On the Convergence Rate of a Distributed Augmented Lagrangian Optimization Algorithm

Nikolaos Chatzipanagiotis and Michael M. Zavlanos

***Abstract*— We consider the Accelerated Distributed Augmented Lagrangians (ADAL) algorithm, a distributed optimization algorithm that was recently developed by the authors to address problems that involve multiple agents optimizing a separable convex objective function subject to convex local constraints and linear coupling constraints. Optimization using augmented Lagrangians (AL) combines low computational complexity with fast convergence speeds due to the regularization terms included in the AL. However, decentralized methods that employ ALs are few, as decomposition of ALs is a particularly challenging task. ADAL is a primal-dual iterative scheme where at every iteration the agents locally optimize a novel separable approximation of the AL and then appropriately update their primal and dual variables, in a way that ensures convergence to their respective optimal sets. In this paper, we prove that ADAL has a worst-case $O(1/k)$ convergence rate, where k denotes the number of iterations. The convergence rate is established in an ergodic sense, i.e., it refers to the ergodic average of the generated sequences of primal variables up to iteration k.**

I. Introduction

Distributed optimization methods have recently received significant attention because they allow us to solve a problem by decomposing it into smaller, more manageable subproblems that can be solved in parallel. For this reason, they are widely used to solve large-scale problems arising in areas as diverse as wireless communications, optimal control, machine learning, artificial intelligence, computational biology, finance and statistics, to name a few. Moreover, distributed algorithms avoid the cost and fragility associated with centralized coordination, and provide better privacy for the autonomous decision makers. These are desirable properties in applications involving networked systems, such as multi-agent robotics, communication or sensor networks, and power distribution systems.

Classic decomposition algorithms utilize the separable structure of the dual function [1, 2]. These methods have low computational cost, however, they suffer from very slow convergence rates and non-uniqueness of solutions, which necessitate the application of advanced techniques from non-smooth optimization in order to ensure the numerical stability and efficiency of the procedure. These drawbacks are alleviated by the application of regularization techniques such as bundle methods and by the augmented Lagrangian (AL) framework. The convergence speed and the numerical advantages of AL methods [1]–[3] provide a strong motivation for creating decomposed versions of them.

Early specialized techniques that allow for decomposition of the AL can be traced back to the works [4]–[6]. More recent literature involves the *Diagonal Quadratic Approximmation* (DQA) algorithm [7, 8] and the *Alternating Direction Method of Multipliers* (ADMM) [1, 9]–[11]. These methods possess some interesting similarities and differences, despite the fact that their convergence proofs follow completely different paths. The DQA method replaces each minimization step in the augmented Lagrangian algorithm [1, 2] by a separable approximation of the AL function. The ADMM methods are based on the relations between splitting methods for monotone operators, such as Douglas-Rachford splitting, and the proximal point algorithm [9, 12].

In this paper, we focus on the *Accelerated Distributed Augmented Lagrangians* (ADAL) method, a distributed algorithm that we have recently developed in [13] for the solution of optimization problems with separable convex objective functions, local convex constraints, and linear coupling constraints. Specifically, ADAL is a primal-dual iterative scheme where at every iteration the agents solve a small local optimization problem, and then appropriately update their primal and dual variables. It was shown in [13] that, under typical convexity assumptions, the ADAL method generates sequences of primal and dual variables that converge to their respective optimal values. The method has been applied to network flow [13, 14], wireless communications [15, 16], and stochastic optimization [17] problems, and numerical results suggest that it compares favorably to the classic dual decomposition methods, as well as the current state-of-the-art AL distributed methods such as the DQA and ADMM.

The contribution of this paper is to show that ADAL has a worst-case $O(1/k)$ convergence rate, where k denotes the iteration number. The convergence rate is established in an ergodic sense, i.e., it refers to the ergodic average of the generated sequence of primal variables up to iteration k. The analysis in this paper is related to that in [18], where an analogous result was shown for the ADMM algorithm. We note that a significant number of works on characterizing the convergence rate of the popular ADMM algorithm has emerged very recently; see, e.g., [19]–[28]. These works make various assumptions, the most common being strong convexity of the objective function.

The rest of this paper is organized as follows: In Section II, we define the optimization problem of interest and state some basic theoretical preliminaries. In Section III, we present the ADAL algorithm and review some of its basic aspects. The main result of this paper is found in Section IV, where the proof of the $O(1/k)$ convergence rate of ADAL is developed.

This work is supported in part by NSF under grants CNS-1261828 and CNS-1302284, and by ONR under grant N000141410479.

Nikolaos Chatzipanagiotis and Michael M. Zavlanos are with the Dept. of Mechanical Engineering and Materials Science, Duke University, Durham, NC, 27708, USA, {n.chatzip,michael.zavlanos}@duke.edu.

Algorithm 1 Augmented Lagrangian Method (ALM)

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

Algorithm 2 Accelerated Distributed Augmented Lagrangians (ADAL)

```
Set $k = 1$ and define initial Lagrange multipliers $\lambda^1$ and initial primal variables $\mathbf{x}^1$.
  1. For fixed Lagrange multipliers $\boldsymbol{\lambda}^k$, determine $\hat{\mathbf{x}}_i^k$ for every $i \in \mathcal{I}$ as the solution of the following problem:
         $\min_{\mathbf{x}_i \in \mathcal{X}_i} \; \Lambda_\rho^i(\mathbf{x}_i, \mathbf{x}^k, \lambda^k).$        (7)
  2. Set for every $i \in \mathcal{I}$
         $\mathbf{x}_i^{k+1} = \mathbf{x}_i^k + \tau(\hat{\mathbf{x}}_i^k - \mathbf{x}_i^k).$        (8)
  3. If the constraints $\sum_{i=1}^N \mathbf{A}_i \mathbf{x}_i^{k+1} = \mathbf{b}$ are satisfied and $\mathbf{A}_i \hat{\mathbf{x}}_i^k = \mathbf{A}_i \mathbf{x}_i^k$ for all $i \in \mathcal{I}$, then stop (optimal solution found). Otherwise, set:
         $\lambda^{k+1} = \lambda^k + \rho\tau\Big(\sum_{i=1}^N \mathbf{A}_i \mathbf{x}_i^{k+1} - \mathbf{b}\Big),$        (9)
     increase $k$ by one and return to Step 1.
```

II. Problem Definition

We are interested in convex optimization problems of the form

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

Problem (1) models situations where a set $\mathcal{I} = \{1, 2, \ldots, N\}$ of decision makers, henceforth referred to as agents, need to determine local decisions $\mathbf{x}_i \in \mathcal{X}_i$ that minimize a collection of functions $f_i(\mathbf{x}_i)$, while respecting a set of affine coupling constraints $\sum_{i=1}^N \mathbf{A}_i \mathbf{x}_i = \mathbf{b}$. Here, for all $i \in \mathcal{I}$, the functions $f_i : \mathbb{R}^{n_i} \to \mathbb{R}$ are convex (not necessarily differentiable), the local sets $\mathcal{X}_i \subseteq \mathbb{R}^{n_i}$ for $i \in \mathcal{I}$ are nonempty closed and convex, and $\mathbf{A}_i$ is $m \times n_i$. Let

