Asynchronous Distributed Optimization via Dual Decomposition and Block Coordinate Subgradient Methods

Yankai Lin, *Student Member, IEEE*, Iman Shames, *Member, IEEE* and Dragan Nešić, *Fellow, IEEE*

***Abstract*— We study the problem of minimizing the sum of potentially non-differentiable convex cost functions with partially overlapping dependences in an asynchronous manner, where communication in the network is not coordinated. We study the behavior of an asynchronous algorithm based on dual decomposition and block coordinate subgradient methods under assumptions weaker than those used in the literature. At the same time, we allow different agents to use local stepsizes with no global coordination. Sufficient conditions are provided for almost sure convergence to the solution of the optimization problem. Under additional assumptions, we establish a sublinear convergence rate that in turn can be strengthened to linear convergence rate if the problem is strongly convex and has Lipschitz gradients. We also extend available results in the literature by allowing multiple and potentially overlapping blocks to be updated at the same time with non-uniform and potentially time varying probabilities assigned to different blocks. A numerical example is provided to illustrate the effectiveness of the algorithm.**

***Index Terms*— Distributed optimization, asynchronous algorithms, networked control systems**

I. INTRODUCTION

DISTRIBUTED optimization algorithms can be roughly classified as primal and dual methods. For primal methods, there is a large body of literature built upon the seminal work by Bertsekas and Tsitsiklis [4] where numerical computations are assigned to a group of processors (or agents) that exchange information with their neighbors. In [23], a distributed (sub)gradient method was proposed, which can be viewed as a first order primal method based on consensus theory. The consensus based algorithm is later extended to other settings including algorithms using constant stepsizes [26], stochastic networks [35], and algorithms with non-uniform stepsizes [19]. The existing dual methods are based on the dual decomposition idea. In [31], a synchronous distributed algorithm for convex problems over a fixed network with bounded communication delays is proposed. The case of time-varying networks is considered in [13]. This class of methods also include the distributed augmented Lagrangian method [14] and the alternating direction method of multipliers (ADMM) [9], [30], [32].

a) The Main Contributions: First, we study an algorithm based on the dual decomposition where agents solve a primal problem locally and independently. Then, they communicate these solutions to a subset of their neighbors whenever they finish the necessary computations and if communication is possible. Particularly, we consider solving optimization problems in an asynchronous environment where the notion of asynchrony is adopted from [4, Chapter 1.4]. Asynchrony refers to both independent computation and asynchronous communication. We investigate the conditions under which the algorithm converges under such asynchronous conditions. In this work, we propose an asynchronous dual supergradient algorithm and give sufficient conditions on the network and stepsizes to guarantee almost sure convergence of the algorithm. As our algorithm is based on dual decomposition, differentiability of the primal problem does not play a role in our analysis.

Second, under mild assumptions, we allow the use of local stepsizes where each agent chooses and updates its own stepsizes. Moreover, we assume that each component is updated infinitely often. This includes commonly used deterministic updating rules such as round robin and persistently exciting [28] in the same unifying framework. Since we assume each block is updated infinitely often, our main result still holds if multiple and overlapping blocks are updating at the same time.

Lastly, we prove our main convergence result using tools from asynchronous stochastic approximation theory [7], [25] which does not rely on strong convexity and Lipschitzness of the gradient of the cost function. To this end, we extend the results in [7, Chapter 7] to the non-smooth case which we believe is an important result in itself. Consequently, we prove our main result for the general case without assuming strict or strong convexity of the problem. This, inevitably, results in our convergence rate estimates to be slower than the results that consider the problem under more stringent conditions (e.g., the aforementioned strong convexity and Lipschitz conditions). However, it can be shown that, if stronger assumptions are imposed, we recover the linear convergence rate of the type shown in [33] as a special case.

This work was supported by the Australian Research Council under the Discovery Project DP170104099.

Y. Lin and D. Nešić are with the Department of Electrical and Electronic Engineering, The University of Melbourne, Parkville, 3010, Victoria, Australia (email: yankail@student.unimelb.edu.au; dnesic@unimelb.edu.au).

I. Shames is with the School of Engineering, The Australian National University, Acton, 0200, ACT, Australia (email:iman.shames@anu.edu.au).

b) Comparison to the Literature: For a recent survey paper on distributed optimization algorithms we refer the reader to [36]. The most relevant ones are those based on dual methods and asynchronous algorithms. In the consensus based primal gradient algorithm considered in [35], the asynchrony is captured by the random communication graph that is drawn i.i.d. from a set of graphs with proper weights to ensure almost sure convergence. Heterogeneous constant stepsizes are used in the asynchronous network under the assumption that the cost functions have Lipschitz gradients. If the cost functions are strongly convex, linear convergence rate can be established.

Other than gradient methods, augmented Lagrangian methods are also used in asynchronous distributed optimization [14]. It is assumed that each agent updates at least once in a finite number of steps [14, Assumption 2]. Another relevant work is [5] on asynchronous primal dual algorithms. The authors in [5] developed a stochastic primal-dual algorithm based on stochastic coordinate descent. Almost sure convergence of the proposed algorithm is established under the assumption of i.i.d. updates. In [31], the authors propose a dual algorithm for solving the optimal consensus algorithm in a synchronous fashion with diminishing stepsizes. In [24], the authors consider the dual algorithm of [31] in an asynchronous network for strongly convex functions with Lipschitz gradients. A dual block coordinate ascent algorithm is proposed with linear convergence rate under the assumption of i.i.d. updates.

In this work, we relax the i.i.d. assumption of the aforementioned works to infinitely often updates and allow agents to have their own local clocks and uncoordinated stepsizes. Consequently, we provide a framework that includes the deterministic update rules of the type presented in [14] and the stochastic update rules used in [11], [20], [35] as special cases. Moreover, [14, Assumption 3] requires that only one agent can update at each time instant, which is relaxed in our work. Additionally, unlike [31] and [24] where strict and strong convexity are assumed to ensure the dual function is differentiable and has a Lipschitz continuous gradient respectively, we prove almost sure convergence without appealing to the differentiability of the dual function. As a result, our dual convergence result holds for general convex primal cost functions, and we recover the result in [24] if strong convexity and i.i.d. updates are assumed. Lastly, compared to [20], we allow the use of uncoordinated local stepsizes where each agent updates its own stepsizes under mild assumptions. Moreover, we only assume that each component is updated infinitely often while in [20] it is assumed that each block component of the subgradient vector is computed with a positive probability at each step.

II. NOTATION AND BASIC DEFINITIONS

Let $\mathbb{R}$ be the set of real numbers and $\mathbb{R}^n$ be the n-dimensional Euclidean space, $\mathbb{R}_{\geq 0}$ (resp. $\mathbb{R}_{>0}$) be the set of non-negative (resp. positive) real numbers. For $x, y \in \mathbb{R}^n$, $\langle x, y\rangle$ denotes the inner product in $\mathbb{R}^n$. The set of non-negative (resp. positive) integers are denoted by $\mathbb{Z}_{\geq 0}$ (resp. $\mathbb{Z}_{>0}$) and the finite set of non-negative integers from 0 to $N \in \mathbb{Z}_{\geq 0}$ is denoted by $\mathbb{Z}_N$. Furthermore, $|\cdot|$ denotes the Euclidean norm of a vector $x \in \mathbb{R}^n$ and the corresponding induced norm of a matrix, respectively. The matrix I_n is used to denote the n-dimensional identity matrix and n will be dropped when the dimension is clear. The vector in $\mathbb{R}^N$ consisting of all ones is denoted by $\mathbf{1}_N$. Let C be a set, $|C|$ denotes its cardinality. The set $\mathbb{B}$ is the closed unit ball in $\mathbb{R}^n$ centred at the origin. For a closed set $S \subset \mathbb{R}^n$ and $\varepsilon > 0$, $S + \varepsilon\mathbb{B}$ is the set $\{x \in \mathbb{R}^n : |x|_S \leq \varepsilon\}$, where $|x|_S := \inf_{y\in S} |x - y|$. Set-valued mappings are indicated by the symbol $\rightrightarrows$. For a convex function $f(x) : \mathbb{R}^n \to \mathbb{R}$, a *subgradient* of $f(x)$ at $\bar{x}$ is a vector denoted by $g(\bar{x}) \in \mathbb{R}^n$ such that $f(x) \geq f(\bar{x}) + \langle g(\bar{x}), (x-\bar{x})\rangle$ for all $x \in \mathbb{R}^n$. The *subdifferential* of $f(x)$ at $\bar{x}$ denoted by $\partial f(\bar{x}) : \mathbb{R}^n \rightrightarrows \mathbb{R}^n$ is the set of all subgradients of $f(x)$ at $\bar{x}$. For a concave function $h(x) : \mathbb{R}^n \to \mathbb{R}$, a vector $g(\bar{x}) \in \mathbb{R}^n$ is a *supergradient* of $h(x)$ at $\bar{x}$ if $-g(\bar{x})$ is a subgradient of $-h(x)$. The *superdifferential* of $h(x)$ at $\bar{x}$ denoted by $\partial^+ h(\bar{x}) : \mathbb{R}^n \rightrightarrows \mathbb{R}^n$ is the set of all supergradients of $h(x)$ at $\bar{x}$. All random variables considered in this paper are defined on a probability space $(\Omega, \mathcal{F}, \mathbb{P})$, where Ω is the sample space, the σ-field $\mathcal{F}$ is the set of events and $\mathbb{P}$ is the probability measure defined on $(\Omega, \mathcal{F})$. For a vector sequence $x \in \mathbb{R}^n$, x^+ denotes its value at the next time instant, $x_{(i)}$ denotes the i^{th} component of x, $x_{[i]}$ denotes the value of x at time i, and $x_{(i),[j]}$ denotes the value of the i^{th} component of x at time j.

III. PROBLEM FORMULATION

We consider an optimization problem, where N agents are used to cooperatively minimize the sum of N individual cost functions that have partially overlapping dependences. The interconnection of the agents is described by an undirected graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$, where $\mathcal{V} = \{1, 2, ..., N\}$ is the set of nodes, $\mathcal{E} \subseteq \mathcal{V} \times \mathcal{V}$ is the set of edges. An edge $(i, j) \in \mathcal{E}$ means that agent i is able to access information from agent j. The (primal) problem we study in the work is given below

$$\begin{aligned} \min_{x\in\mathcal{X}\subseteq\mathbb{R}^n} \quad & F(x) := \sum_{i=1}^{N} F_i(x_{(i)}) \\ \text{s.t.} \quad & E_{ij}x_{(i)} - E_{ji}x_{(j)} = 0, \ \forall (i, j) \in \mathcal{E}, \end{aligned} \tag{1}$$

where $x_{(i)} \in \mathbb{R}^{n_i}$ is the local variable held by agent $i \in \mathcal{V}$, $x = (x_{(1)}, x_{(2)}, \ldots, x_{(N)}) := [x_{(1)}^T \ x_{(2)}^T \ \ldots \ x_{(N)}^T]^T \in \mathbb{R}^n$, $\sum_{i=1}^N n_i = n$ is the collection of all local variables, $F_i : \mathbb{R}^{n_i} \to \mathbb{R}$ is the individual cost function of agent i and the matrices E_{ij} and E_{ji} are selection matrices for agents i, j that characterize how the individual functions share common variables. The constraint set $\mathcal{X} = \mathcal{X}_1 \times \cdots \times \mathcal{X}_N$ is assumed to be convex. Compared to [31], we allow $\mathcal{X}$ to be a non-strict subset of $\mathbb{R}^n$ and non-compact. Let $x^* \in \mathcal{X}$ denote a solution to problem (1) and $F^* := F(x^*)$ be the optimal cost.

