Convergence rate analysis of a subgradient averaging algorithm for distributed optimisation with different constraint sets

Licio Romao, Kostas Margellos, Giuseppe Notarstefano, and Antonis Papachristodoulou

***Abstract*— We consider a multi-agent setting with agents exchanging information over a network to solve a convex constrained optimisation problem in a distributed manner. We propose a new algorithm based on local subgradient exchange under undirected time-varying communication. First, we prove asymptotic convergence of the iterates to a minimum of the given optimisation problem for time-varying step-sizes of the form $c(k) = \frac{\eta}{k+1}$, for some $\eta > 0$. We then restrict attention to step-size choices $c(k) = \frac{\eta}{\sqrt{k+1}}$, $\eta > 0$, and establish a convergence rate of $\mathcal{O}(\frac{\ln(k)}{\sqrt{k}})$ in objective value. Our algorithm extends currently available distributed subgradient/proximal methods by: (i) accounting for different constraint sets at each node, and (ii) enhancing the convergence speed thanks to a subgradient averaging step performed by the agents. A numerical example demonstrates the efficacy of the proposed algorithm.**

I. Introduction

We focus on distributed algorithms to solve convex optimisation problems. They are motivated by applications, such as sensor networks, robust estimation and source localisation [1], that require distribution of computational power and possibly data to alleviate the burden caused by data size. The main challenge is to devise fast and efficient algorithms that converge to an optimal solution of the centralised problem without requiring global information.

In the past decade, motivated by [2], [3], distributed optimisation has drawn the attention of the community because of their relevance to important real-world problems. Indeed, [2], [3] proposed a distributed projected subgradient algorithm that converges under time-varying network for problems with a common constraint set known by all agents. More recently, [4] extended these results to the time-varying directed case by relying on a push-sum consensus protocol [5]. They showed convergence rates of $\mathcal{O}(\frac{\ln(k)}{\sqrt{k}})$ for the function value at the running average of the local iterates. Another contribution in this direction was made by [6], which proposed a proximal algorithm in which the local iterates maintained by the agents converge to a point in the optimal set under time-varying network and for problems with different constraint sets.

A new research direction involves the use of gradient tracking, mainly because sharp convergence results can be obtained for directed and undirected communication networks [7], [8], [9]. To achieve this improved performance, each agent maintains an additional local variable that tracks asymptotically the (sub-)gradient of the global function. Agents then use this additional information, which provides a more accurate direction towards minimising the overall objective function, to update their estimate of the solution.

The contribution of this paper is twofold: 1) unlike the aforementioned literature, we propose a new algorithm based on subgradient averaging that can simultaneously cope with non-differentiable local objective functions, and different constraint sets, while accounting for a time-varying communication network. By showing convergence in iterates for a step-size of the form $c(k) = \frac{\eta}{k+1}$, $\eta > 0$, we set a new framework accounting for the presence of different constraint sets and subgradient exchanges. As a consequence of this result, we expect faster practical convergence when compared to standard projected subgradient algorithms because we use an additional information that better approximate the subgradient of the global function; 2) We build upon the results of [6] and establish a convergence rate of $\mathcal{O}(\frac{\ln(k)}{\sqrt{k}})$, when the step-size is $c(k) = \frac{\eta}{\sqrt{k+1}}$, $\eta > 0$. Even though similar bounds have appeared in the literature, the present analysis offers the first convergence rate for the particular subgradient averaging scheme with the same rate as for standard distributed subgradient methods. We highlight that in our results we allow for different constraint sets, thus extending the scope of existing algorithms in the literature. Note that lifting constraints in the objective via characteristic functions, although possible, is not amenable to algorithms like [3], [10], [11], [12], as this would render the subgradient of the resulting objective unbounded.

The paper is organised as follows. In Section II we present the problem statement, the main assumptions, as well as a description of the proposed algorithm. Section III contains the main results of this paper related to convergence in iterates and a convergence rate analysis as far as the optimal value is concerned. Section IV provides a numerical example to demonstrate the main algorithmic features of our scheme. Finally, some concluding remarks are provided in Section V. All omitted proofs can be found in [13].

II. Problem statement

A. Problem set-up and Assumptions

Consider the following optimisation problem

$$\begin{aligned} &\underset{x}{\text{minimise}} \quad f(x) = \sum_{i=1}^{m} f_i(x) \\ &\text{subject to} \quad x \in \cap_{i=1}^{m} X_i \end{aligned} \tag{1}$$

L. Romao is supported by the Coordination for the Improvement of Higher Education Personnel (CAPES) - Brazil. The work of K. Margellos and A. Papachristodoulou is supported in part by the Engineering and Physical Sciences Research Council (EPSRC) UK under grants EP/P03277X/1 and EP/M002454/1, respectively.

L. Romao, K. Margellos and A. Papachristodoulou are with the Department of Engineering Science, University of Oxford, Parks Road, Oxford OX1 3PJ, UK. {licio.romao, kostas.margellos, antonis}@eng.ox.ac.uk.

G. Notarstefano is with the Department of Electrical, Electronic and Information Engineering, University of Bologna, Bologna, Italy. giuseppe.notarstefano@unibo.it.

where $x \in \mathbb{R}^n$ is the global decision vector, and $f_i : \mathbb{R}^n \to \mathbb{R}$ and $X_i \subset \mathbb{R}^n$, for all $i = 1, \ldots, m$, constitute the local objective function and constraint set for agent i, respectively. We suppose that each agent i possesses as private information the triple (x_i, f_i, X_i), where the first component x_i is a local copy of the global variable x.

The goal is for all agents to agree on the local variables, that is, $x_i = x^\star$, for all $i = 1, \ldots, m$, where $x^\star$ belongs to the optimal set of (1), i.e., the subset of $\mathbb{R}^n$ with the property that $f(x^\star) \le f(x)$ for all $x \in \cap_{i=1}^m X_i$. In this paper, we propose a solution for problem (1) under the following assumptions on f_i and X_i.

Assumption 1: (Convexity, compactness and non-emptiness of the interior)

$i)$ For all $i = 1, \ldots, m$, the function f_i is proper and convex (see [14, Chapter 1] for a definition).

$ii)$ The set $X_i \subset \mathbb{R}^n$ is compact and convex for all $i = 1, \ldots, m$, and $\cap_{i=1}^m X_i$ has a non-empty interior. Moreover, we assume that $X_i \subset \cap_{i=1}^m \mathrm{int}(\mathrm{dom} f_i)$ for each $i = 1, \ldots, m$, where $\mathrm{int}(A)$ stands for the interior of the set A.

