A Consensus-Based Distributed Augmented Lagrangian Method

Yan Zhang and Michael M. Zavlanos

***Abstract*— In this paper, we propose a distributed algorithm to solve multi-agent constrained optimization problems. Specifically, we employ the recently developed Accelerated Distributed Augmented Lagrangian (ADAL) algorithm that has been shown to exhibit faster convergence rates in practice compared to relevant distributed methods. Distributed implementation of ADAL depends on separability of the global coupling constraints. Here we extend ADAL so that it can be implemented distributedly independent of the structure of the coupling constraints. For this, we introduce local estimates of the global constraint functions and multipliers and employ a finite number of consensus steps between iterations of the algorithm to achieve agreement on these estimates. The proposed algorithm can be applied to both undirected or directed networks. Theoretical analysis shows that the algorithm converges at rate $O(1/k)$ and has steady error that is controllable by the number of consensus steps. Our numerical simulation shows that it outperforms existing methods in practice.**

I. INTRODUCTION

Distributed optimization algorithms decompose an optimization problem into smaller, more manageable subproblems that can be solved in parallel by a group of agents or processors. For this reason, they are widely used to solve large-scale problems arising in areas as diverse as wireless communications, optimal control, planning, and machine learning, to name a few. In this paper, we consider the optimization problem

$$\min\ F(x) \triangleq \sum_{i=1}^{N} f_i(x_i) \text{ s.t. } \sum_{i=1}^{N} A_i x_i = b,\ x_i \in \mathcal{X}_i, \forall i \quad (1)$$

where $x = [x_1^T, x_2^T, \ldots, x_N^T]^T$, $x_i \in \mathbb{R}^{n_i}$ is the local decision variable, $f_i(x_i)$ is the local objective function, and N is the total number of agents. The function $f_i(x_i)$ is convex and possibly nonsmooth. All agents are globally coupled by the equality constraint $\sum_{i=1}^{N} A_i x_i = b$. Each agent only has access to its own objective $f_i(x_i)$, constraint matrix $A_i \in \mathbb{R}^{m \times n_i}$, and constraint set $\mathcal{X}_i$. All agents have access to the parameters N and b. We are interested in finding a distributed solution to problem (1) in which all agents communicate only with their $1-$hop neighbors. Several distributed algorithms have been proposed to solve problem (1), including dual decomposition, distributed saddle point methods, and distributed Augmented Lagrangian methods.

Dual decomposition deals with the dual problem

$$\min_{\lambda \in \mathbb{R}^m} \quad -\sum_i \phi^i(\lambda) \triangleq -\sum_i \{\Lambda^i(\lambda) + \frac{1}{N} b^T \lambda\}, \quad (2)$$

where $\Lambda^i(\lambda) = \min_{x_i \in \mathcal{X}_i} f_i(x_i) + \lambda^T A_i x_i$. Problem (2) is a consensus optimization problem and can be solved using, e.g., the methods proposed in [1–4]. While these methods only solve the dual problem (2), the Consensus-Dual Decomposition method in [5] also shows convergence of the primal solution for problem (1). Similarly, Consensus-ADMM in [6,7] introduces consensus constraints and applies ADMM to solve problem (2) and proves convergence of the primal solution for problem (1). Besides dual decomposition, distributed saddle-point methods [8–12] have also been proposed to solve (1). Specifically, in [8,10,12] all agents need to have knowledge of the global constraint, while in [9,11] the agents keep local estimates of the global constraint function and multiplier and employ consensus to agree on those estimates. Since Consensus Saddle Point Dynamics can be viewed as an inexact dual method, this method suffers the same slow convergence rate of the dual method [9].

The Augmented Lagrangian method (ALM) [13] smoothes the dual function and, therefore, converges faster than the dual method. Recently, several distributed ALMs have been proposed [14–17]. Among these methods, the Accelerated Distributed Augmented Lagrangian (ADAL) method, first developed for convex optimization problems [16,17] and later extended to non-convex problems [18] and problems with noise [19], has been shown to converge faster than other distributed ALMs in practice. However, these methods either require separability of the equality constraint in (1), or require local knowledge of the global constraint function, as in the indirect method discussed in [20]. In this paper, we develop a consensus-based ADAL (C-ADAL) method that can be implemented without these requirements. By introducing local estimates of the global constraint function and multipliers and applying a finite number of consensus steps on these local estimates during every iteration, we show that C-ADAL converges at rate $O(1/k)$ and the final primal optimality and feasibility are controllable by the number of consensus steps. In numerical experiments, C-ADAL is demonstrated to outperform existing methods.

The rest of the paper is organized as follows. In Section II, we discuss assumptions on problem (1) and formally present the C-ADAL algorithm. In Section III, we analyze the convergence of the algorithm. In Section IV, we present comparative numerical experiments on a distributed estimation problem. In Section V, we conclude the paper.

II. PROBLEM FORMULATION

In this section, we first discuss some assumptions that are common to consensus-based algorithms; see, e.g., [1,2,5].

Yan Zhang and Michael M. Zavlanos are with the Department of Mechanical Engineering and Materials Science, Duke University, Durham, NC 27708, USA. {yan.zhang2,michael.zavlanos}@duke.edu This work is supported by ONR under grant #N000141410479.

Algorithm 1: ADAL

Require: Initialize the multiplier λ^0 and primal variable x^0. Set $k = 0$.

1: Agent i computes $\hat{x}_i^k = \arg\min_{x \in \mathcal{X}_i} \Lambda_\rho^i(x_i; x_{-i}^k, \lambda^k)$.
2: Agent i computes $x_i^{k+1} = x_i^k + \tau(\hat{x}_i^k - x_i^k)$.
3: Update the multiplier: $\lambda^{k+1} = \lambda^k + \tau\rho(\sum_i A_i x_i^{k+1} - b)$
4: $k \leftarrow k + 1$, go to step 1.

Assumption II.1. The local constraint set $\mathcal{X}_i \subset \mathbb{R}^{n_i}$ is convex and compact for all i. Moreover, there exist constants B_A and B_x such that for all $i = 1, \ldots, N$, we have

$$\|A_i\| \leq B_A \text{ and } \|x_p - x_q\| \leq B_x \text{ for all } x_p, x_q \in \mathcal{X}_i.$$

Assumption II.2. Problem (1) is feasible. That is, there exists at least one optimal solution $x^\star$.