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

where $\mathbf{x} = [\mathbf{x}_1^\top, \ldots, \mathbf{x}_N^\top]^\top \in \mathbb{R}^n$, and $n = \sum_{i=1}^N n_i$. Denoting $\mathbf{A} = [\mathbf{A}_1 \ldots \mathbf{A}_N] \in \mathbb{R}^{m \times n}$, the constraint $\sum_{i=1}^N \mathbf{A}_i \mathbf{x}_i = \mathbf{b}$ in problem (1) becomes $\mathbf{A}\mathbf{x} = \mathbf{b}$. Associating Lagrange multipliers $\lambda \in \mathbb{R}^m$ with that constraint, the Lagrangian for (1) is defined as

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

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

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

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

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

The dual function is decomposable and this gives rise to decomposition methods that address the *dual problem* [1, 2]

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

Such 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 [1]–[3]. The AL associated with problem (1) is

$$\Lambda_\rho(\mathbf{x}, \lambda) \;=\; f(\mathbf{x}) \;+\; \langle \lambda, \mathbf{A}\mathbf{x} - \mathbf{b} \rangle \;+\; \frac{\rho}{2} \|\mathbf{A}\mathbf{x} - \mathbf{b}\|^2, \tag{6}$$

where $\rho > 0$ is a penalty parameter. We recall the standard augmented Lagrangian method, also referred to as the "Method of Multipliers" in the literature [1, 2], in Alg. 1.

A major drawback of the Augmented Lagrangian Method stems from the fact that problem (4) is not in separable form due to the quadratic penalty term in (6). This calls for the development of specialized techniques to decompose the augmented Lagrangian, such as [4]–[11].

III. The ADAL Algorithm

In this section, we describe the *Accelerated Distributed Augmented Lagrangian* (ADAL) method, a specialized AL decomposition technique developed by the authors in [13] to solve problems of the form (1). ADAL is a primal-dual iterative scheme, where each iteration 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. Finally, they update their dual variables based on the values of the communicated primal variables. The method is summarized in Alg. 2.

For every agent $i \in \mathcal{I}$, we define the *local augmented Lagrangian* function $\Lambda_\rho^i : \mathbb{R}^{n_i} \times \mathbb{R}^n \times \mathbb{R}^m \to \mathbb{R}$ as

$$\begin{aligned} \Lambda_\rho^i(\mathbf{x}_i, \mathbf{x}^k, \lambda) \;&=\; f_i(\mathbf{x}_i) \;+\; \langle \lambda, \mathbf{A}_i \mathbf{x}_i \rangle \qquad (10) \\ &\quad+\; \frac{\rho}{2} \|\mathbf{A}_i \mathbf{x}_i + \sum_{j \in \mathcal{I}}^{j \neq i} \mathbf{A}_j \mathbf{x}_j^k - \mathbf{b}\|^2. \end{aligned}$$

At the first step of each iteration, each agent minimizes its local AL subject to its local convex constraints, cf. (7). Note that the variables $\mathbf{A}_j\mathbf{x}_j^k$, appearing in the penalty term of the local AL (10), correspond to the local primal variables of agent j that were communicated to agent i. With respect to agent i, these are considered fixed parameters. The penalty term of each Λ_ρ^i can be equivalently expressed as

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

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 $\sum_{j\in\mathcal{I}}^{j\neq i}\big[\mathbf{A}_j\mathbf{x}_j^k\big]_l - b_l$ are just constant terms in the minimization step, and can be excluded. Here, $[\mathbf{A}_i]_l$ denotes the l-th 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 $\big[\mathbf{A}_j\mathbf{x}_j^k\big]_l$ from all subproblems $j \neq i$ that are involved in the same constraints l as i. Moreover, regarding the term $\langle\lambda, \mathbf{A}_i\mathbf{x}_i\rangle$ in (10), we have that $\langle\lambda, \mathbf{A}_i\mathbf{x}_i\rangle = \sum_{j=1}^m \lambda_j[\mathbf{A}_i\mathbf{x}_i]_j$. Hence, we see that, in order to compute (7), each subproblem i needs access only to those λ_j for which $[\mathbf{A}_i]_j \neq \mathbf{0}$.

The second step of ADAL consists of each agent updating its primal variables by taking a convex combination with the corresponding values from the previous iteration, cf. (8). This update depends on a stepsize τ which must satisfy $\tau \in (0, \frac{1}{q})$ to ensure convergence [13]. Here, q is defined as the *maximum degree*, and is a measure of sparsity of the total constraint matrix $\mathbf{A}$. Specifically, for each constraint $j = 1,\dots,m$, we introduce a measure of involvement. We denote the number of agents i associated with this constraint by q_j, that is, q_j is the number of all $i \in \mathcal{I} : [\mathbf{A}_i]_j \neq \mathbf{0}$. We define q to be the maximum over all q_j, i.e., $q = \max_{1\leq j\leq m} q_j$. Intuitively, q is the number of agents coupled in the "most populated" constraint of the problem.

The third and final step of each ADAL iteration consists of the dual update, cf. (9). This step is distributed by structure, since the Lagrange multiplier of the j-th constraint is updated according to $\lambda_j^{k+1} = \lambda_j^k + \rho\tau\big(\sum_{i=1}^N\big[\mathbf{A}_i\mathbf{x}_i^{k+1}\big]_j - b_j\big)$, which implies that the udpate of λ_j needs only information from those i for which $[\mathbf{A}_i]_j \neq \mathbf{0}$.

The convergence of ADAL relies on the following three assumptions, which are typically required in the analysis of convex optimization methods:

(A1) The functions f_i are convex, and the sets $\mathcal{X}_i$ are nonempty, closed, and convex for all $i \in \mathcal{I}$.

(A2) The Lagrange function L has a saddle point $(\mathbf{x}^*, \lambda^*) \in \mathbb{R}^n \times \mathbb{R}^m$ so that

$$L(\mathbf{x}^*, \lambda) \leq L(\mathbf{x}^*, \lambda^*) \leq L(\mathbf{x}, \lambda^*),\ \forall\, \mathbf{x} \in \mathcal{X},\ \ \lambda \in \mathbb{R}^m.$$

(A3) All subproblems (7) are solvable at every iteration.

The convergence proof of ADAL hinges on showing that the Lyapunov/Merit function $\phi^k = \phi^k(\mathbf{x}^k, \lambda^k)$ defined by

$$\phi^k = \rho\sum_{i\in\mathcal{I}}\|\mathbf{A}_i(\mathbf{x}_i^k - \mathbf{x}_i^*)\|^2 + \frac{1}{\rho}\|\bar{\lambda}^k - \lambda^*\|^2, \quad (11)$$

is strictly decreasing throughout the iterations k. The variables $\bar{\lambda}^k$ are defined as

$$\bar{\lambda}^k = \lambda^k + \rho(1-\tau)\mathbf{r}(\mathbf{x}^k). \quad (12)$$

We recall the main convergence result of ADAL from [13].

Theorem 1 ([13]) *Assume (A1)–(A3). If the stepsize satisfies $0 < \tau < \frac{1}{q}$, then, the sequence $\{\phi(\mathbf{x}^k, \lambda^k)\}$, is strictly decreasing. Moreover, if the sets $\mathcal{X}_i$ are bounded for all $i = 1,\dots N$, then the ADAL method, either stops at an optimal solution of problem* (3)*, or generates a sequence of λ^k converging to an optimal solution of it. Any sequence $\{\mathbf{x}^k\}$ generated by the ADAL algorithm has an accumulation point and any such point is an optimal solution of* (1).