$iii)$ The distance between the set $\cup_{i=1}^m X_i$ and the complement of the interior of the domain of f (which is closed and convex) is strictly greater than zero, i.e.,

$$\mathrm{dist}\big(\cup_{i=1}^m X_i, \big(\mathrm{int}(\mathrm{dom} f)\big)^c\big) = \inf_{x \in \cup_{i=1}^m X_i, y \in \left(\mathrm{int}(\mathrm{dom} f)\right)^c} \|x - y\|_2^2 > 0.$$

As a consequence[1] of Assumption 1, we have that $\cup_{x \in \mathrm{conv}(\cup X_i)} \partial f(x)$ is a bounded set, that is, $\|g\| \le L$, where $g \in \partial f(x)$ for any $x \in \cup_{i=1}^m X_i$. This result is formally stated in the next Lemma.

Lemma 1: Under Assumption 1, we have that

$i)$ The set $\mathrm{conv}(\cup_{i=1}^m X_i)$ is compact, where $\mathrm{conv}(A)$ is the convex hull of the set A;

$ii)$ The set $\cup_{x \in \mathrm{conv}(\cup X_i)} \partial f(x)$ is non-empty and bounded;

$iii)$ The function f is Lipschitz continuous over $\cap_{i=1}^m X_i$, i.e., there exists a positive scalar L such that

$$|f(x) - f(y)| \le L\|x - y\|, \quad \forall\, x, y \in \cap_{i=1}^m X_i.$$

Typical choices of functions that satisfy Assumption 1 are piecewise-linear functions, quadratic convex functions and the logistic regression function $f_i(x) = \ln(1 + \sum_{i=1}^{\ell} e^{-w_i^T x})$, for some $w_i \in \mathbb{R}^n$.

B. Proposed algorithm

The pseudocode of the proposed scheme is shown in Algorithm 1. We initialise each agents' local variable with an arbitrary $x_i(0) \in X_i$, $i = 1, \ldots, m$; such points are not required to belong to $\cap_{i=1}^m X_i$.

At iteration k, agent i receives x_j from the neighbouring agents, and averages them through $A(k)$, which captures the communication network, to obtain $z_i(k)$. Here we represent the element of the j-th row and i-th column of matrix $A(k)$ by $[A(k)]_j^i$. Agent i then calculates a subgradient, g_i, of its own objective function evaluated at $z_i(k)$ and sends this information back to its neighbours. In the sequel, agent i averages the received $g_j(z_j(k))$ in order to compose a proxy for a subgradient of $f(x)$ (Step 3), called $\tilde{z}_i(k)$. Finally, at Step 4, agents use variables $\tilde{z}_i(k)$ and $z_i(k)$ to update their local estimates by projecting $z_i(k) - c(k)\tilde{z}_i(k)$ onto the local set. Indeed, note that Step 4 can be rewritten as

$$x_i(k+1) = \mathcal{P}_{X_i}[z_i(k) - c(k)\tilde{z}_i(k)]$$

where $\mathcal{P}_{X_i}$ denotes the projection operator onto the set X_i.

Algorithm 1 Proposed Scheme

Require: : $x_i(0), \quad i = 1, \ldots, m$

1: **while** Until convergence **do**

2: $\quad z_i(k) = \sum_{j=1}^m [A(k)]_j^i x_j(k), \qquad \forall\, i = 1, \ldots, m$

3: $\quad \tilde{z}_i(k) = \sum_{j=1}^m [A(k)]_j^i g_j(z_j(k)), \qquad \forall\, i = 1, \ldots, m$

4: $\quad x_i(k+1) = \mathrm{argmin}_{\xi \in X_i}\, \tilde{z}_i(k)^T \xi + \frac{1}{2c(k)}\|z_i(k) - \xi\|_2^2, \quad \forall\, i = 1, \ldots, m$

5: $\quad k \leftarrow k + 1$

6: **end while**

We now characterise $A(k)$ that encodes the network in Algorithm 1. To this end, let $\mathcal{G}(k) = (\mathcal{N}, \mathcal{E}(k))$ be a undirected graph, where $\mathcal{N} = \{1, \ldots, m\}$ is the number of agents and $\mathcal{E}(k) \subset \mathcal{N} \times \mathcal{N}$ is the set of edges at iteration k, that is, if node $(j, i) \in \mathcal{E}(k)$ then node j sends information to node i at iteration k. We associate the time-varying matrix $A(k)$ to the edge set $\mathcal{E}(k)$, with $[A(k)]_j^i \neq 0$ if $(j, i) \in \mathcal{E}(k)$ at time k. As the graph is undirected, matrix $A(k)$ can be chosen to be symmetric. We also define the graph $\mathcal{G}_\infty = (\mathcal{N}, \mathcal{E}_\infty)$, in which $(j, i) \in \mathcal{E}_\infty$ if agent j communicates with agent i infinitely many times. Then, we impose the following assumption on the matrix $A(k)$ in Algorithm 1.

Assumption 2: (Network Properties)

$i)$ The graph $(\mathcal{N}, \mathcal{E}_\infty)$ is strongly connected. Moreover, there exits a uniform upper bound on the communication time for all $(j, i) \in \mathcal{E}_\infty$.

$ii)$ There exists an $\eta \in (0, 1)$ such that $[A(k)]_i^i \ge \eta$ and that if $[A(k)]_j^i > 0$ then we have $[A(k)]_j^i \ge \eta$, for all $k \in \mathbb{N}$ and for all $i, j = 1, \ldots, m$.

$iii)$ Matrix $A(k)$ is doubly stochastic, i.e., $1^T A = 1^T$ and $A1 = 1$.

These are standard hypotheses in the distributed optimisation literature. The interested reader is referred to [2], [3], [10], [6] for more details.

The analysis presented in this paper is divided into two parts. First, we prove asymptotic convergence of the local variables, x_i, to some optimal solution of the centralised problem counterpart under square-summable but not summable step-sizes, e.g., $c(k) = \frac{\eta}{k+1}$, $\eta > 0$. We then show convergence rates of $\mathcal{O}(\frac{\ln(k)}{\sqrt{k}})$ in terms of the function value for the time-varying step-sizes of the form $c(k) = \frac{\eta}{\sqrt{k+1}}$, $\eta > 0$.

III. Algorithm Analysis

A. Convergence in iterates

In this subsection, we impose the following assumption on the step-size.

[1] A thorough discussion of Assumption 1 is given in the extended version of this paper [13]. Note that the bound on $\|g\|$ coincides with the Lipschitz constant in Lemma 1 $iii)$.

Assumption 3: (Non-increasing, square-summable step-size) Let $(c(k))_{k\in\mathbb{N}}$ be the sequence of step-sizes adopted in Step 4 of Algorithm 1. We impose that

i) $c(k) \geq 0$, and $c(k) \geq c(r)$, for all $k, r \in \mathbb{N}$ with $r \geq k$,
ii) $\sum_{k=1}^{\infty} c(k) = \infty$ and $\sum_{k=1}^{\infty} c(k)^2 < \infty$.