Remark 1: The problem (1) provides a framework for addressing a range of different problems and it can be shown that it is equivalent to the problem considered in [1]. Consider the optimal consensus problem of minimizing the sum of convex cost functions $\sum_{i=1}^N F_i(x)$, where $x \in \mathbb{R}^n$ is the decision variable. It can be equivalently re-formulated as

$$\min_{\bar{x}\in\mathbb{R}^{nN}} \quad F(\bar{x}) := \sum_{i=1}^{N} F_i(x_{(i)}) \tag{2}$$

$$\text{s.t.} \quad x_{(i)} - x_{(j)} = 0, \ \forall i, j \in \mathcal{V}.$$

It is clear that (2) is equivalent to (1) with all of the selection matrices E_{ij} being identity and $\mathcal{X} = \mathbb{R}^{nN}$. Moreover, some other problems such as distributed MPC, network flow optimization and resource allocation can also be re-formulated in the form of (1), e.g. see [1], [24]. ■

Remark 2: Our problem formulation is similar to the set-up in [17]. However, if the goal is to optimize a global decision variable then the graph should be connected to ensure convergence. ■

We now introduce the dual variables $\lambda_{(ij)}$, for agents $i, j \in \mathcal{V}$ that share common variables, which is associated with the constraint $E_{ij}x_{(i)} = E_{ji}x_{(j)}$. As a result, we use an edge-based communication structure by assigning the task of updating $\lambda_{(ij)}$ corresponding to the constraint $E_{ij}x_{(i)} - E_{ji}x_{(j)} = 0$ and edge (i,j) to agent i if $(i,j) \in \overline{\mathcal{E}} := \{(i,j) \in \mathcal{E} : i < j\}$ and to agent j otherwise. Additionally, we denote the set of neighbors of agent i that are not responsible for updating the corresponding dual vectors by $\mathcal{N}_i := \{j \in \mathcal{V} : (i,j) \in \overline{\mathcal{E}}\}$. Moreover, $\mathcal{I} := \{i \in \mathcal{V} : \mathcal{N}_i \neq \emptyset\}$ denotes the set of agents that are responsible for updating at least one dual vector [1]. Hence, agent $i \in \mathcal{I}$ has access to $(\ldots, \lambda_{(ij)}, \ldots) \in \mathbb{R}^{\bar{n}_i}, j \in \mathcal{N}_i$ and the full dual vector is given by: $\lambda = (\ldots, \lambda_{(ij)}, \ldots) \in \mathbb{R}^{\bar{n}}$ with $(i,j) \in \overline{\mathcal{E}}$. By definition, $\bar{n}_i = \sum_j \bar{n}_{ij} : j \in \mathcal{N}_i$ and $\bar{n} = \sum_i \bar{n}_i : i \in \mathcal{I}$. Here, $\bar{n}_{ij}$ is the dimension of the dual variable $\lambda_{(ij)}$ corresponding to the edge $(i,j) \in \overline{\mathcal{E}}$, $\bar{n}_i$ is the overall dimension of the dual variables assigned to agent $i \in \mathcal{I}$, and $\bar{n}$ is the dimension of the dual vector λ. It can be seen that the number of dual vectors is equal to $|\overline{\mathcal{E}}|$. The Lagrangian L is given by

$$L(x,\lambda) := \sum_{i=1}^{N} F_i(x_{(i)}) + \sum_{(i,j)\in\overline{\mathcal{E}}} \langle \lambda_{(ij)}, E_{ij}x_{(i)} - E_{ji}x_{(j)} \rangle. \tag{3}$$

The corresponding dual function is given by

$$Q(\lambda) := \inf \ L(x,\lambda), \quad \text{s.t.} \quad x \in \mathcal{X}. \tag{4}$$

Due to the finite sum structure of $F(x)$, (4) can be rewritten as $Q(\lambda) = \sum_{i=1}^{N} q_i(\lambda)$, where

$$q_i(\lambda) = \inf \ l_i(x_{(i)}, \lambda), \quad \text{s.t.} \quad x_{(i)} \in \mathcal{X}_i, \tag{5}$$

with $l_i(x_{(i)}, \lambda) := F_i(x_{(i)}) + \sum_{j:(i,j)\in\overline{\mathcal{E}}} \langle \lambda_{(ij)}, E_{ij}x_{(i)} \rangle - \sum_{j:(j,i)\in\overline{\mathcal{E}}} \langle \lambda_{(ji)}, E_{ij}x_{(i)} \rangle$. The functions $q_i(\lambda)$ depend on the dual variable λ only and the infimum is taken over local variables $x_{(i)}$. The set of minimizers of $l_i(x_{(i)}, \lambda)$ with respect to $x_{(i)}$ is denoted by $X_{i,\lambda} \subseteq \mathbb{R}^{n_i}$. This allows the agents to locally compute $q_i(\lambda)$. The dual problem is given by

$$\max_{\lambda \in \mathbb{R}^{\bar{n}}} \ Q(\lambda). \tag{6}$$

The set of maximizers of $Q(\lambda)$ is denoted by Λ_{opt}. The problem (6) can be solved in a standard distributed and asynchronous manner [10]. We make the following assumption.

Assumption 1: The following hold for Problem (1):

(i) Each individual cost function F_i is convex.
(ii) The constraint set $\mathcal{X}$ is such that (1) is strictly feasible.
(iii) The set $X_{i,\lambda} \subseteq \mathbb{R}^{n_i}$ is non-empty for any $\lambda \in \mathbb{R}^{\bar{n}}$.
(iv) The function $Q(\lambda)$ is radially unbounded: $Q(\lambda) \to -\infty$ as $|\lambda| \to \infty$. ■

The convexity of functions F_i in item (i) ensures that F is also convex. Thus, problem (1) is convex with linear equality constraints. This together with item (ii) of Assumption 1 ensures Slater's conditions hold [10]. As a result, $F^* = Q^* := Q(\lambda^*) = \max_{\lambda} Q(\lambda)$, for $\lambda^* \in \Lambda_{\text{opt}}$. Item (iii) of Assumption 1 ensures that each agent can solve its individual problem (5) and return a finite solution. For example, this is the case if $\mathcal{X} \subset \mathbb{R}^n$ is chosen to be compact, see [31] and [24]. Item (iv) of Assumption 1 requires Q to be radially unbounded ensuring that Λ_{opt} is a compact set. The compactness will play a key role in proving Theorems 1 and 2. See Remark 10 for a discussion on relaxing this assumption.

Remark 3: Differentiability of (1) does not play a role in our convergence analysis. This is common to many dual based algorithms analyses (see [31], [24] for example) where the convergence results are provided without taking the intermediate primal minimization steps. Moreover, the proof technique used in this work does not rely on the differentiability of the dual function. This enables us to show the convergence of the algorithm in the absence of strict convexity of $F(x)$ and compactness of $\mathcal{X}$ which are typically used to guarantee differentiability of the dual function in [31]. ■

Remark 4: From a practical point of view, in addition to convergence of the dual variable and the optimal cost F^*, one may be also interested in a solution to (1) that is primal feasible. If this is the case, extra care must be taken to recover optimal primal solutions after the dual problem is solved. This is because for a given $\lambda \in \Lambda_{\text{opt}}$, solving (5) may generate x that is primal-infeasible. To solve this problem, certain averaging techniques can be used as done in [22], [13] and [27]. The simplest way is to add a quadratic regularization term with a small coefficient so that the cost function is strictly convex and the regularization term has negligible effects on the solution to the original optimization problem [34]. ■

Remark 5: The last item of Assumption 1 is needed in proving convergence of the dual variable for our algorithm. It is commonly used in Lyapunov based proofs to conclude *global* stability of a dynamical system as it ensures the existence of a valid Lyapunov function in the proof of Theorem 1 (see Remark 10 for more discussion). ■

In dual algorithms, often the explicit form of the dual function is not available. As a result, checking (iv) of Assumption 1 *a priori* may be hard. Nevertheless, sufficient conditions for radial unboundedness of the dual function as described in the following lemma can be stated.

Lemma 1: Given λ, denote the minimizer of $l_i(x_{(i)}, \lambda)$ in $\mathcal{X}_i$ by $x^*_{\lambda(i)}$. Let for all i, $g_i(x^*_{\lambda(i)}) \in \partial F_i(x^*_{\lambda(i)})$ such that $g_i(x^*_{\lambda(i)}) + \sum_{j:(i,j)\in\overline{\mathcal{E}}} E_{ij}^T \lambda_{(ij)} - \sum_{j:(j,i)\in\overline{\mathcal{E}}} E_{ij}^T \lambda_{(ji)} = 0$. It is further assumed that $\forall (i,j) \in \mathcal{E}$ the constraints $E_{ij}x_{(i)} = E_{ji}x_{(j)}$ are linearly independent, and there exist a $\epsilon > 0$ and

[1] Depending on different ways of writing the equality constraint in (1), there are multiple ways of assigning the task of updating dual variable which leads to different communication graphs and different constructions of $\overline{\mathcal{E}}$. However, these different dual vector assignments result in problems mathematically equivalent to the one discussed in our paper and can be analysed using exactly the same approach used in our paper *mutatis mutandis*.

a non-empty subset of $\bar{\mathcal{X}}_{ij} \bigcap \bar{\mathcal{X}}_{ji}$ denoted by $\mathcal{M}_{ij}$ such that $\mathcal{M}_{ij} + \epsilon\mathbb{B} \subset \bar{\mathcal{X}}_{ij} \bigcap \bar{\mathcal{X}}_{ji}$ with $\bar{\mathcal{X}}_{ij} := \{x \in \mathbb{R}^{\bar{n}_{ij}} | x = E_{ij}y, y \in \mathcal{X}_i\}$ and $\mathbb{B} \subset \mathbb{R}^{\bar{n}_{ij}}$. Then, Q is radially unbounded. ■