In [16,17], ADAL is proposed to solve problem (1), which is presented in Algorithm 1. In step 2 of Algorithm 1, agent i minimizes the local objective function

$$\Lambda_\rho^i(x_i; x_{-i}^k) = f_i(x_i) + \langle \lambda_k, A_i x_i \rangle + \frac{\rho}{2}\|A_i x_i + \sum_{j \neq i} A_j x_j^k - b\|^2,$$

where x_{-i}^k denotes $\{x_j^k\}_{j \neq i}$. From the above definition of $\Lambda_\rho^i(x_i; x_{-i}^k)$, we see that distributed implementation of ADAL depends on separability of the equality constraints in problem (1). Specifically, if two agents are coupled in $\sum_{i=1}^N A_i x_i = b$, then they need to be connected in the communication network. To implement ADAL on more general network structures, we propose consensus-based ADAL, which is presented in Algorithm 2. In line 2 of Algorithm 2, agent i determines $\hat{x}_i^k$ by minimizing the local objective function

$$\begin{aligned} \Lambda_\rho^i(x_i; x_i^k, \tilde{y}_i^k, \tilde{\lambda}_i^k) = & f_i(x_i) + \langle \tilde{\lambda}_i^k, A_i x_i \rangle \\ & + \frac{\rho}{2}\|A_i x_i + N\tilde{y}_i^k - A_i x_i^k - b\|^2, \end{aligned} \tag{3}$$

where C-ADAL uses $\tilde{\lambda}_i^k$ and $N\tilde{y}_i^k$ as the current estimates of the global multiplier and $\sum_{i=1}^N A_i x_i^k$. If at every iteration, all agents reach consensus on $\tilde{\lambda}_i^k$ and $N\tilde{y}_i^k = \sum_{i=1}^N A_i x_i^k$, then (3) reduces to the local objective in ADAL. However, this is only possible if we run infinite consensus steps in (4a). In practice, we run α consensus steps. In the following section, we analyze the convergence of C-ADAL in this case.

Algorithm 2: C-ADAL

Require: Initialize the local multiplier λ_i^0, and primal variable $x_i^0 \in \mathcal{X}_i$. Set $y_i^0 = A_i x_i^0$ and $k = 0$.

1: Agent i communicates with its neighbors and runs α consensus steps on the variables λ_i^k and y_i^k, i.e.,

$$\tilde{\lambda}_i^k = \sum_j [W(k)^\alpha]_{ij} \lambda_j^k, \ \tilde{y}_i^k = \sum_j [W(k)^\alpha]_{ij} y_j^k. \tag{4a}$$

2: Agent i computes $\hat{x}_i^k = \arg\min_{x \in \mathcal{X}_i} \Lambda_\rho^i(x_i; x_i^k, \tilde{y}_i^k, \tilde{\lambda}_i^k)$.
3: Agent i computes $x_i^{k+1} = x_i^k + \tau(\hat{x}_i^k - x_i^k)$.
4: Agent i updates the variables y_i^k and λ_i^k by

$$\begin{aligned} y_i^{k+1} &= \tilde{y}_i^k + A_i x_i^{k+1} - A_i x_i^k, \\ \lambda_i^{k+1} &= \tilde{\lambda}_i^k + \tau\rho(N y_i^{k+1} - b). \end{aligned} \tag{4b}$$

5: $k \leftarrow k + 1$, go to step 1.

III. Convergence Analysis

At each iteration k, let $\lambda_a^k = \frac{1}{N}\sum_i \lambda_i^k$ and $y_a^k = \frac{1}{N}\sum_i y_i^k$ denote the global averages of the corresponding local variables. These variables are not accessible by any local agents, but are simply introduced to facilitate the analysis. In this section, we first show how λ_a^k and y_a^k evolve in C-ADAL. Next, we show that the disagreement errors between the local variables and the global averages can be made arbitrarily small by choosing a large enough number of consensus steps α. Finally, we analyze how the disagreement errors affect the primal optimality and feasiblity of the global solution.

A. *Evolution of λ_a^k and y_a^k*

Let $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ denote the network of agents, where $\mathcal{V}$ is the index set of vertices $\{1, \ldots, N\}$ and $\mathcal{E} \subseteq \mathcal{V} \times \mathcal{V}$ is the set of edges. We assume that the graph $\mathcal{G}$ is fixed and directed. An edge $(i, j) \in \mathcal{E}$ means that node i can receive information from node j. We assign weight W_{ij} to the edge (i, j) so that $W_{ij} > 0$ if $(i, j) \in \mathcal{E}$ and $W_{ij} = 0$ otherwise. We also define the weight matrix W, where W_{ij} denotes its (i, j)th entry. In addition, we make the following assumption:

Assumption III.1. W is doubly stochastic. That is, $W_{ij} \geq 0$, for $\forall i, j$, $W\mathbf{1} = \mathbf{1}$ and $\mathbf{1}^T W = \mathbf{1}^T$, where $\mathbf{1}$ is a column vector with all entries equal to 1. Furthermore, we assume that there exists a $\beta > 0$ such that, $\|W - \frac{\mathbf{1}\mathbf{1}^T}{N}\| \leq \beta < 1$.

