On the Convergence of a Distributed Augmented Lagrangian Method for Nonconvex Optimization

Nikolaos Chatzipanagiotis, *Student Member, IEEE*, and Michael M. Zavlanos, *Member, IEEE*

***Abstract*—In this paper, we propose a distributed algorithm for optimization problems that involve a separable, possibly nonconvex objective function subject to convex local constraints and linear coupling constraints. The method is based on the accelerated distributed augmented Lagrangians (ADAL) algorithm that was recently developed by the authors to address convex problems. Here, we extend this line of work in two ways. First, we establish convergence of the method to a local minimum of the problem, using assumptions that are common in the analysis of nonconvex optimization methods. To the best of our knowledge, this is the first work that shows convergence to local minima specifically for a distributed augmented Lagrangian (AL) method applied to nonconvex optimization problems; distributed AL methods are known to perform very well when used to solve convex problems. Second, we propose a more general and decentralized rule to select the stepsizes of the method. This improves on the authors' original ADAL method, where the stepsize selection used global information at initialization. Numerical results are included to verify the correctness and efficiency of the proposed distributed method.**

***Index Terms*—Augmented Lagrangian (AL), distributed optimization, nonconvex optimization.**

I. Introduction

MANY applications in areas as diverse as wireless communications, machine learning, artificial intelligence, power systems, computational biology, logistics, finance, and statistics involve very large datasets that are obtained, stored, and retrieved in a decentralized manner. Within these areas, a significant number of problems involving, e.g., cellular phone networks, sensor networks, multiagent robotics, power grids, and the Internet, also possess a network structure wherein processors, sensors, actuators, and controllers need to cooperate in a distributed fashion over geographically disparate locations, based only on local information and communication. The increasing size and complexity, and the local nature of information that is particular to these problems has created a need for efficient distributed computation methods.

In this paper, we are particularly interested in distributed optimization algorithms. Such methods have been used recently to address a wide range of modern day problems involving wired and wireless communication networks [1]–[3], multiagent robotic networks [4], [5], machine learning [6], power distribution systems [7], image processing [8], model predictive control [9], statistics [10], and logistics [11].

We propose a distributed algorithm to solve the following class of constrained optimization problems:

$$\begin{aligned} \min \quad & \sum_{i=1}^{N} f_i(\mathbf{x}_i) \\ \text{subject to} \quad & \sum_{i=1}^{N} \mathbf{A}_i \mathbf{x}_i = \mathbf{b}, \\ & \mathbf{x}_i \in \mathcal{X}_i, \quad i = 1, 2, \ldots, N \end{aligned} \tag{1}$$

where for every $i \in \mathcal{I} = \{1, 2, \ldots, N\}$, the function $f_i : \mathbb{R}^{n_i} \to \mathbb{R}$ is twice continuously differentiable, $\mathcal{X}_i \subseteq \mathbb{R}^{n_i}$ denotes a nonempty closed, convex subset of n_i-dimensional Euclidean space, and $\mathbf{A}_i$ is a matrix of dimension $m \times n_i$.

Problem (1) models situations where a set of N decision makers, henceforth referred to as agents, need to determine local decisions $\mathbf{x}_i \in \mathcal{X}_i$ that minimize a collection of local cost functions $f_i(\mathbf{x}_i)$, while respecting a set of affine constraints $\sum_{i=1}^{N} \mathbf{A}_i \mathbf{x}_i = \mathbf{b}$ that couple the local decisions between agents. In previous work [12], [13], we presented the *Accelerated Distributed Augmented Lagrangians* (ADAL) method to solve such problems in a distributed fashion, when the objective functions f_i are convex but not necessarily differentiable. ADAL is a primal-dual iterative scheme based on the augmented Lagrangian (AL) framework [14], [15]. In ADAL, every agent is assumed to know its local problem parameters f_i, $\mathbf{A}_i$, $\mathcal{X}_i$, and is also responsible for determining its own decision variables $\mathbf{x}_i$. Each iteration of ADAL consists of three steps. First, every agent solves a local convex optimization problem based on a separable approximation of the AL, that utilizes only locally available variables. Then, the agents update and communicate their primal variables to neighboring agents. Here, the communication neighbors of agent i are all those agents j that are coupled in the same constraints as i, i.e., the communication requirements between the agents are determined by the structure of the (static) coupling constraint set $\sum_{i=1}^{N} \mathbf{A}_i \mathbf{x}_i = \mathbf{b}$. Finally, in the last step, the dual variables are updated in a distributed fashion based on the new values of the primal variables; the Lagrange multiplier of the jth constraint is updated based on

Manuscript received April 22, 2016; revised July 13, 2016 and January 15, 2017; accepted January 19, 2017. Date of publication January 25, 2017; date of current version August 28, 2017. This work is supported in part by National Science Foundation CNS under Grant #1261828, and in part by Office of Naval Research under Grant #N000141410479. Recommended by Associate Editor A. Nedich.

The authors are with the Department of Mechanical Engineering and Materials Science, Duke University, Durham, NC 27708 USA (e-mail: n.chatzip@duke.edu; michael.zavlanos@duke.edu).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TAC.2017.2658438

communicated information from those agents whose decisions are coupled in this constraint, i.e., those i for which $[\mathbf{A}_i]_j \neq \mathbf{0}$. The computations at each step are performed in parallel. It was shown in [16] that ADAL has a worst-case $O(1/k)$ convergence rate, where k denotes the number of iterations. Moreover, a stochastic convergence framework for ADAL was established in [13] for convex constrained optimization problems that are subject to noise corruption and uncertainties.

In this paper, we extend ADAL in two ways. First, under assumptions that are common in the study of nonconvex optimization methods, we prove the convergence of ADAL to a local minimum of problem (1) when the local cost functions f_i are nonconvex. To the best of our knowledge, this is the first published work that formally establishes the convergence of a distributed augmented Lagrangian method (ALM) for nonconvex optimization problems. Second, we propose a way to select the stepsizes used in the algorithm that is more general compared to [12]. Specifically, it was shown in [12] that ADAL converges to the optimal solution of (1) if the stepsizes satisfy a certain condition that requires knowledge of the global structure of the constraint set at initialization. Here, we lift this requirement for global information and, instead, define m stepsizes associated with each one of the m coupling constraints in (1); each stepsize must adhere to a condition that requires only local information from the corresponding constraint. It is worth noting that these two contributions are independent from each other, meaning that the convergence of the nonconvex ADAL method can still be shown using the stepsizes from [12], and, similarly, convergence of the convex ADAL method can be shown using the stepsizes proposed in this paper.

A. Related Literature

The existing literature on distributed optimization methods mostly focuses on convex problems. The classic approach is that of *dual decomposition* and is based on Lagrangian duality theory [15], [17], [18]. Dual methods are simple and popular, however, they suffer from exceedingly slow convergence rates and require strict convexity of the objective function.

The main drawbacks of simple dual decomposition methods are alleviated by utilizing the AL framework, which has recently received considerable attention as a most efficient approach for distributed optimization in determistic settings; see, e.g., [6], [12], [19]–[23]. The ADAL method [12] considered in this paper belongs in this class of distributed AL algorithms, along with the *Alternating Directions Method of Multipliers* (ADMM) [6], and the *Diagonal Quadratic Approximation* [20] methods. A distributed AL algorithm similar to ADAL that solves deterministic convex problems of the form (1) has been proposed in [22]. The main difference between [22] and [12] lies in the stepsize choice; in [22] the stepsize is determined by the total number of agents in the problem, while in [12] the stepsize is determined by the number of agents coupled in the "most populated" constraint, which naturally leads to larger stepsizes in most cases. Another pertinent method can be found in [23] that also incorporates Bregman divergence factors into the local subproblems and at each iteration only a randomly selected subset of the agents perform updates. Finally, in [24], a similar algorithm to ADAL is proposed, which has a different dual update step, and also uses the additional assumptions that the matrices $\mathbf{A}_i$ are mutually near-orthogonal and have full column-rank.

Apart from AL methods, alternative algorithms for distributed convex optimization include Newton methods [25]–[27], projection-based approaches [28], [29], accelerated-gradient algorithms [30]–[32], online methods [33], [34], primal-dual perturbation approaches [35], reduced-communication algorithms [36], and even continuous-time approaches [37]. On the other hand, there exist only a few works on nonconvex distributed optimization methods, such as parallel variable distribution schemes [38]–[40], successive convex approximation algorithms [41], dual subgradient approaches [42], and fast-Lipschitz methods [43].

In relevant literature regarding distributed AL methods for nonconvex problems, it has been observed that the ADMM can converge in scenarios with nonconvex objective functions; see [44]–[47] for some examples. Nevertheless, the only published work that provides some theoretical justification for such observations is found in [48]. There, Hong and Luo propose a distributed AL method that provably converges to a stationary point of the nonconvex problem, for a certain class of problems and for sufficiently large values of the regularization parameter. The differences between [48] and the current paper are that, for the class of problems considered here, a different algorithm where the agents perform computations sequentially at each iterationis proposed in [48], while in our method the computations are performed in parallel. Moreover, Hong and Luo [48] prove convergence to a stationary point of the nonconvex problem provided that the regularization parameter is chosen large enough, while in this paper we prove convergence to a local minimum, under the additional assumption that the initialization point is sufficiently close to a locally optimal solution. Other relevant work includes [49], where Magnsson *et al.* provide conditions under which certain distributed AL schemes for nonconvex problems are guaranteed to converge, and also [50] where an elaborate distributed AL algorithm with modified gradients and Hessian approximations is proposed, similar in spirit to sequential quadratic programming methods.

The rest of this paper is organized as follows. In Section II, we discuss some essential facts regarding duality and the AL framework. We also provide a description of the ADAL method that utilizes the new local stepsizes, and discuss how it compares to the method using the global stepsizes presented in [12]. In Section III, we analyze the convergence of ADAL for problems of the form (1) under the local stepsize selection rule. Finally, Section IV contains numerical results that validate the effectiveness and efficiency of the proposed algorithm.

II. Preliminaries

We denote

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

where $\mathbf{x} = [\mathbf{x}_1^\top, \ldots, \mathbf{x}_N^\top]^\top \in \mathbb{R}^n$ with $n = \sum_{i=1}^N n_i$. Furthermore, we denote $\mathbf{A} = [\mathbf{A}_1 \ldots \mathbf{A}_N] \in \mathbb{R}^{m \times n}$. The constraint $\sum_{i=1}^N \mathbf{A}_i \mathbf{x}_i = \mathbf{b}$ of problem (1) takes on the form $\mathbf{A}\mathbf{x} = \mathbf{b}$. Associating Lagrange multipliers $\boldsymbol{\lambda} \in \mathbb{R}^m$ with that constraint, the Lagrange function is defined as

$$\begin{aligned} L(\mathbf{x}, \boldsymbol{\lambda}) &= F(\mathbf{x}) + \langle \boldsymbol{\lambda}, \mathbf{A}\mathbf{x} - \mathbf{b} \rangle \qquad (2) \\ &= \sum_{i=1}^N L_i(\mathbf{x}_i, \boldsymbol{\lambda}) - \langle \mathbf{b}, \boldsymbol{\lambda} \rangle \end{aligned}$$

where $L_i(\mathbf{x}_i, \boldsymbol{\lambda}) = f_i(\mathbf{x}_i) + \langle \boldsymbol{\lambda}, \mathbf{A}_i \mathbf{x}_i \rangle$, and $\langle \cdot, \cdot \rangle$ denotes the inner product. Then, the dual function is defined as

$$g(\boldsymbol{\lambda}) = \inf_{\mathbf{x} \in \mathcal{X}} L(\mathbf{x}, \boldsymbol{\lambda}) = \sum_{i=1}^N g_i(\boldsymbol{\lambda}) - \langle \mathbf{b}, \boldsymbol{\lambda} \rangle$$

where $\mathcal{X} = \mathcal{X}_1 \times \mathcal{X}_2 \cdots \times \mathcal{X}_N$, and

$$g_i(\boldsymbol{\lambda}) = \inf_{\mathbf{x}_i \in \mathcal{X}_i} \left[f_i(\mathbf{x}_i) + \langle \boldsymbol{\lambda}, \mathbf{A}_i \mathbf{x}_i \rangle \right].$$

The dual function is decomposable and this gives rise to various decomposition methods addressing the *dual problem*, which is defined by

$$\max_{\boldsymbol{\lambda} \in \mathbb{R}^m} \sum_{i=1}^N g_i(\boldsymbol{\lambda}) - \langle \mathbf{b}, \boldsymbol{\lambda} \rangle. \qquad (3)$$

Algorithm 1: Augmented Lagrangian Method (ALM).

```
Set $k = 1$ and define initial Lagrange multipliers $\boldsymbol{\lambda}^1$.
    1. For a fixed vector $\boldsymbol{\lambda}^k$, calculate $\hat{\mathbf{x}}^k$ as a solution of the problem:
           $\min_{\mathbf{x} \in \mathcal{X}} \Lambda_\rho(\mathbf{x}, \boldsymbol{\lambda}^k)$.    (4)
    2. If the constraints $\sum_{i=1}^N \mathbf{A}_i \hat{\mathbf{x}}_i^k = \mathbf{b}$ are satisfied, then stop (optimal solution found). Otherwise, set :
           $\boldsymbol{\lambda}^{k+1} = \boldsymbol{\lambda}^k + \rho \left( \sum_{i=1}^N \mathbf{A}_i \hat{\mathbf{x}}_i^k - \mathbf{b} \right)$,    (5)
       Increase $k$ by one and return to Step 1.
```

Dual methods suffer from well-documented disadvantages, the most notable ones being their exceedingly slow convergence rates and the requirement for strictly convex objective functions. These drawbacks can be alleviated by the AL framework [14], [15]. The AL associated with problem (1) is given by

$$\Lambda_\rho(\mathbf{x}, \boldsymbol{\lambda}) = F(\mathbf{x}) + \langle \boldsymbol{\lambda}, \mathbf{A}\mathbf{x} - \mathbf{b} \rangle + \frac{\rho}{2} \|\mathbf{A}\mathbf{x} - \mathbf{b}\|^2$$

where $\rho > 0$ is a penalty parameter. We recall the standard ALM, also referred to as the "Method of Multipliers" in the literature [14], [15], in Algorithm 1.

The convergence of the ALM is ensured when problem (3) has an optimal solution independently of the initialization. Under convexity assumptions and a constraint qualification condition, every accumulation point of the sequence $\{\mathbf{x}^k\}$ is an optimal solution of problem (1), cf. [14]. Furthermore, the ALM exhibits convergence properties also in a nonconvex setting, assuming that the functions $f_i, i = 1, \ldots, N$ are twice continuously differentiable and the strong second-order conditions of optimality are satisfied [14]. This fact combined with the known efficiency of distributed AL methods in convex settings provide a strong motivation to develop distributed nonconvex AL schemes, such as the one proposed here.

A. ADAL algorithm

The ADAL method is based on defining the *local augmented Lagrangian* function $\Lambda_\rho^i : \mathbb{R}^{n_i} \times \mathbb{R}^n \times \mathbb{R}^m \to \mathbb{R}$ for every agent $i \in \mathcal{I} = \{1, \ldots, N\}$ at each iteration k, according to

$$\begin{aligned} \Lambda_\rho^i\big(\mathbf{x}_i, \mathbf{A}\mathbf{x}^k, \boldsymbol{\lambda}^k\big) &= f_i(\mathbf{x}_i) + \big\langle \boldsymbol{\lambda}^k, \mathbf{A}_i \mathbf{x}_i \big\rangle \\ &\quad + \frac{\rho}{2} \|\mathbf{A}_i \mathbf{x}_i + \sum_{j \in \mathcal{I}}^{j \neq i} \mathbf{A}_j \mathbf{x}_j^k - \mathbf{b}\|^2 \qquad (6) \end{aligned}$$

where $\rho > 0$ is a scalar penalty parameter defined by the user. Each iteration of ADAL is comprised of three steps.

Algorithm 2: Accelerated Distributed Augmented Lagrangians (ADAL).

```
Set $k = 1$ and define initial Lagrange multipliers $\boldsymbol{\lambda}^1$ and initial primal variables $\mathbf{x}^1$.
    1. For every $i \in \mathcal{I}$, determine $\hat{\mathbf{x}}_i^k$ as the solution of the following problem:
           $\min_{\mathbf{x}_i \in \mathcal{X}_i} \Lambda_\rho^i\big(\mathbf{x}_i, \mathbf{A}\mathbf{x}^k, \boldsymbol{\lambda}^k\big)$.    (7)
    2. Set for every $i \in \mathcal{I}$
           $\mathbf{A}_i \mathbf{x}_i^{k+1} = \mathbf{A}_i \mathbf{x}_i^k + \mathbf{T}\Big(\mathbf{A}_i \hat{\mathbf{x}}_i^k - \mathbf{A}_i \mathbf{x}_i^k\Big)$.    (8)
    3. Set:
           $\boldsymbol{\lambda}^{k+1} = \boldsymbol{\lambda}^k + \rho \mathbf{T} \left( \sum_{i=1}^N \mathbf{A}_i \mathbf{x}_i^{k+1} - \mathbf{b} \right)$,    (9)
       increase $k$ by one and return to Step 1.
```

1) A minimization step of all the local ALs with respect to the primal variables.
2) An update step for the primal variables.
3) An update step for the dual variables. The computations at each step are performed in a parallel fashion, so that ADAL resembles a *Jacobi*-type algorithm; see [15] for more details on Jacobi and Gauss–Seidel-type algorithms. The ADAL method is summarized in Algorithm 2.

