53rd IEEE Conference on Decision and Control
December 15-17, 2014. Los Angeles, California, USA

Online Distributed ADMM via Dual Averaging

Saghar Hosseini, Airlie Chapman, and Mehran Mesbahi

***Abstract*—This paper presents a convergence analysis on a distributed Alternating Direction Method of Multipliers (ADMM) algorithm which solves online convex optimization problems under linear constraints. The goal is to distributively optimize a global objective function over a network of decision makers. 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 to the decision makers and one known *a priori*. We extend an online ADMM algorithm to a distributed setting based on dual-averaging. We then explore 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 regret. 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 is then applied to a formation acquisition problem.**

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

I. Introduction

Many problems in engineering and information sciences can be characterized as a distributed convex optimization over networks such as multi-agent coordination, distributed estimation in sensor networks, decentralized tracking and event localization. These problems often have a composite objective function to be optimized subject to local linear constraints. The well known Alternating Direction Method of Multipliers (ADMM) [1] can solve this class of problems by splitting the variables as

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

where functions f and ϕ are convex functions, and χ and Y are convex sets.

ADMM is often considered 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 $f(x)$ might be varying with time t, denoted as $f_t(x)$, for example due to uncertainties in the environment. Further, the uncertainty in $f_t(x)$ may not be characterized by a known probability distribution. These formulations fall under the class of *online* optimization problems [2]. 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 problems, stochastic ADMM was introduced by Ouyang *et al.* [3], where an identical and independent distribution for the uncertainties in function f_t were considered and a convergence rate of $O(\frac{1}{\sqrt{T}})$ for convex functions is shown. The O-ADMM algorithms in [4], [5] are also able to provide the same convergence rate with no assumptions on the distribution of uncertainties.

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

$$\min_{\substack{x\in\mathcal{X}\\ y_1,\dots y_n\in Y}} \sum_{i=1}^{n}(f_i(x)+\phi_i(y_i)),\ \text{s.t. } A_ix+B_iy_i=c_i,\ \forall i\in[n]\,, \tag{3}$$

and a corresponding *distributed* algorithm implementation involving n agents each cooperatively solving for the global variable x and their respective local variables $y_1,\dots y_n$. Here, the functions that compose problem (3) are distributed, specifically only agent i has access to the functions f_i and ϕ_i. A special case of this formulation is the ADMM form of the consensus problem [6] where agreement is required on each agent's local variable y_i, formally

$$\min_{x\in\chi, y_1,\dots,y_n\in Y} \sum_{i=1}^{n}\phi_i(y_i),\ \text{s.t. } x=y_i,\ \text{ for all } i\in[n]\,. \tag{4}$$

The constraint set can also be reformed to represent the coupling among agents imposed by the underlying network topology. Wei and Ozdaglar [7] have 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 readily achieved distributively when, during each iteration of the algorithm, the constraint set might be violated. In other words, each agent is penalized by its local cost rather than the global cost, facilitating a rapid $O(\frac{1}{T})$ convergence rate reported in [7]. This problem has also been examined in the context of gradient based distributed optimization [8], [9], [10], where, under the global objective (3) and local objective (4), the rate of convergence of $O(\frac{1}{\sqrt{T}})$ and $O(\frac{1}{T})$ can be achieved, respectively.

In addition, Mota *et al.* [11] have studied the consensus problem in connected bipartite graphs based on distributed ADMM. Using quantitative analysis, they have shown that this algorithm requires less communication between agents compared with other algorithms to achieve a given accuracy. Deng *et al.* [12] have proposed a proximal Jacobian ADMM, which is suitable for parallel computation. However, this method requires an all-to-all communication in each iteration.

In this work, by fusing the online and distributed ADMM problems we examine the *online distributed* ADMM (OD-

ADMM) at time T:

$$\min_{\substack{x \in \mathcal{X} \\ y_1, \dots y_n \in Y}} \sum_{t=1}^{T} (\sum_{i=1}^{n} (f_{i,t}(x) + \phi_i(y_i))), \quad (5)$$

$$\text{s.t. } A_i x + B_i y_i = c_i \text{ for all } i \in [n].$$

Inspired by the O-ADMM and our previous work [10], 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), the agents in our setup are required to reach agreement on the global variable. 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 variable is revealed to the decision maker after committing to a decision, while the cost function associated with the local variables is known *a priori*. The rate of convergence of this online algorithm is shown to be $O(\frac{1}{\sqrt{T}})$ and is described by the *regret*. Regret is a metric that measures the difference between the incurred cost and the cost of the best fixed decision in hindsight. In the literature, the average regret of a "good" online algorithm is sub-linear with time.

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

II. Background and Preliminaries

In this section, we review basic concepts from graph theory and online algorithms, as well as the relevant assumptions for our analysis.

