Online Distributed ADMM via Dual Averaging

Saghar Hosseini, Airlie Chapman, and Mehran Mesbahi

***Abstract*—This paper presents a convergence analysis on the distributed Alternating Direction Method of Multipliers (ADMM) algorithm which solves online convex optimization problems under affine constraints. The goal is to distributively optimize a global objective function over a network of agents. The global objective function is composed of convex cost functions associated with each agent. The local cost functions can be broken down into two convex functions, one of which is revealed over time due to uncertainties. We extend an online ADMM algorithm to a distributed setting based on dual-averaging. We examine the rate of convergence of the performance of the sequence of decisions generated by the algorithm to the best fixed decision in hindsight. This performance metric is called the regret of online algorithms. An upper bound on the regret of the proposed algorithm is presented as a function of the underlying network topology and linear constraints. The online distributed ADMM algorithm was applied to a formation acquisition problem.**

***Index Terms*—Online Optimization; Distributed Algorithms; ADMM; Dual-averaging; Formation Algorithms**

I. Introduction

Many problems in engineering and information science applications can be characterized as a distributed convex optimization over networks such as multi-agent coordination, distributed estimation in sensor networks, and decentralized tracking and event localization. These problems often have a composite objective function subject to local linear constraints. A well known algorithm, Alternating Direction Method of Multipliers (ADMM), can solve this class of problems by dividing the problem into two sub-problems:

$$\min_{x\in\chi, y\in Y} f(x)+\phi(y),\ \text{s.t. } Ax+By=c, \tag{1}$$

where functions f and ϕ are convex functions, and χ and Y are convex set. ADMM has been widely used to solve the consensus problem.

The aforementioned problem is an *offline* optimization problem, where the cost function is known *a priori*. However, when the relevant decision is made, one part of the cost function $f_t(x)$ might be arbitrarily varying with time t, for example due to unknown uncertainties in the environment. Furthermore, the uncertainty in $f_t(x)$ may not be characterized by a known probability distribution. Stochastic and online ADMM (O-ADMM) approaches have consequently been proposed to address this scenario which can be posed as the following optimization problem at time T:

$$\min_{x\in\chi, y\in Y} \sum_{t=1}^{T}(f_t(x)+\phi(y)),\ \text{s.t. } Ax+By=c. \tag{2}$$

The research of the authors was supported by the ONR grant N00014-12-1-1002 and AFOSR grant FA9550-12-1-0203-DEF. The authors are with the Department of Aeronautics and Astronautics, University of Washington, WA 98105. Emails: {saghar, airliec, mesbahi}@uw.edu.

For this class of problem, stochastic ADMM has been introduced by Oyang *et al.* in [1] and considers an identical and independent distribution for the uncertainties in function f_t and provides a convergence rate of $O(\frac{1}{\sqrt{T}})$ for convex functions. O-ADMM algorithms in [2], [3] are also able to provide the same rate of convergence with no assumptions placed on the distribution of uncertainties.

Another variation of problem (1) is to consider its *distributed* analog, namely:

$$\min_{x\in\chi, y_i\in Y} \sum_{i=1}^{n}(f_i(x)+\phi_i(y_i)),\ \text{s.t. } A_ix+B_iy_i=c_i,\ \forall i, \tag{3}$$

and a *distributed* implementation solution involving n agents cooperatively solving for x and y. Here, the functions that compose problem (3) are distributed, specifically only agent i has access to the functions $f_i(x)$ and $\phi_i(y)$. A special case of this formulation is the ADMM form of the consensus problem where global agreement is required on each agent's own y_i

$$\min_{x\in\chi, y_i\in Y} \sum_{i=1}^{n}\phi_i(y_i),\ \text{s.t. } x=y_i,\ \forall i. \tag{4}$$

A constraint set can also be formed to represent the coupling among agents imposed by the underlying network topology. In [4], a sequential distributed ADMM was introduced to solve (4), where each agent takes turns to update its local variable. Wei and Ozdaglar [5] also proposed a stochastic asynchronous edge based ADMM algorithm to solve this problem. We observe that the objective in (4) is a local formulation of the global objective in (3) and hence is more easily achieved distributively when, during an algorithm's convergence, the constraint set is not satisfied. In other words, each agent is penalized by its local cost rather than the global cost. This provides a rapid $O(\frac{1}{T})$ convergence rate. This problem has also been examined as a gradient based distributed optimization [6], [7], [8], where under the global objective (3) and local objective (4) the rate of convergence is $O(\frac{1}{\sqrt{T}})$ and $O(\frac{1}{T})$, respectively.

In addition, Mota *et al.* in [9], [10] have studied the consensus problem over connected bipartite graphs based on distributed ADMM. Using quantitative analysis, they have shown that this algorithm requires less communication between nodes than previous algorithms to achieve a given accuracy. In [11], Deng *et al.* have proposed Proximal Jacobian ADMM, which is suitable for parallel computing. However, this method requires an all-to-all communication at each iteration.

Fusing the online and distributed ADMM problem we examine the *online distributed* ADMM (OD-ADMM) problem at time T:

$$\min_{x\in\chi, y\in Y} \sum_{t=1}^{T}(\sum_{i=1}^{n}(f_{i,t}(x)+\phi_i(y_i))), \text{ s.t. } A_i x + B_i y_i = c_i \,\forall i.$$

Inspired by the O-ADMM and our previous work [8], a single loop OD-ADMM based on dual-averaging is proposed in this paper. We consider an online convex optimization over a network of agents where each agent has two sets of variables. Similar to distributed ADMM (D-ADMM) literature, the agents are required to reach agreement on the global variable set. However, each agent keeps a private set of variables satisfying a local linear constraint which presents a relation between the global and local variables. The cost function associated with the global variables is revealed to the decision maker after committing to a decision. Moreover, the cost function associated with the local variables is known *a priori*. The rate of convergence of this online algorithm is $O(\frac{1}{\sqrt{T}})$ and is called *regret*. Regret is a standard metric that measures the difference between the incurred cost and the cost of the best fixed decision in hindsight. The average regret of a "good" online algorithm is sub-linear in time.

The outline of the paper is as follows. In §II, the notation and a brief background on graphs and regret is presented. The optimization problem formulation and the performance measure are stated in §III followed by the description of the algorithm and regret analysis in §IV. Then, in §V a distributed formation acquisition problem is solved based on the algorithm, and the simulation results are presented to reinforce the analysis. Finally, concluding remarks are provided in §VI.

II. Background and Preliminaries

In this section, we briefly review graphs and provide a background on online algorithms and their underlying assumptions.