IV. Rate of Convergence

In this section we show that the ADAL method has a worst-case $O(1/k)$ rate of convergence in an ergodic sense. Our proof relies on defining the ergodic average of the primal variables up to iteration k as $\tilde{\mathbf{x}}^k = \frac{1}{k}\sum_{p=0}^{k-1}\hat{\mathbf{x}}^p$. Specifically, we show that the difference of the Lagrangian $L(\tilde{\mathbf{x}}^k, \lambda^*)$ at iteration k, cf. (2), from the optimal Lagrangian $L(\mathbf{x}^*, \lambda^*)$, i.e., the nonnegative quantity $L(\tilde{\mathbf{x}}^k, \lambda^*) - L(\mathbf{x}^*, \lambda^*) \geq 0$, decreases at a worst-case $O(1/k)$ rate. We note that this is a theoretical result characterizing the worst possible convergence rate; in practice ADAL will converge significanlty faster, as can be seen for example in [13]–[17].

In what follows, we denote the convex subdifferential of a convex function f at a point $\mathbf{x}$ by $\partial f(\mathbf{x})$. We also use $\mathcal{N}_\mathcal{X}(\mathbf{x})$ to denote the normal cone to the set $\mathcal{X}$ at the point $\mathbf{x}$ [2], i.e., $\mathcal{N}_\mathcal{X}(\mathbf{x}) = \{\mathbf{h} \in \mathbb{R}^n : \langle\mathbf{h}, \mathbf{y} - \mathbf{x}\rangle \leq 0,\ \ \forall\, \mathbf{y} \in \mathcal{X}\}$. To avoid cluttering the notation, we will use $\sum_i$ to denote summation over all $i \in \mathcal{I}$, i.e., $\sum_i = \sum_{i=1}^N$, unless explicitly noted otherwise. Define also the *residual* $\mathbf{r}(\mathbf{x}) \in \mathbb{R}^m$ as the vector containing the amount of all constraint violations with respect to primal variable $\mathbf{x}$, i.e., $\mathbf{r}(\mathbf{x}) = \sum_i \mathbf{A}_i\mathbf{x}_i - \mathbf{b}$. Finally, we define the auxiliary variables

$$\hat{\lambda}^k = \lambda^k + \rho\mathbf{r}(\hat{\mathbf{x}}^k). \quad (13)$$

The basic step of the proof is to show that the relation

$$L(\hat{\mathbf{x}}^k, \lambda^*) - L(\mathbf{x}^*, \lambda^*) \leq \frac{1}{2\tau}\big(\phi^k - \phi^{k+1}\big) \quad (14)$$

holds for all iterations k. In the following lemma, we utilize the first order optimality conditions for each local subproblem (7) to obtain a first result towards proving (14).

Lemma 1 *Assume (A1)–(A3). Then, the following holds:*

$$L(\hat{\mathbf{x}}^k, \lambda^*) - L(\mathbf{x}^*, \lambda^*) \leq -\Big\langle\hat{\lambda}^k - \lambda^*, \mathbf{r}(\hat{\mathbf{x}}^k)\Big\rangle \quad (15) \\ + \rho\sum_i\Big\langle\mathbf{A}_i\big(\mathbf{x}_i^* - \hat{\mathbf{x}}_i^k\big), \sum_{j\neq i}\mathbf{A}_j\big(\mathbf{x}_j^k - \hat{\mathbf{x}}_j^k\big)\Big\rangle.$$

Proof: The first order optimality conditions for each local problem (7) imply the following inclusion

$$0 \in \partial f_i(\hat{\mathbf{x}}_i^k) + \rho\mathbf{A}_i^\top\Big(\lambda^k + \mathbf{A}_i\hat{\mathbf{x}}_i^k + \sum_{j\neq i}\mathbf{A}_j\mathbf{x}_j^k - \mathbf{b}\Big) + \mathcal{N}_{\mathcal{X}_i}(\hat{\mathbf{x}}_i^k).$$

We infer that subgradients $\mathbf{s}_i^k \in \partial f_i(\hat{\mathbf{x}}_i^k)$ and normal elements $\mathbf{z}_i^k \in \mathcal{N}_{\mathcal{X}_i}(\hat{\mathbf{x}}_i^k)$ exist such that the above becomes

$$0 \;=\; \mathbf{s}_i^k + \rho\mathbf{A}_i^\top\Big(\lambda^k + \mathbf{A}_i\hat{\mathbf{x}}_i^k + \sum\nolimits_{j\neq i}\mathbf{A}_j\mathbf{x}_j^k - \mathbf{b}\Big) + \mathbf{z}_i^k.$$

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

$$\Big\langle \mathbf{s}_i^k + \rho\mathbf{A}_i^\top\Big(\lambda^k + \mathbf{A}_i\hat{\mathbf{x}}_i^k + \sum_{j\neq i}\mathbf{A}_j\mathbf{x}_j^k - \mathbf{b}\Big), \mathbf{x}_i^* - \hat{\mathbf{x}}_i^k\Big\rangle \;=\; \\ =\; \Big\langle -\mathbf{z}_i^k, \mathbf{x}_i^* - \hat{\mathbf{x}}_i^k\Big\rangle \;\geq\; 0.$$

Substituting λ^k with $\hat{\lambda}^k$, cf. (13), in the above, we get

$$0 \leq \Big\langle \mathbf{s}_i^k + \mathbf{A}_i^\top\Big[\hat{\lambda}^k + \rho\sum_{j\neq i}\mathbf{A}_j(\mathbf{x}_j^k - \hat{\mathbf{x}}_j^k)\Big], \mathbf{x}_i^* - \hat{\mathbf{x}}_i^k\Big\rangle. \quad (16)$$

By the definition of the subgradient of f_i at $\hat{\mathbf{x}}_i^k$, we have the relation

$$f_i(\mathbf{x}_i) - f_i(\hat{\mathbf{x}}_i^k) \;\geq\; \mathbf{s}_i^k\big(\mathbf{x}_i - \hat{\mathbf{x}}_i^k\big), \quad \forall\, \mathbf{x}_i \in \mathcal{X}_i. \qquad (17)$$

Substituting (17) for $\mathbf{x}_i = \mathbf{x}_i^*$ into (16), we get

$$f_i(\mathbf{x}_i^*) - f_i(\hat{\mathbf{x}}_i^k) \;+\; \Big\langle \hat{\lambda}^k, \mathbf{A}_i\big(\mathbf{x}_i^* - \hat{\mathbf{x}}_i^k\big)\Big\rangle \\ +\; \rho\Big\langle \mathbf{A}_i\big(\mathbf{x}_i^* - \hat{\mathbf{x}}_i^k\big), \sum\nolimits_{j\neq i}\mathbf{A}_j\big(\mathbf{x}_j^k - \hat{\mathbf{x}}_i^k\big)\Big\rangle \;\geq\; 0.$$

