Explicit Convergence Rate of a Distributed Alternating Direction Method of Multipliers

F. Iutzeler, P. Bianchi, Ph. Ciblat, and W. Hachem

***Abstract*—Consider a set of N agents seeking to solve distributively the minimization problem $\inf_x \sum_{n=1}^N f_n(x)$ where the convex functions f_n are local to the agents. The popular Alternating Direction Method of Multipliers has the potential to handle distributed optimization problems of this kind. We provide a general reformulation of the problem and obtain a class of distributed algorithms which encompass various network architectures. The rate of convergence of our method is considered. It is assumed that the infimum of the problem is reached at a point $x_\star$, the functions f_n are twice differentiable at this point and $\sum \nabla^2 f_n(x_\star) > 0$ in the positive definite ordering of symmetric matrices. With these assumptions, it is shown that the convergence to the consensus $x_\star$ is linear and the exact rate is provided. Application examples where this rate can be optimized with respect to the ADMM free parameter ρ are also given.**

***Index Terms*—Distributed optimization, Consensus algorithms, Alternating Direction Method of Multipliers, Linear convergence, Convergence rate.**

I. Introduction

Consider a set of $N > 1$ computing agents that seek to solve collectively a minimization problem. Given that Agent n has at its disposal a private convex cost function $f_n : \mathbb{R}^K \to (-\infty, \infty]$ where K is some positive integer, the purpose of the agents is to solve distributively the minimization problem

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

A distributive (or decentralized) scheme is meant here to be an iterative procedure where at a each iteration, each agent updates a local estimate in the parameter space $\mathbb{R}^K$ based on the sole knowledge of this agent's private cost function and on a piece of information it received from its neighbors through some communication network. Eventually, the local estimates will converge to a common value (or consensus) which is a minimizer (assumed to exist) of the aggregate cost function $\sum f_n$.

Instances of this problem appear in learning applications where massive training data sets are distributed over a network and processed by distinct machines [1], [2], in resource allocation problems for communication networks [3], [4], or in statistical estimation problems by sensor networks [5], [6].

The proximal splitting methods [7] have recently attracted a large interest in the fields of statistics, signal processing and communication theory thanks to their convergence properties and to their ability to deal with large scale and decentralized problems. Among these, one of the most emblematic is the Alternate Direction Method of Multipliers (ADMM). In [8], Schizas *et al.* showed that ADMM easily lends itself to a distributed implementation of Problem (1). Since then, distributed versions of ADMM applied to consensus problems have been explored in many works (see [9]–[14] as a non-exhaustive list). In this paper, we provide a general framework inspired from [8] which allows to distribute an optimization problem on a set of agents. From a formal point of view, we do *not* assume the existence of a pre-existing "graph" whose edges would correspond to pairs of communicating agents. Instead, our framework relies on the introduction of components $A_1, \dots, A_L$, each of which is a subset of agents. *i)* In the case where the A_ℓ's are pairs of agents, our algorithm will involve pairwise communications between agents, as in *e.g.* [11]. *ii)* Identifying the A_ℓ's with larger sets of agents (clusters), our algorithm will be distributed at the *cluster* level. For instance, our framework encompasses the case of loosely coupled computer clusters composed of tighly coupled parallel machines. *iii)* Finally, when the collection of components $A_1, \dots, A_L$ is reduced to a single set $A_1 = \{1, \dots, N\}$ (that is, $L = 1$), our algorithm reduces to the parallel ADMM algorithm described in [15, Chapter 7], in which all agents output are reduced in a centralized fashion at each iteration of the algorithm. Otherwise stated, our framework yields a continuum of algorithms ranging from a fully centralized to a fully distributed setting.

The main contribution of this paper deals with the rate of convergence of ADMM in the framework of Problem (1). It is assumed that the infimum of Problem (1) is attained at a point $x_\star$, the functions f_n are twice differentiable at this point, and $\sum \nabla^2 f_n(x_\star) > 0$ in the positive definite ordering of symmetric matrices. With these assumptions, the linear convergence of the ADMM iterates is shown, and most of all, their convergence rate is explicitly provided. Our result potentially allows to evaluate the impact of the communication network on the performance, as well as the effect of the step-size. Application examples where the step-size can be optimized are also given.

The method behind the proof is as follows. We first assume that the functions f_n are quadratic. In that case, an ADMM iteration boils down to an affine transformation that we denote as $\zeta_{k+1} = R\zeta_k + d$. These iterates converge at an exponential rate that can be explicitly obtained through an analysis of the eigenstructure of the matrix R. Turning to the general case, an ADMM iteration for k large enough is shown to be a perturbed version of an affine transformation similar to the quadratic case. A close look at the perturbation terms shows that they

CNRS LTCI; Telecom ParisTech. 46 rue Barrault, 75013 Paris, France. E-mails: forename.name@telecom-paristech.fr. This work was partially granted by the French Defense Agency (DGA) and by the Telecom/Eurecom Carnot Institute.

lie in such an eigenspace of R that the analysis of the quadratic case remains essentially effective.

Beyond the framework of distributed optimization, we believe that our technique can be used to characterize the rate of convergence of ADMM in more general constrained minimization settings where the objective function is smooth in a neighborhood of the solution.

The ADMM rate of convergence was recently investigated in [9], [16]–[19] where the $\mathcal{O}(1/k)$ convergence rate was established in the case where the objective functions are not necessarily smooth.

The authors of [20] consider the problem $\min_{x:Ax+By=c} f(x) + g(y)$ where one of the two objective functions is strongly convex and has a Lipschitz continuous gradient. They establish the linear convergence of the iterates and provide upper bounds on the rate of convergence. The works [21] considers the quadratic or linear problem $\min_{x:Ax=b,x\geq 0} x^*Qx + c^*x$ where Q is a symmetric positive semidefinite matrix that may be equal to zero. The linear convergence of ADMM near the solution is established. A similar problem is investigated in [22] where an upperbound on the decay rate is provided, along with the step size which minimizes the latter upperbound.

The distributed consensus problem considered in this paper was also studied by [11], [12] and [13]. The algorithm studied by [12] strongly relies on the introduction of an inner loop at each iteration of the algorithm. The authors of [13] focus on quadratic programming and introduce a specific type of preconditioning for analysis purposes which also modifies the structure of the algorithm. Hence, both algorithms in [12] and [13] differ from the natural ADMM of interest in [11] and in the present paper. The authors of [11] prove the linear convergence of ADMM in a distributed setting, and provide an upper bound on the norm of the primal error. The bound of [11] is moreover uniform w.r.t. the choice of the functions f_n on a class of strongly convex functions with Lipschitz continuous gradients. However, work is needed to fill the substantial gap between the bound of [11] and the practice. The aim of this paper is to obtain an exact and informative characterization of the convergence rate. In addition, the proof of [11] relies on the assumption that the functions f_n are smooth strongly convex functions with Lipschitz continuous gradients, whereas the present paper relies on weaker assumptions.

Finally, let us mention the recent preprint [23] that considers the non smooth case. Using an approach similar to the one used in [24], the linear convergence of the iterates is established in the case where the step-size for updating the multipliers is small enough. No explicit convergence rate is provided.

After setting our assumptions in Section II, we show how Problem (1) can be distributively solved by ADMM after being adequately reformulated. We then state our main convergence result in Section III. In Section IV, we provide an illustration of our result in some special cases where the rate admits a simple and informative expression. The main result is proven in Section V. In Section VI, some numerical illustrations are provided. The conclusion is provided in Section VII.

II. Assumptions, Algorithm description

A. Assumptions and problem reformulation

Let us denote by $\Gamma_0(\mathbb{R}^K)$ the set of proper, lower semicontinuous, and convex functions from $\mathbb{R}^K$ to $(-\infty,\infty]$ where K is an integer. The proximity operator of a function $h \in \Gamma_0(\mathbb{R}^n)$ is the mapping defined on $\mathbb{R}^n \to \mathbb{R}^n$ by

$$\mathrm{prox}_h(x) = \operatorname*{argmin}_w \Big(h(w) + \frac{1}{2}\|w-x\|^2 \Big).$$

Denote as $\mathcal{A} = \{1,\dots,N\}$ the set of agents. The assumptions on the functions f_n considered in this paper are the following:

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

Assumption 2 *The infimum of the problem* (1) *is attained at a point $x_\star$. At $x_\star$, the functions f_n are twice differentiable and their Hessian matrices satisfy*

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

These assumptions clearly imply that the minimizer $x_\star$ is unique. Observe that the functions f_n are not required to be strictly or strongly convex at an individual level. Moreover, no global property of the gradients such as the existence or the Lipschitz continuity is assumed. We only require the two-fold differentiability of the functions f_n and the strong convexity of $\sum_n f_n(x)$ at a local level.

Along the idea of [8], we now provide another formulation of Problem (1) that will lead us to a distributed optimization algorithm. Thanks to Assumption 3 below, the two formulations will be shown to be equivalent.

We introduce some notations. Given any positive integer ℓ, an element x of $\mathbb{R}^{\ell K}$ will be often denoted as $x = (x(1),\dots,x(\ell))$ where $x(m) \in \mathbb{R}^K$ for $m = 1,\dots,\ell$. Let C_ℓ be the linear subspace of $\mathbb{R}^{\ell K}$ whose elements $x = (x(1),\dots,x(\ell))$ satisfy $x(1) = x(2) = \cdots = x(\ell)$. Denoting by $\mathbf{1}_\ell$ the $\ell \times 1$ vector of ones and by $\otimes$ the Kronecker product, the orthogonal projection matrix on this subspace is $P_\ell = J_\ell \otimes I_K$ where $J_\ell = \ell^{-1}\mathbf{1}_\ell\mathbf{1}_\ell^*$.

Given a positive integer L, let $A_1,\dots,A_L$ be a collection of subsets of $\mathcal{A}$ such that the cardinality of any set A_ℓ satisfies $|A_\ell| > 1$. Define the functions

$$\begin{array}{rccl} f: & \mathbb{R}^{NK} & \longrightarrow & (-\infty,\infty] \\ & x & \longmapsto & f(x) = \sum_1^N f_n(x(n)) \end{array}$$

and

$$\begin{array}{rccl} g: & \mathbb{R}^{|A_1|K} \times \cdots \times \mathbb{R}^{|A_L|K} & \longrightarrow & (-\infty,\infty] \\ & z = (z^{(1)},\dots,z^{(L)}) & \longmapsto & g(z) = \sum_1^L \imath_{C_{|A_\ell|}}(z^{(\ell)}) \end{array}$$

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

For any subset of agents $A \subset \mathcal{A}$, let $\mathcal{S}_A : \mathbb{R}^{NK} \to \mathbb{R}^{|A|K}$ be the selection operator $\mathcal{S}_A x = (x(n))_{n\in A}$. This linear operator

admits the matrix representation $\mathcal{S}_A x = (S_A \otimes I_K)x$ where the matrix S_A is a $|A| \times N$ selection matrix, *i.e.*, its elements are valued in $\{0,1\}$, it has one non zero element per row, and it has one non zero element at most per column. Finally, set $T = \sum_1^L |A_\ell|$ and define the linear operator

$$\begin{array}{rccc} M: & \mathbb{R}^{NK} & \longrightarrow & \mathbb{R}^{TK} \\ & x & \longmapsto & Mx = (\mathcal{S}_{A_\ell}(x))_{\ell=1}^L = (S \otimes I_K)x \end{array}$$

where

$$S = \begin{bmatrix} S_{A_1} \\ \vdots \\ S_{A_L} \end{bmatrix}$$

is a $T \times N$ matrix. Operator M will be identified from now on with the matrix $M = S \otimes I_K$.

