above nonlinear system of equations. An approach for solving (19)-(21) consists of using a first order method (see for instance Section 4.4.1, page 386, [1]), which is given by

$$\mathbf{x}_{k+1} = \mathbf{x}_k - \alpha\left[\nabla\mathbf{F}(\mathbf{x}_k) + \nabla\mathbf{h}(\mathbf{x}_k)\boldsymbol{\mu}_k + \mathbf{S}'\boldsymbol{\lambda}_k\right], \tag{22}$$

$$\boldsymbol{\mu}_{k+1} = \boldsymbol{\mu}_k + \alpha\mathbf{h}(\mathbf{x}_k), \tag{23}$$

$$\boldsymbol{\lambda}_{k+1} = \boldsymbol{\lambda}_k + \alpha\mathbf{S}\mathbf{x}_k, \tag{24}$$

where $\alpha > 0$ is chosen to ensure the stability of the algorithm. By reformulating the above iteration in terms of the n-dimensional components of the vectors $\mathbf{x}_k$ and $\boldsymbol{\lambda}_k$, and in terms of the scalar components of the vector $\boldsymbol{\mu}_k$, we obtain Algorithm ($\mathtt{A}_1$).

Since the matrix $\mathbf{S}$ is not full rank, we cannot apply directly existing results for regular (local) minimizers, such as Proposition 4.4.2, page 388, [1]. Still, for a local minimizer and Lagrange multipliers pair $(\mathbf{x}^*, \boldsymbol{\mu}^*, \boldsymbol{\lambda}^*)$, with $\boldsymbol{\lambda}^* \in \text{Range}(\mathbf{S})$, we show that if the initial values $(\mathbf{x}_0, \boldsymbol{\mu}_0, (\mathbf{I}-\mathbf{J})\boldsymbol{\lambda}_0)$ are close enough to $(\mathbf{x}^*, \boldsymbol{\mu}^*, \boldsymbol{\lambda}^*)$, for a small enough step-size and under some conditions on (the Hessians of) the functions $f_i(x)$ and $h_i(x)$, $i = 1, \ldots, N$, the vectors $\mathbf{x}_k$ and $\boldsymbol{\mu}_k$ do indeed converge to $\mathbf{x}^*$ and $\boldsymbol{\mu}^*$, respectively. However, although under the same conditions $\boldsymbol{\lambda}_k$ does converge, it cannot be guaranteed that it converges to the unique $\boldsymbol{\lambda}^* \in \text{Range}(\mathbf{S})$ but rather to a point in the set $\{\boldsymbol{\lambda}^* + \text{Null}(\mathbf{S}')\}$.

The following theorem addresses the local convergence properties of Algorithm (A_1), which, under some assumptions on the functions $f_i(x)$ and $h_i(x)$, states that provided the initial values used in the Algorithm (A_1) are close enough to a solution of the first order necessary conditions of (P_2), and a small enough step-size α is used, the sequence $\{\mathbf{x}_k, \boldsymbol{\mu}_k, \boldsymbol{\lambda}_k\}$ converges to this solution.

Theorem 5.1: Let Assumptions 2.1 and 2.2 hold and let $(\mathbf{x}^*, \boldsymbol{\mu}^*, \boldsymbol{\lambda}^*)$ with $\boldsymbol{\lambda}^* \in \text{Range}(\mathbf{S})$, be a local minimizer-Lagrange multipliers pair of (P_2). Assume also that $\nabla^2_{\mathbf{xx}}\mathcal{L}(\mathbf{x}^*, \boldsymbol{\mu}^*, \boldsymbol{\lambda}^*)$ is positive definite. Then there exists $\bar{\alpha}$, such that for all $\alpha \in (0, \bar{\alpha}]$, the set $(\mathbf{x}^*, \boldsymbol{\mu}^*, \boldsymbol{\lambda}^* + \text{Null}(\mathbf{S}'))$ is an attractor of iteration (22)-(24) and if the sequence $\{\mathbf{x}_k, \boldsymbol{\mu}_k, \boldsymbol{\lambda}_k\}$ converges to the set $(\mathbf{x}^*, \boldsymbol{\mu}^*, \boldsymbol{\lambda}^* + \text{Null}(\mathbf{S}'))$, the rate of convergence of $\|\mathbf{x}_k - \mathbf{x}^*\|$, $\|\boldsymbol{\mu}_k - \boldsymbol{\mu}^*\|$ and $\left\|\boldsymbol{\lambda}_k - [\boldsymbol{\lambda}^* + \text{Null}(\mathbf{S}')]\right\|$ is linear.

Let us know reformulate the above theorem so that the local convergence result can be applied to problem (P_1).

Corollary 5.1: Let Assumptions 2.1 and 2.2 hold and let (x^*, ψ^*) be a local minimizer-Lagrange multiplier pair of (P_1). Assume also that $\nabla^2 f_i(x^*) + \psi_i^* \nabla^2 h_i(x^*)$ are positive definite for all $i = 1, \ldots, N$. Then there exists $\bar{\alpha}$, such that for all $\alpha \in (0, \bar{\alpha}]$, (x^*, ψ^*) is a point of attraction for iteration (8) and (9), for all $i = 1, \ldots, N$, and if the sequence $\{x_{i,k}, \mu_{i,k}\}$ converges to (x^*, ψ^*), then the rate of convergence of $\|x_{i,k} - x^*\|$ and $\|\mu_{i,k} - \psi^*\|$ is linear.

Proof: By Proposition 4.1 we have that $\mathbf{x}^* = \mathbb{1} \otimes x^*$ is a local minimizer of (P_2) with corresponding Lagrange multipliers $(\boldsymbol{\mu}^*, \boldsymbol{\lambda}^* + \text{Null}(\mathbf{S}'))$, with $\boldsymbol{\lambda}^* \in \text{Range}(\mathbf{S})$. In addition, by Proposition 5.2 we have that $\boldsymbol{\mu}^* = \psi^*$. Using the

definition of the Lagrangian function introduced in (15), we discover that

$$\nabla^2_{\mathbf{xx}}\mathcal{L}(\mathbf{x}^*,\boldsymbol{\mu}^*,\boldsymbol{\lambda}^*) = \mathrm{diag}\Big(\nabla^2 f_i(x^*) + \psi_i^*\nabla^2 h_i(x^*), i = 1,\ldots,N\Big).$$

But since we assumed that $\nabla^2 f_i(x^*) + \psi_i^*\nabla^2 h_i(x^*) \succ 0$ for all i, it follows that $\nabla^2_{\mathbf{xx}}\mathcal{L}(\mathbf{x}^*,\boldsymbol{\mu}^*,\boldsymbol{\lambda}^*) \succ 0$ as well. Using Theorem 5.1, the result follows. ■

Remark 5.1: In the previous corollary the matrices $\nabla^2 f_i(x^*) + \psi_i^*\nabla^2 h_i(x^*)$ were assumed to be positive definite for all $i = 1,\ldots,N$. If we apply directly on (P_1) results from the optimization literature (for instance Proposition 4.4.2, page 388, [1]) concerning convergence of first-order methods used to compute local minimizers and their corresponding Lagrange multipliers, we only require $\sum_{i=1}^N \nabla^2 f_i(x^*) + \psi_i^*\nabla^2 h_i(x^*)$ to be positive definite, and not each element of the sum. Obviously the assumption in Corollary 5.1 does imply the latter, but is not necessary.□

C. Lagrangian methods - Algorithm (A_2)

In Theorem 5.1 we made the assumption that $\nabla^2_{\mathbf{xx}}\mathcal{L}(\mathbf{x}^*,\boldsymbol{\mu}^*,\boldsymbol{\lambda}^*)$ is positive definite. We can weaken this assumption by using an augmented version of the Lagrangian, obtaining an equivalent set of first order necessary conditions and applying again a first order numerical method to solve for the optimal solution.

The first order necessary conditions for (P_2) with respect to the augmented Lagrangian $\mathcal{L}_c(\mathbf{x},\boldsymbol{\mu},\boldsymbol{\lambda})$ are given by

$$\nabla\mathbf{F}(\mathbf{x}) + \nabla\mathbf{h}(\mathbf{x})\boldsymbol{\mu} + \mathbf{S}'\boldsymbol{\lambda} + c\nabla\mathbf{h}(\mathbf{x})\mathbf{h}(\mathbf{x}) + c\mathbf{L}\mathbf{x} = 0, \tag{25}$$

$$\nabla\mathbf{h}(\mathbf{x}) = 0, \tag{26}$$

$$\mathbf{S}\mathbf{x} = 0. \tag{27}$$

A first order numerical method that can be used to solve the necessary conditions (25)-(27) takes the form

$$\begin{aligned}\mathbf{x}_{k+1} &= \mathbf{x}_k - \alpha\big[\nabla\mathbf{F}(\mathbf{x}_k) + \nabla\mathbf{h}(\mathbf{x}_k)\boldsymbol{\mu}_k + \\ &+ c\nabla\mathbf{h}(\mathbf{x}_k)\mathbf{h}(\mathbf{x}_k) + \mathbf{S}'\boldsymbol{\lambda}_k + c\mathbf{L}\mathbf{x}_k\big], \qquad (28)\\ \boldsymbol{\mu}_{k+1} &= \boldsymbol{\mu}_k + \alpha\mathbf{h}(\mathbf{x}_k), \qquad (29)\\ \boldsymbol{\lambda}_{k+1} &= \boldsymbol{\lambda}_k + \alpha\mathbf{S}\mathbf{x}_k. \qquad (30)\end{aligned}$$

Re-writing the above iterations in terms of the components corresponding to each agent, we recover Algorithm $(\mathtt{A}_2)$.

The following result addresses the local convergence properties of the iteration (28)-(30).

Theorem 5.2: Let Assumptions 2.1 and 2.2 hold and let $(\mathbf{x}^*,\boldsymbol{\mu}^*,\boldsymbol{\lambda}^*)$ with $\boldsymbol{\lambda}^* \in \mathrm{Range}(\mathbf{S})$, be a local minimizer-Lagrange multipliers pair of (P_2). Assume also that $\mathbf{x}'\nabla^2_{\mathbf{xx}}\mathcal{L}(\mathbf{x}^*,\boldsymbol{\mu}^*,\boldsymbol{\lambda}^*)\mathbf{x} > 0$ for all $\mathbf{x} \in \mathrm{TC}(\mathbf{x}^*,\boldsymbol{\Omega})$. Then there exists $\bar{c} > 0$ so that for all $c > \bar{c}$ we can find $\bar{\alpha}(c)$ such that for all $\alpha \in (0,\bar{\alpha}(c)]$, the set $(\mathbf{x}^*,\boldsymbol{\mu}^*,\boldsymbol{\lambda}^* + \mathrm{Null}(\mathbf{S}'))$ is an attractor of iteration (28)-(30). In addition, if the sequence $\{\mathbf{x}_k,\boldsymbol{\mu}_k,\boldsymbol{\lambda}_k\}$ converges to the set $(\mathbf{x}^*,\boldsymbol{\mu}^*,\boldsymbol{\lambda}^* + \mathrm{Null}(\mathbf{S}'))$, the rate of convergence of $\|\mathbf{x}_k - \mathbf{x}^*\|$, $\|\boldsymbol{\mu}_k - \boldsymbol{\mu}^*\|$ and $\big\|\boldsymbol{\lambda}_k - [\boldsymbol{\lambda}^* + \mathrm{Null}(\mathbf{S}')]\big\|$ is linear.

The following corollary gives conditions that ensure local convergence to a local minimizer of (P_1) for each agent following Algorithm (A_2).