Element v_i or $[v]_i$ denotes the ith element of a column vector $v \in \mathbb{R}^p$. A unit vector e_i denotes the column vector which contains all zero entries except $[e_i]_i = 1$. The vector of all ones will be denoted by $\mathbf{1}$. For matrix $M \in \mathbb{R}^{p\times q}$, $[M]_{ij}$ denotes the element in its ith row and jth column. A doubly stochastic matrix P is a non-negative matrix with $\sum_{i=1}^{n} P_{ij} = 1$ and $\sum_{j=1}^{n} P_{ij} = 1$. For any positive integer n, the set $\{1, 2, ..., n\}$ is presented by $[n]$. The 2-norm and ∞-norm are denoted by $||.||_2$ and $||.||_\infty$, respectively, and the dual norm of a vector u is defined as $||u||_* = \sup_{||v||=1} \langle u, v\rangle$. For a positive definite matrix Q, let $||x||_Q = \sqrt{\langle Qx, x\rangle}$. The largest and second largest singular values of Q are denoted by $\sigma_1(Q)$ and $\sigma_2(Q)$, respectively. The singular values $\sigma(Q)$ are the square roots of the eigenvalues of Q^TQ. The running average of a variable x_t over time is denoted as $\hat{x}_t = \frac{1}{t}\sum_{s=1}^{t} x_s$. A function $f : \chi \to \mathbb{R}$ is called L-Lipschitz continuous if there exists a positive constant L for which

$$|f(u) - f(v)| \le L\|u - v\| \text{ for all } u, v \in \chi. \tag{5}$$

A. Graphs

A graph is a concise way to represent the interactions among dynamic agents, e.g., sensors and mobile robots. A weighted graph $\mathcal{G} = (V, E, W)$ is defined by a node set V where the number of nodes in the graph is $|V| = n$. Nodes represent the agents in the network, and an edge set E represents the agent's interactions, i.e., agent i communicates with agent j if there is an edge from i to j, i.e., $(i, j) \in E$. In addition, a weight $w_{ji} \in W$ can be associated with every edge $(i, j) \in E$ through a function $W : E \to R$. A way to represent $\mathcal{G}$ is through the adjacency matrix $A(\mathcal{G})$ where $[A(\mathcal{G})]_{ji} = w_{ji}$ for $(i, j) \in E$ and $[A(\mathcal{G})]_{ji} = 0$ otherwise. For a graph $\mathcal{G}$, d_i is the weighted in-degree of node i, defined as $d_i = \sum_{\{j|(j,i)\in E\}} w_{ij}$. Another matrix representation of $\mathcal{G}$ is the weighted graph Laplacian defined as $L(\mathcal{G}) = \Delta(\mathcal{G}) - A(\mathcal{G})$, where $\Delta(\mathcal{G})$ is the diagonal matrix of d_i's. Based on the construction of the weighted graph Laplacian, for every graph $\mathcal{G}$, $L(\mathcal{G})$ has a right eigenvector of $\mathbf{1}$ associated with eigenvalue $\Lambda = 1$ [12]. A graph $\mathcal{G}$ is called strongly connected if there exists a directed path between every pair of distinct vertices.

B. Regret

In online optimization, an online algorithm is used to generate a sequence of decisions $\{x_t\}_{t=1}^{T}$, where T denotes the number of iterations. At iteration t, the convex cost function f_t is unknown before committing to x_t. The feedback available to the algorithm is the last $f_t(x_t)$ and its gradient. We can capture the performance of online algorithms by a standard measure called regret. Regret measures how competitive the algorithm is with respect to the best fixed solution. In addition, the best fixed decision x^* is chosen with the benefit of hindsight. Formally, the regret is defined as the difference between the incurred cost $f_t(x_t)$ and the cost of the best fixed decision $f_t(x^*)$ when running $t = 1, 2, ..., T$ iterations, i.e.,

$$R_T = \sum_{t=1}^{T} \left(f_t(x_t) - f_t(x^*)\right). \tag{6}$$

An online algorithm performs well if its regret grows sub-linearly with the number of iterations, i.e. $\lim_{T\to\infty} R_T/T = 0$. This implies that the average loss of the algorithm tends to the average loss of the best fixed strategy in hindsight, independent of the adversary's moves. We refer to [13], [14], [15], [16] for further discussion on online algorithms and their regret analysis.

III. Problem Statement

In this section, we consider a large scale network of agents cooperatively optimizing a global objective function. Let the communication constraints on the system be denoted by a graph $\mathcal{G} = (V, E)$. The number of nodes is $|V| = n$ and each node $i \in V$ is an agent which is communicating with its neighbor $j \in N(i)$ through edge $(i, j) \in E$. Note that the neighborhood set is defined as $N(i) = \{j \in V|(i, j) \in E\}$. We state the online distributed convex optimization problem as

$$\begin{aligned} \min_{\substack{x\in\chi,\\ y_i\in Y \text{ for } \forall i\in[n]}} \quad & F_t(x, y) = f_t(x) + \sum_{i=1}^{n} \phi_i(y_i) \end{aligned} \tag{7}$$

subject to

$$A_i x + B_i y_i - c_i = 0 \text{ for } \forall i \in [n], \tag{8}$$

where $f_t(x) = \sum_{i=1}^{n} f_{i,t}(x)$, and $f_{i,t}(x) : \mathbb{R}^{d_x} \to \mathbb{R}$ and $\phi_i(y_i) : \mathbb{R}^{d_y} \to \mathbb{R}$ are convex cost functions associated with agent $i \in [n]$. Let $f_{i,t}(x)$ and $\phi_i(y_i)$ be Lipschitz continuous which is defined as

$$|f_{i,t}(u) - f_{i,t}(v)| \leq L\|u - v\|_2 \text{ for all } u, v \in \chi,$$
$$|\phi_i(u^{'}) - \phi_i(v^{'})| \leq L_\phi \|u^{'} - v^{'}\|_2 \text{ for all } u^{'}, v^{'} \in Y.$$

Note that $f_{i,t}(x)$ evolves over time in an unpredictable manner. In other words, $F_t(x_{i,t}, y_{i,t})$ is revealed after each agent i executes $x_{i,t} \in \chi$ and $y_{i,t} \in Y$ based on local information available to it. The local decision made by agent i is represented by the optimization variables x_i and y_i which belong to compact convex sets $\chi \subseteq \mathbb{R}^{d_x}$ and $Y \subseteq \mathbb{R}^{d_y}$, respectively. In addition, for all $x, x' \in \chi$ and $y, y' \in Y$ we have $||x - x'|| \leq D$ and $||y - y'|| \leq D$. The matrices in the local linear constraints can be described as $A_i \in \mathbb{R}^{m_i \times d_x}$, $B_i \in \mathbb{R}^{m_i \times d_y}$, and $c_i \in \mathbb{R}^{m_i}$ at node $i \in [n]$.

A. Regret for Constrained Optimization

In this section we propose a measure for evaluating the performance of OD-ADMM based on variational inequality. This measure is inspired by the convergence analysis of Douglas-Rachford ADMM, first presented in [17].