A sequence that satisfies Assumption 3 is $c(k) = \frac{\eta}{k+1}$, for some $\eta > 0$. Assumption 3 is necessary to prove one of the main results of this paper, namely, the asymptotic convergence for the sequences $(x_i(k))_{k\in\mathbb{N}}$, for all $i = 1, \ldots, m$, to a point in the set optimal set of (1). To streamline the presentation, the proof is deferred to the Appendix.

Theorem 1: Let $(x_i(k))_{k\in\mathbb{N}}$ be the sequences generated by Algorithm 1, for all $i = 1, \ldots, m$. Under Assumptions 1-3, we have that for some minimizer $x^\star$ in the optimal set of problem (1),

$$\lim_{k\to\infty} \|x_i(k) - x^\star\| = 0, \quad \forall\ i = 1, \ldots, m.$$

Proof: See Appendix. ■

Theorem 1 extends the result in [6] by allowing an agent to communicate subgradient information to neighbouring agents, a feature that significantly speeds up practical convergence.

B. Convergence in value and convergence rate

We impose now the following assumption on the step-size.

Assumption 4: The sequence $(c(k))_{k\in\mathbb{N}}$ used in Step 4 of Algorithm (1) is $c(k) = \frac{\eta}{\sqrt{k+1}}$, for some $\eta > 0$.

Our convergence rate results build on the running average of the iterates generated by Algorithm 1, that is, the sequence

$$\hat{x}_i(k+1) = \frac{c(k+1)x_i(k+1) + S(k)\hat{x}_i(k)}{S(k+1)}, \qquad (2)$$

where $S(k) = \sum_{r=1}^{k} c(r)$, and $(x_i(k))_{k\in\mathbb{N}}$ for all $i = 1, \ldots, m$ are the sequences generated by Algorithm 1, with the initial conditions $\hat{x}_i(0) = x_i(0)$ and $S(0) = 1$. By rewriting expression (2) as

$$\hat{x}_i(k) = \frac{1}{S(k)} \sum_{r=1}^{k} c(r) x_i(r),$$

we observe that the running average can interpreted as a convex combination of previous iterates. The next theorem establishes a convergence rate for the function value along the running average defined in (2).

Theorem 2: Consider the running average defined in (2). Under Assumptions 1, 2, and 4 the following inequality holds for all $k \in \mathbb{N}$

$$\sum_{i=1}^{m} f_i(\hat{x}_i(k)) - f(x^\star) \leq B_1 \frac{1}{\sqrt{k}} + B_2 \frac{\ln(k)}{\sqrt{k}}. \qquad (3)$$

where B_1 and B_2 are positive constants defined in the Appendix.

Proof: See Appendix. ■

Theorem 2 asserts convergence of the function value along the running average $\hat{x}_i(k)$, i.e., all limit point of $(\hat{x}_i(k))_{k\in\mathbb{N}}$ are optimal, but the sequence might exhibit an oscillatory behaviour.

We point out that the result of Theorem 2 further extends the work presented in [6] not only by allowing agents to communicate their (sub-)gradients, but by also unveiling how to adapt the proof line in that paper to come up with convergence results that recover traditional rates for distributed subgradient methods.

IV. Numerical Examples

We now demonstrate the results through a numerical example. We consider problem (1) in which the functions $f_i(x)$ are piecewise-linear (hence, non-differentiable) and given by

$$f_i(x) = \max\left\{ |x^{(1)}|, \max_{2\leq\ell\leq n} |x^{(\ell)} - (i+1)x^{(\ell-1)}| \right\},$$

where $x^{(\ell)}$, $\ell = 1, \ldots, n$, represents the ℓ-th component of the vector[2] x. This example was analysed in [15], and was originally adapted from [16] to the distributed case. We consider the case where $n = 20$ and $m = 12$. Note that this function is a convex homogeneous function, therefore the optimal solution and optimal value for this problem are $x^\star = 0$ and $f(x^\star) = 0$, respectively.

As for the communication network, we consider a time-invariant undirected network (notice that the relevant matrices do not depend on the iteration index k) whose topology is given by a line graph. Given the topology, we generate a doubly stochastic such that,

$$[A]^i_j = \frac{1}{1 + \max\{\mathcal{N}_i, \mathcal{N}_j\}}, \quad i \neq j,$$

where $\mathcal{N}_i$ and $\mathcal{N}_j$ are the number of neighbours of agent i and j, respectively. The diagonal elements of A are defined as $[A]^i_i = 1 - \sum_{j\neq i} [A]^i_j$. Note that matrix A is symmetric and doubly stochastic, as the network undirected. Time-varying networks could be accommodated by the proposed algorithm.

Agents' constraint sets are different, and for each one we assumed that the constraint set is encoded by componentwise upper and lower limits on x. These limits were randomly generated from a uniform distribution. In our first numerical investigation, we set the step-sizes to be $c(k) = \frac{1}{k+1}$, aligned with Assumption 3. To investigate the statement of Theorem 1, we monitor the evolution

$$\text{Res}_x(k) = \sum_{i=1}^{12} \|x_i(k) - x^\star\|$$

for $k = 1, \ldots, 10{,}000$ iterations. This result is shown in Figure 1, where the solid blue line corresponds to the iterates generated by Algorithm 1, initialised at the optimal solution $x^\star = 0$. Observe that the iterates do not stay at the optimal solution as the function is not differentiable at the origin, implying that there exists a nonzero subgradient such that the iterate sequence escapes from the optimal solution. However, we also observe that after some initial perturbation, the iterates are steered back towards the optimal solution, thus supporting the results presented in Theorem 1.

In the second part of our numerical investigation, we choose the time-varying step-size according to Assumption 4,

[2] Variable $x^{(\ell)}$ should not be related to x_i which corresponds to a local copy of x maintained by agent i, rather than to a particular component.

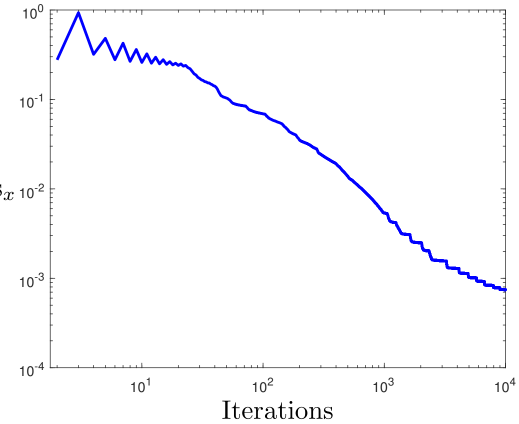

Fig. 1. Evolution of $\mathrm{Res}_x(k)$ for Algorithm 1 (solid, blue line). Both axes are in logarithmic scale.

i.e., $c(k) = \frac{100}{\sqrt{k+1}}$. To investigate the statement of Theorem 2, we monitor the evolution of

$$\mathrm{Res}_f(k) = \sum_{i=1}^{12} f_i(\hat{x}_i(k)) - f(x^\star),$$

where $\hat{x}_i(k)$ is defined in (2), for $k = 1, \ldots, 100{,}000$ iterations. We initialized Algorithm 1 with $x_i^{(\ell)}(0) = 0.1$ for $\ell = 1, \ldots, 19$, and $x_i^{(20)}(0) = 1$, for all $i = 1, \ldots, 12$. The results are illustrated in Figure 2. The theoretical bound $\mathcal{O}(\frac{\ln(k)}{\sqrt{k}})$ is depicted as the dashed-dot black line. The solid blue curve is the sequence $\sum_{i=1}^m f_i(\hat{x}_i(k))$ for the iterates generated by Algorithm 1. Observe that the results comply with the theoretical bound of Theorem 2.