Summing over all i, we get

$$F(\mathbf{x}^*) - F(\hat{\mathbf{x}}^k) \;+\; \Big\langle \hat{\lambda}^k, \sum\nolimits_i \mathbf{A}_i\big(\mathbf{x}_i^* - \hat{\mathbf{x}}_i^k\big)\Big\rangle \\ +\; \rho\sum\nolimits_i\Big\langle \mathbf{A}_i\big(\mathbf{x}_i^* - \hat{\mathbf{x}}_i^k\big), \sum\nolimits_{j\neq i}\mathbf{A}_j\big(\mathbf{x}_j^k - \hat{\mathbf{x}}_j^k\big)\Big\rangle \;\geq\; 0.$$

Substituting $\sum_i \mathbf{A}_i\big(\mathbf{x}_i^* - \hat{\mathbf{x}}_i^k\big) = \mathbf{b} - \sum_i \mathbf{A}_i\hat{\mathbf{x}}_i^k = -\mathbf{r}(\hat{\mathbf{x}}^k)$, and adding and subtracting $\langle\lambda^*, \mathbf{r}(\hat{\mathbf{x}}^k)\rangle$ to the above, we get

$$F(\mathbf{x}^*) - F(\hat{\mathbf{x}}^k) \;-\; \Big\langle \lambda^*, \mathbf{r}(\hat{\mathbf{x}}^k)\Big\rangle \;-\; \Big\langle \hat{\lambda}^k - \lambda^*, \mathbf{r}(\hat{\mathbf{x}}^k)\Big\rangle \\ +\; \rho\sum\nolimits_i\Big\langle \mathbf{A}_i\big(\mathbf{x}_i^* - \hat{\mathbf{x}}_i^k\big), \sum\nolimits_{j\neq i}\mathbf{A}_j\big(\mathbf{x}_j^k - \hat{\mathbf{x}}_j^k\big)\Big\rangle \;\geq\; 0.$$

Rearranging terms in the above inequality, and noting that

$$F(\hat{\mathbf{x}}^k) + \Big\langle \lambda^*, \mathbf{r}(\hat{\mathbf{x}}^k)\Big\rangle - F(\mathbf{x}^*) = L(\hat{\mathbf{x}}^k, \lambda^*) - L(\mathbf{x}^*, \lambda^*),$$

we obtain

$$L(\hat{\mathbf{x}}^k, \lambda^*) - L(\mathbf{x}^*, \lambda^*) \;\leq\; -\Big\langle \hat{\lambda}^k - \lambda^*, \mathbf{r}(\hat{\mathbf{x}}^k)\Big\rangle \\ +\; \rho\sum\nolimits_i\Big\langle \mathbf{A}_i\big(\mathbf{x}_i^* - \hat{\mathbf{x}}_i^k\big), \sum\nolimits_{j\neq i}\mathbf{A}_j\big(\mathbf{x}_j^k - \hat{\mathbf{x}}_j^k\big)\Big\rangle.$$

as required. ■

Next, we further manipulate the result from Lemma 1 to obtain an expression that will help us prove (14).

Lemma 2 *Assume (A1)–(A3). Then, the following holds:*

$$L(\hat{\mathbf{x}}^k, \lambda^*) - L(\mathbf{x}^*, \lambda^*) \qquad (18) \\ +\frac{\rho}{2}\sum_i \|\mathbf{A}_i(\mathbf{x}_i^k - \hat{\mathbf{x}}_i^k)\|^2 \;+\; \rho(\tau - \frac{\tau^2 q}{2})\|\mathbf{r}(\hat{\mathbf{x}}^k)\|^2 \\ \leq \rho\sum\nolimits_i\Big\langle \mathbf{A}_i(\mathbf{x}_i^k - \mathbf{x}_i^*), \mathbf{A}_i(\mathbf{x}_i^k - \hat{\mathbf{x}}_i^k)\Big\rangle - \Big\langle \bar{\lambda}^k - \lambda^*, \mathbf{r}(\hat{\mathbf{x}}^k)\Big\rangle.$$

Proof: Consider (15), and rearrange the terms as

$$-\Big\langle \hat{\lambda}^k - \lambda^*, \mathbf{r}(\hat{\mathbf{x}}^k)\Big\rangle \;\geq\; L(\hat{\mathbf{x}}^k, \lambda^*) - L(\mathbf{x}^*, \lambda^*) \\ +\; \rho\sum\nolimits_i\Big\langle \mathbf{A}_i\big(\hat{\mathbf{x}}_i^k - \mathbf{x}_i^*\big), \sum\nolimits_{j\neq i}\mathbf{A}_j\big(\mathbf{x}_j^k - \hat{\mathbf{x}}_j^k\big)\Big\rangle.$$

To avoid cluttering the notation, we temporarily disregard the $L(\hat{\mathbf{x}}^k, \lambda^*) - L(\mathbf{x}^*, \lambda^*)$ term, i.e., consider only the terms

$$-\Big\langle \hat{\lambda}^k - \lambda^*, \mathbf{r}(\hat{\mathbf{x}}^k)\Big\rangle \;\geq \\ \rho\sum\nolimits_i\Big\langle \mathbf{A}_i\big(\hat{\mathbf{x}}_i^k - \mathbf{x}_i^*\big), \sum\nolimits_{j\neq i}\mathbf{A}_j\big(\mathbf{x}_j^k - \hat{\mathbf{x}}_j^k\big)\Big\rangle.$$

Add the term $\rho\sum_i\langle \mathbf{A}_i(\hat{\mathbf{x}}_i^k - \mathbf{x}_i^*), \mathbf{A}_i(\mathbf{x}_i^k - \hat{\mathbf{x}}_i^k)\rangle$ to both sides of the above inequality, and group the terms at the right-hand side by their common factor to get

$$\rho\sum\nolimits_i\Big\langle \mathbf{A}_i(\hat{\mathbf{x}}_i^k - \mathbf{x}_i^*), \mathbf{A}_i(\mathbf{x}_i^k - \hat{\mathbf{x}}_i^k)\Big\rangle - \Big\langle \hat{\lambda}^k - \lambda^*, \mathbf{r}(\hat{\mathbf{x}}^k)\Big\rangle \\ \geq\; \rho\sum\nolimits_i\Big\langle \mathbf{A}_i(\hat{\mathbf{x}}_i^k - \mathbf{x}_i^*), \sum\nolimits_j \mathbf{A}_j(\mathbf{x}_j^k - \hat{\mathbf{x}}_j^k)\Big\rangle. \qquad (19)$$