We consider the Lagrangian for the constrained optimization problem (7),

$$\mathcal{L}_t = f_t(x) + \sum_{i=1}^{n} \phi_i(y_i) + \sum_{i=1}^{n} \langle \lambda_i, A_i x + B_i y_i - c_i \rangle . \tag{9}$$

Moreover, an augmented Lagrangian method [18] can be applied to further penalize the violation of linear constraints (8),

$$\mathcal{L}_t^\rho = f_t(x) + \sum_{i=1}^{n} \phi_i(y_i) + \sum_{i=1}^{n} (\langle \lambda_i, A_i x + B_i y_i - c_i \rangle + \frac{\rho}{2} ||A_i x + B_i y_i - c_i||^2). \tag{10}$$

Let $\Omega = \chi \times Y \times \mathbb{R}^l$. Based on the first order necessary condition on the Lagrangian, the vector $w^* = (x^*, y_1^*, \cdots, y_n^*, \lambda_1^*, \cdots, \lambda_n^*) \in \Omega$ solves the problem (7) optimally if it satisfies the variational inequality

$$(w - w^*)^T \nabla \mathcal{L}(w^*) \geq 0. \tag{11}$$

The condition in (11) can be expressed as

$$F_t(x, y) - F_t(x^*, y^*) + \sum_{i=1}^{n} \begin{bmatrix} x_{i,t} - x^* \\ y_{i,t} - y_i^* \\ \lambda_{i,t+1} - \lambda_i^* \end{bmatrix}^T \begin{bmatrix} A_i^T \lambda_{i,t+1}^* \\ B_i^T \lambda_{i,t+1}^* \\ -r_i(x_{i,t}^*, y_{i,t}^*) \end{bmatrix} \geq 0,$$

for all $w \in \Omega$, where

$$r_i(x, y) = A_i x + B_i y - c_i, \tag{12}$$

is the residual of the constraint at node $i \in [n]$. Therefore, based on the work in [17], it can be shown that $w_t \in \Omega$ approximately solves problem (7) with accuracy ϵ_t if it satisfies

$$F_t(x_t, y_t) - F_t(x^*, y^*) + \sum_{i=1}^{n} \begin{bmatrix} x_{i,t} - x^* \\ y_{i,t} - y_i^* \\ \lambda_{i,t+1} - \lambda_i^* \end{bmatrix}^T \begin{bmatrix} A_i^T \lambda_{i,t+1} \\ B_i^T \lambda_{i,t+1} \\ -r_i(x_{i,t}, y_{i,t}) \end{bmatrix} \leq \epsilon_t. \tag{13}$$

Similar to the regret definition for O-ADMM algorithm in [19], we can express the regret due to agent i's action as

$$R_{i,T} = \sum_{t=1}^{T} (\frac{1}{n}(f_t(x_{i,t}) - f_t(x^*)) + \phi_i(y_{i,t}) - \phi_i(y_i^*) + \begin{bmatrix} x_{i,t} - x^* \\ y_{i,t} - y_i^* \\ \lambda_{i,t+1} - \lambda_i^* \end{bmatrix}^T \begin{bmatrix} A_i^T \lambda_{i,t+1} \\ B_i^T \lambda_{i,t+1} \\ -r_i(x_{i,t}, y_{i,t}) \end{bmatrix} + \frac{\rho}{2} ||r_i(x_{i,t}, y_{i,t})||^2), \tag{14}$$

with the best fixed decision provided with the full knowledge of f_t in hindsight denoted as x^*, y_i^* for $\forall i \in [n]$. Note that $\lambda^* \in \mathbb{R}^l$ is an arbitrary variable. Moreover, if the best fixed solution to the Lagrangian form of (7) is (x^*, y_i^*), we have $A_i x^* + B_i y_i^* = c_i$. The regret $R_{i,T}$ is the cumulative penalty agent i pays because of its decisions on the global cost sequence $\left\{\frac{1}{n} f_t + \phi_i\right\}$. Therefore, we can define the global regret as $R_T = \sum_{i=1}^{n} R_{i,T}$, or more precisely

$$R_T = \sum_{t=1}^{T} (f_t(\bar{x}_t) - f_t(x^*) + \sum_{i=1}^{n} (\phi_i(y_{i,t}) - \phi_i(y_i^*)) + \sum_{i=1}^{n} \begin{bmatrix} x_{i,t} - x^* \\ y_{i,t} - y_i^* \\ \lambda_{i,t+1} - \lambda_i^* \end{bmatrix}^T \begin{bmatrix} A_i^T \lambda_{i,t+1} \\ B_i^T \lambda_{i,t+1} \\ -r_i(x_{i,t}, y_{i,t}) \end{bmatrix} + \sum_{i=1}^{n} \frac{\rho}{2} ||r_i(x_{i,t}, y_{i,t})||^2), \tag{15}$$

where $\bar{x}_t = \frac{1}{n} \sum_{i=1}^{n} x_{i,t}$. Based on (13) we can say that $(\bar{x}_t, y_{1,t}, \cdots, y_{n,t}, \lambda_{1,t}, \cdots, \lambda_{n,t}) \in \Omega$ approximately solves problem (7) with accuracy ϵ_t if it satisfies $R_T \leq \sum_{t=1}^{T} \epsilon_t$. Therefore, if the global regret is sub-linear with time, the accuracy of the solution will converge to zero. In addition, this property of the regret ensures that the local linear constraints will be satisfied as $T \to \infty$.

IV. Main Result

The main contribution of this paper is adapting O-ADMM [3] and Nesterov's dual averaging algorithm [20] to provide a parallel decision processes for the online optimization problem proposed in §III. The algorithm updates variables $(x_i, y_i, z_i, \lambda_i)$ for each agent $i \in [n]$ by alternately minimizing the Lagrangian and augmented Lagrangian. In addition, the Lagrangian is linearized based on dual averaging update which is a gradient descent method followed by a projection step onto the constraint set χ, specifically, $z_{t+1} = z_t + g_t$, where $g_t = \nabla \mathcal{L}_t(x_t)$; then

$$x_{t+1} = \prod_{\chi}^{\psi} (z_{t+1}, \alpha_t) .$$

The parameter α_t is a non-increasing sequence of positive functions and $\prod_{\chi}^{\psi}(\cdot)$ is the projection operator onto χ defined as

$$\prod_{\chi}^{\psi}(z_{t+1}, \alpha_t) = \arg\min_{x\in\chi}\left\{\langle z_{t+1}, x\rangle + \frac{1}{\alpha_t}\psi(x)\right\}. \quad (16)$$

Note that $\psi(x) : \chi \to \mathbb{R}$ is a positive function where $\psi \geq 0$, and $\psi(0) = 0$. Moreover, $\psi(x)$ is called the proximal function and is a regularizer to avoid wide oscillation in the projection step. Let ψ be a continuously differentiable and strongly convex function, i.e., $\nabla_x^T \nabla_x \psi(x) = G$, where $G > \gamma I$ for $\gamma > 0$.

Finally, we minimize the augmented Lagrangian over y as

$$y_{t+1} = \arg\min_{y\in Y}\left\{\mathcal{L}_t^{\rho}(x_{t+1}, y, z_{t+1}, \lambda_t)\right\},$$

and update the dual variable λ

$$\lambda_{t+1} = \lambda_t + \rho(Ax_{t+1} + By_{t+1} - c).$$

The distributed algorithm can be considered as an approximate ADMM by an agent i via a convex combination of information provided by its neighbors $N(i)$. The underlying communication network can be represented compactly as a doubly stochastic matrix $P \in \mathbb{R}^{n\times n}$ which preserves the zero structure of the Laplacian matrix $L(\mathcal{G})$. It is clear that for all agents to have access to all sub-gradients of $g_{i,t} = \nabla\mathcal{L}_{i,t}(x_{i,t})$ there must be an all-to-all communication. Therefore, graph $\mathcal{G}$ must be strongly connected to attain this requirement. A method to construct a doubly stochastic matrix P of the required form from $L(\mathcal{G})$ is provided in Proposition (3).

The Online Distributed Dual Averaging (ODD) algorithm is presented in Algorithm 1. The projection function used in this algorithm is defined as in (16).