At the first step of each iteration, each agent minimizes its local AL subject to its local convex constraints. This computation step requires only local information. To see this, note that the variables $\mathbf{A}_j \mathbf{x}_j^k$, appearing in the penalty term of the local AL (6), correspond to the local primal variables of agent j that were communicated to agent i for the optimization of its local Lagrangian Λ_ρ^i. With respect to agent i, these are considered fixed parameters. The penalty term of each Λ_ρ^i can be equivalently expressed as

$$\|\mathbf{A}_i\mathbf{x}_i + \sum_{j\in\mathcal{I}}^{j\neq i} \mathbf{A}_j\mathbf{x}_j^k - \mathbf{b}\|^2$$

$$= \sum_{l=1}^{m}\left(\left[\mathbf{A}_i\mathbf{x}_i\right]_l + \sum_{j\in\mathcal{I}}^{j\neq i}\left[\mathbf{A}_j\mathbf{x}_j^k\right]_l - b_l\right)^2$$

where $\left[\mathbf{A}_i\mathbf{x}_i\right]_l$ denotes the lth entry of the vector $\mathbf{A}_i\mathbf{x}_i$. The above penalty term is involved only in the minimization computation (7). Hence, for those l such that $[\mathbf{A}_i]_l = \mathbf{0}$, the terms $\sum_{j\in\mathcal{I}}^{j\neq i}\left[\mathbf{A}_j\mathbf{x}_j^k\right]_l - \mathbf{b}_l$ are just constant terms in the minimization step, and can be excluded. Here, $[\mathbf{A}_i]_l$ denotes the lth row of $\mathbf{A}_i$ and $\mathbf{0}$ stands for a zero vector of proper dimension. This implies that subproblem i needs access only to the decisions $\left[\mathbf{A}_j\mathbf{x}_j^k\right]_l$ from all subproblems $j \neq i$ that are involved in the same constraints l as i. Moreover, regarding the term $\langle\boldsymbol{\lambda}, \mathbf{A}_i\mathbf{x}_i\rangle$ in (6), we have that $\langle\boldsymbol{\lambda}, \mathbf{A}_i\mathbf{x}_i\rangle = \sum_{j=1}^{m}\lambda_j[\mathbf{A}_i\mathbf{x}_i]_j$. Hence, we see that, in order to compute (7), each subproblem i needs access only to those λ_j for which $[\mathbf{A}_i]_j \neq \mathbf{0}$. Intuitively speaking, each agent needs access only to the information that is relevant to the constraints that this agent is involved in.

After the local optimization steps have been carried out, the second step consists of each agent updating its primal variables by taking a convex combination with the corresponding values from the previous iteration. This update depends on a vector of stepsizes $\boldsymbol{\tau} \in \mathbb{R}^m$, where each entry τ_j is the stepsize corresponding to constraint j. For notational purposes, we define the diagonal, square matrix $\mathbf{T}$ of dimension m according to

$$\mathbf{T} = \mathrm{diag}\left(\tau_1, \ldots, \tau_m\right) \tag{10}$$

so that the diagonal entries of $\mathbf{T}$ are the stepsizes for each constraint. To select the appropriate values for $\boldsymbol{\tau}$, we first need to define the *degree* of a constraint for problems of the form (1). Specifically, for each constraint $j = 1, \ldots, m$, let q_j denote the number of individual decision makers i associated with this constraint. That is, q_j is the number of all $i \in \mathcal{I}$ such that $[\mathbf{A}_i]_j \neq \mathbf{0}$. Then, to guarantee the convergence of ADAL, we need to select $\tau_j \in (0, \frac{1}{q_j})$, according to the analysis presented in Section III.

Note that, at the local update steps (8), each agent i does not update the primal variables $\mathbf{x}_i$, but rather the products $\mathbf{A}_i\mathbf{x}_i^k$. Using a more rigorous notation, we could define an auxiliary variable $\mathbf{y}_i^k = \mathbf{A}_i\mathbf{x}_i^k$, so that the update (8) takes the form $\mathbf{y}_i^{k+1} = \mathbf{y}_i^k + \mathbf{T}\Big(\mathbf{A}_i\hat{\mathbf{x}}_i^k - \mathbf{y}_i^k\Big)$. To avoid introducing additional notation, we have chosen not to introduce the variables $\mathbf{y}_i^k$ and, instead, we directly update the terms $\mathbf{A}_i\mathbf{x}_i^k$, slightly abusing notation.

The third and final step of each ADAL iteration consists of the dual update. This step is distributed by structure, since the Lagrange multiplier of the jth constraint is updated according to $\lambda_j^{k+1} = \lambda_j^k + \rho\tau_j\left(\sum_{i=1}^{N}\left[\mathbf{A}_i\mathbf{x}_i^{k+1}\right]_j - b_j\right)$, which implies that the udpdate of λ_j needs only information from those i for which $[\mathbf{A}_i]_j \neq \mathbf{0}$. We can define, without loss of generality, a set $\mathcal{M} \subseteq \{1, \ldots, m\}$ of agents that perform the dual updates, such that an agent $j \in \mathcal{M}$ is responsible for the update of the dual variables corresponding to a subset of the coupling constraint set $\mathbf{A}\mathbf{x} = \mathbf{b}$ (without overlapping agents). For example, if the cardinality of $\mathcal{M}$ is equal to the number of constraints m, then each agent $j \in \mathcal{M}$ is responsible for the update of the dual variable of the jth constraint. In practical settings, $\mathcal{M}$ can be a subset of $\mathcal{I}$, or it can be a separate set of agents, depending on the application.

Remark 1: In the ADAL method presented in [12], the second step of the algorithm has the form

$$\mathbf{x}_i^{k+1} = \mathbf{x}_i^k + \tau\left(\hat{\mathbf{x}}_i^k - \mathbf{x}_i^k\right)$$

where the stepsize τ is a scalar that must satisfy $\tau \in (0, \frac{1}{q})$, for $q = \max_{1\leq j\leq m} q_j$. Intuitively, q is the number of agents coupled in the "most populated" constraint of the problem. Obtaining the parameter q clearly requires global information of the structure of the constraint set at initialization, which may hinder the distributed nature of the algorithm. To remedy this problem, in this paper, we propose the update rule (8), where we update the products $\mathbf{A}_i\mathbf{x}_i^k \in \mathbb{R}^m$, instead of just the variables $\mathbf{x}_i^k \in \mathbb{R}^{n_i}$, using a vector stepsize $\mathbf{T} \in \mathbb{R}^{m\times m}$ (diagonal matrix for notational purposes) that can be locally determined. To see why (8) requires only local information, note that every agent i needs to know only the q_j's that correspond to the constraints that this agent is involved in. Analogous arguments hold for the dual update step (9), also.

III. Convergence of ADAL

In order to prove the convergence of ADAL to a local minimum of (1), we need the following four assumptions.

A1) The sets $\mathcal{X}_i \subseteq \mathbb{R}^{n_i}$, $i = 1, \ldots, N$ are nonempty, closed, and convex.

A2) The functions $f_i : \mathbb{R}^{n_i} \to \mathbb{R}$, $i \in \mathcal{I} = \{1, 2, \ldots, N\}$ are twice continuously differentiable on $\mathcal{X}_i$.

A3) The subproblems (7) are solvable.

A4) There exists a point $\mathbf{x}^*$ satisfying the strong second-order sufficient conditions of optimality for problem (1) with Lagrange multipliers $\boldsymbol{\lambda}^*$.

The assumptions (A1), (A2), and (A4) are common and are used in the convergence proof of the standard ALM for nonconvex optimization problems, cf. [14]. Assumption (A4) implies that there exist Lagrange multipliers $\boldsymbol{\lambda}^* \in \mathbb{R}^m$ that satisfy the first-order optimality conditions for problem (1) at the feasible point $\mathbf{x}^*$, provided that a constraint qualification condition is satisfied at $\mathbf{x}^*$, i.e.

$$\nabla F(\mathbf{x}^*) + \mathbf{A}^\top\boldsymbol{\lambda}^* \in \mathcal{N}_{\mathcal{X}}(\mathbf{x}^*)$$

where we recall that $\mathbf{x} = [\mathbf{x}_1^\top, \ldots, \mathbf{x}_N^\top]^\top \in \mathbb{R}^n$, $F(\mathbf{x}) = \sum_i f_i(\mathbf{x}_i)$, and $\mathbf{A} = [\mathbf{A}_1 \ldots \mathbf{A}_N] \in \mathbb{R}^{m\times n}$. Here, we use $\mathcal{N}_{\mathcal{X}}(\mathbf{x})$ to denote the normal cone to the set $\mathcal{X}$ at point $\mathbf{x}$ [14], i.e.,

$$\mathcal{N}_{\mathcal{X}}(\mathbf{x}) = \{\mathbf{h} \in \mathbb{R}^n : \langle\mathbf{h}, \mathbf{y} - \mathbf{x}\rangle \leq 0, \ \forall\, \mathbf{y} \in \mathcal{X}\}.$$

The strong second-order sufficient conditions of optimality for problem (1) at a point $\mathbf{x}^*$ imply that

$$\left\langle\mathbf{s}, \nabla^2 F(\mathbf{x}^*)\mathbf{s}\right\rangle > 0, \quad \text{for all } \mathbf{s} \neq \mathbf{0}, \text{ such that } \mathbf{A}\mathbf{s} = \mathbf{0}$$

c.f. [14], Lemma 4.32.

Assumption (A3) is satisfied if for every $i = 1, \dots, N$, either the set $\mathcal{X}_i$ is compact, or the function $f_i(\mathbf{x}_i) + \frac{\rho}{2}\|\mathbf{A}_i\mathbf{x}_i - \mathbf{c}\|^2$ is $\inf$-compact for any vector $\mathbf{c}$. The latter condition, means that the level sets of the function are compact sets, implying that the set $\{\mathbf{x}_i \in \mathcal{X}_i : f_i(\mathbf{x}_i) + \frac{\rho}{2}\|\mathbf{A}_i\mathbf{x}_i - \mathbf{c}\|^2 \leq \alpha\}$ is compact for any $\alpha \in \mathbb{R}$.

Define the *residual* $\mathbf{r}(\mathbf{x}) \in \mathbb{R}^m$ as the vector containing the amount of all constraint violations with respect to the primal variable $\mathbf{x}$, i.e.

$$\mathbf{r}(\mathbf{x}) = \sum_{i=1}^{N} \mathbf{A}_i\mathbf{x}_i - \mathbf{b}. \tag{11}$$

Define also the auxiliary variables

$$\hat{\boldsymbol{\lambda}}^k = \boldsymbol{\lambda}^k + \rho\mathbf{r}(\hat{\mathbf{x}}^k) \tag{12}$$

and

$$\bar{\boldsymbol{\lambda}}^k = \boldsymbol{\lambda}^k + \rho(\mathbf{I} - \mathbf{T})\mathbf{r}\big(\mathbf{x}^k\big) \tag{13}$$

where $\mathbf{I}$ is the identity matrix of size m.

The basic idea to show convergence of our method is to introduce the Lyapunov (merit) function

$$\phi(\mathbf{x}^k, \boldsymbol{\lambda}^k) = \sum_{i=1}^{N} \rho\big\|\mathbf{A}_i\mathbf{x}_i^k - \mathbf{A}_i\mathbf{x}_i^*\big\|^2_{\mathbf{T}^{-1}} + \frac{1}{\rho}\big\|\bar{\boldsymbol{\lambda}}^k - \boldsymbol{\lambda}^*\big\|^2_{\mathbf{T}^{-1}} \tag{14}$$

where we use the notation $\big\|\mathbf{x}\big\|_{\mathbf{M}} = \sqrt{\mathbf{x}^\top\mathbf{M}\mathbf{x}}$. We will show in Theorem 1 that this Lyapunov function is strictly decreasing during the execution of the ADAL algorithm (7)–(9), given that the stepsizes τ_j satisfy the condition $0 < \tau_j < 1/q_j$ for all $j = 1, \dots, m$. Then, in Theorem 2, we show that the strictly decreasing property of the Lyapunov function (14) implies the convergence of the primal and dual variables to their respective optimal values defined at a local minimum of problem (1).

We begin the proof by utilizing the first-order optimality conditions of all the subproblems (7) in order to derive some necessary inequalities.

Lemma 1: Assume (A1)–(A4). Then, the followingxbrk inequality holds:

$$\begin{aligned}\sum_i \Big(\nabla f_i(\mathbf{x}_i^*) - \nabla f_i(\hat{\mathbf{x}}_i^k)\Big)^\top \Big(\hat{\mathbf{x}}_i^k - \mathbf{x}_i^*\Big)& \\ + \frac{1}{\rho}\Big(\hat{\boldsymbol{\lambda}}^k - \boldsymbol{\lambda}^*\Big)^\top \Big(\boldsymbol{\lambda}^k - \hat{\boldsymbol{\lambda}}^k\Big)& \\ \geq \rho\sum_i \Big(\mathbf{A}_i\hat{\mathbf{x}}_i^k - \mathbf{A}_i\mathbf{x}_i^*\Big)^\top \Bigg(\sum_{j \neq i}\big(\mathbf{A}_j\mathbf{x}_j^k - \mathbf{A}_j\hat{\mathbf{x}}_j^k\big)\Bigg)&\end{aligned} \tag{15}$$

where $\boldsymbol{\lambda}^k$, $\hat{\boldsymbol{\lambda}}^k$, $\hat{\mathbf{x}}_i^k$, and $\mathbf{x}_j^k$ are calculated at iteration k.

Proof: The first-order optimality conditions for problem (7) imply the following inclusion for the minimizer $\hat{\mathbf{x}}_i^k$:

$$\begin{aligned}\mathbf{0} \in \nabla f_i(\hat{\mathbf{x}}_i^k) + \mathbf{A}_i^\top\boldsymbol{\lambda}^k + \rho\mathbf{A}_i^\top\Bigg(\mathbf{A}_i\hat{\mathbf{x}}_i^k + \sum_{j \neq i}\mathbf{A}_j\mathbf{x}_j^k - \mathbf{b}\Bigg)& \\ + \mathcal{N}_{\mathcal{X}_i}(\hat{\mathbf{x}}_i^k).&\end{aligned} \tag{16}$$

We infer that there exist normal elements $\mathbf{z}_i^k \in \mathcal{N}_{\mathcal{X}_i}(\hat{\mathbf{x}}_i^k)$, such that we can express (16) as follows:

$$\begin{aligned}\mathbf{0} = \nabla f_i(\hat{\mathbf{x}}_i^k) + \mathbf{A}_i^\top\boldsymbol{\lambda}^k + \rho\mathbf{A}_i^\top\Bigg(\mathbf{A}_i\hat{\mathbf{x}}_i^k + \sum_{j \neq i}\mathbf{A}_j\mathbf{x}_j^k - \mathbf{b}\Bigg)& \\ + \mathbf{z}_i^k.&\end{aligned} \tag{17}$$

Taking the inner product with $\mathbf{x}_i^* - \hat{\mathbf{x}}_i^k$ on both sides of this equation and using the definition of a normal cone, we obtain

$$\begin{aligned}&\Bigg\langle \nabla f_i(\hat{\mathbf{x}}_i^k) + \mathbf{A}_i^\top\boldsymbol{\lambda}^k \\ &\quad + \rho\mathbf{A}_i^\top\Bigg(\mathbf{A}_i\hat{\mathbf{x}}_i^k + \sum_{j \neq i}\mathbf{A}_j\mathbf{x}_j^k - \mathbf{b}\Bigg), \mathbf{x}_i^* - \hat{\mathbf{x}}_i^k\Bigg\rangle \\ &\quad = \langle -\mathbf{z}_i^k, \mathbf{x}_i^* - \hat{\mathbf{x}}_i^k\rangle \geq 0.\end{aligned} \tag{18}$$

Using the variables $\hat{\boldsymbol{\lambda}}^k$ defined in (12), we substitute $\boldsymbol{\lambda}^k$ in (18) and obtain

$$\begin{aligned}0 &\leq \Bigg\langle \nabla f_i(\hat{\mathbf{x}}_i^k) + \mathbf{A}_i^\top\Bigg[\hat{\boldsymbol{\lambda}}^k - \rho\Bigg(\sum_j \mathbf{A}_j\hat{\mathbf{x}}_j^k - \mathbf{b}\Bigg) \\ &\quad + \rho\Bigg(\mathbf{A}_i\hat{\mathbf{x}}_i^k + \sum_{j \neq i}\mathbf{A}_j\mathbf{x}_j^k - \mathbf{b}\Bigg)\Bigg], \mathbf{x}_i^* - \hat{\mathbf{x}}_i^k\rangle \\ &= \Bigg\langle \nabla f_i(\hat{\mathbf{x}}_i^k) + \mathbf{A}_i^\top\Bigg[\hat{\boldsymbol{\lambda}}^k \\ &\quad + \rho\Bigg(\sum_{j \neq i}\mathbf{A}_j\mathbf{x}_j^k - \sum_{j \neq i}\mathbf{A}_j\hat{\mathbf{x}}_j^k\Bigg)\Bigg], \mathbf{x}_i^* - \hat{\mathbf{x}}_i^k\Bigg\rangle.\end{aligned} \tag{19}$$