The notation 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 a 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 $\sum_{i=1}^{n} P_{ij} = \sum_{j=1}^{n} P_{ij} = 1$. For any positive integer n, the set $\{1, 2, ..., n\}$ is denoted by $[n]$. The 2-norm,1-norm and infinity norm are denoted by $||.||$, $||.||_1$, and $||.||_\infty$, respectively, and the dual 2-norm of a vector u is defined as $||u||_* = \sup_{||v||=1} \langle u, v \rangle = ||u||$. We denote the largest, second largest, and smallest singular values of Q by $\sigma_1(Q)$, $\sigma_2(Q)$ and $\sigma_n(Q)$, respectively. A function $f : \chi \to \mathbb{R}$ is called L-Lipschitz continuous if there exists a positive constant L for which

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

1) Graphs: A graph is an abstraction for representing the interactions among decision-makers, 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 decision-makers in the network, and the edge set E represents the agents' interactions, that is, 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 the function $W : E \to \mathbb{R}$. The neighborhood set of node i is defined as $N(i) = \{j \in V | (i, j) \in E\}$. One 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 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 node in-degree's d_i. If there exists a directed path between every pair of distinct vertices, the graph $\mathcal{G}$ is referred to as strongly connected.

2) Regret: In online optimization, an online algorithm generates a sequence of decisions $\langle x_t \rangle$. At iteration t, the convex cost function h_t is unknown before committing to x_t. The feedback available to the algorithm is the incurred cost $h_t(x_t)$ 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 x^* chosen with the benefit of hindsight.

Formally, the regret of an online algorithm is defined as the difference between the incurred cost $h_t(x_t)$ and the cost of the best fixed decision $h_t(x^*)$ after T iterations, i.e., $R_T = \sum_{t=1}^{T} (h_t(x_t) - h_t(x^*))$. 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 uncertainties associated with global cost only revealed to each decision-maker after it commits to a decision. We refer to [13], [14] for further discussions on online algorithms and their regret analysis.

III. Problem Statement

In this section, we consider a large scale network of agents cooperatively minimizing a global objective function. Let the communication geometry amongst the n decision-makers, or agents, be denoted by a graph $\mathcal{G} = (V, E, W)$. Each node $i \in V$ is an agent which communicates with its neighbor $j \in N(i)$ through edge $(i, j) \in E$. An equivalent online distributed convex optimization problem to (5) is as follows

$$\min_{\substack{x \in \mathcal{X} \\ y = (y_1, \dots y_n) \in Y^n}} \sum_{t=1}^{T} F_t(x, y) = \sum_{t=1}^{T} (f_t(x) + \frac{1}{n} \sum_{i=1}^{n} \phi_i(y_i)) \quad (6)$$

$$\text{s.t. } r_i(x, y_i) = A_i x + B_i y_i - c_i = 0 \text{ for all } i \in [n],$$

where $f_t(x) = \frac{1}{n} \sum_{i=1}^{n} f_{i,t}(x)$, 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 $i \in [n]$. The *local* decision made by agent i is represented by the optimization variables $x_i \in \mathcal{X}$ and $y_i \in Y$.

The matrices in the local linear constraints are denoted 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 $i \in [n]$. We assume that B_i^T is left invertible, i.e., $\sigma_{m_i}(B_i B_i^T)$ is non-zero, for all $i \in [n]$. Let $f_{i,t}$ and ϕ_i be Lipschitz continuous with Lipschitz constants L_f and L_ϕ, respectively. We assume that an optimal solution to (6) exists.

For the online framework, each decision maker i at time t selects a global variable $x_{i,t} \in \chi$ and local variable $y_{i,t} \in Y$, based on local information. The cost $f_{i,t}(x_{i,t})$ is revealed to the agent i after its local decision $x_{i,t}$ has been executed at time t.

A. Regret for Constrained Optimization

In this section we propose a measure for evaluating the performance of OD-ADMM based on variational inequalities. This measure is inspired by the convergence analysis of Douglas-Rachford ADMM [15].

First, consider the Lagrangian $\mathcal{L}_T$ for the constrained optimization problem (6), as

$$\mathcal{L}_T = \sum_{t=1}^{T}(f_t(x) + \frac{1}{n}\sum_{i=1}^{n}[\phi_i(y_i) + \langle \lambda_i, r_i(x, y_i)\rangle]), \quad (7)$$

where $\lambda = (\lambda_1, \ldots, \lambda_n)$ are the Lagrange multipliers, $x \in \mathcal{X}$ and $y_i \in Y$, and assume $\lambda_i \in \{\lambda : \|\lambda\| \leq D_\lambda\} = \mathcal{Z}$, for all $i \in [n]$. Based on first order necessary optimality conditions on the Lagrangian, the vector $w^* = (x^*, y^*, \lambda^*) \in \chi \times Y^n \times \mathcal{Z}^n = \Omega$ solves problem (6) if it satisfies the variational inequality