Proof: For any $\bar{\lambda}$,

$$\begin{aligned}
&Q(\lambda) - Q(\bar{\lambda}) = \sum_{i=1}^{N} F_i(x^*_{\lambda(i)}) - \sum_{i=1}^{N} F_i(x^*_{\bar{\lambda}(i)}) \\
&+ \sum_{(i,j)\in\overline{\mathcal{E}}} \langle \lambda_{(ij)}, E_{ij}x^*_{\lambda(i)} - E_{ji}x^*_{\lambda(j)} \rangle \\
&- \sum_{(i,j)\in\overline{\mathcal{E}}} \langle \bar{\lambda}_{(ij)}, E_{ij}x^*_{\bar{\lambda}(i)} - E_{ji}x^*_{\bar{\lambda}(j)} \rangle \\
&\quad \le \sum_{i=1}^{N} g_i^T(x^*_{\lambda(i)})(x^*_{\lambda(i)} - x^*_{\bar{\lambda}(i)}) \\
&+ \sum_{j:(i,j)\in\overline{\mathcal{E}}} \langle \lambda_{(ij)}, E_{ij}x^*_{\lambda(i)} \rangle - \sum_{j:(j,i)\in\overline{\mathcal{E}}} \langle \lambda_{(ji)}, E_{ij}x^*_{\lambda(j)} \rangle \\
&- (\sum_{j:(i,j)\in\overline{\mathcal{E}}} \langle \bar{\lambda}_{(ij)}, E_{ij}x^*_{\bar{\lambda}(i)} \rangle - \sum_{j:(j,i)\in\overline{\mathcal{E}}} \langle \bar{\lambda}_{(ji)}, E_{ij}x^*_{\bar{\lambda}(j)} \rangle) \\
&\le - \sum_{i=1}^{N} g_i^T(x^*_{\lambda(i)})x^*_{\bar{\lambda}(i)} - C \\
&= - \sum_{(i,j)\in\overline{\mathcal{E}}} \langle \lambda_{(ij)}, E_{ij}x^*_{\bar{\lambda}(i)} - E_{ji}x^*_{\bar{\lambda}(j)} \rangle - C,
\end{aligned}$$

where the first inequality is due to the definition of subgradients, $C := \sum_{j:(i,j)\in\overline{\mathcal{E}}} \langle \bar{\lambda}_{(ij)}, E_{ij}x^*_{\bar{\lambda}(i)} \rangle - \sum_{j:(j,i)\in\overline{\mathcal{E}}} \langle \bar{\lambda}_{(ji)}, E_{ij}x^*_{\bar{\lambda}(j)} \rangle$, and the last inequality comes from the assumption $g_i(x^*_{\lambda(i)}) + \sum_{j:(i,j)\in\overline{\mathcal{E}}} E_{ij}^T\lambda_{(ij)} - \sum_{j:(j,i)\in\overline{\mathcal{E}}} E_{ij}^T\lambda_{(ji)} = 0$. Due to the assumption that $\mathcal{M}_{ij} + \epsilon\mathbb{B} \subset \bar{\mathcal{X}}_{ij} \bigcap \bar{\mathcal{X}}_{ji}$, the case where a component of λ is identically multiplied by 0 does not exist. It also ensures that there are elements in $\mathcal{X}$ such that the signs for all components of λ can be made negative or positive as one wishes by selecting appropriate $x \in \mathcal{X}$. The question now becomes for such $x \in \mathcal{X}$, can we find the corresponding $\bar{\lambda}$. It is obvious that, all linearly independent constraints can be put in a compact form of $Hx = 0$. Therefore, if for any x, we can find a corresponding $\bar{\lambda}$ such that $g(x) = -H^T\lambda$, where $g(x) \in \partial F(x)$, then the desired $x^*_{\bar{\lambda}(i)}$ always exists for any i. Indeed, since we assume that the constraints are linearly independent, H has full row rank and H^T has full column rank. Therefore, by choosing appropriate $\bar{\lambda}$, $-\sum_{(i,j)\in\overline{\mathcal{E}}} \langle \lambda_{(ij)}, E_{ij}x^*_{\bar{\lambda}(i)} - E_{ji}x^*_{\bar{\lambda}(j)} \rangle$ tends to $-\infty$ as $|\lambda|$ goes to infinity. Thus Q is radially unbounded. ■

IV. The Asynchronous Block Coordinate Subgradient Algorithm

In this section, we present an asynchronous block supergradient algorithm to solve the dual problem (6). First, let us re-write the Lagrangian (3) in a more compact form:

$$L(x, \lambda) := F(x) + \langle \lambda, \mathbf{E}(x) \rangle, \tag{7}$$

where the linear constraint $\mathbf{E}(x) := (\dots, \mathbf{E}_i(x), \dots)$ is given by the concatenation of matrices $\mathbf{E}_i(x) := (\dots, E_{ij}x_{(i)} - E_{ji}x_{(j)}, \dots), (i,j) \in \overline{\mathcal{E}}$ such that $i \in \mathcal{I}$. For $\lambda \in \mathbb{R}^{\bar{n}}$, let $x^*_\lambda = (x^*_{\lambda(1)}, x^*_{\lambda(2)}, \dots, x^*_{\lambda(N)})$ be an arbitrary real minimizer of the Lagrangian $L(x, \lambda)$, where $x^*_{\lambda(i)} \in \mathcal{X}_i, i \in \mathcal{V}$ are minimizers of $l_i(x_{(i)}, \lambda)$. We next state one way to compute a supergradient of $Q(\lambda)$. The proof can be found in standard textbooks, e.g., [3], and is omitted.

Lemma 2: Let $Q(\lambda)$ be the dual function (4), then for any $\lambda \in \mathbb{R}^{\bar{n}}$ and a corresponding $x^*_\lambda \in \mathcal{X}$, $\mathbf{E}(x^*_\lambda) \in \partial^+ Q(\lambda)$. ■

This implies $E_{ij}x^*_{\lambda(i)} - E_{ji}x^*_{\lambda(j)}$ is a component of the supergradient associated with the corresponding multiplier. The synchronous supergradient update for λ is given by

$$\lambda^+ = \lambda + \alpha g(\lambda) = \lambda + \alpha\mathbf{E}(x^*_\lambda), \tag{8}$$

where $\alpha > 0$ is the stepsize to be appropriately chosen and $g \in \partial^+ Q$. As a result, asynchronous version of the dual updates corresponds to the supergradient update (8) for some of its components $\lambda^+_{(ij)} = \lambda_{(ij)} + \alpha(E_{ij}x^*_{\lambda(i)} - E_{ji}x^*_{\lambda(j)})$, for some $(i,j) \in \overline{\mathcal{E}}$. Specifically, each agent minimizes $l_i(x_{(i)}, \lambda)$ with respect to $x_{(i)}$ locally. When a pair of agents finish solving their local problems, they can communicate their solutions to each other to perform a dual update without waiting for others to complete their computations. These communication times may depend on agents' computational capabilities, availability of the network and so on. Based on Lemma 2, the proposed dual asynchronous algorithm that generates the sequence $\{\lambda_{[k]}\}_{k=1}^\infty$ and $\{\alpha_{[k]}\}_{k=1}^\infty$ can be modelled by the following stochastic discrete-time system

$$\lambda^+ \in \lambda + \Phi(\alpha \odot v^+) \odot (\partial^+ Q(\lambda) + e^+) \tag{9a}$$
$$\alpha^+ = A(\alpha) \odot v^+ + \alpha \odot (\mathbf{1}_{|\overline{\mathcal{E}}|} - v^+) \tag{9b}$$
$$\gamma^+ = \gamma + v^+, \tag{9c}$$

where $\odot$ denotes the Hadamard product of two matrices, $\alpha = (\dots, \alpha_{(ij)}, \dots) \in \mathbb{R}^{|\overline{\mathcal{E}}|}_{>0}$, $(i,j) \in \overline{\mathcal{E}}$ is the stepsize vector, $\gamma = (\dots, \gamma_{(ij)}, \dots) \in \mathbb{R}^{|\overline{\mathcal{E}}|}_{\ge 0}$, $(i,j) \in \overline{\mathcal{E}}$ is the vector of the number of dual updates up to the current time instance and its initial condition is therefore set to 0. The function $A(\alpha) = (\dots, A_{(ij)}(\alpha_{(ij)}), \dots) \in \mathbb{R}^{|\overline{\mathcal{E}}|}_{>0}$, $(i,j) \in \overline{\mathcal{E}}$ is the vector representing the dynamics of local stepsize update. To model the asynchrony, we introduce the random input $v = (\dots, v_{(ij)}, \dots)$ such that $v_{(ij),[k+1]}$ is equal to 1 if at time k edge $(i,j) \in \overline{\mathcal{E}}$ is updating and is equal to 0 otherwise. Furthermore, $\Phi(\alpha \odot v^+) := (\dots, \alpha_{(ij)}v^+_{(ij)}\mathbf{1}_{\bar{n}_{ij}}, \dots)$. The other random input $e^+ = (\dots, e_{(ij)}, \dots)$ following some probability distribution satisfying Assumption 4 accounts for supergradient evaluation errors coming from communication noise and the fact that the local problems may not be solved exactly in finite time. For notational convenience, we re-label the vectors λ, α and γ in (9) and give a component-wise representation of (9). Note that there are $|\overline{\mathcal{E}}|$ dual updates and we label the corresponding update variables from 1 to $|\overline{\mathcal{E}}|$, i.e., $\lambda = (\lambda_{(1)}, \dots, \lambda_{(|\overline{\mathcal{E}}|)})$, $\alpha = (\alpha_{(1)}, \dots, \alpha_{(|\overline{\mathcal{E}}|)})$ and $\gamma = (\gamma_{(1)}, \dots, \gamma_{(|\overline{\mathcal{E}}|)})$. The mappings $A(\cdot)$ and $\partial^+ Q(\cdot)$ are partitioned similarly. For $i \in \mathbb{Z}_{|\overline{\mathcal{E}}|} \backslash \{0\}$, the i^{th} component of (9) can be written as

$$\lambda^+_{(i)} \in \lambda_{(i)} + \alpha_{(i)}v^+_{(i)}(\partial^+ Q(\lambda)_{(i)} + e^+_{(i)}) \tag{10a}$$
$$\alpha^+_{(i)} = A_{(i)}(\alpha_{(i)})v^+_{(i)} + \alpha_{(i)}(1 - v^+_{(i)}) \tag{10b}$$

$$\gamma^+_{(i)} = \gamma_{(i)} + v^+_{(i)}. \tag{10c}$$

An implementation of (9) is presented in Algorithm 1. We

Algorithm 1 Asynchronous Block Supergradient Algorithm

1: **Initialization**: $\lambda_{(ij),[0]}$ $\forall (i,j) \in \overline{\mathcal{E}}$.

2: **Primal Update**: At time k, for all agents $i \in \mathcal{V}$,

$$x^*_{(i)} \leftarrow \underset{x_{(i)} \in \mathcal{X}_i}{\arg\min}\ l_i(x_{(i)}, \lambda)$$

3: **Dual Update**: For all $(i,j) \in \overline{\mathcal{E}}$ such that $v_{(ij),[k+1]} = 1$:

$$\lambda^+_{(ij)} \leftarrow \lambda_{(ij)} + \alpha_{(ij)}(E_{ij}x^*_{(i)} - E_{ji}x^*_{(j)} + e^+_{(ij)}),$$
$$\alpha^+_{(ij)} \leftarrow A_{ij}(\alpha_{(ij)}).$$

For all $(i,j) \in \overline{\mathcal{E}}$ such that $v_{(ij),[k+1]} = 0$:

$$\lambda^+_{(ij)} \leftarrow \lambda_{(ij)}, \quad \alpha^+_{(ij)} \leftarrow \alpha_{(ij)}.$$

4: Set $k \leftarrow k+1$ and go to Step 2.

make the following assumption on the update protocol.

Assumption 2: For any $i \in \mathbb{Z}_{|\overline{\mathcal{E}}|}\backslash\{0\}$ and $m \in \mathbb{Z}_{\geq 0}$, there exists a $\Delta > 0$ such that $\liminf_{k\to\infty} \frac{\gamma_{(i),[k+m]}}{k+m} \geq \Delta$ almost surely, where $\gamma_{(i)}$ is generated by (10) with $\gamma_{(i),[0]} = 0$. ■

Remark 6: Assumption 2 requires that each coordinate is updated infinitely often. As shown in [6], it can be satisfied whenever each coordinate, given the past history, has an updating probability lower bounded by a positive constant [32]. Other examples of stochastic updating conditions satisfying Assumption 2 include the case where coordinates are updated in an i.i.d. fashion [20], or when the coordinates are updated based on a controlled Markov chain [7]. Two examples for deterministic rules satisfying Assumption 2 include cyclic updating of coordinates [33] and the case of uniformly persistently exciting updates [28] where each coordinate is updated at least once in N iterations for a given integer N [14]. ■

For $x \in \mathbb{R}$, let $\lfloor x \rfloor$ denote the maximum integer no larger than x. We present next the assumptions on agent's local stepsize.