Algorithm 1: Online Distributed ADMM (OD-ADMM)

1 Input: $\rho > 0$, $\{\alpha_t\}_{t=1}^T$
2 Initilaize $z_{i,1} = \lambda_{i,1} = 0$ and $x_{i,1} = 0$, $y_{i,1} = 0$ for $\forall i = 1, ..., n$
3 **for** $t = 1$ *to* T **do**
4 Adversary reveals $f_t(t) = \{f_{t,i}(t);$ for $\forall i = 1, ..., n\}$
5 Compute subgradient $g_i(t) \in \partial f_{t,i}(x_{t,i})$
6 **for** *Each Agent* i **do**
7 $z_{i,t+1} = \sum_{j\in\{N(i),i\}} P_{j,i} z_{j,t} + g_{i,t}$
8 $x_{i,t+1} = \prod_{\chi}^{\psi}(z_{i,t+1} + A_i^T\hat{\lambda}_{i,t}, \alpha_t)$
9 $r_i(x_{i,t+1}, y) = A_i x_{i,t+1} + B_i y - c_i$
10 $y_{i,t+1} = \text{argmin}_{y\in Y}(\phi_i(y) + \lambda_{i,t}^T r_i(x_{i,t+1}, y) + \frac{\rho}{2}||r_i(x_{i,t+1}, y)||^2)$
11 $\lambda_{i,t+1} = \lambda_{i,t} + \rho(A_i x_{i,t+1} + B_i y_{i,t+1} - c_i)$
12 $\hat{\lambda}_{i,t+1} = \frac{1}{t+1}\sum_{s=1}^{t+1}\lambda_{i,s}$
13 **end**
14 **end**

Before presenting the convergence rate of OD-ADMM algorithm, we provide a few preliminary remarks and definitions. The sequence of average dual sub-gradient $\bar{z}_t$ and average sub-gradient $\bar{g}_t$ are defined as

$$\bar{z}_t = \frac{1}{n}\sum_{i=1}^n z_{i,t}, \quad \bar{g}_t = \frac{1}{n}\sum_{i=1}^n g_{i,t} \quad (17)$$

over all agents in the network. Thus, the following update rule is introduced similar to the standard dual averaging algorithm

$$\bar{z}_{t+1} = \frac{1}{t}(\bar{z}_t + \bar{g}_t), \quad (18)$$

where the primal variable is

$$\theta_{t+1} = \Pi_{\chi}^{\psi}(\bar{z}_{t+1} + A_i^T\hat{\lambda}_{i,t}, \alpha_t). \quad (19)$$

The regret analysis can now be presented as follows.

Theorem 1. *Given the sequence of $x_{i,t}$, $y_{i,t}$, and $\lambda_{i,t}$ generated by lines 8, 10, and 11 in Algorithm 1 for all $i \in [n]$ where $\psi(x^*) \leq K^2$ and $\alpha(t) = k/\sqrt{t}$, we have*

$$R_{i,T} \leq J_{i,c} + kJ_{i,2}\sqrt{T}, \quad (20)$$

where $J_{i,1} = \frac{1}{2\rho}(\frac{L_\phi}{\sigma_1(B_i)})^2 + \frac{DL_\phi}{\sigma_1(B_i)}$ and

$$J_{i,2} = K^2 + \frac{L^2}{\gamma} + 2(3L + \frac{\sigma_1(A_i)}{\sigma_1(B_i)}L_\phi)$$
$$(\frac{\sqrt{n}L}{1-\sigma_2(P)} + 2L + L_\phi(\sum_{j=1}^n \frac{\sigma_1(A_j)}{n\sigma_1(B_j)} + \frac{\sigma_1(A_i)}{\sigma_1(B_i)})).$$

Proof: By optimality of $y_{i,t+1}$ in line 8 of Algorithm 1 we have that for all $y_i^* \in Y$

$$\phi_i(y_{i,t}) - \phi_i(y_i^*) \leq -\langle\lambda_{i,t}, B_i(y_{i,t} - y_i^*)\rangle, \quad (21)$$

which provides a bound on the second part of objective function (7). Since $f_{t,i}(x_{i,t})$ are L-Lipschitz, we have

$$\sum_{t=1}^T(f_t(x_{i,t}) - f_t(x^*)) \leq$$
$$\sum_{t=1}^T(f_t(\theta_t) - f_t(x^*) + L\|x_{i,t} - \theta_t\|). \quad (22)$$

Now, the first term in the right hand side is bounded as follows

$$f_t(\theta_t) - f_t(x^*) = (\sum_{i=1}^n f_{t,i}(x_{i,t}) - f_t(x^*)) + \quad (23)$$
$$(\sum_{i=1}^n [f_{t,i}(\theta_t) - f_{t,i}(x_{i,t})]).$$

Based on (22) and (23) we can bound the first part of objective function (7) as [1]

$$\sum_{t=1}^T \frac{1}{n}(f_t(x_{i,t}) - f_t(x^*)) \leq$$
$$\sum_{t=1}^T(\frac{1}{n}\sum_{i=1}^n\langle g_{i,t}, x_{i,t} - x^*\rangle +$$
$$\frac{L}{n}\sum_{i=1}^n\|x_{i,t} - \theta_t\| + L\|x_{i,t} - \theta_t\|). \quad (24)$$

The first term in the right hand side of (24) is expanded as

[1]Note that $f_{t,i}$'s are convex and $\sum_{t=1}^T\left(\sum_{i=1}^n f_{t,i}(x_{i,t}) - f_t(x^*)\right) \leq \sum_{t=1}^T\left(\sum_{i=1}^n\langle g_{i,t}, x_{i,t} - x^*\rangle\right)$, where $g_i(t) \in \partial f_{t,i}(x_{i,t})$ is the sub-gradient of $f_{t,i}$ at $x_{i,t}$.

$$\sum_{t=1}^{T}(\frac{1}{n}\sum_{i=1}^{n}\langle g_{i,t}, x_{i,t}-x^*\rangle) =$$
$$\sum_{t=1}^{T}(\frac{1}{n}\sum_{i=1}^{n}\langle g_{i,t}, x_{i,t}-\theta_t\rangle$$
$$+\langle \bar{g}_t, \theta_t - x^*\rangle) \quad (25)$$

Now, we need to bound the terms on the right hand side of (25). The first term is bounded based on the convexity and L-Lipschitz conditions on $f_{t,i}$. In other words,

$$\langle g_{i,t}, x_{i,t}-\theta_t\rangle \leq L\|x_{i,t}-\theta_t\|. \quad (26)$$

Therefore, Lemma (26) and (6) imply

$$\sum_{t=1}^{T}\frac{1}{n}(f_t(x_{i,t})-f_t(x^*)) \leq$$
$$\sum_{t=1}^{T}(\frac{2L}{n}\sum_{i=1}^{n}\|x_{i,t}-\theta_t\|+L\|x_{i,t}-\theta_t\|+$$
$$\left\langle A_i^T\lambda_{i,t}, x^*-\theta_{t+1}\right\rangle+\frac{1}{\alpha_T}\psi(x^*)+$$
$$\sum_{t=2}^{T}\frac{\alpha_{t-1}}{2}||\bar{g}_t||^2_{G^{-1}}. \quad (27)$$