$$\mathcal{L}_T(x, y, \lambda^*) - \mathcal{L}_T(x^*, y^*, \lambda) \geq 0, \quad (8)$$

for all $w \in \Omega$ [16]. The inequality (8) can be expressed as

$$\sum_{t=1}^{T} f_t^\Delta(w, w^*) + \frac{1}{n}(\sum_{i=1}^{n} \phi_i^\Delta(w, w^*) + H_i^\Delta(w, w^*)) \geq 0$$

where

$$f_t^\Delta(w, w^*) = f_t(x) - f_t(x^*), \phi_i^\Delta(w, w^*) = \phi_i(y_i) - \phi_i(y_i^*)$$
$$H_i^\Delta(w, w^*) = h_{1i}^\Delta(w, w^*) + h_{2i}^\Delta(w, w^*)$$
$$h_{1i}^\Delta(w, w^*) = \left\langle x - x^*, A_i^T \lambda_i^* \right\rangle + \langle \lambda_i - \lambda_i^*, -r_i(x^*, y_i^*)\rangle$$
$$h_{2i}^\Delta(w, w^*) = \left\langle y_i - y_i^*, B_i^T \lambda_i^* \right\rangle .$$

He and Yuan ([15] Theorem 2.1) showed that a consequence of inequality (8) is that $\widetilde{w} = \left(\widetilde{x}, \widetilde{y}, \widetilde{\lambda}\right) \in \Omega$ approximately solves problem (6) with accuracy $\epsilon_T > 0$ if it satisfies

$$\sum_{t=1}^{T} f_t^\Delta(\widetilde{w}, w) + \frac{1}{n}\sum_{i=1}^{n}[\phi_i^\Delta(\widetilde{w}, w) + H_i^\Delta(\widetilde{w}, w)] \leq \epsilon_T,$$

for all $w \in \Omega$.

Analogous to the regret definition of the O-ADMM algorithm [17], we can consider a sequence of decisions $\langle w_t \rangle$, where $w_t \in \Omega$, instead of a fixed decision $\widetilde{w}$. Consequently, $\langle w_t \rangle$ approximately solves (6) with accuracy ϵ_T if

$$\sum_{t=1}^{T} f_t^\Delta(w_t, w) + \frac{1}{n}\sum_{i=1}^{n}[\phi_i^\Delta(w_t, w) + H_i^\Delta(w_t, w)] \leq \epsilon_T, \quad (9)$$

for all $w \in \Omega$, referred to as *fixed case solutions* to distinguish them from the time-varying online solutions $\langle w_t \rangle$.

In our setting, the sequence $\langle w_t \rangle$ is constructed from the distributed algorithm adopted by the agents $w_t = (x_t, y_t, \lambda_{t+1}) \in \Omega$ at time t, where $x_t = \frac{1}{n}\sum_{i=1}^{n} x_{i,t}$, $y_t = (y_{1,t}, \cdots, y_{n,t})$, and $\lambda_{t+1} = (\lambda_{1,t+1}, \cdots, \lambda_{n,t+1})$. Finally, motivated by the inclusion of regularization terms in the augmented Lagrangian method [16], the term on the left hand side of (9) is supplemented with terms of the form $\frac{\rho}{2}||r_i(x_{i,t}, y_{i,t})||^2$, where $\rho > 0$, to promote agents satisfying their own constraints. The regret is thus defined as

$$R_T = \max_{w \in \Omega} \sum_{t=1}^{T} f_t^\Delta(w_t, w) + \frac{1}{n}\sum_{i=1}^{n}[\phi_i^\Delta(w_t, w) + H_i^\Delta(w_t, w) + \frac{\rho}{2}||r_i(x_{i,t}, y_{i,t})||^2]. \quad (10)$$

Based on (9) we say that $\langle w_t \rangle$ approximately solves (6) with accuracy ϵ_T if it satisfies $R_T \leq \epsilon_T$. The online algorithm will perform as well as the best fixed case decision provided with the full sequence of cost functions *a priori*. 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 [5] and Nesterov's dual averaging algorithm [18] to provide a distributed decision-making processes for the online optimization problem discussed in §III. The proposed 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 the dual averaging update which is a gradient descent method followed by a projection step onto the constraint set.

Consider a centralized online ADMM with dual averaging updates to solve problem (2). The dual sub-gradient update is then $z_{t+1} = z_t + g_t$, where $g_t = \nabla f_t(x_t)$, followed by

$$x_{t+1} = \prod_{\chi}^{\psi}(z_{t+1}, \alpha_t). \quad (11)$$

In this case, 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

$$\prod_{\chi}^{\psi}(z_{t+1}, \alpha_t) = \arg\min_{x \in \chi}\left\{\langle z_{t+1}, x\rangle + \frac{1}{\alpha_t}\psi(x)\right\}. \quad (12)$$

Note that $\psi(x) : \chi \to \mathbb{R}$ is a proximal, continuously differentiable and strongly convex function. It acts as a regularizer to avoid oscillations in the projection step.