Corollary 5.2: Let Assumptions 2.1 and 2.2 hold and let (x^*, ψ^*) be a local minimizer-Lagrange multiplier pair of (P_1). Assume also that $x' \sum_{i=1}^{N} \left[\nabla^2 f_i(x^*) + \psi_i^* \nabla^2 h_i(x^*)\right] x > 0$ for all $x \in \mathrm{TC}(x^*, \Omega)$. Then there exists $\bar{c} > 0$ so that for all $c \geq \bar{c}$ we can find $\bar{\alpha}(c)$ such that for all $\alpha \in (0, \bar{\alpha}(c)]$, (x^*, ψ^*) is a point of attraction for iteration (5)-(7), for all $i = 1, \ldots, N$. In addition, if the sequence $\{x_{i,k}, \mu_{i,k}\}$ converges to (x^*, ψ^*), then the rate of convergence of $\|x_{i,k} - x^*\|$ and $\|\mu_{i,k} - \psi^*\|$ is linear.

Proof: By Proposition 4.1 we have that $\mathbf{x}^* = \mathbb{1} \otimes x^*$ is a local minimizer of (P_2) with corresponding Lagrange multipliers $(\boldsymbol{\mu}^*, \boldsymbol{\lambda}^* + \mathrm{Null}(\mathbf{S}'))$, with $\boldsymbol{\lambda}^* \in \mathrm{Range}(\mathbf{S})$. In addition, by Proposition 5.2 we have that $\boldsymbol{\mu}^* = \psi^*$. Using the definition of the Lagrangian function introduced in (15), we have

$$\nabla^2_{\mathbf{xx}} \boldsymbol{\mathcal{L}}(\mathbf{x}^*, \boldsymbol{\mu}^*, \boldsymbol{\lambda}^*) = \mathrm{diag}\left(\nabla^2 f_i(x^*) + \psi_i^* \nabla^2 h_i(x^*), i = 1, \ldots, N\right).$$

In Proposition 5.1 we showed that

$$\mathrm{TC}(\mathbf{x}^*, \boldsymbol{\Omega}) = \{\mathbb{1} \otimes z \mid z \in \mathrm{TC}(x^*, \Omega)\},$$

and therefore the assumption $x' \sum_{i=1}^{N} \left[\nabla^2 f_i(x^*) + \psi_i^* \nabla^2 h_i(x^*)\right] x > 0$ for all $x \in \mathrm{TC}(x^*, \Omega)$ is equivalent to $\mathbf{x}' \nabla^2_{\mathbf{xx}} \boldsymbol{\mathcal{L}}(\mathbf{x}^*, \boldsymbol{\mu}^*, \boldsymbol{\lambda}^*) \mathbf{x} > 0$ for all $\mathbf{x} \in \mathrm{TC}(\mathbf{x}^*, \boldsymbol{\Omega})$. Using Theorem 5.2, the result follows. ■

Remark 5.2: In the previous corollary we made the assumption that $x' \sum_{i=1}^{N} \left[\nabla^2 f_i(x^*) + \psi_i^* \nabla^2 h_i(x^*)\right] x > 0$ for all $x \in \mathrm{TC}(x^*, \Omega)$. It turns out that the same assumption would appear is we apply directly on (P_1) standard results from the optimization literature concerning convergence of first-order methods applied to solve the first order necessary conditions formulated in terms of the augmented Lagrangian. In other words the assumption is identical when solving (P_1) in a centralized manner.□

VI. Convergence analysis of the distributed algorithms based on the method of multipliers

In this section we prove a set of results that will be used to give conditions under which `Algorithm` $(\mathtt{A}_3)$ converges to a local minimizer of $(\mathtt{P}_1)$. The results are modifications of standard results concerning the method of multipliers (see for example Section 2.2, [2]). In the standard case, a regularity assumption on the minimizers is used to prove convergence of the method of multipliers. In our setup, this is not the case anymore and therefore, the standard results need to be modified accordingly.

Considering the notations introduced so far, Algorithm $(\mathtt{A}_3)$ can be compactly written as

$$\mathbf{x}_k = \arg\min_{\mathbf{x}} \boldsymbol{\mathcal{L}}_{c_k}(\mathbf{x}, \boldsymbol{\mu}_k, \boldsymbol{\lambda}_k), \ \mathbf{x}_0 = \mathbf{x}^0, \tag{31}$$

$$\boldsymbol{\mu}_{k+1} = \boldsymbol{\mu}_k + c_k \mathbf{h}(\mathbf{x}_k), \ \boldsymbol{\mu}_0 = \boldsymbol{\mu}^0, \tag{32}$$

$$\boldsymbol{\lambda}_{k+1} = \boldsymbol{\lambda}_k + c_k \mathbf{S}\mathbf{x}_k, \ \boldsymbol{\lambda}_0 = \boldsymbol{\lambda}^0. \tag{33}$$

where the unconstrained optimization problem (31) is solved using the iteration

$$\mathbf{x}_{k,\tau+1} = \mathbf{x}_{k,\tau} - \alpha_\tau \{\nabla \mathbf{F}(\mathbf{x}_{k,\tau}) + \nabla \mathbf{h}(\mathbf{x}_{k,\tau}) \boldsymbol{\mu}_k + \mathbf{S}' \boldsymbol{\lambda}_k + c_k \nabla \mathbf{h}(\mathbf{x}_{k,\tau}) \mathbf{h}(\mathbf{x}_{k,\tau}) + c_k \mathbf{L}\mathbf{x}_{k,\tau}\}, \mathbf{x}_{k,0} = \mathbf{x}_{k-1}.$$

Therefore, iterations (8)-(10) reflect the method of multipliers applied to Problem (P_2), while the iteration (11) is a gradient descent method for finding a local minimizer of the unconstrained optimization problem $\min_{\mathbf{x}} \mathcal{L}_c(\mathbf{x}, \boldsymbol{\mu}_k, \boldsymbol{\lambda}_k)$.

A. Supporting results

Before stating the conditions under which convergence can be achieved, let us first start by enumerating a set of supporting results used for the convergence analysis. Their complete proofs can be found in the Appendix section. To simplify notations, let us group the Lagrange multipliers in one vector, that is, $\boldsymbol{\eta}' = [\boldsymbol{\mu}', \boldsymbol{\lambda}']$. When we mention a Lagrange multiplier vector $\boldsymbol{\eta}^*$ corresponding to a local minimizer $\mathbf{x}^*$, we will understand that its subcomponent $\boldsymbol{\lambda}^*$ is the unique vector in Range($\mathbf{S}$). In addition, let us group the equality constraints functions of (P_2) into one vector-valued function, that is, $\tilde{\mathbf{h}}(\mathbf{x})' = (\mathbf{h}(\mathbf{x})', \mathbf{x}'\mathbf{S}')$.

The convergence of the distributed algorithm derived from method of multipliers is based on the following result, which is an adaptation of the Proposition 2.4, page 108 of [2] so that it fits our setup.

Proposition 6.1: Let Assumptions 2.1 and 2.2 hold, let $(\mathbf{x}^*, \boldsymbol{\eta}^*)$, be a local minimizer-Lagrange multipliers pair of (P_2), and assume that $\mathbf{z}'\nabla^2_{\mathbf{xx}}\mathcal{L}_0(\mathbf{x}^*, \boldsymbol{\eta}^*)\mathbf{z} > 0$ for all $\mathbf{z} \in \mathrm{TC}(\mathbf{x}^*, \boldsymbol{\Omega})$. In addition let $\bar{c}$ be a positive scalar such that $\nabla^2_{\mathbf{xx}}\mathcal{L}_{\bar{c}}(\mathbf{x}^*, \boldsymbol{\eta}^*) \succ 0$. There exists positive scalars $c_{\max}$, δ, ε and M such that:

(a) For all $(\boldsymbol{\eta}, c)$ in the set $D \subset \mathbb{R}^{N+n\bar{N}+1}$ defined by

$$D = \{(\boldsymbol{\eta}, c) \mid \|\mathbf{T}\boldsymbol{\eta} - \boldsymbol{\eta}^*\| < c\delta,\ \bar{c} \le c \le c_{\max}\ \}, \tag{34}$$

where

$$\mathbf{T} = \begin{bmatrix} \mathbf{I} & \mathbf{0} \\ \mathbf{0} & \mathbf{I} - \mathbf{J} \end{bmatrix},$$

with $\mathbf{J}$ the orthogonal projection operator on $\mathrm{Null}(\mathbf{S}')$, the problem

$$\begin{aligned} \min \quad & \mathcal{L}_c(\mathbf{x}, \boldsymbol{\eta}) \\ \text{subject to} \quad & \mathbf{x} \in \mathcal{S}(\mathbf{x}^*; \varepsilon) \end{aligned} \tag{35}$$

has a unique solution denoted by $\mathbf{x}(\boldsymbol{\eta}, c)$. The function $\mathbf{x}(\cdot, \cdot)$ is continuously differentiable in the interior of D, and for all $(\boldsymbol{\eta}, c) \in D$, we have

$$\|\mathbf{x}(\boldsymbol{\eta}, c) - \mathbf{x}^*\| \le M\|\boldsymbol{\eta} - \boldsymbol{\eta}^*\|/c. \tag{36}$$

(b) For all $(\boldsymbol{\eta}, c) \in D$, we have

$$\|\tilde{\boldsymbol{\eta}}(\boldsymbol{\eta}, c) - \boldsymbol{\eta}^*\| \le M\|\boldsymbol{\eta} - \boldsymbol{\eta}^*\|/c, \tag{37}$$

where

$$\tilde{\boldsymbol{\eta}}(\boldsymbol{\eta}, c) = \mathbf{T}\boldsymbol{\eta} + c\tilde{\mathbf{h}}[\mathbf{x}(\boldsymbol{\eta}, c)]. \tag{38}$$

(c) For all $(\boldsymbol{\eta}, c) \in D$, the matrix $\nabla^2_{\mathbf{xx}}\mathcal{L}_c[\mathbf{x}(\boldsymbol{\eta}, c), \boldsymbol{\eta}]$ is positive definite.

The proof of this proposition can be found in the Appendix section. At this point we would like only the emphasize what is different compared with the original formulation. First recall that the assumption on the Hessian $\nabla^2_{\mathbf{xx}}\mathcal{L}_0(\mathbf{x}^*, \boldsymbol{\eta}^*)$ simply means that $\mathbf{x}^*$ is a strictly local minimizer. Generally speaking, the proof of this result follows the same

lines as the proof of Proposition 2.4, [2]. However, since the local minimizer $\mathbf{x}^*$ is not regular, we made some changes in the statement of the proposition compared to the original result, and consequently we need to adapt the proof accordingly. Compared to Proposition 2.4, page 108 of [2], our results has three main differences. The first difference consists of imposing an upper bound on c, namely $c_{\max}$. The reason we introduced $c_{\max}$ is to ensure that a certain Jacobian matrix that depends on c (explicitly defined in the proof) is invertible. We will see however that $c_{\max}$ can be made arbitrarily large. The price paid for this change is the prevention of reaching a theoretical superlinear rate of convergence for Algorithm ($\mathtt{A}_3$).

The second difference is the introduction of operator $\mathbf{T}$. This operator acting on $\boldsymbol{\eta}$ ensures that $\tilde{\boldsymbol{\lambda}}(\boldsymbol{\eta},c) \in \text{Range}(\mathbf{S})$ for all $(\boldsymbol{\eta},c) \in D$, where $\tilde{\boldsymbol{\eta}} = (\tilde{\boldsymbol{\mu}},\tilde{\boldsymbol{\lambda}})$. In defining the set D, the matrix $\mathbf{T}$ induces a neighborhood around $\boldsymbol{\eta}^*$ were only points $\boldsymbol{\eta} = (\boldsymbol{\mu},\boldsymbol{\lambda})$, with $\boldsymbol{\lambda} \in \text{Range}(\mathbf{S})$ are considered. In this neighborhood, $\boldsymbol{\eta}^* = (\boldsymbol{\mu}^*,\boldsymbol{\lambda}^*)$ is a unique Lagrange multiplier vector corresponding to the local minimizer $\mathbf{x}^*$.