With these definitions, we now consider the optimization problem

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

Let $\mathcal{G} = (\{1,\dots,L\}, \mathcal{E})$ be the non oriented graph with $\{1,\dots,L\}$ as the set of vertices and with the set of edges $\mathcal{E}$ defined as $\{\ell, m\} \in \mathcal{E}$ if $A_\ell \cap A_m \neq \emptyset$. Then, we made the following assumption. Let us remark that our proposed algorithms described later will be distributed at the subset level. A coordination will be needed within each subset A_ℓ, but the exchanges between the subsets are fully distributed.

Assumption 3 *The following facts hold true:*

i) $\bigcup_{\ell=1}^L A_\ell = \mathcal{A}$,

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

We obtain the following lemma.

Lemma 1 *Under Assumption 3, $x_\star$ is a minimizer of Problem* (1) *if and only if $(x_\star, \dots, x_\star)$ is a minimizer of Problem* (2).

Proof: The equivalence stated by this lemma will be established if we prove that $g(Mx) = \sum_{\ell=1}^L \imath_{C_{|A_\ell|}}((x(i))_{i\in A_\ell})$ is finite if and only if $x \in C_N$. Since $\mathcal{G}$ is connected, there exists $\ell_1 \neq 1$ such that $A_1 \cap A_{\ell_1} \neq \emptyset$. Therefore, $\imath_{C_{|A_1|}}((x(i))_{i\in A_1}) + \imath_{C_{|A_{\ell_1}|}}((x(i))_{i\in A_{\ell_1}}) = \imath_{C_{|A_1\cup A_{\ell_1}|}}((x(i))_{i\in A_1\cup A_{\ell_1}})$. Similarly, there exists $\ell_2 \notin \{1,\ell_1\}$ such that $(A_1 \cup A_{\ell_1}) \cap A_{\ell_2} \neq \emptyset$, therefore $\imath_{C_{|A_1|}}((x(i))_{i\in A_1}) + \imath_{C_{|A_{\ell_1}|}}((x(i))_{i\in A_{\ell_1}}) + \imath_{C_{|A_{\ell_2}|}}((x(i))_{i\in A_{\ell_2}}) = \imath_{C_{|A_1\cup A_{\ell_1}\cup A_{\ell_2}|}}((x(i))_{i\in A_1\cup A_{\ell_1}\cup A_{\ell_2}})$. Pursuing, we obtain that $g(Mx) = \imath_{C_{|\cup_\ell A_\ell|}}((x(i))_{i\in\cup_\ell A_\ell})$. By Assumption 3-*i)*, this is equal to $\imath_{C_N}(x)$. ∎

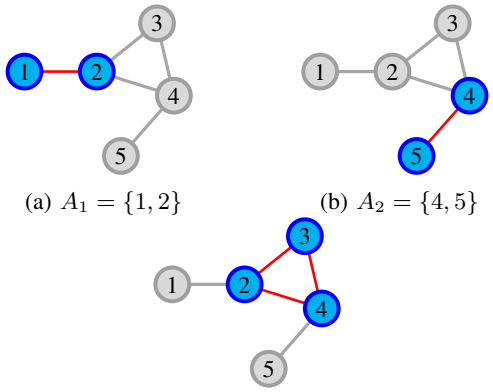
B. An illustration

In order to be less formal and to have some insights on our formulation, consider the example given in Figure 1. In that case, for any $x = (x(1),\dots,x(5))$, the vector Mx has 3 block-components respectively given by $(x(1),x(2))$, $(x(4),x(5))$ and $(x(2),x(3),x(4))$, that is:

$$Mx = (x(1),x(2)\,,\, x(4),x(5)\,,\, x(2),x(3),x(4)) \tag{3}$$

In this example, the function g is the indicator of the linear space composed of all vectors of the form

$$(u,u\,,\, v,v\,,\, w,w,w) \tag{4}$$



(a) $A_1 = \{1,2\}$ (b) $A_2 = \{4,5\}$

(c) $A_3 = \{2,3,4\}$

Fig. 1: An example of with $L = 3$ components.

for any u, v, w. This means that $g(z)$ is equal to zero whenever z has the form (4) and is equal to $+\infty$ otherwise. When $z = Mx$, we obtain that $g(Mx)$ is finite only if the vector (3) has the form (4). This holds if and only if $x(1) = x(2)$, $x(4) = x(5)$, $x(2) = x(3) = x(4)$. Equivalently, all components of x should be equal.

C. Instancianting ADMM

We now recall how ADMM can be used to solve Problem (2) in a distributed manner. ADMM is commonly described by reformulating Problem (2) into the constrained problem

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

and by introducing the so called augmented Lagrangian. This is the function $\mathcal{L}_\rho : \mathbb{R}^{NK} \times \mathbb{R}^{TK} \times \mathbb{R}^{TK} \to (-\infty, \infty]$ 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 constant. ADMM consists in the iterations:

$$x_{k+1} = \operatorname*{argmin}_{x\in\mathbb{R}^{NK}} \mathcal{L}_\rho(x, z_k; \lambda_k) \tag{5a}$$

$$z_{k+1} = \operatorname*{argmin}_{z\in\mathbb{R}^{TK}} \mathcal{L}_\rho(x_{k+1}, z; \lambda_k) \tag{5b}$$

$$\lambda_{k+1} = \lambda_k + \rho(Mx_{k+1} - z_{k+1}). \tag{5c}$$

A proof of the following result can be found in [15, Ch. 3.2 and Appendix A] combined with Lemma 1. Another proof using the so called Douglas Rachford splitting can be found in [25]:

Theorem 1 *Under Assumptions 1 to 3, the set of saddle points of the unaugmented Lagrangian $\mathcal{L}_0(x,z,\lambda)$ is nonempty, and any saddle point is of the form $(\mathbf{1}_N \times x_\star, \mathbf{1}_T \otimes x_\star, \lambda_\star)$ where $x_\star$ is the unique solution of Problem* (1). *Moreover, for any initial value (z_0, λ_0), the sequence of ADMM iterates (x_k, z_k, λ_k) converges to a saddle point.*

We now make the ADMM equations more explicit and show how they lead to a distributed implementation. The x and z – update equations above can be respectively rewritten as

$$x_{k+1} = \underset{x\in\mathbb{R}^{NK}}{\operatorname{argmin}}\, f(x) + \frac{\rho}{2}\|Mx - (z_k - \lambda_k/\rho)\|^2, \quad (6)$$

$$z_{k+1} = \underset{z\in\mathbb{R}^{TK}}{\operatorname{argmin}}\, g(z) + \frac{\rho}{2}\|z - (Mx_{k+1} + \lambda_k/\rho)\|^2. \quad (7)$$

Let us partition z_k as in the definition of the function g above and write $z_k = (z_k^{(1)}, \ldots, z_k^{(L)})$. Accordingly, let us write $\lambda_k = (\lambda_k^{(1)}, \ldots, \lambda_k^{(L)})$ where $\lambda_k^{(\ell)} \in \mathbb{R}^{|A_\ell|K}$ and furthermore, let us write $\lambda_k^{(\ell)} = (\lambda_k^{(\ell)}(n_1), \lambda_k^{(\ell)}(n_2), \ldots, \lambda_k^{(\ell)}(n_{|A_\ell|}))$ where $\lambda_k^{(\ell)}(n_i) \in \mathbb{R}^K$ and where n_i is the column index of the non zero element of the row i of S_{A_ℓ}. The indices n_i of the elements of $\lambda_k^{(\ell)}$ are therefore the indices of the agents belonging to the set A_ℓ. Using these notations, Equation (7) can be parallelized into L equations of the form

$$z_{k+1}^{(\ell)} = \underset{z\in\mathbb{R}^{|A_\ell|K}}{\operatorname{argmin}}\, \imath_{C_{|A_\ell|}}(z) + \frac{\rho}{2}\|z - (\mathcal{S}_{A_\ell}x_{k+1} + \lambda_k^{(\ell)}/\rho)\|^2,$$

whose solution is $z_{k+1}^{(\ell)} = \mathbf{1}_{|A_\ell|} \otimes \bar{z}_k^{(\ell)} \in C_{|A_\ell|}$ with

$$\bar{z}_{k+1}^{(\ell)} = \frac{1}{|A_\ell|}\sum_{n\in A_\ell}\Big(x_{k+1}(n) + \frac{\lambda_k^{(\ell)}(n)}{\rho}\Big).$$

Turning to the λ – update equation in the ADMM iterations and inspecting the structure of the matrix M, this equation can be decomposed into the equations

$$\lambda_{k+1}^{(\ell)}(n) = \lambda_k^{(\ell)}(n) + \rho(x_{k+1}(n) - \bar{z}_{k+1}^{(\ell)}) \quad (8)$$

for $\ell = 1, \ldots, L$ and $n = 1, \ldots, N$. Fixing ℓ and taking the sum of the $\lambda_{k+1}^{(\ell)}(n)$ with respect to n yields $\sum_{n\in A_\ell}\lambda_{k+1}^{(\ell)}(n) = 0$. Therefore, the $\bar{z}_k^{(\ell)}$ update equation can be written after the first iteration as

$$\bar{z}_{k+1}^{(\ell)} = \frac{1}{|A_\ell|}\sum_{n\in A_\ell} x_{k+1}(n). \quad (9)$$

Getting back to Equation (6), we now see that it can be parallelized into N equations of the form

$$x_{k+1}(n) = \underset{w\in\mathbb{R}^K}{\operatorname{argmin}}\, f_n(w) + \sum_{m\in\sigma(n)}\langle\lambda_k^{(m)}(n), w\rangle + \frac{\rho}{2}\|w - \bar{z}_k^{(m)}\|^2 \quad (10)$$

for $n = 1, \ldots, N$, where $\sigma(n) = \{m : n \in A_m\}$. Let us introduce the following aggregate quantities:

$$\Delta_k(n) = \frac{1}{\rho|\sigma(n)|}\sum_{m\in\sigma(n)}\lambda_k^{(m)}(n)$$

$$\chi_k(n) = \frac{1}{|\sigma(n)|}\sum_{m\in\sigma(n)}\bar{z}_k^{(m)}$$

After some algebra, the x-update in (10) simplifies to

$$x_{k+1}(n) = \operatorname{prox}_{\frac{f_n}{\rho|\sigma(n)|}}(\chi_k(n) - \Delta_k(n)). \quad (11)$$

By (8), we have the update equation

$$\Delta_{k+1}(n) = \Delta_k(n) + x_{k+1}(n) - \chi_{k+1}(n). \quad (12)$$

We are now in position to state the main algorithm.