Rather than assuming that the network is undirected and the spectral radius $\rho(W - \frac{\mathbf{1}\mathbf{1}^T}{N}) \leq \gamma < 1$ as in [1,5], here we assume a directed graph with the spectral norm $\|W - \frac{\mathbf{1}\mathbf{1}^T}{N}\| \leq \beta < 1$. This condition implies that $\rho(W - \frac{\mathbf{1}\mathbf{1}^T}{N}) \leq \gamma < 1$, and suggests that W is irreducible and the underlying directed graph is strongly connected, [21]. It is simple to see that under Assumption III.1, Lemma 2 in [1] still holds with γ replaced by β defined here and, furthermore, Lemma 3 in [1] also holds. These results are used in our analysis. Using Assumption III.1, we can show the following lemma:

Lemma III.2. Let Assumption III.1 hold. We have that

$$\lambda_a^k = \frac{1}{N}\sum_i \tilde{\lambda}_i^k \quad \text{and} \quad y_a^k = \frac{1}{N}\sum_i \tilde{y}_i^k. \tag{5}$$

Proof. Recalling that $\lambda_a^k = \frac{1}{N}\sum_i \lambda_i^k$ and the update in (4a), to show the first equality in (5), it suffices to show that

$$\frac{1}{N}\sum_i \lambda_i^k = \frac{1}{N}\sum_i \tilde{\lambda}_i^k = \frac{1}{N}\sum_i \sum_j [W(k)^\alpha]_{ij} \lambda_j^k. \tag{6}$$

To show equation (6), we first prove that under Assumption III.1, $W(k)^\alpha$ is a doubly stochastic matrix, for $\forall \alpha = 1, 2, \ldots$. Since we have that $W\mathbf{1} = \mathbf{1}$, we can also obtain that $W^\alpha \mathbf{1} = W^{\alpha-1}W\mathbf{1} = W^{\alpha-1}\mathbf{1}$. Iterating results in $W^\alpha \mathbf{1} = \mathbf{1}$. Showing that $\mathbf{1}^T W^\alpha = \mathbf{1}^T$ is similar. Meanwhile, it is straightforward to see that since all entries in W are nonnegative, all entries in W^α are also nonnegative. Therefore, W^α is also doubly stochastic.

We can now show (6) by switching the order of the summations on the right hand side of (6) as $\frac{1}{N}\sum_i \sum_j [W(k)^\alpha]_{ij}\lambda_j^k = \frac{1}{N}\sum_j \lambda_j^k (\sum_i [W(k)^\alpha]_{ij}) = \frac{1}{N}\sum_j \lambda_j^k$. The final step follows from the fact that W^α is doubly stochastic. This proves the first equality in (5). The second equality in (5) can be shown in a similar way. □

Next, we show a conservation property on y_a^k.

Lemma III.3. Let Assumption III.1 hold. Then, we have that at each iteration k, $Ny_a^k = \sum_i A_i x_i^k$.

Proof. We show this lemma by mathematical induction. From the initialization of Algorithm 2, we have that $Ny_a^0 = \sum_i y_i^0 = \sum_i A_i x_i^0$. Therefore, the lemma holds for $k = 0$. Next assume that $Ny_a^k = \sum_{i=1}^N A_i x_i^k$ holds for $k \geq 0$. We show that $Ny_a^{k+1} = \sum_i A_i x_i^{k+1}$. We have that

$$Ny_a^{k+1} = \sum_i y_i^{k+1} = \sum_i [\tilde{y}_i^k + A_i(x_i^{k+1} - x_i^k)]. \tag{7}$$

The second equality is due to (4b). According to Lemma III.2 and the induction assumption, we have that $\sum_i \tilde{y}_i^k = Ny_a^k = \sum_{i=1}^N A_i x_i^k$. Combining this with (7), we obtain that

$$Ny_a^{k+1} = \sum_i A_i x_i^k + \sum_i (A_i(x_i^{k+1} - x_i^k)) = \sum_i A_i x_i^{k+1},$$

which completes the proof. □

For the purpose of analysis, we introduce an augmented multiplier $\bar{\lambda}^k = \lambda_a^k + \rho(1-\tau)r(x^k)$, where $r(x^k) = \sum_{i=1}^N A_i x_i^k - b$ is the residual of the constraint. Next, we present how λ_a^k and $\bar{\lambda}^k$ evolve in C-ADAL.

Lemma III.4. Let Assumption III.1 hold. Then we have that

$$\lambda_a^{k+1} = \lambda_a^k + \tau\rho r(x^{k+1}) \text{ and } \bar{\lambda}^{k+1} = \bar{\lambda}^k + \tau\rho r(\hat{x}^k). \tag{8}$$

Proof. First we show that $\lambda_a^{k+1} = \lambda_a^k + \tau\rho r(x^{k+1})$. Recalling the definition of λ_a^{k+1} and the update in (4b), we obtain that

$$\lambda_a^{k+1} = \frac{1}{N}\sum_i [\tilde{\lambda}_i^k + \tau\rho(Ny_i^{k+1} - b)]. \tag{9}$$

Meanwhile, using Lemma III.3, we have that $\frac{1}{N}\sum_i Ny_i^{k+1} = \sum_i y_i^{k+1} = Ny_a^{k+1} = \sum_i A_i x_i^{k+1}$. Substituting this in (9) we obtain that $\lambda_a^{k+1} = \lambda_a^k + \tau\rho(\sum_i A_i x_i^{k+1} - b)$. This proves the first equality in (8). To show the second equality in (8), recall the definition of $\bar{\lambda}^k$, and observe that

$$\begin{aligned} \bar{\lambda}^{k+1} &= \lambda_a^k + \tau\rho r(x^{k+1}) + \rho(1-\tau)r(x^{k+1}) \\ &= \lambda_a^k + \rho r(x^{k+1}). \end{aligned} \tag{10}$$

The first line is due to the first equality in (8). According to the line 3 in Algorithm 2, it is simple to see that $r(x^{k+1}) = (1-\tau)r(x^k) + \tau r(\hat{x}^k)$. Plugging this into (10), we obtain that $\bar{\lambda}^{k+1} = \lambda_a^k + \rho(1-\tau)r(x^k) + \tau\rho r(\hat{x}^k) = \bar{\lambda}^k + \tau\rho r(\hat{x}^k)$. This completes the proof. □

B. Consensus Error

In this section, we show that after a finite number of consensus iterations, the disagreement errors between the local variables y_i^k (respectively, λ_i^k) and the global average y_a^k (respectively, λ_a^k) always remain bounded. Furthermore, this bound can be made arbitrarily small by choosing a proper number of consensus steps α. For this, we need the following assumption on initialization of Algorithm 2.

Assumption III.5. Given a small positive value ϵ, for all i, it holds that $\|\lambda_i^0 - \lambda_a^0\| \leq \epsilon$ and $\|y_i^0 - y_a^0\| \leq \epsilon$.