Assumption 3: For any $i,j \in \mathbb{Z}_{|\overline{\mathcal{E}}|}\backslash\{0\}$ and stepsize sequences $\{\alpha_{(i)k}\}$ and $\{\alpha_{(j)k}\}$ generated by $\alpha^+_{(i)} = A_{(i)}(\alpha_{(i)})$ and $\alpha^+_{(j)} = A_{(j)}(\alpha_{(j)})$ respectively, there exist positive scalars $\bar{\delta}$ and $\bar{\Delta} > 0$ such that *(i)* $\sum_{k=0}^{\infty} \alpha_{(i),[k]} = \infty$ and $\sum_{k=0}^{\infty} \alpha^2_{(i),[k]} < \infty$; *(ii)* $\limsup_{k\to\infty} \sup_{\delta_1 \in [\delta_2,1]} \alpha_{(i),[\lfloor \delta_1 k \rfloor]}/\alpha_{(j),[k]} \leq \bar{\Delta}$, for some $\delta_2 \in (0,\Delta]$; *(iii)* $\liminf_{k\to\infty} \inf_{\delta_1 \in [\delta_2,1]} \alpha_{(i),[\lfloor \delta_1 k \rfloor]}/\alpha_{(j),[k]} \geq \bar{\delta}$, for some $\delta_2 \in (0,\Delta]$. ■

Remark 7: Since we are not working exclusively with Lipschitz continuous gradients, it is reasonable to use diminishing stepsizes to achieve almost sure convergence to the exact solution set Λ_{opt}. Item (ii) of Assumption 3 requires the stepsizes of agents in the network to be roughly in the same "time scale". Moreover, different from [20], we allow agents to use different local stepsizes whose updates are not required to be coordinated as long as they satisfy Assumption 3. They can be easily satisfied by choosing commonly used stepsizes including $\{c_i/k\}$ and $\{c_i/(k \log k)\}$ (with appropriate modifications when $k = 0$) where $c_i > 0$ is a bounded constant that agent $i \in \mathcal{I}$ can choose locally. ■

We denote the history of (9) and (10) up to time $k \in \mathbb{Z}_{\geq 0}$ by $\mathcal{F}_k := \sigma(\lambda_{[m]}, \alpha_{[m]}, e_{[m]}, v_{[m]}, m \in \mathbb{Z}_k)$, which is a σ-subfield of $\mathcal{F}$. The following assumption on the supergradient error states that the error is a martingale difference noise as shown later in the proof of Theorem 1 which is a standard assumption in the stochastic approximation literature [7].

Assumption 4: There exist a sequence $\{\beta_{[k]}\}_{k=0}^{\infty}$ and a $K \geq 0$ such that the random sequence of error vectors e in (9) satisfies $\mathbb{E}[e_{[k+1]}|\mathcal{F}_k] = \beta_{[k]}$ and $\mathbb{E}[|e_{[k+1]}|^2|\mathcal{F}_k] \leq K$. ■

Remark 8: Most stochastic approximation and optimization literature considers zero mean martingale difference noise under which exact almost sure convergence can be proved. However, in some applications, it is natural that the noise may not be zero mean. We introduce the variable $\beta_{[k]}$ to cover both the zero mean ($\beta_{[k]} = 0$) and non-zero mean ($\beta_{[k]} \neq 0$) cases. We prove that if $\beta_{[k]} = 0$ then we can establish the same convergence results as shown in the stochastic approximation literature. If $\beta_{[k]} \neq 0$, then we can show convergence to a neighborhood of Λ_{opt}. ■

Assumption 5: The dual variable $\lambda_{[k]}$ is bounded almost surely for any k. ■

Assumption 5 is standard in stochastic approximation literature [7] to guarantee almost sure convergence. In the next section, we will give a sufficient condition for satisfaction of Assumption 5. The main result of the paper is stated below.

Theorem 1: Let Assumptions 1–5 hold. If $\beta_{[k]} = 0$ for all $k \in \mathbb{Z}_{\geq 0}$, then $\lim_{k\to\infty} |\lambda_{[k]}|_{\Lambda_{\text{opt}}} = 0$ almost surely for any $(\lambda_{[0]}, \alpha_{[0]}, \gamma_{[0]}) \in \mathbb{R}^{\bar{n}} \times \mathbb{R}^{|\overline{\mathcal{E}}|}_{>0} \times \{0\}$. ■

The proof of Theorem 1 is inspired by the asynchronous stochastic approximation idea introduced in [7, Chapter 7] and [6] where similar results are established for the case where the cost function is differentiable. Before presenting it in detail, a brief outline of the proof followed by necessary intermediate results are provided below.

The proof generalises that of [7, Chapter 7, Theorem 2] to the non-smooth case. We first linearly interpolate the asynchronous dual updates to make it a "continuous time" signal. We then show that this continuous time signal almost surely converges to a solution of a non-autonomous differential inclusion (Proposition 3) with Λ_{opt} as its uniformly globally asymptotically stable (UGAS) positively invariant set in the sense of [21, Definition 2.2] (Proposition 4). For notational convenience, we define the radially unbounded convex function $f := -Q$ whose minimizers form a non-empty compact set (Assumption 1). Hence, for $i \in \mathcal{I}$, (10) can be written as

$$\lambda^+_{(i)} = \lambda_{(i)} - \alpha_{(i)}v^+_{(i)}g(\lambda)_{(i)} + \alpha_{(i)}v^+_{(i)}e^+_{(i)} \tag{11a}$$
$$\alpha^+_{(i)} = A_{(i)}(\alpha_{(i)})v^+_{(i)} + \alpha_{(i)}(1 - v^+_{(i)}) \tag{11b}$$
$$\gamma^+_{(i)} = \gamma_{(i)} + v^+_{(i)}, \tag{11c}$$

where $g(\lambda) \in \partial f(\lambda)$, and is calculated by taking the difference of the two corresponding primal solutions of local problems and can be any element of $\partial f(\lambda)$.

Note that for each i, $\lambda_{(i)}$ in (11) can be written as

$$\lambda^+_{(i)} = \lambda_{(i)} - \bar{\alpha}\tau_{(i)}(g(\lambda)_{(i)} - e^+_{(i)}), \tag{12}$$

where $\bar{\alpha} := \max_{i\in\mathcal{I}_a} \alpha_{(i)}$, $\tau_{(i)} := \frac{\alpha_{(i)}}{\bar{\alpha}} v^+_{(i)}$ and $\mathcal{I}_a \subseteq \mathbb{Z}_{|\bar{\mathcal{E}}|}\backslash\{0\}$ is the set of components that are updating (we define $\bar{\alpha}$ and $\tau_{(i)}$ to be 0 when $\mathcal{I}_a = \emptyset$). The next lemma shows that the sequence $\{\bar{\alpha}_{[k]}\}$ is square summable but not summable.

Lemma 3: The sequence $\{\bar{\alpha}_{[k]}\}$ satisfies $\sum_{k=0}^{\infty} \bar{\alpha}^2_{[k]} < \infty$ and $\sum_{k=0}^{\infty} \bar{\alpha}_{[k]} = \infty$ almost surely. ■

Let

$$t_{[k]} := \sum_{m=0}^{k-1} \bar{\alpha}_{[m]},\ k \geq 1,\ t_{[0]} = 0, \tag{13}$$

and for $t \in [t_{[k]}, t_{[k+1]})$

$$\bar{\lambda}(t) := \lambda_{[k]} + \frac{\lambda_{[k+1]} - \lambda_{[k]}}{\bar{\alpha}_{[k]}}(t - t_{[k]}),\ \bar{\lambda}(t_{[k]}) = \lambda_{[k]}. \tag{14}$$

By construction, the linearly interpolated process $\bar{\lambda}(t)$ is continuous and piecewise linear. Let $\theta(t) := \mathrm{diag}(\tau_{(1),[k]}, \ldots, \tau_{(|\bar{\mathcal{E}}|),[k]})$, for $t \in [t_{[k]}, t_{[k+1]})$. Intuitively, the sum of $\tau_{(i)}$ indicates how frequently the agent i is updating. By Lemma 8 (see Appendix), $f(\lambda)$ is guaranteed to be locally Lipschitz. Thus by Rademacher's theorem [12], $f(\lambda)$ is differentiable almost everywhere. Hence, for almost all $\lambda \in \mathbb{R}^{\bar{n}}$, $\nabla f(\lambda)$ is well-defined, continuous, and $\partial f(\lambda) = \{\nabla f(\lambda)\}$. Therefore, $g(\lambda)$ is continuous almost everywhere. Equivalently, by defining $r(t) := \max\{k \in \mathbb{Z}_{\geq 0} : t_{[k]} \leq t\}$:

$$\begin{aligned}\bar{\lambda}(t_{[k]} + u) - \bar{\lambda}(t_{[k]}) = &-\int_0^u \theta(t_{[k]} + s) g(\lambda_{r(t_{[k]}+s)}) ds \\ &+ \zeta_{r(t_{[k]}+u)} - \zeta_{r(t_{[k]})},\end{aligned} \tag{15}$$

where $\zeta_{[r(t_{[k]})]} = \zeta_{[k]} := \sum_{m=0}^{k-1} \bar{\alpha}_{[m]} \theta(t_{[m]}) e_{[m+1]}, k \geq 1$. Compared to $\bar{\lambda}(t)$, which is piecewise linear and continuous, $\lambda_{[r(t)]}$ is piecewise constant and may be discontinuous at $t = t_{[k]}$. From Assumption 4, it can be seen that $(\zeta_{[k]}, \mathcal{F}_k)$ is a martingale process. Moreover, the following proposition states that if $\beta_{[k]} = 0$ for all k, the effect of the term $\zeta_{r(t_{[k]}+u)} - \zeta_{r(t_{[k]})}$ in (15) diminishes as $k \to \infty$.

Proposition 1: If $\beta_{[k]} = 0$ for all k, then $\lim_{k\to\infty} \zeta_{[k]} < \infty$ and the sequence of functions of u, $\{\zeta_{[r(t_{[k]}+u)]} - \zeta_{[r(t_{[k]})]}\}_{k=1}^{\infty}$ converges to 0 on $[0, \infty)$ almost surely. ■

Next, by defining $\phi(t + u, t, \bar{\lambda}(t)) := \bar{\lambda}(t) - \int_0^u \theta(t + s) g(\bar{\lambda}(t + s)) ds$, we have the following result which shows $\phi(t + u, t, \bar{\lambda}(t))$ can be used to approximate $\bar{\lambda}(t + u)$ when t is sufficiently large.

Proposition 2: For any $T > 0$, $\lim_{t\to\infty} \sup_{u\in[0,T]} |\bar{\lambda}(t + u) - \phi(t + u, t, \bar{\lambda}(t))| = 0$ almost surely. ■

The behavior of the limit point of $\{\bar{\lambda}(t_{[k]} + \cdot)\}$ is studied next.