D. Distributed ADMM (General case)

All agents within a subset A_ℓ are assumed to be connected together through a communication network. Recall that for a given n, $|\sigma(n)|$ is the number of clusters to which Agent n belongs.

Before entering the iteration $k+1$, Agent n holds in its memory the values $x_k(n)$, $\chi_k(n)$ and $\Delta_k(n)$.

Distributed-ADMM (General case)
At Iteration $k+1$,

1) The agents $1, \ldots, N$ compute their estimates $x_{k+1}(n)$
$$x_{k+1}(n) = \operatorname{prox}_{\frac{f_n}{\rho|\sigma(n)|}}(\chi_k(n) - \Delta_k(n)).$$
2) For all $\ell = 1, \ldots, L$, the agents belonging to a cluster A_ℓ send their estimates $x_{k+1}(n)$ to a "cluster head" who can be a preselected member of A_ℓ or an independent device. The cluster head computes
$$\bar{z}_{k+1}^{(\ell)} = \frac{1}{|A_\ell|}\sum_{n\in A_\ell} x_{k+1}(n).$$
and sends back this parameter to all the members of the cluster.
3) For $n = 1, \ldots, N$, Agent n computes
$$\chi_k(n) = \frac{1}{|\sigma(n)|}\sum_{m\in\sigma(n)}\bar{z}_k^{(m)}$$
$$\Delta_{k+1}(n) = \Delta_k(n) + x_{k+1}(n) - \chi_{k+1}(n).$$

Note that in the absence of a cluster head, one can think of a distributed computation of $\bar{z}_{k+1}^{(\ell)}$ within the cluster using *e.g.* a gossiping algorithm.

Note also that when $|A_\ell| = 2$ for all $\ell = 1, \ldots, L$, no cluster head nor a gossiping algorithm are needed for the execution of Step 2. Assuming that $A_\ell = \{m, n\}$, Agents m and n exchange the values of $x_{k+1}(m)$ and $x_{k+1}(n)$ then they both compute $\bar{z}_{k+1}^{(\ell)} = (x_{k+1}(m) + x_{k+1}(n))/2$. In this case, the algorithm is fully distributed at the agents level. This point is discussed in the next paragraph.

E. Distributed ADMM (Special cases)

We end this section by two important examples of possible choices of the subsets A_ℓ. We shall come back to these examples later.

Example 1: This is the centralized ADMM described in [15, Chap. 7]. Let $L = 1$ and $A_1 = \mathcal{A}$. Problem (2) becomes $\inf_{x\in\mathbb{R}^{NK}} f(x) + \imath_{C_N}(x)$. At Iteration $k+1$, a dedicated device simply computes $\bar{z}_{k+1} = N^{-1}\sum_{n=1}^N x_{k+1}(n)$ and broadcasts it to all the agents.

Example 2: Here we assume that all the agents belong to a communication network represented by a non oriented graph with no self loops $G = (\mathcal{A}, E)$ where $E = \{\{n_1, m_1\}, \{n_2, m_2\}, \ldots\}$ is the set of edges. Setting $L = |E|$, we consider that any pair of agents $\{n, m\}$ such that $\{n, m\} \in E$ is a set A_ℓ. If the graph G is connected, then Assumption 3 is easily seen to be verified, and Problems (1) and (2) are equivalent. In this situation, $|\sigma(n)|$ simply coincides with the *degree* d_n *of node* n in the graph *i.e.*,

the number of its neigbors. For every edge $A_\ell = \{n, m\}$, $\bar{z}_k^{(\ell)} = (x_k(n)+x_k(m))/2$ is simply the average of the nodes' estimates on that edge. As a consequence, it is straightforward to show that $\chi_k(n) = (x_k(n) + \bar{x}_k(n))/2$ where $\bar{x}_k(n) = \frac{1}{d_n}\sum_{m\in\mathcal{N}_n} x_k(m)$ and $\mathcal{N}_n$ is the neighborhood of node n.

This leads to the following algorithm. An Agent n keeps the variables $x_k(n), \bar{x}_k(n), \Delta_k(n)$ at each time k.

Distributed-ADMM (clusters are edges)
At Iteration $k+1$, Agent n

1) computes its estimate
$$x_{k+1}(n) = \operatorname{prox}_{\frac{f_n}{\rho d_n}}\left(\frac{x_k(n)+\bar{x}_k(n)}{2} - \Delta_k(n)\right),$$
2) receives the estimates $x_k(m)$ of other Agents $m \in \mathcal{N}_n$ in its neighborhood and computes
$$\bar{x}_k(n) = \frac{1}{d_n}\sum_{m\in\mathcal{N}_n} x_k(m),$$
3) updates $\Delta_{k+1}(n) = \Delta_k(n) + (x_{k+1}(n) - \bar{x}_{k+1}(n))/2$.

In this special case, the algorithm boils down to the algorithm of [11].

III. Main result

We now come to the main result of this paper. Define the $T \times T$ orthogonal projection matrix

$$\Pi = \begin{bmatrix} J_{|A_1|} & & \\ & \ddots & \\ & & J_{|A_L|} \end{bmatrix}$$

and consider the $TK \times TK$ orthogonal projection matrix

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

Define the $TK \times TK$ matrix

$$Q = \rho M \left(\begin{bmatrix} \nabla^2 f_1(x_\star) & & \\ & \ddots & \\ & & \nabla^2 f_N(x_\star) \end{bmatrix} + \rho M^* M \right)^{-1} M^*$$
$$= \rho M \left(\nabla^2 f(\mathbf{1}_N \otimes x_\star) + \rho M^* M\right)^{-1} M^* \quad (13)$$

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

Theorem 2 *Let Assumptions 1 to 3 hold true. Let* $\alpha = \boldsymbol{r}((\Pi_{\operatorname{span}(P+Q)} - (P+Q))(I-2P))$ *where* $\Pi_{\operatorname{span}(P+Q)}$ *is the orthogonal projection matrix on* $\operatorname{span}(P+Q)$*. Then the following facts hold true:*

i) $\alpha < 1$,
ii) *For any initial value* (z_0, λ_0) *of ADMM,*
$$\limsup_{k\to\infty} \frac{1}{k}\log\|x_k - \mathbf{1}_N \otimes x_\star\| \le \log\alpha,$$
iii) *The matrix* $R = (I-P-Q)(I-2P)$ *has eigenvalues with absolute value* α*. If* $\operatorname{span}((I-2P)M)$ *is not orthogonal to the invariant subspace of* R *associated with these eigenvalues, then*
$$\limsup_{k\to\infty} \frac{1}{k}\log\|x_k - \mathbf{1}_N \otimes x_\star\| = \log\alpha$$
for any initial value (z_0, λ_0) *outside a set of Lebesgue measure equal to zero.*

This theorem says that $\|x_k - \mathbf{1}_N \otimes x_\star\| < (\alpha + o(1))^k$, and provides a condition under which this rate is tight. It means that, for (z_0, λ_0) outside a set with zero Lebesgue measure, $\|x_k - \mathbf{1}_N \otimes x_\star\| = (\alpha + o(1))^k$. Moreover, Item iii) states that this rate is *tight* as soon as matrices P and M satisfies a technical condition. Although providing deeper insights on this condition is a difficult task, we claim that the latter condition is mild. It is for instance trivially satisfied in the case of a centralized network (see Example 1 in Section II-E).

IV. Special cases

The aim of this section is to provide some examples of the rate $\alpha = \boldsymbol{r}((\Pi_{\operatorname{span}(P+Q)} - (P+Q))(I-2P))$ in simple scenarios. We assume for simplicity that the dimension K of the parameter space is equal to one, in which case we have $M = S$ and $P = \Pi$. Moreover, we investigate the case where $f_n''(x_\star) = \sigma_\star^2 > 0$ is a constant which does not depend on n. Remark that this assumption is not mandatory, but has the benefit of yielding simple and insightful expressions.

A. The centralized network

We consider here the simple configuration of Example 1 of Section II-C, where $L = 1$ and $A_1 = \mathcal{A}$. This case amounts to assuming that $M = I_N$, $P = N^{-1}\mathbf{1}_N\mathbf{1}_N^*$ and $Q = \frac{\rho}{\sigma_\star^2+\rho} I_N$. The projector $\Pi_{\operatorname{span}(P+Q)}$ is the identity and the rate of convergence α of ADMM coincides with the spectral radius of

$$R = \Big(I_N - P - Q\Big)\Big(I_N - 2P\Big)$$
$$= \frac{\sigma_\star^2}{\sigma_\star^2+\rho}(I_N - P) + \frac{\rho}{\sigma_\star^2+\rho}P.$$

Matrix R has two possibly distinct eigenvalues $\frac{\sigma_\star^2}{\sigma_\star^2+\rho}$ and $\frac{\rho}{\sigma_\star^2+\rho}$ both of them less than one. We have the following corollary.

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

$$\alpha = \frac{\max(\rho, \sigma_\star^2)}{\rho + \sigma_\star^2}. \quad (14)$$

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

Corollary 1 states that the optimal convergence rate is $1/2$ and that this rate is attained when the step-size of the algorithm is equal to $\sigma_\star^2$. In general, $\sigma_\star^2$ is unknown, but the result nevertheless provides useful guidelines on the way to select parameter ρ, and potentially allows to design adaptive step-size selection strategies.

It is worth noting that a closed-form expression of the rate α can as well be obtained in the case where the second order

derivatives $f''(x_\star)$ are distinct. The analysis is however somewhat tedious, as the latter closed-form expression depends on the location of ρ on the real axis. We shall discuss this case below in the numerical section.

B. The ring network

We now assume that $N \geq 3$ and consider the framework of Example 2 of Section II-C. In that framework, the graph $G = (\mathcal{A}, E)$ that we study here is the ring network with $E = \{\{1,2\}, \{2,3\}, \ldots, \{N-1,N\}, \{N,1\}\}$ as the set of edges. We therefore have $L = N$ and

$$A_\ell = \begin{cases} \{\ell, \ell+1\} & \text{if } \ell < N \\ \{1, N\} & \text{otherwise} \end{cases}$$

as shown in Figure 2.

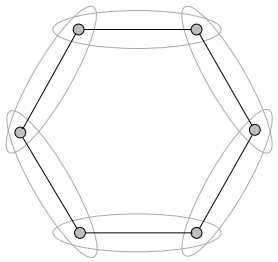


Fig. 2: Example of a ring network with $N = 6$. Sets A_ℓ are represented by the ellipses.

We define for simplicity $s_N = \sin(2\pi/N)$, $c_N = \cos(2\pi/N)$ and $t_N = \tan(2\pi/N)$.