To satisfy the Assumption III.5, we can run enough consensus steps to initialize the algorithm. Since such results are well-known, we refer the reader to [2,21,22]. A consequence of Assumption III.5 is that $\|\tilde{\lambda}_i^0 - \lambda_a^0\| \leq \epsilon$ and $\|\tilde{y}_i^0 - y_a^0\| \leq \epsilon$. This can be easily seen by the convexity of the $\|\cdot\|$ function. This result will be useful in the following analysis.

Next we study the boundness of the disagreement errors of y_i^k (or λ_i^k). The update of y_i^k (or λ_i^k) can be viewed as a consensus step with local perturbation $\eta_i^k = A_i x_i^{k+1} - A_i x_i^k$ (or $\eta_i^k = \tau\rho(Ny_i^{k+1} - b)$) at iteration k. According to Lemma 3 in [1], if $\|\eta_i^k\|_\infty \leq B$, $\|\tilde{y}_i^k - y_a^k\| \leq \epsilon$ for $\forall i$, and $\alpha \geq (\log(\epsilon) - \log(4\sqrt{N}\sqrt{m}(\epsilon + B))/\log(\beta)$, then we have that $\|\tilde{y}_i^{k+1} - y_a^{k+1}\| \leq \epsilon$ for all i ($\tilde{\lambda}_i^k$ is the same). Therefore, in order to upper bound the disagreement error, we show that η_i^k is bounded for all $k \geq 0$ in the next lemma.

Lemma III.6. Let Assumptions II.1 and III.1 hold. Then, for all i and $k \geq 0$, there exists a constant B_y that satisfies

$$\|A_i x_i^{k+1} - A_i x_i^k\|_\infty \leq B_y. \tag{11}$$

Meanwhile, for any $\epsilon > 0$, let α satisfy $\alpha \geq (\log(\epsilon) - \log(4\sqrt{N}\sqrt{m}(\epsilon + B_y))/\log(\beta)$. Then, there exists a constant B_λ that, for all i and $k \geq 0$, satisfies

$$\|\tau\rho(Ny_i^{k+1} - b)\|_\infty \leq B_\lambda. \tag{12}$$

Proof. First we show the bound in (11). For this, it suffices to show that $x_i^k \in \mathcal{X}_i$ for all $k \geq 0$. Then, because $\mathcal{X}_i$ is compact, it is straightforward to see the result. According to line 3 in Algorithm 2, we have that $x_i^{k+1} = (1-\tau)x_i^k + \tau\hat{x}_i^k$. Therefore, we need to show that $x_i^{k+1} \in \mathcal{X}_i$ if both x_i^k and $\hat{x}_i^k$ belong to $\mathcal{X}_i$. The latter condition is satisfied by line 2 in Algorithm 2, while the former one is satisfied when $k = 0$ at initialization of Algorithm 2. Therefore, it is simple to show that $x_i^k \in \mathcal{X}_i$ for all $k \geq 0$ using mathematical induction.

Next, we prove the bound in (12). We have that

$$\begin{aligned} &\|Ny_i^{k+1} - b\|_\infty = \|N(\tilde{y}_i^k + A_i(x_i^{k+1} - x_i^k)) - b\|_\infty \\ &\leq \|N\tilde{y}_i^k - b\|_\infty + N\|A_i(x_i^{k+1} - x_i^k)\|_\infty, \end{aligned} \tag{13}$$

where the inequality is due to the triangle inequality. The second term in (13) is upper bounded due to Assumption II.1. Meanwhile, we can expand the first term in (13) as $\|N\tilde{y}_i^k - b\|_\infty = \|N(\tilde{y}_i^k - y_a^k) + (Ny_a^k - b)\|_\infty \leq N\|\tilde{y}_i^k - y_a^k\|_\infty + \|Ny_a^k - b\|_\infty$, where the inequality is due to the triangle inequality. Choosing α specified in Lemma III.6, according to Lemma 3 in [1] and Lemma III.3, we can show the bound in (12). This completes the proof. □

The following theorem provides upper bounds on the disagreement errors of $\tilde{y}_i^k$ and $\tilde{\lambda}_i^k$:

Theorem III.7. *Let Assumptions III.1 and III.5 hold, and define $B_p = \max\{B_y, B_\lambda\}$. If we choose α to satisfy $\alpha \geq (\log(\epsilon) - \log(4\sqrt{N}\sqrt{m}(\epsilon + B_p))/\log(\beta)$, then for all i and $k \geq 0$, we have $\|\tilde{y}_i^k - y_a^k\|$, $\|\tilde{\lambda}_i^k - \lambda_a^k\| \leq \epsilon$.*

Proof. The proof follows from combining Lemma 3 in [1] and Lemma III.6 and using Assumption III.5. □

C. Primal optimality and feasibility

In this section, we use a Lyapunov function to obtain the convergence rate of Algorithm 2 as well as the primal optimality and feasibility of the global solution. Specifically, we define the Lyapunov function as

$$\phi^k(\lambda) = \rho \sum_i \|A_i(x_i^k - x_i^\star)\|^2 + \frac{1}{\rho}\|\bar{\lambda}^k - \lambda\|^2. \tag{14}$$

The following lemma characterizes the decrease of the Lyapunov function at every iteration k.

Lemma III.8. Let Assumption III.1 hold and define $q = \max\{q_1, q_2, \dots, q_m\}$, where q_l is the degree of the l-th constraint, that is the number of agents involved in the l-th row of A_i. Moreover, let $\tau \in (1, \frac{1}{q})$ and assume that α satisfies the condition in Theorem III.7. Then, for any $\lambda \in \mathbb{R}^m$ and $k \geq 0$, we have that

$$F(\hat{x}^k) - F(x^\star) + \langle \lambda, r(\hat{x}^k)\rangle \leq \frac{1}{2\tau}(\phi^k(\lambda) - \phi^{k+1}(\lambda)) + C\epsilon, \tag{15}$$

where $C = N(1 + \rho N)B_A B_x$.

