Linear Convergence Rate for Distributed Optimization with the Alternating Direction Method of Multipliers

F. Iutzeler[1], P. Bianchi[2], Ph. Ciblat[2], and W. Hachem[2]

***Abstract*— Consider the problem of distributed optimization where a network of N agents cooperate to solve a minimization problem of the form $\inf_x \sum_{n=1}^N f_n(x)$ where function f_n is convex and known only by agent n. The Alternating Direction Method of Multipliers (ADMM) has shown to be particularly efficient to solve this kind of problem. In this paper, we assume that there exists a unique minimum $x_\star$ and that the functions f_n are twice differentiable at $x_\star$ and verify $\sum_{n=1}^N \nabla^2 f_n(x_\star) > 0$ where the inequality is taken in the positive definite ordering. Under these assumptions, we prove the linear convergence of the distributed ADMM to the consensus over $x_\star$ and derive a tight convergence rate. Finally, we give examples where one can derive the ADMM hyper-parameter ρ corresponding to the optimal rate.**

***Index Terms*— Distributed optimization, Consensus algorithms, Alternating Direction Method of Multipliers.**

I. Introduction

The problem of distributed optimization arises in a large variety of applications ranging from learning in massive datasets distributed over distant machines [1], [2], to resource allocation in communicating networks [3], [4], or statistical estimation by sensor networks [5], [6].

Consider a group of N agents seeking to solve a distributed optimization problem. Each agent n has a private convex function $f_n : \mathbb{R}^K \to \overline{\mathbb{R}} \triangleq \mathbb{R} \cup \{+\infty\}$ where $K \in \mathbb{N}^*$ is the dimension of the sensors' variables space. The goal of these agents is to solve

$$\inf_{x \in \mathbb{R}^K} \sum_{n=1}^{N} f_n(x) \tag{1}$$

in a distributed fashion. This means that the agents update a local estimate belonging to the parameter space $\mathbb{R}^K$ according to their own private function and the information it received from some other agents. The global objective is then that every sensor converges to a common value (we say that they reach consensus) which is a solution of the above problem if any.

The above mentioned computation is thus composed of two different parts: the local updates performed by the sensors using their functions and the communications between the agents to reach an agreement. It is thus usual in distributed optimization to minimize the sum of two functions, one depending on the agents cost functions and the other encompassing the communications between them, by using splitting methods [7]. Recently, the Alternating Direction Method of Multipliers (ADMM), popularized by the monograph [8], was shown to be particularly suited for distributed implementation of Problem (1) [9].

In this paper, we assume that the infimum of our problem is attained at a point $x_\star \in \mathbb{R}^K$ and that each agent function are twice differentiable at this point, and finally that $\sum_{n=1}^N \nabla^2 f_n(x_\star) > 0$ where the inequality is taken in the positive definite ordering. Under these assumptions, we show the linear (exponential) convergence of a distributed optimization algorithm based on the ADMM and, most importantly, we explicitly provide the rate of convergence. This result enables us to evaluate the impact of the ADMM hyper-parameter as well as the effects of the communication network.

In the literature, the convergence speed of the ADMM was recently shown to be $\mathcal{O}(1/k)$ when the objective functions are not necessarily smooth [10]–[14]. In the case where the function are strongly convex and have a Lipschitz continuous gradient, the linear convergence of the ADMM and ADMM-based distributed optimization algorithms were proven in [15]–[17] and some upper bounds on the convergence rate were given. [18] proved the linear convergence of ADMM-based distributed optimization under lighter assumptions but only when the step-size is small enough and no explicit rate was provided. The aim of this paper is thus i) to prove the linear convergence of an ADMM-based distributed optimization algorithm for any positive step-size, and ii) to obtain a tight asymptotic convergence rate.

First, we will state our assumptions and derive a distributed optimization algorithm based on the ADMM in Section II. Then, in Section III, we will state our main result about the linear convergence rate of the previously derived algorithm. Finally, in Section V, we will give examples of instances of the considered algorithm over some particular communication graphs. When possible, we will provide simple closed form expression of our theoretical rate with respect to the ADMM step size ρ and the second-order derivatives of the agents functions. We will also give numerical simulations of the convergence rate along with our bound.

II. Algorithm and Assumptions

A. Assumptions about the agents functions

We note $\Gamma_0(\mathbb{R}^K)$ the set of proper lower semi-continuous convex $\mathbb{R}^K \longrightarrow \overline{\mathbb{R}}$ functions [19, Chapter 9]. We will assume that the agents functions verify the following properties.

[1] Alcatel-Lucent Chair on Flexible Radio - SUPÉLEC, Gif-sur-Yvette, France. E-mail: franck.iutzeler@supelec.fr.

[2] CNRS LTCI; Telecom ParisTech, Paris, France. E-mails: forename.name@telecom-paristech.fr.

This work was partially funded by the ANR ODISSEE (ASTRID Program) and the French Defense Agency (DGA).