The third difference is the definition of $\tilde{\boldsymbol{\eta}}$ in (38). Compared to the original statement[1] of Proposition 2.4, [2], we introduce the operator $\mathbf{T}$ that multiplies $\boldsymbol{\eta}$, to deal with the fact as $\mathbf{x}^*$ is not regular local minimizer. As a consequence, we will have that $\tilde{\boldsymbol{\lambda}}(\boldsymbol{\eta},c) \in \text{Range}(\mathbf{S})$ for all $(\boldsymbol{\eta},c) \in D$, where $\tilde{\boldsymbol{\eta}} = (\tilde{\boldsymbol{\mu}},\tilde{\boldsymbol{\lambda}})$.□

Given a minimizer-Lagrange multiplier pair $(\mathbf{x}^*,\boldsymbol{\eta}^*)$ of (P_2), let us define the following matrix

$$\mathbf{M} = \left\{\nabla\tilde{\mathbf{h}}(\mathbf{x}^*)'\left[\nabla^2_{\mathbf{xx}}\mathcal{L}_c(\mathbf{x}^*,\boldsymbol{\eta}^*)\right]^{-1}\nabla\tilde{\mathbf{h}}(\mathbf{x}^*)\right\}^{-1} - c\mathbf{I} \tag{39}$$

for any c for which $\nabla^2_{\mathbf{xx}}\mathcal{L}_c(\mathbf{x}^*,\boldsymbol{\eta}^*)$ is invertible. In addition, it can be shown that if $\left[\nabla^2_{\mathbf{xx}}\mathcal{L}_0(\mathbf{x}^*,\boldsymbol{\eta}^*)\right]^{-1}$ exist, then

$$\mathbf{M} = \left\{\nabla\tilde{\mathbf{h}}(\mathbf{x}^*)'\left[\nabla^2_{\mathbf{xx}}\mathcal{L}_0(\mathbf{x}^*,\boldsymbol{\eta}^*)\right]^{-1}\nabla\tilde{\mathbf{h}}(\mathbf{x}^*)\right\}^{-1}, \tag{40}$$

respectively.

B. Method of multipliers - Algorithm ($\mathtt{A}_3$)

The following Theorem is the equivalent of Proposition 2.7 of [2] and if formulated keeping in mind the distributed setup of the problem. Our result does not include the case of superlinear rate of convergence since we upper bound the scalar c and its statement is adapted so that it fits to the characteristics of Problem (P_2).

Theorem 6.1: Let Assumptions 2.1 and 2.2 hold, let $(\mathbf{x}^*,\boldsymbol{\eta}^*)$, with $\boldsymbol{\eta}^* = (\boldsymbol{\mu}^*,\boldsymbol{\lambda}^*)$ be a local minimizer-Lagrange multipliers pair of (P_2) that satisfies $\mathbf{z}'\nabla^2_{\mathbf{xx}}\mathcal{L}_0(\mathbf{x}^*,\boldsymbol{\eta}^*)\mathbf{z} > 0$ for all $\mathbf{z} \in \text{TC}(\mathbf{x}^*,\boldsymbol{\Omega})$. In addition, let $\bar{c}$, δ and $c_{\max}$ be as in Proposition 6.1 with $\bar{c} > \max\{-2\mathbf{e}_1,\ldots,-2\mathbf{e}_{nN}\}$, where $\mathbf{e}_1,\ldots,\mathbf{e}_{nN}$ are the eigenvalues of $\mathbf{M}$ defined in (39). Then there exists δ_1 with $0 < \delta_1 \le \delta$ such that if $\{c_k\}$ and $\boldsymbol{\eta}_0$ satisfy

$$\|\mathbf{T}\boldsymbol{\eta}_0 - \boldsymbol{\eta}^*\| < \delta_1 c_0,\ \bar{c} \le c_k \le c_{k+1} \le c_{\max}\ \forall k, \tag{41}$$

then for the sequence $\{\boldsymbol{\eta}_k\}$ generated by

$$\boldsymbol{\eta}_{k+1} = \mathbf{T}\boldsymbol{\eta}_k + c_k\tilde{\mathbf{h}}\left[\mathbf{x}(\boldsymbol{\eta}_k,c_k)\right], \tag{42}$$

[1] $\boldsymbol{\eta}$ corresponds to λ in Proposition 2.4, [2].

we have that $\|\boldsymbol{\eta}_k - \boldsymbol{\eta}^*\|$ and $\|\mathbf{x}(\boldsymbol{\eta}_k, c_k) - \mathbf{x}^*\|$ converge to zero. Furthermore if $\boldsymbol{\eta}_k \neq \boldsymbol{\eta}^*$ for all k, there holds

$$\limsup_{k\to\infty} \frac{\|\boldsymbol{\eta}_{k+1} - \boldsymbol{\eta}^*\|}{\|\boldsymbol{\eta}_k - \boldsymbol{\eta}^*\|} \leq \max_{i=1\ldots nN} \left| \frac{\mathbf{e}_i}{\mathbf{e}_i + c_{\max}} \right|. \square \tag{43}$$

In the following we show that under some conditions, `Algorithm` ($\mathtt{A}_3$) ensures the convergence of each $x_{i,k}$ to x^* and the convergence of $\mu_{i,k}$ to μ_i^*, where $\boldsymbol{\mu}^* = (\mu_i^*)$. The algorithm does not guarantee convergence of $\lambda_{ij,k}$ to λ_{ij}^*, where $\boldsymbol{\lambda}^* = (\lambda_{ij}^*)$, for $j \in \mathcal{N}_i$ and $i = 1, \ldots, N$. In fact, if convergence is achieved, $\boldsymbol{\lambda}_k$ converges to the set $\boldsymbol{\lambda}^* + \mathrm{Null}(\mathbf{S}')$.

Corollary 6.1: Let Assumptions 2.1 and 2.2 hold, let (x^*, ψ^*) be a local minimizer-Lagrange multipliers pair of (P_1). In addition, let $\mathbf{x}^* = \mathbb{1} \otimes x^*$ be a local minimizer of (P_2) (as stated by Proposition 4.1), with unique Lagrange multipliers $(\boldsymbol{\mu}^*, \boldsymbol{\lambda}^*)$ and $\boldsymbol{\lambda}^* \in \mathrm{Range}(\mathbf{S})$. Assume also that $z' \sum_{i=1}^N \left(\nabla^2 f_i(x^*) + \psi_i^* \nabla^2 h_i(x^*)\right) z > 0$ for all $z \in \mathrm{TC}(x^*, \Omega)$. There exist scalars $c_{\max}$, δ, δ_1 with $0 < \delta_1 \leq \delta$ and $c_{\max}$ satisfying $\bar{c} > \max\{-2\mathbf{e}_1, \ldots, -2\mathbf{e}_{nN}\}$ where $\mathbf{e}_1, \ldots, \mathbf{e}_{nN}$ are the eigenvalues of $\mathbf{M}$ defined in (39) such that if the sequence $\{c_k\}$, $\boldsymbol{\mu}_0$ and $\boldsymbol{\lambda}_0$ satisfy

$$\left(\|\boldsymbol{\mu}_0 - \psi^*\|^2 + \|(\mathbf{I} - \mathbf{J})\boldsymbol{\lambda}_0 - \boldsymbol{\lambda}^*\|^2\right)^{1/2} < \delta_1 c_0, \tag{44}$$

$$\bar{c} \leq c_k \leq c_{k+1} \leq c_{\max}, \tag{45}$$

then for the sequences $\{x_{i,k}\}$ and $\{\mu_{i,k}\}$ generated by the iteration (31)-(33), we have that $\|x_{i,k} - x^*\|$ and $\|\mu_{i,k} - \psi_i^*\|$ converge to zero. Furthermore if $\mu_{i,k} \neq \psi_i^*$ for all k, the rate of convergence of $\{\mu_{i,k}\}$ is linear.

Proof: By Proposition 5.2 we have that $\boldsymbol{\mu}^* = \psi^*$. Using the definition of the Lagrangian function introduced in (15), we have $\nabla^2_{\mathbf{xx}} \mathcal{L}(\mathbf{x}^*, \boldsymbol{\mu}^*, \boldsymbol{\lambda}^*) = \mathrm{diag}\left(\nabla^2 f_i(x^*) + \mu_i^* \nabla^2 h_i(x^*), i = 1, \ldots, N\right)$. In Proposition 5.1 we showed that $\mathrm{TC}(\mathbf{x}^*, \boldsymbol{\Omega}) = \{\mathbb{1} \otimes h \mid h \in \mathrm{TC}(x^*, \Omega)\}$, and therefore the assumption $z' \sum_{i=1}^N \left[\nabla^2 f_i(x^*) + \mu_i^* \nabla^2 h_i(x^*)\right] z > 0$ for all $x \in \mathrm{TC}(x^*, \Omega)$ is equivalent to $\mathbf{x}' \nabla^2_{\mathbf{xx}} \mathcal{L}_0(\mathbf{x}^*, \boldsymbol{\mu}^*, \boldsymbol{\lambda}^*) \mathbf{x} > 0 \ \forall \mathbf{x} \in \mathrm{TC}(\mathbf{x}^*, \boldsymbol{\Omega})$. Pick $\bar{c}$ such that $\nabla^2_{\mathbf{xx}} \mathcal{L}_{\bar{c}}(\mathbf{x}^*, \boldsymbol{\mu}^*, \boldsymbol{\lambda}^*) \succ 0$ and $\bar{c} > \max\{-2\mathbf{e}_1, \ldots, -2\mathbf{e}_{nN}\}$. Let $c_{\max}$, δ and δ_1 as in Proposition 6.1. As before, we make the following transformation: $\bar{\boldsymbol{\lambda}}_k = (\mathbf{I} - \mathbf{J})\boldsymbol{\lambda}_k$, where $\mathbf{J}$ is the projection operator on $\mathrm{Null}(\mathbf{S}')$. Making the observation that $\mathbf{S}' = \mathbf{S}'(\mathbf{I} - \mathbf{J})$ and that $(\mathbf{I} - \mathbf{J})\bar{\boldsymbol{\lambda}}_k = (\mathbf{I} - \mathbf{J})\boldsymbol{\lambda}_k$, iterations (31)-(33) become

$$\mathbf{x}_k = \arg\min_{\mathbf{x}} \mathcal{L}_{c_k}(\mathbf{x}, \boldsymbol{\mu}_k, \bar{\boldsymbol{\lambda}}_k), \ \mathbf{x}_0 = \mathbf{x}^0, \tag{46}$$

$$\boldsymbol{\mu}_{k+1} = \boldsymbol{\mu}_k + c_k \mathbf{h}(\mathbf{x}_k), \ \boldsymbol{\mu}_0 = \boldsymbol{\mu}^0, \tag{47}$$

$$\bar{\boldsymbol{\lambda}}_{k+1} = (\mathbf{I} - \mathbf{J})\bar{\boldsymbol{\lambda}}_k + c_k \mathbf{S}\mathbf{x}_k, \ \bar{\boldsymbol{\lambda}}_0 = \bar{\boldsymbol{\lambda}}^0. \tag{48}$$

that are exactly the iterations found in Theorem 6.1. All assumptions of Theorem 6.1 are satisfied and the result follows. ■

Remark 6.1: At each step of Algorithm ($\mathtt{A}_3$), we use iteration (11) to obtain the solution of (8), and therefore the convergence is dependent on solving

$$\begin{aligned} \min \quad & \mathcal{L}_c(\mathbf{x}, \boldsymbol{\eta}_k) \\ \text{subject to} \quad & \mathbf{x} \in \mathcal{S}(\mathbf{x}^*; \varepsilon). \end{aligned}$$

The solution of the above problem is well defined if it is "close enough" to the local minimizer $\mathbf{x}^*$. However, the unconstrained optimization problem may have multiple local minimizers. Thus for the algorithm to converge to the correct solution, $\mathbf{x}_k$ must remain in a neighborhood of the same local minimizer, at least after some time instant k. Practice showed that using $\mathbf{x}_k$ as starting point in (11) to compute $\mathbf{x}_{k+1}$ tends to ensure that the solutions

of the unconstrained optimization problems remain in a neighborhood of the same local minimizer. In addition to starting closed enough from $\mathbf{x}^*$, appropriate step-sizes α_τ must be used so that (11) converges. A "sufficiently small" constant sequence ($\alpha_\tau = \alpha$) or a slowly diminishing sequence ($\alpha_\tau \to 0, \sum_\tau \alpha_\tau = \infty$) can be chosen. Conditions on the stepsize sequence that ensure convergence can be found in [1] (Propositions 1.2.3 and 1.2.4).□