Finally, the algorithm minimizes the augmented Lagrangian over y as

$$y_{t+1} = \arg\min_{y \in Y}\left\{\mathcal{L}_t^{'}(x_{t+1}, y, \lambda_{t+1}) + \frac{\rho}{2}||r(x_{t+1}, y)||^2\right\},$$

where $\mathcal{L}_t^{'}$ is the Lagrangian of (2), $\rho > 0$, and $r(x, y) = Ax + By - c$. Then, the dual variable λ is updated as

$$\lambda_{t+2} = \lambda_{t+1} + \rho(Ax_{t+1} + By_{t+1} - c).$$

The distributed algorithm can be considered as an approximate ADMM by agent i via a convex combination of information provided by its neighbors $N(i)$. Specifically, the

global update step (11) can be reformulated with a distributed dual averaging method. 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 $f_{i,t}(x_{i,t})$ there must be an information flow between the agents. Therefore, we assume that the graph $\mathcal{G}$ is strongly connected. A method to construct a doubly stochastic matrix P of the required form from the Laplacian of the network is provided in our previous work ([10] Proposition 1).

The Online Distributed ADMM (OD-ADMM) is presented in Algorithm 1. The projection operator $\prod_{\chi}^{\psi}(\cdot)$ is defined as in (12).

Algorithm 1: Online Distributed ADMM (OD-ADMM)

1 Input: $\rho > 0$, $\{\alpha_t\}_{t=1}^{T}$
2 Initialize $z_{i,1} = \lambda_{i,1} = x_{i,1} = y_{i,1} = 0$ for all $i \in [n]$
3 **for** $t = 1$ **to** T **do**
4 Compute $g_i(t) \in \partial f_{i,t}(x_{i,t})$ for all $i \in [n]$
5 **foreach** *Agent* i **do**
6 $\lambda_{i,t+1} = \lambda_{i,t} + \rho r_i(x_{i,t}, y_{i,t})$
7 $z_{i,t+1} = \sum_{j\in\{N(i),i\}} P_{j,i} z_{j,t} + g_{i,t} + A_i^T \lambda_{i,t+1}$
8 $x_{i,t+1} = \prod_{\chi}^{\psi}(z_{i,t+1}, \alpha_t)$
9 $y_{i,t+1} = \operatorname{argmin}_{y\in Y}(\phi_i(y) + \lambda_{i,t+1}^T r_i(x_{i,t+1}, y) + \frac{\rho}{2}||r_i(x_{i,t+1}, y)||^2)$
10 **end**
11 **end**

Before presenting the convergence rate of the proposed OD-ADMM algorithm we provide a few preliminary remarks and definitions. The sequences $\langle z_t \rangle$, the average dual sub-gradient, and $\langle g_t \rangle$, the average sub-gradient. are defined as

$$z_t = \frac{1}{n}\sum_{i=1}^{n} z_{i,t}, \quad g_t = \frac{1}{n}\sum_{i=1}^{n} g_{i,t}. \tag{13}$$

Thus, the following update rule is introduced similar to the standard dual averaging algorithm

$$z_{t+1} = z_t + g_t + \sum_{i=1}^{n} A_i^T \lambda_{i,t+1}, \tag{14}$$

where the primal variable is

$$\theta_{t+1} = \Pi_{\chi}^{\psi}(z_{t+1}, \alpha_t). \tag{15}$$

The regret analysis can now be presented as follows with intermediate results required for the theorem's proof relegated to the Appendix, as Lemma 2, 3 and 4.

Theorem 1. *Given the sequence $\langle w_t \rangle$ generated by Algorithm 1 with $\psi(x^*) \le K^2$ and $\alpha(t) = k/\sqrt{t}$, we have*

$$R_T \le J_1 + J_2 k\sqrt{T}, \tag{16}$$

where $J_1 = \frac{D_\lambda L_\phi}{\rho n}\sum_{i=1}^{n}\frac{1}{\sigma_{m_i}(B_i)}$, $\mathcal{Q} = \frac{\sqrt{n}}{1-\sigma_2(P)}$, $\mathcal{K}^{\max} = L_\phi \max_{i\in[n]}\frac{\sigma_1(A_i)}{\sigma_{m_i}(B_i^T)}$, *and* $J_2 = (L_f + \mathcal{K}^{\max})^2(5+2\mathcal{Q}) + \frac{K^2}{k^2}$.