Based on (21) and (27), the regret associated with cost function is bounded as

$$\sum_{t=1}^{T}(\frac{1}{n}(f_t(x_{i,t})-f_t(x^*))+\phi_i(y_{i,t})-\phi_i(y_i^*)) \leq$$
$$\frac{T}{\alpha_T}\psi(x^*)+\sum_{t=2}^{T}\frac{\alpha_{t-1}}{2}||\bar{g}_t||^2_{G^{-1}}+$$
$$\sum_{t=1}^{T}(-\langle\lambda_{i,t}, B_i(y_{i,t}-y_i^*)\rangle+$$
$$\frac{2L}{n}\sum_{i=1}^{n}\|x_{i,t}-\theta_t\|+L\|x_{i,t}-\theta_t\|+$$
$$\left\langle A_i^T\lambda_{i,t}, x^*-\theta_{t+1}\right\rangle). \quad (28)$$

We can expand the last term on the right hand side of (28) as

$$\left\langle A_i^T\lambda_{i,t}, x^*-\theta_{t+1}\right\rangle = -\langle\lambda_{i,t}, A_i(x_{i,t}-x^*)\rangle+$$
$$\langle\lambda_{i,t}, A_i(x_{i,t}-\theta_{t+1})\rangle. \quad (29)$$

After carrying on several algebraic operations[2], the last term in (29) can be bounded as

$$\langle\lambda_{i,t}, A_i(x_{i,t}-\theta_{t+1})\rangle \leq \langle\lambda_{i,t}-\lambda_i^*, r_{i,t}\rangle+$$
$$\frac{1}{2\rho}(||\lambda_i^*-\lambda_{i,t}||^2-||\lambda_i^*-\lambda_{i,t+1}||^2)-$$
$$\frac{\rho}{2}||r_i(x_{i,t},y_{i,t})||^2+\langle\lambda_i^*, A_i(x_{i,t}-\theta_t)\rangle$$
$$+\langle\lambda_i^*, A_i(\theta_t-\theta_{t+1})\rangle, \quad (30)$$

where [3]

$$\langle\lambda_i^*, A_i(x_{i,t}-\theta_t)\rangle \leq ||\lambda_i^*||_*\|A_i(x_{i,t}-\theta_t)\|$$
$$\leq \sigma_1(A_i)||\lambda_i^*||_*\|x_{i,t}-\theta_t\|. \quad (31)$$

Now we can use (28), (29), (30), and (31) to express the regret defined in (14)

$$R_{i,T} \leq \frac{1}{\alpha_T}\psi(x^*)+\sum_{t=2}^{T}\frac{\alpha_{t-1}}{2}||\bar{g}_t||^2_{G^{-1}}+$$
$$\sum_{t=1}^{T}(\frac{2L}{n}\sum_{i=1}^{n}\|x_{i,t}-\theta_t\|+L\|x_{i,t}-\theta_t\|+$$
$$\frac{1}{2\rho}\left(||\lambda_i^*-\lambda_{i,t}||^2-||\lambda_i^*-\lambda_{i,t+1}||^2\right)+$$
$$\sigma_1(A_i)||\lambda_i^*||_*\|x_{i,t}-\theta_t\|+\langle\lambda_i^*, A_i(\theta_t-\theta_{t+1})\rangle). \quad (32)$$

In order to further bound the regret in (32) we can use Lemma (4) which implies

$$\|x_{i,t}-\theta_t\| \leq \alpha_{t-1}\|\bar{q}_t-q_{i,t}\|_*. \quad (33)$$

Moreover, Lemma (5) provides a bound on $\|\bar{q}_t-q_{i,t}\|_*$ and we can further bound (33) as

$$\|x_{i,t}-\theta_t\| \leq \alpha_{t-1}(\frac{\sqrt{n}L}{1-\sigma_2(P)}+2L+$$
$$L_\phi(\sum_{j=1}^{n}\frac{\sigma_1(A_j)}{n\sigma_1(B_j)}+\frac{\sigma_1(A_i)}{\sigma_1(B_i)})). \quad (34)$$

Note that $\lambda_i(1)=0$ and $||\bar{g}_t||_{G^{-1}} \leq \frac{L}{\gamma}$. Thus, inequalities (32) and (34) imply

$$R_{i,T} \leq \frac{T}{\alpha_T}\psi(x^*)+\frac{L^2}{2\gamma^2}\sum_{t=1}^{T-1}\alpha_t+$$
$$(3L+\sigma_1(A_i)||\lambda_i^*||_*)(\frac{\sqrt{n}L}{1-\sigma_2(P)}+2L+$$
$$L_\phi(\sum_{j=1}^{n}\frac{\sigma_1(A_j)}{n\sigma_1(B_j)}+\frac{\sigma_1(A_i)}{\sigma_1(B_i)}))\sum_{t=1}^{T-1}\alpha_t+$$
$$\frac{||\lambda_i^*||^2}{2\rho}+\langle\lambda_i^*, A_i(\theta_1-\theta_{T+1})\rangle. \quad (35)$$

Moreover, $\theta \in \chi$ is a compact convex set with diameter D, thus, the last term on the right hand side is bounded as

$$\langle\lambda_i^*, A_i(\theta_1-\theta_{T+1})\rangle \leq ||\lambda_i^*|| \times ||\theta_1-\theta_{T+1}||$$
$$\leq D||\lambda_i^*||. \quad (36)$$

Now we need to find an upper bound for $||\lambda_i^*||$. From line 8 of Algorithm 1 we have $\nabla_y\phi_i(y_i^*) = -B_i^T\lambda_i^*$, and since $||\nabla_y\phi_i(y^*)|| \leq L_\phi$, we have $||B_i^T\lambda_i^*|| \leq L_\phi$. Thus, $||\lambda_i^*||$ is bounded as

$$||\lambda_i^*|| \leq \frac{L_\phi}{\sigma_1(B_i)}. \quad (37)$$

Equations (36) and (37) impose an upper bound on the right hand side of (35) which can be expressed as

[2]Note that $\langle v_1-v_2, v_3+v_4\rangle = \frac{1}{2}(||v_4-v_2||^2-||v_4-v_1||^2+||v_3+v_1||^2-||v_3+v_2||^2$.

[3]Note that $\|Qx\| \leq \sigma_1(Q)||x||$ for any matrix $Q \in \mathbb{R}^{m\times n}$ and vector $x \in \mathbb{R}^n$.

$$R_{i,T} \leq \frac{T}{\alpha_T}\psi(x^*) + \frac{L^2}{2\gamma}\sum_{t=1}^{T-1}\alpha_t + (3L + \sigma_1(A_i)||\lambda_i^*||_*)(\frac{\sqrt{n}L}{1-\sigma_2(P)} + 2L + L_\phi(\sum_{j=1}^{n}\frac{\sigma_1(A_j)}{n\sigma_1(B_j)} + \frac{\sigma_1(A_i)}{\sigma_1(B_i)}))\sum_{t=1}^{T-1}\alpha_t + \frac{1}{2\rho}(\frac{L_\phi}{\sigma_1(B_i)})^2 + \frac{DL_\phi}{\sigma_1(B_i)}. \quad (38)$$

Since $\psi(x^*) \leq K^2$ and $\alpha_t = k\sqrt{t}$, applying the integral test on the $\sum_{t=1}^{T-1}\frac{\alpha_t}{t}$ [4] in (38) leads to (20) and the theorem follows. ∎

The result validates the "good" performance of OD-ADMM by showing sub-linear regret. In addition, it highlights the importance of the underlying topology through $\sigma_2(P)$ and the local linear constraints through $\sigma_1(A_i)$ and $\sigma_1(B_i)$. Note that a well known measure of network connectivity is the second smallest eigenvalue of the graph Laplacian $L(\mathcal{G})$ denoted by $\Lambda_2(\mathcal{G})$. Since the communication matrix P is formed as proposed in Proposition 3, $1-\sigma_2(P)$ is proportional to $\Lambda_2(\mathcal{G})$ implying high network connectivity promotes good performance of the algorithm.