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,

$$\alpha \geq \alpha_{opt} := \frac{1}{\sqrt{2}}\sqrt{\frac{1+c_N}{1+s_N}}$$

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

Proof: See Appendix. ∎

The optimal step-size $\rho_{\text{opt}} = \frac{\sigma_\star^2}{2s_N}$ is equal to $\frac{\sigma_\star^2 N}{4\pi} + o(N)$ which suggests that the step-size should increase at the rate N. In that case, we get

$$\alpha_{\text{opt}} = 1 - \frac{\pi}{N} + o(\frac{1}{N}). \tag{15}$$

Before closing this section, it is interesting to compare these results with the speeds of some well-known algorithms used for solving the so-called *average consensus* problem (see the seminal work of [26] for more details). Given a network of N agents each holding a measurement θ_n, the purpose of these agents is to reach a consensus over the average $N^{-1}\sum_{n=1}^N \theta_n$. To attain this consensus, one idea is to solve the problem $\min_x \sum_{n=1}^N (x-\theta_n)^2$ by means of ADMM. This approach has been undertaken in [27], where the authors also obtained Equation (15) in the large N asymptotic regime of the ring network. When synchronous *gossip* algorithms for average consensus is considered, it is proven in [28] that $\alpha \simeq 1 - 2\pi^2/N^2$ for large N ring graphs.

V. Proof of Theorem 2

The proof is centered around the vector sequence $\zeta_k = \lambda_k + \rho z_k$ who can be shown to govern the evolution of ADMM. Starting with the case where all the functions f_n are quadratic, we show that an ADMM iteration boils down to the affine transformation $\zeta_{k+1} = R\zeta_k + d$ where R is a $TK \times TK$ matrix with a certain structure. A spectral analysis of R shows then that the rate of convergence of ADMM is provided by the largest modulus of the eigenvalues of R different from one (this number is smaller that one). Part of the proof consists in characterizing the eigenspace associated with these eigenvalues. We then generalize our results to the case where the functions f_n are not necessarily quadratic, but remain twice differentiable in the vicinity of the minimum. In this case, we obtain a perturbed version of an affine transformation similar to the quadratic case. The perturbation terms will be shown to lie in such a subspace of R that the analysis of the quadratic case remains essentially effective.

We start our proof by providing preliminary results that describe some simple algebraic properties of the matrices S and Π who will help us study the eigenstructure of R. We then recall some well known properties of the *proximity operator* which is known to be tightly related with ADMM. We then establish Theorem 2 in the quadratic case, ending our proof with the general case.

A. Preliminary results

We shall need to reformulate Assumption 3 in a form that will be more conveniently used in the proof:

Lemma 2 *Assumption 3 is equivalent to:*

i) $\operatorname{rank}(S) = N$, *and*
ii) $\operatorname{span}(S) \cap \operatorname{span}(\Pi) = \operatorname{span}(\mathbf{1}_T)$.

Proof: Since the matrix S has one non zero element per row, its non zero columns are linearly independent. Assumption 3-*i)* is equivalent to the fact that no column of S is zero. Therefore, Assumptions 3-*i)* and Item $i)$ in the statement of the lemma are equivalent.

Denoting any vector $v \in \mathbb{R}^N$ as $v = (v_1, \dots, v_N)$, we have

$$y \in \operatorname{span}(S) \cap \operatorname{span}(\Pi) \Leftrightarrow \exists v \in \mathbb{R}^N : y = ((v_n)_{n\in A_1}, \dots, (v_n)_{n\in A_L}) \text{ and } \sum_{\ell=1}^{L} \imath_{\operatorname{span}(\mathbf{1}_{|A_\ell|})}((v_n)_{n\in A_\ell}) = 0.$$

If Assumption 3-*ii)* is satisfied, then the proof of Lemma 1 shows that $\sum_\ell \imath_{\operatorname{span}(\mathbf{1}_{|A_\ell|})}((v_n)_{n\in A_\ell}) = \imath_{\operatorname{span}(\mathbf{1}_{|\cup_\ell A_\ell|})}((v_n)_{n\in \cup_\ell A_\ell})$, which shows that $y \in \operatorname{span}(\mathbf{1}_T)$. Conversely, suppose that Assumption 3-*ii)* is not satisfied. Then there exists a non empty set $\mathcal{C}$ strictly included in $\{1, \dots, L\}$ such that $\cup_{\ell\in\mathcal{C}} A_\ell$ and $\cup_{\ell\in\{1,\dots,L\}-\mathcal{C}} A_\ell$ are disjoint. Let $v \in \mathbb{R}^N$ be defined as $(v_n)_{n\in\cup_{\ell\in\mathcal{C}} A_\ell} = \alpha \mathbf{1}_{|\cup_{\ell\in\mathcal{C}} A_\ell|}$ and $(v_n)_{n\in\cup_{\ell\in\{1,\dots,L\}-\mathcal{C}} A_\ell} = \beta \mathbf{1}_{|\cup_{\ell\in\{1,\dots,L\}-\mathcal{C}} A_\ell|}$ with $\alpha \neq \beta$. Then it is easy to see that $y = ((v_n)_{n\in A_1}, \dots, (v_n)_{n\in A_L}) \in \operatorname{span}(S) \cap \operatorname{span}(\Pi)$ but $y \notin \operatorname{span}(\mathbf{1}_T)$. In conclusion, Assumption 3-*ii)* and Item *ii)* are equivalent. ∎

Other properties of S will be needed:

Lemma 3 *The matrix S satisfies the following properties:* (i) $S\mathbf{1}_N = \mathbf{1}_T$, (ii) $(I_T - \Pi)S\mathbf{1}_N = 0$.

Proof: The property *(i)* is due to the fact that S contains one non zero element per row, and this element is equal to 1. Since $\Pi\mathbf{1}_T = \mathbf{1}_T$, we obtain *(ii)*. ∎

B. The proximity operator

Let $h \in \Gamma_0(\mathbb{R}^n)$. The following two lemmas are well known, see *e.g.* [29]:

Lemma 4 *The operator* prox_h *is non expansive,* i.e.*,* $\|\operatorname{prox}_h(x) - \operatorname{prox}_h(x')\| \leq \|x - x'\|$.

Lemma 5 *Let $\mathcal{C} \subset \mathbb{R}^n$ be a closed convex set. Then* $\operatorname{prox}_{\imath_\mathcal{C}}(x)$ *coincides with the orthogonal projection of x on $\mathcal{C}$.*

C. The quadratic case

We consider herein the case where the functions f_n are quadratic: $f_n(x(n)) = 0.5\, x(n)^* \Phi_n x(n) + c_n^* x(n) + d_n$ where the $K \times K$ matrices Φ_n are symmetric and nonnegative, the c_n are some given $K \times 1$ vectors, and the d_n are some given scalars. Assumption 2 reads in this case $\sum_1^N \Phi_n > 0$. We immediately observe that the solution of Problem (1) is attained at a unique point $x_\star$. Observe also that the function $f(x)$ is quadratic with the gradient $\nabla f(x) = \Phi x + c$ where

$$\Phi = \begin{bmatrix} \Phi_1 & & \\ & \ddots & \\ & & \Phi_N \end{bmatrix}, \quad \text{and} \quad c = \begin{bmatrix} c_1 \\ \vdots \\ c_N \end{bmatrix}. \tag{16}$$

Our first task is formulate an ADMM iteration as a single line affine transformation, as alluded to in the introduction of this paper. We start with the z – update equation (7). Writing $\zeta_{k+1} = \rho M x_{k+1} + \lambda_k$, Equation (7) can be rewritten $z_{k+1} = \operatorname{prox}_{\rho^{-1}g}(\zeta_{k+1}/\rho)$. Recall now that the matrices $P_{|A_\ell|}$ introduced in Section II-A are the orthogonal projection matrices on the subspaces $C_{|A_\ell|} \subset \mathbb{R}^{|A_\ell|K}$. Therefore, the matrix P is the orthogonal projection matrix on the subspace of $\mathbb{R}^{TK}$ that coincides with the domain of g. By Lemma 5, we get that $z_{k+1} = \rho^{-1} P \zeta_{k+1}$.

Letting $P_\perp = I_{TK} - P$ be the orthogonal projection matrix on the orthogonal complement of $\operatorname{span}(P)$, we get that $P_\perp \zeta_{k+1} = \zeta_{k+1} - P\zeta_{k+1} = \zeta_{k+1} - \rho z_{k+1} = \lambda_k + \rho(M x_{k+1} - z_{k+1}) = \lambda_{k+1}$. Summarizing, we have $\zeta_k = \lambda_k + \rho z_k$, $\rho z_k = P\zeta_k$ and $\lambda_k = P_\perp \zeta_k$ for any $k \in \mathbb{N}$.

We now turn to the x – update equation. Since f is differentiable, Equation (6) can be rewritten as

$$\nabla f(x_{k+1}) + \rho M^*(M x_{k+1} + \lambda_k/\rho - z_k) = 0,$$

or equivalently,

$$\nabla f(x_{k+1}) + \rho M^* M x_{k+1} + M^*(I - 2P)\zeta_k = 0.$$

As $\nabla f(x) = \Phi x + c$, we get $(\Phi + \rho M^* M) x_{k+1} = -M^*(I - 2P)\zeta_k - c$. Since $M^*M = (S^*S) \otimes I_K > 0$ by Lemma 2, the matrix $H = \Phi + \rho M^* M$ is invertible, and we end up with

$$x_{k+1} = -H^{-1} M^* (I - 2P)\zeta_k - H^{-1} c. \tag{17}$$

Recalling that $\zeta_{k+1} = \rho M x_{k+1} + \lambda_k$ and observing that the matrix Q defined in (13) coincides with $\rho M H^{-1} M^*$ in the quadratic case, we finally obtain

$$\begin{aligned} \zeta_{k+1} &= -Q(I - 2P)\zeta_k + \lambda_k - \rho M H^{-1} c \\ &= (P_\perp - Q)(I - 2P)\zeta_k - \rho M H^{-1} c \\ &\stackrel{\text{def}}{=} R\zeta_k + d \end{aligned} \tag{18}$$

where $R = (P_\perp - Q)(I - 2P)$ and $d = -\rho M H^{-1} c$.

Remark 1 *These derivations show that the sequence ζ_k is autonomous and completely characterizes ADMM. This phenomenon is in fact general and shows up naturally when ADMM is interpreted as a particular case of the Douglas-Rachford splitting algorithm [25], [30].*

The following lemma provides some important spectral properties of R. In fact, part of the results shown in its statement can be deduced from Theorem 1. Indeed, a consequence of this theorem is that the iterations $\zeta_{k+1} = R\zeta_k + d$ converge for any initial value ζ_0. Yet, the direct proof provided below provides a finer understanding of the spectral properties of R:

Lemma 6 *$\boldsymbol{r}(R) \leq 1$. Moreover, for any $\theta \in (0, 2\pi)$, $\exp(\imath\theta)$ is not an eigenvalue of R. Finally, the algebraic and geometric multiplicities of any eigenvalue of R coincide.*

Proof: We have $\boldsymbol{r}(R) \leq \|R\| = \|P_\perp - Q\|$. Upon noting that $|x^* P_\perp x - x^* Q x| \leq \|x\|^2$ for any $x \in \mathbb{R}^{TK}$, all eigenvalues of the real symmetric matrix $P_\perp - Q$ have their absolute value no larger than one. Thus, the same holds for $(P_\perp - Q)^2$ from which we obtain $\|P_\perp - Q\| \leq 1$. This proves the first point of the Lemma.