Proof. From the first order optimality condition of $\hat{x}_i^k$, we have that there exists an $s_i^k \in \partial f_i(\hat{x}_i^k)$,[1] so that $\langle s_i^k + A_i\tilde{\lambda}_i^k + \rho A_i^T(A_i\hat{x}_i^k - A_i x_i^k + N\tilde{y}_i^k - b), x_i^\star - \hat{x}_i^k\rangle \geq 0$. Since $f_i(x)$ is convex, we have that $f_i(x_i^\star) - f_i(\hat{x}_i^k) \geq \langle s_i^k, x_i^\star - \hat{x}_i^k\rangle$. Combining this with the optimality condition and rearranging terms, we obtain

$$\begin{aligned} &f_i(x_i^\star) - f_i(\hat{x}_i^k) + \langle \tilde{\lambda}_i^k, A_i(x_i^\star - \hat{x}_i^k)\rangle \\ &+ \rho\langle A_i\hat{x}_i^k - A_i x_i^k + N\tilde{y}_i^k - b, A_i(x_i^\star - \hat{x}_i^k)\rangle \geq 0. \end{aligned} \tag{16}$$

Next, we substitute the term $\tilde{\lambda}_i^k$ in (16) with $\lambda_a^k + (\tilde{\lambda}_i^k - \lambda_a^k)$ and the term $-A_i x_i^k + N\tilde{y}_i^k$ with $N(\tilde{y}_i^k - y_a^k) + \sum_{j\neq i} A_j x_j^k$. The second substitution is due to the fact that $-A_i x_i^k + N\tilde{y}_i^k = -A_i x_i^k + N y_a^k + N(\tilde{y}_i^k - y_a^k)$ and Lemma III.3. Therefore, inequality (16) becomes $f_i(x_i^\star) - f_i(\hat{x}_i^k) + \langle \lambda_a^k, A_i(x_i^\star - \hat{x}_i^k)\rangle + \rho\langle A_i\hat{x}_i^k + \sum_{j\neq i} A_j x_j^k - b, A_i(x_i^\star - \hat{x}_i^k)\rangle \geq -\langle \tilde{\lambda}_i^k - \lambda_a^k, A_i(x_i^\star - \hat{x}_i^k)\rangle - \rho N\langle \tilde{y}_i^k - y_a^k, A_i(x_i^\star - \hat{x}_i^k)\rangle$. Applying the Cauchy-Schwartz inequality, and using Assumption II.1 and Theorem III.7, we can lower bound the left hand side (LHS) of above inequality as LHS $\geq -B_A B_x(1 + \rho N)\epsilon$. Summing this inequality over all agents i, we obtain

$$\begin{aligned} &F(x^\star) - F(\hat{x}^k) + \langle \lambda_a^k, \sum_i A_i(x_i^\star - \hat{x}_i^k)\rangle + \rho\sum_i\langle A_i\hat{x}_i^k - b \\ &+ \sum_{j\neq i} A_j x_j^k, A_i(x_i^\star - \hat{x}_i^k)\rangle \geq -C\epsilon \end{aligned} \tag{17}$$

[1] $\partial f(x)$ is the subdifferential of the nonsmooth function $f(x)$ at x and $s \in \partial f(x)$ is the subgradient. For more details see [13].

where $C = NB_A B_x(1 + \rho N)$. Adding and subtracting $\sum_{j\neq i} A_j\hat{x}_j^k$ to the term $A_i\hat{x}_i^k + \sum_{j\neq i} A_j x_j^k - b$ on the left hand side of (17), adding $\langle \lambda, r(\hat{x}^k)\rangle$ to both sides of this inequality, recalling that $\sum_i A_i x_i^\star = b$ and $r(x) = \sum_i A_i x_i - b$, and rearranging terms, we have that

$$\begin{aligned} &\langle \lambda - \lambda_a^k, r(\hat{x}^k)\rangle - \rho\|r(\hat{x}^k)\|^2 + \rho\sum_i\langle\sum_{j\neq i} A_j x_j^k - A_j\hat{x}_j^k, \\ &A_i(x_i^\star - \hat{x}_i^k)\rangle + C\epsilon \geq F(\hat{x}^k) - F(x^\star) + \langle \lambda, r(\hat{x}^k)\rangle. \end{aligned} \tag{18}$$

Therefore, to show (15), it suffices to prove that the left hand side in (18) is upper bounded by $\frac{1}{2\tau}(\phi^k(\lambda) - \phi^{k+1}(\lambda)) + C\epsilon$. To show this, expanding $\phi^k(\lambda) - \phi^{k+1}(\lambda)$, applying Lemma III.4, and dividing both sides by 2τ, we have that

$$\begin{aligned} &\frac{1}{2\tau}(\phi^k(\lambda) - \phi^{k+1}(\lambda)) = \rho\sum_i\langle A_i(x_i^k - x_i^\star), A_i(x_i^k - \hat{x}_i^k)\rangle \\ &- \langle \bar{\lambda}^k - \lambda, r(\hat{x}^k)\rangle - \frac{\tau\rho}{2}(\sum_i\|A_i(\hat{x}_i^k - x_i^k)\|^2 + \|r(\hat{x}^k)\|^2). \end{aligned} \tag{19}$$

Next we manipulate the left hand side in (18) so that it can be related to $\frac{1}{2\tau}(\phi^k(\lambda) - \phi^{k+1}(\lambda))$ in (19). Specifically, adding and subtracting $-(1-\tau)\rho\langle r(x^k), r(\hat{x}^k)\rangle$ to the term $\langle \lambda - \lambda_a^k, r(\hat{x}^k)\rangle$, we have that $\langle \lambda - \lambda_a^k, r(\hat{x}^k)\rangle$ can be replaced by

$$\langle \lambda - \bar{\lambda}^k, r(\hat{x}^k)\rangle + (1-\tau)\rho\langle r(x^k), r(\hat{x}^k)\rangle. \tag{20}$$