The assumption (A4) entails that the following first-order optimality conditions are satisfied at the point $(\mathbf{x}^*, \boldsymbol{\lambda}^*)$, i.e.

$$\mathbf{0} \in \nabla f_i(\mathbf{x}_i^*) + \mathbf{A}_i^\top\boldsymbol{\lambda}^* + \mathcal{N}_{\mathcal{X}_i}(\mathbf{x}_i^*) \quad \text{for all } i = 1, \dots, N. \tag{20}$$

After using the definition of the normal cone and taking the inner product with $\hat{\mathbf{x}}_i^k - \mathbf{x}_i^*$ on both sides of this equation (as before), we obtain the equivalent expression for the above inclusion

$$\Big\langle \nabla f_i(\mathbf{x}_i^*) + \mathbf{A}_i^\top\boldsymbol{\lambda}^*, \hat{\mathbf{x}}_i^k - \mathbf{x}_i^*\Big\rangle \geq 0, \quad \text{for all } i = 1, \dots, N. \tag{21}$$

Adding together (19) and (21), we obtain the following inequalities for all $i = 1, \dots, N$:

$$\begin{aligned}&\Big\langle \nabla f_i(\mathbf{x}_i^*) - \nabla f_i(\hat{\mathbf{x}}_i^k) + \mathbf{A}_i^\top(\boldsymbol{\lambda}^* - \hat{\boldsymbol{\lambda}}^k) \\ &\quad - \rho\mathbf{A}_i^\top\Bigg(\sum_{j \neq i}\mathbf{A}_j\mathbf{x}_j^k - \sum_{j \neq i}\mathbf{A}_j\hat{\mathbf{x}}_j^k\Bigg), \hat{\mathbf{x}}_i^k - \mathbf{x}_i^*\Big\rangle \geq 0.\end{aligned}$$

Adding the inequalities for all $i = 1, \dots, N$ and rearranging terms, we obtain

$$\begin{aligned}&\sum_i \Big(\nabla f_i(\mathbf{x}_i^*) - \nabla f_i(\hat{\mathbf{x}}_i^k)\Big)^\top \Big(\hat{\mathbf{x}}_i^k - \mathbf{x}_i^*\Big)\\ &\quad + \Big(\boldsymbol{\lambda}^* - \hat{\boldsymbol{\lambda}}^k\Big)^\top \bigg(\sum_i \Big(\mathbf{A}_i\hat{\mathbf{x}}_i^k - \mathbf{A}_i\mathbf{x}_i^*\Big)\bigg)\\ &\quad \geq \rho \sum_i \Big(\mathbf{A}_i\hat{\mathbf{x}}_i^k - \mathbf{A}_i\mathbf{x}_i^*\Big)^\top \bigg(\sum_{j\neq i} \Big(\mathbf{A}_j\mathbf{x}_j^k - \mathbf{A}_j\hat{\mathbf{x}}_j^k\Big)\bigg).\end{aligned}$$

Substituting $\sum_{i=1}^N \mathbf{A}_i\mathbf{x}_i^* = \mathbf{b}$ and $\sum_{i=1}^N \mathbf{A}_i\hat{\mathbf{x}}_i^k - \mathbf{b} = \frac{1}{\rho}(\hat{\boldsymbol{\lambda}}^k - \boldsymbol{\lambda}^k)$ from (12), we conclude that

$$\begin{aligned}&\sum_i \Big(\nabla f_i(\mathbf{x}_i^*) - \nabla f_i(\hat{\mathbf{x}}_i^k)\Big)^\top \Big(\hat{\mathbf{x}}_i^k - \mathbf{x}_i^*\Big)\\ &\quad + \frac{1}{\rho}\Big(\hat{\boldsymbol{\lambda}}^k - \boldsymbol{\lambda}^*\Big)^\top \Big(\boldsymbol{\lambda}^k - \hat{\boldsymbol{\lambda}}^k\Big)\\ &\quad \geq \rho \sum_i \Big(\mathbf{A}_i\hat{\mathbf{x}}_i^k - \mathbf{A}_i\mathbf{x}_i^*\Big)^\top \bigg(\sum_{j\neq i} \Big(\mathbf{A}_j\mathbf{x}_j^k - \mathbf{A}_j\hat{\mathbf{x}}_j^k\Big)\bigg)\end{aligned}$$

as required. ■

The following lemma is similar to Lemma 2 presented in [12]. The difference is that here the statement of the lemma includes also the gradient terms of the objective functions; in the convex case studied in [12], these terms are factored out due to the monotonicity property of the convex subdifferential. The proof of the lemma is given in the Appendix.

Lemma 2: Under assumptions (A1)–(A4), the following relation holds:

$$\begin{aligned}&\sum_i \Big(\nabla f_i(\mathbf{x}_i^*) - \nabla f_i(\hat{\mathbf{x}}_i^k)\Big)^\top \Big(\hat{\mathbf{x}}_i^k - \mathbf{x}_i^*\Big)\\ &\quad + \rho \sum_i \Big(\mathbf{A}_i\mathbf{x}_i{}^k - \mathbf{A}_i\mathbf{x}_i^*\Big)^\top \Big(\mathbf{A}_i\mathbf{x}_i^k - \mathbf{A}_i\hat{\mathbf{x}}_i^k\Big)\\ &\quad + \frac{1}{\rho}\Big(\boldsymbol{\lambda}^k - \boldsymbol{\lambda}^*\Big)^\top \Big(\boldsymbol{\lambda}^k - \hat{\boldsymbol{\lambda}}^k\Big) \qquad (22)\\ &\quad \geq \sum_i \rho\|\mathbf{A}_i(\mathbf{x}_i^k - \hat{\mathbf{x}}_i^k)\|^2 + \frac{1}{\rho}\|\hat{\boldsymbol{\lambda}}^k - \boldsymbol{\lambda}^k\|^2\\ &\quad + \Big(\hat{\boldsymbol{\lambda}}^k - \boldsymbol{\lambda}^k\Big)^\top \Big(\mathbf{r}\big(\mathbf{x}^k\big) - \mathbf{r}\big(\hat{\mathbf{x}}^k\big)\Big).\end{aligned}$$

In the next lemma, we obtain a modified version of (22) whose right-hand side is nonnegative.

Lemma 3: Under the assumptions (A1)–(A4), the following relation holds:

$$\begin{aligned}&\sum_i \Big(\nabla f_i(\mathbf{x}_i^*) - \nabla f_i(\hat{\mathbf{x}}_i^k)\Big)^\top \Big(\hat{\mathbf{x}}_i^k - \mathbf{x}_i^*\Big)\\ &\quad + \rho \sum_i \Big(\mathbf{A}_i\mathbf{x}_i{}^k - \mathbf{A}_i\mathbf{x}_i^*\Big)^\top \Big(\mathbf{A}_i\mathbf{x}_i^k - \mathbf{A}_i\hat{\mathbf{x}}_i^k\Big)\\ &\quad + \frac{1}{\rho}\Big(\bar{\boldsymbol{\lambda}}^k - \boldsymbol{\lambda}^*\Big)^\top \Big(\boldsymbol{\lambda}^k - \hat{\boldsymbol{\lambda}}^k\Big) \qquad (23)\\ &\quad \geq \rho \sum_i \|\mathbf{A}_i\Big(\mathbf{x}_i^k - \hat{\mathbf{x}}_i^k\Big)\|^2 + \rho\Big\|\mathbf{r}\Big(\hat{\mathbf{x}}^k\Big)\Big\|_{\mathbf{T}-\frac{1}{2}\mathbf{D}}^2\end{aligned}$$

where $\mathbf{D} = \operatorname{diag}\big(q_1\tau_1^2, \dots, q_m\tau_m^2\big)$, and the variable $\bar{\boldsymbol{\lambda}}^k$ is defined in (13).

Proof: The first term in the left-hand side of (22) that includes the gradients of the objective functions will not be altered in what follows, so we neglect it temporarily for simplicity of notation. Add the term

$$\begin{aligned}&\frac{1}{\rho}\Big(\rho(\mathbf{I} - \mathbf{T})\mathbf{r}\big(\mathbf{x}^k\big)\Big)^\top \Big(\boldsymbol{\lambda}^k - \hat{\boldsymbol{\lambda}}^k\Big)\\ &\quad = \rho\Big((\mathbf{I} - \mathbf{T})\mathbf{r}\big(\mathbf{x}^k\big)\Big)^\top \Big(-\mathbf{r}\big(\hat{\mathbf{x}}^k\big)\Big)\end{aligned}$$

to both sides of inequality (22). Recalling the definition of $\bar{\boldsymbol{\lambda}}^k$ from (13), we obtain

$$\begin{aligned}&\rho \sum_i \Big(\mathbf{A}_i\mathbf{x}_i{}^k - \mathbf{A}_i\mathbf{x}_i^*\Big)^\top \Big(\mathbf{A}_i\mathbf{x}_i^k - \mathbf{A}_i\hat{\mathbf{x}}_i^k\Big)\\ &\quad + \frac{1}{\rho}\Big(\bar{\boldsymbol{\lambda}}^k - \boldsymbol{\lambda}^*\Big)^\top \Big(\boldsymbol{\lambda}^k - \hat{\boldsymbol{\lambda}}^k\Big)\\ &\quad \geq \rho \sum_i \|\mathbf{A}_i\Big(\mathbf{x}_i^k - \hat{\mathbf{x}}_i^k\Big)\|^2 + \rho\|\mathbf{r}(\hat{\mathbf{x}}^k)\|^2\\ &\quad + \Big(\hat{\boldsymbol{\lambda}}^k - \boldsymbol{\lambda}^k\Big)^\top \Big(\mathbf{r}\big(\mathbf{x}^k\big) - \mathbf{r}\big(\hat{\mathbf{x}}^k\big)\Big) - \rho\Big((\mathbf{I}\\ &\quad - \mathbf{T})\mathbf{r}\big(\mathbf{x}^k\big)\Big)^\top \mathbf{r}\big(\hat{\mathbf{x}}^k\big). \qquad (24)\end{aligned}$$

Consider the term $\Big(\hat{\boldsymbol{\lambda}}^k - \boldsymbol{\lambda}^k\Big)^\top \Big(\mathbf{r}\big(\mathbf{x}^k\big) - \mathbf{r}\big(\hat{\mathbf{x}}^k\big)\Big) - \rho\Big((\mathbf{I} - \mathbf{T})\mathbf{r}\big(\mathbf{x}^k\big)\Big)^\top \mathbf{r}\big(\hat{\mathbf{x}}^k\big)$ in the right-hand side of (24). We manipulate it to obtain

$$\begin{aligned}&\Big(\hat{\boldsymbol{\lambda}}^k - \boldsymbol{\lambda}^k\Big)^\top \Big(\mathbf{r}\big(\mathbf{x}^k\big) - \mathbf{r}\big(\hat{\mathbf{x}}^k\big)\Big) - \rho\Big((\mathbf{I} - \mathbf{T})\mathbf{r}\big(\mathbf{x}^k\big)\Big)^\top \mathbf{r}\big(\hat{\mathbf{x}}^k\big)\\ &\quad = \rho\mathbf{r}\big(\hat{\mathbf{x}}^k\big)^\top \Big(\mathbf{r}\big(\mathbf{x}^k\big) - \mathbf{r}\big(\hat{\mathbf{x}}^k\big)\Big) - \rho\Big((\mathbf{I} - \mathbf{T})\mathbf{r}\big(\mathbf{x}^k\big)\Big)^\top \mathbf{r}\big(\hat{\mathbf{x}}^k\big)\\ &\quad = \rho\mathbf{r}\big(\hat{\mathbf{x}}^k\big)^\top \Big(\mathbf{r}\big(\mathbf{x}^k\big) - \mathbf{r}\big(\hat{\mathbf{x}}^k\big)\Big)\\ &\quad - \rho\Big((\mathbf{I} - \mathbf{T})\big[\mathbf{r}\big(\mathbf{x}^k\big) - \mathbf{r}\big(\hat{\mathbf{x}}^k\big) + \mathbf{r}\big(\hat{\mathbf{x}}^k\big)\big]\Big)^\top \mathbf{r}\big(\hat{\mathbf{x}}^k\big)\\ &\quad = \rho\Big(\mathbf{T}\,\mathbf{r}\big(\hat{\mathbf{x}}^k\big)\Big)^\top \Big(\mathbf{r}\big(\mathbf{x}^k\big) - \mathbf{r}\big(\hat{\mathbf{x}}^k\big)\Big) - \rho\mathbf{r}\big(\hat{\mathbf{x}}^k\big)^\top(\mathbf{I} - \mathbf{T})\mathbf{r}\big(\hat{\mathbf{x}}^k\big)\\ &\quad = \rho\Big(\mathbf{T}\,\mathbf{r}\big(\hat{\mathbf{x}}^k\big)\Big)^\top \Big(\sum_i \mathbf{A}_i\big(\mathbf{x}_i^k - \hat{\mathbf{x}}_i^k\big)\Big) - \rho\mathbf{r}\big(\hat{\mathbf{x}}^k\big)^\top(\mathbf{I}\\ &\quad - \mathbf{T})\mathbf{r}\big(\hat{\mathbf{x}}^k\big). \qquad (25)\end{aligned}$$

Substituting back in (24), we obtain

$$\begin{aligned}&\rho \sum_i \Big(\mathbf{A}_i\mathbf{x}_i{}^k - \mathbf{A}_i\mathbf{x}_i^*\Big)^\top \Big(\mathbf{A}_i\mathbf{x}_i^k - \mathbf{A}_i\hat{\mathbf{x}}_i^k\Big)\\ &\quad + \frac{1}{\rho}\Big(\bar{\boldsymbol{\lambda}}^k - \boldsymbol{\lambda}^*\Big)^\top \Big(\boldsymbol{\lambda}^k - \hat{\boldsymbol{\lambda}}^k\Big)\\ &\quad \geq \rho \sum_i \|\mathbf{A}_i(\mathbf{x}_i^k - \hat{\mathbf{x}}_i^k)\|^2 + \rho\mathbf{r}\big(\hat{\mathbf{x}}^k\big)^\top \mathbf{T}\,\mathbf{r}\big(\hat{\mathbf{x}}^k\big) \qquad (26)\\ &\quad + \rho\Big(\mathbf{T}\,\mathbf{r}\big(\hat{\mathbf{x}}^k\big)\Big)^\top \bigg(\sum_i \mathbf{A}_i\big(\mathbf{x}_i^k - \hat{\mathbf{x}}_i^k\big)\bigg).\end{aligned}$$

Using the basic inequality $\|a\|_2^2 + 2ab + \|b\|_2^2 \geq 0$ for any vectors a and b, each of the terms $\rho\Big(\mathbf{T}\,\mathbf{r}(\hat{\mathbf{x}}^k)\Big)^\top\Big(\mathbf{A}_i(\mathbf{x}_i^k - \hat{\mathbf{x}}_i^k)\Big)$ in the right-hand side of (26) can be bounded below by considering

$$\begin{aligned}
&\rho\Big(\mathbf{T}\,\mathbf{r}(\hat{\mathbf{x}}^k)\Big)^\top\Big(\mathbf{A}_i\mathbf{x}_i^k - \mathbf{A}_i\hat{\mathbf{x}}_i^k\Big)\\
&\quad= \rho\sum_{j=1}^{m}\Big(\tau_j\Big[\mathbf{r}\Big(\hat{\mathbf{x}}^k\Big)\Big]_j\Big)\Big(\big[\mathbf{A}_i(\mathbf{x}_i^k - \hat{\mathbf{x}}_i^k)\big]_j\Big)\\
&\quad\geq -\frac{\rho}{2}\sum_{j=1}^{m}\bigg(\Big[\mathbf{A}_i(\mathbf{x}_i^k - \hat{\mathbf{x}}_i^k)\Big]_j^2 + \tau_j^2\Big[\mathbf{r}\Big(\hat{\mathbf{x}}^k\Big)\Big]_j^2\bigg)
\end{aligned}$$

where $\big[\,\cdot\,\big]_j$ denotes the jth entry of a vector. Note, however, that some of the rows of $\mathbf{A}_i$ might be zero. If $[\mathbf{A}_i]_j = \mathbf{0}$, then it follows that $\Big[\mathbf{r}\Big(\hat{\mathbf{x}}^k\Big)\Big]_j\big[\mathbf{A}_i(\mathbf{x}_i^k - \hat{\mathbf{x}}_i^k)\big]_j = 0$. Hence, denoting the set of nonzero rows of $\mathbf{A}_i$ as $\mathcal{Q}_i$, i.e., $\mathcal{Q}_i = \{j = 1, \ldots, m : [\mathbf{A}_i]_j \neq \mathbf{0}\}$, we can obtain a tighter lower bound for each term $\rho\Big(\mathbf{T}\,\mathbf{r}(\hat{\mathbf{x}}^k)\Big)^\top\Big(\mathbf{A}_i\mathbf{x}_i^k - \mathbf{A}_i\hat{\mathbf{x}}_i^k\Big)$ as

$$\begin{aligned}
&\rho\Big(\mathbf{T}\,\mathbf{r}(\hat{\mathbf{x}}^k)\Big)^\top\Big(\mathbf{A}_i\mathbf{x}_i^k - \mathbf{A}_i\hat{\mathbf{x}}_i^k\Big)\\
&\quad\geq -\frac{\rho}{2}\sum_{j\in\mathcal{Q}_i}\bigg(\Big[\mathbf{A}_i(\mathbf{x}_i^k - \hat{\mathbf{x}}_i^k)\Big]_j^2 + \tau_j^2\Big[\mathbf{r}\Big(\hat{\mathbf{x}}^k\Big)\Big]_j^2\bigg). \qquad (27)
\end{aligned}$$

Now, recall that q_j denotes the number of nonzero blocks $[\mathbf{A}_i]_j$ over all $i = 1, \ldots, N$, in other words, q_j is the number of decision makers i that are involved in the constraint j. Then, summing inequality (27) over all i, we observe that each term $\tau_j^2\Big[\mathbf{r}\Big(\hat{\mathbf{x}}^k\Big)\Big]_j^2$ is included in the summation at most q_j times.