Remark 6.2: Corollary 6.1 shows that we can use the method of multipliers to compute a local minimizer for the Problem (P_1). Note the change in the condition the initial values $\boldsymbol{\eta}_0$ must satisfy, compared to the original result, namely the projection of $\boldsymbol{\lambda}_0$ on Range($\mathbf{S}$). This change was necessary as a result of the lack of regularity of the local minimizer; lack of regularity that also prevented us from showing that Algorithm (A_1) can achieve a (theoretical) superlinear rate of convergence due to the upper-bound imposed on the sequence $\{c_k\}$.□

Although we assumed the sequence $\{c_k\}$ to be globally known by all agents, there are simple strategies to choosing such sequences that do not require significant initial communication overhead. For example, each agent can compute c_k according to the scheme $c_{k+1} = \max\{\beta c_k, c_{\max}\}$ for some scalar $\beta > 1$, and where only the initial value c_0, the upper-bound $c_{\max}$ and the scalar β must be known by all agents. As seen earlier, the method of multipliers involves solving at each time instant the unconstrained optimization problem (8). To solve this step, we use a gradient method (iteration (11)), since, due to the nature of the cost function, it can be implemented in a distributed manner. In practice, the step (8) is not solved exactly, and usually the iteration (11) is stopped as some stopping criterion is satisfied. Section 2.5 of [2] introduces the "asymptotically exact minimization in methods of multipliers", which basically shows several strategies for solving approximately the step (8), and still obtain convergence to a local minimizer.

References

[1] D. P. Bertsekas. *Nonlinear Programming*. Athena Scientific, Belmont, MA, 1999.

[2] D.P. Bertsekas. *Constrained Optimization and Langrange Multipliers Methods*. Athena Scientific, Belmont, MA, 1982.

[3] S. Boyd, N. Parikh, E. Chu, B. Peleato, and J. Eckstein. Distributed optimization and statistical learning via the alternating direction method of multipliers. *Found. Trends Mach. Learn.*, 3(1):1–122, January 2011.

[4] J. Chen and A.H. Sayed. Diffusion adaptation strategies for distributed optimization and learning over networks. *Signal Processing, IEEE Transactions on*, 60(8):4289 –4305, aug. 2012.

[5] A. Cherukuri and J. Cortés. Distributed generator coordination for initialization and anytime optimization in economic dispatch. *IEEE Transactions on Control of Network Systems (Submitted)*, 2013.

[6] M.R. Hestenes. Multiplier and gradient methods. *Journal of Optimization Theory and Applications*, 4:303–320, 1969.

[7] V.I. Istratescu. *Fixed Point Theory, An Introduction*. D.Reidel, the Netherlands., 1981.

[8] D. Jakovetic, M.F.J. Moura, and J. Xavier. Linear convergence rate of a class of distributed augmented lagrangian algorithms. *CoRR*, abs/1307.2482, 2013.

[9] D. Jakovetic, J. Xavier, and J.M.F. Moura. Cooperative convex optimization in networked systems: Augmented lagrangian algorithms with directed gossip communication. *Signal Processing, IEEE Transactions on*, 59(8):3889–3902, Aug.

[10] B. Johansson, T. Keviczky, M. Johansson, and K.H. Johansson. Subgradient methods and consensus algorithms for solving convex optimization problems. *Proceedings of the 47th IEEE Conference on Decision and Control*, pages 4185–4190, Dec 2008.

[11] B. Johansson, M. Rabi, and K.H. Johansson. A randomized incremental subgradient method for distributed optimization in networked systems. *SIAM Journal on Optimization*, 20(3):1157–1170, 2009.

[12] I. Lobel and A. Ozdaglar. Distributed subgradient methods for convex optimization over random networks. *Automatic Control, IEEE Transactions on*, 56(6):1291–1306, june 2011.

[13] I. Lobel, A. Ozdaglar, and D. Feijer. Distributed multi-agent optimization with state-dependent communication. *Mathematical Programming*, 129(2):255–284, 2011.

[14] I. Matei and J.S. Baras. Performance evaluation of the consensus-based distributed subgradient method under random communication topologies. *Selected Topics in Signal Processing, IEEE Journal of*, 5(4):754–771, aug. 2011.

[15] I. Matei and J.S. Baras. Distributed algorithms for optimization problems with equality constraints. In *Decision and Control (CDC), 2013 IEEE 52nd Annual Conference on*, pages 2352–2357, Dec 2013.

[16] I. Matei, J.S. Baras, M. Nabi, and T. Kurtoglu. An extension of the method of multipliers for distributed nonlinear programming. In *Decision and Control (CDC), 2014 IEEE 53nd Annual Conference on*, pages 6951–6956, Dec 2014.

[17] A. Nedic. Asynchronous broadcast-based convex optimization over a network. *Automatic Control, IEEE Transactions on*, 56(6):1337 –1351, june 2011.

[18] A. Nedic, A. Ozdaglar, and P.A. Parrilo. Constrained consensus and optimization in multi-agent networks. *Automatic Control, IEEE Transactions on*, 55(4):922 –938, april 2010.

[19] A. Nedic, A. Ozdaglar, and P.A. Parrilo. Constrained consensus and optimization in multi-agent networks. *IEEE Trans. Autom. Control*, 55(4):922–938, Apr 2010.

[20] A. Nedic and A. Ozdalgar. Distributed subgradient methods for multi-agent optimization. *IEEE Trans. Autom. Control*, 54(1):48–61, Jan 2009.

[21] S. Sundhar Ram, A. Nedic, and V. V. Veeravalli. Distributed stochastic subgradient projection algorithms for convex optimization. *Journal of Optimization Theory and Applications*, pages 516–545, 2010.

[22] S.S. Ram, V.V. Veeravalli, and A. Nedic. Distributed and recursive parameter estimation in parametrized linear state-space models. *Automatic Control, IEEE Transactions on*, 55(2):488–492, feb. 2010.

[23] W. Shi, Q. Ling, K. Yuan, G Wu, and W. Yin. On the linear convergence of the admm in decentralized consensus optimization. *Signal Processing, IEEE Transactions on*, 62(7):1750–1761, April 2014.

[24] K. Srivastava and A. Nedic. Distributed asynchronous constrained stochastic optimization. *Selected Topics in Signal Processing, IEEE Journal of*, 5(4):772 –790, aug. 2011.

[25] H. Terelius, U. Topcu, and R. M. Murray. Decentralized multi-agent optimization via dual decomposition. In *18th World Congress of the International Federation of Automatic Control (IFAC)*, August 2011.

[26] P.P. Varaiya. *Notes on optimization*. Van Nostrand Reinhold Co., 1972.

[27] E. Wei and A. Ozdaglar. Distributed alternating direction method of multipliers. In *Decision and Control (CDC), 2012 IEEE 51st Annual Conference on*, pages 5445–5450, Dec 2012.

[28] M. Zhu and S. Martinez. An approximate dual subgradient algorithm for multi-agent non-convex optimization. *Automatic Control, IEEE Transactions on*, 58(6):1534–1539, June 2013.

Appendix

A. *Proofs of the supporting results for the distributed algorithm based on the Lagrangian methods*

This section includes the missing proofs of the supporting results in Section V plus the auxiliary results necessary to achieve this goal. We start with a well known result on the properties of the tangent cone to the constraint set at a local minimizer of (P_1).

Proposition A.1: Let Assumptions 2.1-(a) and 2.2 hold, let x^* be a local minimizer of (P_1) and let Ω denote the constraint set, that is, $\Omega = \{x \mid h(x) = 0\}$. Then the tangent cone to Ω at x^* is given by $\mathrm{TC}(x^*, \Omega) = \mathrm{Null}(\nabla h(x^*)')$, where $\nabla h(x^*) \triangleq [\nabla h_1(x^*), \nabla h_2(x^*), \ldots, \nabla h_N(x^*)]$.

Let $\mathbf{x}^* = \mathbb{1} \otimes x^*$ denote a local minimizer of (P_2) and let $\nabla \mathbf{h}(\mathbf{x}^*)$ denote the matrix $\nabla \mathbf{h}(\mathbf{x}^*) \triangleq [\nabla \mathbf{h}_1(\mathbf{x}^*), \nabla \mathbf{h}_2(\mathbf{x}^*), \ldots, \nabla \mathbf{h}_N(\mathbf{x}^*)]$. The vectors $\nabla \mathbf{h}_i(\mathbf{x}^*)$ are the gradients of the functions $\mathbf{h}_i(\mathbf{x})$ at $\mathbf{x}^*$ with a structure given by

$$\nabla \mathbf{h}_i(\mathbf{x}^*)' = \left[\underbrace{0, \ldots, 0}_{n \text{ zeros}}, \ldots, \underbrace{0, \ldots, 0}_{n \text{ zeros}}, \underbrace{\nabla h_i(x^*)'}_{i^{\text{th}} \text{ component}}, \underbrace{0, \ldots, 0}_{n \text{ zeros}}, \ldots, \underbrace{0, \ldots, 0}_{n \text{ zeros}} \right], \tag{49}$$

as per definition of the function $\mathbf{h}_i(\mathbf{x})$.

The second result of this section is concerned with the nullspace of the matrix $[\nabla \mathbf{h}(\mathbf{x}^*), \mathbf{S}']$, which will be used to characterize the tangent cone at a local minimizer of (P_2).

Proposition A.2: Let Assumptions 2.1 and 2.2 hold. The nullspace of the matrix $[\nabla \mathbf{h}(\mathbf{x}^*), \mathbf{S}']$ is given by $\mathrm{Null}([\nabla \mathbf{h}(\mathbf{x}^*), \mathbf{S}']) = \{(\mathbf{0}', \mathbf{v}')' \mid \mathbf{v} \in \mathrm{Null}(\mathbf{S}')\}$.

Proof: Let $\mathbf{u} \in \mathbb{R}^N$ and $\mathbf{v} \in \mathbb{R}^{nN}$ be two vectors. To characterize the nullspace of $[\nabla \mathbf{h}(\mathbf{x}^*), \mathbf{S}']$ we need to check for what values of $\mathbf{u}$ and $\mathbf{v}$ the equation

$$\nabla \mathbf{h}(\mathbf{x}^*)\mathbf{u} + \mathbf{S}'\mathbf{v} = \mathbf{0} \tag{50}$$

is satisfied. Using the definition of $\nabla \mathbf{h}_i(\mathbf{x}^*)$ shown in (49), equation (50) can be equivalently written as

$$\nabla h_i(x^*)\mathbf{u}_i + \sum_{j \in \mathcal{N}_i} \left(s_{ij} v_{ij} - s_{ji} v_{ji} \right) = 0, \ i = 1, \ldots, N,$$

where $\mathbf{u} = (\mathbf{u}_i)$ with $\mathbf{u}_i \in \mathbb{R}$ and $i = 1, \ldots, N$, and $\mathbf{v} = (\mathbf{v}_i)$ with $\mathbf{v}_i \in \mathbb{R}^{n|\mathcal{N}_i|}$ and $\mathbf{v}_i = (v_{ij})$ with $v_{ij} \in \mathbb{R}^n$ and $j \in \mathcal{N}_i$. Summing the above equations over i we obtain that

$$\sum_{i=1}^{N} \mathbf{u}_i \nabla h_i(x^*) = 0,$$

and since $\nabla h(x^*)$ is assumed full rank we must have that $\mathbf{u} = 0$ and the result follows. ■

We now have all the machinery necessary to characterize the tangent cone at a local minimizer of (P_2).

