Automatica 131 (2021) 109738

Contents lists available at ScienceDirect

Automatica

journal homepage: www.elsevier.com/locate/automatica

Subgradient averaging for multi-agent optimisation with different constraint sets☆

Licio Romao [a,*], Kostas Margellos [a], Giuseppe Notarstefano [b], Antonis Papachristodoulou [a]

[a] Department of Engineering Science, University of Oxford, Parks Road, Oxford OX1 3PJ, UK
[b] Department of Electrical, Electronic, and Information Engineering G. Marconi at Alma Mater Studiorum Università di Bologna, Italy

ARTICLE INFO

Article history:
Received 28 October 2019
Received in revised form 4 February 2021
Accepted 13 April 2021
Available online 2 June 2021

Keywords:
Distributed optimisation
Multi-agent networks
Parallel algorithms
Subgradient methods
Consensus

ABSTRACT

We consider a multi-agent setting with agents exchanging information over a possibly time-varying network, aiming at minimising a separable objective function subject to constraints. To achieve this objective we propose a novel subgradient averaging algorithm that allows for non-differentiable objective functions and different constraint sets per agent. Allowing different constraints per agent simultaneously with a time-varying communication network constitutes a distinctive feature of our approach, extending existing results on distributed subgradient methods. To highlight the necessity of dealing with a different constraint set within a distributed optimisation context, we analyse a problem instance where an existing algorithm does not exhibit a convergent behaviour if adapted to account for different constraint sets. For our proposed iterative scheme we show asymptotic convergence of the iterates to a minimum of the underlying optimisation problem for step sizes of the form $\frac{\eta}{k+1}$, $\eta > 0$. We also analyse this scheme under a step size choice of $\frac{\eta}{\sqrt{k+1}}$, $\eta > 0$, and establish a convergence rate of $\mathcal{O}(\frac{\ln k}{\sqrt{k}})$ in objective value. To demonstrate the efficacy of the proposed method, we investigate a robust regression problem and an ℓ_2 regression problem with regularisation.

1. Introduction

Distributed optimisation deals with multiple agents interacting over a network and has found numerous applications in different domains, such as wireless sensor networks (Baingana, Mateos, & Giannakis, 2014; Mateos & Giannakis, 2012), robotics (Martinez, Bullo, Cortes, & Frazzoli, 2007), and power systems (Bolognani, Carli, Cavraro, & Zampieri, 2015), due to its ability to parallelise computation and prevent agents from sharing information considered as private. Typically, distributed algorithms are based on an iterative process in which agents maintain some estimate about the decision vector in an optimisation context, exchange this information with neighbouring agents according to an underlying communication protocol/network, and update their estimate on the basis of the received information.

Despite the intense research activity in this area, only a few algorithms can simultaneously deal with time-varying networks, non-differentiable objective functions and account for the presence of constraints (Liang, Wang, & Yin, 2019; Margellos, Falsone, Garatti, & Prandini, 2018; Nedić & Olshevsky, 2015; Xi & Khan, 2017; Zhu & Martinez, 2012), features that are often treated separately in the literature. Several of the commonly employed methods are based on a projected subgradient or a proximal step and their analysis consists of selecting the step size underlying these algorithms, establishing a convergence rate analysis, and quantifying practical convergence for (near-)real time applications.

In this paper, we study a class of optimisation problems that involves a separable objective function, while the feasible set can be decomposed as an intersection of different compact convex sets. A centralised version of this class of problems has been studied under a stochastic setting in Bianchi (2016) and Patrascu and Necoara (2018). Distributed algorithms for this class have been proposed in Johansson, Keviczky, Johansson, and Johansson (2008), Lee and Nedić (2013), Lin, Ren, and Song (2016), Mai and Abed (2019), Margellos et al. (2018), Nedic and Ozdaglar (2009), Nedic, Ozdaglar, and Parrilo (2010) and Zhu and Martinez (2012). References Johansson et al. (2008), Nedic and Ozdaglar (2009) and Nedic et al. (2010) rely on Bertsekas and Tsitsiklis (1989) and Tsitsiklis, Bertsekas, and Athans (1986) to propose a distributed strategy based on projected sub-gradient methods. These results consist of an averaging step followed by a local

☆ The material in this paper was presented at the 58th IEEE Conference on Decision and Control, December 11–13, 2019, Nice, France. This paper was recommended for publication in revised form by Associate Editor Julien M. Hendrickx under the direction of Editor Christos G. Cassandras.

* Corresponding author.
E-mail addresses: licio.romao@eng.ox.ac.uk (L. Romao), kostas.margellos@eng.ox.ac.uk (K. Margellos), giuseppe.notarstefano@unibo.it (G. Notarstefano), antonis@eng.ox.ac.uk (A. Papachristodoulou).

https://doi.org/10.1016/j.automatica.2021.109738
0005-1098/

sub-gradient projection update. In Margellos et al. (2018) a distributed scheme based on a proximal update is proposed, thus extending Johansson et al. (2008) and Nedic et al. (2010) to the case where different local constraint sets and an arbitrarily time-varying network are considered. The authors in Zhu and Martinez (2012) provide asymptotic convergence for a primal–dual algorithm that allows coupling between agents' local estimates. We discuss additional related results in Section 4, after the proposed algorithm is presented and some notation introduced.

We motivate our approach by constructing an example showing that extending available algorithms to the case of different constraint sets might not exhibit a convergent behaviour for all problem instances. Hence, a direct adaptation of existing schemes is not always possible when dealing with different constraint sets. Notice also that distributed algorithms developed for the unconstrained case cannot be trivially adapted to our setting, as lifting the constraints in the objective (e.g., via characteristic functions) would violate boundedness of the subgradient, a typical requirement for such algorithms (Duchi, Agarwal, & Wainwright, 2012; Margellos et al., 2018; Nedić & Olshevsky, 2015; Nedic et al., 2010).

The main contribution of this paper is the introduction and the characterisation of the convergence rate for a new subgradient averaging algorithm. The proposed scheme allows us to account for time-varying networks, non-differentiable objective functions and different constraint sets per agent as in Margellos et al. (2018), while achieving faster practical convergence as it is based on subgradient averaging as in Duchi et al. (2012), Johansson et al. (2008) and Mai and Abed (2019). Note that allowing simultaneously for different constraint sets per agent and time-varying communication network by means of a subgradient averaging scheme is a distinct feature of the algorithm in this paper. Preliminary results related to this paper appeared in Romao, Margellos, Notarstefano, and Papachristodoulou (2019), where several proofs have been omitted. Moreover, the construction of Section 2.2 that motivates the analysis of algorithms with different constraint sets is novel, and offers insight on the limitations of existing algorithms. We also provide detailed numerical examples, not included in the conference version.

The paper is organised as follows. In Section 2 we present the problem statement, the network communication structure, and the main assumptions adopted in this paper, followed by a numerical construction that motivates the algorithm of this paper. In Section 3 we present the proposed scheme and the main convergence results, namely, asymptotic convergence in iterates and a convergence rate as far as the optimal value is concerned. Section 4 provides detailed discussion and comparison of our scheme with other results in the literature. In Section 5 we study the robust linear regression problem and ℓ_2 regression with regularisation to demonstrate the main algorithmic features of our scheme and to compare our strategy against existing methods. Finally, some concluding remarks and future research directions are provided in Section 6. To ease the reader all proofs have been deferred to the Appendix).

Notation: We denote by $\mathbb{R}$ the set of real numbers and $\mathbb{N}$ the set of natural numbers (excluding zero). The symbol $\mathbb{R}^n$ stands for the Cartesian product $\mathbb{R}\times\cdots\times\mathbb{R}$ with n terms. A sequence of elements in $\mathbb{R}^n$ is denoted by $(x(k))_{k\in\mathbb{N}}$. For any set $X\subset\mathbb{R}^n$, we denote its interior, relative interior and convex hull by $\text{int}(X)$, $\text{ri}(X)$, and $\text{conv}(X)$, respectively. We also denote by $f(X)$ as the image of the set X over a function f. The subdifferential of f at a point $x\in\text{dom}f$ is denoted by $\partial f(x)$. For any point $x\in\mathbb{R}^n$, $\|x\|_2$ stands for the Euclidean norm of x and $\|x\|_1$ for the ℓ_1 norm of $x\in\mathbb{R}^n$, which are reduced to $|x|$ if x is scalar.

2. Problem statement and a motivating example

2.1. Problem set-up and network communication

Consider the optimisation problem

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

where $x\in\mathbb{R}^n$ is the vector of decision variables, and $f_i:\mathbb{R}^n\to\mathbb{R}$ and $X_i\subset\mathbb{R}^n$ constitute the local objective function and constraint set, respectively, for agent i, $i=1,\dots,m$. We suppose that each agent i possesses as private information the pair (f_i,X_i) and maintains a local estimate x_i of the common decision vector x.

The goal is for all agents to agree on the local variables, that is, $x_i=x^\star$, for all $i=1,\dots,m$, where $x^\star$ is an optimiser of (1), i.e., a feasible point such that $f(x^\star)\le f(x)$ for all $x\in\cap_{i=1}^{m}X_i$. We impose the following assumption.

Assumption 1. We assume that:

(*i*) For all $i=1,\dots,m$, the function f_i is convex.
(*ii*) The set $X_i\subset\mathbb{R}^n$ is compact and convex for all $i=1,\dots,m$, and $\cap_{i=1}^{m}X_i$ has a non-empty interior.
(*iii*) The subgradient of the function $f(x)$ is bounded on $\cup_{i=1}^{m}X_i$, that is, $L=\max_{\substack{\xi\in\partial f(x),\\ x\in\cup_{i=1}^{m}X_j}}\|\xi\|_2<\infty$.

Assumption 1 imposes standard restriction for constrained non-smooth optimisation. Item (*ii*) implies informally that $\cup_{i=1}^{m}X_i$ has volume in $\mathbb{R}^n$, i.e., that the affine hull of $\cup_{i=1}X_i$ has dimension n. Moreover, the compactness assumption of item (*ii*) guarantees that the optimal set of problem (1) is non-empty. Item (*iii*) is an assumption that is needed to prove convergence of sub-gradient methods applied to problem (1). Under item (*iii*), the sub-gradient of the function f can be evaluated at points that belong to $\cup_{i=1}^{m}X_i$. We provide in Appendix A.2 a technical condition on the domain of the functions f_i that is sufficient to guarantee that Assumption 1, item (*iii*), holds. An important consequence of Assumption 1 is given in the following lemma.

Lemma 1. *Under Assumption 1, we have that:*

(*i*) *The set* $\text{conv}(\cup_{i=1}^{m}X_i)$ *is compact.*
(*ii*) *The function f is Lipschitz continuous over $\cap_{i=1}^{m}X_i$, i.e., the following inequality hods*

$$|f(x)-f(y)|\le L\|x-y\|_2,\quad\forall\,x,y\in\cap_{i=1}^{m}X_i,$$

where L is the constant defined in Assumption 1.

Typical choices of functions that satisfy Assumption 1 are piecewise-linear functions, quadratic convex functions and the logistic regression function.

In this paper, we aim to solve problem (1) through a network of agents that use only the available local information, namely, the pair (f_i,X_i) and the current estimate for the optimal solution, $x_i(k)$, $i=1,\dots,m$, maintained by agent i at a given instance k. We will show how $x_i(k)$, $i=1,\dots,m$, can be constructed and updated in Section 3, with k playing the role of iteration index. To this end, we now characterise the underlying communication network. Let $\mathcal{G}(k)=(\mathcal{N},\mathcal{E}(k))$ be an undirected graph, where $\mathcal{N}=\{1,\dots,m\}$ 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, only 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 $[A(k)]_j^i>0$ only if $(j,i)\in\mathcal{E}(k)$ at time k. As the graph is undirected, the 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 often. We impose the following assumption on the matrix $A(k)$.

Assumption 2. We assume that:

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

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

(*iii*) Matrix $A(k)$ is doubly stochastic.

These are standard requirements in the distributed optimisation literature. We refer the reader to Duchi et al. (2012), Margellos et al. (2018), Nedić and Ozdaglar (2009) and Nedic et al. (2010) for more details.

2.2. Dealing with different constraint sets

In this section, we highlight the necessity of developing a new algorithmic scheme to deal with the presence of a different constraint sets per agent. To this end, consider the iterative scheme[1]

$$z_i(k+1) = \sum_{j=1}^{m} [A]_j^i z_j(k) + g_i(k) \tag{2a}$$

$$x_i(k+1) = \operatorname*{argmin}_{\xi \in X_i} z_i(k+1)^T \xi + \frac{1}{c(k)} \|\xi\|_2^2, \tag{2b}$$

which consists of a modified version of the algorithm considered in Duchi et al. (2012), adapted to account for different constraint sets in each agent's local optimisation problem. In the setting of the previous section, notice that matrix A in (2a) corresponds to a time-invariant network $\mathcal{G}(k) = (\mathcal{N}, \mathcal{E})$, for all $k \in \mathbb{N}$. Assumption 2 is satisfied if the graph $(\mathcal{N}, \mathcal{E})$ is connected and matrix A is doubly-stochastic.