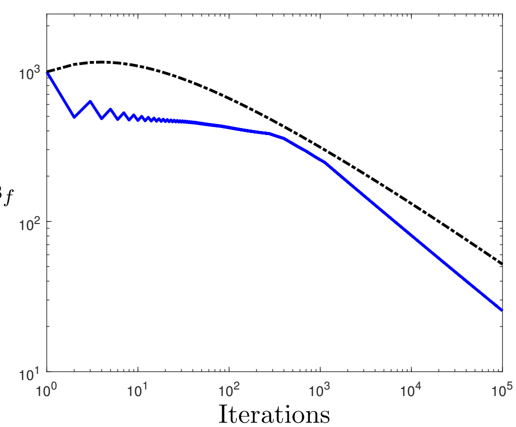

Fig. 2. Evolution of $\mathrm{Res}_f(k)$ for Algorithm 1 (solid, blue line). The solid black line represents an estimate of the theoretical bound (up to a constant factor) predicted by means of Theorem 2. Both axes are in logarithmic scale.

V. Conclusion

In this paper, we presented and analysed a new algorithm that uses subgradient averaging to solve an optimisation problem in a distributed manner. The presented analysis captures possibly different constraint sets and time-varying undirected communication network. We presented two main results: (1) we considered a time-varying step-size of the form $c(k) = \frac{\eta}{k+1}, \eta > 0$, and proved convergence of the generated iterates to an optimiser of the centralised problem counterpart. This outcome generalises current results in the literature as it deals simultaneously with different constraint sets and subgradient exchange between neighbouring agents; (2) for time-varying step-sizes of the form $c(k) = \frac{\eta}{\sqrt{k+1}}, \eta > 0$, we established a convergence rate of $\mathcal{O}(\frac{\ln(k)}{\sqrt{k}})$ as far as the function value of the running average of the local iterates is concerned. We recovered standard convergence rates of distributed subgradient methods; however, extended them to the more general case in which agents are allowed to have their own constraint set. A numerical example has been presented to demonstrate the obtained results.

Appendix

A. Auxiliary results and proofs of Section III-A

Let

$$v(k) = \frac{1}{m} \sum_{i=1}^m x_i(k), \tag{4}$$

be the average of the estimates at time k. Since this quantity might not necessarily belong to the feasible set $\cap_{i=1}^m X_i$, we also define

$$\bar{v}(k) = \frac{\rho}{\epsilon(k) + \rho} v(k) + \frac{\epsilon(k)}{\epsilon(k) + \rho} \bar{x}, \tag{5}$$

where $\bar{x}$ is a point in the interior of the feasible set (which is non-empty by Assumption 1), and $\epsilon(k) = \sum_{i=1}^m \mathrm{dist}(v(k), X_i)$. As shown in [3], the point $\bar{v}(k)$ is in $\cap_{i=1}^m X_i$ for all $k \in \mathbb{N}$. Define $e_i(k+1) = x_i(k+1) - z_i(k)$, and note that Step 2 of Algorithm 1 can be written as

$$x_i(k+1) = \sum_{j=1}^m [A(k)]_j^i x_j(k) + e_i(k+1), \tag{6}$$

which can be interpreted as perturbed consensus protocol.

Lemma 2: The following relations hold.

i) Let $(x_i(k))_{k\in\mathbb{N}}$ for all $i = 1, \ldots, m$ be the sequences generated by Algorithm 1, and $(v(k))_{k\in\mathbb{N}}$ and $(\bar{v}(k))_{k\in\mathbb{N}}$ be defined as in (4) and (5), respectively. Then, under Assumption 1,

$$\sum_{i=1}^m \|x_i(k+1) - \bar{v}(k)\| \leq \mu \sum_{i=1}^m \|x_i(k) - v(k)\|,$$

where $\mu = \frac{2}{\rho} mD + 1$, and D is the diameter of the set $\cup_{i=1}^m X_i$ (which is well-defined by Lemma 1 *i)*).

ii) Let $(x_i(k))_{k\in\mathbb{N}}$ for all $i = 1, \ldots, m$ and $(v(k))_{k\in\mathbb{N}}$ be as in item *i)*, and consider the definition of the error $e_i(k)$. Then, under Assumption 2, we have that

$$\|x_i(k+1) - v(k+1)\| \leq \lambda q^k \sum_{j=1}^m \|x_j(0)\| + \|e_i(k+1)\|$$
$$+ \sum_{r=0}^{k-1} \lambda q^{k-r-1} \sum_{j=1}^m \|e_j(r+1)\| + \frac{1}{m} \sum_{j=1}^m \|e_j(k+1)\|,$$

where $\lambda = 2(1 + \eta^{-(m-1)T})/(1 - \eta^{(m-1)T}) \in \mathbb{R}_+$ and $q = (1 - \eta^{(m-1)T})^{\frac{1}{(m-1)T}} \in (0, 1)$, holds for all $k \in \mathbb{N}$ and for all $i = 1, \ldots, m$.

iii) Let $(c(k))_{k\in\mathbb{N}}$ be a non-increasing and non-negative sequence, and $\bar{L}$ be a positive scalar. If Assumption 2 holds, then for all $k \in \mathbb{N}$ we have that

$$2\bar{L}\sum_{k=0}^{N} c(k)\sum_{i=1}^{m}\|x_i(k+1)-\bar{v}(k+1)\| < \beta_1\sum_{k=0}^{N}\sum_{i=1}^{m}\|e_i(k+1)\|^2 + \beta_2\sum_{k=0}^{N}c(k)^2 + \beta_3,$$

where $\beta_1 \in (0,1)$, and β_2 and β_3 are positive constants.