This observation leads us to the bound

$$\begin{aligned}
&\rho\sum_i\Big(\mathbf{T}\,\mathbf{r}(\hat{\mathbf{x}}^k)\Big)^\top\Big(\mathbf{A}_i\mathbf{x}_i^k - \mathbf{A}_i\hat{\mathbf{x}}_i^k\Big)\\
&\quad\geq -\frac{\rho}{2}\bigg(\sum_i\|\mathbf{A}_i(\mathbf{x}_i^k - \hat{\mathbf{x}}_i^k)\|^2 + \sum_{j=1}^{m} q_j\tau_j^2\Big[\mathbf{r}\Big(\hat{\mathbf{x}}^k\Big)\Big]_j^2\bigg)
\end{aligned}$$

or, equivalently

$$\begin{aligned}
&\rho\sum_i\Big(\mathbf{T}\,\mathbf{r}(\hat{\mathbf{x}}^k)\Big)^\top\Big(\mathbf{A}_i\mathbf{x}_i^k - \mathbf{A}_i\hat{\mathbf{x}}_i^k\Big)\\
&\quad\geq -\frac{\rho}{2}\sum_i\|\mathbf{A}_i(\mathbf{x}_i^k - \hat{\mathbf{x}}_i^k)\|^2 - \frac{\rho}{2}\mathbf{r}(\hat{\mathbf{x}}^k)^\top\mathbf{D}\mathbf{r}(\hat{\mathbf{x}}^k) \qquad (28)
\end{aligned}$$

where $\mathbf{D} = \mathrm{diag}\big(q_1\tau_1^2, \ldots, q_m\tau_m^2\big)$. Finally, we substitute (28) into (26) to obtain

$$\begin{aligned}
&\rho\sum_i\Big(\mathbf{A}_i\mathbf{x}_i{}^k - \mathbf{A}_i\mathbf{x}_i^*\Big)^\top\Big(\mathbf{A}_i\mathbf{x}_i^k - \mathbf{A}_i\hat{\mathbf{x}}_i^k\Big)\\
&\quad+ \frac{1}{\rho}\Big(\bar{\boldsymbol{\lambda}}^k - \boldsymbol{\lambda}^*\Big)^\top\Big(\boldsymbol{\lambda}^k - \hat{\boldsymbol{\lambda}}^k\Big)\\
&\quad\geq \frac{\rho}{2}\sum_i\|\mathbf{A}_i\Big(\mathbf{x}_i^k - \hat{\mathbf{x}}_i^k\Big)\|^2 + \rho\big\|\mathbf{r}(\hat{\mathbf{x}}^k)\big\|^2_{\mathbf{T}-\frac{1}{2}\mathbf{D}}.
\end{aligned}$$

After reinstating the gradient terms that we have neglected thus far, we obtain the required result. ■

Next, our goal is to find a lower bound for the gradient terms appearing in (23). To do so, let $\mathbf{C}$ denote any diagonal matrix with strictly positive diagonal entries, and consider the function

$$G_\rho(\mathbf{x}) = F(\mathbf{x}) + \frac{\rho}{2}\mathbf{x}^\top\mathbf{A}^\top\mathbf{C}\mathbf{A}\mathbf{x} \qquad (29)$$

where we recall that $\mathbf{x} = [\mathbf{x}_1^\top, \ldots, \mathbf{x}_N^\top]^\top \in \mathbb{R}^n$, $F(\mathbf{x}) = \sum_i f_i(\mathbf{x}_i)$, and $\mathbf{A} = [\mathbf{A}_1 \ldots \mathbf{A}_N] \in \mathbb{R}^{m\times n}$. In the next lemmas, we will make use of the fact that, for sufficiently large ρ, the function $G_\rho(\mathbf{x})$ is strongly convex in a neighborhood around the optimal solution $\mathbf{x}^*$ of (1). For this, we will make use of the following result.

Lemma 4 ([14], Lemma 4.28): Assume that a symmetric matrix $\mathbf{Q}$ of dimension n and a matrix $\mathbf{B}$ of dimension $m\times n$ are such that

$$\big\langle\mathbf{x}, \mathbf{Q}\mathbf{x}\big\rangle > 0, \quad \text{for all } \mathbf{x} \neq \mathbf{0} \ \text{ such that } \ \mathbf{B}\mathbf{x} = \mathbf{0}.$$

Then, there exists ρ_0, such that for all $\rho > \rho_0$ the matrix $\mathbf{Q} + \rho\mathbf{B}^\top\mathbf{B}$ is positive definite.

Using Lemma 4, we can obtain an important relation involving the gradient terms that appear in (23).

Lemma 5: Assume (A1)–(A4). Then, for any diagonal matrix $\mathbf{C}$ with strictly positive diagonal entries, there exists some $\kappa > 0$, such that the following relation holds:

$$\begin{aligned}
&\rho\sum_i\Big(\mathbf{A}_i\mathbf{x}_i{}^k - \mathbf{A}_i\mathbf{x}_i^*\Big)^\top\Big(\mathbf{A}_i\mathbf{x}_i^k - \mathbf{A}_i\hat{\mathbf{x}}_i^k\Big)\\
&\quad+ \frac{1}{\rho}\Big(\bar{\boldsymbol{\lambda}}^k - \boldsymbol{\lambda}^*\Big)^\top\Big(\boldsymbol{\lambda}^k - \hat{\boldsymbol{\lambda}}^k\Big) \qquad (30)\\
&\quad\geq \frac{\rho}{2}\sum_i\|\mathbf{A}_i(\mathbf{x}_i^k - \hat{\mathbf{x}}_i^k)\|^2 + \rho\big\|\mathbf{r}(\hat{\mathbf{x}}^k)\big\|^2_{\mathbf{T}-\frac{1}{2}\mathbf{D}-\mathbf{C}}\\
&\quad+ \kappa\|\hat{\mathbf{x}}^k - \mathbf{x}^*\|^2
\end{aligned}$$

provided that ρ is sufficiently large, and that for all iterations k the terms $\boldsymbol{\lambda}^k + \rho\sum_{j\neq i}(\mathbf{A}_j\mathbf{x}_j^k - \mathbf{A}_j\mathbf{x}_j^*)$ are sufficiently close to $\boldsymbol{\lambda}^*$ for all $i = 1, \ldots, N$.

Proof: From assumption (A4), we have that there exists a point $\mathbf{x}^*$ satisfying the strong second-order sufficient conditions of optimality for problem (1). These conditions imply that

$$\big\langle\mathbf{s}, \nabla^2F(\mathbf{x}^*)\mathbf{s}\big\rangle > 0, \quad \text{for all } \mathbf{s} \neq \mathbf{0}, \ \text{such that } \mathbf{A}\mathbf{s} = \mathbf{0}.$$

Now, combine this with the result of Lemma 4 for

$$\mathbf{Q} = \nabla^2F(\mathbf{x}^*), \ \text{ and } \ \mathbf{B} = \mathbf{C}^{1/2}\mathbf{A}.$$

It follows that there exists ρ_0, such that for all $\rho > \rho_0$ the matrix $\nabla^2F(\mathbf{x}^*) + \rho\mathbf{A}^\top\mathbf{C}\mathbf{A}$ is positive definite with some modulus $\kappa_0 > 0$. Moreover, from assumption (A2) the matrix $\nabla^2F(\mathbf{x}) + \rho\mathbf{A}^\top\mathbf{C}\mathbf{A}$ (note that this matrix is defined w.r.t. $\mathbf{x}$ instead of $\mathbf{x}^*$) is also continuous, hence there exists sufficiently large ρ such that, for all $\mathbf{x} \in \mathcal{X}$ sufficiently close to $\mathbf{x}^*$, i.e., for $\|\mathbf{x} - \mathbf{x}^*\| \leq \beta$, all the eigenvalues of $\nabla^2F(\mathbf{x}) + \rho\mathbf{A}^\top\mathbf{C}\mathbf{A}$ lie above some $\kappa > 0$. To see this, observe that from Schwarz's theorem [51] we have that the continuous differentiability assumption (A2) which means that the Hessian matrix $H(\mathbf{x}) =$

$\nabla^2 F(\mathbf{x})$ is symmetric within $\mathcal{X}$. According to the eigenvalue perturbation theory [52], the symmetry of the Hessian entails that for a perturbation δH of the matrix H, the perturbation $\delta\epsilon$ of its smallest eigenvalue ϵ is bounded by δH, i.e., $|\delta\epsilon| \leq \|\delta H\|$. By the continuity of the Hessian, we infer that there exists some neighborhood $\|\mathbf{x}-\mathbf{x}^*\| \leq \beta$ around $\mathbf{x}^*$ such that $|\delta\epsilon| < \kappa_0$, which in turn means that within this neighborhood the matrix $\nabla^2 F(\mathbf{x}) + \rho\mathbf{A}^\top\mathbf{C}\mathbf{A}$ remains positive definite with a modulus at least $\kappa = \kappa_0 - |\delta\epsilon| > 0$.

Since the positive definite matrix $\nabla^2 F(\mathbf{x}) + \rho\mathbf{A}^\top\mathbf{C}\mathbf{A}$ is the Hessian of the function $G_\rho(\mathbf{x}) = F(\mathbf{x}) + \frac{\rho}{2}\mathbf{x}^\top\mathbf{A}^\top\mathbf{C}\mathbf{A}\mathbf{x}$ and $\mathcal{X}$ is a convex closed set, we infer that, for sufficiently large ρ, there exists some β, such that the function $G_\rho(\mathbf{x})$ is strongly convex with modulus κ for every $\mathbf{x}$ belonging in the set $\{\mathbf{x} \in \mathcal{X} : \|\mathbf{x}-\mathbf{x}^*\| \leq \beta\}$. From the definition of strongly convex functions, we get that the following holds for all $\mathbf{x}$ that are sufficiently close to $\mathbf{x}^*$:

$$\Big(\nabla G_\rho(\mathbf{x}) - \nabla G_\rho(\mathbf{x}^*)\Big)^\top\Big(\mathbf{x}-\mathbf{x}^*\Big) \;\geq\; \kappa\|\mathbf{x}-\mathbf{x}^*\|^2.$$

For the term $(\nabla G_\rho(\mathbf{x}) - \nabla G_\rho(\mathbf{x}^*))^\top(\mathbf{x}-\mathbf{x}^*)$, we have

$$\begin{aligned}
&\Big(\nabla G_\rho(\mathbf{x}) - \nabla G_\rho(\mathbf{x}^*)\Big)^\top\Big(\mathbf{x}-\mathbf{x}^*\Big)\\
&\quad= \Big(\nabla F(\mathbf{x}) - \nabla F(\mathbf{x}^*) + \rho\mathbf{A}^\top\mathbf{C}\mathbf{A}(\mathbf{x}-\mathbf{x}^*)\Big)^\top\Big(\mathbf{x}-\mathbf{x}^*\Big)\\
&\quad= \Big(\nabla F(\mathbf{x}) - \nabla F(\mathbf{x}^*) + \rho\mathbf{A}^\top\mathbf{C}(\mathbf{A}\mathbf{x}-\mathbf{b})\Big)^\top\Big(\mathbf{x}-\mathbf{x}^*\Big)\\
&\quad= \sum_i \Big(\nabla f_i(\mathbf{x}_i) - \nabla f_i(\mathbf{x}_i^*)\Big)^\top\Big(\mathbf{x}_i-\mathbf{x}_i^*\Big) + \rho\mathbf{r}(\mathbf{x})^\top\mathbf{C}\mathbf{r}(\mathbf{x})
\end{aligned}$$

where we have used the fact that $\mathbf{A}\mathbf{x}^* = \mathbf{b}$, and $\mathbf{r}(\mathbf{x}) = \mathbf{A}\mathbf{x}-\mathbf{b}$. It follows that

$$\begin{aligned}
&\sum_i \Big(\nabla f_i(\mathbf{x}_i) - \nabla f_i(\mathbf{x}_i^*)\Big)^\top\Big(\mathbf{x}_i-\mathbf{x}_i^*\Big) + \rho\mathbf{r}(\mathbf{x})^\top\mathbf{C}\mathbf{r}(\mathbf{x})\\
&\qquad\geq\; \kappa\|\mathbf{x}-\mathbf{x}^*\|^2.
\end{aligned}\tag{31}$$

Now, substitute $\mathbf{x} = \hat{\mathbf{x}}^k$ in (31), and add it to (23). We get the following relation:

$$\begin{aligned}
&\rho\sum_i \Big(\mathbf{A}_i{\mathbf{x}_i}^k - \mathbf{A}_i\mathbf{x}_i^*\Big)^\top\Big(\mathbf{A}_i\mathbf{x}_i^k - \mathbf{A}_i\hat{\mathbf{x}}_i^k\Big)\\
&\quad+ \frac{1}{\rho}\Big(\bar{\boldsymbol{\lambda}}^k - \boldsymbol{\lambda}^*\Big)^\top\Big(\boldsymbol{\lambda}^k - \hat{\boldsymbol{\lambda}}^k\Big)\\
&\quad\geq \frac{\rho}{2}\sum_i \|\mathbf{A}_i(\mathbf{x}_i^k - \hat{\mathbf{x}}_i^k)\|^2 + \rho\mathbf{r}\big(\hat{\mathbf{x}}^k\big)^\top\\
&\quad\times\Big(\mathbf{T} - \frac{1}{2}\mathbf{D} - \mathbf{C}\Big)\mathbf{r}\big(\hat{\mathbf{x}}^k\big) + \kappa\|\hat{\mathbf{x}}^k - \mathbf{x}^*\|^2
\end{aligned}$$

which is the required result.

Note that, in order to substitute $\mathbf{x} = \hat{\mathbf{x}}^k$ in (31), it is necessary that the $\hat{\mathbf{x}}^k$ are sufficiently close to $\mathbf{x}^*$ at iteration k, i.e., that they belong to the set $\{\hat{\mathbf{x}} \in \mathcal{X} : \|\hat{\mathbf{x}}-\mathbf{x}^*\| \leq \beta\}$. To see when this condition holds, note that the local AL for each i can be expressed as

$$\begin{aligned}
\Lambda_\rho^i\big(\mathbf{x}_i, \mathbf{A}\mathbf{x}^k, \boldsymbol{\lambda}^k\big) &= f_i(\mathbf{x}_i)\\
&+ \Big\langle \boldsymbol{\lambda}^k + \rho\sum_{j\neq i}\Big(\mathbf{A}_j\mathbf{x}_j^k - \mathbf{A}_j\mathbf{x}_j^*\Big), \mathbf{A}_i\mathbf{x}_i \Big\rangle\\
&+ \frac{\rho}{2}\|\mathbf{A}_i\mathbf{x}_i + \sum_{j\neq i}\mathbf{A}_j\mathbf{x}_j^* - \mathbf{b}\|^2 + \frac{\rho}{2}\|\\
&\times \sum_{j\neq i}\Big(\mathbf{A}_j\mathbf{x}_j^k - \mathbf{A}_j\mathbf{x}_j^*\Big)\|^2\\
&+ \rho\Big\langle \sum_{j\neq i}\Big(\mathbf{A}_j\mathbf{x}_j^k - \mathbf{A}_j\mathbf{x}_j^*\Big), \sum_{j\neq i}\mathbf{A}_j\mathbf{x}_j^* - \mathbf{b} \Big\rangle
\end{aligned}\tag{32}$$

where we have added the zero terms $\sum_{j\neq i}\mathbf{A}_j\mathbf{x}_j^* - \sum_{j\neq i}\mathbf{A}_j\mathbf{x}_j^*$ in the penalty term of the AL, and expanded it. The last two terms can be disregarded when minimizing with respect to $\mathbf{x}_i$. Recalling a well-known result on the sensitivity analysis of ALs, c.f. [14], [53], [54], we have that, given assumptions (A1)–(A4), if ρ is sufficiently large and the terms $\xi_i = \boldsymbol{\lambda}^k + \rho\sum_{j\neq i}(\mathbf{A}_j\mathbf{x}_j^k - \mathbf{A}_j\mathbf{x}_j^*)$ are sufficiently close to $\boldsymbol{\lambda}^*$ for all $i = 1,\ldots,N$, then $\sup_{\hat{\mathbf{x}}_i}\|\hat{\mathbf{x}}_i - \mathbf{x}_i^*\| = O(\|\xi_i - \boldsymbol{\lambda}^*\|)$ holds, i.e., $\hat{\mathbf{x}}_i^k$ will be sufficiently close to $\mathbf{x}_i^*$, as required for (31) to hold. ■

We are now ready to prove the key result pertaining to the convergence of our method. We will show that the function ϕ defined in (14) is a strictly decreasing Lyapunov function for ADAL. The results from Lemmas 3 and 5 will help us characterize the decrease of ϕ at each iteration.