Proof: By optimality of Algorithm 1 line 9 and applying line 6, we have

$$\nabla_y \phi_i(y_{i,t}) = -B_i^T(\lambda_{i,t} + \rho r_i(x_{i,t}, y_{i,t})) = -B_i^T \lambda_{i,t+1},$$

and since $||\nabla_y \phi_i(y_{i,t})|| \le L_\phi$, then $||B_i^T \lambda_{i,t+1}|| \le L_\phi$. Thus,

$$||A_i^T \lambda_{i,t}|| \le \frac{\sigma_1(A_i)||B_i^T\lambda_{i,t}||}{\sigma_{m_i}(B_i^T)} \le \frac{L_\phi \sigma_1(A_i)}{\sigma_{m_i}(B_i^T)} \le \mathcal{K}^{\max}. \tag{17}$$

Since f_t is convex,

$$f_t^{\Delta}(w_t, w) \le \frac{1}{n}\sum_{i=1}^{n}(f_t(x_{i,t}) - f_t(x)) = \frac{1}{n}\sum_{i=1}^{n} f_t^{\Delta}(w_{i,t}, w),$$

where $w_{i,t} = (x_{i,t}, y_{i,t}, \lambda_{i,t+1})$ and so $R_T \le \frac{1}{n}\times$
$\sum_{i,t} f_t^{\Delta}(w_{i,t}, w) + \phi_i^{\Delta}(w_t, w) + H_i^{\Delta}(w_t, w) + \frac{\rho}{2}||r_i(x_{i,t}, y_{i,t})||^2$.

Bounding the first two terms separately, by convexity of ϕ_i, then

$$\begin{aligned}\phi_i^{\Delta}(w_t, w) &\le \langle \nabla_y \phi_i(y_{i,t+1}), y_{i,t} - y_i\rangle \\ &= -\left\langle B_i^T \lambda_{i,t+1}, y_{i,t} - y_i\right\rangle = -h_{2i}^{\Delta}(w_t, w),\end{aligned} \tag{18}$$

and as f_t is L-Lipschitz and convex, we have

$$\begin{aligned}f_t^{\Delta}(w_{i,t}, w) &= f_t(x_{i,t}) - f_t(\theta_t) + f_t(\theta_t) - f_t(x) \\ &\le L_f \left\|x_{i,t} - \theta_t\right\| + \langle g_t, \theta_t - x\rangle .\end{aligned} \tag{19}$$

In order to further bound the first term in (19) we can use Lemma 2 which implies that

$$\|x_{i,t} - \theta_t\| \le \alpha_{t-1}\|z_t - z_{i,t}\|_*. \tag{20}$$

From the integral test with $\alpha_t = k/\sqrt{t}$ it follows that[1]

$$\sum_{t=2}^{T}\alpha_{t-1} \le \sum_{t=1}^{T}\alpha_t \le 2k\sqrt{T}. \tag{21}$$

Moreover, applying Lemma 4, from (20) - (21) it follows that

$$\sum_{t=1}^{T}\|x_{i,t} - \theta_t\| \le 2k\sqrt{T}(L_f + \mathcal{K}^{\max})(\mathcal{Q}+2). \tag{22}$$

Applying Lemma 3 to the second term in (19) with (14) and (15) then

$$\begin{aligned}\sum_{t=1}^{T}\langle g_t, \theta_t - x\rangle \le \sum_{t=1}^{T}[&\frac{\alpha_t}{2}\left\| g_{t+1} + \frac{1}{n}\sum_{i=1}^{n} A_i^T\lambda_{i,t+2}\right\|_*^2 \\ &+ \frac{1}{n}\sum_{i=1}^{n}\langle \lambda_{i,t+1}, A_i(x - \theta_t)\rangle] + \frac{1}{\alpha_T}\psi(x).\end{aligned} \tag{23}$$

The first term on the right hand side of (23) is bounded by applying (17) then (21) as

$$\sum_{t=1}^{T}\frac{\alpha_t}{2}\left\| g_{t+1} + \frac{1}{n}\sum_{i=1}^{n} A_i^T\lambda_{i,t+2}\right\|_*^2 \le (L_f + \mathcal{K}^{\max})^2 k\sqrt{T}. \tag{24}$$

The second term in the bound of (23) is expanded as

$$\begin{aligned}&\langle \lambda_{i,t+1}, A_i(x - \theta_t)\rangle \\ &= \langle \lambda_{i,t+1}, A_i(x - x_t)\rangle + \langle \lambda_{i,t+1}, A_i(x_t - \theta_t)\rangle \\ &= \langle \lambda_{i,t+1}, A_i(x - x_t)\rangle + \langle \lambda_i - \lambda_{i,t+1}, -r_i(x_t, y_{i,t})\rangle \\ &\quad + \langle \lambda_{i,t+1}, A_i(x_{i,t} - \theta_t)\rangle + \langle \lambda_i - \lambda_{i,t+1}, r_i(x_{i,t}, y_{i,t})\rangle \\ &= -h_{1i}^{\Delta}(w, w^*) + \left\langle A_i^T\lambda_{i,t+1}, (x_{i,t} - \theta_t)\right\rangle \\ &\quad + \langle \lambda_i - \lambda_{i,t+1}, r_i(x_{i,t}, y_{i,t})\rangle .\end{aligned} \tag{25}$$