Proposition 3: For any positive scalar T and positive integer $\bar{n}$, any limit point $\tilde{\lambda}(\cdot)$ of $\{\bar{\lambda}(t_{[k]} + \cdot)\}$ defined in (14) in the space of continuous functions from $[0, T]$ to $\mathbb{R}^n$ denoted by $C([0, T]; \mathbb{R}^{\bar{n}})$ is almost surely a solution to the non-autonomous differential inclusion

$$\dot{\tilde{\lambda}} \in -\Theta(t) \partial f(\tilde{\lambda}), \tag{16}$$

where $\Theta(t)$ is a diagonal matrix-valued measurable function with diagonal entries in $[\varepsilon, 1]$, with $0 < \varepsilon \leq 1$. ■

Lastly, the following result states stability properties of the limiting differential inclusion (16) which also ensures that no finite escape time exists for (16) by compactness of Λ_{opt}.

Proposition 4: Suppose $\Theta(t)$ is a diagonal matrix-valued measurable function with diagonal entries in $[\varepsilon, 1]$, then (16) is UGAS with respect to the compact set Λ_{opt}. ■

Now, we are ready to prove Theorem 1.

Proof of Theorem 1: By Lemma 3, $t_{[k]} \to \infty$ (defined in (13)) as $k \to \infty$. Thus, $\bar{\lambda}(t)$ is defined over all $t \in [0, \infty)$. From (14), the limit $\lambda_{[k]}$ as $k \to \infty$ exists if and only if the limit of $\bar{\lambda}(t)$ as $t \to \infty$ exists. These limits are the same. By Lemma 8, $\partial f(\lambda)$ is locally bounded. Moreover, since the sequence $\{\lambda_{[k]}\}_{k=0}^{\infty}$ generated by (11) is almost surely bounded, $g(\lambda_{[k]})$ is almost surely bounded for all $g(\lambda_{[k]}) \in \partial f(\lambda_{[k]})$, $k \in \mathbb{Z}_{\geq 0}$. Furthermore, $\theta(t)$ is uniformly bounded for all $t \in [0, \infty)$. Combining these facts and Proposition 1, it can be seen that, almost surely, $\bar{\lambda}(t_{[k]} + u) - \bar{\lambda}(t_{[k]})$ in (15) is norm bounded. Hence, the sequence of continuous time processes $\{\bar{\lambda}(t_{[k]} + \cdot)\}_{k=0}^{\infty}$ is equicontinuous (see Preliminaries in the appendix for the definition) and bounded. Note that, for a given k, $\bar{\lambda}(t_{[k]} + \cdot)$ is a function defined on $[0, \infty)$ shifted by $t_{[k]}$, thus the sequence of functions $\{\bar{\lambda}(t_{[k]} + \cdot)\}_{k=0}^{\infty}$ are in fact defined on the same interval, but shifted differently. Then by the Arzelà-Ascoli Theorem (see Lemma 5 in Appendix), for any $T > 0$, $\{\bar{\lambda}(t_{[k]} + \cdot)\}_{k=0}^{\infty}$ is relatively compact in $C([0, T]; \mathbb{R}^{\bar{n}})$. Thus, there exists a subsequence of $\{\bar{\lambda}(t_{[k]} + \cdot)\}_{k=0}^{\infty}$ that converges to a limit point. Following similar arguments, it can be shown that, for $t \geq 0$, $\{\phi(t_{[k]} + \cdot, t_{[k]}, \bar{\lambda}(t_{[k]}))\}$ is also an equicontinuous, point-wise bounded family of trajectories. Therefore, it follows from Proposition 2 that $\{\phi(t_{[k]} + \cdot, t_{[k]}, \bar{\lambda}(t_{[k]}))\}$ and $\{\bar{\lambda}(t_{[k]} + \cdot)\}$ have the same limit points in $C([0, T]; \mathbb{R}^{\bar{n}})$. As a result, studying the limiting behavior of $\{\bar{\lambda}(t_{[k]} + \cdot)\}$ can be done via looking at the corresponding common subsequence of $\{\phi(t_{[k]} + \cdot, t_{[k]}, \bar{\lambda}(t_{[k]}))\}$. By Propositions 2 and 3, almost surely, all limit points of the interpolated process (14) are solutions of (16) which is UGAS with respect to Λ_{opt} as shown in Proposition 4. Thus, from [7, Chapter 5, Corollary 4], $\lambda_{[k]}$ in (11) almost surely converges to Λ_{opt}. ■

By imposing the following additional assumption on the stepsize sequence, we are able to state almost sure convergence of $\lambda_{[k]}$ to a neighborhood of Λ_{opt} if $\beta_{[k]}$ is norm-bounded.

Assumption 6: For any $i, j \in \mathbb{Z}_{|\bar{\mathcal{E}}|}\backslash\{0\}$, the stepsize sequences $\{\alpha_{(i),[k]}\}$ and $\{\alpha_{(j),[k]}\}$ where $\alpha^+_{(i)} = A_{(i)}(\alpha_{(i)})$ and $\alpha^+_{(j)} = A_{(j)}(\alpha_{(j)})$ satisfy the following:

(i) $(\exists c_{ij} > 0)$ $\lim_{n\to\infty} \frac{\sum_{k=0}^{n} \alpha_{(i),[k]}}{\sum_{k=0}^{n} \alpha_{(j),[k]}} = c_{ij}$;

(ii) $(\exists b_1 \in (0, \Delta])$ $(\forall b_2 \in [b_1, 1])$ $(\exists c \in (0, 1])$ $\lim_{n\to\infty} \frac{\sum_{k=0}^{\lfloor b_2 n\rfloor} \alpha_{(i),[k]}}{\sum_{k=0}^{n} \alpha_{(i),[k]}} = c$. ■

Theorem 2: Let Assumptions 1–6 hold. Suppose that for any $\delta_e > 0$ and $\Delta_e > 0$ there exists $\varepsilon_e > 0$, such that $|\beta_{[k]}| \leq \varepsilon_e$ and the sequence $\{\lambda_{[k]}\}_{k=1}^{\infty}$ generated by (9) is bounded almost surely and $\lambda_{[k]} \in \Delta_e \mathbb{B}$ infinitely often. Then $\lim_{k\to\infty} |\lambda_{[k]}|_{\Lambda_{\mathrm{opt}}+\delta_e\mathbb{B}} = 0$ almost surely for any $(\lambda_{[0]}, \alpha_{[0]}, \gamma_{[0]}) \in \mathbb{R}^{\bar{n}} \times \mathbb{R}^{|\bar{\mathcal{E}}|}_{>0} \times \{0\}$. ■

Proof: Note that, we can split the non-zero-mean error into two terms with one being the martingale difference noise considered in Theorem 1 and the other one being a noise term bounded by ε_e. Following similar arguments in the proof of Theorem 1, it can be seen that the limiting differential inclusion becomes $\dot{\tilde{\lambda}} \in -\Theta(t)(\partial f(\tilde{\lambda}) + \varepsilon_e \mathbb{B})$. Since there exists $c_{ij} > 0$ such that $\lim_{n\to\infty} \frac{\sum_{k=0}^{n} \alpha_{(i),[k]}}{\sum_{k=0}^{n} \alpha_{(j),[k]}} = c_{ij}$, the stepsize sequences used by agents belong to the class of balanced stepsizes defined in [6, Section 3]. Thus, following the arguments in [6, Theorem 3.2], $\Theta(t) = \Theta$ and, hence, the deterministic limiting differential inclusion becomes

$$\dot{\tilde{\lambda}} \in -\Theta(\partial f(\tilde{\lambda}) + \varepsilon_e \mathbb{B}), \tag{17}$$

which is autonomous. By Lemma 8 (see Appendix), $\partial f(\tilde{\lambda})$ is outer semicontinuous, convex and locally bounded[2]. Thus, the continuous time system $\dot{\tilde{\lambda}} \in -\Theta\partial f(\tilde{\lambda})$ satisfies [15, Assumption 6.5] and is well-posed. Since $\Theta(\partial f(\tilde{\lambda}) + \varepsilon_e \mathbb{B}) \subseteq \Theta(\partial f(\tilde{\lambda})) + \varepsilon_e \mathbb{B}$ and Λ_{opt} is a compact set by Assumption 1, the proof of Theorem 2 is completed by applying [16, Theorem 6.6]. The result is based on semi-global practical stability of (17) and the recurrence condition on the set $\Delta_e \mathbb{B}$ is needed since the stability property is not global. ■

Remark 9: Since the compact set Λ_{opt} is UGAS, it follows from [15, Theorem 7.21] that Λ_{opt} is also robustly globally asymptotically stable. Namely, there exists a continuous function $\varsigma : \mathbb{R}^{\bar{n}} \to \mathbb{R}_{\geq 0}$, such that Λ_{opt} is UGAS for $\dot{\tilde{\lambda}} \in -\overline{\text{co}}(\Theta\partial f(\tilde{\lambda} + \varsigma(\tilde{\lambda})\mathbb{B})) + \varsigma(\tilde{\lambda})\mathbb{B}$, where for a set S, $\overline{\text{co}}S$ denotes the closed convex hull of S. The function ρ can be used to cover some uncertainties commonly seen in applications such as quantization errors, measurement noise. Moreover, a smooth Lyapunov function is guaranteed to exist by a converse theorem given in [15] which can also be used as a tool for further robustness analysis. ■

Remark 10: If, in addition to balanced stepsizes [6, Section 3], identical stepsize sequences are used, it can be shown that $\Theta = I$ [7, Chapter 7]. In this case one can use $|\lambda|^2_{\Lambda_{\text{opt}}}$ as a Lyapunov function in the proof of Theorem 1 without relying on radial unboundedness of the dual function. ■

V. Additional Results

We first discuss a special case where an almost sure convergence of $\lambda_{[k]}$ in (9) without requiring the iteration to be almost surely bounded is obtained. Then, some convergence rates analysis under more restrictive assumptions are provided.

A. Bounded Supergradient

Almost sure bounded iteration assumption is only used to ensure that $g(\cdot)$ in (15) is also almost surely bounded. Thus, it can be established that $\{\bar{\lambda}(t_{[k]} + \cdot)\}$ is equicontinuous and bounded. Thus, if we find another way to ensure the boundedness of $g(\cdot)$, then we can show almost sure convergence without imposing the condition that the iteration is almost surely bounded. One possible approach is to assume that the set $\mathcal{X}$ is compact. In most situations, this is a reasonable assumption since the decision variables are meaningful only when they are finite. The boundedness of the supergradient of $Q(\lambda)$ comes from the fact that the supergradient in the algorithm is formed by the difference between two elements in $\mathcal{X}$ which is bounded.

Corollary 1: Under Assumptions 1–4, if $\mathcal{X}$ is a compact set and $\beta_{[k]} = 0$ for all $k \in \mathbb{Z}_{\geq 0}$, then the sequence $\{\lambda_{[k]}\}_{k=1}^{\infty}$ generated by (9) satisfies $\lim_{k\to\infty} |\lambda_{[k]}|_{\Lambda_{\text{opt}}} = 0$ almost surely, for any $(\lambda_0, \alpha_0, \gamma_0) \in \mathbb{R}^{\bar{n}} \times \mathbb{R}^{|\bar{\mathcal{E}}|}_{>0} \times \{0\}$. ■

Remark 11: The assumption on the compactness of $\mathcal{X}$ is a simplifying assumption to ensure the almost sure boundedness of the dual variables which is a crucial element of our proof as described in Section IV. This is in contrast with [31] where compactness is to guarantee a differentiable dual function. We do not rely on the differentiability of the dual function in the analysis and the compactness assumption can be replaced by any assumption that results in the boundedness of the dual variable (see [8] for example). Investigating other sufficient conditions for boundedness of λ is left for future work. ■

B. Convergence Under Stronger Update Assumptions

In this part, we give another set of sufficient conditions to achieve convergence based on [20]. We show that under an assumption on the update rule that is stronger than Assumption 2 and global stepsizes[3], the algorithm almost surely converges if all other assumptions in Theorem 1 are satisfied and there is an appropriate upper bound on the supergradient of the dual function during updates.