Theorem 1: Assume (A1)–(A4). Assume also that ρ is sufficiently large, and that the initial iterates $\mathbf{x}^1, \boldsymbol{\lambda}^1$ are chosen such that $\phi(\mathbf{x}^1, \boldsymbol{\lambda}^1)$ is sufficiently small. If the ADAL method uses stepsizes τ_j satisfying

$$0 < \tau_j < \frac{1}{q_j},\ \forall\, j = 1,\ldots,m$$

then the sequence $\{\phi(\mathbf{x}^k, \boldsymbol{\lambda}^k)\}$, with $\phi(\mathbf{x}^k, \boldsymbol{\lambda}^k)$ defined in (14), is strictly decreasing.

Proof: First, we show that the dual update step (9) in the ADAL method results in the following update rule for the variables $\bar{\boldsymbol{\lambda}}^k$, which are defined in (13):

$$\bar{\boldsymbol{\lambda}}^{k+1} = \bar{\boldsymbol{\lambda}}^k + \rho\mathbf{T}\mathbf{r}\big(\hat{\mathbf{x}}^k\big)\tag{33}$$

Indeed

$$\begin{aligned}
\boldsymbol{\lambda}^{k+1} &= \boldsymbol{\lambda}^k + \rho\mathbf{T}\mathbf{r}(\mathbf{x}^{k+1})\\
\boldsymbol{\lambda}^{k+1} + \rho\mathbf{r}(\mathbf{x}^{k+1}) &= \boldsymbol{\lambda}^k + \rho\mathbf{T}\mathbf{r}(\mathbf{x}^{k+1}) + \rho\mathbf{r}(\mathbf{x}^{k+1})\\
\boldsymbol{\lambda}^{k+1} + \rho(\mathbf{I}-\mathbf{T})\mathbf{r}(\mathbf{x}^{k+1}) &= \boldsymbol{\lambda}^k + \rho\mathbf{r}(\mathbf{x}^{k+1})\\
\bar{\boldsymbol{\lambda}}^{k+1} &= \boldsymbol{\lambda}^k + \rho(\mathbf{I}-\mathbf{T})\mathbf{r}\big(\mathbf{x}^k\big) + \rho\mathbf{T}\mathbf{r}\big(\hat{\mathbf{x}}^k\big)\\
\bar{\boldsymbol{\lambda}}^{k+1} &= \bar{\boldsymbol{\lambda}}^k + \rho\mathbf{T}\mathbf{r}\big(\hat{\mathbf{x}}^k\big)
\end{aligned}$$

as required.

We define the progress at each iteration k of the ADAL method as

$$\theta_k(\mathbf{T}) = \phi(\mathbf{x}^k, \boldsymbol{\lambda}^k) - \phi(\mathbf{x}^{k+1}, \boldsymbol{\lambda}^{k+1}).$$

We substitute $\bar{\boldsymbol{\lambda}}^k$ in the formula for calculating the function ϕ and use relation (33). The progress $\theta_k(\mathbf{T})$ can be evaluated as follows:

$$\begin{aligned}\theta_k(\mathbf{T}) &= \sum_{i=1}^{N} \rho \big\| \mathbf{A}_i \big(\mathbf{x}_i^k - \mathbf{x}_i^*\big) \big\|^2_{\mathbf{T}^{-1}} + \frac{1}{\rho} \big\| \bar{\boldsymbol{\lambda}}^k - \boldsymbol{\lambda}^* \big\|^2_{\mathbf{T}^{-1}} \\ &\quad - \sum_{i=1}^{N} \rho \big\| \mathbf{A}_i \big(\mathbf{x}_i^{k+1} - \mathbf{x}_i^*\big) \big\|^2_{\mathbf{T}^{-1}} - \frac{1}{\rho} \big\| \bar{\boldsymbol{\lambda}}^{k+1} - \boldsymbol{\lambda}^* \big\|^2_{\mathbf{T}^{-1}}.\end{aligned} \tag{34}$$

First, consider the term $\big\|\mathbf{A}_i(\mathbf{x}_i^{k+1} - \mathbf{x}_i^*)\big\|^2_{\mathbf{T}^{-1}}$. We have that

$$\begin{aligned}&\rho \big\| \mathbf{A}_i (\mathbf{x}_i^{k+1} - \mathbf{x}_i^*) \big\|^2_{\mathbf{T}^{-1}} \\ &\quad = \rho \Big(\mathbf{A}_i \mathbf{x}_i^{k+1} - \mathbf{A}_i \mathbf{x}_i^* \Big)^\top \mathbf{T}^{-1} \Big(\mathbf{A}_i \mathbf{x}_i^{k+1} - \mathbf{A}_i \mathbf{x}_i^* \Big) \\ &\quad = \rho \Big(\mathbf{A}_i \mathbf{x}_i^{k} - \mathbf{A}_i \mathbf{x}_i^* \Big)^\top \mathbf{T}^{-1} \Big(\mathbf{A}_i \mathbf{x}_i^{k} - \mathbf{A}_i \mathbf{x}_i^* \Big) \\ &\quad + \rho \Big(\mathbf{T}\big(\mathbf{A}_i \hat{\mathbf{x}}_i^k - \mathbf{A}_i \mathbf{x}_i^k\big) \Big)^\top \mathbf{T}^{-1} \Big(\mathbf{T}\big(\mathbf{A}_i \hat{\mathbf{x}}_i^k - \mathbf{A}_i \mathbf{x}_i^k\big) \Big) \\ &\quad + 2\rho \Big(\mathbf{A}_i \mathbf{x}_i^{k} - \mathbf{A}_i \mathbf{x}_i^* \Big)^\top \mathbf{T}^{-1} \Big(\mathbf{T}\big(\mathbf{A}_i \hat{\mathbf{x}}_i^k - \mathbf{A}_i \mathbf{x}_i^k\big) \Big)\end{aligned}$$

where we substituted $\mathbf{A}_i \mathbf{x}_i^{k+1} = \mathbf{A}_i \mathbf{x}_i^k + \mathbf{T}(\mathbf{A}_i \hat{\mathbf{x}}_i^k - \mathbf{A}_i \mathbf{x}_i^k)$ from (8) and expanded the terms. The last equation in the above can be written as

$$\begin{aligned}&\rho \big\| \mathbf{A}_i (\mathbf{x}_i^{k+1} - \mathbf{x}_i^*) \big\|^2_{\mathbf{T}^{-1}} \\ &\quad = \rho \big\| \mathbf{A}_i (\mathbf{x}_i^{k} - \mathbf{x}_i^*) \big\|^2_{\mathbf{T}^{-1}} + \rho \big\| \mathbf{A}_i \hat{\mathbf{x}}_i^k - \mathbf{A}_i \mathbf{x}_i^k \big\|^2_{\mathbf{T}} \\ &\quad + 2\rho \Big(\mathbf{A}_i \mathbf{x}_i^{k} - \mathbf{A}_i \mathbf{x}_i^* \Big)^\top \Big(\mathbf{A}_i \hat{\mathbf{x}}_i^k - \mathbf{A}_i \mathbf{x}_i^k \Big).\end{aligned} \tag{35}$$

Similarly, for the term $\big\|\bar{\boldsymbol{\lambda}}^{k+1} - \boldsymbol{\lambda}^*\big\|^2_{\mathbf{T}^{-1}}$, we have

$$\begin{aligned}&\frac{1}{\rho} \big\| \bar{\boldsymbol{\lambda}}^{k+1} - \boldsymbol{\lambda}^* \big\|^2_{\mathbf{T}^{-1}} = \\ &\quad = \frac{1}{\rho} \Big(\bar{\boldsymbol{\lambda}}^{k+1} - \boldsymbol{\lambda}^* \Big)^\top \mathbf{T}^{-1} \Big(\bar{\boldsymbol{\lambda}}^{k+1} - \boldsymbol{\lambda}^* \Big) \\ &\quad = \frac{1}{\rho} \Big(\bar{\boldsymbol{\lambda}}^{k} - \boldsymbol{\lambda}^* + \rho \mathbf{T}\mathbf{r}\big(\hat{\mathbf{x}}^k\big) \Big)^\top \mathbf{T}^{-1} \Big(\bar{\boldsymbol{\lambda}}^{k} - \boldsymbol{\lambda}^* + \rho \mathbf{T}\mathbf{r}\big(\hat{\mathbf{x}}^k\big) \Big) \\ &\quad = \frac{1}{\rho} \big\| \bar{\boldsymbol{\lambda}}^{k} - \boldsymbol{\lambda}^* \big\|^2_{\mathbf{T}^{-1}} + \rho \big\| \mathbf{r}\big(\hat{\mathbf{x}}^k\big) \big\|^2_{\mathbf{T}} \\ &\quad + \frac{2}{\rho} \Big(\bar{\boldsymbol{\lambda}}^{k} - \boldsymbol{\lambda}^* \Big)^\top \Big(\rho \mathbf{r}\big(\hat{\mathbf{x}}^k\big) \Big)\end{aligned} \tag{36}$$

where in the second equality we have used relation (33).

Hence, substituting (35) and (36) into (34), and recalling that $\hat{\boldsymbol{\lambda}}^k - \boldsymbol{\lambda}^k = \rho \mathbf{r}\big(\hat{\mathbf{x}}^k\big)$ we get that the progress $\theta_k(\mathbf{T})$ at each iteration is given by

$$\begin{aligned}\theta_k(\mathbf{T}) &= 2\rho \sum_i \Big(\mathbf{A}_i \mathbf{x}_i^k - \mathbf{A}_i \mathbf{x}_i^* \Big)^\top \Big(\mathbf{A}_i \mathbf{x}_i^k - \mathbf{A}_i \hat{\mathbf{x}}_i^k \Big) \\ &\quad + \frac{2}{\rho} \Big(\bar{\boldsymbol{\lambda}}^k - \boldsymbol{\lambda}^* \Big)^\top \Big(\boldsymbol{\lambda}^k - \hat{\boldsymbol{\lambda}}^k \Big) \\ &\quad - \rho \sum_i \big\| \mathbf{A}_i \hat{\mathbf{x}}_i^k - \mathbf{A}_i \mathbf{x}_i^k \big\|^2_{\mathbf{T}} - \rho \big\| \mathbf{r}\big(\hat{\mathbf{x}}^k\big) \big\|^2_{\mathbf{T}}.\end{aligned} \tag{37}$$

The last two (quadratic) terms in (37) are always negative, due to $\mathbf{T}$ being positive definite by construction. Hence, in order to show that ϕ is strictly decreasing, we need to show that the first two terms in (37) are always "more positive" than the last two terms. This is exactly what Lemma 5 and (30) enable us to do. In particular, using (30), we obtain a lower bound for the first two terms in (37), which gives us that

$$\begin{aligned}\theta_k(\mathbf{T}) \;&\geq\; \rho \sum_i \big\| \mathbf{A}_i \hat{\mathbf{x}}_i^k - \mathbf{A}_i \mathbf{x}_i^k \big\|^2 + \rho \big\| \mathbf{r}\big(\hat{\mathbf{x}}^k\big) \big\|^2_{2\mathbf{T} - \mathbf{D} - 2\mathbf{C}} \\ &\quad + 2\kappa \|\hat{\mathbf{x}}^k - \mathbf{x}^*\|^2 - \rho \sum_i \big\| \mathbf{A}_i \hat{\mathbf{x}}_i^k - \mathbf{A}_i \mathbf{x}_i^k \big\|^2_{\mathbf{T}} - \rho \big\| \mathbf{r}\big(\hat{\mathbf{x}}^k\big) \big\|^2_{\mathbf{T}} \\ &\quad = \rho \sum_i \big\| \mathbf{A}_i \hat{\mathbf{x}}_i^k - \mathbf{A}_i \mathbf{x}_i^k \big\|^2_{\mathbf{I} - \mathbf{T}} + \rho \big\| \mathbf{r}\big(\hat{\mathbf{x}}^k\big) \big\|^2_{\mathbf{T} - \mathbf{D} - 2\mathbf{C}} \\ &\quad + 2\kappa \|\hat{\mathbf{x}}^k - \mathbf{x}^*\|^2.\end{aligned} \tag{38}$$

The above relation suggests that we can choose $\mathbf{T}$ appropriately in order to guarantee that ϕ is strictly decreasing. Specifically, it is sufficient to ensure that the matrices $\mathbf{I} - \mathbf{T}$ and $\mathbf{T} - \mathbf{D} - 2\mathbf{C}$ are positive definite. From the condition $\mathbf{I} - \mathbf{T} > 0$, we infer that the diagonal elements of $\mathbf{T}$ must be strictly less than one. To ensure that $\mathbf{T} - \mathbf{D} - 2\mathbf{C} > 0$, recall that $\mathbf{D} = \mathrm{diag}\big(q_1\tau_1^2, \ldots, q_m\tau_m^2\big)$ by construction. Also, according to Lemma 5, the matrix $\mathbf{C}$ can be any diagonal matrix with strictly positive diagonal entries. Let $\mathbf{C} = \frac{1}{2}\mathbf{T}\mathbf{E}$, where $\mathbf{E} = \mathrm{diag}\big(\epsilon_1, \ldots, \epsilon_m\big)$, and each ϵ_j, $j = 1, \ldots, m$ is an arbitrarily small, positive number. Then, if we can choose $\mathbf{T}$, such that

$$\tau_j - q_j\tau_j^2 - \epsilon_j\tau_j > 0, \ \forall j = 1, \ldots, m$$

the diagonal matrix $\mathbf{T} - \mathbf{D} - 2\mathbf{C}$ is guaranteed to be positive definite. The above relation has solution

$$\tau_j < \frac{1 - \epsilon_j}{q_j}, \ \forall j = 1, \ldots, m. \tag{39}$$

Hence, if we select τ_j according to (39), then $\theta_k > 0$ during the execution of ADAL, which in turn means that the sequence $\{\phi(\mathbf{x}^k, \boldsymbol{\lambda}^k)\}$ is strictly decreasing, as required. Since ϵ_j can be as small as we want, we obtain the corresponding condition of the theorem.

Note that to arrive at (38), we have used the result of Lemma 5, which requires that the terms $\boldsymbol{\lambda}^k + \rho\sum_{j\neq i}(\mathbf{A}_j\mathbf{x}_j^k - \mathbf{A}_j\mathbf{x}_j^*) - \boldsymbol{\lambda}^*$ are sufficiently close to zero for all $i = 1, \ldots, N$ at iteration k; recall that the purpose of this condition is to guarantee that the $\hat{\mathbf{x}}_i^k$ will fall into the strong convexity region of G_ρ, which allows us to use (31). Suppose also that the terms $\mathbf{A}_i\mathbf{x}_i^k - \mathbf{A}_i\mathbf{x}_i^*$ are sufficiently close to zero for all $i = 1, \ldots, N$ at iteration k.

Adding $\mathbf{A}_i\mathbf{x}_i^k - \mathbf{A}_i\mathbf{x}_i^*$ to $\boldsymbol{\lambda}^k + \rho\sum_{j\neq i}(\mathbf{A}_j\mathbf{x}_j^k - \mathbf{A}_j\mathbf{x}_j^*) - \boldsymbol{\lambda}^*$, the assumption that the $\boldsymbol{\lambda}^k + \rho\sum_{j\neq i}(\mathbf{A}_j\mathbf{x}_j^k - \mathbf{A}_j\mathbf{x}_j^*) - \boldsymbol{\lambda}^*$ are sufficiently close to zero at iteration k for all $i = 1,\ldots,N$ in Lemma 5, becomes equivalent to the condition that the terms $\boldsymbol{\lambda}^k + \rho\mathbf{r}(\mathbf{x}^k) - \boldsymbol{\lambda}^*$ and $\mathbf{A}_i\mathbf{x}_i^k - \mathbf{A}_i\mathbf{x}_i^*$ for all $i = 1,\ldots,N$ are sufficiently close to zero at iteration k.