Assume that $\exp(\imath\theta)$ is an eigenvalue of R for some $\theta \in [0, 2\pi)$, and let w be an associated eigenvector with $\|w\| = 1$. Since $Rw = \exp(\imath\theta) w$, we have $\|Rw\| = 1$, which implies that $w_{\mathrm{R}}^* (P_\perp - Q)^2 w_{\mathrm{R}} = 1 = \|w_{\mathrm{R}}\|^2$ where $w_{\mathrm{R}} = (I - 2P)w$. This implies that w_{R} lies in the eigenspace of $(P_\perp - Q)^2$ corresponding to the unit eigenvalue. Any such vector can be written as the sum $w_{\mathrm{R}} = u + v$ of two orthogonal vectors

satisfying $(P_\perp - Q)u = u$ and $(P_\perp - Q)v = -v$. As $u^*(P_\perp - Q)u = \|u\|^2$, one has $\|u\|^2 \geq u^* P_\perp u = \|u\|^2 + u^* Q u$. Since Q is non negative, we obtain $Qu = 0$ and, consequently, $Pu = 0$. Using similar arguments, $P_\perp v = 0$. Now, equation $Rw = \exp(\imath\theta)w$ reads $(P_\perp - Q)(u+v) = \exp(\imath\theta)(I-2P)(u+v)$. In the light of the above properties of u and v, this is equivalent to $u - v = \exp(\imath\theta)(u - v)$. Thus $\theta = 0$ and the second point is proved.

The eigenvalue λ has the same algebraic and geometric multiplicities if and only if $\mathrm{rank}((R-\lambda I)^2) = \mathrm{rank}(R-\lambda I)$ (indeed, any Jordan block $\mathcal{J}_\lambda$ with size > 1 associated with λ would satisfy $\mathrm{rank}((\mathcal{J}_\lambda - \lambda I)^2) = \mathrm{rank}(\mathcal{J}_\lambda - \lambda I) - 1$). Using the identities $\mathrm{rank}(AB) = \mathrm{rank}(BA)$ and $\mathrm{rank}(AB) = \mathrm{rank}(B)$ if A is invertible, we obtain

$$\mathrm{rank}(R - \lambda I) = \mathrm{rank}(P_\perp - Q - \lambda(P_\perp - P))$$

since $(P_\perp - P)^2 = I$, and

$$\begin{aligned}\mathrm{rank}((R-\lambda I)^2) &= \mathrm{rank}([(P_\perp - Q - \lambda(P_\perp - P))(P_\perp - P)]^2) \\ &= \mathrm{rank}([P_\perp - Q - \lambda(P_\perp - P)]^2).\end{aligned}$$

Since $P_\perp - Q - \lambda(P_\perp - P)$ is symmetric, these two ranks coincide. ∎

When it exists, the eigenspace of R associated with the eigenvalue 1 plays an important role. In order to characterize this subspace, we start with a preliminary result:

Lemma 7 *Let Assumptions 2 and 3 hold true, and let $\mathcal{N} = \ker(Q - P)$ be the null space of $Q - P$. Then*

$$\mathcal{N} = \ker(Q) \cap \ker(P) = \ker(Q + P).$$

Proof: We provide the proof of the first equality $\mathcal{N} = \ker(Q) \cap \ker(P)$ (the second equality follows from the non negativity of Q and P). Note that we only need to prove that $\mathcal{N} \subset \ker(Q) \cap \ker(P)$, the other inclusion being trivial. For any $\zeta \in \mathcal{N}$, we set $\zeta = \lambda + \rho z$ where $\lambda = P_\perp \zeta$ and $\rho z = P\zeta$. Vector ζ satisfies $Q\zeta = P\zeta = \rho z$. Observe that $\mathrm{span}(Q) = \mathrm{span}(M) = \mathrm{span}(S \otimes I_K) = \mathrm{span}(S) \otimes \mathbb{R}^K$ where the second $\otimes$ denotes the tensor product. And since $\mathrm{span}(P) = \mathrm{span}(\Pi \otimes I_K) = \mathrm{span}(\Pi) \otimes \mathbb{R}^K$, we obtain that $z \in \mathrm{span}(M) \cap \mathrm{span}(P) = (\mathrm{span}(S) \cap \mathrm{span}(\Pi)) \otimes \mathbb{R}^K = \mathrm{span}(\mathbf{1}_T) \otimes \mathbb{R}^K$ by Lemma 2. Hence $z = \mathbf{1}_T \otimes q$ where $q \in \mathbb{R}^K$ needs to be determined. Replacing in the equation $Q\zeta = \rho z$, we obtain $MH^{-1}\left(M^*\lambda + \rho M^*(\mathbf{1}_T \otimes q)\right) = \mathbf{1}_T \otimes q$. But $M(\mathbf{1}_N \otimes q) = (S \otimes I_K)(\mathbf{1}_N \otimes q) = S\mathbf{1}_N \otimes q = \mathbf{1}_T \otimes q$ by Lemma 3. Since M is full column rank by Lemma 2, we obtain $M^*\lambda + \rho M^*(\mathbf{1}_T \otimes q) = H(\mathbf{1}_N \otimes q)$, or

$$\begin{aligned}M^*\lambda &= \Phi(\mathbf{1}_N \otimes q) + \rho M^*(S \otimes I_K)(\mathbf{1}_N \otimes q) - \rho M^*(\mathbf{1}_T \otimes q) \\ &= \Phi(\mathbf{1}_N \otimes q)\end{aligned}$$

by Lemma 3. This lemma also shows that $(\mathbf{1}_N^* \otimes I_K)M^*\lambda = (\mathbf{1}_N^* S^* \otimes I_K)((I_T - \Pi) \otimes I_K)\lambda = 0$. Hence $(\mathbf{1}_N^* \otimes I_K)\Phi(\mathbf{1}_N \otimes q) = \sum_1^N \Phi_n q = 0$, which implies $q = 0$ by Assumption 2. This shows that $Q\zeta = P\zeta = 0$, in other words $\zeta \in \ker(Q) \cap \ker(P)$. ∎

Recall from Lemma 6 that all the Jordan blocks of R are trivial. Denoting as $\dim(\cdot)$ the dimension of a vector space, we have:

Lemma 8 *The matrix R has an eigenvalue equal to 1 if and only if $\dim(\mathcal{N}) > 0$. In that case, let*

$$R = W\Lambda W^{-1} = \begin{bmatrix} W_1 & W_2 \end{bmatrix} \begin{bmatrix} I & \\ & \tilde{\Lambda} \end{bmatrix} \begin{bmatrix} \underline{W}_1^* \\ \underline{W}_2^* \end{bmatrix}$$

be a spectral factorization of R. Then $W_1\underline{W}_1^* = \Pi_{\mathcal{N}}$, the orthogonal projection matrix onto $\mathcal{N}$. Whether R has or has not an eigenvalue equal to 1 (in which case we set $\Pi_{\mathcal{N}} = 0$),

$$R - \Pi_{\mathcal{N}} = (\Pi_{\mathrm{span}(P+Q)} - (P+Q))(I - 2P).$$

Proof: When R has an eigenvalue equal to 1, any vector w of the associated right eigenspace $\mathcal{E}_{\text{right}}(1)$ satisfies $(P_\perp - Q)(I-2P)w = w$. Writing $w_{\mathrm{R}} = (I-2P)w$ and recalling that $(I-2P)^2 = I$, we obtain $w \in \mathcal{E}_{\text{right}}(1) \Leftrightarrow (Q-P)w_{\mathrm{R}} = 0$, in other words $w_{\mathrm{R}} \in \mathcal{N}$. But since $\mathcal{N} \subset \ker(P)$ by Lemma 7, we obtain that $\mathcal{E}_{\text{right}}(1) = \mathcal{N}$. We show similarly that the left eigenspace associated with the eigenvalue 1 is $\mathcal{N}$.

Turning to the spectral factorization, since $\mathrm{span}(W_1) = \mathrm{span}(\underline{W}_1) = \mathcal{N}$, we can write $\underline{W}_1 = W_1 U$ where U is an invertible matrix. Since $\underline{W}_1^* W_1 = I_{\dim(\mathcal{N})}$, we have $U = (W_1^* W_1)^{-1}$ which shows that $W_1\underline{W}_1^* = \Pi_{\mathcal{N}}$.

Finally, since $(I-2P)^2 = I$ and $\mathcal{N} \subset \ker(P)$,

$$\begin{aligned}R - \Pi_{\mathcal{N}} &= (I - (P+Q) - \Pi_{\mathcal{N}}(I-2P))(I-2P) \\ &= (I - (P + Q + \Pi_{\mathcal{N}}))(I-2P) \\ &= (\Pi_{\mathrm{span}(P+Q)} - (P+Q))(I-2P).\end{aligned}$$

∎

The spectral properties of R that we just established lead us to the following lemma. We denote by $A^\sharp$ the Moore-Penrose pseudo-inverse of the matrix A.

Lemma 9 *Define $\overline{\boldsymbol{\zeta}}_\star = \rho(I-2P)(P-Q)^\sharp MH^{-1}c$. The set of fixed points of the transformation $\zeta \mapsto R\zeta + d$ coincides with $\{\overline{\boldsymbol{\zeta}}_\star\} + \mathcal{N}$.*

Proof: Let $\bar{\zeta}$ be a fixed point of the transformation $\zeta' = R\zeta + d$. Using the identity $(I-2P)^2 = I$, the equation $(I-R)\bar{\zeta} = d$ reads $(Q-P)\bar{\zeta}_{\mathrm{R}} = -\rho MH^{-1}c$ where $\bar{\zeta}_{\mathrm{R}} = (I-2P)\bar{\zeta}$. Note that $MH^{-1}c \in \mathrm{span}\, Q$. By Lemma 7, $\mathcal{N} = \ker(Q-P) \subset \ker Q$, thus $u^T MH^{-1}c = 0$ for any $u \in \mathcal{N}$. This means that $MH^{-1}c \in \mathrm{span}\,(Q-P)$. Consequently, the set of solutions to $(Q-P)\bar{\zeta}_{\mathrm{R}} = -\rho MH^{-1}c$ is nonempty and reads $\bar{\zeta}_{\mathrm{R}} \in \rho(P-Q)^{\#} MH^{-1}c + \mathcal{N}$. Note that multiplication by $(I-2P)$ leaves the space $\mathcal{N}$ invariant by Lemma 7. By multiplying both sides of the above equality by $(I-2P)$, we obtain $\bar{\zeta} \in \rho(I-2P)(P-Q)^{\#} MH^{-1}c + \mathcal{N}$ which proves Lemma 9. ∎

We are now in position to prove Theorem 2 in the quadratic case. Item *i)* in the statement of this theorem was shown by Lemma 8.