Observe that (2a) constitutes a subgradient update step, with neighbouring local variables $z_j(k)$ being "mixed" according to the matrix A and added to $g_i(k) \in \partial f_i(x_i(k))$, i.e., a subgradient of f_i evaluated at $x_i(k)$, $i = 1, \ldots, m$. Step (2b) is an optimisation programme with the objective function being the sum (weighted via $c(k)$) of

$z_i(k+1)^T \xi$: linear "proxy" of f_i,

and a regularisation term $\|\xi\|_2^2$. To comply with Duchi et al. (2012), we set $c(k) = \frac{1}{\sqrt{k+1}}$. Recall that the algorithm in Duchi et al. (2012) involves the same constraint set in the update rule of (2b), that is $X_i = X$ for all $i = 1, \ldots, m$, and possesses a guaranteed convergence rate of $\mathcal{O}(\frac{\ln k}{\sqrt{k}})$ for the running averages of the iterates $x_i(k)$; here, we introduce a different set X_i per agent and show that this (natural) modification may lead to erroneous results.

Consider a two-agent instance of (1), i.e., $m = 2$ with $x \in \mathbb{R}^2$, $f_i = x^T Q x + q_i^T x + r_i$, for $i = 1, 2$ and

$$Q = \begin{bmatrix} 1.2 & 0.4 \\ 0.4 & 1.8 \end{bmatrix}, \quad q_1 = \begin{bmatrix} 8 \\ -4 \end{bmatrix}, \quad q_2 = \begin{bmatrix} 2.93 \\ -11.46 \end{bmatrix},$$
$$r_1 = 20, \quad r_2 = 25. \tag{3}$$

The local constraint sets are given by $X_1 = [-1, 1] \times [-1, 1]$ and $X_2 = [0.5, 2.5] \times [0.5, 2.5]$. The feasible set $X_1 \cap X_2$ is the box $[0.5, 1] \times [0.5, 1]$. Fig. 1 depicts the level curves of the quadratic functions $f_1(x)$ (dashed-red lines), f_2 (double-dashed lines), and $f = f_1 + f_2$ (solid-black lines). The red and blue boxes represent the sets X_1 and X_2 respectively, with the feasible set, $X_1 \cap X_2$, being also indicated in the figure in black.

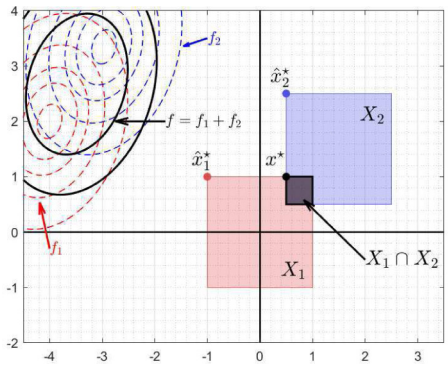

Fig. 1. Geometric representation of problem instance encoded by (3). The red ellipsoids (dashed lines) correspond to the level curves of f_1, the blue ellipsoids (double-dashed lines) represent the function f_2, while the black (solid lines) ellipsoids correspond to the ones of $f = f_1 + f_2$. The shaded red box illustrates the constraint set X_1, while the shaded blue box illustrates X_2. Vectors $\hat{x}_1^\star = [-1, 1]^T$ and $\hat{x}_2^\star = [0.5, 2.5]^T$ are the optimal solutions of $f_1(x)$ and $f_2(x)$ under the constraints X_1 and X_2, respectively. The global optimal solution of $f = f_1 + f_2$ with matrices given by (3) subject to $x \in X_1 \cap X_2$ is denoted by $x^\star$. This construction shows that $\hat{x}_1^\star$ and $\hat{x}_2^\star$ constitute fixed-points of (2) thus preventing the iteration from reaching $x^\star$ if initialised at those points.

By inspection the optimal solution of f_1 under the constraint $x \in X_1$ is $\hat{x}_1^\star = [-1, 1]^T$. Similarly, the optimal solution for f_2 under $x \in X_2$ is $\hat{x}_2^\star = [0.5, 2.5]^T$. We then have the following proposition.

Proposition 1. *Let $(z_i(k))_{k\in\mathbb{N}}$, $(x_i(k))_{k\in\mathbb{N}}$, $i = 1, 2$, be the sequences generated by algorithm (2) when applied to problem (3) with initial conditions $x_i(0) = \hat{x}_i^\star$, $i = 1, 2$, and with $A = \frac{1}{2}\mathbf{1}\mathbf{1}^T$ and $c(k) = \frac{1}{\sqrt{k+1}}$. We have that*

$$x_1(k) = \hat{x}_1^\star, \quad x_2(k) = \hat{x}_2^\star, \quad \forall k \in \mathbb{N}.$$

Proposition 1 shows that $\hat{x}_1^\star$ and $\hat{x}_2^\star$ constitute fixed points of (2), hence the iteration cannot reach $x^\star$ if initialised from these points. This highlights the necessity of devising a new algorithm to deal with the presence of a different constraint set per agent.

3. Distributed methodology

3.1. Proposed algorithm

The main steps of the proposed scheme are summarised 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 $z_i(k)$. Recall that we denote 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 broadcasts 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 $f(x)$, namely, $d_i(k)$. Finally, agents minimise a linear proxy $d_i(k)^T \xi$ of $f(\xi)$ plus a regularisation term weighted by $\frac{1}{c(k)}$. An alternative interpretation of this last computation is that agents update their local estimates by performing a subgradient step with step size $c(k)$

[1] It should be noted that z_i, $i = 1, \ldots, m$, in (2a) should not be confused with that of Step 2 in Algorithm 1 presented in the sequel; we use the same symbol to match the notation in Duchi et al. (2012) and ease the reader.

and projecting $z_i(k) - c(k)d_i(k)$ onto their local set. Indeed, this local update can be rewritten as

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

where $\mathcal{P}_{X_i}[\cdot]$ denotes projection onto the set X_i.

Algorithm 1: Proposed distributed algorithm

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

For $i = 1, \dots, m$, **repeat until convergence**

1: Compute $z_i(k) = \sum_{j=1}^{m} [A(k)]_j^i x_j(k)$,

2: Pick $g_i(z_i(k)) \in \partial f_i(z_i(k))$,

3: Compute $d_i(k) = \sum_{j=1}^{m} [A(k)]_j^i g_j(z_j(k))$,

4: Compute $x_i(k+1) = \operatorname{argmin}_{\xi \in X_i} d_i(k)^T \xi + \frac{1}{2c(k)} \|z_i(k) - \xi\|_2^2$,

5: Set $k \leftarrow k+1$

end

3.2. Algorithm analysis

3.2.1. Convergence in iterates

In this subsection, we impose the following assumption on the step size $c(k)$.

Assumption 3. Let $(c(k))_{k\in\mathbb{N}}$ be the sequence adopted in Algorithm 1. We require that:

(*i*) $c(k)$ is non-negative and non-increasing;
(*ii*) $\sum_{k=1}^{\infty} c(k) = \infty$ and $\sum_{k=1}^{\infty} c(k)^2 < \infty$.

A sequence satisfying Assumption 3 is $c(k) = \frac{\eta}{k+1}$, for $\eta > 0$.

Theorem 1. *Let $(x_i(k))_{k\in\mathbb{N}}$ be the sequences generated by Algorithm 1, for all $i = 1, \dots, m$. Under Assumptions 1–3, we have that for some minimiser $x^\star$ of (1),*

$$\lim_{k\to\infty} \|x_i(k) - x^\star\|_2 = 0, \quad \forall\, i = 1, \dots, m.$$

The proof of Theorem 1, as well as of Theorem 2 presented in the sequel, is based on some auxiliary technical results presented in Appendix A.4. Theorem 1 extends the result in Margellos et al. (2018) by allowing an agent to communicate subgradient information to neighbouring agents, a feature that, as illustrated in Section 5, can speed up practical convergence.

3.2.2. Convergence in objective value and convergence rate

Throughout this section, we impose the following assumption on the step size $c(k)$.

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

Our convergence rate results build on the following related sequence generated by Algorithm 1,

$$\hat{x}_i(k) = \frac{1}{S(k)} \sum_{r=1}^{k} c(r) x_i(r), \tag{4}$$

where $S(k) = \sum_{r=1}^{k} c(r)$, and $(x_i(k))_{k\in\mathbb{N}}$, for all $i = 1, \dots, m$, are the sequences generated by Algorithm 1, with initial condition $\hat{x}_i(0) = x_i(0)$. Note that (4) is a convex combination of past iterates.

Theorem 2. *Consider the running average defined in (4). Under Assumptions 1, 2, and 4, we have that:*

(*i*) *For all $i, j = 1, \dots, m$, the sequence $(\|\hat{x}_i(k) - \hat{x}_j(k)\|)_{k\in\mathbb{N}}$ converges to zero at a rate $\mathcal{O}(\frac{\ln k}{\sqrt{k}})$.*
(*ii*) *All accumulation points of the sequence $(\hat{x}_i(k))_{k\in\mathbb{N}}$ are feasible.*
(*iii*) *The sequence $\left(\left|\sum_{i=1}^{m} f_i(\hat{x}_i(k)) - f(x^\star)\right|\right)_{k\in\mathbb{N}}$ converges to zero at a rate $\mathcal{O}(\frac{\ln k}{\sqrt{k}})$.*

Note that 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, however, the iterates might exhibit an oscillatory behaviour. For the exact expression of B_1 and B_2, we refer the reader to Appendix A.6. The absolute value in Theorem 2 is due to the fact that $\hat{x}_i(k)$ may not be necessarily feasible; however, item (*ii*) in Theorem 2 implies that all accumulation points of $(\hat{x}_i(k))_{k\in\mathbb{N}}$, $i = 1, \dots, m$, are feasible. Item (*i*) states the rate at which consensus is achieved for the sequences $(\hat{x}_i(k))_{k\in\mathbb{N}}$, $i = 1, \dots, m$. Similar rates can be obtained with more general choices for the step size, e.g., $c(k) = \frac{1}{k^a}$, for $a \in [0.5, 1)$.

It should be noted that the result of Theorem 2 further extends the work presented in Margellos et al. (2018) not only by allowing agents to communicate their (sub-) gradients, but by also unveiling how to (non trivially) adapt the proof line in that paper to come up with convergence results that recover traditional rates for distributed subgradient methods. This is the first convergence rate result under the scenario considered in this paper.

4. Comparison with related algorithms

In this section we provide a detailed comparison of the proposed algorithm with other results in the literature. To this end, note that in Johansson et al. (2008) a similar distributed subgradient scheme is mentioned, but no analysis of such a scheme is presented. References Lee and Nedić (2013) and Lin et al. (2016) characterise the convergence rate of a sub-gradient algorithm under different constraint sets per agent that does not possess subgradient averaging. References Margellos et al. (2018) and Zhu and Martinez (2012) show asymptotic convergence of distributed algorithms with different constraint sets and time-varying communication network. Hence, by combining (sub)-gradient averaging and providing an analysis that yields convergence rates under time-varying communication networks and different constraint sets per agent, the results in this paper are distinct from all the above literature.

Closely related algorithms to the one presented here are Mai and Abed (2019) and Wang et al. (2019). Paper Mai and Abed (2019) provides convergence rates assuming a regularity condition on the local sets (weaker than compactness) and requiring the network to be row-stochastic; however, it does not analyse the case where the communication network is time-varying. This requires different analysis arguments, thus complementing the results in Mai and Abed (2019), extending them to allow for time-varying networks. Meanwhile, paper Wang et al. (2019) proposes a subgradient-free algorithm that converges under different constraint sets and undirected time-varying network, which, however, does not involve any subgradient averaging when specialised to use subgradients. Moreover, the example of Section 2.2 highlights the need for developing a different analysis when agents possess different constraints sets.

Although only marginally related to the results of this paper, it is worth mentioning distributed algorithms that deal with similar optimisation problems (Qu & Li, 2018; Scutari & Sun, 2019; Shi et al., 2015). Paper Scutari and Sun (2019) proposes an algorithm whose convergence is valid for non-convex objectives and directed communication network, while Qu and Li (2018) and Shi et al. (2015) use a constant step size to establish linear convergence rates for strongly convex functions. Moreover, distributed algorithms based on proximal methods with constant step sizes

Table 1
Summary of distributed schemes for smooth and non-smooth optimisation.

	Smooth + Constant step-size				Non-smooth + Diminishing step-size			
	Common sets		Different sets		Common sets		Different sets	
	Convex	Strongly Convex	Convex	Strongly Convex	Convex	Strongly Convex	Convex	Strongly Convex
No (sub)grad. avg.	Jakovetic, Moura, and Xavier (2012), Nedic and Ozdaglar (2009), Yuan, Ling, and Yin (2016)	Yuan et al. (2016)	Lei, Chen, and Fang (2016), Xie, You, Tempo, Song, and Wu (2018)	–	Nedić and Olshevsky (2015), Scutari and Sun (2019)	Liu, Qiu, and Xie (2017), Tsianos, Lawlor, and Rabbat (2012)	Lee and Nedić (2013), Lin et al. (2016), Margellos et al. (2018), Wang, Zhao, Hong, and Zamani (2019), Zhu and Martinez (2012)	–
(Sub)grad. avg.	Qu and Li (2018), Scutari and Sun (2019), Shi, Ling, Wu, and Yin (2015), Zanella, Varagnolo, Cenedese, Pillonetto, and Schenato (2016)	Qu and Li (2018), Scutari and Sun (2019), Shi et al. (2015)	–	–	Duchi et al. (2012), Liang et al. (2019), Scutari and Sun (2019), Xi and Khan (2017)	–	our work, Mai and Abed (2019)	–

have been proposed in Chen and Ozdaglar (2012). In this setting, the objective function is assumed to be differentiable to obtain convergence to the optimal solution of problem (1), and the size of the allowable step-size is upper bounded by a quantity related to the Lipschitz constant of the objective function. Unlike these results, we allow for non-differentiable objective functions.