Proof: The proof of *i)* is presented in [6, Lemma 1]. For *ii)*, see [6, Lemma 2]. Finally, the proof of *iii)* follows the line of [6, Lemma 3]. ■

Observe that the values of λ and q in Lemma 2 *ii)*, depend on the parameter T that characterises the uniform bound in Assumption 2; and on η, the lower bound for the elements of $A(k)$. In fact, these parameters also depend on the connectivity of the communication network. Studying this dependence is an interesting question *per se*, but one that is not pursued in this paper. The reader is referred to [10] for a thorough discussion on this aspect. Moreover, it is important to notice that in Lemma 2 *iii)*, we can choose any value for the $\beta_1 \in (0,1)$, at the price of increasing the value β_2 and modifying β_3. For the presented analysis, the specific values for β_2 and β_3 are not important, provided these are positive.

Item *ii)* of the following lemma is a novel derivation, allowing the auxiliary sequences $\alpha_1(k)$ and $\alpha_2(k)$ to be iteration dependent. This is instrumental for the proof of Theorem 2.

Lemma 3: Let $(x_i(k))_{k\in\mathbb{N}}, (z_i(k))_{k\in\mathbb{N}}$ and $(\tilde{z}_i(k))_{k\in\mathbb{N}}$, $i = 1,\dots,m$, be the sequences generated by Algorithm 1, and $x^\star$ by any point in the set of optimal solutions of problem (1). Suppose Assumptions 1 and 2 hold. Then,

i) For all $k \in \mathbb{N}$ we have that

$$2c(k)\sum_{i=1}^{m}\tilde{z}_i(k)^T(x_i(k+1)-x^\star) + \sum_{i=1}^{m}\|e_i(k+1)\|^2 + \sum_{i=1}^{m}\|x_i(k+1)-x^\star\|^2 \le \sum_{i=1}^{m}\|x_i(k)-x^\star\|^2. \quad (7)$$

ii) For any $\beta_1 \in (0,1)$, there exist sequences $(\alpha_1(k))_{k\in\mathbb{N}}$ and $(\alpha_2(k))_{k\in\mathbb{N}}$ such that $1-\beta_1-\alpha_1(k)-\alpha_2(k) \ge 0$ for all $k \in \mathbb{N}$ and that

$$\begin{aligned}&2\sum_{k=0}^{N}c(k)\sum_{i=1}^{m}(f_i(\bar{v}(k+1))-f_i(x^\star))\\&+\sum_{k=0}^{N}(1-\alpha_1(k)-\alpha_2(k)-\beta_1)\sum_{i=1}^{m}\|e_i(k+1)\|_2^2\\&+\sum_{k=0}^{N}\sum_{i=1}^{m}\|x_i(k+1)-x^\star\|^2 \le \sum_{k=0}^{N}\sum_{i=1}^{m}\|x_i(k)-x^\star\|^2\\&+\sum_{k=0}^{N}\Big(mL^2\frac{\alpha_1(k)+\alpha_2(k)}{\alpha_1(k)\alpha_2(k)}+\beta_2\Big)c(k)^2+\beta_3\end{aligned}$$

Proof: The proof of *i)* is omitted for brevity, as it follows from similar arguments of Lemma 5 in [6], defining $\tilde{z}_i(k)^T(x_i(k)-x^\star)$ in place of the f_i in that paper. This proof can also be founded in [13].

To prove *ii)* we use *i)*. Indeed, consider the first term of the left-hand side in inequality (7), and rewrite it as

$$\begin{aligned}2c(k)\sum_{i=1}^{m}\tilde{z}_i(k)^T(x_i(k+1)-x^\star) = \\ 2c(k)\sum_{i=1}^{m}\tilde{z}_i(k)^T(x_i(k+1)-\bar{v}(k+1))\\ +2c(k)\sum_{i=1}^{m}\tilde{z}_i(k)^T(\bar{v}(k+1)-x^\star)\end{aligned} \quad (8)$$

by adding and subtracting $\bar{v}(k+1)$. Now, let us consider the terms of the right hand-side of (8) separately. First, observe that

$$\begin{aligned}&2c(k)\sum_{i=1}^{m}\tilde{z}_i(k)^T(x_i(k+1)-\bar{v}(k+1))\\&\ge -2c(k)L\sum_{i=1}^{m}\|x_i(k+1)-\bar{v}(k+1)\|,\end{aligned} \quad (9)$$

by Cauchy-Schwartz, triangle inequality, and $L = \max_{\xi\in\cup_{i=1}^m X_j}\|g_j(\xi)\|$, which is well-defined by Lemma 1. Second, we use the definition of $\tilde{z}_i(k)$ – Step 3 in Algorithm 1 – into the second term of the right-hand side of (8) to obtain (via double stochasticity of A)

$$\begin{aligned}&2c(k)\sum_{i=1}^{m}\tilde{z}_i(k)^T(\bar{v}(k+1)-x^\star)\\&=2c(k)\sum_{i=1}^{m}g_i(z_i(k))^T(\bar{v}(k+1)-x^\star).\end{aligned} \quad (10)$$

Moreover, if we add and subtract $x_i(k+1)$ and $z_i(k)$ for all $i = 1,\dots,m$ into the right-hand side of (10) we obtain

$$\begin{aligned}&2c(k)\sum_{i=1}^{m}g_i(z_i(k))^T(\bar{v}(k+1)-x^\star)\\&=2c(k)\sum_{i=1}^{m}g_i(z_i(k))^T(\bar{v}(k+1)-x_i(k+1))\\&+2c(k)\sum_{i=1}^{m}g_i(z_i(k))^T(x_i(k+1)-z_i(k))\\&+2c(k)\sum_{i=1}^{m}g_i(z_i(k))^T(z_i(k)-x^\star).\end{aligned} \quad (11)$$

Let us focus on the right-hand side of (11). The left-most term can be lower-bounded as

$$\begin{aligned}&2c(k)\sum_{i=1}^{m}g_i(z_i(k))^T(\bar{v}(k+1)-x_i(k+1))\\&\ge -2c(k)L\sum_{i=1}^{m}\|\bar{v}(k+1))-x_i(k+1)\|,\end{aligned} \quad (12)$$

by Cauchy-Schwartz. As for the middle term, we have that

$$
\begin{aligned}
&2c(k)\sum_{i=1}^{m} g_i(z_i(k))^T(x_i(k+1)-z_i(k)) \\
&\geq -2c(k)L\sum_{i=1}^{m}\|e_i(k+1)\| \\
&\geq -\alpha_1(k)\sum_{i=1}^{m}\|e_i(k+1)\|^2 - m\frac{L^2}{\alpha_1(k)}c(k)^2 \qquad (13)
\end{aligned}
$$

where the first inequality follows from Cauchy-Schwartz and the definition $e_i(k)$ given in (6). For the second inequality, we applied the relation $2xy \leq x^2+y^2$ with $x = \frac{L}{\sqrt{\alpha_1(k)}}c(k)$ and $y = \sqrt{\alpha_1(k)}\|e_i(k+1)\|$ for some $\alpha_1(k) \in (0,1)$ for all $k \in \mathbb{N}$.