Proof of Proposition 5.1: All we have to show is that any vector in $\mathrm{Null}([\nabla \mathbf{h}(\mathbf{x}^*), \mathbf{S}']')$ belongs to $\mathrm{TC}(\mathbf{x}^*, \mathbf{\Omega})$ as well, since it is well known that (the closure of the convex hull of) $\mathrm{TC}(\mathbf{x}^*, \mathbf{\Omega})$ is included in $\mathrm{Null}([\nabla \mathbf{h}(\mathbf{x}^*), \mathbf{S}']')$. Let $\mathbf{u}$ be a vector in $\mathrm{Null}([\nabla \mathbf{h}(\mathbf{x}^*), \mathbf{S}']')$ and therefore it must satisfy

$$\nabla \mathbf{h}(\mathbf{x}^*)'\mathbf{u} = 0 \text{ and } \mathbf{S}\mathbf{u} = 0. \tag{51}$$

From the second equation of (51), $\mathbf{u}$ must be of the form $\mathbf{u} = \mathbb{1} \otimes u$, for some $u \in \mathbb{R}^n$. From the first equation of (51), using the definition of $\nabla \mathbf{h}_i(\mathbf{x}^*)$ in (49) together with the particular structure of $\mathbf{u}$, we obtain that

$$\nabla h_i(x^*)'u = 0 \ \forall i = 1, \ldots, N,$$

or equivalently

$$u \in \mathrm{Null}(\nabla h(x^*)').$$

We need to show that a vector $\mathbf{u} = \mathbb{1} \otimes u$, with $u \in \text{Null}(\nabla h(x^*)')$ belongs to $\text{TC}(\mathbf{x}^*, \mathbf{\Omega})$. More explicitly, using the definition of the tangent cone, we must find a function $\mathbf{o} : \mathbb{R} \to \mathbb{R}^{nN}$, with $\lim_{t \to 0, t>0} \frac{\mathbf{o}(t)}{t} = 0$, so that

$$\mathbf{x}^* + t\mathbf{u} + \mathbf{o}(t) \in \mathbf{\Omega} \ \forall t > 0.$$

Choosing $\mathbf{o}(t) = \mathbb{1}_N \otimes o(t)$, where $o : \mathbb{R} \to \mathbb{R}^n$ is a function so that $\lim_{t \to 0, t>0} \frac{o(t)}{t} = 0$, we note that

$$\mathbf{g}\big(\mathbf{x}^* + t\mathbf{u} + \mathbf{o}(t)\big) = 0 \ \forall t > 0,$$

and therefore, all we are left to do is to check that

$$\mathbf{h}\big(\mathbf{x}^* + t\mathbf{u} + \mathbf{o}(t)\big) = 0 \ \forall t > 0, \tag{52}$$

as well. Making the observation that $\mathbf{x}^* + t\mathbf{u} + \mathbf{o}(t) = \mathbb{1} \otimes (x^* + tu + o(t))$, (52) is equivalent to showing that

$$h\big(x^* + tu + o(t)\big) = 0 \ \forall t > 0. \tag{53}$$

However, we showed previously that $u \in \text{Null}(\nabla h(x^*)')$, and therefore by Proposition A.2 $u \in TC(x^*, \Omega)$, as well. Therefore there exits a function $o(t)$ so that (53) is satisfied, which shows that indeed

$$\text{TC}(\mathbf{x}^*, \mathbf{\Omega}) = \text{Null}\Big([\nabla \mathbf{h}(\mathbf{x}^*), \mathbf{S}']'\Big),$$

and consequently $\text{TC}(\mathbf{x}^*, \mathbf{\Omega})$ is a closed and convex subspace. ■

Let $\mathbf{x}^* = \mathbb{1} \otimes x^*$ denote a local minimizer of (P_2). From the theory concerning optimization problems with equality constraints (see for example Chapter 3, page 15 of [26], or Chapter 3, page 253 of [1]), the first order necessary conditions for (P_2) ensure the existence of $\lambda_0^* \in \mathbb{R}$, $\boldsymbol{\mu}^* \in \mathbb{R}^N$ and $\boldsymbol{\lambda}^* \in \mathbb{R}^{n\bar{N}}$ so that $\lambda_0^* \nabla \mathbf{F}(\mathbf{x}^*) + \nabla \mathbf{h}(x^*)\boldsymbol{\mu}^* + \mathbf{S}'\boldsymbol{\lambda}^* = 0$. Since $\mathbf{S}$ is not full rank, and therefore the matrix $[\nabla \mathbf{h}(\mathbf{x}^*), \mathbf{S}']$ is not full rank either, the uniqueness of $\boldsymbol{\mu}^*$ and $\boldsymbol{\lambda}^*$ cannot be guaranteed. The following result characterizes the set of Lagrange multipliers verifying the first order necessary conditions of (P_2).

Proposition A.3: Let Assumptions 2.1 and 2.2 hold and let $\mathbf{x}^* = \mathbb{1} \otimes x^*$ be a local minimizer for problem (P_2). There exist unique vectors $\boldsymbol{\mu}^*$ and $\boldsymbol{\lambda}^* \in \text{Range}(\mathbf{S})$ so that $\nabla \mathbf{F}(\mathbf{x}^*) + \nabla \mathbf{h}(\mathbf{x}^*)\boldsymbol{\mu}^* + \mathbf{S}'\boldsymbol{\lambda} = 0$ for all $\boldsymbol{\lambda} \in \{\boldsymbol{\lambda}^* + \boldsymbol{\lambda}_\perp \mid \boldsymbol{\lambda}_\perp \in \text{Null}(\mathbf{S}')\}$.

Proof: By Lemma 1[2] , page 50 of [26] we have that $\nabla \mathbf{F}(\mathbf{x}^*)$ is orthogonal on $\text{TC}(\mathbf{x}^*, \mathbf{\Omega})$ and therefore, by Proposition 5.1, $\nabla \mathbf{F}(\mathbf{x}^*)$ must belong to $\text{Range}([\nabla \mathbf{h}(\mathbf{x}^*), \mathbf{S}'])$. Consequently, there exist the vectors $\boldsymbol{\mu}^*$ and $\boldsymbol{\lambda}$ so that

$$-\nabla \mathbf{F}(\mathbf{x}^*) = \nabla \mathbf{h}(\mathbf{x}^*)\boldsymbol{\mu}^* + \mathbf{S}'\boldsymbol{\lambda}. \tag{54}$$

Noting that $\mathbb{R}^{nN}$ can be written as a direct sum between the nullspace of $\mathbf{S}'$ and the range of $\mathbf{S}$, there exist the orthogonal vectors $\boldsymbol{\lambda}^* \in \text{Range}(\mathbf{S})$ and $\boldsymbol{\lambda}_\perp \in \text{Null}(\mathbf{S}')$ so that $\boldsymbol{\lambda} = \boldsymbol{\lambda}^* + \boldsymbol{\lambda}_\perp$. Note that we can replace $\boldsymbol{\lambda}_\perp$ by any vector in $\text{Null}(\mathbf{S}')$ and (54) still holds. The only thing left to do is to prove the uniqueness of $\boldsymbol{\mu}^*$ and $\boldsymbol{\lambda}^*$. We use a

[2]The result states that given a local minimizer x^* of a function $f(x)$, $h'\nabla f(x^*) \geq 0$ for all $h \in TC(x^*, \Omega)$. When $TC(x^*, \Omega)$ is a (closed, convex) subspace, orthogonality follows.

contradiction argument. Let $\tilde{\boldsymbol{\mu}} \neq \boldsymbol{\mu}^*$ and $\tilde{\boldsymbol{\lambda}} \neq \boldsymbol{\lambda}^*$ with $\tilde{\boldsymbol{\lambda}} \in \text{Range}(\mathbf{S})$ be two vectors so that (54) is satisfied. Hence we have that

$$-\nabla \mathbf{F}(\mathbf{x}^*) = \nabla \mathbf{h}(\mathbf{x}^*)\boldsymbol{\mu}^* + \mathbf{S}'\boldsymbol{\lambda}^* \text{ and } -\nabla \mathbf{F}(\mathbf{x}^*) = \nabla \mathbf{h}(\mathbf{x}^*)\tilde{\boldsymbol{\mu}} + \mathbf{S}'\tilde{\boldsymbol{\lambda}},$$

and therefore

$$0 = \nabla \mathbf{h}(\mathbf{x}^*)(\boldsymbol{\mu}^* - \tilde{\boldsymbol{\mu}}) + \mathbf{S}'\left(\boldsymbol{\lambda}^* - \tilde{\boldsymbol{\lambda}}\right).$$

By Proposition A.2 we have that

$$\text{Null}([\nabla \mathbf{h}(\mathbf{x}^*), \mathbf{S}']) = \left\{(\mathbf{0}', \mathbf{v}')' \mid \mathbf{v} \in \text{Null}(\mathbf{S}')\right\},$$

and therefore $\boldsymbol{\mu}^* = \tilde{\boldsymbol{\mu}}$ and $\boldsymbol{\lambda}^* = \tilde{\boldsymbol{\lambda}}$ since $\boldsymbol{\lambda}^* - \tilde{\boldsymbol{\lambda}} \in \text{Range}(\mathbf{S})$, and the result follows. ■

We can now proceed with the proofs of Proposition 5.2 and Lemma A.1.

Proof of Proposition 5.2:

By Proposition A.3, there exist two unique vector $\boldsymbol{\mu}^*$ and $\boldsymbol{\lambda}^* \in \text{Range}(\mathbf{L})$ so that

$$\nabla \mathbf{F}(\mathbf{x}^*) + \nabla \mathbf{h}(\mathbf{x}^*)\boldsymbol{\mu}^* + \mathbf{S}'\boldsymbol{\lambda}^* = 0.$$

Using the structure of $\nabla \mathbf{F}(\mathbf{x}^*)$, $\mathbf{h}(\mathbf{x}^*)$ and $\mathbf{S}'$, the above equation can be equivalently expressed as

$$\nabla f_i(x^*) + \boldsymbol{\mu}_i^* \nabla h_i(x^*) + \sum_{j \in \mathcal{N}_i} \left(s_{ij}\lambda_{ij}^* - s_{ji}\lambda_{ji}^*\right), \; i = 1, \ldots, N, \tag{55}$$

where $\boldsymbol{\mu}_i^*$ are the scalar entries of $\boldsymbol{\mu}^*$ and $\boldsymbol{\lambda}_i^* = (\lambda_{ij}^*)$ are the $n|\mathcal{N}_i|$-dimensional sub-vectors of $\boldsymbol{\lambda}^*$. Summing up equations (55) over i, we obtain

$$\sum_{i=1}^{N} \nabla f_i(x^*) + \sum_{i=1}^{N} \nabla h_i(x^*)\boldsymbol{\mu}_i^* = 0.$$

Equivalently,

$$\nabla f(x^*) + \nabla h(x^*)\boldsymbol{\mu}^* = 0,$$

which is just the first order necessary condition for (P_1). But since $\boldsymbol{\mu}^*$ must be unique, it follows that $\boldsymbol{\mu}^* = \psi^*$. ■

The convergence properties of the first two the distributed algorithms depend on the spectral properties of a particular matrix; properties analyzed in the following result.

Lemma A.1: Let Assumptions 2.1 and 2.2 hold, let α be a positive scalar, and let $\mathbf{x}^*$ be a local minimizer of (P_2). Then the eigenvalues of the matrix

$$\mathbf{B} = \begin{pmatrix} \mathbf{H} & \nabla \mathbf{h}(\mathbf{x}^*) & \mathbf{S}' \\ -\nabla \mathbf{h}(\mathbf{x}^*)' & \mathbf{0} & \mathbf{0} \\ -\mathbf{J} & \mathbf{0} & \frac{1}{\alpha}\mathbf{J} \end{pmatrix},$$

have positive real parts, where $\mathbf{H}$ is a positive definite matrix and $\mathbf{J}$ is the orthogonal projection operator on $\text{Null}(\mathbf{S}')$.