Instead of Assumption 2, we make the following assumption that can be satisfied if edges are updated in an i.i.d. fashion:

Assumption 7: Considering system (9), for any $(i, j) \in \bar{\mathcal{E}}$, $\mathbb{P}(v_{(ij),[k]} = 1|\mathcal{F}_{k-1}) = p_{ij} > 0$ for all $k \geq 1$. ■

Note that, Assumption 7 is sufficient but not necessary to ensure Assumption 2 as it rules out situations where deterministic protocols are used. In addition to Assumption 7, we assume the following upper bound on the supergradient of the dual function Q.

Assumption 8: There exists a scalar c, such that $c^2(1 + |\lambda_{[k]}|^2_{\Lambda_{\text{opt}}}) \geq |g(\lambda_{[k]})|^2$, where $g(\lambda) \in \partial^+ Q(\lambda)$, $\forall k \in \mathbb{Z}_{\geq 0}$, almost surely. ■

Remark 12: This assumption is weaker than both the bounded subgradient assumption in [31] and the Lipschitz continuous gradient assumption in [24]. For example, it can be satisfied if $\mathcal{X}$ is compact. Moreover, if the cost is quadratic which can be commonly seen in linear MPC related applications, then the gradient will be linear which also satisfies Assumption 8. ■

The following lemma will be used to establish almost sure convergence of λ in (9):

Lemma 4: [3, Proposition A.4.5] Let $\mathcal{F}_k$ be the history of (9) and (10) up to time k. In addition, let $p_{[k]}$, $a_{[k]}$, $w_{[k]}$ and $u_{[k]}$ be sequences of non-negative random variables and let the following relation hold almost surely for any $k \in \mathbb{Z}_{\geq 0}$

$$\mathbb{E}[p_{[k+1]}|\mathcal{F}_k] \leq (1 + a_{[k]})p_{[k]} - u_{[k]} + w_{[k]}, \tag{18}$$

[2]A set-valued mapping $M : \mathbb{R}^p \rightrightarrows \mathbb{R}^n$ is outer semicontinuous if for each $(x_i, y_i) \to (x, y) \in \mathbb{R}^p \times \mathbb{R}^n$ satisfying $y_i \in M_i(x_i)$ for all $i \in \mathbb{Z}_{\geq 0}$, $y \in M(x)$. It is locally bounded if for each bounded $K \subset \mathbb{R}^p$, $M(K) := \bigcup_{x\in K} M(x)$ is bounded.

[3]We require all dual updates to use the same stepsize. It can be local in situations where all updating edges share a common positive end at each time.

where $\sum_{k=0}^{\infty} a_{[k]} < \infty$ and $\sum_{k=0}^{\infty} w_{[k]} < \infty$ almost surely. Then almost surely $p_{[k]}$ will converge to some random variable p and $\sum_{k=0}^{\infty} u_{[k]} < \infty$. ■

Proposition 5: Consider Assumptions 1, 4, 7, and 8. Let $\beta_{[k]} = 0$ for all $k \in \mathbb{Z}_{\geq 0}$ and all edges use the same stepsize sequence $\{\alpha_{[k]}\}$ satisfying $\sum_{k=0}^{\infty} \alpha_{[k]} = \infty$ and $\sum_{k=0}^{\infty} \alpha_{[k]}^2 < \infty$ during dual updates. Then $\lambda_{[k]}$ generated by (9) almost surely converges to a point in Λ_{opt} as $k \to \infty$ for any $(\lambda_{[0]}, \alpha_{[0]}, \gamma_{[0]}) \in \mathbb{R}^{\bar{n}} \times \mathbb{R}_{>0}^{|\bar{\mathcal{E}}|} \times \{0\}$. ■

Proof: For any $\lambda^* \in \Lambda_{\text{opt}}$, and $g \in \partial^+ Q$:

$$\begin{aligned}
&|\lambda_{[k+1]} - \lambda^*|^2 = |\lambda_{[k+1]} - \lambda_{[k]} + \lambda_{[k]} - \lambda^*|^2 \\
&= |\lambda_{[k]} - \lambda^*|^2 + |\alpha_{[k]} \odot (g(\lambda_{[k]}) + e_{[k+1]}) \odot v_{[k+1]}|^2 \\
&\quad + 2\langle \alpha_{[k]} \odot (g(\lambda_{[k]}) + e_{[k+1]}) \odot v_{[k+1]}, \lambda_{[k]} - \lambda^* \rangle.
\end{aligned}$$

Since $\mathbb{E}[e_{[k+1]}|\mathcal{F}_k] = 0$ and $\mathbb{E}[|e_{[k+1]}|^2|\mathcal{F}_k] \leq K$, taking expectation of $|\lambda_{[k+1]} - \lambda^*|^2$ conditioned on past history yields

$$\begin{aligned}
&\mathbb{E}[|\lambda_{[k+1]} - \lambda^*|^2|\mathcal{F}_k] = \mathbb{E}[|\alpha_{[k]} \odot e_{[k+1]} \odot v_{[k+1]}|^2|\mathcal{F}_k] \\
&\quad + |\lambda_{[k]} - \lambda^*|^2 + \mathbb{E}[|\alpha_{[k]} \odot g(\lambda_{[k]}) \odot v_{[k+1]}|^2|\mathcal{F}_k] \\
&\quad + 2\mathbb{E}[\langle \alpha_{[k]} \odot g(\lambda_{[k]}) \odot v_{[k+1]}, \lambda_{[k]} - \lambda^* \rangle|\mathcal{F}_k] \\
&\leq |\lambda_{[k]} - \lambda^*|^2 + \alpha_{[k]}^2 K + +\alpha_{[k]}^2 c^2 (1 + |\lambda_{[k]} - \lambda^*|^2) \\
&\quad + 2\alpha_{[k]} \langle W g(\lambda_{[k]}), \lambda_{[k]} - \lambda^* \rangle \\
&= (1 + \alpha_{[k]}^2 c^2)|\lambda_{[k]} - \lambda^*|^2 + \alpha_{[k]}^2 (K + c^2) \\
&\quad + 2\alpha_{[k]} \langle W g(\lambda_{[k]}), \lambda_{[k]} - \lambda^* \rangle, \qquad (19)
\end{aligned}$$

where $W := \operatorname{diag}(\ldots, p_{ij} I, \ldots) \in \mathbb{R}^{\bar{n}} \times \mathbb{R}^{\bar{n}}$ is the block diagonal matrix that contains p_{ij} for all $(i,j) \in \bar{\mathcal{E}}$ and is positive definite due to Assumption 7. Thus, $W^{\frac{1}{2}}$ and $W^{-\frac{1}{2}}$ are well-defined. Define $m_1 := |W^{-\frac{1}{2}}|^2$, then

$$\begin{aligned}
&\mathbb{E}[|W^{-\frac{1}{2}}(\lambda_{[k+1]} - \lambda^*)|^2|\mathcal{F}_k] \leq \\
&(1 + \alpha_{[k]}^2 c^2)|W^{-\frac{1}{2}}(\lambda_{[k]} - \lambda^*)|^2 + \alpha_{[k]}^2 (|W^{-\frac{1}{2}}|^2 K + c^2) \\
&+ 2\alpha_{[k]} \langle W^{\frac{1}{2}} g(\lambda_{[k]}), W^{-\frac{1}{2}}(\lambda_{[k]} - \lambda^*) \rangle \\
&\leq (1 + \alpha_{[k]}^2 c^2)|W^{-\frac{1}{2}}(\lambda_{[k]} - \lambda^*)|^2 + \alpha_{[k]}^2 (m_1 K + c^2) \\
&- 2\alpha_{[k]} (Q(\lambda^*) - Q(\lambda_{[k]}))
\end{aligned}$$

Since $\alpha_{[k]}$ is square summable, we can apply Lemma 4 by taking $p_{[k]} = |W^{-\frac{1}{2}}(\lambda_{[k]} - \lambda^*)|^2$, $a_{[k]} = \alpha_{[k]}^2 c^2$, $u_{[k]} = 2\alpha_{[k]}(Q(\lambda^*) - Q(\lambda_{[k]}))$ and $w_{[k]} = \alpha_{[k]}^2 (m_1 K + c^2)$ to arrive at the conclusion that p_k almost surely converges to a random variable which means the algorithm is almost surely bounded. To see this, define S_A as the set of all bounded sequences in a metric space and S_B as the set of all convergent sequences in the same metric space. A convergent sequence in a metric space is bounded. Thus, $S_B \subseteq S_A$. Consider a proper probability measure P such that $P(S_B) \leq P(S_A)$. Since $P(S_B) = 1$, we must have $1 = P(S_B) \leq P(S_A) \leq 1$ which means $P(S_A) = 1$. The rest of the proof follows from the proof of Theorem 1. ■

C. Convergence Rates

Since we focus on asynchronous algorithms with diminishing stepsizes, one may not expect a convergence rate that is as competitive as those that assume strongly convex functions with Lipschitz gradient. Nevertheless, we are able to state the following result under the hypotheses of Proposition 5, whose proof follows from the proof of Proposition 1 in [20].

Proposition 6: Consider Assumptions 1, 4, 7, and 8. Let $\beta_{[k]} = 0$ for all $k \in \mathbb{Z}_{\geq 0}$ and all edges use the same stepsize sequence $\{\alpha_{[k]}\}$ satisfying $\sum_{k=0}^{\infty} \alpha_{[k]} = \infty$ and $\sum_{k=0}^{\infty} \alpha_{[k]}^2 < \infty$ during dual updates. Then almost surely $\liminf_{k \to \infty} k\alpha_{[k]}(Q^* - Q(\lambda_{[k]})) = 0$ for any $(\lambda_{[0]}, \alpha_{[0]}, \gamma_{[0]}) \in \mathbb{R}^{\bar{n}} \times \mathbb{R}_{>0}^{|\bar{\mathcal{E}}|} \times \{0\}$. ■

Remark 13: After sufficiently many iterations it is possible to observe at least one iteration where the difference between the value obtained at this iteration is arbitrarily close to Q^*. Moreover, the convergence rate depends on how the stepsizes are chosen. Let $\alpha_{[k]} = \frac{1}{(k+1)^q}$, where $\frac{1}{2} < q \leq 1$ for all agents. Define $b_{[k]} := \inf_{i \leq k} (Q(\lambda^*) - Q(\lambda_{[k]}))$. If $\liminf_{k \to \infty} b_k < \epsilon$. Hence, $\liminf_{k \to \infty} k^{1-q} b_{[k]} = 0$. It in turn yields a sublinear convergence rate of $O(\frac{1}{k^{1-q}})$ almost surely. ■

We state another sublinear convergence estimate in a stochastic setting under the uniform boundedness of the subgradient at $\lambda_{[k]}$ for all $k \in \mathbb{Z}_{\geq 0}$.