Similarly, the right-most term of (11) can be manipulated to yield

$$
\begin{aligned}
&2c(k)\sum_{i=1}^{m} g_i(z_i(k))^T(z_i(k)-x^\star) \geq -\alpha_2(k)\sum_{i=1}^{m}\|e_i(k+1)\|^2 \\
&- m\frac{L^2}{\alpha_2(k)}c(k)^2 - 2c(k)L\sum_{i=1}^{m}\|x_i(k+1)-\bar{v}(k+1)\| \\
&+ 2c(k)\sum_{i=1}^{m}\Big(f_i(\bar{v}(k+1)) - f_i(x^\star)\Big). \qquad (14)
\end{aligned}
$$

for some sequence $\alpha_2(k) \in (0,1)$ for all $k \in \mathbb{N}$. The details that led to inequality (14) resemble the ones in (13), and are omitted for brevity. See [13] for details. Substituting inequalities (9), (12), (13) and (14) into (7) we obtain

$$
\begin{aligned}
&2c(k)\sum_{i=1}^{m}(f_i(\bar{v}(k+1)) - f_i(x^\star)) + \sum_{i=1}^{m}\|x_i(k+1)-x^\star\|^2 \\
&+\Big(1-\alpha_1(k)-\alpha_2(k)\Big)\sum_{i=1}^{m}\|e_i(k+1)\|_2^2 \\
&\leq \sum_{i=1}^{m}\|x_i(k)-x^\star\|^2 + mL^2\Big(\frac{\alpha_1(k)+\alpha_2(k)}{\alpha_1(k)\alpha_2(k)}\Big)c(k)^2 \\
&\quad + 6c(k)L\sum_{i=1}^{m}\|x_i(k+1)-\bar{v}(k+1)\|. \qquad (15)
\end{aligned}
$$

We use the result of Lemma 2 iii), with $\bar{L} = 3L$. Note that for any $\beta_1 \in (0,1)$, the sequences $(\alpha_1(k))_{k\in\mathbb{N}}$ and $(\alpha_2(k))_{k\in\mathbb{N}}$ can be chosen to guarantee that $1-\alpha_1(k)-\alpha_2(k)-\beta_1 \geq 0$ for all $k \in \mathbb{N}$. For instance, one particular choice is $\alpha_1(k) = \alpha_2(k) = \alpha$ with $1-\beta_1-2\alpha > 0$. The result follows by summing up inequality (15) from $k=0$ to $k=N$. This concludes the proof of ii), thus also concluding the proof of the lemma. ■

Two immediate consequences of Lemma 3 are presented in the following proposition.

Proposition 1: Consider the result of Lemma 3 ii), and suppose Assumptions 1–3 hold. Then we obtain that

i) The error sequence $(e_i(k))_{k\in\mathbb{N}}$ converges to zero for all $i = 1,\dots,m$.

ii) Consensus is asymptotically achieved, that is, for all $i = 1,\dots,m$,

$$\lim_{k\to\infty}\|x_i(k) - v(k)\| = 0,$$

where $v(k)$ is defined in (4).

Proof: To prove i), consider the inequality of item ii) in Lemma 3. Let $\beta_1 \in (0,1)$, choose $\alpha_1(k) = \alpha_2(k) = \alpha$ so that $1-2\alpha-\beta_1 > 0$. Then the proof follows the arguments presented in [6, Proposition 2], and is omitted for brevity. We also omitted the proof of ii) as it also follows from the arguments presented in [6, Proposition 3]. This concludes the proof of the proposition. ■

We are now in a position to prove Theorem 1. To this end, we use the inequality of item ii) in Lemma 3 and leverage on a deterministic version of the supermartingale theorem in order to establish convergence of the sequences $(\|x_i(k) - x^\star\|)_{k\in\mathbb{N}}$, $i = 1,\dots,m$, to zero for some minimiser $x^\star$ of the centralised problem. Let us first present the supermartingale result.

Lemma 4: Given non-negative scalar sequences $(\ell(k))_{k\in\mathbb{N}}$, $(u(k))_{k\in\mathbb{N}}$ and $(\zeta(k))_{k\in\mathbb{N}}$ that obey the recursion

$$\ell(k+1) \leq \ell(k) - u(k) + \zeta(k).$$

If $\sum_{k=1}^{\infty}\zeta(k) < \infty$, then the sequence $(\ell(k))_{k\in\mathbb{N}}$ converges and the sequence $(u(k))_{k\in\mathbb{N}}$ is summable.

Proof of Theorem 1

Consider inequality of item ii) in Lemma 3, and choose $\alpha_1(k), \alpha_2(k)$ and β_1 as in the proof of Proposition 1. With reference to Lemma 4 set[3]

$$\ell(k) = \sum_{i=1}^{m}\|x_i(k)-x^\star\|^2, \quad \zeta(k) = \bar{L}c(k)^2,$$

$$u(k) = 2c(k)\big(f(\bar{v}(k+1)) - f(x^\star)\big),$$

where $\bar{L} = \left(\frac{2mL^2}{\alpha} + \beta_2\right) > 0$. As a consequence, we have that the sequence $(\sum_i\|x_i(k+1)-x^\star\|)_{k\in\mathbb{N}}$ converges and that

$$\sum_{k=1}^{\infty} c(k)\big(f(\bar{v}(k+1)) - f(x^\star)\big) < \infty.$$

The latter result gives us that $\liminf_{k\to\infty}(f(\bar{v}(k+1)) - f(x^\star)) = 0$. Therefore, there exists a subsequence of $(f(\bar{v}(k+1)) - f(x^\star))_{k\in\mathbb{N}}$ that converges to zero. Since the function $f(x)$ is continuous (by convexity) for all $i = 1,\dots,m$, there exists some minimizer $x^\star$ such that a subsequence of $(\|\bar{v}(k)-x^\star\|)_{k\in\mathbb{N}}$ converges to zero. Hence, for each $i = 1,\dots,m$, we obtain

$$
\begin{aligned}
\|x_i(k)-x^\star\| &\leq \|\bar{v}(k)-x^\star\| + \|x_i(k)-\bar{v}(k)\| \\
&\leq \|\bar{v}(k)-x^\star\| + \mu\sum_{i=1}^{m}\|x_i(k)-v(k)\|.
\end{aligned}
$$

by triangle inequality. Moreover, we can find a subsequence of $(\|x_i(k)-x^\star\|)_{k\in\mathbb{N}}$ that converges to zero because of Proposition 1 ii) applied to the term involving $\|x_i(k) - v(k)\|$. As we already know that the sequence $(\|x_i(k) - x^\star\|)_{k\in\mathbb{N}}$ converges, the proof is concluded as every Cauchy sequence has the same limit point. ■

[3]Note that the assumptions of Lemma 4 are satisfied because the sequence of step-sizes is square-summable under Assumption 3.