Now, the global regret analysis can be stated exhibiting similar dependence on the parameters of the network and local linear constraints.

Corollary 2. *Given the sequence of $x_{i,t}$, $y_{i,t}$, and $\lambda_{i,t}$ generated by lines 8, 10, and 11 in Algorithm 1 for all $i \in [n]$ where $\psi(x^*) \leq K^2$ and $\alpha(t) = k/\sqrt{t}$, we have*

$$R_T \leq \sum_{i=1}^{n}(J_{i,c} + kJ_{i,2}\sqrt{T}), \quad (39)$$

where $J_{i,1} = \frac{1}{2\rho}(\frac{L_\phi}{\sigma_1(B_i)})^2 + \frac{DL_\phi}{\sigma_1(B_i)}$ *and*

$$J_{i,2} = K^2 + \frac{L^2}{\gamma} + 2(3L + \frac{\sigma_1(A_i)}{\sigma_1(B_i)}L_\phi)\left(\frac{\sqrt{n}L}{1-\sigma_2(P)} + 2L + L_\phi(\sum_{j=1}^{n}\frac{\sigma_1(A_j)}{n\sigma_1(B_j)} + \frac{\sigma_1(A_i)}{\sigma_1(B_i)})\right).$$

Proof: The global regret is defined as $R_T = \sum_{i=1}^{n} R_{i,T}$ where an upper bound on the distributed regret is provided in (20). This leads to the statement of the corollary. ∎

Note that both distributed and global regrets have the convergence rate of $O(\sqrt{T})$.

V. Example - Formation Acquisition with Points of Interest and Boundary Constraints

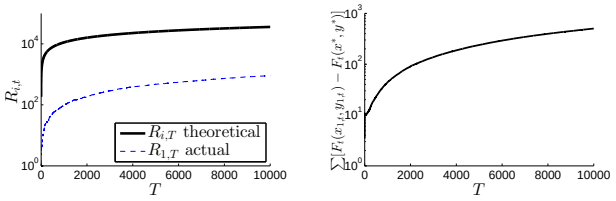
Consider a formation acquisition problem with n agents, where agent i has position $y_i \in \mathbb{R}^2$ restricted to a convex set $\mathcal{X} = [-1,1]^2$. The centroid of the formation is $x \in \mathbb{R}^2$ which is similarly constrained to $Y = \mathcal{X}$. The formation shape is defined for each agent by its offset c_i from the centroid, namely $x - y_i = c_i$. There is a known boundary S of the convex environment which agents are required to avoid by increasing the distance to the boundary $\text{dist}(y_i, S) = \inf_{x\in S}\|x - y_i\|$. This is achieved with a penalty function $\phi_i(y_i) = (\text{dist}(y_i, S) + 1)^{-1}$ associated with agent y_i's proximity S. Assuming $\text{int}\,(S \bigcap \mathcal{X}) = \emptyset$ then $\phi_i(y_i)$ is convex. For each time step t, each agent i obtains a point of interest $w_{t,i}$. The centroid is ideally located close to these points of interest, and is promoted through the minimization of a function $f_{t,i}(x) = \frac{1}{2}\|x - w_{t,i}\|_2^2$.

$$\min_{x\in\mathcal{X}, y\in Y} \quad \sum_t\sum_i f_{t,i}(x) + \sum_t \phi_{t,i}(y_i)$$
$$\text{s.t.} \quad A_i x + B_i y_i = c_i \qquad \text{for all } i \in [n]$$

where $A_i = I_2$ and $B_i = -I_2$ for all i.

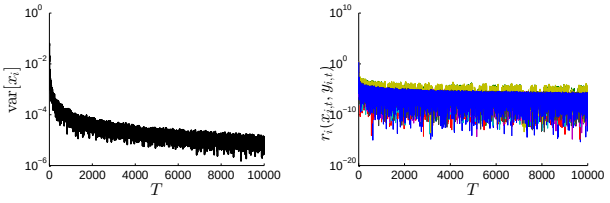
Consider $S = \{x|\,\|x\| = 1\}$ be so $\phi_i(y_i) = (2 - \|y_i\|_\infty)^{-1}$. The relevant parameters of the ADMM algorithm are $g_i(t) = \nabla f_{t,i}(x_i) = x_i - w_{t,i}$, $k = \rho = 1$, and $\Theta(x) = \|x\|_2^2$. Therefore, $\|\nabla\phi_i(y_i)\|_2 \leq 1 = L_\phi, \|g_i\| = \|w_{i,t} - y_i\| \leq = 2 = L$. The remaining terms of the regret bound are $\sigma_1(B_i) = 1$, $\sigma_1(A_i) = 1$, $D = \gamma = 2$ and $K = 1$.

The algorithm was applied to $n = 8$ agents connected over a random graph with $\sigma_2(P) = 0.43$ with c_i's selected to acquire a formation with the n agents equidistant apart on the circumference of a circle of radius 0.4. Points of interest randomly switch, for each agent i, and each time t, between a fixed point $(-1,-1)$, a fixed point $(-1,1)$, and uniformly spread over the area $[0,1]^2$. The actual regret for agent 1 compared to the theoretical appear in Figure 1a and the general accumulative regret in Figure 1b. The convergence of the global variables x_i to agreement as well as the reduction of the residue are displayed in Figure 2.



(a) (b)

Figure 1: (a) Actual distributed regret $R_{1,T}$ of the formation acquisition problem compared with its theoretical bound (14). (b) General accumulative regret of the formation acquisition problem.



(a) (b)

Figure 2: (a) The variance of the global variable x_i and (b) the residue for each agent over times

[4] Note that $\frac{1}{\sqrt{t}}$ is a non increasing positive function and the integral test leads to $\sum_{t=1}^{T}\frac{1}{\sqrt{t}} \leq 2\sqrt{T} - 1$.

VI. Conclusion

A fully decentralized online algorithm was developed to study a convex optimization problem with a distributable objective function and linear constraints. This problem setup is applicable to a network of agents, where each agent optimizes the global objective function with access to its privately known local objective function and linear constraint. In this algorithm the decisions are made in parallel based on local information and communication with neighbors. A sub-linear regret bound of $O(\sqrt{T})$ is attained for the objective function and linear local constraints violation. In particular, the online algorithm is competitive with respect to the best fixed decision performance in hindsight. In addition, we highlight the role of the underlying network topology in achieving "good" regret, i.e., the regret bound improves with increased connectivity in the network.

The proposed algorithm was applied to a formation acquisition problem showing sound agreement with the theoretical results.

Future work of particular interest includes exploring regret bound over time varying network topologies, and investigating favorable graph characteristics for the online ADMM framework.