Assumption 9: There exists a scalar $G > 0$, such that $G \geq |g(\lambda_{[k]})|$, $\forall k \in \mathbb{Z}_{\geq 0}, g(\lambda) \in \partial^+ Q(\lambda)$ almost surely. ■

Proposition 7: Consider the sequence of random variables $\lambda_{[k]}$ generated by (9). Under Assumptions 1, 4, 7, and 8, assume $\beta_{[k]} = 0$ for all $k \in \mathbb{Z}_{\geq 0}$ and all edges use the same stepsize sequence $\{\alpha_{[k]}\}$ satisfying $\sum_{k=0}^{\infty} \alpha_{[k]} = \infty$ and $\sum_{k=0}^{\infty} \alpha_{[k]}^2 < \infty$ during dual updates. Then for any $(\lambda_{[0]}, \alpha_{[0]}, \gamma_{[0]}) \in \mathbb{R}^{\bar{n}} \times \mathbb{R}_{>0}^{|\bar{\mathcal{E}}|} \times \{0\}$:

$$Q^* - Q_{\text{best}[k]} \leq \frac{m_1 R^2 + (m_1 K + G^2) \sum_{i=1}^{k} \alpha_{[i]}^2}{2 \sum_{i=0}^{k} \alpha_{[i]}},$$

almost surely, where $Q_{\text{best}[k]} = \max_{i=0,1,\ldots,k} \mathbb{E}[Q(\lambda_{[k]})|\mathcal{F}_{i-1}]$ and $R := \sup_{\lambda^* \in \Lambda_{\text{opt}}} |\lambda_{[0]} - \lambda^*|$. ■

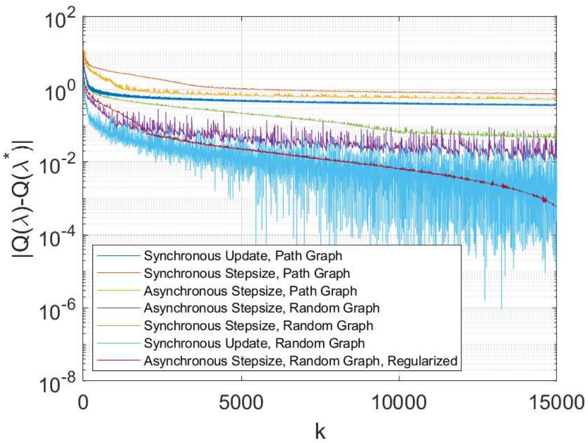
Proof: Similar to the proof of Proposition 5:

$$\begin{aligned}
&\mathbb{E}[|W^{-\frac{1}{2}}(\lambda_{[k+1]} - \lambda^*)|^2|\mathcal{F}_k] \leq |W^{-\frac{1}{2}}(\lambda_{[k]} - \lambda^*)|^2 \\
&\quad + \alpha_{[k]}^2 (m_1 K + G^2) - 2\alpha_{[k]} (Q(\lambda^*) - Q(\lambda_{[k]})) \qquad (20)
\end{aligned}$$

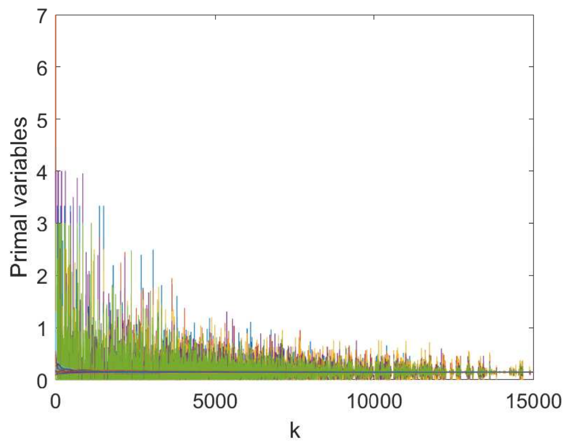
The rest of the proof follows from the proof of Theorem 2 in [20] and is omitted for the sake of brevity. ■

Remark 14: Since $\sum_{t=0}^{k} \frac{1}{t+1} \geq \int_0^k \frac{1}{1+t} dt = \log(t+1)|_0^k = \log(k+1)$, a convergence rate of $O(\frac{1}{\log k})$ for $\alpha_{[k]} = \frac{1}{k+1}$ is established using Proposition 7. ■

Remark 15: Proposition 7 establishes the sublinear convergence of the dual algorithm. If F_i in (1) is strongly convex and has Lipschitz gradient, then it can be shown via Fenchel duality that the dual function Q is strongly concave and has Lipschitz continuous gradient. Then, if we further assume that i.i.d. local timers are used in the updates, we recover the random block coordinate descent algorithm of [33, Theorem 1] where a stochastic linear convergence rate can be achieved. As a result, while the main focus of this work is the asynchronous aspect of the algorithm rather than achieving the fastest possible convergence rate, linear convergence can be recovered if we make more stringent assumptions about the problem and update rules in the network. ■



(a) Plot of $|Q(\lambda^*) - Q(\lambda)|$ versus the number of iterations



(b) Plot of primal variables $x_{(i)}$ versus the number of iterations

Fig. 1: Fig. 1(a) shows the convergence of the dual function under different settings, while Fig. 1(b) shows the primal variables indeed approach consensus as k grows.

VI. A Numerical Example

We consider the problem of minimizing $F(x) := \sum_{i=1}^{50} F_i(x)$ over a network of 50 agents. In this case, all selection matrices E_{ij} are identity matrices and for simplicity we choose e to be 0 so that the superdradient of the dual is error free. The functions $F_i(x)$ are given by $F_i(x) = \max\{-w_i(x - a_i) + b_s, b_s\}$ (from equation (5) in [18]) for $1 \le i \le 5$, where b_s is selected to be 0 while w_i and $a_i > 0$ are randomly chosen in the intervals $[0.2, 1]$ and $[2, 8]$ respectively. Note that these functions are non-smooth convex but not strictly convex. For $6 \le i \le 50$, $F_i(x) = x \log p_i x$, where for each $1 \le i \le 50$, p_i is a real number chosen in $(1, 5)$ uniformly at random. We consider two communication graphs: a path graph and a connected random geometric graph with 358 edges with all dual variables initialized at 0. Each agent $i \in \mathcal{V}$ is given a local decision variable $x_{(i)}$. Since for $1 \le i \le 5$, $F_i(x)$ is not strictly convex, as discussed in Remark 4, only convergence of the dual variable is guaranteed. We consider the following update scheme. We assign to each $i \in \mathcal{V}$ a real number $\tilde{p}_i$ chosen in $(0.5, 1)$ uniformly at random, a local counter ϖ_i that counts the number of times that there is at least one update at edges connecting i to its neighbors in the last 10 iterations in the whole network. Then for $(i, j) \in \overline{\mathcal{E}}$, $\lambda_{(ij)}$ gets updated with a probability $\tilde{p}_i \tilde{p}_j 0.7^{\varpi_i \varpi_j}$. The probability of updating for each node $i \in \mathcal{I}$ reduces as the number of updates grows in the last 10 global iterations. It is worth mentioning that although ϖ_i uses global information in the network, the introduction of ϖ_i is merely to simulate the algorithm with time varying updating probabilities. The implementation of the algorithm does not require global information in the network as long as the assumptions of the main results are satisfied. The stepsizes for dual updates are assumed to be the same, initialized at 0.15, and generated by $A(\alpha) = 0.15(1 + \log_{0.51} 0.15 - \log_{0.51} \alpha)^{-0.51}$ (This stepsize rule corresponds to $\alpha_{[k]} = 0.15(1 + k)^{-0.51}$ in the case of a global time counter k). For each of the two networks, we simulate the algorithm for the case where all communication and updates are synchronous, the updates are asynchronous with global stepsize, and the updates are asynchronous with local stepsize. To compare, we also consider the case where for $1 \le i \le 5$, $F_i(x)$ is regularized by $0.005x^2$ so that we should have all $x_{(i)}$, $i \in \mathcal{V}$ converge to the same x^* that minimizes $F(x)$. Dual variables are all initialized at the zero vector. Furthermore, since we work with non-strictly convex functions, to satisfy Assumption 1, we work with x over compact sets. By doing so, we can ensure a real solution to the local problem always exists and the upper bound of x does not need to be known, The plots for dual functions and primal variables when regularizers are used are given in Fig. 1.

As predicted by our results, $Q(\lambda)$ tends to its maximum and all agents reach consensus as the number of iterations tends to infinity if regularizers are used. Moreover, the use of local stepsizes are observed to provide better convergence rates. This can be explained by the fact that the stepsize for the edges that are not updating does not decrease while in the global stepsize case the stepsize decreases whenever there is an update anywhere in the network. Although not theoretically explored in the paper, adding more communication links in the network is also observed to accelerate the convergence of both dual and primal variables which may also induce more fluctuations caused by the numerical errors. It can be seen from Fig. 1(a) that the three dual functions corresponding to the random graph converge faster than the three dual functions corresponding to the path graph but with more fluctuations. It shows an interesting trade-off between the number of communication links and a faster convergence rate. The investigation of this interesting relationship between communication and convergence is left as future work. On the other hand, Fig. 1(b) shows that the primal variables reach consensus asymptotically as the functions are regularized.

VII. Conclusion and Future Work

We have proposed an edge based asynchronous distributed dual algorithm to minimize the sum of convex cost functions that have partially overlapping dependences. The analysis was done by viewing the asynchronous algorithm as block coordinate supergradient update for a concave function and applying techniques from stochastic approximation literature.

Under relatively weak assumptions on the cost function, sufficient conditions on the communications between agents and stepsizes were provided to prove almost sure convergence of the algorithm to the solution of the dual problem. We gave a sublinear convergence rate estimate for the algorithm under stronger assumptions. A numerical example was given to illustrate our main result. Future work will focus on providing convergence rate analysis under weaker assumptions on the communication between agents and the use of constant stepsizes in the algorithm to achieve faster convergence.

References

[1] S. A. Alghunaim and A. H. Sayed, "Distributed coupled multiagent stochastic optimization," *IEEE Trans. Autom. Control*, vol. 65, no. 1, pp. 175–190, Jan. 2020.

[2] D. Angeli, E. D. Sontag, and Y. Wang, "A characterization of integral input-to-state stability," *IEEE Trans. Autom. Control*, vol. 45, no. 6, pp. 1082–1097, Jun. 2000.

[3] D. P. Bertsekas, *Convex Optimization Algorithms*. Belmont, MA, USA: Athena Scientific, 2015.

[4] D. P. Bertsekas and J. N. Tsitsiklis, *Parallel and Distributed Computation: Numerical Methods*. Englewood Cliffs, NJ, USA: Prentice-Hall, 1989.

[5] P. Bianchi, W. Hachem, and F. Iutzeler, "A coordinate descent primal-dual algorithm and application to distributed asynchronous optimization," *IEEE Trans. Autom. Control*, vol. 61, no. 10, pp. 2947–2957, Oct. 2016.

[6] V. S. Borkar, "Asynchronous stochastic approximation," *SIAM J. Control Optim.*, vol. 36, no. 3, pp. 840–851, 1998, (Correction note in ibid. vol. 38, no. 3, pp. 662–663, 2000).

[7] V. S. Borkar, *Stochastic Approximation: A Dynamical Systems Viewpoint*. Cambridge Univ. Press, 2008.