Assumption 1 *For any $n = 1, ..., N$, $f_n \in \Gamma_0(\mathbb{R}^K)$.*

Assumption 2 *The infimum of the problem* (1) *is attained at a point $x_\star$. Furthermore, at $x_\star$, the functions $\{f_n\}_{n=1,...,N}$ are twice differentiable and satisfy*

$$\sum_{n=1}^{N} \nabla^2 f_n(x_\star) > 0.$$

We remark that these assumptions imply that the problem minimizer $x_\star$ is unique. Please note that, contrary to most papers of the literature, the agents functions are not required to be strictly nor strongly convex. Moreover, no assumption is made towards the gradients of these functions. We only require two-fold differentiability for the agents functions and strong convexity for $\sum_{n=1}^{N} f_n$ at a local level.

B. Problem Reformulation

Now, we reformulate Problem (1) using the idea of [9] in order to take into account the network. We start by introducing some simple mathematical objects along with some notations.

First, we give each agent $n = 1, ..., N$ a variable $x(n) \in \mathbb{R}^K$, we note $x = (x(1), \dots, x(N))$ and define

$$\begin{array}{rccc} f: & \mathbb{R}^{NK} & \longrightarrow & \overline{\mathbb{R}} \\ & x & \longmapsto & f(x) = \sum_{n=1}^{N} f_n(x(n)) \end{array}$$

Given any positive integer L, let $A_1, \dots, A_L$ be a collection of subsets of $\{1, ..., N\}$ so that for all $\ell = 1, ..., L$ the cardinality of A_ℓ verifies $|A_\ell| > 1$ and we note $A_\ell = (a_{\ell,n})_{n=1,..,|A_\ell|}$. The idea of [9] is to ensure consensus over each of these subsets. Let us define for every $\ell = 1, ..., L$, $z^{(\ell)} = (z^{(\ell)}(a_{\ell,n}))_{n=1,..,|A_\ell|} \in \mathbb{R}^{|A_\ell|K}$ and $z = (z^{(1)}, ..., z^{(L)}) \in \mathbb{R}^{TK}$ with $T = \sum_{\ell=1}^{L} |A_\ell|$. For every subset A_ℓ, we define $C_{|A_\ell|}$ as the linear subspace of $\mathbb{R}^{|A_\ell|K}$ whose elements $z^{(\ell)}$ verify $z^{(\ell)}(a_{\ell,1}) = z^{(\ell)}(a_{\ell,2}) = \cdots = z^{(\ell)}(a_{\ell,|A_\ell|})$. We define

$$\begin{array}{rccc} g: & \mathbb{R}^{TK} & \longrightarrow & \overline{\mathbb{R}} \\ & z = (z^{(1)}, \dots, z^{(L)}) & \longmapsto & g(z) = \sum_1^L \imath_{C_{|A_\ell|}}(z^{(\ell)}) \end{array}$$

where $\imath_C$ is the indicator function of set C, defined to be equal to zero on C and to ∞ outside this set.

Now, for any subset A_ℓ, we define S_{A_ℓ} as the $|A_\ell| \times N$ selection matrix (its entries are zeros except for one 1 per row), the i-th row non-null coefficient being in column $j = a_{\ell,i}$. This way, for all $x \in \mathbb{R}^{NK}$, we have $(S_{A_\ell} \otimes I_K)x = (x(n))_{n \in A_\ell}$ where '$\otimes$' stands for the Kronecker product. Finally, we define M as the $T \times N$ matrix such that

$$M = \underbrace{\begin{bmatrix} S_{A_1} \\ \vdots \\ S_{A_L} \end{bmatrix}}_{\triangleq S} \otimes I_K.$$

We will now consider the following problem

$$\inf_{x \in \mathbb{R}^{NK}} f(x) + g(Mx). \tag{2}$$

Let us define the graph $\mathcal{G}$ of the relations between the subsets as $\mathcal{G} = (\{1, \dots, L\}, \mathcal{E})$ where $\{1, \dots, L\}$ are the vertices/subsets and $\mathcal{E}$ is the set of the bidirectional links between the subsets. More precisely, $\{\ell, m\} \in \mathcal{E}$ if $A_\ell \cap A_m \neq \emptyset$. Then, with the natural assumption

Assumption 3 *The following facts hold true:*

i) $\bigcup_{\ell=1}^{L} A_\ell = \{1, ..., N\}$;

ii) *The graph $\mathcal{G}$ is connected.*

we immediately get the next result with the notation $\mathbf{1}_N$ standing for the size-N vector of ones.

Lemma 1 *Under Assumption 3, $x_\star$ is a minimizer of Problem* (1) *if and only if $(x_\star, \dots, x_\star) = \mathbf{1}_N \otimes x_\star \triangleq \boldsymbol{x}_\star$ is a minimizer of Problem* (2)*.*

C. The distributed ADMM algorithm

We now recall the ADMM and derive a distributed optimization algorithm by applying it to Problem (2).