VII. Appendix

The following Lemma 4 and Proposition 3 can be found in [7] and [8], respectively. Thus, they are presented here without proof.

Proposition 3. *If graph $\mathcal{G}$ is strongly connected then the matrix $P = I - \frac{1}{\epsilon} diag\,(v)\, L(\mathcal{G})$ is doubly stochastic, where $v^T L\,(\mathcal{G}) = 0$ with positive vector $v = [v_1, v_2, \ldots, v_n]^T$ and $\epsilon \in (\max_{i\in V}\,(v_i d_i), \infty)$. If graph $\mathcal{G}$ is balanced then the matrix $P = I - \frac{1}{\epsilon} L(\mathcal{G})$ is doubly stochastic, where $\epsilon \in (d_{\max}, \infty)$.*

Lemma 4. *For any $u, v \in \mathbb{R}^m$, and under the conditions stated for proximal function ψ and step size α, we have*

$$\| \prod_{\chi}^{\psi}(u,\alpha) - \prod_{\chi}^{\psi}(v,\alpha)\| \leq \alpha \|u - v\|_*.$$

Lemma 5. *For any sequences of $q_{i,t} = z_{i,t} + A^T \hat{\lambda}_{i,t-1}$ and $\bar{q}_t = \bar{z}_t + \frac{1}{n}\sum_{i=1}^n A_i^T \hat{\lambda}_{i,t-1}$ generated by Algorithm 1, we have $\|\bar{q}_t - q_{i,t}\|_* \leq \frac{\sqrt{n}L}{1-\sigma_2(P)} + 2L + L_\phi(\sum_{j=1}^n \frac{\sigma_1(A_j)}{n\sigma_1(B_j)} + \frac{\sigma_1(A_i)}{\sigma_1(B_i)})$ for all $i \in [n]$ and $t \in [T]$.*

Proof: Based on line 7 of Algorithm 1 we have

$$z_{i,t} = \sum_{j=1}^n P_{ji}^s z_{j,t-s} + \sum_{k=t-s}^{t-2}\sum_{j=1}^n P_{ji}^{t-k-1} g_{j,k} + g_{i,t-1}.$$

In addition, $\bar{z}_t$ evolves as

$$\bar{z}_t = \bar{z}_{t-s} + \sum_{k=t-s}^{t-1} \bar{g}_k. \tag{40}$$

Assuming $t - s = 1$, and $z_{i,1} = 0$ for all $i \in [n]$ and based on (40) we have

$$\bar{z}_t - z_{i,t} = \sum_{k=1}^{t-2}(\sum_{j=1}^n (\frac{1}{n} - P_{ji}^{t-k-1}) g_j(k)) + (\bar{g}_{t-1} - g_{i,t-1}). \tag{41}$$

We define $p_{i,t-1} = A_i^T \hat{\lambda}_{i,t-1}$ and $\bar{p}_{t-1} = \frac{1}{n}\sum_{i=1}^n p_{i,t-1}$. Based on Algorithm 1, we have

$$\bar{p}_{t-1} - p_{i,t-1} = \frac{1}{t-1}\sum_{k=1}^{t-1}(\frac{1}{n}\sum_{j=1}^n A_j^T \lambda_{j,k} - A_i^T \lambda_{i,k}). \tag{42}$$

Thus, the dual norm of $\bar{q}_t - q_{i,t}$ can be bounded as

$$\|\bar{q}_t - q_{i,t}\|_* \leq \sum_{k=1}^{t-2} \|P^{t-k-1} e_i - \frac{\mathbf{1}}{n}\|_1 ||g_{j,k}||_* + \frac{1}{t-1}\sum_{k=1}^{t-1}(\frac{1}{n}\sum_{j=1}^n \sigma_1(A_j)||\lambda_{j,k}||_* + \sigma_1(A_i)||\lambda_{i,k}||_*) + ||\bar{g}_{t-1} - g_{i,t-1}||_*. \tag{43}$$

Since $\|g_i(t)\|_\star \leq L$ and $||\lambda_{i,t}||_* \leq \frac{L_\phi}{\sigma_1(B_i)}$, the dual norm of $\bar{q}_t - q_{i,t}$ is further bounded as [5]

$$\|\bar{q}_t - q_{i,t}\|_* \leq \sqrt{n}L\sum_{k=1}^{t-1}\sigma_2(P)^k + 2L + L_\phi(\sum_{j=1}^n \frac{\sigma_1(A_j)}{n\sigma_1(B_j)} + \frac{\sigma_1(A_i)}{\sigma_1(B_i)}) \tag{44}$$

In addition, the second largest singular value of P is $\sigma_2(P) \leq 1$, where P is a doubly stochastic matrix [21]. The inequality (44) is thus bounded as

$$\|\bar{q}_t - q_{i,t}\|_* \leq \frac{\sqrt{n}L}{1-\sigma_2(P)} + 2L + L_\phi(\sum_{j=1}^n \frac{\sigma_1(A_j)}{n\sigma_1(B_j)} + \frac{\sigma_1(A_i)}{\sigma_1(B_i)}).$$

■

Lemma 6. *For any positive and non-increasing sequence $\alpha(t)$ and $x^* \in \chi$*

$$\sum_{t=1}^T \langle g_t, x_t - x^*\rangle \leq \sum_{t=1}^T \langle A_i^T \lambda_{i,t}, x^* - x_{t+1}\rangle + \frac{1}{\alpha_T}\psi(x^*) + \sum_{t=2}^T \frac{\alpha_{t-1}}{2}||g_t||^2_{G^{-1}},$$

where the sequence of x_t is generated by

$$z_{t+1} = \frac{1}{t}(z_t + g_t)$$
$$x_{t+1} = \Pi_\chi^\psi(z_{t+1} + A_i^T \hat{\lambda}_{i,t}, \alpha_t).$$

Proof: We define ϱ_t and V_t as follows

$$\varrho_t(z,x) = \langle z, x\rangle - \left\langle A_i^T \hat{\lambda}_{i,t-1}, x\right\rangle - \frac{1}{\alpha_{t-1}}\psi(x), \tag{45}$$

[5] Note that $\|P^t x - \frac{\mathbf{1}}{n}\|_1 \leq \sigma_2(P)^t\sqrt{n}$, where the vector x belongs to $\{x \in \mathbb{R}^n | x \geq 0, \sum_{i=1}^n x_i = 1\}$. This is a property of stochastic matrix P introduced by Duchi *et al.* [7].

$$V_t(z) = \max_{x \in \chi} \{\varrho_t(z, x)\}. \quad (46)$$

From line 8 in Algorithm 1, we have $V_t(-z_t) = \varrho_t(-z_t, x_t)$ and $\langle x - x_t, \nabla_x \varrho_t(-z_t, x_t)\rangle \leq 0$. Since $\varrho_{t-1}(-z(t-1), x)$ is concave and differentiable, then it is bounded above by its first-order Taylor approximation and we have

$$\varrho_t(-z_t, x_{t+1}) \leq$$
$$\varrho_t(-z_t, x_t) + \langle x_{t+1} - x_t, \nabla_x \varrho_t(-z_t, x_t)\rangle +$$
$$\frac{1}{2}(x_{t+1} - x_t)^T \nabla_x^T \nabla_x \varrho_t(-z_t, x_t)(x_{t+1} - x_t). \quad (47)$$