Proof: Let β be an eigenvalue of $\mathbf{B}$ and let $(\mathbf{u}', \mathbf{v}', \mathbf{z}')' \neq 0$ be the corresponding eigenvector, where $\mathbf{u}$, $\mathbf{v}$ and $\mathbf{z}$ are complex vectors of appropriate dimensions. Denoting by $\hat{\mathbf{u}}$, $\hat{\mathbf{v}}$ and $\hat{\mathbf{z}}$ the conjugates of $\mathbf{u}$, $\mathbf{v}$ and $\mathbf{z}$, respectively we have

$$\mathrm{Re}(\beta)\left(||\mathbf{u}||^2 + ||\mathbf{v}||^2 + ||\mathbf{z}||^2\right) = \mathrm{Re}\left\{(\hat{\mathbf{u}}', \hat{\mathbf{v}}', \hat{\mathbf{z}}')\mathbf{B}\begin{pmatrix} \mathbf{u} \\ \mathbf{v} \\ \mathbf{z} \end{pmatrix}\right\} =, \tag{56}$$

$$\mathrm{Re}\left\{\hat{\mathbf{u}}'\mathbf{H}\mathbf{u} + \hat{\mathbf{u}}'\mathbf{S}'\mathbf{z} - \hat{\mathbf{z}}'\mathbf{S}\mathbf{u} + \hat{\mathbf{u}}'\nabla\mathbf{h}(\mathbf{x}^*)\mathbf{v} - \hat{\mathbf{v}}'\nabla\mathbf{h}(\mathbf{x}^*)'\mathbf{u} + \hat{\mathbf{z}}'\frac{1}{\alpha}\mathbf{J}\mathbf{z}\right\}$$

$$= \mathrm{Re}\left\{\hat{\mathbf{u}}^T\mathbf{H}\mathbf{u} + \hat{\mathbf{z}}'\frac{1}{\alpha}\mathbf{J}\mathbf{z}\right\}.$$

Since $\mathbf{J}$ is a semi-positive definite matrix and $\mathbf{H}$ is positive definite we have that

$$\mathrm{Re}(\beta)\left(||\mathbf{u}||^2 + ||\mathbf{v}||^2 + ||\mathbf{z}||^2\right) > 0,$$

as long as $\mathbf{u} \neq 0$ or $\mathbf{z} \notin \mathrm{Range}(\mathbf{S})$ and therefore $\mathrm{Re}(\beta) > 0$. In the case $\mathbf{u} = 0$ and $\mathbf{z} \in \mathrm{Range}(\mathbf{S})$ we get

$$\mathbf{B}\begin{pmatrix} 0 \\ \mathbf{v} \\ \mathbf{z} \end{pmatrix} = \beta\begin{pmatrix} 0 \\ \mathbf{v} \\ \mathbf{z} \end{pmatrix},$$

from where we obtain

$$\nabla\mathbf{h}(\mathbf{x}^*)\mathbf{v} + \mathbf{S}'\mathbf{z} = 0.$$

But from Proposition A.2, we have that $\mathbf{v} = 0$ and $\mathbf{z} \in \mathrm{Null}(\mathbf{S}')$ and since $\mathbf{z} \in \mathrm{Range}(\mathbf{S})$ as well, it must be that $\mathbf{z} = 0$. Hence we have a contradiction since we assumed that $(\mathbf{u}', \mathbf{v}', \mathbf{z}') \neq \mathbf{0}'$ and therefore the real part of β must be positive. In addition, it can be easily checked that the matrix $\mathbf{B}$ has n eigenvalues equal to $\frac{1}{\alpha}$ and their corresponding eigenspace is $\{(0', 0', \mathbf{z}')' \mid \mathbf{z} \in \mathrm{Null}(\mathbf{S}')\}$. ■

We finalize this section by providing the proofs of the theorems describing the convergence properties of Algorithms ($\mathtt{A}_1$) and ($\mathtt{A}_2$) for solving ($\mathtt{P}_2$).

Proof of Theorem 5.1: Using the Lagrangian function defined in (15), iteration (22)-(24) can be equivalently expressed as

$$\begin{pmatrix} \mathbf{x}_{k+1} \\ \boldsymbol{\mu}_{k+1} \\ \boldsymbol{\lambda}_{k+1} \end{pmatrix} = \bar{\mathbf{M}}_\alpha(\mathbf{x}_k, \boldsymbol{\mu}_k, \boldsymbol{\lambda}_k), \tag{57}$$

with

$$\bar{\mathbf{M}}_\alpha(\mathbf{x}, \boldsymbol{\mu}, \boldsymbol{\lambda}) = \begin{pmatrix} \mathbf{x} - \alpha\nabla_{\mathbf{x}}\mathcal{L}(\mathbf{x}, \boldsymbol{\mu}, \boldsymbol{\lambda}) \\ \boldsymbol{\mu} + \alpha\nabla_{\boldsymbol{\mu}}\mathcal{L}(\mathbf{x}, \boldsymbol{\mu}, \boldsymbol{\lambda}) \\ \boldsymbol{\lambda} + \alpha\nabla_{\boldsymbol{\lambda}}\mathcal{L}(\mathbf{x}, \boldsymbol{\mu}, \boldsymbol{\lambda}) \end{pmatrix}.$$

It can be easily checked that $(\mathbf{x}^*, \boldsymbol{\mu}^*, \boldsymbol{\lambda}^* + \text{Null}(\mathbf{S}'))$ is a set of fixed points of $\bar{\mathbf{M}}_\alpha$. Let us now consider the transformation $\tilde{\boldsymbol{\lambda}} = (\mathbf{I} - \mathbf{J})\boldsymbol{\lambda}$, where $\mathbf{J}$ is the orthogonal projection operator on Null($\mathbf{S}'$). This transformation extracts the projection of $\boldsymbol{\lambda}$ on the nullspace of $\mathbf{S}'$ from $\boldsymbol{\lambda}$ and therefore $\tilde{\boldsymbol{\lambda}}$ is the error between $\boldsymbol{\lambda}$ and its orthogonal projection on $\text{Null}(\mathbf{S}')$. Under this transformation, iteration (57) becomes

$$\begin{pmatrix} \mathbf{x}_{k+1} \\ \boldsymbol{\mu}_{k+1} \\ \tilde{\boldsymbol{\lambda}}_{k+1} \end{pmatrix} = \mathbf{M}_\alpha(\mathbf{x}_k, \boldsymbol{\mu}_k, \tilde{\boldsymbol{\lambda}}_k)$$

with

$$\mathbf{M}_\alpha(\mathbf{x}, \boldsymbol{\mu}, \tilde{\boldsymbol{\lambda}}) = \begin{pmatrix} \mathbf{x} - \alpha \nabla_{\mathbf{x}} \mathcal{L}(\mathbf{x}, \boldsymbol{\mu}, \tilde{\boldsymbol{\lambda}}) \\ \boldsymbol{\mu} + \alpha \nabla_{\boldsymbol{\mu}} \mathcal{L}(\mathbf{x}, \boldsymbol{\mu}, \tilde{\boldsymbol{\lambda}}) \\ (\mathbf{I} - \mathbf{J})\tilde{\boldsymbol{\lambda}} + \alpha \nabla_{\tilde{\boldsymbol{\lambda}}} \mathcal{L}(\mathbf{x}, \boldsymbol{\mu}, \tilde{\boldsymbol{\lambda}}) \end{pmatrix},$$

where we used the fact that $(\mathbf{I} - \mathbf{J})\tilde{\boldsymbol{\lambda}} = (\mathbf{I} - \mathbf{J})\boldsymbol{\lambda}$, $(\mathbf{I} - \mathbf{J})\mathbf{S}\mathbf{x} = \mathbf{S}\mathbf{x}$, since $\mathbf{S}\mathbf{x} \in \text{Range}(\mathbf{S})$, and $\mathbf{S}'\boldsymbol{\lambda} = \mathbf{S}'(\tilde{\boldsymbol{\lambda}} + \mathbf{J}\boldsymbol{\lambda}) = \mathbf{S}'\tilde{\boldsymbol{\lambda}}$. Clearly $(\mathbf{x}^*, \boldsymbol{\mu}^*, \boldsymbol{\lambda}^*)$ is a fixed point for $\mathbf{M}_\alpha$ and if $\left(\mathbf{x}_k, \boldsymbol{\mu}_k, \tilde{\boldsymbol{\lambda}}_k\right)$ converges to $(\mathbf{x}^*, \boldsymbol{\mu}^*, \boldsymbol{\lambda}^*)$, we in fact show that $(\mathbf{x}_k, \boldsymbol{\mu}_k, \boldsymbol{\lambda}_k)$ converges to $(\mathbf{x}^*, \boldsymbol{\mu}^*, \boldsymbol{\lambda}^* + \text{Null}(\mathbf{S}'))$. The derivative of the mapping $\mathbf{M}_\alpha(\mathbf{x}, \boldsymbol{\mu}, \boldsymbol{\lambda})$ at $(\mathbf{x}^*, \boldsymbol{\mu}^*, \boldsymbol{\lambda}^*)$ is given by

$$\nabla \mathbf{M}_\alpha(\mathbf{x}^*, \boldsymbol{\mu}^*, \boldsymbol{\lambda}^*) = \mathbf{I} - \alpha \mathbf{B},$$

where

$$\mathbf{B} = \begin{pmatrix} \nabla^2_{\mathbf{xx}} \mathcal{L}(\mathbf{x}^*, \boldsymbol{\mu}^*, \boldsymbol{\lambda}^*) & \nabla \mathbf{h}(\mathbf{x}^*) & \mathbf{L}' \\ -\nabla \mathbf{h}(\mathbf{x}^*)' & \mathbf{0} & \mathbf{0} \\ -\mathbf{L} & \mathbf{0} & \frac{1}{\alpha}\mathbf{J} \end{pmatrix}.$$

By Lemma A.1 we have that the real parts of the eigenvalues of $\mathbf{B}$ are positive and therefore we can find an $\bar{\alpha}$ so that for all $\alpha \in (0.\bar{\alpha}]$ the eigenvalues of $\nabla \mathbf{M}_\alpha(\mathbf{x}^*, \boldsymbol{\mu}^*, \boldsymbol{\lambda}^*)$ are strictly within the unit circle. Using a similar argument as in Proposition 4.4.1, page 387, [1], there exist a norm $\|\cdot\|$ and a sphere $\mathcal{S}_\epsilon = \left\{(\mathbf{x}', \boldsymbol{\mu}', \boldsymbol{\lambda}')' \mid \|(\mathbf{x}', \boldsymbol{\mu}', \boldsymbol{\lambda}')' - \left(\mathbf{x}^{*\prime}, \boldsymbol{\mu}^{*\prime}, \boldsymbol{\lambda}^{*\prime}\right)'\| < \epsilon\right\}$ for some $\epsilon > 0$ so that the induced norm of $\nabla \mathbf{M}_\alpha(\mathbf{x}, \boldsymbol{\mu}, \boldsymbol{\lambda})$ is less than one within the sphere $\mathcal{S}_\epsilon$. Therefore, using the mean value theorem, it follows that $\mathbf{M}_\alpha(\mathbf{x}, \boldsymbol{\mu}, \boldsymbol{\lambda})$ is a contraction map for any vector in the sphere $\mathcal{S}_\epsilon$. By invoking the contraction map theorem (see for example Chapter 7 of [7]) it follows that $\left(\mathbf{x}_k, \boldsymbol{\mu}_k, \tilde{\boldsymbol{\lambda}}_k\right)$ converges to $(\mathbf{x}^*, \boldsymbol{\mu}^*, \boldsymbol{\lambda}^*)$ for any initial value in S_ϵ. ■