To better position this paper within the recent literature, we summarise the main distributed algorithms that are amenable to smooth and non-smooth constrained optimisation in Table 1. We highlight both scenarios of common and different local constraint sets, which are indicated in the table by common sets and different sets, respectively. In this brief summary, we restrict our attention to algorithms that use constant step size for smooth optimisation, and to those that use diminishing step sizes for the non-smooth case. We also present a categorisation of these schemes between those that have results for general convex functions and strongly convex functions. In row entitled "No (sub)grad. avg.", we include distributed algorithms based on projected (sub)gradient, proximal minimisation, and primal–dual update that do not leverage on averaging first-order information from neighbouring agents. In contrast, row "(Sub)grad. avg." includes algorithms that exploit (sub) gradient averaging. Among the few papers that are suitable for different local sets, this is the first result to establish a convergence rate that matches that of the common local sets case, and simultaneously allows agents to use first-order information of their neighbours under time-varying communication networks, thus speeding up practical convergence.

5. Numerical examples

5.1. Problem instance of Section 2.2 — revisited

We revisit the two-agent problem in (3), for which the iterative scheme in (2) is not guaranteed to converge, and apply this time our algorithm. Note that the optimal solution of (3) is given by

$$x^\star = \mathcal{P}_{[0.5,1]^2}\left[-\frac{1}{8}Q^{-1}(q_1+q_2)\right] = \begin{bmatrix}0.5\\1\end{bmatrix}$$

where $\mathcal{P}_{[0.5,1]^2}[\cdot]$ represents the projection onto the feasible set of problem (3). Pictorially $x^\star$ is shown in Fig. 1. To illustrate the convergence properties of Algorithm 1 we monitor the evolution of $\sqrt{\sum_{i=1}^2 \|x_i(k)-x^\star\|_2^2}$, where $(x_i(k))_{k\in\mathbb{N}}$, $i=1,2$, are the iterates generated by Algorithm 1. We use $c(k)=\frac{1}{\sqrt{k+1}}$ similarly to Duchi et al. (2012), $A=\frac{1}{2}\mathbf{1}\mathbf{1}^T$ and $x_i(0)=\hat{x}_i^\star$, where $\hat{x}_i^\star$, $i=1,2$, are defined in Section 2.2. Observe that our initial condition is the same as in Proposition 1. In contrast, as shown in Fig. 2, the iterates generated by Algorithm 1 converge to the optimal solution of (3).

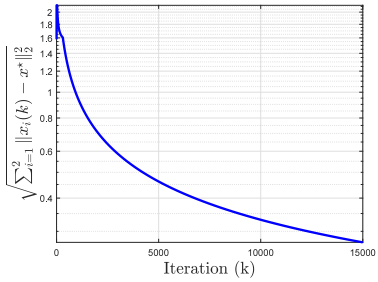

Fig. 2. Evolution of $\sqrt{\sum_{i=1}^2 \|x_i(k)-x^\star\|_2^2}$ for (3), where $(x_i(k))_{k\in\mathbb{N}}$, $i=1,2$, are the iterates generated by Algorithm 1.

5.2. Example 2: robust linear regression

We consider the problem of estimating an unknown (but deterministic) vector $x\in\mathbb{R}^n$ from m noisy measurements y_i by means of the linear model

$$y_i = b_i^T x + v_i, \quad i=1,\dots,m,$$

with $b_i\in\mathbb{R}^n$, and v_i are independent random variables drawn from a Laplacian distribution, that is, for each i the density of v_i is given by $h_{v_i}(z)=\frac{1}{2a}\exp^{-|z|/a}$, for all $z\in\mathbb{R}$.

A common strategy is to impose a norm constraint of the form $\|x\|_2\le c$, for some $c>0$, to reflect some prior knowledge on the unknown vector x, and solve the second order conic programme

$$\hat{x}\in \operatorname*{argmin}_{\|x\|_2\le c} \|y-Bx\|_1. \tag{5}$$

Typically, (5) is referred to as robust regression in the literature, as the ℓ_1-norm penalises relatively less outliers than other convex

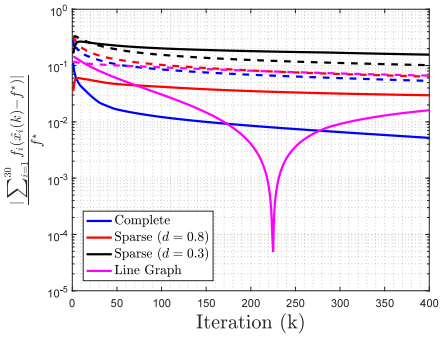

Fig. 3. Evolution of $\frac{|\sum_{i=1}^{30} f_i(x_i(k))-f^\star|}{f^\star}$ for Algorithm 1 (solid lines) and the one in Duchi et al. (2012) (dashed lines) when applied to the robust regression problem given by (5). The different colours correspond to the different network connectivities.

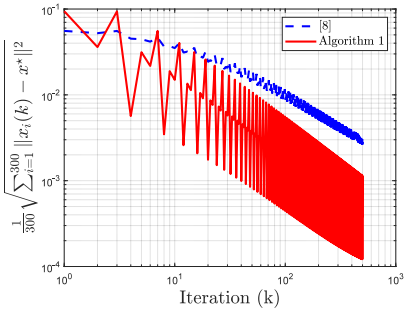

Fig. 4. Evolution of the average distance to the optimal solution given by $\frac{\sqrt{\sum_{i=1}^{300}\|x_i(k)-x^\star\|_2^2}}{300}$ for Algorithm 1 (solid-red line) and that of Margellos et al. (2018) (dashed-blue line).

metrics (e.g., quadratic penalties). In our set-up, we consider the case where data are collected locally and agents are not willing to share their measurements with a central processing unit.

Observe that (5) has the format of (1) by setting $X_i = X = \{x \in \mathbb{R}^n : \|x\|_2 \leq 5\}$ and $f_i(x) = |y_i - b_i^T x|$, $i = 1, \ldots, m$. Moreover, the constraint sets X_i and the objective functions f_i, $i = 1, \ldots, m$, trivially satisfy Assumption 1. Hence, we can apply the proposed scheme to obtain a solution to (5). We consider $m = 30$ and $n = 4$ and generate y independently from a standard Gaussian distribution, and matrix B from a uniform distribution with support $[0, 1]$.

We solve (5) in a distributed manner, and compare Algorithm 1 with the one proposed in Duchi et al. (2012) under four different network connectivity structures: (i) complete network graph (which corresponds to the centralised version of the problem); (ii) line network graph; (iii) sparse network graph with sparsity degree $d = 0.3$; (iv) sparse network graph with sparsity degree $d = 0.8$. We say that a network with m agents has a sparsity degree $d \in (0, 1)$ if the number of connections among the network nodes is given by dm^2, where m^2 indicates the number of connections of a complete graph.

We assess the performance of Algorithm 1 for each of the aforementioned networks in Fig. 3. Solid lines correspond to Algorithm 1, whereas dashed lines correspond to the algorithm proposed in Duchi et al. (2012). Different colours correspond to the different network connectivities. For each case, we monitor the evolution of $\frac{|\sum_{i=1}^{30} f_i(x_i(k))-f^\star|}{f^\star}$, where $f^\star$ is the optimal value of (5). The proposed scheme exhibits similar and often favourable performance with the one in Duchi et al. (2012), in particular for cases where the underlying graph is not sparse. It should be noted, however, that Algorithm 1 possesses more general convergence properties, i.e., the proposed scheme is guaranteed to converge under non-identical local sets.

Note that due to the fact that Algorithm 1 requires two rounds of communication per iteration, the results presented in Fig. 3 should be rescaled by a factor of two if we use communication rounds instead of the iteration index.

5.3. Example 3: ℓ_2 linear regression with regularisation

In this example, we consider a variation of the regression problem where we assume v_i, $i = 1, \ldots, m$, to be independent and Gaussian, i.e., the density function is given by $h_{v_i}(z) = \frac{1}{\sqrt{2\pi}} e^{-\frac{z^2}{2}}$, for all $z \in \mathbb{R}$, for all $i = 1, \ldots, m$, and we assume that x is sparse. A common relaxation of this problem is to choose the maximum likelihood estimator $\hat{x}$ such that

$$\hat{x} = \operatorname*{argmin}_{x \in X} \|y - Bx\|_2^2 + \lambda \|x\|_1, \tag{6}$$

where X can be interpreted as a set including prior beliefs, e.g., $\|x\|_2 \leq c$ or $\underline{x} \leq x \leq \bar{x}$ for some vectors $\underline{x}, \bar{x} \in \mathbb{R}^n$. The estimator $\hat{x}$ obtained by solving (6) depends on the value of the parameter λ. In fact, the larger the value of λ, the worse the performance is in terms of the error and the sparser the obtained solution is.

In this example, we aim to verify the performance of Algorithm 1 under step size choices $c(k) \propto \frac{1}{k+1}$ and a time-varying communication network. Similar to the previous example, the vector y is generated according to a standard normal distribution and matrix B from a uniform distribution on the interval $[0, 1]$. We assume $m > n$ and consider the case where agents possess private, local information, encoded by $X_i = [\underline{x}_i, \bar{x}_i]$ $i = 1, \ldots, m$, such that $X = \cap_{i=1}^m X_i = [\underline{x}, \bar{x}]$.

The algorithm presented in Duchi et al. (2012) does not necessarily converge in the set-up of problem (6), as we have different constraint sets per agent. We thus compare our algorithm against the one proposed in Margellos et al. (2018), which converges under similar conditions but does not leverage on subgradient averaging. This allows us to assess the impact of averaging subgradients on practical convergence.

We now investigate the behaviour of the proposed algorithm in the presence of time-varying communication networks. To this end, we set $m = 300$ and $n = 10$, and generate four network configurations with different sparsity patterns, alternating cyclically among these. We also set $c(k) = \frac{0.2}{k+1}$ for both Algorithm 1 and the one in Margellos et al. (2018). Fig. 4 shows the evolution for the average distance to the optimal solution for Algorithm 1 (solid-red line) and the one in Margellos et al. (2018) (dashed-blue line). We observe that Algorithm 1 consistently outperforms the one proposed in Margellos et al. (2018); this is mainly due to the sub-gradient averaging step of Algorithm 1.

6. Conclusion

In this paper we proposed a subgradient averaging algorithm for multi-agent optimisation problems involving non-differentiable objective functions and different constraint sets per agent. For this set-up we showed by means of a geometric construction that available schemes involving subgradient averaging cannot be used. For the proposed scheme we showed convergence of the algorithm iterates to some minimiser of a centralised problem counterpart. Moreover, we have also established a convergence rate under a particular choice for the underlying step size. The performance of our approach was illustrated by means of several numerical examples, quantifying also the improvement in terms of practical convergence with respect to other algorithms that are not based on (sub)gradient exchange.

Future work will concentrate towards replacing the diminishing step size employed by our approach with a constant one, showing convergence rates to a neighbourhood of the set of optimal solutions. A more detailed study on the communication requirements, and an investigation on how we could reduce the two rounds of communication required by the proposed algorithm are also topics of current work.

Acknowledgements

L. Romao is supported by the Coordination for the Improvement of Higher Education Personnel (CAPES) - Brazil. The work of K. Margellos and A. Papachristodoulou has been supported by EPSRC UK under grants EP/P03277X/1 and EP/M002454/1, respectively. Giuseppe Notarstefano is supported by the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation program (Grant 638992-OPT4SMART).

Appendix

A.1. Proof of Lemma 1

We start by proving item (i). Consider the continuous mapping $\phi : \mathbb{R}^m \times \prod_{i=1}^m \mathbb{R}^n \to \mathbb{R}^n$, defined as $\phi(\gamma, x_1, \dots, x_m) = \sum_{i=1}^m \gamma_i x_i$, where $\gamma = (\gamma_1, \dots, \gamma_m)$ belongs to the simplex in $\mathbb{R}^m$, denoted by Γ. Consider $K = \phi(\Gamma, \prod_{i=1}^m X_i)$, and note that K is compact, as it is the image of the compact set $\Gamma \times \prod_{i=1}^m X_i$ under the continuous map ϕ. Moreover, note that by definition we have $K \subseteq \text{conv}(\cup_{i=1}^m X_i)$, as any element in K is a convex combination of elements in $\cup_{i=1}^m X_i$. To conclude the argument, we need to show that $\text{conv}(\cup_{i=1}^m X_i) \subseteq K$. To this end, it suffices to show that K is a convex set, due to the fact that the convex hull is the smallest convex set containing a given set. Let $z, w \in K$, i.e., $z = \sum_{i=1}^m \gamma_i z_i$ and $w = \sum_{i=1}^m \beta_i w_i$, with $z_i, w_i \in X_i$, and $\gamma = (\gamma_1, \dots, \gamma_m)$, $\beta = (\beta_1, \dots, \beta_m) \in \Gamma$. Fix an $\alpha \in (0, 1)$, and note that $\alpha z + (1-\alpha)w = \sum_{i=1}^m (\alpha\gamma_i + (1-\alpha)\beta_i)x_i$, where $x_i = c_i z_i + (1-c_i)w_i \in A_i$, with $c_i = \frac{\alpha\gamma_i}{\alpha\gamma_i + (1-\alpha)\beta_i}$.