B. Proofs of Section III-B

Throughout this proof, suppose Assumption 4 holds. We drop the constant η for simplicity of exposition, but general choices $\frac{\eta}{\sqrt{k+1}}$, $\eta > 0$, are also applicable. Let $\hat{v}(k)$ be similarly defined as $\hat{x}_i(k)$ in Theorem 2 from the sequence $(\bar{v}(k))_{k\in\mathbb{N}}$, and consider the relation

$$\sum_{i=1}^{m} f_i(\hat{x}_i(k+1)) - f(x^\star) \le f(\hat{v}(k+1)) - f(x^\star) + L\sum_{i=1}^{m} \|\hat{x}_i(k+1) - \hat{v}(k+1)\|, \quad (16)$$

which follows from Lemma 1 $iii)$. In order to simplify our presentation, we change the notation in Lemma 3 $ii)$ by replacing k by r, and N by k. Besides, we split the proof into two parts: we first consider that (17) and (18) are satisfied for positive constants d_1, d_2, d_3 and d_4, and prove the statement of Theorem 2. Then we return to (17) and (18), and prove the existence of such constants. To this end, consider

$$f(\hat{v}(k+1)) - f(x^\star) \le d_1 \frac{1}{S(k+1)} + d_2 \frac{\sum_{r=0}^{k} c(r)^2}{S(k+1)} \quad (17)$$

$$L\sum_{i=1}^{m} \|\hat{x}_i(k+1) - \hat{v}(k+1)\| \le d_3 \frac{1}{S(k+1)} + d_4 \frac{\sum_{r=0}^{k} c(r)^2}{S(k+1)}. \quad (18)$$

Notice that $S(k+1)$ can be lower-bounded as

$$S(k+1) = \sum_{r=1}^{k+1} \frac{1}{\sqrt{r+1}} \ge \int_2^{k+3} \frac{1}{\sqrt{x}} dx = 2(\sqrt{k+3} - \sqrt{2}) \ge \nu\sqrt{k+3} \ge \nu\sqrt{k+1}, \quad (19)$$

where $\nu = 2 - \sqrt{2} = 0.5858$. Besides, we have that

$$\sum_{r=0}^{k} c(r)^2 = \sum_{r=0}^{k} \frac{1}{r+1} = \sum_{r=2}^{k+1} \frac{1}{r} + 1 \le \int_1^{k+1} \frac{1}{x} dx + 1 \le \ln(k+1) + 1. \quad (20)$$

The result of the Theorem 2 would then follow by substituting (19) and (20) into (17) and (18) with constants $B_1 = \sum_{i=1}^{4} \frac{d_i}{\nu}$ and $B_2 = \frac{d_2}{\nu} + \frac{d_4}{\nu}$. This concludes the proof of Theorem 2, provided inequalities (17) and (18) hold.

We now prove these inequalities. Let us start with (17). Indeed, in light of Lemma 3 $ii)$, given any $\beta_1 \in (0,1)$, a valid choice for the sequences $(\alpha_1(k))_{k\in\mathbb{N}}$ and $(\alpha_2(k))_{k\in\mathbb{N}}$ is $\alpha_1(k) = \alpha_2(k) = \alpha(k)$, where $\alpha(k) = a\Big(1 - \frac{1}{\sqrt{k+1}}\Big)$, with $a = (1-\beta_1)/2$. Under these choices, note that

$$1 - \beta_1 - 2\alpha(k) = \frac{1-\beta_1}{\sqrt{k+1}} = (1-\beta_1)c(k). \quad (21)$$

Consider now inequality $ii)$ of Lemma 3 with the above choices for $\alpha_1(k)$ and $\alpha_2(k)$. Note that the series $\sum_{r=0}^{k}\sum_{i=1}^{m} \|x_i(r+1) - x^\star\|$ and $\sum_{r=0}^{k}\sum_{i=1}^{m} \|x_i(r) - x^\star\|$ are telescopic so all the middle terms cancel. We now drop the terms involving $\|e_i(r+1)\|^2$ and $\|x_i(k+1) - x^\star\|$, and then divide the resulting expression by $2S(k+1) = 2\sum_{r=1}^{k+1} \frac{1}{\sqrt{r+1}}$ to obtain the following upper bound on the left-hand side of (17)

$$\sum_{r=0}^{k} \frac{c(r+1)}{S(k+1)} \sum_{i=1}^{m} (f_i(\bar{v}(r+1)) - f_i(x^\star)) \le \frac{\sum_{i=1}^{m} \|x_i(0) - x^\star\|^2}{S(k+1)} + \frac{\beta_3}{S(k+1)} + \beta_2 \sum_{r=0}^{k} \frac{c(r)^2}{S(k+1)} + 2mL^2 \frac{1}{S(k+1)} \sum_{r=0}^{k} \frac{c(r)^2}{\alpha(r)}. \quad (22)$$

The left-hand side of (22) is an upper bound on that of (17) due to convexity of f_i and the definition of $\hat{v}(k+1)$. Besides, note that in the right-hand side of (22) we have $c(r)$ rather than $c(r+1)$ due to the fact that $c(r+1) \le c(r)$. By the right-hand side of (22), we obtain inequality (17) with

$$d_1 = 4mD^2 + \beta_3, \quad d_2 = \beta_2 + \frac{8mL^2}{a}.$$

where we have used Assumption 1 so that $\sum_{i=1}^{m} \|x_i(0) - x^\star\|^2 \le 4mD^2$, where D is defined as in Lemma 2 $i)$, and the fact that

$$\frac{c(r)^2}{\alpha(r)} = \frac{1}{a} \frac{\sqrt{r+1}}{\sqrt{r+1} - 1} \frac{1}{r+1} \le \frac{4}{a} c(r)^2$$

by monotonicity of $\frac{\sqrt{x+1}}{\sqrt{x+1}-1}$.