Proof of Theorem 5.2: First note that the assumption on the Hessian of $\mathcal{L}(\mathbf{x}, \boldsymbol{\mu}, \boldsymbol{\lambda})$ basically means that $\mathbf{x}^*$ is a strictly local minimizer of (P_2). Proceeding as in the case of the proof of Theorem 5.1, iteration (28)-(30) can be compactly expressed as

$$\begin{pmatrix} \mathbf{x}_{k+1} \\ \boldsymbol{\mu}_{k+1} \\ \boldsymbol{\lambda}_{k+1} \end{pmatrix} = \bar{\mathbf{M}}_{\alpha,c}(\mathbf{x}_k, \boldsymbol{\mu}_k, \boldsymbol{\lambda}_k), \tag{58}$$

with

$$\bar{\mathbf{M}}_{\alpha,c}(\mathbf{x},\boldsymbol{\mu},\boldsymbol{\lambda}) = \begin{pmatrix} \mathbf{x} - \alpha\nabla_{\mathbf{x}}\mathcal{L}_c(\mathbf{x},\boldsymbol{\mu},\boldsymbol{\lambda}) \\ \boldsymbol{\mu} + \alpha\nabla_{\boldsymbol{\mu}}\mathcal{L}_c(\mathbf{x},\boldsymbol{\mu},\boldsymbol{\lambda}) \\ \boldsymbol{\lambda} + \alpha\nabla_{\boldsymbol{\lambda}}\mathcal{L}_c(\mathbf{x},\boldsymbol{\mu},\boldsymbol{\lambda}), \end{pmatrix},$$

or, expressing (30) in terms of the error between $\boldsymbol{\lambda}_k$ and its projection on $\mathrm{Null}(\mathbf{S}')$, we further have

$$\begin{pmatrix} \mathbf{x}_{k+1} \\ \boldsymbol{\mu}_{k+1} \\ \tilde{\boldsymbol{\lambda}}_{k+1} \end{pmatrix} = \mathbf{M}_{\alpha,c}(\mathbf{x}_k,\boldsymbol{\mu}_k,\tilde{\boldsymbol{\lambda}}_k)$$

with

$$\mathbf{M}_{\alpha,c}(\mathbf{x},\boldsymbol{\mu},\tilde{\boldsymbol{\lambda}}) = \begin{pmatrix} \mathbf{x} - \alpha\nabla_{\mathbf{x}}\mathcal{L}_c(\mathbf{x},\boldsymbol{\mu},\tilde{\boldsymbol{\lambda}}) \\ \boldsymbol{\mu} + \alpha\nabla_{\boldsymbol{\mu}}\mathcal{L}_c(\mathbf{x},\boldsymbol{\mu},\tilde{\boldsymbol{\lambda}}) \\ (\mathbf{I}-\mathbf{J})\tilde{\boldsymbol{\lambda}} + \alpha\nabla_{\tilde{\boldsymbol{\lambda}}}\mathcal{L}_c(\mathbf{x},\boldsymbol{\mu},\tilde{\boldsymbol{\lambda}}) \end{pmatrix},$$

and $\tilde{\boldsymbol{\lambda}}_k = (\mathbf{I}-\mathbf{J})\boldsymbol{\lambda}_k$. Clearly $(\mathbf{x}^*,\boldsymbol{\mu}^*,\boldsymbol{\lambda}^*)$ is a fixed point for $\mathbf{M}_{\alpha,c}$ and we have that

$$\nabla\mathbf{M}_{\alpha,c}(\mathbf{x}^*,\boldsymbol{\mu}^*,\boldsymbol{\lambda}^*) = \mathbf{I} - \alpha\mathbf{B}_c,$$

where

$$\mathbf{B}_c = \begin{pmatrix} \nabla^2_{\mathbf{xx}}\mathcal{L}_c(\mathbf{x}^*,\boldsymbol{\mu}^*,\boldsymbol{\lambda}^*) & \nabla\mathbf{h}(\mathbf{x}^*) & \mathbf{S}' \\ -\nabla\mathbf{h}(\mathbf{x}^*)' & \mathbf{0} & \mathbf{0} \\ -\mathbf{S} & \mathbf{0} & \frac{1}{\alpha}\mathbf{J} \end{pmatrix}.$$

Since $\mathrm{Null}(\mathbf{S}) = \mathrm{Null}(\mathbf{S}'\mathbf{S}) = \mathrm{Null}(\mathbf{L})$, it can be easily checked that $\mathrm{Null}([\nabla\mathbf{h}(\mathbf{x}^*),\mathbf{L}']') = \mathrm{Null}([\nabla\mathbf{h}(\mathbf{x}^*),\mathbf{S}']')$. Using the assumption that $\mathbf{x}'\nabla^2_{\mathbf{xx}}\mathcal{L}(\mathbf{x}^*,\boldsymbol{\mu}^*,\boldsymbol{\lambda}^*)\mathbf{x} > 0$ for all $\mathbf{x} \in \mathrm{TC}(\mathbf{x}^*,\boldsymbol{\Omega}) = \mathrm{Null}([\nabla\mathbf{h}(\mathbf{x}^*),\mathbf{L}']')$, according to Proposition 5.3 there exists a positive scalar $\bar{c}$ such that $\nabla^2_{\mathbf{xx}}\mathcal{L}_c(\mathbf{x}^*,\boldsymbol{\mu}^*,\boldsymbol{\lambda}^*) \succ 0$ for all $c \geq \bar{c}$. Therefore, by Lemma A.1, the real parts of the eigenvalues of $\mathbf{B}_c$ are positive and consequently we can find an $\bar{\alpha}(c)$ so that for all $\alpha \in (0,\bar{\alpha}(c)]$ the eigenvalues of $\nabla\mathbf{M}_{\alpha,c}(\mathbf{x}^*,\boldsymbol{\mu}^*,\boldsymbol{\lambda}^*)$ are strictly within the unit circle. Mimicking the last part of the proof of Theorem 5.1, we find that $\mathbf{M}_{\alpha,c}(\mathbf{x},\boldsymbol{\mu},\boldsymbol{\lambda})$ is a contraction map within a sphere centered at $(\mathbf{x}^*,\boldsymbol{\mu}^*,\boldsymbol{\lambda}^*)$, and the result follows from the contraction map theorem. ■

B. Proofs of the supporting results for the distributed algorithm based on the method of multipliers

The key result behind Algorithm ($\mathtt{A}_3$) is Proposition 6.1, whose proof is provided in what follows.

Proof of Proposition 6.1: By Proposition 5.3 there exists a positive scalar $\bar{c}$ such that $\nabla^2_{\mathbf{xx}}\mathcal{L}_c(\mathbf{x}^*,\boldsymbol{\eta}^*) \succ 0$ for all $c \geq \bar{c}$. Following the same idea as in [2], for $c > 0$ we define the system of equations

$$\nabla\mathbf{F}(\mathbf{x}) + \nabla\tilde{\mathbf{h}}(\mathbf{x})\tilde{\boldsymbol{\eta}} = 0, \tag{59}$$

$$\tilde{\mathbf{h}}(\mathbf{x}) + \gamma(\mathbf{T}\boldsymbol{\eta} - \tilde{\boldsymbol{\eta}})/c = 0. \tag{60}$$

Note that compared to Proposition 2.4, [2], we introduced the operator $\mathbf{T}$ acting on $\boldsymbol{\eta}$. From the equation (60) we

can also note that $\tilde{\lambda} \in \text{Range}(\mathbf{S})$ (consider the structure of $\mathbf{T}$ and $\tilde{\mathbf{h}}(\mathbf{x})$). By introducing the variables

$$t = \frac{\mathbf{T}(\boldsymbol{\eta} - \boldsymbol{\eta}^*)}{c}, \ \gamma = \frac{1}{c},$$

the system (59)-(60) becomes

$$\nabla \mathbf{F}(\mathbf{x}) + \nabla \tilde{\mathbf{h}}(\mathbf{x}) \tilde{\boldsymbol{\eta}} = 0, \tag{61}$$

$$\tilde{\mathbf{h}}(\mathbf{x}) + t + \gamma(\boldsymbol{\eta}^* - \tilde{\boldsymbol{\eta}}) = 0, \tag{62}$$

where we used the fact that $\mathbf{T}\boldsymbol{\eta}^* = \boldsymbol{\eta}^*$. For $t = 0$ and $\gamma \in [0, 1/\bar{c}]$, system (61)-(62) has the solution $\mathbf{x} = \mathbf{x}^*$ and $\tilde{\boldsymbol{\eta}} = \boldsymbol{\eta}^*$. Note that thanks to the manner we defined (60), $(\mathbf{x}^*, \boldsymbol{\eta}^*)$ is the solution of (61)-(62) and not $(\mathbf{x}^*, \boldsymbol{\eta}^* + \boldsymbol{\eta}_\perp)$, with $\boldsymbol{\eta}_\perp = (0, \text{Null}(\mathbf{S}'))$. Basically the entire proof of this proposition is based on the properties of the Jacobian of (61)-(62), with respect of $(\mathbf{x}, \tilde{\boldsymbol{\eta}})$ at the solution $(\mathbf{x}^*, \boldsymbol{\eta}^*)$. This Jacobian is given by

$$\begin{bmatrix} \nabla^2_{\mathbf{xx}} \mathcal{L}_0(\mathbf{x}^*, \boldsymbol{\eta}^*) & \nabla \tilde{\mathbf{h}}(\mathbf{x}^*) \\ \nabla \tilde{\mathbf{h}}(\mathbf{x}^*)' & \gamma \mathbf{I} \end{bmatrix}. \tag{63}$$

Using Proposition 5.3, it can be check that for any $\gamma > 0$, the Jacobian defined in (63) is invertible. For $\gamma = 0$, however, it turns out that the nullspace of the matrix (63) is given by $\{(\mathbf{0}', \mathbf{0}', \mathbf{w}')' \mid \mathbf{w} \in \text{Null}(\mathbf{S}')\}$, and therefore, unlike Proposition 2.4, [2], the Jacobian is not invertible. By choosing an arbitrarily large positive scalar $c_{\max}$ so that $\bar{c} \leq c \leq c_{\max}$, we in fact make sure that the matrix (63) is invertible for all considered values of c.

By defining the compact set $K = \{(0, \gamma) \mid \gamma \in [1/c_{\max}, 1/\bar{c}]\}$ and applying the implicit function theorem (with respect to a compact set), there exist $\delta > 0$, $\varepsilon > 0$, and unique continuously differentiable functions $\hat{\mathbf{x}}(t, \gamma)$ and $\hat{\boldsymbol{\eta}}(t, \gamma)$, defined on $S(K; \delta)$ such that

$$\left(\|\hat{\mathbf{x}}(t, \gamma) - \mathbf{x}^*\|^* + \|\hat{\boldsymbol{\eta}}(t, \gamma) - \boldsymbol{\eta}^*\|^2 \right)^{1/2} < \varepsilon \ \forall (t, \gamma) \in S(K; \delta),$$

and satisfying

$$\nabla \mathbf{F}[\hat{\mathbf{x}}(t, \gamma)] + \nabla \tilde{\mathbf{h}}[\hat{\mathbf{x}}(t, \gamma)] \hat{\boldsymbol{\eta}}(t, \gamma) = 0, \tag{64}$$

$$\tilde{\mathbf{h}}[\hat{\mathbf{x}}(\mathbf{t}, \gamma)] + t + \gamma \boldsymbol{\eta}^* - \gamma \hat{\boldsymbol{\eta}}(t, \gamma) = 0. \tag{65}$$

In addition, from the continuity of the Hessian of the augmented Lagrangian and using the fact that

$$\nabla^2_{\mathbf{xx}} \mathcal{L}_c(\mathbf{x}^*, \boldsymbol{\eta}^*) = \nabla^2_{\mathbf{xx}} \mathcal{L}_0(\mathbf{x}^*, \boldsymbol{\eta}^*) + \frac{1}{\gamma} \nabla \tilde{\mathbf{h}}(\mathbf{x}^*) \nabla \tilde{\mathbf{h}}(\mathbf{x}^*)' \succ 0, \ \forall c \geq \bar{c}$$