The ADMM is well suited for solving convex optimization problems of the form

$$\inf_{Mx=z} f(x) + g(z),$$

by alternately minimizing the augmented Lagrangian of the problem which is the function $\mathcal{L}_\rho : \mathbb{R}^{NK} \times \mathbb{R}^{TK} \times \mathbb{R}^{TK} \to \overline{\mathbb{R}}$ defined as

$$\mathcal{L}_\rho(x, z, \lambda) = f(x) + g(z) + \langle \lambda, Mx - z \rangle + \frac{\rho}{2} \|Mx - z\|^2$$

where $\rho > 0$ is a free parameter. The ADMM then simply consists in a alternated minimization of the Lagrangian followed by a dual gradient ascent:

$$x_{k+1} = \operatorname*{argmin}_{x \in \mathbb{R}^{NK}} \mathcal{L}_\rho(x, z_k; \lambda_k) \tag{3a}$$

$$z_{k+1} = \operatorname*{argmin}_{z \in \mathbb{R}^{TK}} \mathcal{L}_\rho(x_{k+1}, z; \lambda_k) \tag{3b}$$

$$\lambda_{k+1} = \lambda_k + \rho(Mx_{k+1} - z_{k+1}). \tag{3c}$$

The convergence of this algorithm under suitable assumption is stated in the following theorem. Proofs of this result can be found in [8, Chap. 3.2 and App. A], or in [20] with the use of monotone operators theory.

Theorem 1 *Under Assumptions 1 to 3, the sequence $\{x_k\}_{k>0}$ generated by* (3) *converges to $(x_\star, \dots, x_\star) = \boldsymbol{x}_\star$ where $x_\star$ is the solution of Problem* (1)*.*

Now, we apply the iterations of Equation (3) to Problem (2). One can see from the definition of f and $\mathcal{L}_\rho$ that the updates in the variable x can be done agent by agent depending only on its own functions and on local variables. For agent n we denote by $\sigma(n) = \{m : n \in A_m\}$ the set of the blocks it takes part in. We remark that the second step can be decomposed subset by subset; and in each subset ℓ, it enforces a consensus around a value $\bar{z}_{k+1}^{(\ell)}$. Finally, the last step can be done independently by each sensor, we will use the same notation for λ as for z. Finally, after some manipulations including the introduction of a (size $\mathbb{R}^k$) variable $\Phi_k(n)$ at each sensor n (a more detailed treatment

can be found in [8], [21], or [9]), we get the algorithm in the box below.

Distributed Optimization using the ADMM

For each agent n, initialize $\Phi_0(n) = 0$.
At iteration $k+1$:

- Every agent $n = 1, ..., N$ compute

$$x_{k+1}(n) = \underset{w\in\mathbb{R}^K}{\operatorname{argmin}} \left\{ f_n(w) + \frac{\rho|\sigma(n)|}{2} \|w - \Phi_k(n)\|^2 \right\} \tag{4}$$

- For each $\ell = 1, ..., L$, the agents of A_ℓ compute the average of the newly computed variables

$$\bar{z}_{k+1}^{(\ell)} = \frac{1}{|A_\ell|} \sum_{n\in A_\ell} x_{k+1}(n) \tag{5}$$

- Every agent $n = 1, ..., N$ computes $\forall m \in \sigma(n)$

$$\lambda_{k+1}^{(m)}(n) = \lambda_k^{(m)}(n) + \rho(x_{k+1}(n) - \bar{z}_{k+1}^{(m)}), \tag{6}$$
$$\text{and} \quad \Phi_{k+1}(n) = \frac{-1}{\rho|\sigma(n)|} \sum_{m\in\sigma(n)} \lambda_{k+1}^{(m)}(n) - \rho\bar{z}_{k+1}^{(m)}$$

As we can see in the iterations of the box, this algorithm requires the subsets to compute the average of its agents at each iteration so the agents of a same subset must be connected by some underlying communication network. In Section V, we will see examples of subsets design and associated algorithms. Finally, we also remark that each agent n only has to store its $|\sigma(m)|$ lagrange multipliers $\{\lambda_k^{(m)}\}_{m\in\sigma(n)}$ and $\Phi_{k+1}(n)$.

III. Main Result

We now state the main result of this paper after some other definitions. For any positive integer d, we note $J_d = 1/d\mathbf{1}_d\mathbf{1}_d^*$ and then, the orthogonal projection matrix to consensus in subset A_ℓ, $C_{|A_\ell|}$, is $\Pi_{|A_\ell|} = J_{|A_\ell|} \otimes I_K$. We define P as the $TK \times TK$ orthogonal projection matrix such that

$$P = \begin{bmatrix} J_{|A_1|} & & \\ & \ddots & \\ & & J_{|A_L|} \end{bmatrix} \otimes I_K.$$

We also define the $TK \times TK$ matrix

$$Q = \rho M \left(\begin{bmatrix} \nabla^2 f_1(x_\star) & & \\ & \ddots & \\ & & \nabla^2 f_N(x_\star) \end{bmatrix} + \rho M^* M \right)^{-1} M^*$$
$$= \rho M \left(\nabla^2 f(\mathbf{1}_N \otimes x_\star) + \rho M^* M \right)^{-1} M^* \tag{7}$$

Finally, we denote by $\operatorname{span}(\cdot)$ and by $\boldsymbol{r}(\cdot)$ the column space and the spectral radius of a matrix respectively.