Note that $\boldsymbol{\lambda}^k + \rho\mathbf{r}(\mathbf{x}^k) - \boldsymbol{\lambda}^*$ is sufficiently close to zero if and only if $\boldsymbol{\lambda}^k + \rho(\mathbf{I}-\mathbf{T})\mathbf{r}(\mathbf{x}^k) - \boldsymbol{\lambda}^*$ is sufficiently close to zero at iteration k. This is because $(\mathbf{I}-\mathbf{T})$ is a finite multiplicative factor on $\mathbf{r}(\mathbf{x}^k)$ and $\mathbf{r}(\mathbf{x}^k)$ is close to zero, since the $\mathbf{A}_i\mathbf{x}_i^k$ are close to $\mathbf{A}_i\mathbf{x}_i^*$ for all $i = 1,\ldots,N$. Now, recall the definition of $\phi(\mathbf{x}^k,\boldsymbol{\lambda}^k) = \sum_{i=1}^N \rho\big\|\mathbf{A}_i\mathbf{x}_i^k - \mathbf{A}_i\mathbf{x}_i^*\big\|^2_{\mathbf{T}^{-1}} + \frac{1}{\rho}\big\|\boldsymbol{\lambda}^k + \rho(\mathbf{I}-\mathbf{T})\mathbf{r}(\mathbf{x}^k) - \boldsymbol{\lambda}^*\big\|^2_{\mathbf{T}^{-1}}$, and observe that the terms in the right-hand side of $\phi(\mathbf{x}^k,\boldsymbol{\lambda}^k)$ are exactly the terms that we need to be sufficiently close to zero in order to apply the result of Lemma 5. Since $\mathbf{0} < \mathbf{T} < \mathbf{I}$, it follows that the terms $\boldsymbol{\lambda}^k + \rho(\mathbf{I}-\mathbf{T})\mathbf{r}(\mathbf{x}^k) - \boldsymbol{\lambda}^*$ and $\mathbf{A}_i\mathbf{x}_i^k - \mathbf{A}_i\mathbf{x}_i^*$ for all $i = 1,\ldots,N$ are sufficiently close to zero if $\phi(\mathbf{x}^k,\boldsymbol{\lambda}^k)$ is sufficiently small. To see this, observe that all terms in the expression for $\phi(\mathbf{x}^k,\boldsymbol{\lambda}^k)$ are individually upper bounded by the value of $\phi(\mathbf{x}^k,\boldsymbol{\lambda}^k)$, e.g., $\|\mathbf{A}_i\mathbf{x}_i^k - \mathbf{A}_i\mathbf{x}_i^*\|^2 < \|\mathbf{A}_i\mathbf{x}_i^k - \mathbf{A}_i\mathbf{x}_i^*\|^2_{\mathbf{T}^{-1}} \leq \frac{1}{\rho}\phi(\mathbf{x}^k,\boldsymbol{\lambda}^k)$. Hence, if we choose initial values $\mathbf{x}^1,\boldsymbol{\lambda}^1$ such that $\phi(\mathbf{x}^1,\boldsymbol{\lambda}^1)$ is sufficiently small, then $\theta_1 > 0$, which implies that $\phi(\mathbf{x}^2,\boldsymbol{\lambda}^2) < \phi(\mathbf{x}^1,\boldsymbol{\lambda}^1)$. Since $\phi(\mathbf{x}^1,\boldsymbol{\lambda}^1)$ is sufficiently small and $\phi(\mathbf{x}^2,\boldsymbol{\lambda}^2)$ is even smaller, we can infer that the iterates $\boldsymbol{\lambda}^k + \rho\sum_{j\neq i}(\mathbf{A}_j\mathbf{x}_j^k - \mathbf{A}_j\mathbf{x}_j^*) - \boldsymbol{\lambda}^*$ will be sufficiently close to zero for all iterations k. Therefore, the result of Lemma 5 can be used, as required. ■

Remark 2: In the statement of Theorem 1, we assume that the initial iterates $\mathbf{x}^1,\boldsymbol{\lambda}^1$ are chosen such that $\phi(\mathbf{x}^1,\boldsymbol{\lambda}^1)$ is sufficiently small. For comparison, in the convergence proof of the standard ALM described in Algorithm 1, the assumption that the dual iterates $\boldsymbol{\lambda}^k$ are sufficiently close to $\boldsymbol{\lambda}^*$ for all iterations is used. Following a similar argument as in Theorem 1, this condition holds true if the initial values $\boldsymbol{\lambda}^1$ are sufficiently close to $\boldsymbol{\lambda}^*$; see [14], [55], [56] for more details. Here, we cannot simply require that the dual variables alone are close to their optimal values. Instead, we need to consider the terms $\boldsymbol{\lambda}^k + \rho\sum_{j\neq i}(\mathbf{A}_j\mathbf{x}_j^k - \mathbf{A}_j\mathbf{x}_j^*)$ for all $i = 1,\ldots,N$, due to the structure of the local ALs, cf. (32), and the distributed nature of the algorithm. This difference gives rise to the condition that $\phi(\mathbf{x}^1,\boldsymbol{\lambda}^1)$ is sufficiently small, which replaces the requirement that $\boldsymbol{\lambda}^1$ is sufficiently close to $\boldsymbol{\lambda}^*$ as is the case in the ALM.

We are now ready to prove the main result of this paper.

Theorem 2: Assume(A1)–(A4). Assume also that ρ is sufficiently large, and that the initial iterates $\mathbf{x}^1,\boldsymbol{\lambda}^1$ are chosen such that $\phi(\mathbf{x}^1,\boldsymbol{\lambda}^1)$ is sufficiently small. Then, the ADAL method generates sequences of primal variables $\{\hat{\mathbf{x}}^k\}$ and dual variables $\{\boldsymbol{\lambda}^k\}$ that converge to a local minimum $\mathbf{x}^*$ of problem (1) and the corresponding optimal Lagrange multipliers $\boldsymbol{\lambda}^*$, respectively.

Proof: Relation (38) implies that

$$\phi(\mathbf{x}^{k+1},\boldsymbol{\lambda}^{k+1}) \leq \phi(\mathbf{x}^k,\boldsymbol{\lambda}^k) - \rho\sum_i \big\|\mathbf{A}_i\hat{\mathbf{x}}_i^k - \mathbf{A}_i\mathbf{x}_i^k\big\|^2_{\mathbf{I}-\mathbf{T}} - \rho\big\|\mathbf{r}(\hat{\mathbf{x}}^k)\big\|^2_{\mathbf{T}-\mathbf{D}-2\mathbf{C}} - 2\kappa\|\hat{\mathbf{x}}^k - \mathbf{x}^*\|^2.$$

Summing the above inequality for $k = 1,2,\ldots$, we obtain

$$\sum_{k=1}^{\infty}\Big[\rho\sum_i \big\|\mathbf{A}_i\hat{\mathbf{x}}_i^k - \mathbf{A}_i\mathbf{x}_i^k\big\|^2_{\mathbf{I}-\mathbf{T}} + \rho\big\|\mathbf{r}(\hat{\mathbf{x}}^k)\big\|^2_{\mathbf{T}-\mathbf{D}-2\mathbf{C}} + 2\kappa\|\hat{\mathbf{x}}^k - \mathbf{x}^*\|^2\Big] < \phi(\mathbf{x}^1,\boldsymbol{\lambda}^1). \tag{40}$$

Since $\phi(\mathbf{x}^1,\boldsymbol{\lambda}^1)$ is bounded, this implies that the sequences $\{\mathbf{r}(\hat{\mathbf{x}}^k)\}$, $\{\hat{\mathbf{x}}_i^k - \mathbf{x}_i^*\}$, and $\{\mathbf{A}_i\hat{\mathbf{x}}_i^k - \mathbf{A}_i\mathbf{x}_i^k\}$ for all $i = 1\ldots,N$, converge to zero as $k \to \infty$. It follows that $\{\mathbf{r}(\mathbf{x}^k)\}$ converges to zero as well. By the monotonicity and boundedness properties of $\phi(\mathbf{x}^k,\boldsymbol{\lambda}^k)$, we conclude that the sequence $\{\boldsymbol{\lambda}^k\}$ is also convergent. We denote $\lim_{k\to\infty}\boldsymbol{\lambda}^k = \boldsymbol{\mu}$.

From assumption (A2), the gradients of the functions f_i are continuous on $\mathcal{X}_i$. Therefore, the sequences $\{\nabla f_i(\hat{\mathbf{x}}_i^k)\}$ converge to $\nabla f_i(\mathbf{x}_i^*)$ for all $i = 1,\ldots,N$. Passing to the limit in equation (17), we infer that each sequence $\{\mathbf{z}_i^k\}$ converges to a point $\tilde{\mathbf{z}}_i$, $i = 1,\ldots,N$. The mapping $\mathbf{x}_i \rightrightarrows \mathcal{N}_{\mathcal{X}_i}(\mathbf{x}_i)$ has closed graph and, hence, $\tilde{\mathbf{z}}_i \in \mathcal{N}_{\mathcal{X}_i}(\mathbf{x}_i^*)$.

After the limit pass in (17), we conclude that

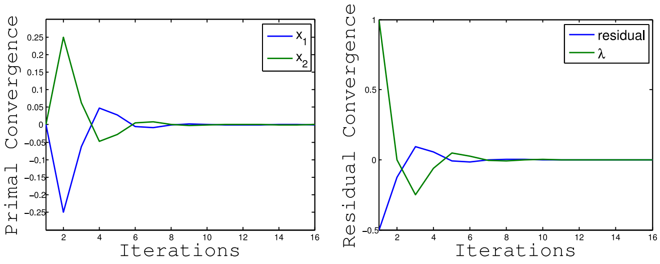
$$0 = \nabla f_i(\mathbf{x}_i^*) + \mathbf{A}_i^\top\boldsymbol{\mu} + \tilde{\mathbf{z}}_i, \quad \forall\ i = 1\ldots,N.$$

Hence, $\boldsymbol{\mu}$ satisfies the first-order optimality conditions for problem (1). Since $\mathbf{x}^*$ is a feasible point that satisfies the strong second-order sufficient conditions of optimality for problem (1), we conclude that ADAL generates primal sequences $\{\hat{\mathbf{x}}^k\}$ that converge to a local minimum $\mathbf{x}^*$ of (1), and dual sequences $\{\boldsymbol{\lambda}^k\}$ that converge to their optimal values $\boldsymbol{\lambda}^*$ for the point $\mathbf{x}^*$, as required. ■

Remark 3: The sufficient closeness assumption used in this paper, cf. Lemma 5 and Theorem 1, is required to establish strong convexity of the local ALs and, subsequently, local convergence of the proposed distributed AL method. Analogous proximity assumptions are used to show convergence of the centralized AL method for nonconvex problems in [14]. Nevertheless, for problems where the constraint matrices $\mathbf{A}_i$ are full column rank, the sufficient closeness assumption is no longer necessary. Instead, for problems with this structure the strong convexity of the local ALs can be established by selecting a large enough value for ρ, which can be computed based on bounds on the gradients of the nonconvex functions f_i at all points in the constraint space, in a spirit similar to the analysis presented in [48].

IV. Numerical Experiments

In order to illustrate the proposed method, in this section, we present numerical results of ADAL applied to nonconvex optimization problems. The main objectives here are two. First, we verify the correctness of the theoretical analysis developed in Section III by showing that the proposed distributed method converges to a local minimum. We also show that the Lyapunov function defined in (14) is indeed strictly decreasing for all iterations, as expected. Second, we examine how sensitive ADAL is to the choice of the user-defined penalty coefficient ρ, and also to different initialization points.



(a) (b)

Fig. 1. Simulation results for ADAL applied to problem (41): (a) Evolution of the primal variables x_1 and x_2, and (b) evolution of the dual variable λ and the constraint residual $x_1 - x_2$.

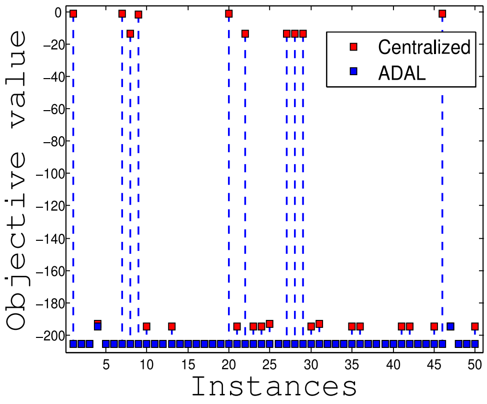


Fig. 2. Simulation results for ADAL and the centralized solver applied to problem (42). The results correspond to 50 different initialization instances. At each instance, the initialization point is the same for both ADAL and the centralized solver. The red and blue squares indicate the objective function value at the point of convergence for the centralized method and ADAL, respectively. A blue dashed line indicates that ADAL converged to a better (or the same) local minimum, while a red dashed line indicates the opposite.

Since the problems are nonconvex, ADAL will converge to some local minimum. To evaluate the quality of this local minimum, we use the solution that is obtained by directly solving the nonconvex problems with a commercial nonlinear optimization solver; we refer to that solution as "centralized," as we do not enforce any decomposition when using this solver. Note that the goal here is not to compare the centralized solution to the solution that is returned by ADAL, but rather to establish that ADAL does not converge to trivial solutions. In comparison, in the convex case, we would compare the solution of ADAL to the global optimal solution and show that they are the same. The simulations were carried out in MATLAB, using the $fmincon$ command to solve the centralized problem, as well as the nonconvex local subproblems (7) at each iteration of ADAL.[1] The results correspond to the *"active-set"* solver option of $fmincon$, which performed better than all other options, in terms of optimality and computation time.

First, we examine a simple nonconvex optimization problem with $N = 2$ agents that control their decision variables x_1 and x_2, respectively. The problem is

$$\min_{x_1, x_2} x_1 \cdot x_2, \quad s.t.\ x_1 - x_2 = 0. \tag{41}$$

This problem is particularly interesting because the straightforward application of the popular ADMM algorithm fails to converge, as discussed in [50]. The problem has an obvious optimal solution at $x_1^* = x_2^* = \lambda^* = 0$. It is shown in [50] that initializing ADMM at $x_1^1 = x_2^1 = 0$ and $\lambda^1 \neq 0$ for this problem gives iterates of the form $\mathbf{x}^{k+1} = \mathbf{0}$ and $\lambda^{k+1} = -2\lambda^k$, and we can see how the latter update produces a diverging dual sequence. On the other hand, the proposed ADAL method is convergent for the same initialization, as can be seen in Fig. 1.

Next, we consider a nonconvex problem with $N = 6$ agents, where each agent controls a scalar decision variable x_i, $i = 1, \ldots, 6$ that is subject to box constraints. Each agent has a different nonconvex objective function and all decisions are coupled in a single linear constraint:

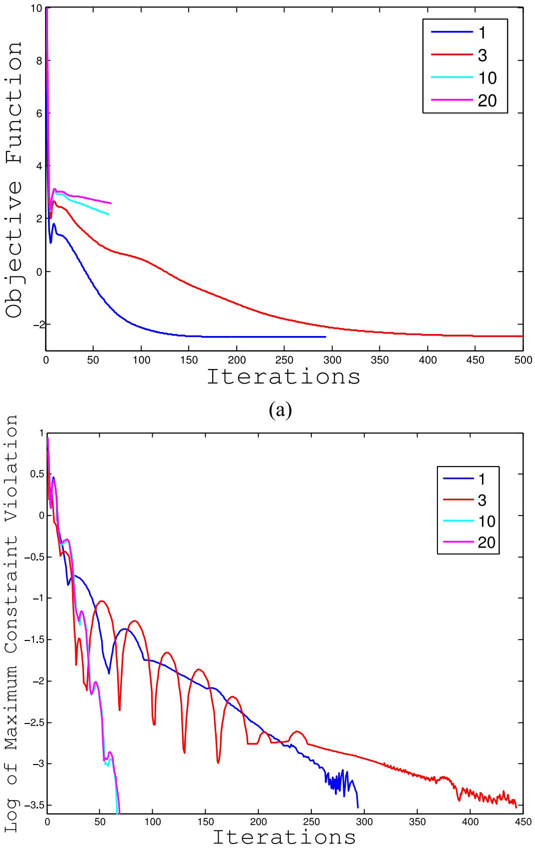
$$\begin{aligned}\min_{\mathbf{x}} \quad & \cos(x_1) + \sin(x_2) + e^{x_3} + 0.1x_4^3 \\ & + \frac{1}{1 + e^{-x_5}} + 0.05\left(x_6^5 - x_6 - x_6^4 + x_6^3\right) \\ \text{s.t.} \quad & x_1 + x_2 + x_3 + x_4 + x_5 + x_6 = 4, \\ & -5 \leq x_i \leq 5, \ \ \forall\, i = 1, \ldots, 6.\end{aligned} \tag{42}$$

The simulation results for this problem are depicted in Fig. 2, where we compare the solutions of ADAL and the centralized solver for 50 different initialization instances. For each instance, the initialization points for each x_i, $i = 1, \ldots, 6$, are generated by sampling from the uniform distribution with support $[-5, 5]$. We set $\rho = 1$, and terminate ADAL after the maximum residual $\max_j \|\mathbf{r}_j(\mathbf{x}^k)\|$, i.e., the maximum constraint violation among all constraints $j = 1, \ldots, m$, reached a value of 1e− 4. We note that this termination criterion was satisfied at around 100 iterations for practically all instances. Also note that for this case $m = 1$ and $q = 6$, hence, the stepsize is simply a scalar that is set to $\tau = 1/6$. For this problem, we observe an interesting behavior: ADAL converges to the "best" local minimum of the problem in almost all cases, which is not always true for the centralized solver. Both schemes are initialized at the same point at each instance.