Since $x_i \in A_i$ due to convexity of A_i and $\alpha\gamma + (1-\alpha)\beta \in \Gamma$, we conclude that $\alpha z + (1-\alpha)w \in K$ for any $\alpha \in (0, 1)$, thus showing that K is a convex set. This implies then that $K = \text{conv}(\cup_{i=1}^m X_i)$ as we have established that $K \subseteq \text{conv}(\cup_{i=1}^m X_i)$ and $\text{conv}(\cup_{i=1}^m X_i) \subseteq K$. Since K was shown to be compact, we have that $\text{conv}(\cup_{i=1}^m X_i)$ is also compact. This concludes the proof of item (i). An alternative proof can be found at Bertsekas (2009, Prop. 1.2.2). The proof of item (ii) follows from Proposition 5.4.2, p. 186, in Bertsekas (2009), and is omitted for brevity. This concludes the proof of the lemma.

A.2. Sufficient condition for Assumption 1 , item (iii).

The goal of this subsection is to provide a sufficient condition for Assumption 1, item (iii). The subsequent arguments can be found in standard optimisation books, such as Rockafellar (1972, Theorem 24.7); however we present here a more direct proof.

Assumption 5. Let X_i, $i = 1, \dots, m$, be the level sets of problem (1) and $\text{dom} f$ the domain of f. We suppose that:

(i) 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.,

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

(ii) $X_i \subset \cap_{i=1}^m \text{int}(\text{dom} f_i)$ for each $i = 1, \dots, m$.

As a consequence of Assumption 5, and since $\text{dom} f = \cap_{i=1}^m \text{dom} f_i$, $\text{ri}(\text{dom} f) = \cap_{i=1}^m \text{ri}(\text{dom} f_i)$ and $\text{ri}(\text{dom} f_i) \subset \text{dom} f_i$ we have that the subdifferential $\partial f(x)$ is non empty for each $x \in \cap_{i=1}^m X_i$, as by item (ii) of Assumption 5 every feasible solution of (1) belongs to the interior of the domain of f. Furthermore, $\partial f(x)$ is compact by Bertsekas (2009, Proposition 5.4.1) since the affine hull of $\text{dom} f$ has dimension n due to Assumption 1, item (ii).

We use this fact to show that $\cup_{x \in \text{conv}(\cup X_i)} \partial f(x)$ is a bounded set, that is, $\|g\|_2 \leq L$, 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 2. *Under Assumptions 1, items (i) and (ii), and 5, we have that the set $\cup_{x \in \text{conv}(\cup X_i)} \partial f(x)$ is non-empty and bounded.*

Proof. The proof of the lemma relies on Assumption 5, item (ii), that is, $X_i \subset \cap_{j=1}^m \text{ri}(\text{dom} f_j)$, for all $i = 1, \dots, m$. This implies that $\text{conv}(\cup_{i=1}^m X_i) \subset \cap_{j=1}^m \text{ri}(\text{dom} f_j)$, as $\cap_{j=1}^m \text{ri}(\text{dom} f_j)$ is convex and contains $\cup_{i=1}^m X_i$. Suppose, by contradiction, that $\cup_{x \in \text{conv}(\cup X_i)} \partial f(x)$ is unbounded. Then there exists a sequence $(x_k)_{k \in \mathbb{N}} \subset \text{conv}(\cup_{i=1}^m X_i)$ such that $(g_k)_{k \in \mathbb{N}}$, with $g_k \in \partial f(x_k)$, satisfies $\|g_k\|_2 < \|g_{k+1}\|_2$, $\forall\, k \in \mathbb{N}$.

Notice that $x_k \in \cap_{i=1}^m \text{int}(\text{dom} f_i)$ by Assumption 5, item (ii). By item (i) of Assumption 5, we can construct a sequence $(\beta_k)_{k \in \mathbb{N}}$ such that $x_k + \beta_k d_k \in \cap_{i=1}^m \text{dom} f_i$. with $d_k = g_k / \|g_k\|_2$. Let $\beta = \inf_{k \in \mathbb{N}} \beta_k$ and notice that $\beta > 0$ (i.e., it is bounded away from zero) due to Assumption 5, item (i). By the definition of g_k we have that

$$\frac{f(x_k + \beta d_k) - f(x_k)}{\beta} \geq \|g_k\|_2, \quad \forall\, k \in \mathbb{N}. \tag{7}$$

As inequality (7) is valid for all $k \in \mathbb{N}$, we take the limit superior on both sides to obtain

$$\limsup_{k \to \infty} \|g_k\|_2 \leq \limsup_{k \to \infty} \frac{f(x_i + \gamma d_k) - f(x_k)}{\gamma} < \infty, \tag{8}$$

where the right-hand side of (8) is finite as the sequences $(x_k)_{k \in \mathbb{N}}$ and $(d_k)_{k \in \mathbb{N}}$ are bounded (notice that d_k is a normalised subgradient), and since f is continuous on its domain (f is convex). This establishes a contradiction, as we assumed $(g_k)_{k \in \mathbb{N}}$ were unbounded, thus concluding the proof of item (ii).

A.3. Proof of Proposition 1

The proof is based on an induction argument.

Base case

We show that $z_i(1)^T(\xi - \hat{x}_j^\star) \geq 0$, for all $\xi \in X_j$, for all $i, j = 1, 2$, and also that $x_i(1) = \hat{x}^\star$, for all $i = 1, 2$. Consider the inequalities

$$\begin{aligned} &\nabla f_1(\hat{x}_1^\star)^T(\xi - \hat{x}_i^\star) \geq 0, \\ &\nabla f_2(\hat{x}_2^\star)^T(\xi - \hat{x}_i^\star) \geq 0, \quad \forall \xi \in X_i, \quad i = 1, 2. \end{aligned} \tag{9}$$

Fix $i = 1$. The first inequality in (9) holds due to optimality of $\hat{x}_1^\star$ (Bertsekas, 2009). To show the second inequality observe that $\nabla f_2(\hat{x}_2^\star) = [13.68, -3.94]^T$, and that $\xi - \hat{x}_1^\star = [a_1, a_2]^T$ with $a_1 \geq 0$ and $a_2 \leq 0$, for all $\xi \in X_1$.

Since $\nabla f_1(\hat{x}_1^\star) = [12, -4]^T$, using a symmetric argument we show that

$$\begin{aligned} &\nabla f_2(\hat{x}_2^\star)^T(\xi - \hat{x}_2^\star) \geq 0, \\ &\nabla f_1(\hat{x}_1^\star)^T(\xi - \hat{x}_2^\star) \geq 0, \quad \forall \xi \in X_2. \end{aligned} \tag{10}$$

By (2a), and under our choice for A,

$$z_i(1) = \frac{1}{2}\left(\nabla f_1(\hat{x}_1^\star) + \nabla f_2(\hat{x}_2^\star)\right) + \nabla f_i(\hat{x}_i^\star), \tag{11}$$

for $i = 1, 2$, hence inequalities (9) and (10) imply that $z_i(1)^T(\xi - \hat{x}_j^\star) \geq 0$, $\forall \xi \in X_j$, for all $i, j = 1, 2$.

We will now prove that $x_i(1) = \hat{x}_i^\star$, for $i = 1, 2$. Fix $i = 1$. Since $z_1(1)^T\xi + \frac{2}{c(k)}\|\xi\|_2^2$ is strictly convex, there is a unique point satisfying

$$\Big(z_1(1) + 2x_1(1)\Big)^T(\xi - x_1(1)) \geq 0, \quad \forall \xi \in X_1, \tag{12}$$

where $(z_1(1) + 2x_1(1))$ is the gradient of the objective function in (2b) evaluated at $x_1(1)$, with $c(1) = 1$. Therefore, it suffices to show that

$$\Big(z_1(1) + 2\hat{x}_1^\star\Big)^T(\xi - \hat{x}_1^\star) \geq 0, \quad \forall \xi \in X_1. \tag{13}$$

By substituting (3) into (11), we observe that $z_1(1) + 2\hat{x}_1^\star = [22.8414, -5.9708]^T$, and due to the structure of $\xi - \hat{x}_1^\star$, (13) holds, thus proving that $x_1(1) = \hat{x}_1^\star$. A symmetric argument yields that $x_2(1) = x_2^\star$.

Induction hypothesis

Assume that $z_i(k)^T(\xi - \hat{x}_j^\star) \geq 0$ for all $\xi \in X_j$, for $i, j = 1, 2$, and that $x_i(k) = x_i^\star$ for $i = 1, 2$. We aim to show that the aforementioned relations remain true for the step $k + 1$.

Proof for iteration $k + 1$

Fix $i = 1$. Following a similar reasoning with the base case, observe that $x_1(k + 1) = x_1^\star$ if

$$\left[z_1(k+1) + \frac{2}{c(k)}\hat{x}_1^\star\right]^T(\xi - \hat{x}_1^\star) \geq 0, \quad \forall \xi \in X_1. \tag{14}$$

As the sequence $(z_i(k))_{k\in\mathbb{N}}$ is generated by (2a), we propagate the dynamical system in (2a) by $k + 1$ steps to obtain

$$z_i(k+1) = \frac{1}{2}\Big(\nabla f_1(\hat{x}_1^\star) + \nabla f_2(\hat{x}_2^\star)\Big)(k+1) + \nabla f_1(\hat{x}_1^\star),$$

where we have used the fact that $A = \frac{1}{m}\mathbf{1}\mathbf{1}^T$ and $c(k) = \frac{1}{\sqrt{k+1}}$. A sufficient condition for Eq. (14) to hold is that

$$\left[\frac{1}{2}\Big(\nabla f_1(\hat{x}_1^\star) + \nabla f_2(\hat{x}_2^\star)\Big)(k+1) + 2\hat{x}_1^\star\sqrt{k+1}\right]^T(\xi - \hat{x}_1^\star) \geq 0, \quad \forall \xi \in X_1, \tag{15}$$

since $\nabla f_1(\hat{x}_1^\star)^T(\xi - \hat{x}_1^\star) \geq 0$ by optimality of $\hat{x}_1^\star$. Recall that $(\xi - \hat{x}_1^\star) = [a_1, a_2]$ with $a_1 \geq 0$ and $a_2 \leq 0$ for all $\xi \in X_1$. To prove (15) we will show that the left-most vector in the same equation can be written as $[b_1, b_2]$ for some $b_1 \geq 0$ and $b_2 \leq 0$. To achieve this, notice that $k + 1 \geq \sqrt{2}\sqrt{k+1}$, for all $k \geq 1$, and let e_i denote the unit vector with 1 in the i-th position, $i = 1, 2$. We then have that

$$e_1^T\left[\frac{1}{2}\Big(\nabla f_1(\hat{x}_1^\star) + \nabla f_2(\hat{x}_2^\star)\Big)\right](k+1) \geq e_1^T\left[\frac{\sqrt{2}}{2}\Big(\nabla f_1(\hat{x}_1^\star) + \nabla f_2(\hat{x}_2^\star)\Big)\right]\sqrt{k+1}, \tag{16}$$

and

$$2e_2^T\hat{x}_1^\star\sqrt{k+1} \leq \sqrt{2}e_2^T\hat{x}_1^\star(k+1), \tag{17}$$

since the first component of the averaged gradient and the second component of $\hat{x}_1^\star$ are both positive. Therefore, for all $k \in \mathbb{N}$,

$$b_1 \geq 16.1604\sqrt{k+1} > 0, \quad b_2 \leq -2.5566(k+1) < 0. \tag{18}$$

Inequalities (16), (17) and (18), together with the structure of $\xi - \hat{x}_1^*$, imply that (15) holds, so we can conclude that $x_1(k+1) = \hat{x}_1^\star$. A symmetric argument shows that $x_2(k+1) = \hat{x}_2^\star$.

To complete the proof it remains to show that $z_i(k+1)^T(\xi - \hat{x}_j^\star) \geq 0$ for all $\xi \in X_j$, for all $i, j = 1, 2$, where $z_i(k+1) = \frac{1}{2}\Big(z_1(k) + z_2(k)\Big) + \nabla f_i(x_i(k))$, due to (2a) and our choice for A. By our induction hypothesis, $z_i(k)(\xi - \hat{x}_j^\star) \geq 0$, for all $i, j = 1, 2$, hence it suffices to show that $\nabla f_i(x_i(k))^T(\xi - \hat{x}_j^\star) \geq 0$, $\forall \xi \in X_j$, $\forall i = 1, 2$. Since $x_i(k) = \hat{x}_i^\star$ for $i = 1, 2$, due to our induction hypothesis, the claim follows from (9) and (10), thus concluding the proof.