Given any fixed point $\zeta_\star$ of the transformation $\zeta' = R\zeta + d$, we have

$$\zeta_k - \zeta_\star = R(\zeta_{k-1} - \zeta_\star) = \cdots = R^k(\zeta_0 - \zeta_\star). \qquad (19)$$

Defining $\boldsymbol{x}_\star = -H^{-1}M^*(I-2P)\zeta_\star - H^{-1}c$ and recalling Eq. (17), we have

$$\begin{aligned} x_{k+1} - \boldsymbol{x}_\star &= -H^{-1}M^*(I-2P)(\zeta_k - \zeta_\star) \\ &= -H^{-1}M^*(I-2P)R^k(\zeta_0 - \zeta_\star) \\ &= -H^{-1}M^*(I-2P)\left(\Pi_{\mathcal{N}} + W_2\tilde{\Lambda}^k\underline{W}_2^*\right)(\zeta_0 - \zeta_\star)\,. \end{aligned}$$

By Lemma 7, note that $P\Pi_{\mathcal{N}} = Q\Pi_{\mathcal{N}} = 0$. Since $\mathrm{span}(Q) = \mathrm{span}(M)$, we have $M^*(I-2P)\Pi_{\mathcal{N}} = 0$. Consequently,

$$x_{k+1} - \boldsymbol{x}_\star = -H^{-1}M^*(I-2P)W_2\tilde{\Lambda}^k\underline{W}_2^*(\zeta_0 - \zeta_\star)\,. \quad (20)$$

Therefore, (x_k) converges to $\boldsymbol{x}_\star$ as $k \to \infty$. Since we know already by Theorem 1 that (x_k) converges to $\mathbf{1}_N \otimes x_\star$, this implies that $\boldsymbol{x}_\star = \mathbf{1}_N \otimes x_\star$. It is worth noting that this identity could have been be derived directly from the mere definition of $\boldsymbol{x}_\star$ with no need to use Theorem 1. As a sanity check, the reader may indeed verify that $\boldsymbol{x}_\star = \mathbf{1}_N \otimes x_\star$ using direct algebra. We skip this verification here as it is not mandatory for the proof.

Equality (20) yields Theorem 2-*ii)* in the quadratic case.

To show Theorem 2-*iii)*, write

$$W_2\tilde{\Lambda}^k\underline{W}_2^* = \begin{bmatrix} W_{2,1} & W_{2,2} \end{bmatrix} \begin{bmatrix} \tilde{\Lambda}_1^k & \\ & \tilde{\Lambda}_2^k \end{bmatrix} \begin{bmatrix} \underline{W}_{2,1}^* \\ \underline{W}_{2,2}^* \end{bmatrix}$$

where $\tilde{\Lambda}_1$ collects on its diagonal the eigenvalues of $W_2\tilde{\Lambda}\underline{W}_2^*$ with the absolute value α, and write

$$\begin{aligned} x_{k+1} - \boldsymbol{x}_\star &= -H^{-1}M^*(I-2P)W_{2,1}\tilde{\Lambda}_1^k\underline{W}_{2,1}^*(\zeta_0 - \zeta_\star) + \xi_k \\ &= G\tilde{\Lambda}_1^k v + \xi_k \end{aligned}$$

where $G = -H^{-1}M^*(I-2P)W_{2,1}$, $v = \underline{W}_{2,1}^*(\zeta_0 - \zeta_\star)$ and $\xi_k = -H^{-1}M^*(I-2P)W_{2,2}\Lambda_2^k\underline{W}_{2,2}^*(\zeta_0 - \zeta_\star)$. The condition on $\mathrm{span}((I-2P)M)$ in the statement of Theorem 2 asserts that $G \neq 0$. On the other hand, $v = \underline{W}_{2,1}^*(\zeta_0 - \overline{\zeta}_\star)$ by Lemmas 8 and 9. Hence, when ζ_0 lies outside a set with zero Lebesgue measure as we shall assume, $Gv \neq 0$. Denote by $\alpha e^{\imath\theta_1}, \ldots, \alpha e^{\imath\theta_L}$ the distinct elements on the diagonal of $\tilde{\Lambda}_1$. For any $k \geq 0$, the vector $G\tilde{\Lambda}_1^k v$ coincides with $\alpha^k f(k)$ where $f(k) = \sum_{\ell=1}^L a_\ell e^{\imath\theta_\ell k}$ for some coefficients $a_1, \ldots, a_L$. Since $Gv \neq 0$, at least one of these coefficients is non zero. Hence, $\limsup_k |f(k)| > 0$. Indeed, one can easily show that $n^{-1}\sum_{k=0}^{n-1}|f(k)|^2 \xrightarrow[n\to\infty]{} \sum |a_\ell|^2 > 0$. This would not be possible if $\limsup_k |f(k)| = 0$.

By construction, $\boldsymbol{r}(\tilde{\Lambda}_2) < \alpha$, hence $\|\xi_k\| \leq C\beta^k$ where C is a constant and where $0 \leq \beta < \alpha$. We therefore have

$$\|x_{k+1} - \boldsymbol{x}_\star\| \geq \|G\tilde{\Lambda}_1^k v\| - \|\xi_k\| \geq \alpha^k(|f(k)| - C(\beta/\alpha)^k)$$

and we obtain that $\|x_{k+1} - \boldsymbol{x}_\star\| \geq \alpha^k g(k)$ where $g(k) = \max(|f(k)| - C(\beta/\alpha)^k, 0)$. Observing that $\limsup_k g(k) > 0$ and using the convention $\log 0 = -\infty$, we obtain

$$\begin{aligned} \limsup_k \frac{1}{k}\log\|x_{k+1} - \boldsymbol{x}_\star\| &\geq \log\alpha + \limsup_k \frac{1}{k}\log g(k) \\ &= \log\alpha. \end{aligned}$$

Combining this lower bound with the already established upper bound, we obtain the result.

D. The general case

We now assume that the functions f_n satisfy Assumptions 1 and 2 in full generality. Theorem 1 shows that the iterates x_k converge towards the unique minimizer $\boldsymbol{x}_\star = \mathbf{1}_N \otimes x_\star$ of the problem. For k large enough, the iterates x_k are in a neighborhood of $\boldsymbol{x}_\star$ where the functions f_n are differentiable, and the x – update equation (Eq. (6)) boils down to the equation $\nabla f(x_{k+1}) + \rho M^* M x_{k+1} = -M^*(\lambda_k - \rho z_k) = -M^*(I-2P)\zeta_k$. This equation can be rewritten in two different manners. On the one hand, we have

$$x_{k+1} = \mathrm{prox}_h(-(\rho M^*M)^{-1}M^*(I-2P)\zeta_k)$$

where $h(x) = \sum_{n=1}^N [(\rho M^*M)^{-1}]_{nn} f_n(x(n))$, and on the other hand, we have for any x close enough to $\boldsymbol{x}_\star$

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

where $\|E(x)\|/\|x\| \to 0$ as $x \to 0$. With this relation, the update equation for x becomes

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

where $c = \nabla f(\boldsymbol{x}_\star) - \nabla^2 f(\boldsymbol{x}_\star)\boldsymbol{x}_\star$, or equivalently

$$\begin{aligned} x_{k+1} = &-H^{-1}M^*(I-2P)\zeta_k - H^{-1}c \\ &- H^{-1}E(x_{k+1} - \boldsymbol{x}_\star) \end{aligned} \quad (21)$$

where $H = \nabla^2 f(\boldsymbol{x}_\star) + \rho M^*M$. Mimicking the derivation made before Remark 1, this equation leads to

$$\zeta_{k+1} = R\zeta_k + d - \rho M H^{-1}E(x_{k+1} - \boldsymbol{x}_\star) \quad (22)$$

where $R = (P_\perp - Q)(I-2P)$ with $Q = \rho M H^{-1} M^*$ as in Equation (13), and where $d = -\rho M H^{-1} c$.

By replacing the matrix Φ defined in (16) with $\nabla^2 f(\boldsymbol{x}_\star)$, we notice that the lemmas 6–9 remain true for the matrices R and Q just introduced. Moreover, Equation (22) shows that the sequence ζ_k converges to a fixed point of the transformation $\zeta' = R\zeta + d$. Making $k \to \infty$ in (21), we also notice that

$$\begin{aligned} \boldsymbol{x}_\star &= -H^{-1}M^*(I-2P)\zeta_\star - H^{-1}c \\ &= \mathrm{prox}_h(-(\rho M^*M)^{-1}M^*(I-2P)\zeta_\star) \end{aligned} \quad (23)$$

where $\zeta_\star$ is any fixed point of the transformation $\zeta' = R\zeta + d$. We now have the elements to establish the Theorem 2-*ii)*. Given one fixed point $\zeta_\star$, the analogue of Eq. (19) is

$$\zeta_k - \zeta_\star = R^k(\zeta_0 - \zeta_\star) - \rho\sum_{\ell=1}^k R^{k-\ell}MH^{-1}E(x_\ell - \boldsymbol{x}_\star). \quad (24)$$

Lemma 4 shows now that

$$\begin{aligned} &\|x_{k+1} - \boldsymbol{x}_\star\| \\ =&\|\mathrm{prox}_h(-(\rho M^*M)^{-1}M^*(I-2P)\zeta_k) \\ &\quad - \mathrm{prox}_h(-(\rho M^*M)^{-1}M^*(I-2P)\zeta_\star)\| \\ \leq&\|(\rho M^*M)^{-1}M^*(I-2P)(\zeta_k - \zeta_\star)\| \\ \leq&\|(\rho M^*M)^{-1}\|\,\Big(\|M^*(I-2P)R^k(\zeta_0 - \zeta_\star)\| \\ &+ \rho\sum_{\ell=1}^k \|M^*(I-2P)R^{k-\ell}MH^{-1}E(x_\ell - \boldsymbol{x}_\star)\|\Big). \end{aligned}$$

Our purpose is to show that

$$\forall \varepsilon > 0,\ \sup_k (\alpha+\varepsilon)^{-k}\|x_{k+1}-\boldsymbol{x}_\star\| < \infty. \qquad (25)$$

This shows indeed that

$$\limsup_k \frac{\log\|x_{k+1}-\boldsymbol{x}_\star\|}{k} \leq \log\alpha + \log(1+\frac{\varepsilon}{\alpha})$$

for any $\varepsilon > 0$, which is equivalent to Theorem 2-*ii)*.

Fix $\varepsilon > 0$. Recall that $\|M^*(I-2P)R^k\| \leq C\alpha^k$ where C is a constant, $x_k \to \boldsymbol{x}_\star$, and $\|E(x_k-\boldsymbol{x}_\star)\| = o(\|x_k-\boldsymbol{x}_\star\|)$. By delaying the time origin as much as needed, we can assume that for any $k > 0$ and any $\ell \in \{1,\dots,k\}$,

$$\|(\rho M^*M)^{-1}\|\,\|M^*(I-2P)R^k(\zeta_0-\zeta_\star)\| \leq \alpha^{k+1},$$
$$\rho\|(\rho M^*M)^{-1}\|\,\|M^*(I-2P)R^{k-\ell}MH^{-1}\| \leq \alpha^{k+1-\ell},$$
$$\|E(x_\ell-\boldsymbol{x}_\star)\| \leq \delta\|x_\ell-\boldsymbol{x}_\star\|, \text{ and}$$
$$\|x_0-\boldsymbol{x}_\star\| \leq B = \frac{\varepsilon}{\varepsilon-\delta(\alpha+\varepsilon)}$$

where we choose $\delta < \varepsilon/(\alpha+\varepsilon)$. With this choice of the time origin, we have

$$\|x_{k+1}-\boldsymbol{x}_\star\| \leq \alpha^{k+1} + \delta\sum_{\ell=1}^{k}\alpha^{k+1-\ell}\|x_\ell-\boldsymbol{x}_\star\|.$$

Putting $w_k = (\alpha+\varepsilon)^{-k}\|x_k-\boldsymbol{x}_\star\|$, this inequality is rewritten

$$w_{k+1} \leq \Big(\frac{\alpha}{\alpha+\varepsilon}\Big)^{k+1} + \delta\sum_{\ell=0}^{k}\Big(\frac{\alpha}{\alpha+\varepsilon}\Big)^{k+1-\ell} w_\ell.$$

We know that $w_0 \leq B$. Assume that $w_1,\dots,w_k \leq B$. Then

$$w_{k+1} < 1 + \delta B\sum_{\ell=0}^{k}\Big(\frac{\alpha}{\alpha+\varepsilon}\Big)^{k+1-\ell} < 1 + \frac{\delta B}{1-\frac{\alpha}{\alpha+\varepsilon}} = B$$

and Inequality (25) is established.