Theorem 2 *Let Assumptions 1 to 3 hold true. Then the following facts hold true:*

i) $\alpha = \boldsymbol{r}((\Pi_{\operatorname{span}(P+Q)} - P - Q)(I - 2P)) < 1$ *where* $\Pi_{\operatorname{span}(P+Q)}$ *is the orthogonal projection matrix on* $\operatorname{span}(P + Q)$*;*

ii) *For any initial value* (z_0, λ_0) *of the Distributed Optimization using the ADMM, one has*

$$\limsup_{k\to\infty} \frac{1}{k} \log \|x_k - \boldsymbol{x}_\star\| \leq \log \alpha;$$

This theorem states that when running Distributed Optimization using the ADMM under assumptions 1 to 3, the error between the iterates $\{x_k\}_{k>0}$ and the searched optimum $\boldsymbol{x}_\star$ decreases exponentially. It also provides an upper-bound α on the convergence rate which will be showed to be tight in the numerical illustrations.

IV. Proof of Theorem 2

As advocated in [20] (or more recently in [21]), monotone operators are particularly suited for the analysis of convex optimization algorithms including the ADMM. In our context, this advocates to see our minimization problem as a fixed-point problem on $\zeta = \lambda + \rho z$. Note that there is a one-to-one correspondence between ζ and the couple (λ, z) as $z = 1/\rho P\zeta$ and $\lambda = (I - P)\zeta$.

A. Recursion in the quadratic case

Using Fermat's rule [19, Chap. 26-27], one can derive that the first update of the k-th iteration of Distributed Optimization using the ADMM is obtained by solving

$$0 \in \rho M^*(Mx_{k+1} - z_k) + \partial f(x_{k+1}) + M^*\lambda_k.$$

In the quadratic case, $\partial f(x) = \nabla f(x) = \nabla^2 f(\boldsymbol{x}_\star)x + c$ with $c = \nabla^2 f(\boldsymbol{x}_\star)\boldsymbol{x}_\star$ thus

$$\begin{aligned} 0 &= \rho M^* M x_{k+1} + \nabla^2 f(\boldsymbol{x}_\star)x_{k+1} + c + M^*(\lambda_k - \rho z_k) \\ x_{k+1} &= -H^{-1}M^*(\lambda_k - \rho z_k) - H^{-1}c \\ &= -H^{-1}M^*(I - 2P)\zeta_k - H^{-1}c \end{aligned} \tag{8}$$

where $H = \rho M^* M + \nabla^2 f(\boldsymbol{x}_\star)$ is a symmetric positive definite (thus invertible) matrix. Indeed, $M^*M = (S^*S) \otimes I_K > 0$ as it is diagonal from the fact that S is a selection matrix and $(S^*S)_{i,i} > 0$ from Assumption 3ii. Furthermore, $\nabla^2 f(\boldsymbol{x}_\star) \geq 0$ as f is convex from Assumption 1.

Now, as we can see in Eq. (6), $\zeta_{k+1} = \lambda_{k+1} + \rho z_{k+1} = \lambda_k + \rho M x_{k+1}$ so we have

$$\begin{aligned} \zeta_{k+1} &= \lambda_k + \rho M x_{k+1} \\ &= (I - P)\zeta_k - \rho M H^{-1} M^*(I - 2P)\zeta_k - \rho M H^{-1} c \\ &= (I - P - \rho M H^{-1} M^*)(I - 2P)\zeta_k - \rho M H^{-1} c \\ &= (I - P - Q)(I - 2P)\zeta_k - \rho M H^{-1} c \end{aligned}$$

as $(I-P)(I-2P) = I-P$. Taking $R = (I-(P+Q))(I-2P)$ and $d = -\rho M H^{-1} c$, we have to study the fixed points of the transformation

$$\zeta_{k+1} = R\zeta_k + d. \tag{9}$$

We know from [22] that under Assumptions 1 to 3 such a fixed point exists and that the sequence $\{\zeta_k\}_{k>0}$ converges to a fixed point $\bar{\zeta} = \bar{\lambda} + \rho\bar{z}$, and that $\bar{x}$ obtained by Eq. (4) is necessarily equal to $\boldsymbol{x}_\star$.

Hence, for any fixed point $\bar{\zeta}$ of (9), we get

$$\zeta_k - \bar{\zeta} = R(\zeta_{k-1} - \bar{\zeta}) = ... = R^k(\zeta_0 - \bar{\zeta}) \qquad (10)$$

and thus

$$\begin{aligned} x_{k+1} - \boldsymbol{x}_\star &= -H^{-1}M^*(I-2P)(\zeta_k - \bar{\zeta}) \\ &= -H^{-1}M^*(I-2P)R^k(\zeta_0 - \bar{\zeta}). \end{aligned} \qquad (11)$$

Let us now analyse the properties of the recursion matrix R.