Next, we consider a problem with multiple constraints $m = 5$, more agents $N = 8$, and larger box constraint sets

$$\begin{aligned}\min_{\mathbf{x}} \quad & \cos(x_1) + \sin(x_2) + e^{x_3} + 0.1x_4^3 \\ & + 0.1/(1 + e^{-x_5}) + 0.01\left(x_6^5 - x_6 - x_6^4 + x_6^3\right) \\ & + \sqrt{x_7 + 15}\sin(x_7/10) + e^{x_8}/(x_8^2 + e^{x_8}) \\ \text{s.t.} \quad & \mathbf{A}\mathbf{x} = \mathbf{b}, \\ & -10 \leq x_i \leq 10, \ \ \forall\, i = 1, \ldots, 8\end{aligned} \tag{43}$$

[1] We note that, for the problems considered here, the $fmincon$ solver of Matlab returned the same solutions as other solvers such as MINOS, LANCELOT, SNOPT, and IPOPT in AMPL for the vast majority of cases. Since the purpose of this paper is not to compare the performance of nonlinear optimization solvers, we have focused just on the $fmincon$.



(a)

(b)

Fig. 3. Simulation results of ADAL applied to problem (43) for different values of the penalty parameter $\rho = 1, 3, 10, 20$: (a) Objective function convergence, and (b) constraint violation convergence.

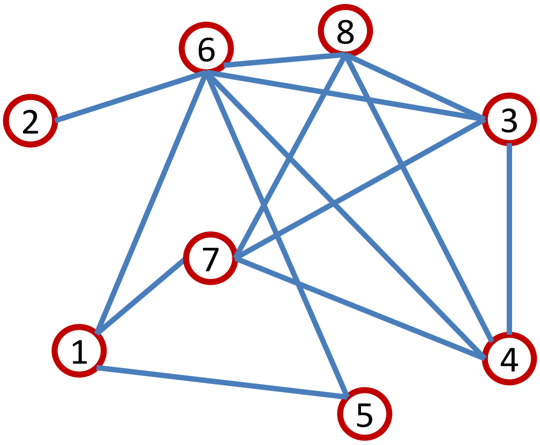


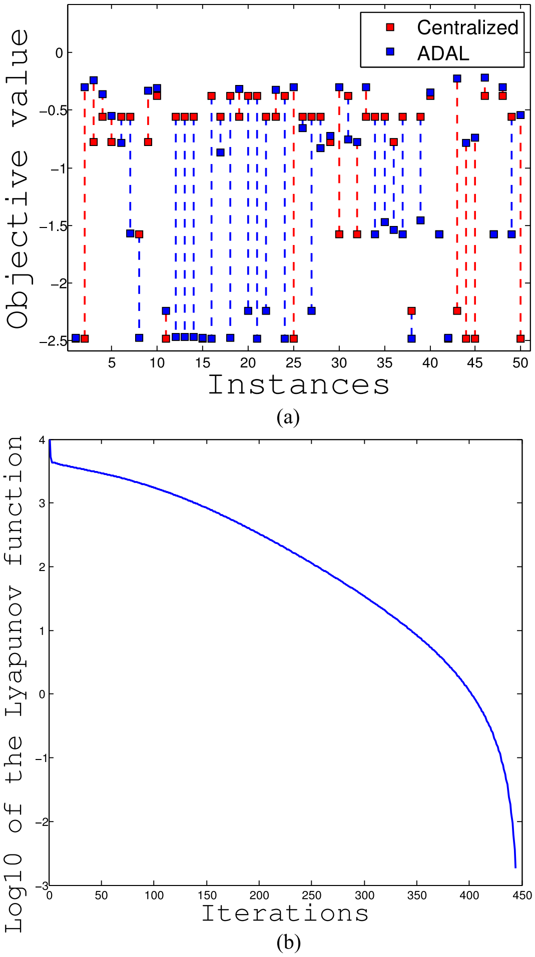
Fig. 4. Structure of the ADAL communication network that needs to be established between the agents of problem (43). The red circles indicate agents, while the blue lines depict two-way message exchanges between the corresponding agents.

where the constraint parameters $\mathbf{A} \in \mathbb{R}^{5\times 8}$ and $\mathbf{b} \in \mathbb{R}^5$ are randomly generated with entries sampled from the standard normal distribution (such that the problem is feasible). When generating $\mathbf{A}$, we always ensure that it has full row rank (to prevent trivial constraint sets), and that at least two decision variables are coupled in each constraint.

Fig. 3 depicts the convergence results of ADAL applied to problem (43), where the generated matrix $\mathbf{A}$ is equation shown at the bottom of this page, and $\mathbf{b} = [-0.0579, -1.6883, 0.8465, 0.1843, 0.6025]^\top$. In this case, the stepsizes are set to $\mathbf{T} = \text{diag}(1/5, 1/2, 1/3, 1/2, 1/3)$. To examine how sensitive ADAL is to the choice of the user-defined penalty coefficient ρ, we present convergence results for four different choices $\rho = 1, 3, 10, 20$. We terminate ADAL after reaching a maximum constraint violation of 3e− 4. Two significant observations can be made based on these results. On one hand, choosing larger values of ρ, e.g., 10 or 20, leads to faster convergence, albeit at the cost of converging to a worse local minimum in terms of objective function value. On the other hand, choosing small ρ, e.g., 1 or 3, allows ADAL to find a better solution, however, convergence of the constraint violation slows down significantly after reaching accuracy levels of about 1e− 3. Furthermore, to clarify the necessary communication pattern between agents during the execution of ADAL, cf. the pertinent discussion in Section II, Fig. 4 illustrates the communication network that needs to be established for this particular problem. For example, agent 2 is coupled only in the second constraint with agent 6, hence, it only needs to communicate with 6.

In order to test the sensitivity of ADAL to initialization for problem (43), we test it for 50 different initialization instances. The results are depicted in Fig. 5(a), where we also include the solutions obtained from the centralized scheme for the same initializations as ADAL. We observe that, for this problem, the choice of initialization point plays a more significant role in determining which local minimum ADAL will converge to, as compared to the corresponding results for the previous problem (42) where ADAL converged to the same local minimum for the vast majority of initializations. Moreover, in Fig. 5(b), we plot the evolution of the Lyapunov function $\phi(\mathbf{x}^k, \boldsymbol{\lambda}^k)$, cf. (14), for an instance of problem (43), where ADAL is initialized (randomly) at $\mathbf{x}^0 = [4.993, -5.904, -4.087, 2.292, -1.648, -2.883, 6.388, 7.331]$ and $\boldsymbol{\lambda} = \mathbf{0}$ with $\rho = 1$. We observe that ϕ is strictly decreasing at each iteration, as expected.

$$\begin{pmatrix} 0 & 0 & 1.2634 & 0.9864 & 0 & 0.4970 & -0.2259 & -0.2783 \\ 0 & 1.6995 & 0 & 0 & 0 & 1.9616 & 0 & 0 \\ -1.8780 & 0 & 0 & 0 & 0 & -2.5970 & -0.8325 & 0 \\ 0 & 0 & 0 & -0.3894 & 0 & 0 & 0 & 0.8270 \\ -0.8666 & 0 & 0 & 0 & 0.2461 & -0.1226 & 0 & 0 \end{pmatrix}$$



(a)

(b)

Fig. 5. (a) Simulation results for ADAL and the centralized solver applied to problem (43). The results correspond to 50 different initialization instances. At each instance, the initialization point is the same for both ADAL and the centralized solver. The red and blue squares indicate the objective function value at the point of convergence for the centralized method and ADAL, respectively. A blue dashed line indicates that ADAL converged to a better (or the same) local minimum, while a red dashed line indicates the opposite. (b) Evolution of the Lyapunov function $\phi(\mathbf{x}^k, \boldsymbol{\lambda}^k)$.

Next, we test ADAL on problems of the form (43) for 50 different instances of the problem parameters $\mathbf{A}$ and $\mathbf{b}$. The objective of this experiment is to examine the behavior of ADAL with a predefined value of ρ for a wide range of problems, instead of finding the best ρ for a given problem as in Fig. 3. This is important for practical applications, where we need to choose a value for ρ without knowing the exact problem parameters. In order to ensure that ρ is sufficiently large for all problem realizations, in this experiment, we set $\rho = 5$. We terminate ADAL after reaching a maximum constraint violation of 3e−4. The results are shown in Fig. 6. We observe that overall the performance of ADAL is satisfactory, judging by the fact that it converges to the same local minimum as the centralized solver for most of the cases.

In the theoretical analysis of Section III, we used the assumptions that the initialization point is sufficiently close to a locally optimal solution and that ρ is large enough. Here, we perform numerical experiments to explore more thoroughly how these conditions affect the convergence of the proposed method. Towards this goal, we consider the following optimal consensus problem, where 25 agents have different versions of the Rosenbrock function and all need to agree on a common optimal decision that minimizes the sum of the individual objectives:

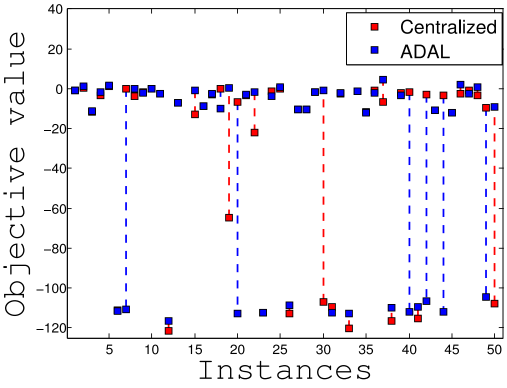


Fig. 6. Simulation results for ADAL and the centralized solver applied to problem (43). The results correspond to 50 different choices of the problem parameters $\mathbf{A}$ and $\mathbf{b}$. At each instance, the initialization point is the same for both ADAL and the centralized solver. The red and blue squares indicate the objective function value at the point of convergence for the centralized method and ADAL, respectively. A blue dashed line indicates that ADAL converged to a better (or the same) local minimum, while a red dashed line indicates the opposite.

TABLE I
Convergence Results for Problem (44)

Value of ρ	50	100	250	500
Converged cases	820 (41%)	960 (48%)	220 (11%)	N/A
Mean obj. value	346.83	817.04	2003.9	N/A
Min obj. value	257.78	505.02	1483.8	N/A
Max obj. value	479.66	1319.4	2250	N/A

$$\min \sum_{i=1}^{25} (a_i - x_i)^2 + b_i (y_i - x_i^2)^2$$

$$\begin{aligned} &\text{subject to } x_i = x_{i+1},\ \forall i = 1, \dots, 24,\\ &y_i = y_{i+1},\ \forall i = 1, \dots, 24,\\ &x_i, y_i \in [-4, 4], \quad \forall i = 1, \dots, 25. \end{aligned} \tag{44}$$

We generate 2000 instances of the problem; for each instance the parameters $a_i \in [1, 6]$, $b_i \in [40, 120]$, and the primal variables x_i, y_i are randomly sampled from a uniform distribution for each agent i, while the dual variables are initialized uniformly randomly within the $[-10, 10]$ interval. We consider values of $\rho \in \{50, 100, 250, 500\}$. For each instance, we start with $\rho = 50$ and if the algorithm does not converge (maximum absolute constraint violation of 10^{-3}) within 1000 iterations, we increase ρ to the next value and restart the algorithm from the same initialization point. The convergence results are summarized in Table I, where we include the percentage of converged cases for each value of ρ (note that they sum to 100%), and the average, minimum, and, maximum objective function values at convergence for each value of ρ over the 2000 instances. We observe that, for large enough ρ, the proposed method always converges, regardless of the initialization. Nevertheless, it appears that for larger values of ρ the algorithm consistently converges to points with relatively larger objective function value; an interesting result that warrants further investigation on how the value of ρ affects the convergence properties of the proposed method.

The aforementioned results suggest that certain heuristics can be used to appropriately tune ADAL. For example, we can perform an online hyperparameter search by running in parallel multiple instances of ADAL, each one for a different value of ρ and a different initialization, and then selecting the best solution. If running multiple problem instances in parallel is not possible due to limited resources, we can alternatively perform a dynamic-update search where we run one instance of ADAL each time, starting with small values of ρ, and increasing ρ if the solution does not yield a reasonable reduction in the constraint violations within a pre-specified number of iterations. Note that the theoretical analysis does not allow for varying ρ during the execution of a single ADAL instance, i.e., if we change ρ between iterations there is no guarantee that ADAL will converge. This is a typical characteristic of all ALMs, distributed or not.

V. Conclusion

In this paper, we have investigated a distributed solution technique for a certain class of nonconvex constrained optimization problems. In particular, we have considered the problem of minimizing the sum of, possibly nonconvex, local objective functions whose arguments are local variables that are constrained to lie in closed, convex sets. The local variables are also globally coupled via a set of affine constraints. We have proposed an iterative distributed algorithm and established its convergence to a local minimum of the problem under assumptions that are commonly used for the convergence of nonconvex optimization methods. The proposed method is based on the AL framework and is an extension of previous work that considered only convex problems. To the best of our knowledge, this is the first paper that formally establishes the convergence to local minima for a distributed ALM in nonconvex settings. Moreover, compared to our previous work, in this paper, we have proposed a more general and fully decentralized rule to select the stepsizes involved in the method. We have verified the theoretical convergence analysis via numerical simulations.

VI. Appendix

Proof of Lemma 2: Consider the result of Lemma 1 and add the term $\rho\sum_i\Big(\mathbf{A}_i\hat{\mathbf{x}}_i^k-\mathbf{A}_i\mathbf{x}_i^*\Big)^\top\Big(\mathbf{A}_i\mathbf{x}_i^k-\mathbf{A}_i\hat{\mathbf{x}}_i^k\Big)$ to both sides of inequality (15), which gives us

$$\sum_i\Big(\nabla f_i(\mathbf{x}_i^*)-\nabla f_i(\hat{\mathbf{x}}_i^k)\Big)^\top\Big(\hat{\mathbf{x}}_i^k-\mathbf{x}_i^*\Big)$$

$$+\ \rho\sum_i\Big(\mathbf{A}_i\hat{\mathbf{x}}_i^k-\mathbf{A}_i\mathbf{x}_i^*\Big)^\top\Big(\mathbf{A}_i\mathbf{x}_i^k-\mathbf{A}_i\hat{\mathbf{x}}_i^k\Big)$$

$$+\ \frac{1}{\rho}\Big(\hat{\boldsymbol{\lambda}}^k-\boldsymbol{\lambda}^*\Big)^\top\Big(\boldsymbol{\lambda}^k-\hat{\boldsymbol{\lambda}}^k\Big)$$

$$\geq\ \rho\sum_i\Big(\mathbf{A}_i\hat{\mathbf{x}}_i^k-\mathbf{A}_i\mathbf{x}_i^*\Big)^\top\Big(\sum_{j\neq i}(\mathbf{A}_j\mathbf{x}_j^k-\mathbf{A}_j\hat{\mathbf{x}}_j^k)\Big)$$

$$+\ \rho\sum_i\Big(\mathbf{A}_i\hat{\mathbf{x}}_i^k-\mathbf{A}_i\mathbf{x}_i^*\Big)^\top\Big(\mathbf{A}_i\mathbf{x}_i^k-\mathbf{A}_i\hat{\mathbf{x}}_i^k\Big).$$

Grouping the terms at the right-hand side of the inequality by their common factor, we transform the estimate as follows:

$$\sum_i\Big(\nabla f_i(\mathbf{x}_i^*)-\nabla f_i(\hat{\mathbf{x}}_i^k)\Big)^\top\Big(\hat{\mathbf{x}}_i^k-\mathbf{x}_i^*\Big)$$

$$+\ \rho\sum_i\Big(\mathbf{A}_i\hat{\mathbf{x}}_i^k-\mathbf{A}_i\mathbf{x}_i^*\Big)^\top\Big(\mathbf{A}_i\mathbf{x}_i^k-\mathbf{A}_i\hat{\mathbf{x}}_i^k\Big)$$

$$+\ \frac{1}{\rho}\Big(\hat{\boldsymbol{\lambda}}^k-\boldsymbol{\lambda}^*\Big)^\top\Big(\boldsymbol{\lambda}^k-\hat{\boldsymbol{\lambda}}^k\Big)$$

$$\geq\ \rho\sum_i\Big(\mathbf{A}_i\hat{\mathbf{x}}_i^k-\mathbf{A}_i\mathbf{x}_i^*\Big)^\top\sum_j\Big(\mathbf{A}_j\mathbf{x}_j^k-\mathbf{A}_j\hat{\mathbf{x}}_j^k\Big).$$