A.4. Auxiliary Lemmas for the proofs of Theorems 1 and 2

Let

$$v(k) = \frac{1}{m}\sum_{i=1}^m x_i(k), \tag{19}$$

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

$$\bar{v}(k) = \frac{\rho}{\epsilon(k) + \rho}v(k) + \frac{\epsilon(k)}{\epsilon(k) + \rho}\bar{x}, \tag{20}$$

where $\bar{x}$ is a point in the interior of the feasible set (which is non-empty by Assumption 1, item (*ii*)), $\rho > 0$ is such that the 2-norm ball of centre $\bar{x}$ and radius ρ is contained in $\cap_{i=1}^m X_i$, and $\epsilon(k) = \sum_{i=1}^m \text{dist}(v(k), X_i)$. As shown in Nedic et al. (2010), $\bar{v}(k) \in \cap_{i=1}^m X_i$, for all $k \in \mathbb{N}$. We also define $e_i(k+1) = x_i(k+1) - z_i(k)$, and note that the z_i-update in Algorithm 1 can be written as

$$x_i(k+1) = \sum_{j=1}^m [A(k)]_j^i x_j(k) + e_i(k+1). \tag{21}$$

Lemma 3. *The following relations hold.*

(*i*) *Let $(x_i(k))_{k\in\mathbb{N}}$, $i = 1, \ldots, m$, be the sequences generated by Algorithm 1, and $(v(k))_{k\in\mathbb{N}}$ and $(\bar{v}(k))_{k\in\mathbb{N}}$ defined by (19) and (20), respectively. Under Assumption 1, we have that for all $k \geq 0$,*

$$\sum_{i=1}^m \|x_i(k+1) - \bar{v}(k)\|_2 \leq \mu\sum_{i=1}^m \|x_i(k) - v(k)\|_2,$$

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, item (i)).

(*ii*) *Let $(x_i(k))_{k\in\mathbb{N}}$, $i = 1, \ldots, m$, and $(v(k))_{k\in\mathbb{N}}$ be as in item (i). Under Assumption 2, we have that for all $i = 1, \ldots, m$, for all $k \geq 0$,*

$$\|x_i(k+1) - v(k+1)\|_2 \leq \lambda q^k\sum_{j=1}^m \|x_j(0)\|_2 + \|e_i(k+1)\|_2 + \sum_{r=0}^{k-1}\lambda q^{k-r-1}\sum_{j=1}^m \|e_j(r+1)\|_2 + \frac{1}{m}\sum_{j=1}^m \|e_j(k+1)\|_2,$$

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

(*iii*) *Given a non-increasing and non-negative sequence $(c(k))_{k\in\mathbb{N}}$, and a scalar $\bar{L} > 0$, we have that*

$$2\bar{L}\sum_{k=0}^N c(k)\sum_{i=1}^m \|x_i(k+1) - \bar{v}(k+1)\|_2 < \beta_1\sum_{k=0}^N\sum_{i=1}^m \|e_i(k+1)\|_2^2 + \beta_2\sum_{k=0}^N c(k)^2 + \beta_3,$$

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

Proof. The proof of item (*i*) is presented in Margellos et al. (2018, Lemma 1). For item (*ii*), see Margellos et al. (2018, Lemma 2). Finally, the proof of item (*iii*) follows the line of Margellos et al. (2018, Lemma 3).

Observe that the values of λ and q in Lemma 3, item (*ii*), depend on the parameter T that characterises the uniform bound in Assumption 2, item (*i*); and on η, the lower bound for the elements of $A(k)$, Assumption 2, item (*ii*). The following lemma is instrumental for the proof of Theorem 2. In particular, Lemma 4, item (*ii*), constitutes a non-trivial extension of the result in Margellos et al. (2018), allowing some sequences to be iteration-varying.

Lemma 4. *Let $(x_i(k))_{k\in\mathbb{N}}, (z_i(k))_{k\in\mathbb{N}}$ and $(d_i(k))_{k\in\mathbb{N}}$, $i = 1,\dots,m$, be the sequences generated by Algorithm 1, and $x^\star$ by any optimal solution of (1). Under Assumptions 1 and 2, we have that:*

(*i*) *For all $k \in \mathbb{N}$,*

$$2c(k)\sum_{i=1}^{m} d_i(k)^T(x_i(k+1)-x^\star)+\sum_{i=1}^{m}\|e_i(k+1)\|_2^2 + \sum_{i=1}^{m}\|x_i(k+1)-x^\star\|_2^2 \le \sum_{i=1}^{m}\|x_i(k)-x^\star\|_2^2. \tag{22}$$

(*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, for all $k \in \mathbb{N}$, $\alpha_1(k) \in (0,1)$, $\alpha_2(k) \in (0,1)$, $1-\beta_1-\alpha_1(k)-\alpha_2(k) \ge 0$ and*

$$\begin{aligned} &2\sum_{k=0}^{N} c(k)\sum_{i=1}^{m}(f_i(\bar v(k+1))-f_i(x^\star)) \\ &+\sum_{k=0}^{N}(1-\alpha_1(k)-\alpha_2(k)-\beta_1)\sum_{i=1}^{m}\|e_i(k+1)\|_2^2 \\ &+\sum_{k=0}^{N}\sum_{i=1}^{m}\|x_i(k+1)-x^\star\|_2^2 \le \sum_{k=0}^{N}\sum_{i=1}^{m}\|x_i(k)-x^\star\|_2^2 \\ &+\sum_{k=0}^{N}\Big(mL^2\frac{\alpha_1(k)+\alpha_2(k)}{\alpha_1(k)\alpha_2(k)}+\beta_2\Big)c(k)^2+\beta_3. \end{aligned} \tag{23}$$

Proof. Item (*i*): Fix any $i \in \{1,\dots,m\}$ and consider the sequence $(x_i(k))_{k\in\mathbb{N}}$. By optimality of $x_i(k+1)$ (see Algorithm 1), for any $\xi \in X_i$,

$$\begin{aligned} &d_i(k)^T x_i(k+1)-\frac{1}{c(k)}(z_i(k)-x_i(k+1))^T x_i(k+1) \\ &\quad\le d_i(k)^T\xi-\frac{1}{c(k)}(z_i(k)-x_i(k+1))^T\xi, \end{aligned} \tag{24}$$

where $d_i(k)-\frac{1}{c(k)}(z_i(k)-x_i(k+1))$ constitutes the gradient of the objective function in the x_i-update of Algorithm 1, evaluated at $x_i(k+1)$. Fix any optimal solution of (1), $x^\star \in \cap_{i=1}^m X_i$, and consider the following identity

$$\begin{aligned} &-\frac{1}{c(k)}(z_i(k)-x_i(k+1))^T(x_i(k+1)-x^\star) \\ &=\frac{1}{2c(k)}\|x_i(k+1)-z_i(k)\|_2^2+\frac{1}{2c(k)}\|x_i(k+1)-x^\star\|_2^2 \\ &-\frac{1}{2c(k)}\|z_i(k)-x^\star\|_2^2. \end{aligned} \tag{25}$$

Combining (24) and (25) with $\xi = x^\star$, we obtain

$$\begin{aligned} &d_i(k)^T x_i(k+1)+\frac{1}{2c(k)}\|x_i(k+1)-z_i(k)\|_2^2 \\ &\quad+\frac{1}{2c(k)}\|x_i(k+1)-x^\star\|_2^2 \\ &\le d_i(k)^T x^\star+\frac{1}{2c(k)}\|z_i(k)-x^\star\|_2^2 \\ &\le d_i(k)^T x^\star+\frac{1}{2c(k)}\sum_{j=1}^{m}[A(k)]_j^i\|x_j(k)-x^\star\|_2^2, \end{aligned} \tag{26}$$

where the last inequality follows from double stochasticity of $A(k)$ and convexity of $\|\cdot\|^2$.

We now multiply both sides of (26) by $2c(k)$ and sum the result for all $i = 1,\dots,m$, to obtain

$$\begin{aligned} &2c(k)\sum_{i=1}^{m} d_i(k)^T x_i(k+1)+\sum_{i=1}^{m}\|x_i(k+1)-z_i(k)\|_2^2 \\ &\quad+\sum_{i=1}^{m}\|x_i(k+1)-x^\star\|_2^2 \le 2c(k)\sum_{i=1}^{m} d_i(k)^T x^\star \\ &\quad+\sum_{i=1}^{m}\|x_i(k)-x^\star\|_2^2, \end{aligned} \tag{27}$$

where $\sum_{i=1}^m\sum_{j=1}^m[A(k)]_j^i\|x_j(k)-x^\star\|_2^2 = \sum_{i=1}^m\|x_i(k)-x^\star\|_2^2$ by exchanging the order of summation, and due to double stochasticity of $A(k)$. The result follows from (27) by recalling that $e(k+1) = x_i(k+1)-z_i(k)$ and moving the first term on the right-hand side of (27) to the left one. This concludes the proof of item (*i*).

Item (*ii*): Consider the first term on the left-hand side of (22), and rewrite it as

$$\begin{aligned} &2c(k)\sum_{i=1}^{m} d_i(k)^T(x_i(k+1)-x^\star) = \\ &\quad 2c(k)\sum_{i=1}^{m} d_i(k)^T(x_i(k+1)-\bar v(k+1)) \\ &\quad+2c(k)\sum_{i=1}^{m} d_i(k)^T(\bar v(k+1)-x^\star) \end{aligned} \tag{28}$$

by adding and subtracting $\bar v(k+1)$. We next consider the terms on the right hand-side of (28) separately. First, observe that

$$\begin{aligned} &2c(k)\sum_{i=1}^{m} d_i(k)^T(x_i(k+1)-\bar v(k+1)) \\ &\quad\ge -2c(k)L\sum_{i=1}^{m}\|x_i(k+1)-\bar v(k+1)\|_2, \end{aligned} \tag{29}$$

by the Cauchy-Schwartz inequality, where $L = \max_{\xi\in\cup_{i=1}^m X_i}\|g_j(\xi)\|_2$, which is well-defined due to Lemma 1. Using the definition of $d_i(k)$ – see Algorithm 1 – in the second term on the right-hand side of (28), we then have that (via double stochasticity of A)

$$\begin{aligned} &2c(k)\sum_{i=1}^{m} d_i(k)^T(\bar v(k+1)-x^\star) \\ &\quad=2c(k)\sum_{i=1}^{m} g_i(z_i(k))^T(\bar v(k+1)-x^\star). \end{aligned} \tag{30}$$

Moreover, by adding and subtracting $x_i(k+1)$ and $z_i(k)$ for all $i = 1,\dots,m$, into the right-hand side of (30) we obtain

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

$$+2c(k)\sum_{i=1}^{m} g_i(z_i(k))^T(x_i(k+1)-z_i(k))$$
$$+2c(k)\sum_{i=1}^{m} g_i(z_i(k))^T(z_i(k)-x^\star). \tag{31}$$

Consider now the right-hand side of (31). The left-most term can be lower-bounded as

$$2c(k)\sum_{i=1}^{m} g_i(z_i(k))^T(\bar{v}(k+1)-x_i(k+1))$$
$$\geq -2c(k)L\sum_{i=1}^{m}\|(\bar{v}(k+1))-x_i(k+1)\|_2, \tag{32}$$

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

$$2c(k)\sum_{i=1}^{m} g_i(z_i(k))^T(x_i(k+1)-z_i(k))$$
$$\geq -2c(k)L\sum_{i=1}^{m}\|e_i(k+1)\|_2$$
$$\geq -\alpha_1(k)\sum_{i=1}^{m}\|e_i(k+1)\|_2^2 - m\frac{L^2}{\alpha_1(k)}c(k)^2 \tag{33}$$

where the first inequality follows from the Cauchy-Schwartz inequality and the definition $e_i(k)$ in (21). For the second inequality, we employed the relation $2xy \leq x^2+y^2$ with $x=\frac{L}{\sqrt{\alpha_1(k)}}c(k)$ and $y=\sqrt{\alpha_1(k)}\|e_i(k+1)\|_2$ for some $\alpha_1(k)\in(0,1)$, $k\in\mathbb{N}$.