The term $\sum_j \mathbf{A}_j(\mathbf{x}_j^k - \hat{\mathbf{x}}_j^k) = \mathbf{r}(\mathbf{x}^k) - \mathbf{r}(\hat{\mathbf{x}}^k)$ is a constant factor with respect to the summation over i in the right hand side of the above, and $\sum_i \mathbf{A}_i\mathbf{x}_i^* = \mathbf{b}$. Hence, we have that

$$\rho\sum\nolimits_i\Big\langle \mathbf{A}_i(\hat{\mathbf{x}}_i^k - \mathbf{x}_i^*), \sum\nolimits_j \mathbf{A}_j(\mathbf{x}_j^k - \hat{\mathbf{x}}_j^k)\Big\rangle \;= \\ =\; \rho\Big\langle \mathbf{r}(\hat{\mathbf{x}}^k), \mathbf{r}(\mathbf{x}^k) - \mathbf{r}(\hat{\mathbf{x}}^k)\Big\rangle,$$

and (19) becomes

$$\rho\sum\nolimits_i\Big\langle \mathbf{A}_i(\hat{\mathbf{x}}_i^k - \mathbf{x}_i^*), \mathbf{A}_i(\mathbf{x}_i^k - \hat{\mathbf{x}}_i^k)\Big\rangle - \Big\langle \hat{\lambda}^k - \lambda^*, \mathbf{r}(\hat{\mathbf{x}}^k)\Big\rangle \\ \geq\; \rho\Big\langle \mathbf{r}(\hat{\mathbf{x}}^k), \mathbf{r}(\mathbf{x}^k) - \mathbf{r}(\hat{\mathbf{x}}^k)\Big\rangle. \qquad (20)$$

Next, we represent

$$\mathbf{A}_i\hat{\mathbf{x}}_i^k - \mathbf{A}_i\mathbf{x}_i^* = (\mathbf{A}_i\mathbf{x}_i^k - \mathbf{A}_i\mathbf{x}_i^*) + (\mathbf{A}_i\hat{\mathbf{x}}_i^k - \mathbf{A}_i\mathbf{x}_i^k) \;\text{ and} \\ \hat{\lambda}^k - \lambda^* = (\lambda^k - \lambda^*) + (\hat{\lambda}^k - \lambda^k) = (\lambda^k - \lambda^*) + \rho\mathbf{r}(\hat{\mathbf{x}}^k),$$

in the left-hand side of (20). We obtain

$$\rho\sum\nolimits_i\Big\langle \mathbf{A}_i(\mathbf{x}_i^k - \mathbf{x}_i^*), \mathbf{A}_i(\mathbf{x}_i^k - \hat{\mathbf{x}}_i^k)\Big\rangle - \Big\langle \lambda^k - \lambda^*, \mathbf{r}(\hat{\mathbf{x}}^k)\Big\rangle \\ \geq\; \rho\sum\nolimits_i\|\mathbf{A}_i(\mathbf{x}_i^k - \hat{\mathbf{x}}_i^k)\|^2 \;+\; \rho\|\mathbf{r}(\hat{\mathbf{x}}^k)\|^2 \\ +\; \rho\Big\langle \mathbf{r}(\hat{\mathbf{x}}^k), \mathbf{r}(\mathbf{x}^k) - \mathbf{r}(\hat{\mathbf{x}}^k)\Big\rangle.$$

Adding $-(1-\tau)\rho\big\langle \mathbf{r}(\mathbf{x}^k), \mathbf{r}(\hat{\mathbf{x}}^k)\big\rangle$ to both sides of the above inequality, and recalling the definition of $\bar{\lambda}^k$ in (12), we get

$$\rho\sum\nolimits_i\Big\langle \mathbf{A}_i(\mathbf{x}_i^k - \mathbf{x}_i^*), \mathbf{A}_i(\mathbf{x}_i^k - \hat{\mathbf{x}}_i^k)\Big\rangle - \Big\langle \bar{\lambda}^k - \lambda^*, \mathbf{r}(\hat{\mathbf{x}}^k)\Big\rangle \\ \geq\; \rho\sum\nolimits_i\|\mathbf{A}_i(\mathbf{x}_i^k - \hat{\mathbf{x}}_i^k)\|^2 \;+\; \rho\|\mathbf{r}(\hat{\mathbf{x}}^k)\|^2 \qquad (21) \\ +\; \rho\Big\langle \mathbf{r}(\hat{\mathbf{x}}^k), \mathbf{r}(\mathbf{x}^k) - \mathbf{r}(\hat{\mathbf{x}}^k)\Big\rangle - (1-\tau)\rho\Big\langle \mathbf{r}(\mathbf{x}^k), \mathbf{r}(\hat{\mathbf{x}}^k)\Big\rangle.$$

Consider only the last two terms $\rho\Big\langle \mathbf{r}(\hat{\mathbf{x}}^k), \mathbf{r}(\mathbf{x}^k) - \mathbf{r}(\hat{\mathbf{x}}^k)\Big\rangle - (1-\tau)\rho\Big\langle \mathbf{r}(\mathbf{x}^k), \mathbf{r}(\hat{\mathbf{x}}^k)\Big\rangle$ at the right hand side of (21). We can manipulate them to obtain

$$\begin{aligned}
&\rho\Big\langle \mathbf{r}(\hat{\mathbf{x}}^k), \mathbf{r}(\mathbf{x}^k)-\mathbf{r}(\hat{\mathbf{x}}^k)\Big\rangle - (1-\tau)\rho\Big\langle \mathbf{r}(\mathbf{x}^k), \mathbf{r}(\hat{\mathbf{x}}^k)\Big\rangle = \\
&= \rho\Big\langle \mathbf{r}(\hat{\mathbf{x}}^k), \mathbf{r}(\mathbf{x}^k)-\mathbf{r}(\hat{\mathbf{x}}^k)\Big\rangle \\
&\qquad - (1-\tau)\rho\Big\langle \mathbf{r}(\mathbf{x}^k)-\mathbf{r}(\hat{\mathbf{x}}^k)+\mathbf{r}(\hat{\mathbf{x}}^k), \mathbf{r}(\hat{\mathbf{x}}^k)\Big\rangle \\
&= \tau\rho\Big\langle \mathbf{r}(\hat{\mathbf{x}}^k), \mathbf{r}(\mathbf{x}^k)-\mathbf{r}(\hat{\mathbf{x}}^k)\Big\rangle - (1-\tau)\rho\|\mathbf{r}(\hat{\mathbf{x}}^k)\|^2.
\end{aligned}$$

Substituting back in (21), we obtain

$$\begin{aligned}
&\rho\sum\nolimits_i \Big\langle \mathbf{A}_i(\mathbf{x}_i^k-\mathbf{x}_i^*), \mathbf{A}_i(\mathbf{x}_i^k-\hat{\mathbf{x}}_i^k)\Big\rangle - \Big\langle \bar{\lambda}^k-\lambda^*, \mathbf{r}(\hat{\mathbf{x}}^k)\Big\rangle \\
&\geq \rho\sum\nolimits_i \|\mathbf{A}_i(\mathbf{x}_i^k-\hat{\mathbf{x}}_i^k)\|^2 + \tau\rho\|\mathbf{r}(\hat{\mathbf{x}}^k)\|^2 \qquad (22) \\
&\qquad + \tau\rho\Big\langle \mathbf{r}(\hat{\mathbf{x}}^k), \mathbf{r}(\mathbf{x}^k)-\mathbf{r}(\hat{\mathbf{x}}^k)\Big\rangle.
\end{aligned}$$