B. Analysis of the recursion matrix R

In this part, we will i) prove that the spectral radius of R is less or equal to 1, with 1 the only possible eigenvalue of modulus 1; ii) characterize this stable space; and iii) show that $H^{-1}M^*(I-2P)R^k$ vanishes exponentially.

▷ *i)* We have $\boldsymbol{r}(R) \leq \|R\| = \|I - P - Q\|$ as $I - 2P$ is a reflexion. For any $w \in \mathbb{R}^{TK}$, we have $|w^*(I-P)w - w^*Qw| \leq \|w\|^2$ thus all the eigenvalues of $(I - P - Q)$ are lower than 1 in absolute value, hence its singular values are lower than 1 and so $\boldsymbol{r}(R) \leq \|I - P - Q\| \leq 1$. Let $e^{i\theta}$ $\theta \in [0, 2\pi[$ be an eigenvalue of R and w be an associated eigenvector. We have $\|Rw\| = 1$ so $w^*(I-2P)(I-P-Q)^2(I-2P)w = 1$ so $w' = (I-2P)w$ can be decomposed as $u + v$ where u (resp. v) is an eigenvector of $(I - P - Q)$ associated to eigenvalue 1 (resp. -1) as it is a real symmetric matrix. We obviously have $u^*Pu + u^*Qu = 0$ hence, as P and U are positive semi-definite, $Pu = Qu = 0$. By the same argument, one can have that $Pv = Qv = v$. Then, $Rw = e^{i\theta}w$ corresponds to $(I - P - Q)(u + v) = e^{i\theta}(I-2P)(u+v)$ so $u - v = e^{i\theta}(u - v)$ which means that $e^{i\theta} = 1$. So, the spectral radius of R is at most 1 and the only possible eigenvalue with this modulus is 1.

▷ *ii)* Let us denote by $\mathcal{N}$ the eigenspace of R corresponding to eigenvalue 1. Let $w \in \mathcal{N}$, we have from above that $w = (I - 2P)w' = u - v$ with $Pu = Qu = 0$ and $Pv = Qv = v$. Let us now remark that according to the last equality v belong to the span of P and the span of Q (which is also the span of M) so the proof of Lemma 1 can be reproduced leading to $v = a\mathbf{1}_{TK}$ for some $a \in \mathbb{R}$ and as $Pv = v$, $a = 0$ so that $v = 0$. This mean that $\mathcal{N} \subset \ker(P) \cap \ker(Q) = \ker(P + Q)$. The converse is obviously true so we have $\mathcal{N} = \ker(P + Q)$. Finally, defining $\Pi_{\mathcal{N}}$ as the orthogonal projection to $\mathcal{N}$, we have $R - \Pi_{\mathcal{N}} = (\Pi_{\text{span}(Q+P)} - (P + Q))(I - 2P)$ with $\Pi_{\text{span}(Q+P)}$ the orthogonal projection to the span of $P + Q$ and $\alpha \triangleq \boldsymbol{r}[(\Pi_{\text{span}(Q+P)} - (P + Q))(I - 2P)] < 1$.

▷ *iii)* We just have to remind that as $\mathcal{N} = \ker(P + Q)$ and $\text{span}(Q) = \text{span}(M)$, $M^*(I - 2P)\Pi_{\mathcal{N}} = 0$, thus we have

$$\begin{aligned} x_{k+1} - \boldsymbol{x}_\star &= -H^{-1}M^*(I-2P)R(\zeta_{k-1} - \bar{\zeta}) \\ &= -H^{-1}M^*(I-2P)(\Pi_{\mathcal{N}} + (R - \Pi_{\mathcal{N}}))(\zeta_{k-1} - \bar{\zeta}) \\ &= -H^{-1}M^*(I-2P)(R - \Pi_{\mathcal{N}})^k(\zeta_0 - \bar{\zeta}) \end{aligned}$$

hence $\|x_{k+1} - \boldsymbol{x}_\star\|$ will vanish exponentially at rate $\alpha < 1$.

C. General case

Due to the lack of space only the main steps of the proof are given here. It relies on the fact that for any x close enough to $\boldsymbol{x}_\star$, we have

$$\nabla f(x) = \nabla f(\boldsymbol{x}_\star) + \nabla^2 f(\boldsymbol{x}_\star)(x - \boldsymbol{x}_\star) + E(x - \boldsymbol{x}_\star)$$

where $\|E(x)\|/\|x\| \to 0$ as $x \to 0$. Eq. (8) then becomes

$$x_{k+1} = -H^{-1}M^*(I-2P)\zeta_k - H^{-1}c - H^{-1}E(x_{k+1} - \boldsymbol{x}_\star)$$

with $c = \nabla f(\boldsymbol{x}_\star) - \nabla^2 f(\boldsymbol{x}_\star)\boldsymbol{x}_\star$. Thus, similar to Eq. (9), we have

$$\zeta_{k+1} = R\zeta_k + d - \rho M H^{-1}E(x_{k+1} - \boldsymbol{x}_\star)$$

and so

$$\zeta_k - \zeta_\star = R^k(\zeta_0 - \zeta_\star) - \rho \sum_{\ell=1}^{k} R^{k-\ell} M H^{-1} E(x_\ell - \boldsymbol{x}_\star).$$

By applying the same reasoning, and with some more analysis, we are able to prove that same result holds for non-quadratic functions.