As for (18), note that its left-hand side can be upper-bounded by

$$\frac{L\mu}{S(k+1)} \sum_{r=1}^{k+1} c(r) \sum_{i=1}^{m} \|x_i(r) - v(r)\| = \frac{L\mu c(1)}{S(k+1)} \sum_{i=1}^{m} \|x_i(1) - v(1)\| + \frac{L\mu}{S(k+1)} \sum_{r=2}^{k+1} c(r) \sum_{i=1}^{m} \|x_i(r) - v(r)\|, \quad (23)$$

which follows from the definition of $\hat{v}(k)$, triangle inequality, and Lemma 2 $i)$. We now invoke Lemma 2 $ii)$ – with r in the place of k, and t in the place of r – for the last term on the right-hand side of (23) so that

$$\begin{aligned}
&\sum_{r=2}^{k+1} c(r) \sum_{i=1}^{m} \|x_i(r) - v(r)\| \\
&= \sum_{r=1}^{k} c(r+1) \sum_{i=1}^{m} \|x_i(r+1) - v(r+1)\| \\
&\le 2\sum_{r=0}^{k} c(r) \sum_{i=1}^{m} \|e_i(r+1)\| + m\lambda \sum_{i=1}^{m} \|x_i(0)\| \sum_{r=0}^{k} c(r) q^r \\
&\quad + m\lambda \sum_{r=1}^{k} c(r+1) \sum_{t=0}^{r-1} q^{r-t-1} \sum_{i=1}^{m} \|e_i(t+1)\|, \quad (24)
\end{aligned}$$

where we added the term corresponding to $r = 0$ in first two terms on the right-hand side of (24), and used the fact that $c(r+1) \leq c(r)$ for all $r \in \mathbb{N}$. By using properties of geometric series and series convolution, we can manipulate the last term on the right-hand side of (24) to obtain

$$\leq \sum_{i=1}^{m} \|x_i(0)\| \frac{m\lambda c(1)}{1-q} + 2\left(1 + \frac{m\lambda}{2(1-q)}\right) \sum_{r=0}^{k} c(r) \sum_{i=1}^{m} \|e_i(r+1)\|. \quad (25)$$

We now consider the last term in inequality (25) and upper-bound it as

$$2\sum_{r=0}^{k} c(r) \sum_{i=1}^{m} \|e_i(r+1)\| \leq \sum_{r=0}^{k} c(r)^2 + \sum_{r=0}^{k} \sum_{i=1}^{m} \|e_i(r+1)\|^2, \quad (26)$$

by the inequality $2xy \leq x^2 + y^2$, with $x = c(r)$ and $y = \|e_i(r+1)\|$ for all $i = 1, \ldots, m$. To finalise the argument, we need to provide bounds on the term involving $\|e_i(r+1)\|^2$ in inequality (26). To do so, we invoke Lemma 3 $ii)$ once more with the same β_1 as in (21), but for different sequences $(\alpha_1(k))_{k\in\mathbb{N}}$ and $(\alpha_2(k))_{k\in\mathbb{N}}$. In fact, choose $\alpha_1(k) = \alpha_2(k) = \alpha$ in the same manner as in the Proposition 1 to show that the following inequality holds

$$\sum_{r=0}^{k} \sum_{i=1}^{m} \|e_i(r+1)\|^2 \leq \frac{\sum_{i=1}^{m} \|x(0) - x^\star\|^2 + \beta_3}{1 - \beta_1 - 2\alpha} + \frac{1}{1-\beta_1-2\alpha}\Big(mL^2 \frac{2}{\alpha} + \beta_2\Big) \sum_{r=0}^{k} c(r)^2, \quad (27)$$

this proves inequality (18) with

$$d_3 = 2mDL\mu c(1) + L\mu \sum_{i=1}^{m} \|x_i(0)\| \frac{m\lambda c(1)}{1-q} + L\mu\left(1 + \frac{m\lambda}{2(1-q)}\right) \frac{4mD^2 + \beta_3}{1-\beta_1-2\alpha},$$

$$d_4 = L\mu\left(1 + \frac{m\lambda}{2(1-q)}\right)\left(1 + \frac{1}{1-\beta_1-2\alpha}\Big(mL^2\frac{2}{\alpha} + \beta_2\Big)\right).$$

This concludes the proof of inequality (18), thus concluding the proof of Theorem 2. ■

REFERENCES

[1] M. G. Rabbat and R. Nowak, "Distributed Optimization in Sensor Networks," in *Inf. Process. Sens. Networks*, 2004, pp. 20–27.

[2] A. Nedić and A. Ozdaglar, "Distributed Subgradient Methods for Multi-Agent Optimization," *IEEE Trans. Autom. Control*, vol. 54, no. 1, pp. 48–61, 2009.

[3] A. Nedic, A. Ozdaglar, and P. A. Parrilo, "Constrained Consensus and Optimization in Multi-Agent Networks," *IEEE Trans. Automat. Contr.*, vol. 55, no. 4, pp. 922–938, 2010.

[4] A. Nedić and A. Olshevsky, "Distributed optimization over time-varying directed graphs," *IEEE Trans. Automat. Contr.*, vol. 60, no. 3, pp. 601–615, 2015.

[5] D. Kempe, A. Dobra, and J. Gehrke, "Gossip-Based Computation of Aggregate Information David," in *IEEE Symp. Found. Comput. Sci.* IEEE Computer Society, 2003, p. 482.

[6] K. Margellos, A. Falsone, S. Garatti, and M. Prandini, "Distributed Constrained Optimization and Consensus in Uncertain Networks via Proximal Minimization," *IEEE Trans. Automat. Contr.*, vol. 63, no. 5, pp. 1372–1387, 2018.

[7] W. Shi, Q. Ling, G. Wu, and W. Yin, "EXTRA: An Exact First-Order Algorithm for Decentralized Consensus Optimization," *SIAM J. Optim.*, vol. 25, no. 2, pp. 944–966, 2015.

[8] A. Nedić, A. Olshevsky, and W. Shi, "Achieving Geometric Convergence for Distributed Optimization over Time-Varying Graphs," *SIAM J. Optim.*, vol. 55, no. 2, pp. 664–690, 2017.

[9] G. Qu and N. Li, "Harnessing smoothness to accelerate distributed optimization," *IEEE Trans. Control Netw. Syst.*, vol. 5, no. 3, pp. 1245–1260, 2018.

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

[11] K. I. Tsianos and M. G. Rabbat, "Distributed strongly convex optimization," in *50th Annu. Allert. Conf. Commun. Control. Comput.*, 2012, pp. 593–600.

[12] K. I. Tsianos, S. Lawlor, and M. G. Rabbat, "Push-Sum Distributed Dual Averaging for convex optimization," *2012 IEEE 51st IEEE Conf. Decis. Control*, pp. 5453–5458, 2012. [Online]. Available: http://ieeexplore.ieee.org/document/6426375/

[13] L. Romao, K. Margellos, G. Notarstefano, and A. Papachristodoulou, "Convergence rate analysis of a subgradient averaging algorithm for distribtued optimisation with different constraint sets," 2019, pp. 1—-12. [Online]. Available: http://sysos.eng.ox.ac.uk/wiki/index.php/User:Licio

[14] D. P. Bertsekas, *Convex Optimization Theory.* Athena Scientific, 2009.

[15] S. Liang, L. Wang, and G. Yin, "Distributed quasi-monotone subgradient algorithm for nonsmooth convex optimization over directed graphs," *Automatica*, vol. 101, pp. 175–181, 2019.

[16] Y. Nesterov and V. Shikhman, "Quasi-monotone Subgradient Methods for Nonsmooth Convex Minimization," *J. Optim. Theory Appl.*, vol. 165, no. 3, pp. 917–940, 2015.