Similarly, the right-most term of (31) can be manipulated to yield

$$2c(k)\sum_{i=1}^{m} g_i(z_i(k))^T(z_i(k)-x^\star)$$
$$\geq 2c(k)\sum_{i=1}^{m}\Big(f_i(z_i(k))-f_i(x^\star)\Big)$$
$$= 2c(k)\sum_{i=1}^{m}\Big(f_i(z_i(k))-f_i(\bar{v}(k+1))\Big)$$
$$+\ 2c(k)\sum_{i=1}^{m}\Big(f_i(\bar{v}(k+1))-f_i(x^\star)\Big) \tag{34}$$

where the inequality follows from the definition of the subgradient for a convex function, and the equality by adding and subtracting $f_i(\bar{v}(k+1))$. The first term on the right-hand side of (34) can be lower bounded as

$$2c(k)\sum_{i=1}^{m}\Big(f_i(z_i(k))-f_i(\bar{v}(k+1))\Big)$$
$$\geq -2c(k)L\sum_{i=1}^{m}\|z_i(k)-\bar{v}(k+1)\|_2$$
$$\geq -2c(k)L\left(\sum_{i=1}^{m}(\|e_i(k+1)\|_2+\|x_i(k+1)-\bar{v}(k+1)\|_2)\right)$$
$$\geq -\alpha_2(k)\sum_{i=1}^{m}\|e_i(k+1)\|_2^2 - m\frac{L^2}{\alpha_2(k)}c(k)^2$$
$$-\ 2c(k)L\sum_{i=1}^{m}\|x_i(k+1)-\bar{v}(k+1)\|_2 \tag{35}$$

where the first inequality follows from the relation $x\geq -|x|$, for all $x\in\mathbb{R}$, and from item (*iii*) of Lemma 1, and the second inequality by adding and subtracting $x_i(k+1)$, for all $i=1,\dots,m$, and then using triangle inequality. The last inequality follows from $2xy\leq x^2+y^2$ with $x=\frac{L}{\sqrt{\alpha_2(k)}}c(k)$ and $y=\sqrt{\alpha_2(k)}\|e_i(k+1)\|_2$ for some $\alpha_2(k)\in(0,1)$, $k\in\mathbb{N}$. Substituting (35) into (34)

$$2c(k)\sum_{i=1}^{m} g_i(z_i(k))^T(z_i(k)-x^\star)$$
$$\geq -\alpha_2(k)\sum_{i=1}^{m}\|e_i(k+1)\|_2^2 - m\frac{L^2}{\alpha_2(k)}c(k)^2$$
$$-\ 2c(k)L\sum_{i=1}^{m}\|x_i(k+1)-\bar{v}(k+1)\|_2 \tag{36}$$

$$+\ 2c(k)\sum_{i=1}^{m}\Big(f_i(\bar{v}(k+1))-f_i(x^\star)\Big). \tag{37}$$

Substituting (28), (29), (32), (33), (37) into (22)

$$2c(k)\sum_{i=1}^{m}(f_i(\bar{v}(k+1))-f_i(x^\star))+\sum_{i=1}^{m}\|x_i(k+1)-x^\star\|_2^2$$
$$+\Big(1-\alpha_1(k)-\alpha_2(k)\Big)\sum_{i=1}^{m}\|e_i(k+1)\|_2^2$$
$$\leq\sum_{i=1}^{m}\|x_i(k)-x^\star\|_2^2+mL^2\Big(\frac{\alpha_1(k)+\alpha_2(k)}{\alpha_1(k)\alpha_2(k)}\Big)c(k)^2$$
$$+\ 6c(k)L\sum_{i=1}^{m}\|x_i(k+1)-\bar{v}(k+1)\|_2. \tag{38}$$

Summing (38) from $k=0$ to $k=N$, and using Lemma 4, item (*iii*), with $\bar{L}=3L$, the desired inequality (23) follows. This concludes the proof of item (*ii*).

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 $1-\beta_1-2\alpha>0$. Three immediate consequences of Lemma 4 are presented in the following proposition.

Proposition 2. *Consider Assumptions 1–3. The following statements hold*

(*i*) *We have that* $\sum_{k=0}^{\infty}\sum_{i=1}^{m}\|e_i(k)\|_2^2<\infty$;
(*ii*) *For all* $i=1,\dots,m$, *we have that* $\lim_{k\to\infty}\|e_i(k)\|_2=0$;
(*iii*) *For all* $i=1,\dots,m$, $\lim_{k\to\infty}\|x_i(k)-v(k)\|_2=0$.

Proof. Item (*i*): Consider Lemma 4, item (*ii*). Note that $\sum_{k=0}^{N}\sum_{i=1}^{m}\|x_i(k+1)-x^\star\|_2$ and $\sum_{k=0}^{N}\sum_{i=1}^{m}\|x_i(k)-x^\star\|_2$ form a telescopic series, so they can be replaced by $\sum_{i=1}^{m}\|x_i(N+1)-x^\star\|_2$ and $\sum_{i=1}^{m}\|x_i(0)-x^\star\|_2$, respectively. Let $\beta_1\in(0,1)$, choose $\alpha_1(k)=\alpha_2(k)=\alpha$ so that $1-2\alpha-\beta_1>0$. Observe that $\sum_{i=1}^{m}(f_i(\bar{v}(k+1))-f_i(x^\star))\geq 0$ for all $k\in\mathbb{N}$, due to optimality of $x^\star$, so this term can be dropped from (23). Besides, we can also drop the term $\sum_{i=1}^{m}\|x_i(N+1)-x^\star\|_2^2\geq 0$ since it is non-negative and appears on the left-hand side of (23). This yields

$$(1-2\alpha-\beta_1)\sum_{k=0}^{N}\sum_{i=1}^{m}\|e_i(k+1)\|_2^2\leq\sum_{i=1}^{m}\|x_i(0)-x^\star\|_2^2$$
$$+\left(mL^2\frac{2}{\alpha}+\beta_2\right)\sum_{k=0}^{N}c(k)^2+\beta_3.$$

Letting $N \to \infty$, we conclude that $\sum_{k=0}^{\infty}\sum_{i=1}^{m}\|e_i(k)\|_2^2$ is finite since the sequence $(c(k))_{k\in\mathbb{N}}$ is square-summable under Assumption 3 and the feasible set is compact. This concludes the proof of item (*i*).

Item (*ii*): Follows directly from item (*i*).

Item (*iii*): This proof follows directly from the arguments presented in Margellos et al. (2018, Proposition 3), and is omitted for brevity.

A.5. Proof of Theorem 1

We are now in a position to prove Theorem 1. To this end, we use the inequality (38) and leverage on Lemma 3.4 in Bertsekas and Tsitsiklis (1996) to establish convergence of the sequences $(\|x_i(k)-x^\star\|_2)_{k\in\mathbb{N}}$, $i=1,\dots,m$, to zero for some minimiser $x^\star$ of (1). We first present Lemma 3.4 in Bertsekas and Tsitsiklis (1996).

Lemma 5 (*Bertsekas & Tsitsiklis, 1996*)**.** *Consider non-negative scalar sequences $(\ell(k))_{k\in\mathbb{N}}$, $(u(k))_{k\in\mathbb{N}}$ and $(\zeta(k))_{k\in\mathbb{N}}$ that satisfy the recursion $\ell(k+1)\le\ell(k)-u(k)+\zeta(k)$. If $\sum_{k=0}^{\infty}\zeta(k)<\infty$, then the sequence $(\ell(k))_{k\in\mathbb{N}}$ converges and the sequence $(u(k))_{k\in\mathbb{N}}$ is summable.*

Consider inequality (38), and choose $\alpha_1(k),\alpha_2(k)$ and β_1 as in the proof of Proposition 2 item (*i*). We now drop the term involving $(1-2\alpha)\sum_{i=1}^{m}\|e_i(k+1)\|_2^2$ as it appears on the left-hand side of the inequality and is non-negative so that we obtain

$$\begin{aligned}\sum_{i=1}^{m}\|x_i(k+1)-x^\star\|_2^2 &\le \sum_{i=1}^{m}\|x_i(k)-x^\star\|_2^2\\ &-2c(k)\sum_{i=1}^{m}(f_i(\bar v(k+1))-f_i(x^\star))+\frac{2mL^2}{\alpha}c(k)^2\\ &+6c(k)L\sum_{i=1}^{m}\|x_i(k+1)-\bar v(k+1)\|_2. \end{aligned}\tag{39}$$

With reference to Lemma 5 and considering inequality (39), we set $\ell(k)=\sum_{i=1}^{m}\|x_i(k)-x^\star\|_2^2$, and

$$\begin{aligned}\zeta(k)&=\frac{2mL^2}{\alpha}c(k)^2+6c(k)L\sum_{i=1}^{m}\|x_i(k+1)-\bar v(k+1)\|_2,\\ u(k)&=2c(k)\big(f(\bar v(k+1))-f(x^\star)\big).\end{aligned}\tag{40}$$

By Lemma 3, item (iii), with $\bar L=3L$, and by Proposition 2, item (i), it follows that $6L\sum_{k=1}^{\infty}c(k)\sum_{i=1}^{m}\|x_i(k+1)-\bar v(k+1)\|<\infty$, hence, $\sum_{k=1}^{\infty}\zeta(k)<\infty$, as $c(k)$ is square-summable due to Assumption 3, which implies that the assumptions of Lemma 5 hold.

Therefore, we have that the sequence $(\sum_{i=1}^{m}\|x_i(k)-x^\star\|_2^2)_{k\in\mathbb{N}}$ converges, which implies that $(\sum_i\|x_i(k)-x^\star\|_2)_{k\in\mathbb{N}}$ also converges. To see this, note that, by continuity of the square-root function, $(\sum_{i=1}^{m}\|x_i(k)-x^\star\|_2^2)_{k\in\mathbb{N}}$ being a convergent sequence implies that $(\|X(k)-x^\star\otimes\mathbf{1}^T\|_F)_{k\in\mathbb{N}}$ also converges, where, for a fixed $k\in\mathbb{N}$, $X(k)$ is a $n\times m$ matrix whose i-th column is given by $x_i(k)$, and $\otimes$ represents the Kronecker product. Moreover, note that the set of $n\times m$ matrices can be equipped with the norm $\sum_{i=1}^{m}\|x_i\|_2$, where x_i, $i=1,\dots,m$, is the i-th column of a generic element $X\in\mathbb{R}^{n\times m}$. Since all norms in finite-dimensional spaces are equivalent, we conclude that the sequence $(\sum_{i=1}^{m}\|x_i(k)-x^\star\|_2)_{k\in\mathbb{N}}$ also converges. An alternative but more tedious justification of this argument can be found in Margellos et al. (2018).

By Lemma 5, we also have that $\sum_{k=1}^{\infty}c(k)\big(f(\bar v(k+1))-f(x^\star)\big)<\infty$.The latter implies 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) there exists some minimiser $x^\star$ such that a subsequence of $(\|\bar v(k)-x^\star\|_2)_{k\in\mathbb{N}}$ converges to zero. Moreover, we obtain $\sum_{i=1}^{m}\|x_i(k)-x^\star\|_2\le\sum_{i=1}^{m}\|\bar v(k)-x^\star\|_2+\mu\sum_{i=1}^{m}\|x_i(k)-v(k)\|_2$. by adding and subtracting $\bar v(k)$, then applying triangle inequality and invoking Lemma 3, item (*i*).

Note that $(\|\bar v(k)-x^\star\|_2)_{k\in\mathbb{N}}$ converges to zero across a subsequence and $(\sum_{i=1}^{m}\|x_i(k)-v(k)\|_2)_{k\in\mathbb{N}}$ converges to zero (due to Proposition 2, item (*iii*)) hence we can find a subsequence of $(\sum_{i=1}^{m}\|x_i(k)-x^\star\|_2)_{k\in\mathbb{N}}$ that converges to zero. However, we have shown by means of Lemma 5 that the sequence $(\sum_{i=1}^{m}\|x_i(k)-x^\star\|_2)_{k\in\mathbb{N}}$ converges; as a result it should converge to zero since every Cauchy sequence has a unique limit point. To conclude the proof, note that, for all $k\in\mathbb{N}$ and for all $j=1,\dots,m$, $\|x_j(k)-x^\star\|_2\le\sum_{i=1}^{m}\|x_i(k)-x^\star\|_2$, so we conclude that the sequences $(\|x_j(k)-x^\star\|_2)_{k\in\mathbb{N}}$, $j=1,\dots,m$, converge to zero. This concludes the proof.

A.6. Proof of Theorem 2

Consider Assumption 4. We drop the constant η for simplicity of exposition, but general choices $\frac{\eta}{\sqrt{k+1}}$, $\eta>0$, are also applicable. Let $(\hat v(k))_{k\in\mathbb{N}}$ be the running average sequence associated with $(\bar v(k))_{k\in\mathbb{N}}$ (definition is analogous to $(\hat x_i(k))_{k\in\mathbb{N}}$ in (4)). Note that since $\cap_{i=1}^{m}X_i$ is assumed to be convex, we have that $\hat v(k)$ is feasible for all $k\in\mathbb{N}$ (see also the discussion below (20)). We have that

$$\begin{aligned}\left|\sum_{i=1}^{m}f_i(\hat x_i(k+1))-f(x^\star)\right| &\le f(\hat v(k+1))-f(x^\star)\\ &+L\sum_{i=1}^{m}\|\hat x_i(k+1)-\hat v(k+1)\|_2,\end{aligned}\tag{41}$$

which follows from triangle inequality and Lemma 1, item (*iii*). Note that the first term on the right-hand side of (41) does not involve an absolute value due to feasibility of the sequence $(\hat v(k))_{k\in\mathbb{N}}$, which in turn implies that $f(\hat v(k+1))\ge f(x^\star)$.