Adding and subtracting $\rho\sum_i\langle A_i x_i^k - A_i\hat{x}_i^k, A_i(x_i^\star - \hat{x}_i^k)\rangle$ to the third term of (18) and recalling that $r(x) = \sum_i A_i x_i - b$, we can replace the third term in (18) with

$$\begin{aligned} &-\rho\langle r(x^k) - r(\hat{x}^k), r(\hat{x}^k)\rangle - \rho[\sum_i\langle A_i(x_i^k - \hat{x}_i^k), \\ &A_i(x_i^\star - x_i^k)\rangle + \sum_i\|A_i(x_i^k - \hat{x}_i^k)\|^2]. \end{aligned} \tag{21}$$

Substituting (20) and (21) into (18) and rearranging terms, we have that $\langle \lambda - \bar{\lambda}^k, r(\hat{x}^k)\rangle - \tau\rho\langle r(\hat{x}^k), r(x^k) - r(\hat{x}^k)\rangle + (1-\tau)\rho\|r(\hat{x}^k)\|^2 + \rho\sum_i\langle A_i(x_i^k - \hat{x}_i^k), A_i(x_i^k - x_i^\star)\rangle - \rho\sum_i\|A_i(x_i^k - \hat{x}_i^k)\|^2 - \rho\|r(\hat{x}^k)\|^2 + C\epsilon \geq F(\hat{x}^k) - F(x^\star) + \langle \lambda, r(\hat{x}^k)\rangle$. Recalling the bound (35) in [17], we have that $-\tau\rho\langle r(\hat{x}^k),\ r(x^k) - r(\hat{x}^k)\rangle \leq \frac{\rho}{2}\sum_i\|A_i(x_i^k - \hat{x}_i^k)\|^2 + \frac{\tau^2\rho q}{2}\|r(\hat{x}^k)\|^2$. Therefore, we obtain that

$$\begin{aligned} &\langle \lambda - \bar{\lambda}^k, r(\hat{x}^k)\rangle + \rho\sum_i\langle A_i(x_i^k - \hat{x}_i^k), A_i(x_i^k - x_i^\star)\rangle - \frac{\rho}{2} \\ &\sum_i\|A_i(x_i^k - \hat{x}_i^k)\|^2 - \tau\rho(1 - \frac{\tau q}{2})\|r(\hat{x}^k)\|^2 + C\epsilon \\ &\geq F(\hat{x}^k) - F(x^\star) + \langle \lambda, r(\hat{x}^k)\rangle \end{aligned} \tag{22}$$

Finally, to show (15), we upper bound the left hand side in (22) by $\frac{1}{2\tau}(\phi^k(\lambda) - \phi^{k+1}(\lambda)) + C\epsilon$. Recalling the expression for $\frac{1}{2\tau}(\phi^k(\lambda) - \phi^{k+1}(\lambda))$ in (19), it is straghtforward to see that this upper bound holds when $\frac{\tau\rho}{2} \leq \frac{\rho}{2}$ and $\frac{\tau\rho}{2} \leq \tau\rho(1 - \frac{\tau q}{2})$. Since $\tau \in (0, \frac{1}{q})$, $q \geq 1$, both conditions above are satisfied. This completes the proof. □

Denote the global running average by $\tilde{x}^K = \frac{1}{K}\sum_{k=0}^{K-1}\hat{x}^k$. The following theorem provides bounds on primal optimality and feasibility of $\tilde{x}^K$.

Theorem III.9. *Let Assumptions II.1, II.2, III.1 and III.5 hold and let α satisfy the condition in Theorem III.7. Then, we have the following bound on primal optimality of $\tilde{x}^K$:*

$$-(\frac{1}{2\tau K}\phi^0(2\lambda^\star)+C\epsilon) \le F(\tilde{x}^K)-F(x^\star) \le \frac{1}{2\tau K}\phi^0(0)+C\epsilon. \quad (23)$$

Moreover, we have the following bound on primal feasibility:

$$\begin{aligned}\|r(\tilde{x}^K)\| \le & \frac{1}{2\tau K}[\rho \sum_i \|A_i(x_i^0 - x_i^\star)\|^2 \\ & + \frac{2}{\rho}(\|\bar{\lambda}^0 - \lambda^\star\|^2 + 1)] + C\epsilon.\end{aligned} \quad (24)$$

Proof. Summing both sides of (15) in Lemma III.8 for $k = 0, 1, \ldots, K-1$ and dividing both sides by K, we obtain $\frac{1}{K}\sum_k F(\hat{x}^k) - F(x^\star) + \langle \lambda, r(\tilde{x}^K)\rangle \le \frac{1}{2\tau}(\phi^0(\lambda) - \phi^K(\lambda)) + C\epsilon$. Since $\phi^K(\lambda) \ge 0$, we can neglect this term. Using $\frac{1}{K}\sum_k F(\hat{x}^k) \ge F(\tilde{x}^K)$ from convexity of the function $F(x)$, we have that

$$F(\tilde{x}^K) - F(x^\star) + \langle \lambda, r(\tilde{x}^K)\rangle \le \frac{1}{2\tau K}\phi^0(\lambda) + C\epsilon. \quad (25)$$

Next we use (25) to show the bound (23) and (24). To show the upper bound in (23), since $\lambda \in \mathbb{R}^m$ in (25), we select $\lambda = 0$ and obtain $F(\tilde{x}^K) - F(x^\star) \le \frac{1}{2\tau K}\phi^0(0) + C\epsilon$. To prove the lower bound in (23), we recall that $(x^\star, \lambda^\star)$ is a saddle point that satisfies $F(x^\star) \le F(\tilde{x}^K) + \langle \lambda^\star, r(\tilde{x}^K)\rangle$. Adding $\langle \lambda^\star, r(\tilde{x}^K)\rangle$ to both sides of the saddle point inequality and rearranging terms, we get

$$\langle \lambda^\star, r(\tilde{x}^K)\rangle \le F(\tilde{x}^K) - F(x^\star) + \langle 2\lambda^\star, r(\tilde{x}^K)\rangle. \quad (26)$$

The right hand side in (26) can be upper bounded by selecting $\lambda = 2\lambda^\star$ in (25). Specifically, we have that $F(x^\star) - F(\tilde{x}^K) \le \langle \lambda^\star, r(\tilde{x}^K)\rangle \le \frac{1}{2\tau K}\phi^0(2\lambda^\star) + C\epsilon$, which completes the proof of (23).