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

Bounding the second term of (25) by applying (17) and (22), it follows that $\sum_{t=1}^{T}\left\langle A_i^T\lambda_{i,t+1},(x_{i,t}-\theta_t)\right\rangle \leq$

$$2k\sqrt{T}\mathcal{K}^{\max}(L_f+\mathcal{K}^{\max})(\mathcal{Q}+2). \quad (26)$$

Expanding the final term of (25) by applying line 7 of Algorithm 1 and an inner product equality,[2] we obtain

$$\langle\lambda_i-\lambda_{i,t+1},r_i(x_{i,t},y_{i,t})\rangle=\frac{1}{\rho}\langle\lambda_i-\lambda_{i,t+1},\lambda_{i,t+1}-\lambda_{i,t}\rangle$$
$$=\frac{1}{2\rho}(-\|\lambda_{i,t+1}-\lambda_{i,t}\|^2+\|\lambda_i-\lambda_{i,t}\|^2-\|\lambda_i-\lambda_{i,t+1}\|^2)$$
$$=\frac{1}{2\rho}(\|\lambda_i-\lambda_{i,t}\|^2-\|\lambda_i-\lambda_{i,t+1}\|^2)-\frac{\rho}{2}\|r_i(x_{i,t},y_{i,t})\|^2.$$

Resolving the telescoping sum $\sum_{t=1}^{T}\|\lambda_i-\lambda_{i,t}\|^2-\|\lambda_i-\lambda_{i,t+1}\|^2$, using the fact $\lambda_{i,1}=0$, it follows that $\sum_{t=1}^{T}\langle\lambda_i-\lambda_{i,t+1},r_i(x_{i,t},y_{i,t})\rangle$

$$\leq\frac{1}{2\rho}(\|\lambda_i\|^2-\|\lambda_i-\lambda_{i,T+1}\|^2)-\frac{\rho}{2}\sum_{t=1}^{T}\|r_i(x_{i,t},y_{i,t})\|^2$$
$$\leq\frac{D_\lambda L_\phi}{\rho\sigma_{m_i}(B_i)}-\frac{\rho}{2}\sum_{t=1}^{T}\|r_i(x_{i,t},y_{i,t})\|^2. \quad (27)$$

Here, the final inequality follows from applying (17) and the assumption $\|\lambda_i\|\leq D_\lambda$. Substituting (26) and (27) into (25), $\frac{1}{n}\sum_{i=1}^{n}\sum_{t=1}^{T}\langle\lambda_{i,t+1},A_i(x-\theta_t)\rangle\leq$

$$J_1-\frac{1}{n}\sum_{t=1}^{T}\sum_{i=1}^{n}[h_{1i}^{\Delta}(w,w^*)+\frac{\rho}{2}||r_i(x_{i,t},y_{i,t})||^2]+$$
$$2k\sqrt{T}\mathcal{K}^{\max}(L_f+\mathcal{K}^{\max})(\mathcal{Q}+2). \quad (28)$$

Applying $\psi(x)\leq K^2$, $\alpha_T=k/\sqrt{T}$, (24) and (28) into (23) and simplifying, it then follows that $\sum_{t=1}^{T}\langle g_t,\theta_t-x\rangle\leq$

$$J_1-\frac{1}{n}\sum_{t=1}^{T}\sum_{i=1}^{n}[h_{1i}^{\Delta}(w,w^*)+\frac{\rho}{2}||r_i(x_{i,t},y_{i,t})||^2]+$$
$$k\sqrt{T}((L_f+\mathcal{K}^{\max})[5\mathcal{K}^{\max}+2\mathcal{K}^{\max}\mathcal{Q}+L_f]+\frac{K^2}{k^2}). \quad (29)$$

Combining (29) and (22) into (19) and adding (18), the result follows. ∎

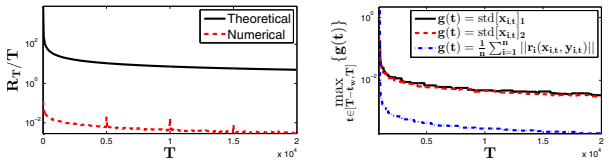
The theorem validates the "good" performance of OD-ADMM by demonstrating a sub-linear regret. In addition, it highlights the importance of the underlying topology through $\sigma_2(P)$ and the local linear constraints through $\sigma_1(A_i)$ and $\sigma_{m_i}(B_i)$. Further, a well known measure of network connectivity is the second smallest eigenvalue of the graph Laplacian $L(\mathcal{G})$ denoted by $\Lambda_2(\mathcal{G})$. Since the communication matrix P can be formed from $L(\mathcal{G})$ ([10] Proposition 1), $1-\sigma_2(P)$ is proportional to $\Lambda_2(\mathcal{G})$ implying that high network connectivity promotes good performance of the algorithm.