Based on (45) and (46), we can further bound (47) as

$$\varrho_t(-z_t, x_{t+1}) \leq V_t(-z_t) - \frac{1}{2\alpha_{t-1}}||x_{t+1} - x_t||_G^2, \quad (48)$$

where $\nabla_x^T \nabla_x \psi(x) = G$. Using (45) and $\alpha_{t+1} \leq \alpha_t \Rightarrow \frac{1}{\alpha_t} \leq \frac{1}{\alpha_{t+1}}$, we have

$$\varrho_{t+1}(-z_{t+1}, x_{t+1}) \leq \varrho_t(-z_t, x_{t+1}) -$$
$$\langle g_t, x_{t+1}\rangle - \left\langle A_i^T \lambda_{i,t}, x_{t+1}\right\rangle. \quad (49)$$

Note that

$$V_{t+1}(-z_{t+1}) = \varrho_{t+1}(-z_{t+1}, x_{t+1}). \quad (50)$$

Thus, using (48) and (49) we can compare $V_{t+1}(-z_{t+1})$ to $V_t(-z_t)$ as

$$V_{t+1}(-z_{t+1}) \leq V_t(-z_t) - \frac{1}{2\alpha_{t-1}}||x_{t+1} - x_t||_G^2 -$$
$$\langle g_t, x_t\rangle - \left\langle A_i^T \lambda_{i,t}, x_{t+1}\right\rangle + \langle g_t, x_t - x_{t+1}\rangle. \quad (51)$$

The last term on the right hand side of (51) can be bounded as

$$\langle g_t, x_t - x_{t+1}\rangle \leq$$
$$||g_t||_{G^{-1}} ||x_{t+1} - x_t||_G \leq$$
$$\frac{1}{2\alpha_{t-1}}||x_{t+1} - x_t||_G^2 + \frac{\alpha_{t-1}}{2}||g_t||_{G^{-1}}^2. \quad (52)$$

Therefore, (51) and (52) implies

$$\langle g_t, x_t\rangle \leq V_t(-z_t) - V_{t+1}(-z_{t+1}) -$$
$$\left\langle A_i^T \lambda_{i,t}, x_{t+1}\right\rangle + \frac{\alpha_{t-1}}{2}||g_t||_{G^{-1}}^2. \quad (53)$$

From (53) we have

$$\sum_{t=1}^{T} \langle g_t, x_t\rangle \leq -V(-z_{T+1}) -$$
$$\sum_{t=1}^{T} \left\langle A_i^T \lambda_{i,t}, x_{t+1}\right\rangle + \sum_{t=2}^{T} \frac{\alpha_{t-1}}{2}||g_t||_{G^{-1}}^2. \quad (54)$$

Note that $V_1(-z_1) = 0$ since $z_1 = 0$ and $\lambda_i(1) = 0$. Moreover, given (45) and (46), we note that for any $x^* \in \chi$

$$-\sum_{t=1}^{T} \langle g_t, x^*\rangle =$$
$$\varrho_{T+1}(-\sum_{t=1}^{T} g_t, x^*) + \left\langle A_i^T \hat{\lambda}_{i,T}, x^*\right\rangle + \frac{1}{\alpha_T}\psi(x^*) \leq$$
$$V_{T+1}(-\sum_{t=1}^{T} g_t) + \sum_{t=1}^{T} \left\langle A_i^T \lambda_{i,t}, x^*\right\rangle + \frac{1}{\alpha_T}\psi(x^*). \quad (55)$$

Since $z_{T+1} = \sum_{t=1}^{T} g_t$, we have

$$-\sum_{t=1}^{T} \langle g_t, x^*\rangle \leq V_{T+1}(-z_{T+1}) +$$
$$\sum_{t=1}^{T} \left\langle A_i^T \lambda_{i,t}, x^*\right\rangle + \frac{1}{\alpha_T}\psi(x^*). \quad (56)$$

Finally, we combine (54) and (56) and the lemma will follow. ■

REFERENCES

[1] H. Ouyang, N. He, and A. Gray, "Stochastic ADMM for nonsmooth optimization," *arXiv preprint arXiv:1211.0632*, pp. 1–11, 2012.

[2] H. Wang and A. Banerjee, "Online Alternating Direction Method," in *International Conference on Machine Learning*, no. 1, 2012, pp. 1–40.

[3] T. Suzuki, "Dual averaging and proximal gradient descent for online alternating direction multiplier method," *International Conference on Machine Learning*, vol. 28, 2013.

[4] E. Wei and A. Ozdaglar, "Distributed Alternating Direction Method of Multipliers," *IEEE Conference on Decision and Control*, pp. 5445–5450, Dec. 2012.

[5] ——, "On the O (1/k) Convergence of Asynchronous Distributed Alternating Direction Method of Multipliers," in *IEEE Global Conference on Signal and Information Processing*, 2013, pp. 1–30.

[6] F. Yan, S. Sundaram, S. V. N. Vishwanathan, and Y. Qi, "Distributed autonomous online learning: Regrets and intrinsic privacy-preserving properties," *IEEE Transactions on Knowledge and Data Engineering*, vol. 99, p. 1, 2012.

[7] 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.

[8] S. Hosseini, A. Chapman, and M. Mesbahi, "Online Distributed Optimization via Dual Averaging," *IEEE Conference on Decision and Control*, 2013.

[9] J. Mota and J. Xavier, "D-ADMM: A distributed algorithm for compressed sensing and other separable optimization problems," *IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 2869–2872, 2012.

[10] J. Mota, J. Xavier, P. Aguiar, and M. Puschel, "D-ADMM: A communication-efficient distributed algorithm for separable optimization," *IEEE Transactions on Signal Processing*, vol. 61, no. 10, pp. 2718–2723, 2013.

[11] W. Deng, M. Lai, and W. Yin, "On the o (1/k) Convergence and Parallelization of the Alternating Direction Method of Multipliers," *arXiv preprint arXiv:1312.3040*, pp. 1–23, 2013.

[12] M. Mesbahi and M. Egerstedt, *Graph Theoretic Methods in Multiagent Networks*. NJ: Princeton University Press, 2010.

[13] S. Shalev-Shwartz, "Online learning and online convex optimization," *Foundations and Trends in Machine Learning*, vol. 4, pp. 107–194, 2012.

[14] S. Bubeck, "Introduction to Online Optimization," *Lecture Notes*, 2011.

[15] E. Hazan, "The Convex Optimization Approach to Regret Minimization," *Optimization for machine learning*, pp. 287–294, 2011.

[16] E. Hazan, A. Agarwal, and S. Kale, "Logarithmic regret algorithms for online convex optimization," *Machine Learning*, vol. 69, pp. 169–192, 2007.

[17] 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.

[18] D. Bertsekas, *Nonlinear programming*. Athena Scientific, 1999.

[19] T. Suzuki, "Stochastic Dual Coordinate Ascent with Alternating Direction Multiplier Method," *arXiv preprint arXiv:1311.0622*, pp. 1–26, 2013.

[20] Y. Nesterov, "Primal-dual subgradient methods for convex problems," *Mathematical Programming*, vol. 120, no. 1, pp. 221–259, Jun. 2007.

[21] A. Berman and R. J. Plemmons, *Nonnegative Matrices in the Mathematical Sciences*. Academic Press, 1979.