Next we show the bound (24) on primal feasibility. Letting $\lambda = \lambda^\star + \frac{r(\tilde{x}^K)}{\|r(\tilde{x}^K)\|}$ in (25), we have that

$$\begin{aligned}& F(\tilde{x}^K) - F(x^\star) + \langle \lambda^\star, r(\tilde{x}^K)\rangle + \|r(\tilde{x}^K)\| \\ & \le \frac{1}{2\tau K}\phi^0(\lambda^\star + \frac{r(\tilde{x}^K)}{\|r(\tilde{x}^K)\|}) + C\epsilon.\end{aligned} \quad (27)$$

Since $(x^\star, \lambda^\star)$ is the saddle point, we have that $F(\tilde{x}^K) - F(x^\star) + \langle \lambda^\star, r(\tilde{x}^K)\rangle \ge 0$ and, therefore, these terms can be neglected. This gives $\|r(\tilde{x}^K)\| \le \frac{1}{2\tau K}\phi^0(\lambda^\star + \frac{r(\tilde{x}^K)}{\|r(\tilde{x}^K)\|}) + C\epsilon$. Plugging the expression of $\phi^k(\lambda)$ in (14) into this upper bound completes the proof. □

The error terms in Theorem III.9 are due to the disagreement errors $\|\tilde{\lambda}_i^k - \lambda_a^k\|$ and $\|\tilde{y}_i^k - y_a^k\|$. Similar to the results in Lemma 3 in [1], as long as the perturbation in (4b) is uniformly bounded, it is straightforward to make the error term in Theorem III.9 uniformly bounded by an arbitrarily small value by choosing α large enough. Exponential convergence of consensus under time-varying digraphs with doubly stochastic weight matrices has been proved in [2,22]. Therefore, our results can be easily modified to fit into the time-varying network topology in [2,22].

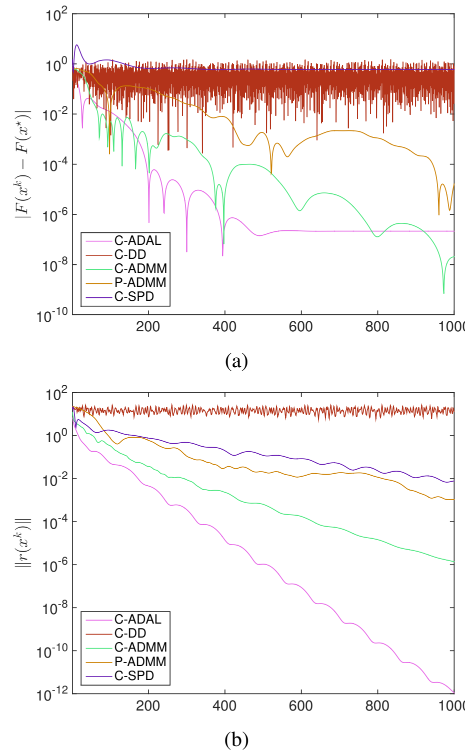

Fig. 1: Primal optimality (a) and feasibility (b) of the direct iterate of C-ADAL (magenta), C-DD (brown), C-SPD (purple), C-ADMM (green) and P-ADMM (yellow).

IV. Numerical Simulation

In this section, we conduct numerical experiments to verify our theoretical analysis and compare the performance of C-ADAL to other existing methods. Specifically, we consider a network of N agents and apply C-ADAL to solve the following distributed estimation problem:

$$\begin{aligned}\min \quad & \sum_i \|M_i x_i - y_i\|^2 \\ \text{s.t.} \quad & \sum_i A_i x_i = b \text{ and } l_i \le x_i \le u_i, \text{for all } i,\end{aligned} \quad (28)$$

where $M_i \in \mathbb{R}^{n_i \times p}$ is the design matrix, $y_i \in \mathbb{R}^{n_i}$ is the observation vector, $x_i \in \mathbb{R}^p$, $A_i \in \mathbb{R}^{m \times p}$ and $b \in \mathbb{R}^m$.

To demonstrate the performance of C-ADAL, we compare it with consensus Dual Decomposition (C-DD) in [5], consensus Saddle Point Dynamics (C-SPD) in [9], consensus ADMM (C-ADMM) in [6], and an indirect version of ADMM (P-ADMM) in [15]. C-ADMM and P-ADMM can only be implemented on undirected graphs, while C-ADAL, C-DD, and C-SPD can be applied to directed graph. We conduct simulations on an undirected graph, so that C-ADAL can be compared with C-ADMM and P-ADMM. We study the case where we have $N = 10$ agents, $n_i = 5$, $p = 10$, and $m = 20$. The problem data are randomly generated. A chain graph is applied and the weight matrix W is determined as in [21]. According to Theorem III.7, if $\epsilon = 0.1$, then $\alpha \approx 300$. However, in numerical experiments we have observed that C-ADAL behaves well for much smaller values of α for most randomly generated problems. Therefore, we select $\alpha = 10$

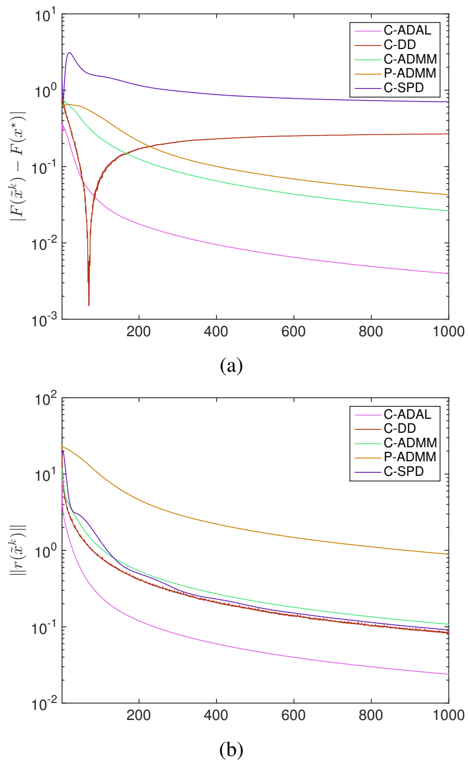

Fig. 2: Primal optimality (a) and feasibility (b) of the running average of C-ADAL (magenta), C-DD (brown), C-SPD (purple), C-ADMM (green) and P-ADMM (yellow).

for C-ADAL, C-DD, and C-SPD. The other parameters in these methods are optimized by trials. The primal optimality and feasibility for the direct iterate (or running average) are presented in Figures 1 (or Figure 2).