V. Examples and Numerical Illustrations

The aim of this section is threefold: i) giving examples of instances of Distributed Optimization with ADMM on particular communication network topologies; ii) providing simple forms of the convergence rate over these topologies; and iii) showing that this rate is tight in our examples. For simplicity, we will assume that the sensors variable space size K is equal to one, so that $M = S \otimes I_K = S$ and $P = \Pi$. Moreover, we will generally assume that $\nabla^2 f(x_\star) = \sigma_\star^2 I_N$ in order to obtain more simple and insightful expressions.

A. The centralized network

We consider here the case where $L = 1$, so that $A_1 = \{1, ..., N\}$. With this setup, every agent computes its new variable x with Eq. (4) then a dedicated fusion center computes and broadcast the average of these values which gives the parallel optimization algorithm with a centralized communication step described in [8, Chap. 7].

In this case, we have $M = I_N$, $P = N^{-1}\mathbf{1}_N\mathbf{1}_N^*$ and $Q = \frac{\rho}{\sigma_\star^2+\rho} I_N$; obviously, $\Pi_{\text{span}(P+Q)} = I_N$. Thus, our rate of convergence α is the specral radius of

$$\begin{aligned} R &\triangleq \Big(I_N - P - Q\Big)\Big(I_N - 2P\Big) \\ &= \frac{\sigma_\star^2}{\sigma_\star^2 + \rho}(I_N - P) + \frac{\rho}{\sigma_\star^2 + \rho}P\,. \end{aligned}$$

It is easy to see that R has two distinct eigenvalues: one equal to $\frac{\sigma_\star^2}{\sigma_\star^2+\rho}$ and one equal to $\frac{\rho}{\sigma_\star^2+\rho}$. We thus have the following corollary.

Corollary 1 (Centralized network) *Under the stated assumptions, the rate is given by:*

$$\alpha = \frac{\max(\rho, \sigma_\star^2)}{\rho + \sigma_\star^2}\,. \qquad (12)$$

In particular, $\alpha \geq \frac{1}{2}$ *with equality iff* $\rho = \sigma_\star^2$.

This result means that the optimal rate is $1/2$ and is obtained when the parameter $\rho = \sigma_\star^2$. We also remark that α goes to 1 when ρ goes to 0 or to ∞. Thus, there is a trade-off in the choice of ρ that enables to obtain significantly better convergence rate, as discussed in [23] for quadratic problems. And even if $\sigma_\star^2$ is generally unknown, this result provides us a guideline in the choosing of step-size.

In Figure 1, we plot the rate α as a function of the step-size ρ of the algorithm in the case of a 5-agent centralized communication network when $\nabla^2 f(x_\star) = \sigma_\star^2 I_N$ and $\sigma_\star^2 = 16$. We observe the above-mentioned trade-off in the step size ρ.

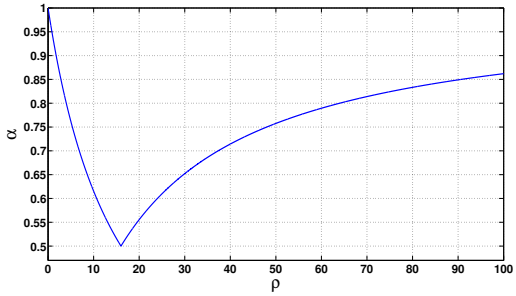

Fig. 1. Rate α as a function of ρ - Centralized network - $\sigma_\star^2 = 16$.

We now address in Figure 2 the case where the second order derivatives are not equal. Still with a 5-agent centralized communication network, we take the values of $f_n''(x_\star)$ for all agents n equal to 4, 9, 16, 25 and 39. We still observe a trade-off in the step-size but now the optimal parameter is not as clear as before.

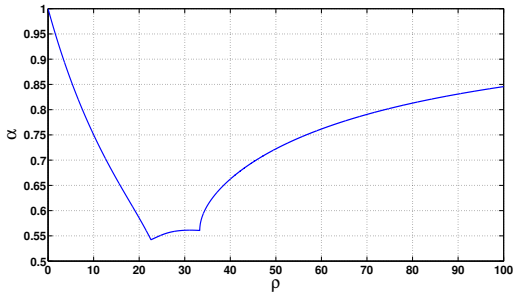

Fig. 2. Rate α as a function of ρ - Centralized network - $N = 5$ - Distinct second order derivatives.

B. The ring network

We now consider the case where the agents communicate through a ring communication network (see Figure 3). In this framework, the $N \geq 3$ agents are connected with two agents such that $E = \{\{1,2\}, \{2,3\}, \ldots, \{N-1, N\}, \{N, 1\}\}$ is the set of edges of the underlying communication network. We thus have $L = N$, $A_\ell = \{\ell, \ell+1\}$ for all $\ell = 1, ..., L-1$, and $A_L = \{N, 1\}$.