Our aim now is to show that the right hand side of (22) is nonnegative at all times. For this, consider the term $\tau\rho\Big\langle \mathbf{r}(\hat{\mathbf{x}}^k), \mathbf{r}(\mathbf{x}^k)-\mathbf{r}(\hat{\mathbf{x}}^k)\Big\rangle = \tau\rho\Big\langle \mathbf{r}(\hat{\mathbf{x}}^k), \sum_i \mathbf{A}_i(\mathbf{x}_i^k-\hat{\mathbf{x}}_i^k)\Big\rangle$. Each one of the summands in this term is bounded below by

$$\begin{aligned}
&\tau\rho\Big\langle \mathbf{r}(\hat{\mathbf{x}}^k), \mathbf{A}_i(\mathbf{x}_i^k-\hat{\mathbf{x}}_i^k)\Big\rangle = \tau\rho\sum_{j=1}^m \big[\mathbf{r}(\hat{\mathbf{x}}^k)\big]_j \big[\mathbf{A}_i(\mathbf{x}_i^k-\hat{\mathbf{x}}_i^k)\big]_j \\
&\geq -\frac{1}{2}\sum\nolimits_{j=1}^m \Big(\rho\Big[\mathbf{A}_i(\mathbf{x}_i^k-\hat{\mathbf{x}}_i^k)\Big]_j^2 + \tau^2\rho\big[\mathbf{r}(\hat{\mathbf{x}}^k)\big]_j^2\Big).
\end{aligned}$$

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 $\big[\mathbf{r}(\hat{\mathbf{x}}^k)\big]_j \big[\mathbf{A}_i(\mathbf{x}_i^k-\hat{\mathbf{x}}_i^k)\big]_j = 0$. 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 $\tau\rho\Big\langle \mathbf{r}(\hat{\mathbf{x}}^k), \mathbf{A}_i(\mathbf{x}_i^k-\hat{\mathbf{x}}_i^k)\Big\rangle$ term as

$$\begin{aligned}
&\tau\rho\Big\langle \mathbf{r}(\hat{\mathbf{x}}^k), \mathbf{A}_i(\mathbf{x}_i^k-\hat{\mathbf{x}}_i^k)\Big\rangle \geq \qquad (23) \\
&-\frac{1}{2}\sum\nolimits_{j\in\mathcal{Q}_i} \Big(\rho\Big[\mathbf{A}_i(\mathbf{x}_i^k-\hat{\mathbf{x}}_i^k)\Big]_j^2 + \tau^2\rho\big[\mathbf{r}(\hat{\mathbf{x}}^k)\big]_j^2\Big).
\end{aligned}$$

Recalling that q denotes the maximum number of non-zero blocks $[\mathbf{A}_i]_j$ over all j, and summing inequality (23) over all i, we observe that each quantity $[\mathbf{r}(\hat{\mathbf{x}}^k)]_j^2$ is included in the summation at most q times. This leads us to the bound

$$\begin{aligned}
&\tau\rho\Big\langle \mathbf{r}(\hat{\mathbf{x}}^k), \mathbf{r}(\mathbf{x}^k)-\mathbf{r}(\hat{\mathbf{x}}^k)\Big\rangle \\
&\quad = \sum\nolimits_i \tau\rho\Big\langle \mathbf{r}(\hat{\mathbf{x}}^k), \mathbf{A}_i(\mathbf{x}_i^k-\hat{\mathbf{x}}_i^k)\Big\rangle \qquad (24) \\
&\quad \geq -\frac{\rho}{2}\sum\nolimits_i \|\mathbf{A}_i(\mathbf{x}_i^k-\hat{\mathbf{x}}_i^k)\|^2 - \frac{\tau^2 q\rho}{2}\|\mathbf{r}(\hat{\mathbf{x}}^k)\|^2.
\end{aligned}$$

We substitute (24) back into (22) and arrive at

$$\begin{aligned}
&\rho\sum\nolimits_i \Big\langle \mathbf{A}_i(\mathbf{x}_i^k-\mathbf{x}_i^*), \mathbf{A}_i(\mathbf{x}_i^k-\hat{\mathbf{x}}_i^k)\Big\rangle - \Big\langle \bar{\lambda}^k-\lambda^*, \mathbf{r}(\hat{\mathbf{x}}^k)\Big\rangle \\
&\geq \frac{\rho}{2}\sum\nolimits_i \|\mathbf{A}_i(\mathbf{x}_i^k-\hat{\mathbf{x}}_i^k)\|^2 + \rho(\tau - \frac{\tau^2 q}{2})\|\mathbf{r}(\hat{\mathbf{x}}^k)\|^2. \qquad (25)
\end{aligned}$$

Recall that until now we have disregarded the term $L(\hat{\mathbf{x}}^k, \lambda^*) - L(\mathbf{x}^*, \lambda^*)$. Reinstating this term in (25), we get

$$\begin{aligned}
&L(\hat{\mathbf{x}}^k, \lambda^*) - L(\mathbf{x}^*, \lambda^*) + \frac{\rho}{2}\sum\nolimits_i \|\mathbf{A}_i(\mathbf{x}_i^k-\hat{\mathbf{x}}_i^k)\|^2 \\
&\qquad + \rho(\tau - \frac{\tau^2 q}{2})\|\mathbf{r}(\hat{\mathbf{x}}^k)\|^2 \\
&\leq \rho\sum\nolimits_i \Big\langle \mathbf{A}_i(\mathbf{x}_i^k-\mathbf{x}_i^*), \mathbf{A}_i(\mathbf{x}_i^k-\hat{\mathbf{x}}_i^k)\Big\rangle - \Big\langle \bar{\lambda}^k-\lambda^*, \mathbf{r}(\hat{\mathbf{x}}^k)\Big\rangle,
\end{aligned}$$

as required. ■

Now, we are ready to prove the key relation (14).

Lemma 3 *Assume (A1)–(A3). Then, the following holds:*

$$L(\hat{\mathbf{x}}^k, \lambda^*) - L(\mathbf{x}^*, \lambda^*) \leq \frac{1}{2\tau}\big(\phi^k - \phi^{k+1}\big). \qquad (26)$$

Proof: First, we show that the dual update step (9) of ADAL is equivalent to the update rule

$$\bar{\lambda}^{k+1} = \bar{\lambda}^k + \tau\rho\mathbf{r}(\hat{\mathbf{x}}^k). \qquad (27)$$

for the variables $\bar{\lambda}^k$. Indeed, we have

$$\begin{aligned}
\lambda^{k+1} &= \lambda^k + \tau\rho\mathbf{r}(\mathbf{x}^{k+1}) \\
&= \lambda^k + \tau\rho\Big[(1-\tau)\mathbf{r}(\mathbf{x}^k) + \tau\mathbf{r}(\hat{\mathbf{x}}^k)\Big] \\
&= \lambda^k + \tau\Big[-(1-\tau)\rho\Big(\mathbf{r}(\hat{\mathbf{x}}^k) - \mathbf{r}(\mathbf{x}^k)\Big) + \rho\mathbf{r}(\hat{\mathbf{x}}^k)\Big] \\
&= \lambda^k - (1-\tau)\rho\tau\Big(\mathbf{r}(\hat{\mathbf{x}}^k) - \mathbf{r}(\mathbf{x}^k)\Big) + \tau\rho\mathbf{r}(\hat{\mathbf{x}}^k).
\end{aligned}$$