We now show Theorem 2-*iii)*. From the equations (21), (23) and (24), we have

$$\begin{aligned} x_{k+1}-\boldsymbol{x}_\star &= -H^{-1}M^*(I-2P)R^k(\zeta_0-\zeta_\star) \\ &\quad + \rho\sum_{\ell=1}^{k} H^{-1}M^*(I-2P)R^{k-\ell}MH^{-1}E(x_\ell-\boldsymbol{x}_\star) \\ &\quad - H^{-1}E(x_{k+1}-\boldsymbol{x}_\star) \\ &= X_k + Y_k + Z_k. \end{aligned}$$

By the argument establishing Theorem 2-*iii)* in the quadratic case, for any ζ_0 outside a set of Lebesgue measure zero, there is a function $g(k)$ such that $\|X_k\| \geq \alpha^k g(k)$ and $a = \limsup_k g(k) > 0$. For the sake of contradiction, assume that $\limsup_k(k^{-1}\log\|x_k-\boldsymbol{x}_\star\|) < \log\alpha$ for this ζ_0. Then $\|x_k-\boldsymbol{x}_\star\| \leq C\beta^k$ for some $C > 0$ and some $\beta \in (0,\alpha)$. By delaying the time origin as much as needed, we can assume that

$$\frac{\|E(x_\ell-\boldsymbol{x}_\star)\|}{\|x_\ell-\boldsymbol{x}_\star\|} \leq \delta = \frac{a(\alpha-\beta)}{2\alpha} \text{ for any } \ell > 0, \text{ and}$$
$$\|Y_k+Z_k\| \leq \delta\sum_{\ell=1}^{k}\alpha^{k-\ell}\beta^\ell \ + \ \delta\beta^k$$
$$< \delta\frac{\alpha}{\alpha-\beta}\alpha^k.$$

We therefore have

$$\|x_{k+1}-\boldsymbol{x}_\star\| \geq \|X_k\| - \|Y_k+Z_k\| \geq \alpha^k\Big(g(k) - \delta\frac{\alpha}{\alpha-\beta}\Big).$$

Since $\limsup_k(g(k)-\delta\alpha/(\alpha-\beta)) = a/2 > 0$, we obtain $\limsup_k(k^{-1}\log\|x_k-\boldsymbol{x}_\star\|) \geq \log\alpha$. Theorem 2-*iii)* is proven.

VI. Numerical Illustrations

We first provide a numerical illustration in the special cases described in Section IV-A and IV-B. The second order derivative $\sigma_\star^2$ of the functions f_n at the minimum is set to 16. Figures 3 and 4 represent the rate α as a function of the step-size ρ of the algorithm in the case of a centralized network and a ring network respectively. In the centralized case, the optimal value of ρ coincides with σ^2 and is thus independent of N. In the ring network, the rate α depends on both ρ and N.

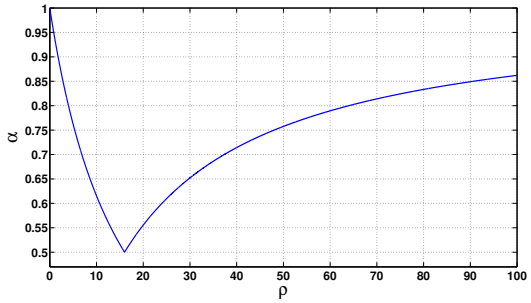


Fig. 3: Rate α as a function of ρ - Centralized network - $\sigma_\star^2 = 16$.

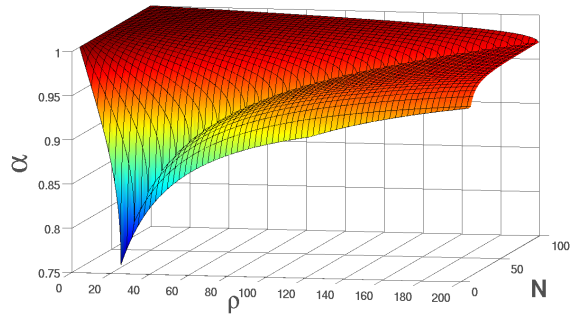


Fig. 4: Rate α as a function of ρ and N - Ring network - $\sigma_\star^2 = 16$.

We now address the case where the second order derivatives are not necessarily equal. We set $N = 5$ and assume that the values of $f_n''(x_\star)$ for all agents n are respectively equal to 4, 9, 16, 25 and 39. Figure 5 represents the rate α as a function of ρ.

Finally, we compare our theoretical result with the performance of ADMM observed by simulation; we also compare with the bound of Shi et al. [11]. These simulations were conducted on a 20 nodes Random Geometric Graph with radius 0.2. The sets $\{A_\ell\}_\ell$ are taken equal to the pairs of connected agents as in Example 2 of Section II-E. We plot

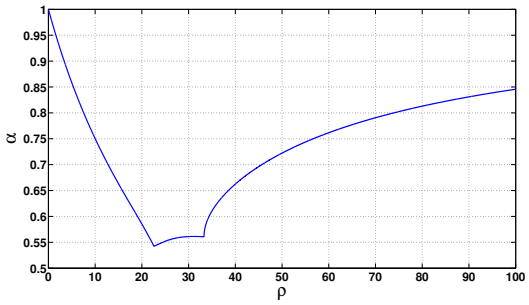


Fig. 5: Rate α as a function of ρ - Centralized network - $N = 5$ - Distinct second order derivatives.

$-k^{-1} \log \|x_k - x_\star\|$ as a function of the number of iterations k.

In Figure 6, the functions are taken as $f_n(x) = \exp(\beta_n x)$ where the β_n's are drawn uniformly in $[-10, 10]$ then centered (in that case $\sum_n f_n$ admits $x_\star = 0$ as unique minimizer) and ρ is set to 20. Here, the bound of Shi et al. [11] is not defined (*i.e.*, its log is equal to zero). As expected, Figure 6 shows that the rate α is tight in the sense that $-k^{-1} \log \|x_k - x_\star\|$ numerically converges to $-\log \alpha$.

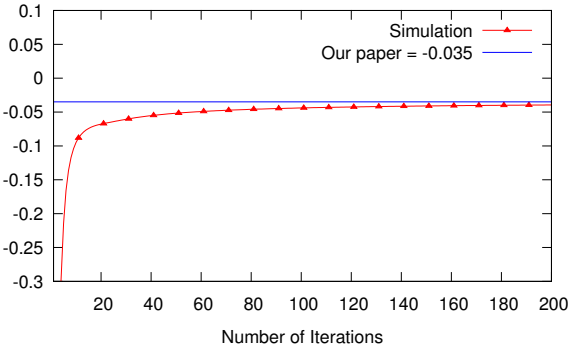


Fig. 6: $k^{-1} \log \|x_k - x_\star\|$ as a function of the number of iterations k - $N = 20$ - Exponential functions

We also investigate the case of quadratic functions. In that case, the bound of Shi et al. [11] is well defined and plotted in Figure 7. The functions f_n are defined as $f_n(x) = a_n(x - b_n)^2$ where the a_n's are drawn uniformly in $[1, 100]$ and the b_n's are drawn from a Gaussian distribution with mean 5 and variance 100. The parameter ρ has been set to 100 as this seems to be a good choice to take it around the second order derivatives from the above simulations and derivations. We observe that our characterization of the convergence rate is tight in the sense that it fits the empirical performance of ADMM, whereas a gap exists between the latter and the bound of [11].

VII. Conclusion

In this paper, we addressed the rate of convergence of ADMM to solve distributively the minimization problem $\inf_x \sum_{n=1}^N f_n(x)$ where the f_n's are private convex functions. Letting $\boldsymbol{x}_\star$ be the minimizer and assuming that the functions are twice differentiable at $\boldsymbol{x}_\star$, we obtained an explicit characterization of the linear convergence rate under the form

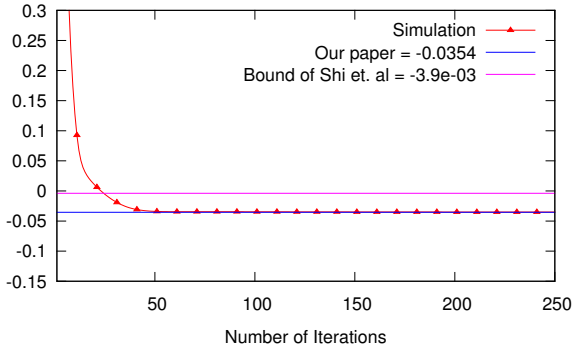


Fig. 7: $k^{-1} \log \|x_k - x_\star\|$ as a function of the number of iterations k - $N = 20$ - Quadratic functions

of the spectral radius of a matrix depending to the Hessian $\sum_n \nabla^2 f_n(\boldsymbol{x}_\star)$ and the communication network. Under mild conditions, it is shown that the obtained rate is tight in the sense that the actual convergence rate is no faster than the one obtained.

In practice, our analysis is useful to accurately predict the performance of ADMM and to optimize various parameters, such as the step-size of the algorithm and the communication graph. Our method potentially allows to design augmented ADM methods with enhanced convergence rate.

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 preprint arXiv:1110.4198*, 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 Transactions on Automatic Control*, vol. 58, no. 2, pp. 391 – 405, February 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 Transactions 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] 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.

[9] E. Wei and A. Ozdaglar, "On the $O(1/k)$ convergence of asynchronous distributed Alternating Direction Method of Multipliers," *arXiv preprint arXiv:1307.8254*, 2013.

[10] 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.

[11] 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.

[12] 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.

[13] A. Teixeira, E. Ghadimi, I. Shames, H. Sandberg, and M. Johansson, "Optimal scaling of the admm algorithm for distributed quadratic programming," in *IEEE Conference on Decision and Control (CDC)*, 2013.

[14] P. Bianchi, W. Hachem, and F. Iutzeler, "A stochastic coordinate descent primal-dual algorithm and applications to large-scale composite optimization," *arXiv preprint arXiv:1407.0898*, 2014.

[15] 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.

[16] 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.

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

[19] 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.