Recall that $\sum_j\mathbf{A}_j(\mathbf{x}_j^k-\hat{\mathbf{x}}_j^k)=\mathbf{r}\big(\mathbf{x}^k\big)-\mathbf{r}\big(\hat{\mathbf{x}}^k\big)$, which means that this term is a constant factor with respect to the summation over i in the right-hand side of the previous relation. Moreover, $\sum_i\mathbf{A}_i\hat{\mathbf{x}}_i^k-\sum_i\mathbf{A}_i\mathbf{x}_i^*=\sum_i\mathbf{A}_i\hat{\mathbf{x}}_i^k-\mathbf{b}=\mathbf{r}\big(\hat{\mathbf{x}}^k\big)$. Substituting these terms at the right-hand side of the previous relation, gives us

$$\sum_i\Big(\nabla f_i(\mathbf{x}_i^*)-\nabla f_i(\hat{\mathbf{x}}_i^k)\Big)^\top\Big(\hat{\mathbf{x}}_i^k-\mathbf{x}_i^*\Big)$$

$$+\ \rho\sum_i\Big(\mathbf{A}_i\hat{\mathbf{x}}_i^k-\mathbf{A}_i\mathbf{x}_i^*\Big)^\top\Big(\mathbf{A}_i\mathbf{x}_i^k-\mathbf{A}_i\hat{\mathbf{x}}_i^k\Big)$$

$$+\ \frac{1}{\rho}\Big(\hat{\boldsymbol{\lambda}}^k-\boldsymbol{\lambda}^*\Big)^\top\Big(\boldsymbol{\lambda}^k-\hat{\boldsymbol{\lambda}}^k\Big)$$

$$\geq\ \rho\mathbf{r}\big(\hat{\mathbf{x}}^k\big)^\top\Big(\mathbf{r}\big(\mathbf{x}^k\big)-\mathbf{r}\big(\hat{\mathbf{x}}^k\big)\Big)$$

$$=\ \Big(\hat{\boldsymbol{\lambda}}^k-\boldsymbol{\lambda}^k\Big)^\top\Big(\mathbf{r}\big(\mathbf{x}^k\big)-\mathbf{r}\big(\hat{\mathbf{x}}^k\big)\Big). \tag{45}$$

Next, we substitute the expressions

$$(\mathbf{A}_i\hat{\mathbf{x}}_i^k-\mathbf{A}_i\mathbf{x}_i^*)=(\mathbf{A}_i\mathbf{x}_i^k-\mathbf{A}_i\mathbf{x}_i^*)+(\mathbf{A}_i\hat{\mathbf{x}}_i^k-\mathbf{A}_i\mathbf{x}_i^k)$$

$$\text{and}\ \ \hat{\boldsymbol{\lambda}}^k-\boldsymbol{\lambda}^*=(\boldsymbol{\lambda}^k-\boldsymbol{\lambda}^*)+(\hat{\boldsymbol{\lambda}}^k-\boldsymbol{\lambda}^k)$$

in the left-hand side of (45). We obtain

$$\sum_i\Big(\nabla f_i(\mathbf{x}_i^*)-\nabla f_i(\hat{\mathbf{x}}_i^k)\Big)^\top\Big(\hat{\mathbf{x}}_i^k-\mathbf{x}_i^*\Big)$$

$$+\ \rho\sum_i\Big(\mathbf{A}_i\mathbf{x}_i{}^k-\mathbf{A}_i\mathbf{x}_i^*\Big)^\top\Big(\mathbf{A}_i\mathbf{x}_i^k-\mathbf{A}_i\hat{\mathbf{x}}_i^k\Big)$$

$$+\ \frac{1}{\rho}\Big(\boldsymbol{\lambda}^k-\boldsymbol{\lambda}^*\Big)^\top\Big(\boldsymbol{\lambda}^k-\hat{\boldsymbol{\lambda}}^k\Big)$$

$$\geq\ \sum_i\rho\|\mathbf{A}_i(\mathbf{x}_i^k-\hat{\mathbf{x}}_i^k)\|^2+\frac{1}{\rho}\|\hat{\boldsymbol{\lambda}}^k-\boldsymbol{\lambda}^k\|^2$$

$$+\ \Big(\hat{\boldsymbol{\lambda}}^k-\boldsymbol{\lambda}^k\Big)^\top\Big(\mathbf{r}\big(\mathbf{x}^k\big)-\mathbf{r}\big(\hat{\mathbf{x}}^k\big)\Big)$$

which completes the proof.

References

[1] M. Chiang, S. H. Low, A. R. Calderbank, and J. C. Doyle, “Layering as optimization decomposition: A mathematical theory of network architectures,” *Proc. IEEE*, vol. 9, no. 1, pp. 255–312, Jan. 2007.

[2] A. Ribeiro, N. Sidiropoulos, and G. Giannakis, “Optimal distributed stochastic routing algorithms for wireless multihop networks,” *IEEE Trans. Wireless Commun.*, vol. 7, no. 11, pp. 4261–4272, Nov. 2008.

[3] N. Chatzipanagiotis, Y. Liu, A. Petropulu, and M. M. Zavlanos, “Distributed cooperative beamforming in multi-source multi-destination clustered systems,” *IEEE Trans. Signal Process.*, vol. 62, no. 23, pp. 6105–6117, Dec. 2014.

[4] A. Nedic and A. Ozdaglar, “Distributed subgradient methods for multi-agent optimization,” *IEEE Trans. Autom. Control*, vol. 54, no. 1, pp. 58–61, Jan. 2009.

[5] M. M. Zavlanos, A. Ribeiro, and G. J. Pappas, “Mobility and routing control in networks of robots,” in *Proc. 49th IEEE Conf. Decision Control*, Atlanta, GA, USA, Dec. 2010, pp. 7545–7550.

[6] 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.*, vol. 3, no. 1, pp. 1–122, 2011.

[7] N. Gatsis and G. Giannakis, “Decomposition algorithms for market clearing with large-scale demand response,” *IEEE Trans. Smart Grid*, vol. 4, no. 4, pp. 1976–1987, Dec. 2013.

[8] M. Figueiredo and J. Bioucas-Dias, “Restoration of poissonian images using alternating direction optimization,” *IEEE Trans. Image Process.*, vol. 19, no. 12, pp. 3133–3145, Dec. 2010.

[9] P. Giselsson and A. Rantzer, “Distributed model predictive control with suboptimality and stability guarantees,” in *Proc. 49th IEEE Conf. Decision Control*, 2010, pp. 7272–7277.

[10] Z. Lu, T. K. Pong, and Y. Zhang, “An alternating direction method for finding Dantzig selectors,” *Comput. Statist. Data Anal.*, vol. 56, no. 12, pp. 4037–4046, 2012.

[11] L. Tang, W. Jiang, and G. Saharidis, “An improved Benders decomposition algorithm for the logistics facility location problem with capacity expansions,” *Ann. Oper. Res.*, vol. 210, no. 1, pp. 165–190, 2013.

[12] N. Chatzipanagiotis, D. Dentcheva, and M. M. Zavlanos, “An augmented Lagrangian method for distributed optimization,” *Math. Programm.*, vol. 152, no. 1, pp. 405–434, 2014.

[13] N. Chatzipanagiotis and M. Zavlanos, “A distributed algorithm for convex constrained optimization under noise,” *IEEE Trans. Autom. Control*, vol. 61, no. 9, pp. 2496–2511, Sep. 2016, doi: 10.1109/TAC.2015.2504932.

[14] A. Ruszczyński, *Nonlinear Optimization*. Princeton, NJ, USA: Princeton Univ. Press, 2006.

[15] D. P. Bertsekas and J. N. Tsitsiklis, *Parallel and Distributed Computation: Numerical Methods*. Belmont, MA, USA: Athena Scientific, 1997.

[16] N. Chatzipanagiotis and M. Zavlanos, “On the convergence rate of a distributed augmented Lagrangian optimization algorithm,” in *Proc. IEEE Amer Control Conf.*, Jul. 2015, pp. 541–546.

[17] A. Nedic and A. Ozdaglar, “Approximate primal solutions and rate analysis for dual subgradient methods,” *SIAM J. Optim.*, vol. 19, no. 4, pp. 1757–1780, 2009.

[18] I. Necoara and J. Suykens, “Application of a smoothing technique to decomposition in convex optimization,” *IEEE Trans. Autom. Control*, vol. 53, no. 11, pp. 2674–2679, Dec. 2008.

[19] J. Eckstein and D. P. Bertsekas, “On the Douglas-Rachford splitting method and the proximal point algorithm for maximal monotone operators,” *Math. Programm.*, vol. 55, pp. 293–318, 1992.

[20] A. Ruszczyński, “On convergence of an augmented Lagrangian decomposition method for sparse convex optimization,” *Math. Oper. Res.*, vol. 20, pp. 634–656, 1995.

[21] J. Mulvey and A. Ruszczyński, “A diagonal quadratic approximation method for large scale linear programs,” *Oper. Res. Lett.*, vol. 12, pp. 205–215, 1992.

[22] B. He, L. Hou, and X. Yuan, “On full Jacobian decomposition of the augmented Lagrangian method for separable convex programming,” *SIAM J. Optim.*, vol. 25, no. 4, pp. 2274–2312, 2015.

[23] H. Wang, A. Banerjee, and Z.-Q. Luo, “Parallel direction method of multipliers,” in *Proc. Adv. Neural Inf. Process. Syst.*, 2014, pp. 181–189.

[24] W. Deng, M.-J. Lai, Z. Peng, and W. Yin, “Parallel multi-block ADMM with o(1/k) convergence,” *J. Scientific Comput.*, 2016, doi: 10.1007/s10915-016-0318-2.

[25] D. Bickson, Y. Tock, A. Zymnis, S. Boyd, and D. Dolev, “Distributed large scale network utility maximization,” in *Proc. IEEE Int. Conf. Symp. Inf. Theory*, 2009, pp. 829–833.

[26] E. Wei, A. Ozdaglar, and A. Jadbabaie, “A distributed Newton method for network utility maximization,” in *Proc. 49th IEEE Conf. Decision Control*, Dec. 2010, pp. 1816–1821.

[27] M. Zargham, A. Ribeiro, A. Jadbabaie, and A. Ozdaglar, “Accelerated dual descent for network optimization,” in *Proc. Amer. Control Conf.*, San Francisco, CA, USA, 2011, pp. 2663–2668.

[28] S. Lee and A. Nedic, “Distributed random projection algorithm for convex optimization,” *IEEE J. Sel. Topics Signal Process.*, vol. 7, no. 2, pp. 221–229, Apr. 2013.

[29] S. Lee and A. Nedic, “Gossip-based random projection algorithm for distributed optimization: Error bound,” in *Proc. IEEE 52nd Annu. Conf. Decision Control*, Dec. 2013, pp. 6874–6879.

[30] Y. Nesterov, “Subgradient methods for huge-scale optimization problems,” *Math. Program.*, vol. 146, nos. 1/2, pp. 275–297, 2014.

[31] D. Jakovetic, J. Freitas Xavier, and J. Moura, “Convergence rates of distributed Nesterov-like gradient methods on random networks,” *IEEE Trans. Signal Process.*, vol. 62, no. 4, pp. 868–882, Feb. 2014.

[32] A. Beck, A. Nedic, A. Ozdaglar, and M. Teboulle, “An O(1/k) gradient method for network resource allocation problems,” *IEEE Trans. Control Netw. Syst.*, vol. 1, no. 1, pp. 64–73, Mar. 2014.

[33] S. Hosseini, A. Chapman, and M. Mesbahi, “Online distributed ADMM via dual averaging,” in *Proc. IEEE 53rd Annu. Conf. Decision Control*, Dec. 2014, pp. 904–909.

[34] A. Koppel, F. Jakubiec, and A. Ribeiro, “A saddle point algorithm for networked online convex optimization,” in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, May 2014, pp. 8292–8296.

[35] T.-H. Chang, A. Nedic, and A. Scaglione, “Distributed constrained optimization by consensus-based primal-dual perturbation method,” *IEEE Trans. Autom. Control*, vol. 59, no. 6, pp. 1524–1538, Jun. 2014.

[36] J. F. C. Mota, J. M. F. Xavier, P. M. Q. Aguiar, and M. Pschel, “Distributed optimization with local domains: Applications in MPC and network flows,” *IEEE Trans. Autom. Control*, vol. 60, no. 7, pp. 2004–2009, Jul. 2015.

[37] S. S. Kia, J. Corts, and S. Martnez, “Distributed convex optimization via continuous-time coordination algorithms with discrete-time communication,” *Automatica*, vol. 55, pp. 254–264, 2015.

[38] M. C. Ferris and O. L. Mangasarian, “Parallel variable distribution,” *SIAM J. Optim.*, vol. 4, pp. 815–832, 1994.

[39] M. V. Solodov, “On the convergence of constrained parallel variable distribution algorithms,” *SIAM J. Optim.*, vol. 8, no. 1, pp. 187–196, 1998.

[40] C. Sagastizbal and M. Solodov, “Parallel variable distribution for constrained optimization,” *Comput. Optim. Appl.*, vol. 22, no. 1, pp. 111–131, 2002.

[41] A. Alvarado, G. Scutari, and J.-S. Pang, “A new decomposition method for multiuser DC-programming and its applications,” *IEEE Trans. Signal Process.*, vol. 62, no. 11, pp. 2984–2998, Jun. 2014.

[42] M. Zhu and S. Martinez, “An approximate dual subgradient algorithm for multi-agent non-convex optimization,” *IEEE Trans. Autom. Control*, vol. 58, no. 6, pp. 1534–1539, Jun. 2013.

[43] M. Jakobsson and C. Fischione, “Extensions of fast-Lipschitz optimization for convex and non-convex problems,” in *Proc. 3rd IFAC Workshop Distrib. Estimation Control Networked Syst.*, 2012, pp. 162–167.

[44] S. Marchesini, A. Schirotzek, C. Yang, H. T. Wu, and F. Maia, “Augmented projections for ptychographic imaging,” *Inverse Probl.*, vol. 29, no. 11, 2013, Art. no. 115009.

[45] R. Zhang and J. Kwok, “Asynchronous distributed ADMM for consensus optimization,” in *Proc. 31st Int. Conf. Mach. Learn.*, 2014, pp. 1701–1709.

[46] P. Forero, A. Cano, and G. Giannakis, “Distributed clustering using wireless sensor networks,” *IEEE J. Sel. Topics Signal Process.*, vol. 5, no. 4, pp. 707–724, Aug. 2011.

[47] Y. Shen, Z. Wen, and Y. Zhang, “Augmented Lagrangian alternating direction method for matrix separation based on low-rank factorization,” *Optim. Methods Softw.*, vol. 29, no. 2, pp. 239–263, Mar. 2014.

[48] M. Hong and Z.-Q. Luo, “On the linear convergence of the alternating direction method of multipliers,” *Math. Programm.*, 2016, doi: 10.1007/s10107-016-1034-2.

[49] S. MagnAºsson, P. Weeraddana, M. Rabbat, and C. Fischione, “On the convergence of alternating direction Lagrangian methods for nonconvex structured optimization problems,” in *IEEE Trans. Control Netw. Syst.*, vol. 3, no. 3, pp. 296–309, Sept. 2016,.

[50] B. Houska, J. Frasch, and M. Diehl, “An augmented Lagrangian based algorithm for distributed non-convex optimization,” *SIAM J. Optim.*, vol. 26, pp. 1101–1127, 2016.

[51] E. K. P. Chong and S. H. Zak, *An Introduction to Optimization*. Hoboken, NJ, USA: Wiley, 2013.

[52] L. N. Trefethen and D. Bau III, *Nonlinear Linear Algebra*. Philadelphia, PA, USA: SIAM, 1997.

[53] D. P. Bertsekas, "Constrained optimization and Lagrange multiplier methods." Belmont, MA, USA: Athena Scientific, 1982.

[54] A. Shapiro and J. Sun, "Some properties of the augmented Lagrangian in cone constrained optimization," *Math. Oper. Res.*, vol. 29, no. 3, pp. 479–491, 2004.

[55] J. Nocedal and S. J. Wright, *Numerical Optimization*, 2nd ed. New York, NY, USA: Springer, 2006.

[56] D. P. Bertsekas, "On penalty and multiplier methods for constrained minimization," *SIAM J. Control Optim.*, vol. 14, no. 2, pp. 216–235, 1976.

Nikolaos Chatzipanagiotis (S'12) received the Diploma degree in mechanical engineering, and the M.Sc. degree in microsystems and nanodevices from the National Technical University of Athens, Athens, Greece, in 2006 and 2008, respectively, and the Ph.D. degree in mechanical engineering from Duke University, Durham, NC, USA, in 2015.

His research interests include optimization theory and algorithms with applications on networked control systems, wired and wireless communications, and multiagent mobile robotic networks.

Michael M. Zavlanos (S'05–M'09) received the Diploma degree in mechanical engineering from the National Technical University of Athens, Athens, Greece, in 2002, and the M.S.E. and Ph.D. degrees in electrical and systems engineering from the University of Pennsylvania, Philadelphia, PA, USA, in 2005 and 2008, respectively.

From 2008 to 2009, he was a Postdoctoral Researcher in the Department of Electrical and Systems Engineering, University of Pennsylvania. He then joined the Stevens Institute of Technology, Hoboken, NJ, USA, as an Assistant Professor of Mechanical Engineering, where he remained until 2012. He is currently an Assistant Professor of mechanical engineering and materials science at Duke University, Durham, NC, USA. He also holds a secondary appointment in the Department of Electrical and Computer Engineering. His research interests include a wide range of topics in the emerging discipline of networked systems, with applications in robotic, sensor, communication, and biomolecular networks. He is particularly interested in hybrid solution techniques, on the interface of control theory, distributed optimization, estimation, and networking.

Dr. Zavlanos received the 2014 Office of Naval Research Young Investigator Program Award, the 2011 National Science Foundation Faculty Early Career Development (CAREER) Award, and was a finalist for the Best Student Paper Award at CDC 2006.