We define for simplicity $s_N = \sin(2\pi/N)$, $c_N = \cos(2\pi/N)$ and $t_N = \tan(2\pi/N)$. After some omitted computations based on the fact that R is a block-circulant matrix, we get the following result.

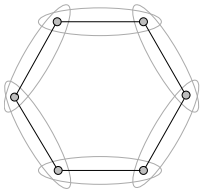

Fig. 3. Example of a ring network with $N = 6$. Sets A_ℓ are represented by the ellipses.

Corollary 2 (Ring network) *Under the stated assumptions, the rate $\alpha = \alpha(\rho)$ is given by the following expression.*

- *If* $\rho \leq \dfrac{\sigma_\star^2}{2s_N}$, *then*

$$\alpha = \frac{\sigma_\star^2 + 2\rho\,(1 + c_N) + \sqrt{\sigma_\star^4 - 4\rho^2 s_N^2}}{2(\sigma_\star^2 + 2\rho)},$$

- *If* $\rho \in \left[\dfrac{\sigma_\star^2}{2s_N}, \dfrac{\sigma_\star^2}{2t_N^2}\right]$, *then*

$$\alpha = \sqrt{\frac{\rho(1 + c_N)}{\sigma_\star^2 + 2\rho}},$$

- *If* $\rho \geq \frac{\sigma_\star^2}{2t_N^2}$ *then*

$$\alpha = \frac{2\rho}{\sigma_\star^2 + 2\rho}.$$

For any $N \geq 3$, the function $\rho \mapsto \alpha(\rho)$ is continuous, decreasing on $(0, \frac{\sigma_\star^2}{2s_N}]$, increasing on $[\frac{\sigma_\star^2}{2s_N}, +\infty)$. Finally,

$$\alpha \geq \frac{1}{\sqrt{2}} \sqrt{\frac{1 + c_N}{1 + s_N}}$$

with equality iff $\rho = \frac{\sigma_\star^2}{2s_N}$.

The optimal step-size $\rho_{opt} = \frac{\sigma_\star^2}{2s_N}$ is equal to $\frac{\sigma_\star^2 N}{4\pi} + o(N)$ which suggests that the step-size should increase at the same rate as N, contrary to the centralized case above where the optimal step-size was equal to $\sigma_\star^2$ for any number of agents.

In Figure 4, we plot the rate α as a function of the step-size ρ and the number of a agents N in the case of a ring communication network when $\nabla^2 f(x_\star) = \sigma_\star^2 I_N$ and $\sigma_\star^2 = 16$. We observe the above-mentioned trade-off in the step size ρ and the fact that the optimal step-size varies with the number of agents.

C. Arbitrary Network

Finally, let us compare our theoretical result with the performance of Distributed Optimization with the ADMM by simulation. We set $N = 5$ and consider edges $E = \{\{1,2\}, \{2,3\}, \{3,4\}, \{4,5\}, \{5,3\}\}$. For each edge, we consider the block formed by the two linked vertices. We now illustrate our theoretical rate α with numerical simulations.

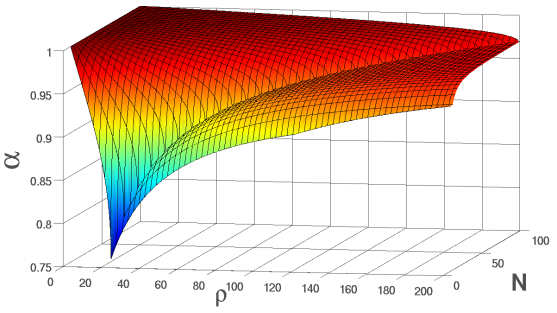

Fig. 4. Rate α as a function of ρ and N - Ring network - $\sigma_\star^2 = 16$.

In Figure 5, we plot $k^{-1}\log\|x_k - x_\star\|$ as a function of the number of iterations k along with our bound α for $\rho = 1$. The agents functions are taken quadratic. In that case, we compare ourselves with the bound of Shi et al. [17]. The functions f_n are defined as $f_n(x) = a_n(x - b_n)^2$ where $a_1, \ldots, a_n$ and $b_1, \ldots, b_n$ are respectively equal to 0.1, 0.5, 1, 2, 10 and -2, -1, 0, 1, 2. We observe that our convergence rate is tight as it fits the empirical performance of the considered algorithm, whereas a gap exists between the latter and the bound of [17]. We also observe that the asymptotic regime is attained after a very moderate number of iterations (around 40).

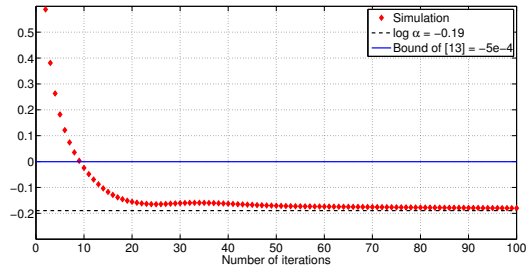

Fig. 5. $k^{-1}\log\|x_k - x_\star\|$ as a function of the number of iterations k - $N = 5$ - Quadratic functions

VI. Conclusion

In this paper, we focused on the convergence rate of a distributed optimization algorithm based on the ADMM to find a minimizer $\boldsymbol{x}_\star$ of the problem $\inf_x \sum_{n=1}^N f_n(x)$ where f_n is agent n convex private function. Assuming that the functions are twice differentiable at $\boldsymbol{x}_\star$, we gave an explicit characterization of the linear convergence rate through the spectral radius of a matrix depending on the Hessian of the functions at $\boldsymbol{x}_\star$ and on the communication network.