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

[21] D. Boley, "Local linear convergence of the Alternating Direction Method of Multipliers on quadratic or linear programs," *SIAM Journal on Optimization*, vol. 23, no. 4, pp. 2183–2207, 2013.

[22] E. Ghadimi, A. Teixeira, I. Shames, and M. Johansson, "Optimal parameter selection for the alternating direction method of multipliers (admm): Quadratic problems," *arXiv preprint arXiv:1306.2454*, 2013.

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

[24] Z.-Q. Luo and P. Tseng, "On the convergence rate of dual ascent methods for linearly constrained convex minimization," *Mathematics of Operations Research*, vol. 18, no. 4, pp. 846–867, 1993.

[25] 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.

[26] S. Boyd, A. Ghosh, B. Prabhakar, and D. Shah, "Randomized gossip algorithms," *IEEE Trans. on Information Theory*, vol. 52, no. 6, pp. 2508–2530, June 2006.

[27] T. Erseghe, D. Zennaro, E. Dall'Anese, and L. Vangelista, "Fast consensus by the Alternating Direction Multipliers Method," *IEEE Trans. on Signal Processing*, vol. 59, no. 11, pp. 5523–5537, Nov 2011.

[28] R. Carli, F. Fagnani, A. Speranzon, and S. Zampieri, "Communication constraints in the average consensus problem," *Automatica*, vol. 44, no. 3, pp. 671–684, 2008.

[29] 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.

[30] D. Gabay and B. Mercier, "A dual algorithm for the solution of nonlinear variational problems via finite element approximation," *Computers & Mathematics with Applications*, vol. 2, no. 1, pp. 17–40, 1976.

[31] R. M. Gray, *Toeplitz and circulant matrices: A review*, Now Pub, 2006.

APPENDIX

We start by introducing a notation and by recalling a known fact. Given three 2×2 matrices G_{-1}, G_0 and G_1, we denote as $A = \mathrm{circ}_N(G_0 + e^{\imath\lambda}G_1 + e^{-\imath\lambda}G_{-1})$ the $2N \times 2N$ block circulant matrix

$$A = \begin{bmatrix} G_0 & G_1 & & G_{-1} \\ G_{-1} & G_0 & G_1 & \\ & & \ddots & \\ G_1 & & G_{-1} & G_0 \end{bmatrix}.$$

By generalizing a well known result for circulant matrices (see *e.g.* [31]), we know that the eigenvalue spectrum of A coincides with the set of eigenvalues of the trigonometric matrix polynomial $\mathcal{G}(e^{\imath\lambda}) = G_0 + \exp(-\imath\lambda)G_{-1} + \exp(\imath\lambda)G_1$ taken at $\lambda = 2\pi k/N$ for $k = 0, \dots, N-1$.
Getting back to our model, we have $T = 2N$,

$$M = \begin{bmatrix} 1 & & & & & & \\ & 1 & & & & & \\ & 1 & & & & & \\ & & 1 & & & & \\ & & 1 & & & & \\ & & & \ddots & & & \\ & & & & 1 & & \\ & & & & & 1 & \\ & & & & & 1 & \\ 1 & & & & & & \end{bmatrix}$$

and

$$P = \frac{1}{2}\begin{bmatrix} 1 & 1 & & & & & & & & \\ 1 & 1 & & & & & & & & \\ & & 1 & 1 & & & & & & \\ & & 1 & 1 & & & & & & \\ & & & & \ddots & & & & & \\ & & & & & 1 & 1 & & & \\ & & & & & 1 & 1 & & & \\ & & & & & & & 1 & 1 & \\ & & & & & & & 1 & 1 & \end{bmatrix}.$$

Since $M^*M = 2I_N$, we also have

$$\begin{aligned} Q &= \frac{\rho}{\sigma_\star^2 + 2\rho}\begin{bmatrix} 1 & & & & & & & 1 \\ & 1 & 1 & & & & & \\ & 1 & 1 & & & & & \\ & & & \ddots & & & & \\ & & & & 1 & 1 & & \\ & & & & 1 & 1 & & \\ 1 & & & & & & & 1 \end{bmatrix} \\ &= \frac{\rho}{\sigma_\star^2 + 2\rho}\mathrm{circ}_N\Big(I_2 + e^{\imath\lambda}\begin{bmatrix} 0 & 0 \\ 1 & 0 \end{bmatrix} + e^{-\imath\lambda}\begin{bmatrix} 0 & 1 \\ 0 & 0 \end{bmatrix}\Big). \end{aligned}$$

Noticing that $R = I - P - Q + 2QP$ and writing $a = \rho/(\sigma_\star^2 + 2\rho)$, we have

$$\begin{aligned} 2QP &= a\ \mathrm{circ}_N\Big(\Big(I_2 + e^{\imath\lambda}\begin{bmatrix} 0 & 0 \\ 1 & 0 \end{bmatrix} + e^{-\imath\lambda}\begin{bmatrix} 0 & 1 \\ 0 & 0 \end{bmatrix}\Big)\mathbf{1}\mathbf{1}^*\Big) \\ &= a\ \mathrm{circ}_N\Big(\begin{bmatrix} 1 & 1 \\ 1 & 1 \end{bmatrix} + e^{\imath\lambda}\begin{bmatrix} 0 & 0 \\ 1 & 1 \end{bmatrix} + e^{-\imath\lambda}\begin{bmatrix} 1 & 1 \\ 0 & 0 \end{bmatrix}\Big). \end{aligned}$$

Therefore, $R = \mathrm{circ}_N(G_0 + e^{\imath\lambda}G_1 + e^{-\imath\lambda}G_{-1})$ where

$$\begin{aligned} G_0 &= I_2 - \frac{1}{2}\mathbf{1}\mathbf{1}^* - aI_2 + a\mathbf{1}\mathbf{1}^* = \frac{1}{2}\begin{bmatrix} 1 & -\frac{\sigma_\star^2}{\sigma_\star^2+2\rho} \\ -\frac{\sigma_\star^2}{\sigma_\star^2+2\rho} & 1 \end{bmatrix}, \\ G_1 &= \frac{\rho}{\sigma_\star^2 + 2\rho}\Big(\begin{bmatrix} 0 & 0 \\ -1 & 0 \end{bmatrix} + \begin{bmatrix} 0 & 0 \\ 1 & 1 \end{bmatrix}\Big) = \frac{\rho}{\sigma_\star^2 + 2\rho}\begin{bmatrix} 0 & 0 \\ 0 & 1 \end{bmatrix}, \\ G_{-1} &= \frac{\rho}{\sigma_\star^2 + 2\rho}\Big(\begin{bmatrix} 0 & -1 \\ 0 & 0 \end{bmatrix} + \begin{bmatrix} 1 & 1 \\ 0 & 0 \end{bmatrix}\Big) = \frac{\rho}{\sigma_\star^2 + 2\rho}\begin{bmatrix} 1 & 0 \\ 0 & 0 \end{bmatrix}. \end{aligned}$$

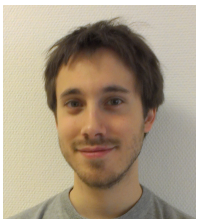
The eigenvalues of $\mathcal{G}(e^{2\imath\pi k/N}) = G_0 + e^{2\imath\pi k/N}G_1 + e^{-2\imath\pi k/N}G_{-1}$ are the solutions of the equation $\lambda^2 - \lambda s_k + D_k = 0$ where

$$s_k = \operatorname{tr}\mathcal{G}(e^{2\imath\pi k/N}) = \frac{\sigma_\star^2 + 2\rho\,(1+\cos(2\pi k/N))}{\sigma_\star^2 + 2\rho}$$

and

$$D_k = \det\mathcal{G}(e^{2\imath\pi k/N}) = \frac{\rho}{\sigma_\star^2 + 2\rho}\,(1+\cos(2\pi k/N))\,.$$

The analysis of these solutions for all $k = 0, \ldots, N-1$, which is tedious but straightforward, directly leads to the expression of α in Corollary 2. By simple algebra, the rate $\alpha = \alpha(\rho)$ is shown to be a continuous function of ρ which is decreasing on the interval $(0, \frac{\sigma_\star^2}{2s_N}]$ and increasing on $[\frac{\sigma_\star^2}{2s_N}, +\infty)$.



Franck Iutzeler was born in Besançon, France, in 1987. He received the Engineering Degree from Telecom ParisTech and the M.Sc. degree from U.P.M.C. (University of Paris VI) both in 2010, and the Ph.D. degree from Telecom ParisTech in 2013. Then, he joined the LANEAS group in Supélec as a post-doctoral researcher. He is currently a post-doctoral researcher at the ICTEAM Institute, Université Catholique de Louvain. His research interests are in the area of optimization and distributed algorithms, particularly distributed estimation and distributed/randomized optimization.

Pascal Bianchi was born in 1977 in Nancy, France. He received the M.Sc. degree of the University of Paris XI and Supélec in 2000 and the Ph.D. degree of the University of Marne-la-Vallée in 2003. From 2003 to 2009, he was an Associate Professor at the Telecommunication Department of Supélec. In 2009, he joined the Statistics and Applications group at LTCI-Telecom ParisTech. His current research interests are in the area of continuous optimization with application to signal processing and distributed systems.

Philippe Ciblat was born in Paris, France, in 1973. He received the Engineering degree from Télécom ParisTech (formerly, Ecole Nationale Supérieure des Télécommunications - ENST) and the M.Sc. degree (DEA, in french) in automatic control and signal processing from the Université de Paris XI, Orsay, France, both in 1996, and the Ph.D. degree from University of Marne-la-Vallée, France, in 2000. He eventually received the HDR degree from the University of Marne-la-Vallée, France, in 2007. In 2001, he was a Postdoctoral Researcher with University of Louvain, Belgium. At the end of 2001, he joined the Communications and Electronics Department at Télécom ParisTech, Paris, France, as an Associate Professor. Since 2011, he has been (full) Professor in the same institution. He served as Associate Editor for the IEEE Communications Letters from 2004 to 2007. Since 2008, he has served as Associate Editor and Area Editor for the IEEE Transactions on Signal Processing.

His research areas include statistical signal processing (blind equalization, frequency estimation, and distributed estimation), signal processing for digital communications (synchronization, localization, estimation for cooperative communications), and resource allocation (multiple access technique optimization, power allocation in cooperative communication, and system designs).

Walid Hachem was born in Bhamdoun, Lebanon, in 1967. He received the Engineering degree in telecommunications from St Joseph University (ESIB), Beirut, Lebanon, in 1989, the Masters degree from Telecom ParisTech, France, in 1990, the PhD degree in signal processing from Université de Marne-La-Vallée in 2000 and he Habilitation à Diriger des Recherches from Université de Paris-Sud in 2006.

Between 1990 and 2000 he worked in the telecommunications industry with Philips, then with FERMA, as a signal processing engineer. In 2001 he joined the academia as a faculty member at Supélec, France. In 2006, he joined the CNRS (Centre national de la Recherche Scientifique), where he is now a research director at Telecom ParisTech. His teaching and research interests concern the analysis of communication systems using large random matrix theory, and the detection and estimation in sensor networks. He served as an associate editor for the IEEE Transactions on Signal Processing between 2007 and 2010.