Adding $(1-\tau)\rho\mathbf{r}(\mathbf{x}^k)$ to both sides of the above equation and rearranging terms we obtain

$$\begin{aligned}
&\lambda^{k+1} + (1-\tau)\rho\Big[\mathbf{r}(\mathbf{x}^k) + \tau\Big(\mathbf{r}(\hat{\mathbf{x}}^k) - \mathbf{r}(\mathbf{x}^k)\Big)\Big] \\
&\qquad\qquad = \lambda^k + (1-\tau)\rho\mathbf{r}(\mathbf{x}^k) + \tau\rho\mathbf{r}(\hat{\mathbf{x}}^k).
\end{aligned}$$

This is equivalent to

$$\lambda^{k+1} + (1-\tau)\rho\mathbf{r}(\mathbf{x}^{k+1}) = \lambda^k + (1-\tau)\rho\mathbf{r}(\mathbf{x}^k) + \tau\rho\mathbf{r}(\hat{\mathbf{x}}^k),$$

which is exactly (27). With that in mind, we have that

$$\begin{aligned}
\phi^k - \phi^{k+1} &= \sum\nolimits_{i=1}^N \rho\|\mathbf{A}_i(\mathbf{x}_i^k-\mathbf{x}_i^*)\|^2 + \frac{1}{\rho}\|\bar{\lambda}^k-\lambda^*\|^2 \\
&\quad - \sum\nolimits_{i=1}^N \rho\|\mathbf{A}_i(\mathbf{x}_i^{k+1}-\mathbf{x}_i^*)\|^2 - \frac{1}{\rho}\|\bar{\lambda}^{k+1}-\lambda^*\|^2 \\
&= 2\tau\Big[\rho\sum\nolimits_i \Big\langle \mathbf{A}_i(\mathbf{x}_i^k-\mathbf{x}_i^*), \mathbf{A}_i(\mathbf{x}_i^k-\hat{\mathbf{x}}_i^k)\Big\rangle - \\
&\quad \Big\langle \bar{\lambda}^k-\lambda^*, \mathbf{r}(\hat{\mathbf{x}}^k)\Big\rangle\Big] - \tau^2\Big[\sum_i \rho\|\mathbf{A}_i(\hat{\mathbf{x}}_i^k-\mathbf{x}_i^k)\|^2 + \rho\|\mathbf{r}(\hat{\mathbf{x}}^k)\|^2\Big].
\end{aligned}$$

Rearranging terms in the above equation, we get that

$$\begin{aligned}
&\rho\sum\nolimits_i \Big\langle \mathbf{A}_i(\mathbf{x}_i^k-\mathbf{x}_i^*), \mathbf{A}_i(\mathbf{x}_i^k-\hat{\mathbf{x}}_i^k)\Big\rangle - \Big\langle \bar{\lambda}^k-\lambda^*, \mathbf{r}(\hat{\mathbf{x}}^k)\Big\rangle \\
&= \frac{1}{2\tau}\big(\phi^k - \phi^{k+1}\big) + \frac{\tau\rho}{2}\Big(\sum\nolimits_i \|\mathbf{A}_i(\hat{\mathbf{x}}_i^k-\mathbf{x}_i^k)\|^2 + \|\mathbf{r}(\hat{\mathbf{x}}^k)\|^2\Big).
\end{aligned}$$

From the relation (18) of Lemma 2, we can substiture the term $\rho\sum_i \Big\langle \mathbf{A}_i(\mathbf{x}_i^k-\mathbf{x}_i^*), \mathbf{A}_i(\mathbf{x}_i^k-\hat{\mathbf{x}}_i^k)\Big\rangle - \Big\langle \bar{\lambda}^k-\lambda^*, \mathbf{r}(\hat{\mathbf{x}}^k)\Big\rangle$ on the left hand side of the above equality, to arrive at

$$\begin{aligned}
&L(\hat{\mathbf{x}}^k, \lambda^*) - L(\mathbf{x}^*, \lambda^*) \\
&\qquad + \frac{\rho}{2}\sum\nolimits_i \|\mathbf{A}_i(\mathbf{x}_i^k-\hat{\mathbf{x}}_i^k)\|^2 + \rho(\tau - \frac{\tau^2 q}{2})\|\mathbf{r}(\hat{\mathbf{x}}^k)\|^2 \\
&\leq \frac{1}{2\tau}\Big(\phi^k - \phi^{k+1}\Big) + \frac{\tau\rho}{2}\Big(\sum\nolimits_i \|\mathbf{A}_i(\hat{\mathbf{x}}_i^k-\mathbf{x}_i^k)\|^2 + \|\mathbf{r}(\hat{\mathbf{x}}^k)\|^2\Big),
\end{aligned}$$

or, equivalently, at

$$\begin{aligned}
&L(\hat{\mathbf{x}}^k, \lambda^*) - L(\mathbf{x}^*, \lambda^*) + \frac{\rho(1-\tau)}{2}\sum\nolimits_i \|\mathbf{A}_i(\mathbf{x}_i^k-\hat{\mathbf{x}}_i^k)\|^2 \\
&\quad + \frac{\rho(\tau-\tau^2 q)}{2}\|\mathbf{r}(\hat{\mathbf{x}}^k)\|^2 \leq \frac{1}{2\tau}\big(\phi^k - \phi^{k+1}\big).
\end{aligned}$$

Since $0 < \tau < \frac{1}{q}$, we have that $1 - \tau > 0$ and $\tau - \tau^2 q > 0$. Thus, the term $\frac{\rho(1-\tau)}{2}\sum_i \|\mathbf{A}_i(\mathbf{x}_i^k - \hat{\mathbf{x}}_i^k)\|^2 + \frac{\rho(\tau-\tau^2 q)}{2}\|\mathbf{r}(\hat{\mathbf{x}}^k)\|^2$ is always nonnegative, which gives us $L(\hat{\mathbf{x}}^k, \lambda^*) - L(\mathbf{x}^*, \lambda^*) \leq \frac{1}{2\tau}\big(\phi^k - \phi^{k+1}\big)$. ■

We can now prove the worst-case $O(1/k)$ convergence rate using (14) together with the properties of convex functions.

Theorem 2 *Assume (A1)–(A3). Let* $\tilde{\mathbf{x}}^k = \frac{1}{k}\sum_{p=0}^{k-1}\hat{\mathbf{x}}^p$ *denote the ergodic average of the primal variable sequence generated by ADAL up to iteration* k*. Then, for all* k

$$0 \;\leq\; L(\tilde{\mathbf{x}}^k, \lambda^*) - L(\mathbf{x}^*, \lambda^*) \;\leq\; \frac{1}{2k\tau}\phi^0. \tag{28}$$

Proof: The first inequality $0 \leq L(\tilde{\mathbf{x}}^k, \lambda^*) - L(\mathbf{x}^*, \lambda^*)$ follows directly from the definition of a saddle point, cf. Assumption (A2). To prove the other inequality, we proceed as follows. Summing (26) for all $p = 0, \dots, k-1$, we get

$$\sum_{p=0}^{k-1} F(\hat{\mathbf{x}}^p) \;+\; \sum_{p=0}^{k-1}\Big\langle \lambda^*, \mathbf{r}(\hat{\mathbf{x}}^p)\Big\rangle \;-\; \sum_{p=0}^{k-1} F(\mathbf{x}^*) \leq \frac{1}{2\tau}\Big(\phi^0 - \phi^k\Big). \tag{29}$$