In Figure 1, the direct iterate of C-DD oscillates because this algorithm applies a subgradient method to maximize a nonsmooth dual function of (28). C-ADAL and C-ADMM outperform other methods in terms of primal optimality, while C-ADAL outperforms other methods in terms of primal feasibility. In Figure 2, we see that the running average of all methods converge. Specifically, in Figure 2(a), we observe that C-DD and C-SPD have relatively large errors due to the constant stepsize, [1,2,5]. C-ADAL outperforms other methods both in terms of primal optimality and feasibility.

V. CONCLUSIONS

In this paper, we proposed a distributed algorithm to solve multi-agent constrained optimization problems. Specifically, we employed the recently developed Accelerated Distributed Augmented Lagrangian (ADAL) algorithm that has been shown to exhibit faster convergence rates in practice compared to relevant distributed methods. Distributed implementation of ADAL depends on separability of the global coupling constraints. Here we extended ADAL so that it can be implemented distributedly independent of the structure of the coupling constraints. Our proposed algorithm can be applied to both undirected and directed networks. We showed that our algorithm converges at rate $O(1/k)$ and has steady error that is controllable by the number of consensus steps. Moreover, we provided numerical simulations showing that our algorithm outperforms existing methods in practice.

REFERENCES

[1] B. Johansson, T. Keviczky, M. Johansson, and K. H. Johansson, "Subgradient methods and consensus algorithms for solving convex optimization problems," in *Decision and Control, 2008. CDC 2008. 47th IEEE Conference on*. IEEE, 2008, pp. 4185–4190.

[2] A. Nedic, A. Ozdaglar, and P. A. Parrilo, "Constrained consensus and optimization in multi-agent networks," *IEEE Transactions on Automatic Control*, vol. 55, no. 4, pp. 922–938, 2010.

[3] J. C. Duchi, A. Agarwal, and M. J. Wainwright, "Dual averaging for distributed optimization: Convergence analysis and network scaling," *IEEE Transactions on Automatic control*, vol. 57, no. 3, pp. 592–606, 2012.

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

[5] A. Simonetto and H. Jamali-Rad, "Primal recovery from consensus-based dual decomposition for distributed convex optimization," *Journal of Optimization Theory and Applications*, vol. 168, no. 1, pp. 172–197, 2016.

[6] T.-H. Chang, M. Hong, and X. Wang, "Multi-agent distributed optimization via inexact consensus admm," *IEEE Transactions on Signal Processing*, vol. 63, no. 2, pp. 482–497, 2015.

[7] T.-H. Chang, "A proximal dual consensus admm method for multi-agent constrained optimization," *IEEE Transactions on Signal Processing*, vol. 64, no. 14, pp. 3719–3734, 2016.

[8] M. Zhu and S. Martínez, "On distributed convex optimization under inequality and equality constraints," *IEEE Transactions on Automatic Control*, vol. 57, no. 1, pp. 151–164, 2012.

[9] D. Mateos-Núnez and J. Cortés, "Distributed subgradient methods for saddle-point problems," in *Decision and Control (CDC), 2015 IEEE 54th Annual Conference on*. IEEE, 2015, pp. 5462–5467.

[10] S. Yang, Q. Liu, and J. Wang, "A multi-agent system with a proportional-integral protocol for distributed constrained optimization," *IEEE Transactions on Automatic Control*, vol. 62, no. 7, pp. 3461–3467, 2017.

[11] T.-H. Chang, A. Nedić, and A. Scaglione, "Distributed constrained optimization by consensus-based primal-dual perturbation method," *IEEE Transactions on Automatic Control*, vol. 59, no. 6, pp. 1524–1538, 2014.

[12] D. Yuan, D. W. Ho, and S. Xu, "Regularized primal–dual subgradient method for distributed constrained optimization," *IEEE transactions on cybernetics*, vol. 46, no. 9, pp. 2109–2118, 2016.

[13] D. P. Bertsekas, *Nonlinear programming*. Athena scientific Belmont, 1999.

[14] J. M. Mulvey and A. Ruszczyński, "A diagonal quadratic approximation method for large scale linear programs," *Operations Research Letters*, vol. 12, no. 4, pp. 205–215, 1992.

[15] J. Eckstein, "The alternating step method for monotropic programming on the connection machine cm-2," *ORSA Journal on Computing*, vol. 5, no. 1, pp. 84–96, 1993.

[16] N. Chatzipanagiotis, D. Dentcheva, and M. M. Zavlanos, "An augmented lagrangian method for distributed optimization," *Mathematical Programming*, vol. 152, no. 1-2, pp. 405–434, 2015.

[17] S. Lee, N. Chatzipanagiotis, and M. M. Zavlanos, "Complexity certification of a distributed augmented lagrangian method," *IEEE Transactions on Automatic Control*, vol. 63, no. 3, pp. 827–834, 2018.

[18] N. Chatzipanagiotis and M. M. Zavlanos, "On the convergence of a distributed augmented lagrangian method for nonconvex optimization," *IEEE Transactions on Automatic Control*, vol. 62, no. 9, pp. 4405–4420, 2017.

[19] ——, "A distributed algorithm for convex constrained optimization under noise," *IEEE Transactions on Automatic Control*, vol. 61, no. 9, pp. 2496–2511, 2016.

[20] 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, 2014.

[21] L. Xiao and S. Boyd, "Fast linear iterations for distributed averaging," *Systems & Control Letters*, vol. 53, no. 1, pp. 65–78, 2004.

[22] A. Nedic, A. Olshevsky, A. Ozdaglar, and J. N. Tsitsiklis, "On distributed averaging algorithms and quantization effects," *IEEE Transactions on Automatic Control*, vol. 54, no. 11, pp. 2506–2517, 2009.