To facilitate subsequent statements, we change the notation in Lemma 4, item (*ii*), by replacing k by r, and N by k. The inequality with this modified notation is repeated here for clarity. Indeed, we have that for all $k\in\mathbb{N}$

$$\begin{aligned}&2\sum_{r=0}^{k}c(r)\sum_{i=1}^{m}(f_i(\bar v(r+1))-f_i(x^\star))\\ &+\sum_{r=0}^{k}(1-\alpha_1(r)-\alpha_2(r)-\beta_1)\sum_{i=1}^{m}\|e_i(r+1)\|_2^2\\ &+\sum_{r=0}^{k}\sum_{i=1}^{m}\|x_i(r+1)-x^\star\|_2^2\le\sum_{r=0}^{k}\sum_{i=1}^{m}\|x_i(r)-x^\star\|_2^2\\ &+\sum_{r=0}^{k}\Big(mL^2\frac{\alpha_1(r)+\alpha_2(r)}{\alpha_1(r)\alpha_2(r)}+\beta_2\Big)c(r)^2+\beta_3,\end{aligned}\tag{42}$$

where $(\alpha_1(r))_{r\in\mathbb{N}}$ and $(\alpha_2(r))_{r\in\mathbb{N}}$ are sequences such that $1-\beta_1-\alpha_1(r)-\alpha_2(r)\ge0$ for all $r\in\mathbb{N}$.

The proofs of items (*i*), (*ii*) and (*iii*) of Theorem 2 are intertwined and will be composed into two parts: we first assume that there exist constants $d_1,d_2,d_3,d_4>0$ such that (43) and (44) below are satisfied, and on this basis prove the claims of the theorem; we then return to (43) and (44), and prove the existence of such constants. To this end, consider

$$f(\hat v(k+1))-f(x^\star)\le d_1\frac{1}{S(k+1)}+d_2\frac{\sum_{r=0}^{k}c(r)^2}{S(k+1)}\tag{43}$$

$$L\sum_{i=1}^{m}\|\hat{x}_i(k+1)-\hat{v}(k+1)\|_2 \le \frac{d_3}{S(k+1)}+d_4\frac{\sum_{r=0}^{k}c(r)^2}{S(k+1)}. \quad (44)$$

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

$$\begin{aligned}S(k+1)&=\sum_{r=1}^{k+1}\frac{1}{\sqrt{r+1}}\ge\int_2^{k+3}\frac{1}{\sqrt{x}}dx\\&=2(\sqrt{k+3}-\sqrt{2})\ge\nu\sqrt{k+3}\ge\nu\sqrt{k+1},\end{aligned} \quad (45)$$

with $\nu=2-\sqrt{2}$, and where we employed monotonicity of $\frac{\sqrt{x+3}-\sqrt{2}}{\sqrt{x+1}}$ for $x\ge 1$. Moreover, we have that

$$\begin{aligned}\sum_{r=0}^{k}c(r)^2&=\sum_{r=0}^{k}\frac{1}{r+1}=\sum_{r=1}^{k+1}\frac{1}{r}\\&\le\int_1^{k+1}\frac{1}{x}dx+1\le\ln(k+1)+1.\end{aligned} \quad (46)$$

The result of Theorem 2, item (*iii*), follows then from (41) by substituting (43)–(46), and setting $B_1=\sum_{i=1}^{4}\frac{d_i}{\nu}$ and $B_2=\frac{d_2}{\nu}+\frac{d_4}{\nu}$. Since (44) is valid for all $i=1,\dots,m$, we have that (via a direct application of triangle inequality) $\|\hat{x}_i(k)-\hat{x}_j(k)\|_2\le\sum_{i=1}^{m}\|\hat{x}_i(k)-\hat{v}(k)\|+\sum_{i=1}^{m}\|\hat{x}_j(k)-\hat{v}(k)\|$, which due to (45) and (46) then implies that the sequence $(\|\hat{x}_i(k)-\hat{x}_j(k)\|_2)_{k\in\mathbb{N}}$ converges to zero at a rate $\mathcal{O}(\frac{\ln k}{\sqrt{k}})$. This concludes the proof of item (*i*).

Moreover, these relations also imply that the set of accumulation points of the sequence $(\hat{v}(k))_{k\in\mathbb{N}}$ coincides to that of the sequences $(\hat{x}_i(k))_{k\in\mathbb{N}}$, $i=1,\dots,m$. Hence, we conclude that all accumulation points of $(\hat{x}_i(k))_{k\in\mathbb{N}}$ are feasible due to the fact that all accumulation points of $(\hat{v}(k))_{k\in\mathbb{N}}$ are in $\cap_{i=1}^{m}X_i$ and the latter is a closed set, thus concluding the proof of item (*ii*). This concludes the proof of Theorem 2.

Derivation of (43)

We first construct an upper-bound for the term on the left-hand side of (43). In fact, observe that

$$\begin{aligned}f(\hat{v}(k+1))-f(x^\star)&=f\left(\frac{1}{S(k+1)}\sum_{r=1}^{k+1}c(r)\bar{v}(r)\right)-f(x^\star)\\&\le\sum_{r=1}^{k+1}\frac{c(r)}{S(k+1)}f(\bar{v}(r))-f(x^\star)\\&=\sum_{r=0}^{k}\frac{c(r+1)}{S(k+1)}\sum_{i=1}^{m}(f_i(\bar{v}(r+1))-f_i(x^\star))\\&\le\sum_{r=0}^{k}\frac{c(r)}{S(k+1)}\sum_{i=1}^{m}(f_i(\bar{v}(r+1))-f_i(x^\star)),\end{aligned} \quad (47)$$

where the first equality follows by definition of $\hat{v}(k+1)$, the first inequality by convexity of f, the second equality by using the fact that $f=\sum_{i=1}^{m}f_i$ and changing the summation index, and the second inequality by using the fact that $c(r+1)=\frac{1}{\sqrt{r+1}}\le\frac{1}{\sqrt{r}}=c(r)$ for all $r\in\mathbb{N}$.

In light of (42), for 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 $\alpha(k)=a\left(1-\frac{1}{\sqrt{k+1}}\right)$; to ensure that $1-\beta_1-\alpha_1(k)-\alpha_2(k)\ge 0$ as required by Lemma 4, item (*ii*), it suffices to set $a=(1-\beta_1)/2$. Under these choices we have that

$$1-\beta_1-2\alpha(k)=\frac{1-\beta_1}{\sqrt{k+1}}=(1-\beta_1)c(k). \quad (48)$$

Consider now (42) with the above choices for $\alpha_1(k)$ and $\alpha_2(k)$. Note that the series $\sum_{r=0}^{k}\sum_{i=1}^{m}\|x_i(r+1)-x^\star\|_2$ and $\sum_{r=0}^{k}\sum_{i=1}^{m}\|x_i(r)-x^\star\|_2$ are telescopic, thus all intermediate terms cancel. We now drop the terms involving $\|e_i(r+1)\|_2^2$ and $\|x_i(k+1)-x^\star\|_2$ as they are non-negative, 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 right-hand side of (47)

$$\begin{aligned}&\sum_{r=0}^{k}\frac{c(r)}{S(k+1)}\sum_{i=1}^{m}(f_i(\bar{v}(r+1))-f_i(x^\star))\\&\le\frac{\sum_{i=1}^{m}\|x_i(0)-x^\star\|_2^2}{2S(k+1)}+\frac{\beta_3}{2S(k+1)}\\&+\frac{\beta_2}{2}\sum_{r=0}^{k}\frac{c(r)^2}{S(k+1)}+mL^2\frac{1}{S(k+1)}\sum_{r=0}^{k}\frac{c(r)^2}{\alpha(r)}.\end{aligned} \quad (49)$$

By the right-hand side of (49), we obtain (43) with $d_1=\frac{4mD^2+\beta_3}{2}$, $d_2=\frac{\beta_2}{2}+\frac{4mL^2}{a}$. where, by Assumption 1, $\sum_{i=1}^{m}\|x_i(0)-x^\star\|_2^2\le 4mD^2$, with D defined as in Lemma 3, item (*i*). Moreover, we used the fact that $\frac{c(r)^2}{\alpha(r)}=\frac{1}{a}\frac{\sqrt{r+1}}{\sqrt{r+1}-1}\frac{1}{r+1}\le\frac{4}{a}c(r)^2$, due to monotonicity of $\frac{\sqrt{x+1}}{\sqrt{x+1}-1}$.

Derivation of (44)

Similarly to the derivation of (43), we apply the definition of both $\hat{x}_i(k)$, $i=1,\dots,m$, and $\hat{v}(k)$ to upper-bound the left-hand side of (44) as

$$\begin{aligned}&L\sum_{i=1}^{m}\|\hat{x}_i(k+1)-\hat{v}(k+1)\|_2\\&=L\sum_{i=1}^{m}\left\|\frac{1}{S(k+1)}\sum_{r=1}^{k+1}c(r)\Big(x_i(r)-\bar{v}(r)\Big)\right\|_2\\&\le\frac{L\mu}{S(k+1)}\sum_{r=1}^{k+1}c(r)\sum_{i=1}^{m}\|x_i(r)-v(r)\|_2,\end{aligned} \quad (50)$$

where the inequality follows from convexity of the norm. We will now construct an upper-bound on the right-hand side of (50). To this end, note that

$$\begin{aligned}&\frac{L\mu}{S(k+1)}\sum_{r=1}^{k+1}c(r)\sum_{i=1}^{m}\|x_i(r)-v(r)\|_2\\&=\frac{L\mu c(1)}{S(k+1)}\sum_{i=1}^{m}\|x_i(1)-v(1)\|_2\\&+\frac{L\mu}{S(k+1)}\sum_{r=2}^{k+1}c(r)\sum_{i=1}^{m}\|x_i(r)-v(r)\|_2.\end{aligned} \quad (51)$$

We now invoke Lemma 3, item (*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 (51) so that

$$\begin{aligned}&\sum_{r=2}^{k+1}c(r)\sum_{i=1}^{m}\|x_i(r)-v(r)\|_2\\&=\sum_{r=1}^{k}c(r+1)\sum_{i=1}^{m}\|x_i(r+1)-v(r+1)\|_2\\&\le 2\sum_{r=0}^{k}c(r)\sum_{i=1}^{m}\|e_i(r+1)\|_2+m\lambda\sum_{i=1}^{m}\|x_i(0)\|_2\sum_{r=0}^{k}c(r)q^r\\&+m\lambda\sum_{r=1}^{k}c(r+1)\sum_{t=0}^{r-1}q^{r-t-1}\sum_{i=1}^{m}\|e_i(t+1)\|_2\end{aligned} \quad (52)$$

where we added the term corresponding to $r = 0$ and used the fact that $c(r+1) \leq c(r)$ for all $r \in \mathbb{N}$, in first two terms on the right-hand side of (52). We analyse each term on the right-hand side of (52) separately. First, observe that

$$2\sum_{r=0}^{k} c(r) \sum_{i=1}^{m} \|e_i(r+1)\|_2 \leq \sum_{r=0}^{k} c(r)^2 + \sum_{i=1}^{m} \|e_i(r+1)\|_2^2, \tag{53}$$

using the identity $2xy \leq x^2 + y^2$. The intermediate term on the right-hand side of (52) can be manipulated to yield

$$m\lambda \sum_{i=1}^{m} \|x_i(0)\|_2 \sum_{r=0}^{k} c(r) q^r \leq \frac{m^2 \lambda D}{1-q}, \tag{54}$$

since $c(r) \leq 1$ for all $r \in \mathbb{N} \cup \{0\}$, $\|x_i(0)\|_2 \leq D$ (Lemma 1) for all $i = 1, \ldots, m$, and using the closed-form expression for the sum of geometric series as $q \in (0, 1)$. We deal with the last term in (52) in several steps. We start by expanding the terms to obtain

$$\begin{aligned}
&\sum_{r=1}^{k} c(r+1) \sum_{t=0}^{r-1} q^{r-t-1} \sum_{i=1}^{m} \|e_i(t+1)\|_2 \\
&= c(2) \sum_{i=1}^{m} \|e_i(1)\|_2 + c(3) \left(q \sum_{i=1}^{m} \|e_i(1)\|_2 \sum_{i=1}^{m} \|e_i(2)\|_2 \right) \\
&+ \ldots + c(k+1) \left(\sum_{t=1}^{k} q^{k-t} \sum_{i=1}^{m} \|e_i(t)\|_2 \right).
\end{aligned} \tag{55}$$