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

Consider a formation acquisition problem amongst n agents where the position of agent i, denoted as y_i, is restricted to a convex set $Y=[-1,1]^2$. The centroid of the formation is $x\in\mathbb{R}^2$ which is similarly constrained to $\mathcal{X}=Y$. 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 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 $\text{int}\,(S\bigcap\mathcal{X})$ is an empty set then $\phi_i(y_i)$ is convex. At each time step t, each agent i obtains a location of interest $q_{i,t}$ and the centroid is ideally located close to these locations of interest promoted through the minimization of the function $f_{i,t}(x)=\frac{1}{2}\|x-q_{i,t}\|_2^2$. The example takes the form of problem (6), where $A_i=-B_i=I_2$ for all $i\in[n]$.

Consider $S=\{(x,y)\in\mathbb{R}|\,|x|=1.5,|y|=1.5\}$ and so $\phi_i(y_i)=(2.5-\|y_i\|_\infty)^{-1}$. The relevant parameters of the ADMM algorithm are $g_i(t)=\nabla f_{i,t}(x_i)=x_i-q_{i,t}$, $k=2$, $\rho=0.5$, and $\psi(x)=\|x\|_2^2$. The remaining terms of the regret bound are $L_\phi=4/9$, $L_f=\sqrt{2}$, $\sigma_1(A_i)=\sigma_{m_i}(B_i)=1$, $D_\lambda=2$, and $K=1$.

The algorithm was applied to $n=8$ agents connected over a random graph with $\sigma_2(P)=0.78$ with c_i's selected to acquire a formation with n agents equidistant apart on the circumference of a circle of radius 0.4. Locations of interest switch at each time step between a uniform distribution over the area of a length 0.5 square centered at $(-0.75,0)$ and a Gaussian distribution with mean $(0,-0.75)$ and standard deviation $0.01I_2$, with bounds outside of $\mathcal{X}$ ignored. The sub-linear numerical regret, where w is the optimal solution of problem (6), compared to the theoretical regret appears in Figure 1a. The convergence of the global variables $x_{i,t}$ to agreement as well as the reduction of the residue over time are displayed in Figure 1b.



(a) (b)

Figure 1: (a) The regret per time R_T/T compared with its theoretical bound in Theorem 1. (b) The standard deviation of the global variable x_i and the average residue for each agent over times smoothed by taking the maximum over a $t_w=1000$ sliding window.

VI. Conclusion

In this paper, a decentralized online ADMM algorithm was developed. The problem setup conforms to a network of agents, where each agent optimizes the global objective function with access to its privately known local objective function and a linear constraint. In this algorithm the decisions are made in parallel based on local information and communication with neighboring agents. A sub-linear regret bound of $O(\sqrt{T})$ is attained for the objective function and local constraint violations. In particular, the online algorithm is competitive with respect to the best fixed decision performance in hindsight. Moreover, we highlight the role of the underlying network topology in achieving a "good" regret. The proposed algorithm was then applied to a formation

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

acquisition problem showing agreement with the theoretical results. Future work of particular interest includes exploring regret bound over a time-varying network topology, and investigating favorable graph characteristics for the online ADMM framework.

VII. Appendix

The following results can be found in [9] and [10], respectively. Thus, they are presented here with absent or abridged proofs.

Lemma 2. *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 3. *For any positive and non-increasing sequence $\alpha(t)$ and $x^* \in \chi$, $\sum_{t=1}^{T} \langle g_t, \theta_t - x^* \rangle \leq \frac{1}{\alpha_T}\psi(x^*)+$*

$$\sum_{t=1}^{T} \left\langle A_i^T \lambda_{i,t+1}, x^* - \theta_t \right\rangle + \sum_{t=2}^{T} \frac{\alpha_{t-1}}{2} ||g_t + A_i^T \lambda_{i,t+1}||_*^2,$$

where the sequence $\langle \theta_t \rangle$ is generated by (14)-(15)*.*

Proof: Applying Lemma 3 in [9], it follows that $\sum_{t=1}^{T} \left\langle g_t + A_i^T \lambda_{i,t+1}, \theta_t - x^* \right\rangle \leq \frac{1}{\alpha_T}\psi(x^*) + \sum_{t=2}^{T} \frac{\alpha_{t-1}}{2} ||g_t + A_i^T \lambda_{i,t+1}||_*^2$, and the statement of the lemma follows. ∎

Lemma 4. *For any sequences of $z_{i,t}$ and z_t generated by Algorithm* 1*, we have*

$$\|z_t - z_{i,t}\|_* \leq (L_f + \mathcal{K}^{\max})(\mathcal{Q}+2)$$

for all $i \in [n]$ and $t \in [T]$, where terms are defined in statement of Theorem 1.