References

[1] P. A. Forero, A. Cano, and G. B. Giannakis, "Consensus-based distributed support vector machines," *The Journal of Machine Learning Research*, vol. 99, pp. 1663–1707, 2010.

[2] A. Agarwal, O. Chapelle, M. Dudík, and J. Langford, "A reliable effective terascale linear learning system," *arXiv e-print*, 2011.

[3] M. Chiang, S. H. Low, A. R. Calderbank, and J. C. Doyle, "Layering as optimization decomposition: A mathematical theory of network architectures," *Proceedings of the IEEE*, vol. 95, no. 1, pp. 255–312, 2007.

[4] P. Bianchi and J. Jakubowicz, "Convergence of a multi-agent projected stochastic gradient algorithm for non-convex optimization," *IEEE Trans. on Automatic Control*, vol. 58, no. 2, pp. 391–405, Feb. 2013.

[5] S.S. Ram, V.V. Veeravalli, and A. Nedic, "Distributed and recursive parameter estimation in parametrized linear state-space models," *IEEE Trans. on Automatic Control*, vol. 55, no. 2, pp. 488–492, 2010.

[6] P. Bianchi, G. Fort, and W. Hachem, "Performance of a distributed stochastic approximation algorithm," *IEEE Trans. on Information Theory*, vol. 59, no. 11, pp. 7405–7418, November 2013.

[7] P. L. Combettes and J.-C. Pesquet, "Proximal splitting methods in signal processing," in *Fixed-Point Algorithms for Inverse Problems in Science and Engineering*, pp. 185–212. Springer, 2011.

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

[9] I.D. Schizas, A. Ribeiro, and G.B. Giannakis, "Consensus in ad hoc WSNs with noisy links - Part I: Distributed estimation of deterministic signals," *IEEE Trans. on Signal Processing*, vol. 56, no. 1, pp. 350–364, 2008.

[10] R. D. C. Monteiro and B. F. Svaiter, "Iteration-complexity of block-decomposition algorithms and the Alternating Direction Method of Multipliers," *Optimization-online preprint*, vol. 2713, pp. 1, 2010.

[11] 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.

[12] E. Wei and A. Ozdaglar, "Distributed Alternating Direction Method of Multipliers," in *Proc. IEEE 51st Annual Conference on Decision and Control (CDC)*, 2012, pp. 5445–5450.

[13] E. Wei and A. Ozdaglar, "On the $O(1/k)$ convergence of asynchronous distributed Alternating Direction Method of Multipliers," *arXiv e-prints*, 2013.

[14] D. Goldfarb, S. Ma, and K. Scheinberg, "Fast alternating linearization methods for minimizing the sum of two convex functions," *Mathematical Programming*, pp. 1–34, 2012.

[15] W. Deng and W. Yin, "On the global and linear convergence of the generalized Alternating Direction Method of Multipliers," Tech. Rep., Rice University, 2012.

[16] D. Jakovetic, J. M. F. Moura, and J. Xavier, "Linear Convergence Rate of a Class of Distributed Augmented Lagrangian Algorithms," *ArXiv e-prints*, July 2013.

[17] W. Shi, Q. Ling, K. Yuan, G. Wu, and W. Yin, "On the Linear Convergence of the ADMM in Decentralized Consensus Optimization," *ArXiv e-prints*, July 2013.

[18] M. Hong and Z.-Q. Luo, "On the Linear Convergence of the Alternating Direction Method of Multipliers," *ArXiv e-prints*, Mar. 2013.

[19] H. H. Bauschke and P. L. Combettes, *Convex analysis and monotone operator theory in Hilbert spaces*, CMS Books in Mathematics/Ouvrages de Mathématiques de la SMC. Springer, New York, 2011.

[20] J. Eckstein and D. P. Bertsekas, "On the Douglas-Rachford splitting method and the proximal point algorithm for maximal monotone operators," *Mathematical Programming*, vol. 55, no. 1-3, pp. 293–318, 1992.

[21] F. Iutzeler, P. Bianchi, Ph. Ciblat, and W. Hachem, "Asynchronous distributed optimization using a randomized Alternating Direction Method of Multipliers," in *Proc. IEEE Conf. Decision and Control (CDC)*, Florence, Italy, Dec. 2013.

[22] P.-L. 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.

[23] E. Ghadimi, A. Teixeira, I. Shames, and M. Johansson, "Optimal parameter selection for the alternating direction method of multipliers (ADMM): quadratic problems," *ArXiv e-print*, 2013.