We now collect the terms containing the error vector $e_i(r)$, $r = 1, \ldots, k$, to obtain

$$\begin{aligned}
&m\lambda \sum_{r=1}^{k} c(r+1) \sum_{t=0}^{r-1} q^{r-t-1} \sum_{i=1}^{m} \|e_i(t+1)\|_2 \\
&= m\lambda \sum_{i=1}^{m} \|e_i(1)\|_2 \Big(c(2) + qc(3) + \cdots \\
&+ q^{k-1} c(k+1) \Big) + \cdots + \sum_{i=1}^{m} \|e_i(k)\|_2 c(k+1) \\
&\leq \frac{m\lambda}{1-q} \sum_{r=1}^{k} c(r+1) \sum_{i=1}^{m} \|e_i(r)\|_2 \\
&\leq \frac{m\lambda}{1-q} \sum_{r=1}^{k} c(r) \sum_{i=1}^{m} \|e_i(r)\|_2 \leq \frac{m\lambda}{2(1-q)} \sum_{r=0}^{k} c(r)^2 \\
&\quad + \frac{m\lambda}{2(1-q)} \sum_{r=0}^{k} \sum_{i=1}^{m} \|e_i(r+1)\|_2^2
\end{aligned} \tag{56}$$

where in the first inequality we used the fact that $q \leq \frac{1}{1-q}$ and $1 \leq \frac{1}{1-q}$ for any $q \in (0, 1)$, while in the second inequality we used the fact that $c(r+1) \leq c(r)$. To obtain the last inequality we applied the relation $2xy \leq x^2 + y^2$ with $x = c(r)$ and $y = \|e_i(r+1)\|_2$, and then added the non-negative terms involving $c(0)^2$ and $\sum_{i=1}^{m} \|e_i(k+1)\|_2^2$. Substituting (51)–(54) and (56) into (50) we have that

$$\begin{aligned}
&L \sum_{i=1}^{m} \|\hat{x}_i(k+1) - \hat{v}(k+1)\|_2 \\
&\leq L\mu \left(1 + \frac{m\lambda}{2(1-q)} \right) \frac{\sum_{r=0}^{k} c(r)^2}{S(k+1)} \\
&\quad + \left(m\lambda + 2c(1) \right) \frac{L\mu m D}{S(k+1)} \\
&\quad + \frac{L\mu}{S(k+1)} \left(1 + \frac{m\lambda}{2(1-q)} \right) \sum_{r=0}^{k} \sum_{i=1}^{m} \|e_i(r+1)\|_2^2.
\end{aligned} \tag{57}$$

To obtain the result, we need to manipulate the last term on the right-hand side of (57). To this end, we invoke (42) with the same β_1 as in (48), but with $(\alpha_1(k))_{k\in\mathbb{N}}$ and $(\alpha_2(k))_{k\in\mathbb{N}}$ such that $\alpha_1(k) = \alpha_2(k) = \alpha$, for all $k \in \mathbb{N}$, following the same rationale as in Proposition 2 to obtain

$$\begin{aligned}
\sum_{r=0}^{k} \sum_{i=1}^{m} \|e_i(r+1)\|_2^2 &\leq \frac{\sum_{i=1}^{m} \|x(0) - x^\star\|_2^2 + \beta_3}{1 - \beta_1 - 2\alpha} \\
&+ \frac{1}{1 - \beta_1 - 2\alpha} \left(mL^2 \frac{2}{\alpha} + \beta_2 \right) \sum_{r=0}^{k} c(r)^2 \\
&\leq \frac{4mD^2 + \beta_3}{1 - \beta_1 - 2\alpha} \\
&+ \frac{1}{1 - \beta_1 - 2\alpha} \left(mL^2 \frac{2}{\alpha} + \beta_2 \right) \sum_{r=0}^{k} c(r)^2.
\end{aligned} \tag{58}$$

Substituting (58) into (57) we obtain (44) with constants

$$d_3 = L\mu \left[\left(1 + \frac{m\lambda}{2(1-q)} \right) \frac{4mD^2 + \beta_3}{1 - \beta_1 - 2\alpha} + mD \left(m\lambda + 2c(1) \right) \right],$$

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

thus concluding the proof of Theorem 2.

References

Baingana, B., Mateos, G., & Giannakis, G. B. (2014). Proximal-gradient algorithms for tracking cascades over social networks. *IEEE Journal of Selected Topics in Signal Processing*, *8*(4), 563–575.

Bertsekas, D. P. (2009). *Convex optimization theory*. Athena Scientific.

Bertsekas, D. P., & Tsitsiklis, J. N. (1989). *Parallel and Distributed Computation: Numerical Methods*. Athena: Scientific.

Bertsekas, D. P., & Tsitsiklis, J. (1996). *Neuro-dynamic programming*. Athena Scientific.

Bianchi, P. (2016). Ergodic convergence of a stochastic proximal point algorithm. *SIAM Journal on Optimization*, *26*(4), 2235–2260.

Bolognani, S., Carli, R., Cavraro, G., & Zampieri, S. (2015). Distributed reactive power feedback control for voltage regulation and loss minimization. *IEEE Transactions on Automatic Control*, *60*(4), 966–981.

Chen, A. I., & Ozdaglar, A. (2012). A fast distributed proximal-gradient method. In *50th annual allerton conference on communication, control, and computing* (pp. 601–608).

Duchi, J. C., Agarwal, A., & Wainwright, M. J. (2012). Dual averaging for distributed optimization: Convergence analysis and network scaling. *IEEE Transactions on Automatic Control*, *57*(3), 592–606.

Jakovetic, D., Moura, J. M., & Xavier, J. (2012). Distributed Nesterov-like gradient algorithms. In *51st IEEE conference on decision and control* (pp. 5459–5464).

Johansson, B., Keviczky, T., Johansson, M., & Johansson, K. H. (2008). Subgradient methods and consensus algorithms for solving convex optimization problems. *Proceedings of the IEEE Conference on Decision and Control*, (5), 4185–4190.

Lee, S., & Nedić, A. (2013). Distributed random projection algorithm for convex optimization. *IEEE Journal on Selected Topics in Signal Processing*, *7*(2), 221–229.

Lei, J., Chen, H.-F., & Fang, H.-T. (2016). Primal–dual algorithm for distributed constrained optimization. *Systems & Control Letters*, *96*, 110–117.

Liang, S., Wang, L., & Yin, G. (2019). Distributed quasi-monotone subgradient algorithm for nonsmooth convex optimization over directed graphs. *Automatica*, *101*, 175–181.

Lin, P., Ren, W., & Song, Y. (2016). Distributed multi-agent optimization subject to nonidentical constraints and communication delays. *Automatica*, *65*, 120–131.

Liu, S., Qiu, Z., & Xie, L. (2017). Convergence rate analysis of distributed optimization with projected subgradient algorithm. *Automatica*, *83*, 162–169.

Mai, V. S., & Abed, E. H. (2019). Distributed optimization over directed graphs with row stochasticity and constraint regularity. *Automatica*, *102*, 94–104.

Margellos, K., Falsone, A., Garatti, S., & Prandini, M. (2018). Distributed constrained optimization and consensus in uncertain networks via proximal minimization. *IEEE Transactions on Automatic Control*, *63*(5), 1372–1387.

Martinez, S., Bullo, F., Cortes, J., & Frazzoli, E. (2007). On synchronous robotic networks - Part I: Models, tasks and complexity. *IEEE Transactions on Automatic Control*, *52*(12), 2199–2213.

Mateos, G., & Giannakis, G. B. (2012). Distributed recursive least-squares: Stability and performance analysis. *IEEE Transactions on Signal Processing*, *60*(7), 3740–3754.

Nedić, A., & Olshevsky, A. (2015). Distributed optimization over time-varying directed graphs. *IEEE Transactions on Automatic Control*, *60*(3), 601–615.

Nedic, A., & Ozdaglar, A. (2009). Approximate primal solutions and rate analysis for dual subgradients methods. *SIAM Journal on Optimization*, *33*(5), 2295–2317.

Nedić, A., & Ozdaglar, A. (2009). Distributed subgradient methods for multi-agent optimization. *IEEE Transactions on Automatic Control*, *54*(1), 48–61.

Nedic, A., Ozdaglar, A., & Parrilo, P. A. (2010). Constrained consensus and optimization in multi-agent networks. *IEEE Transactions on Automatic Control*, *55*(4), 922–938.

Patrascu, A., & Necoara, I. (2018). Nonasymptotic convergence of stochastic proximal point methods for constrained convex optimization. *Journal of Machine Learning Research*, *18*, 1–42.

Qu, G., & Li, N. (2018). Harnessing smoothness to accelerate distributed optimization. *IEEE Transactions on Control of Network Systems*, *5*(3), 1245–1260.

Rockafellar, R. T. (1972). *Convex analysis*. Princeton University Press.

Romao, L., Margellos, K., Notarstefano, G., & Papachristodoulou, A. (2019). Convergence rate analysis of a subgradient averaging algorithm for distributed optimisation with different constraint sets. In *58th conference on decision and control* (pp. 7448–7453).

Scutari, G., & Sun, Y. (2019). Distributed nonconvex constrained optimization over time-varying digraphs. In *Mathematical programming* (pp. 497–544).

Shi, W., Ling, Q., Wu, G., & Yin, W. (2015). EXTRA: An exact first-order algorithm for decentralized consensus optimization. *SIAM Journal on Optimization*, *25*(2), 944–966.

Tsianos, K. I., Lawlor, S., & Rabbat, M. G. (2012). Push-sum distributed dual averaging for convex optimization. In *2012 IEEE 51st IEEE conference on decision and control* (pp. 5453–5458).

Tsitsiklis, J. N., Bertsekas, D. P., & Athans, M. (1986). Distributed asynchronous deterministic and stochastic gradient optimization algorithms. *IEEE Transactions on Automatic Control*, *31*(9), 803–812.

Wang, Y., Zhao, W., Hong, Y., & Zamani, M. (2019). Distributed subgradient-free stochastic optimization algorithm for nonsmooth convex functions over time-varying networks. *SIAM Journal on Control and Optimization*, *57*(4), 2821–2842.

Xi, C., & Khan, U. A. (2017). Distributed subgradient projection algorithm over directed graphs. *IEEE Transactions on Automatic Control*, *62*(8), 3986–3992.

Xie, P., You, K., Tempo, R., Song, S., & Wu, C. (2018). Distributed convex optimization with inequality constraints over time-varying unbalanced digraphs. *IEEE Transactions on Automatic Control*, *63*(12), 4331–4337.

Yuan, K., Ling, Q., & Yin, W. (2016). On the convergence of decentralized gradient descent. *SIAM Journal on Optimization*, *26*(5), 1835–1854.

Zanella, F., Varagnolo, D., Cenedese, A., Pillonetto, G., & Schenato, L. (2016). Newton–Raphson Consensus for distributed convex optimization. *IEEE Transactions on Automatic Control*, *61*(4), 994–1009.

Zhu, M., & Martinez, S. (2012). On distributed convex optimization under inequality and equality constraints. *IEEE Transactions on Automatic Control*, *57*(1), 151–164.

Licio Romao received the B.S. degree in Electrical Engineering from the Universidade Federal de Campina Grande, Brazil, in 2014, and M.S. degree in Electrical Engineering from the University of Campinas, Brazil, in 2017. He is currently pursuing the Ph.D. degree in Control Engineering at the University of Oxford.

He visited the University of Rome Tor Vergata in 2012, the University of California San Diego in 2015, and the University of Bologna in 2019. His research interests include optimisation algorithms and control strategies applied to large-scale, uncertain systems.

Kostas Margellos received the Diploma in Electrical Engineering from the University of Patras, Patras, Greece, in 2008, and the Ph.D. degree in control engineering from ETH Zurich, Zurich, Switzerland, in 2012.

He spent 2013, 2014 and 2015 as a Postdoctoral Researcher at ETH Zurich, UC Berkeley, and Politecnico di Milano, respectively. In 2016, he joined the Control Group, Department of Engineering Science, University of Oxford, Oxford, U.K., where he is currently an Associate Professor. He is also a Lecturer at Worcester College, Oxford. His research interests include optimisation and control of complex uncertain systems, with applications to generation and load side control for power networks.

Giuseppe Notarstefano is a Professor in the Department of Electrical, Electronic, and Information Engineering G. Marconi at Alma Mater Studiorum Università di Bologna. He was Associate Professor (June '16–June '18) and previously Assistant Professor, Ricercatore, (from Feb '07) at the Università del Salento, Lecce, Italy. He received the Laurea degree "summa cum laude" in Electronics Engineering from the Università di Pisa in 2003 and the Ph.D. degree in Automation and Operation Research from the Università di Padova in 2007. He has been visiting scholar at the University of Stuttgart, University of California Santa Barbara and University of Colorado Boulder. His research interests include distributed optimisation, cooperative control in complex networks, applied nonlinear optimal control, and trajectory optimisation and manoeuvring of aerial and car vehicles. He serves as an Associate Editor for IEEE Transactions on Automatic Control, IEEE Transactions on Control Systems Technology and IEEE Control Systems Letters. He has been part of the Conference Editorial Board of IEEE Control Systems Society and EUCA. He is recipient of an ERC Starting Grant.

Antonis Papachristodoulou FIEEE received the M.A./M.Eng. degrees in electrical and information sciences from the University of Cambridge, Cambridge, U.K., and the Ph.D. degree in Control and Dynamical systems (with a minor in aeronautics) from the California Institute of Technology, Pasadena, CA, USA. He is currently Professor of Engineering Science at the University of Oxford, Oxford, U.K., and a Tutorial Fellow at Worcester College, Oxford. He is also an EPSRC Fellow for Growth in Synthetic Biology and the Director of the EPSRC & BBSRC Centre for Doctoral Training in Synthetic Biology. His research interests include large-scale nonlinear systems analysis, sum of squares programming, synthetic and systems biology, networked systems, and flow control. Professor Papachristodoulou received the 2015 European Control Award for his contributions to robustness analysis and applications to networked control systems and systems biology. In the same year, he received the O. Hugo Schuck Best Paper Award.