By the convexity of F, we have that

$$\sum_{p=0}^{k-1}\frac{1}{k}F(\hat{\mathbf{x}}^p) \;\geq\; F\Big(\sum_{p=0}^{k-1}\frac{1}{k}\hat{\mathbf{x}}^p\Big),$$

which implies that $\sum_{p=0}^{k-1} F(\hat{\mathbf{x}}^p) \;\geq\; kF(\tilde{\mathbf{x}}^k)$. The analogous relation holds for $\sum_{p=0}^{k-1}\mathbf{r}(\hat{\mathbf{x}}^p) \;\geq\; k\mathbf{r}(\tilde{\mathbf{x}}^k)$, since it is a linear (convex) mapping. We also have that $\sum_{p=0}^{k-1} F(\mathbf{x}^*) = kF(\mathbf{x}^*)$. Hence, (29) can be expressed as

$$kF(\tilde{\mathbf{x}}^k) + k\langle \lambda^*, \mathbf{r}(\tilde{\mathbf{x}}^k)\rangle - kF(\mathbf{x}^*) \;\leq\; \frac{1}{2\tau}\Big(\phi^0 - \phi^k\Big),$$

or, equivalently,

$$F(\tilde{\mathbf{x}}^k) + \langle \lambda^*, \mathbf{r}(\tilde{\mathbf{x}}^k)\rangle - F(\mathbf{x}^*) \;+\; \frac{1}{2k\tau}\phi^k \;\leq\; \frac{1}{2k\tau}\phi^0.$$

Since $\phi^k \geq 0$, we infer that $L(\tilde{\mathbf{x}}^k, \lambda^*) - L(\mathbf{x}^*, \lambda^*) \;\leq\; \frac{1}{2k\tau}\phi^0$, as required. ■

V. Conclusions

We have considered the ADAL algorithm, an augmented Lagrangian decomposition method for convex optimization problems with linear coupling constraints. ADAL is a distributed iterative scheme, wherein at each iteration all agents update their local decisions based only on local computations and message exchanges with other neighboring agents. In this paper, we have characterized the convergence rate of ADAL by showing that the algorithm generates sequences of primal variables whose ergodic averages converge to their respective optimal values at an $O(1/k)$ rate in the worst-case.

References

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

[2] A. Ruszczyński, *Nonlinear Optimization*. Princeton, NJ, USA: Princeton University Press, 2006.

[3] R. Rockafellar, "Augmented Lagrange multiplier functions and duality in nonconvex programming," *SIAM Journal on Control*, vol. 12, pp. 268–285, 1973.

[4] P. Tatjewski, "New dual-type decomposition algorithm for nonconvex separable optimization problems," *Automatica*, vol. 25, no. 2, pp. 233–242, 1989.

[5] N. Watanabe, Y. Nishimura, and M. Matsubara, "Decomposition in large system optimization using the method of multipliers," *Journal of Optimiz. Theory and Applic.*, vol. 25, no. 2, pp. 181–193, 1978.

[6] G. Chen and M. Teboulle, "A proximal-based decomposition method for convex minimization problems," *Mathematical Programming*, vol. 64, pp. 81–101, 1994.

[7] J. Mulvey and A. Ruszczyński, "A diagonal quadratic approximation method for large scale linear programs," *Operations Research Letters*, vol. 12, pp. 205–215, 1992.

[8] A. Ruszczyński, "On convergence of an Augmented Lagrangian decomposition method for sparse convex optimization," *Mathematics of Operations Research*, vol. 20, pp. 634–656, 1995.

[9] J. Eckstein and D. P. Bertsekas, "On the Douglas-Rachford splitting method and the proximal point algorithm for maximal monotone operators," *Mathematical Programming,*, vol. 55, pp. 293–318, 1992.

[10] ——, "An alternating direction method for linear programming." LIDS, MIT, 1990.

[11] S. Boyd, N. Parikh, E. Chu, B. Peleato, and J. Eckstein, "Distributed optimization and statistical learning via the alternating direction method of multipliers," *Foundations and Trends in Machine Learning*, vol. 3, no. 1, pp. 1–122, 2011.

[12] M. Fortin and R. Glowinski, *Augmented Lagrangian Methods: Applications to the Numerical Solution of Boundary-Value Problems.* Amsterdam: North-Holland, 1983.

[13] N. Chatzipanagiotis, D. Dentcheva, and M. Zavlanos, "An augmented Lagrangian method for distributed optimization," *Mathematical Programming*, 2014.

[14] ——, "Approximate augmented lagrangians for distributed network optimization," in *Decision and Control (CDC), 2012 IEEE 51st Annual Conference on*, Dec 2012, pp. 5840–5845.

[15] N. Chatzipanagiotis, A. Petropulu, and M. Zavlanos, "A distributed algorithm for cooperative relay beamforming," in *American Control Conference (ACC), 2013*, June 2013, pp. 3796–3801.

[16] N. Chatzipanagiotis, Y. Liu, A. Petropulu, and M. M. Zavlanos, "Distributed cooperative beamforming in multi-source multi-destination clustered systems," *IEEE Transactions on Signal Processing*, vol. 62, no. 23, pp. 6105–6117, Dec. 2014.

[17] N. Chatzipanagiotis and M. Zavlanos, "Distributed stochastic multicommodity flow optimization," in *Global Conf. on Signal and Inform. Processing (GlobalSIP), 2013 IEEE*, Dec 2013, pp. 883–886.

[18] B. He and X. Yuan, "On the $O(1/n)$ convergence rate of the Douglas–Rachford alternating direction method," *SIAM Journal on Numerical Analysis*, vol. 50, no. 2, pp. 700–709, 2012.

[19] D. Boley, "Local linear convergence of the alternating direction method of multipliers on quadratic or linear programs," *SIAM Journal on Optimization*, vol. 23, no. 4, pp. 2183–2207, 2013.

[20] R. Monteiro and B. Svaiter, "Iteration-complexity of block-decomposition algorithms and the alternating direction method of multipliers," *SIAM Journal on Optimization*, vol. 23, no. 1, pp. 475–507, 2013.

[21] W. Shi, Q. Ling, K. Yuan, G. Wu, and W. Yin, "On the linear convergence of the admm in decentralized consensus optimization," *Signal Processing, IEEE Transactions on*, vol. 62, no. 7, pp. 1750–1761, April 2014.

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

[23] ——, "On the $O(1/k)$ convergence of asynchronous distributed alternating direction method of multipliers," arXiv:1307.8254.

[24] D. Jakovetic, J. M. F. Moura, and J. Xavier, "Linear convergence rate of a class of distributed augmented lagrangian algorithms," arXiv:1307.2482.

[25] W. Deng and W. Yin, "On the global and linear convergence of the generalized alternating direction method of multipliers," Rice CAAM technical report, 2012.

[26] F. Iutzeler, P. Bianchi, P. Ciblat, and W. Hachem, "Explicit convergence rate of a distributed alternating direction method of multipliers," arXiv:1312.1085.

[27] M. Hong and Z.-Q. Luo, "On the linear convergence of the alternating direction method of multipliers," arXiv:1208.3922.

[28] T. Lin, S. Ma, and S. Zhang, "On the global linear convergence of the ADMM with multi-block variables," arXiv:1408.4266.