Proof: Based on line 7 of Algorithm 1 we have

$$z_{i,t} = \sum_{j=1}^{n} \left[P^s\right]_{ji} z_{j,t-s} + g_{i,t-1} + A_i^T \lambda_{i,t} + \sum_{k=t-s}^{t-2} \sum_{j=1}^{n} \left[P^{t-k-1}\right]_{ji} \left(g_{j,k} + A_j^T \lambda_{j,k+1}\right).$$

In addition, z_t evolves as

$$z_t = z_{t-s} + \sum_{k=t-s}^{t-1} \sum_{j=1}^{n} \frac{1}{n}(g_{j,k} + A_j^T \lambda_{j,k+1}). \quad (30)$$

Assuming $t-s=1$, and $z_{i,1}=0$ for all $i \in [n]$ and based on (30) we have

$$z_t - z_{i,t} = \sum_{k=1}^{t-2}\sum_{j=1}^{n} (\frac{1}{n} - \left[P^{t-k-1}\right]_{ji})(g_{j,k} + A_j^T \lambda_{j,k+1}) + \frac{1}{n}\sum_{j=1}^{n} A_j^T \lambda_{j,t} - A_i^T \lambda_{i,t} + g_{t-1} - g_{i,t-1}. \quad (31)$$

Thus, the dual norm of $z_t - z_{i,t}$ can be bounded as $\|z_t - z_{i,t}\|_*$

$$\leq \sum_{k=1}^{t-2}\sum_{j=1}^{n} \left\|g_{j,k} + A_j^T \lambda_{j,k+1}\right\|_* |\frac{1}{n} - \left[P^{t-k-1}\right]_{ji}| + \frac{1}{n}\sum_{j=1}^{n} \left\|A_j^T \lambda_{j,t}\right\|_* + \left\|A_i^T \lambda_{i,t}\right\|_* + ||g_{t-1} - g_{i,t-1}||_*$$
$$\leq (\mathcal{K}^{\max} + L_f)(\sum_{k=1}^{t-2} \|P^{t-k-1} e_i - \frac{\mathbf{1}}{n}\|_1 + 2), \quad (32)$$

since $\|g_{i,t}\|_* \leq L_f$ and $||A_i^T \lambda_{i,t}||_* \leq \mathcal{K}^{\max}$ from (17), the dual norm of $z_t - z_{i,t}$ is further bounded as[3]

$$\|z_t - z_{i,t}\|_* \leq (L_f + \mathcal{K}^{\max})(\sqrt{n}\sum_{k=1}^{t-2} \sigma_2(P)^k + 2). \quad (33)$$

In addition, the second largest singular value of P is $\sigma_2(P) \leq 1$, as P is a doubly stochastic matrix [19]. Thus, $\sum_{k=1}^{t-2} \sigma_2(P)^k \leq (1-\sigma_2(P))^{-1}$ and the result follows. ∎

References

[1] P. Lions and B. Mercier, "Splitting algorithms for the sum of two nonlinear operators," *SIAM Journal on Numerical Analysis*, vol. 16, no. 6, pp. 964–979, 1979.

[2] M. Zinkevich, "Online convex programming and generalized infinitesimal gradient ascent," *International Conference on Machine Learning*, pp. 421–422, 2003.

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

[4] H. Wang and A. Banerjee, "Online alternating direction method," in *International Conference on Machine Learning*, no. 1, 2012, pp. 1–40.

[5] T. Suzuki, "Dual averaging and proximal gradient descent for online alternating direction multiplier method," *International Conference on Machine Learning*, vol. 28, pp. 392–400, 2013.

[6] S. Boyd, "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, 2010.

[7] E. Wei and A. Ozdaglar, "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.

[8] 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. 25, pp. 1041–4347, 2013.

[9] 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.

[10] S. Hosseini, A. Chapman, and M. Mesbahi, "Online distributed optimization via dual averaging," *IEEE Conference on Decision and Control*, pp. 1484–1489, 2013.

[11] 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.

[12] 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.

[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] 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.

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

[17] T. Suzuki, "Stochastic dual coordinate ascent with alternating direction multiplier method," *arXiv preprint arXiv:1311.0622*, pp. 1–26, 2013.

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

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

[3] Note for stochastic P, $\|P^t x - \frac{\mathbf{1}}{n}\|_1 \leq \sigma_2(P)^t \sqrt{n}$, where the vector x belongs to $\{x \in \mathbb{R}^n | x \geq 0, \sum_{i=1}^n x_i = 1\}$ [9].