δ and ε can be chosen so that

$$\nabla^2_{\mathbf{xx}} \mathcal{L}_0[\hat{\mathbf{x}}(t, \gamma), \hat{\boldsymbol{\eta}}(t, \gamma)] + \frac{1}{\gamma} \nabla \tilde{\mathbf{h}}[\hat{\mathbf{x}}(t, \gamma)] \nabla \tilde{\mathbf{h}}[\hat{\mathbf{x}}(t, \gamma)]' \succ 0 \ \forall (t, \gamma) \in S(K; \delta), \ \bar{c} \leq c \leq c_{\max}.$$

For $\bar{c} \leq c \leq c_{\max}$ and $\|\mathbf{T}\boldsymbol{\eta} - \boldsymbol{\eta}^*\| < c\delta$ we define

$$\mathbf{x}(\boldsymbol{\eta}, c) = \hat{\mathbf{x}}\left(\frac{\mathbf{T}\boldsymbol{\eta} - \boldsymbol{\eta}^*}{c}, \frac{1}{c}\right), \tilde{\boldsymbol{\eta}}(\boldsymbol{\eta}, c) = \hat{\boldsymbol{\eta}}\left(\frac{\mathbf{T}\boldsymbol{\eta} - \boldsymbol{\eta}^*}{c}, \frac{1}{c}\right),$$

and we obtain that for $(\boldsymbol{\eta}, c) \in D$

$$\nabla \mathbf{F}[\mathbf{x}(\boldsymbol{\eta}, c)] + \nabla \tilde{\mathbf{h}}[\mathbf{x}(\boldsymbol{\eta}, c)]\tilde{\boldsymbol{\eta}}(\boldsymbol{\eta}, c) = 0, \tag{66}$$

$$\tilde{\boldsymbol{\eta}}(\boldsymbol{\eta}, c) = \mathbf{T}\boldsymbol{\eta} + c\tilde{\mathbf{h}}[\mathbf{x}(\boldsymbol{\eta}, c)], \tag{67}$$

$$\nabla^2_{\mathbf{xx}} \mathcal{L}_c[\mathbf{x}(\boldsymbol{\eta}, c), \boldsymbol{\eta}] \succ 0. \tag{68}$$

which basically proves the result, except (36) and (37).

The inequalities (36)-(37) are based on the differentiation of (61)-(62), with respect to t and γ. Defining $\gamma_{\min} = 1/c_{\max}$, we obtain that for all (t, γ) such $\|t\| < \delta$ and $\gamma \in [\gamma_{\min}, 1/\bar{c}]$

$$\begin{bmatrix} \hat{\mathbf{x}}(t,\gamma) - \mathbf{x}^* \\ \hat{\boldsymbol{\eta}}(t,\gamma) - \boldsymbol{\eta}^* \end{bmatrix} = \begin{bmatrix} \hat{\mathbf{x}}(t,\gamma) - \hat{\mathbf{x}}(0,\gamma_{\min}) \\ \hat{\boldsymbol{\eta}}(t,\gamma) - \hat{\boldsymbol{\eta}}(0,\gamma_{\min}) \end{bmatrix} = \tag{69}$$

$$\int_0^1 A\left(\xi t, (\gamma - \gamma_{\min})\xi + \gamma_{\min}\right) \begin{bmatrix} 0 & 0 \\ -I & \hat{\boldsymbol{\eta}}(\xi t, (\gamma - \gamma_{\min})\xi + \gamma_{\min}) \end{bmatrix} \begin{bmatrix} t \\ \gamma - \gamma_{\min} \end{bmatrix} d\xi,$$

where

$$A(t,\gamma) = \begin{bmatrix} \nabla^2_{\mathbf{xx}} \mathcal{L}_0\left[\hat{\mathbf{x}}(t,\gamma), \hat{\boldsymbol{\eta}}(t,\gamma)\right] & \nabla \tilde{\mathbf{h}}\left[\hat{\mathbf{x}}(t,\gamma)\right] \\ \nabla \tilde{\mathbf{h}}\left[\hat{\mathbf{x}}(t,\gamma)\right]' & \gamma \mathbf{I} \end{bmatrix}^{-1}. \tag{70}$$

In the above expressions we used the fact that $\hat{\mathbf{x}}(0,\gamma) = \mathbf{x}^*$ and $\hat{\boldsymbol{\eta}}(0,\gamma) = \boldsymbol{\eta}^*$ for all $\gamma \in [\gamma_{\min}, 1/\bar{c}]$ and that $A(t,\gamma)$ is well defined for all $(t,\gamma) \in D$. Note that compared to the standard case we introduced $\gamma_{\min}$ in (69) to cope with the fact that c is upper bounded by $c_{\max}$. The rest of the proof follows pretty much identical steps as in Proposition 2.4, page 108 of [2]. It is based on applying the norm operator on (69) and on the fact that for all $(t,\gamma) \in D$, $\|A(t,\gamma)\|$ is uniformly bounded, and where we also use the fact that $\|\mathbf{T}\boldsymbol{\eta} - \boldsymbol{\eta}^*\| \leq \|\boldsymbol{\eta} - \boldsymbol{\eta}^*\|$. ■

The relevant quantity to study convergence is $\tilde{\boldsymbol{\eta}}(\boldsymbol{\eta}, c) - \boldsymbol{\eta}^*$. The following result, which is the counterpart of Proposition 2.6 of [2], permits the calculation of an upper bound for this quantity that will be used for the convergence analysis.

Proposition A.4: Let Assumptions 2.1 and 2.2 hold, let $(\mathbf{x}^*, \boldsymbol{\eta}^*)$, be a local minimizer-Lagrange multipliers pair of (P_2), and let $\bar{c}$ and δ as in Proposition 6.1. For all $(\boldsymbol{\eta}, c)$ in the set D defined by (34), there holds

$$\tilde{\boldsymbol{\eta}}(\boldsymbol{\eta}, c) - \boldsymbol{\eta}^* = \int_0^1 N_c\left[\boldsymbol{\eta}^* + \zeta \mathbf{T}(\boldsymbol{\eta} - \boldsymbol{\eta}^*)\right] \mathbf{T}(\boldsymbol{\eta} - \boldsymbol{\eta}^*) d\zeta, \tag{71}$$

where for all $(\boldsymbol{\eta}, c) \in D$, the matrix N_c is given by

$$N_c(\boldsymbol{\eta}) = \mathbf{I} - c\nabla \tilde{\mathbf{h}}\left[\mathbf{x}(\boldsymbol{\eta}, c)\right]' \left\{\nabla^2_{\mathbf{xx}} \mathcal{L}_c\left[\mathbf{x}(\boldsymbol{\eta}, c), \boldsymbol{\eta}\right]\right\}^{-1} \nabla \tilde{\mathbf{h}}\left[\mathbf{x}(\boldsymbol{\eta}, c)\right]. \tag{72}$$

Proof: The proof is similar to the proof of Proposition 2.6, page 115, [2], with the difference that in (71) we

introduce the operator $\mathbf{T}$, due the definition of $\tilde{\boldsymbol{\eta}}(\boldsymbol{\eta},c)$ in (67). In addition, in the change of variable in (71), we made use of the fact that $\mathbf{T}\boldsymbol{\eta}^* = \boldsymbol{\eta}^*$. Note that for $(\boldsymbol{\eta},c) \in D$, by (68) we guarantee that $\nabla^2_{\mathbf{xx}}\mathcal{L}_c\left[\mathbf{x}(\boldsymbol{\eta},c),\boldsymbol{\eta}\right] \succ 0$, and therefore its inverse is well defined. ■

We are now ready to provide the proof of Theorem 6.1.

Proof of Theorem 6.1: The proof follows the same lines as the proof of Proposition 2.7 of [2]. The proof is based on the results introduced in Proposition 6.1 which states that the matrix $\nabla^2_{\mathbf{xx}}\mathcal{L}_c\left[\mathbf{x}(\boldsymbol{\eta},c),\boldsymbol{\eta}\right]$ is invertible for all $(\boldsymbol{\eta},c) \in D$ and therefore, the matrix $N_c(\boldsymbol{\eta})$ defined in (72) of Proposition A.4 is well defined.

Following the foot steps of the proof of Proposition 2.7 of [2], we have that the eigenvalues of $N_c(\boldsymbol{\eta}^*)$ can be expressed as

$$\boldsymbol{\sigma}_i(c) = \frac{\mathbf{e}_i}{\mathbf{e}_i + c},\ i = 1,\ldots,nN,$$

and from inequality $\bar{c} > \max\{-2\mathbf{e}_1,\ldots,-2\mathbf{e}_{nN}\}$, we have that

$$\max_{i=1,\ldots nN} |\boldsymbol{\sigma}_i(c)| < 1,\ \forall c \geq \bar{c}. \tag{73}$$

By Proposition 6.1, for all $(\boldsymbol{\eta},c) \in D$ we have that $\mathbf{x}(\boldsymbol{\eta},c)$ and $\tilde{\boldsymbol{\eta}}(\boldsymbol{\eta},c)$ satisfying (66)-(67), are continuously differentiable, and therefore for any $\varepsilon_1 > 0$ there exists a $\delta_1 \in (0,\delta]$ such that, for all $(\boldsymbol{\eta},c) \in D_1 = \{(\boldsymbol{\eta},c) \mid \|\mathbf{T}\boldsymbol{\eta} - \boldsymbol{\eta}^*\|/c < \delta_1, \bar{c} \leq c \leq c_{\max}\}$ we have

$$\|N_c(\boldsymbol{\eta})\| \leq \|N_c(\boldsymbol{\eta}^*)\| + \varepsilon_1 = \max_{i=1,\ldots nN} |\boldsymbol{\sigma}_i(c)| + \varepsilon_1,$$

where the matrix $N_c(\boldsymbol{\eta})$ was defined in (72).

Using (71) of Proposition A.4 together with the above inequality we have that for all $(\boldsymbol{\eta},c) \in D_1$

$$\begin{aligned}\|\tilde{\boldsymbol{\eta}}(\boldsymbol{\eta},c) - \boldsymbol{\eta}^*\| &\leq \int_0^1 \left\|N_c\left[\boldsymbol{\eta}^* + \zeta\mathbf{T}(\boldsymbol{\eta} - \boldsymbol{\eta}^*)\right]\right\| \|\mathbf{T}(\boldsymbol{\eta} - \boldsymbol{\eta}^*)\| d\zeta \leq \\ &\leq \left(\max_{i=1,\ldots nN} |\boldsymbol{\sigma}_i(c)| + \varepsilon_1\right) \|\mathbf{T}\| \|\boldsymbol{\eta} - \boldsymbol{\eta}^*\| \leq \left(\max_{i=1,\ldots nN} |\boldsymbol{\sigma}_i(c)| + \varepsilon_1\right) \|\boldsymbol{\eta} - \boldsymbol{\eta}^*\|,\end{aligned} \tag{74}$$

where the last inequality followed from the fact that $\mathbf{T}\boldsymbol{\eta}^* = \boldsymbol{\eta}^*$ and $\|\mathbf{T}\| = 1$. Keeping in mind that ε_1 can be chosen arbitrarily small, we have that there exists $\rho(\varepsilon_1,c) = \max_{i=1,\ldots nN} |\boldsymbol{\sigma}_i(c)| + \varepsilon_1$ so that $\rho(\varepsilon_1,c) \in (0,1)$ for all $(\boldsymbol{\eta},c) \in D_1$, and therefore

$$\|\tilde{\boldsymbol{\eta}}(\boldsymbol{\eta},c) - \boldsymbol{\eta}^*\| \leq \rho(\varepsilon_1,c)\|\boldsymbol{\eta} - \boldsymbol{\eta}^*\|.$$

From the above, together with (36) of Proposition 6.1 we have that $\boldsymbol{\eta}_k \to \boldsymbol{\eta}^*$ and $\mathbf{x}(\boldsymbol{\eta}_k,c_k) \to \mathbf{x}^*$, and the results for the rates of convergence follow from (74). ■