[8] V. S. Borkar and S. P. Meyn, "The O.D.E. method for convergence of stochastic approximation and reinforcement learning," *SIAM J. Control Optim.*, vol. 38, no. 2, pp. 447–469, 2000.

[9] S. Boyd, N. Parikh, E. Chu, B. Peleato, and J. Eckstein, "Distributed optimization and statistical learning via the alternating direction method of multipliers," *Found. Trends Mach. Learn.*, vol. 3, no. 1, pp. 1–122, 2010.

[10] S. Boyd and L. Vandenberghe, *Convex Optimization*. Cambridge, UK: Cambridge Univ. Press, 2004.

[11] A. Camisa, F. Farina, I. Notarnicola, and G. Notarstefano, "Distributed constraint-coupled optimization over random time-varying graphs via primal decomposition and block subgradient approaches," in *Proc. IEEE 58th Conf. Decis. Control*, pp. 6374–6379, 2019.

[12] F. H. Clarke, "Optimization and nonsmooth analysis," in *Classics in Applied Mathematics*, vol. 5. Philadelphia, PA, USA: SIAM, 1990.

[13] A. Falsone, K. Margellos, S. Garatti, and M. Prandini, "Dual decomposition for multi-agent distributed optimization with coupling constraints," *Automatica*, vol. 84, pp. 149–158, 2017.

[14] F. Farina, A. Garulli, A. Giannitrapani, and G. Notarstefano, "A distributed asynchronous method of multipliers for constrained nonconvex optimization," *Automatica*, vol. 103, pp. 243–253, 2019.

[15] R. Goebel, R. G. Sanfelice, and A. R. Teel, *Hybrid Dynamical Systems: Modelling, Stability, and Robustness*, Princeton, NJ, USA: Princeton Univ. Press, 2012.

[16] R. Goebel and A. R. Teel, "Solutions to hybrid inclusions via set and graphical convergence with stability theory applications," *Automatica*, vol. 42, pp. 573–587, 2006.

[17] M. T. Hale, A. Nedić, and M. Egerstedt, "Asynchronous multiagent primal-dual optimization," *IEEE Trans. Autom. Control*, vol. 62, no. 9, pp. 4421–4435, Sep. 2017.

[18] H. Iiduka, "Distributed optimization for network resource allocation With nonsmooth utility functions," *IEEE Trans. Control Netw. Syst.*, vol. 6, no. 4, pp. 1354–1365, Dec. 2019.

[19] P. Lin, W. Ren, C. Yang, and W. Gui, "Distributed optimization with nonconvex velocity constraints, nonuniform position constraints, and nonuniform stepsizes," *IEEE Trans. Autom. Control*, vol. 64, no. 6, pp. 2575–2582, Jun. 2019.

[20] Y. Lin, I. Shames, and D. Nešić, "Asynchronous distributed optimization via dual decomposition and block coordinate ascent," in *Proc. IEEE 58th Conf. Decis. Control*, pp. 6380–6385, 2019.

[21] Y. Lin, E. D. Sontag, and Y. Wang, "A smooth converse Lyapunov Theorem for robust stability," *SIAM J. Control Optim.*, vol. 34, no. 1, pp. 124–160, 1996.

[22] A. Nedić and A. Ozdaglar, "Approximate primal solutions and rate analysis for dual subgradient methods," *SIAM J. Optim.*, vol. 19, no. 4, pp. 1757–1780, 2009.

[23] A. Nedić and A. Ozdaglar, "Distributed subgradient methods for multi-agent optimization," *IEEE Trans. Autom. Control*, vol. 54, no. 1, pp. 48–61, Jan. 2009.

[24] I. Notarnicola, R. Carli, and G. Notarstefano, "Distributed partitioned big-data optimization via asynchronous dual decomposition," *IEEE Trans. Control Netw. Syst.*, vol. 5, no. 4, pp. 1910–1919, Dec. 2018.

[25] A. Ramaswamy, S. Bhatnagar, and D. E. Quevedo, "Asynchronous stochastic approximations with asymptotically biased errors and deep multi-agent learning," arXiv preprint arXiv:1802.07935, 2018.

[26] 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.

[27] A. Simonetto and H. Jamali-Rad,"Primal recovery from consensus based dual decomposition for distributed convex optimization," *J. Optim. Theory Appl.*, vol. 168, no. 1, pp. 172–197, 2016.

[28] M. Tabbara and D. Nešić, "Input-output stability of networked control systems with stochastic protocols and channels," *IEEE Trans. Autom. Control*, vol. 53, no. 5, pp. 1160–1175, May. 2008.

[29] A. R. Teel and L. Praly, "On assigning the derivative of a disturbance attenuation control Lyapunov function," *Math. Control Signals Syst.*, vol. 13, pp. 95–124, 2000.

[30] A. Teixeira, E. Ghadimi, I. Shames, H. Sandberg, and M. Johansson, "The ADMM algorithm for distributed quadratic problems: Parameter selection and constraint preconditioning," *IEEE Trans. Signal Process.*, vol. 64, no. 2, pp. 290–305, Feb. 2016.

[31] H. Terelius, U. Topcu, and R. M. Murray, "Decentralized multi-agent optimization via dual decomposition," in *Proc. 18th IFAC World Congr.*, pp. 11245–11251, 2011.

[32] E. Wei and A. Ozdaglar, "On the $O(1/k)$ convergence of asynchronous distributed alternating direction method of multipliers," in *Proc. IEEE Global. Conf. Signal Inf. Process.*, pp. 551–554, 2013.

[33] S. J. Wright, "Coordinate descent algorithms," *Math. Program.*, vol. 151, no. 1, pp. 3–34, 2015.

[34] L. Xiao, M. Johansson, and S. Boyd, "Simultaneous routing and resource allocation via dual decomposition," *IEEE Trans. Commun.*, vol. 52, no. 7, pp. 1136–1144, Jul. 2004.

[35] J. Xu, S. Zhu, Y. C. Soh, and L. Xie, "Convergence of asynchronous distributed gradient methods over stochastic networks," *IEEE Trans. Autom. Control*, vol. 63, no. 2, pp. 434–448, Feb. 2018.

[36] T. Yang, X. Yi, J. Wu, Y. Yuan, D. Wu, Z. Meng, Y. Hong, H. Wang, Z. Lin, and K. H. Johansson, "A survey of distributed optimization," *Annual Reviews in Control*, vol. 47, pp. 278–305, 2019.

Appendix

A. Background Definitions and Lemmas

For any $T > 0$ and $n \in \mathbb{Z}_{>0}$, the space of continuous functions from $[0, T]$ to $\mathbb{R}^n$ is denoted by $C([0, T]; \mathbb{R}^n)$ and the space of measurable functions f from $[0, T]$ to $\mathbb{R}^n$ satisfying $\int_0^T |f(t)|^2 dt < \infty$ is denoted by $\mathcal{L}_2([0, T]; \mathbb{R}^n)$. A subset of a topological space is *relatively compact* if its closure is compact. A set $B \subseteq C([0, T]; \mathbb{R}^n)$ is *equicontinuous at* $t \in [0, T]$ if for any $\epsilon > 0$, there exists a $\delta > 0$ such that $|t-s| < \delta$, $s \in [0, T]$ implies $\sup_{f \in B} ||f(t) - f(s)||_{\infty,T} < \epsilon$, where $||f||_{\infty,T} := \sup_{s\in[0,T]} |f(s)|$. It is *equicontinuous* if it is equicontinuous at all $t \in [0, T]$. It is *pointwise bounded* if for any $t \in [0, T]$, $\sup_{f \in B} ||f(t)||_{\infty,T} < \infty$. The following result known as the Arzelà-Ascoli Theorem characterises relative compactness in $C([0, T]; \mathbb{R}^n)$.

Lemma 5: [7, Appendix A, Theorem 1] A set $B \subseteq C([0, T]; \mathbb{R}^n)$ is relatively compact if and only if it is equicontinuous and pointwise bounded. ■

Consider the Hilbert space $\mathcal{L}_2([0, T]; \mathbb{R}^n)$ with the inner product defined by $\langle f, g\rangle_T := \int_0^T \langle f(t), g(t)\rangle dt, \quad f, g \in$

$\mathcal{L}_2([0,T];\mathbb{R}^n)$ and the norm $||f||_{2,T} := \sqrt{\langle f,f\rangle_T}$. The *weak* topology on $\mathcal{L}_2([0,T];\mathbb{R}^n)$ is defined as the coarsest topology with respect to which the functions $f \to \langle f,g\rangle_T$ are continuous for all $g \in \mathcal{L}_2([0,T];\mathbb{R}^n)$. Then, we say $f_n \to f$ weakly in $\mathcal{L}_2([0,T];\mathbb{R}^n)$ if and only if $\langle f_n,g\rangle_T \to \langle f,g\rangle_T$ for all $g \in \mathcal{L}_2([0,T];\mathbb{R}^n)$. The following result which is a special instance of the Banach-Alaoglu Theorem characterises relative compactness in $\mathcal{L}_2([0,T];\mathbb{R}^n)$.

Lemma 6: [7, Appendix A, Theorem 3] A $||\cdot||_{2,T}$-bounded set $B \subseteq \mathcal{L}_2([0,T];\mathbb{R}^n)$ is relatively compact in the weak topology. ■

Next, we state a martingale convergence theorem that is critical in the proof of Theorem 1. Let $(\Omega,\mathcal{F},\mathbb{P})$ be a probability space and $\{\mathcal{F}_n\}$ be a family of increasing σ-subfields of $\mathcal{F}$. For a martingale $\{X_n\}$ with respect to $\{\mathcal{F}_n\}$, we have the following convergence theorem.

Lemma 7: [7, Appendix C, Theorem 11] If $\mathbb{E}[|X_n|^2] < \infty$ for any $n \in \mathbb{Z}_{\geq 0}$ and $\sum_n \mathbb{E}[|X_{n+1}-X_n|^2|\mathcal{F}_n] < \infty$, then $\{X_n\}$ converges almost surely. ■

Some useful properties of convex functions are stated below.

Lemma 8: Let $f : \mathbb{R}^n \to \mathbb{R}$ be a real-valued convex function. Then, the following statements hold for any $x \in \mathbb{R}^n$:

(i) $\partial f(x)$ is non-empty, compact and convex.
(ii) f is locally Lipschitz.
(iii) $\partial f(x)$ is locally bounded.
(iv) $\partial f(x)$ is upper semicontinuous[4].
(v) $\partial f(x)$ is outer semicontinuous. ■

Proof: Item (i) follows from [3, Proposition 3.1.1]. Items (ii) and (iii) follow from [3, Proposition 3.1.2]. By [12, Proposition 2.2.7], if $f(x)$ is convex, then $\partial f(x)$ coincides with the generalized gradient which is upper semicontinuous by [12, item(d), Proposition 2.1.5]. Lastly, by [15, Lemma 5.15], when $\partial f(x)$ is closed and locally bounded, upper semicontinuity of $\partial f(x)$ is equivalent to outer semicontinuity of $\partial f(x)$. Since $\partial f(x)$ is indeed closed and locally bounded by item (i) and (iii) of Lemma 8, item (v) follows. ■

A function $\psi : \mathbb{R}_{\geq 0} \to \mathbb{R}_{\geq 0}$ is *of class* $\mathcal{K}$ if it is continuous, strictly increasing and $\psi(0) = 0$. It is *of class* $\mathcal{K}_\infty$ if it is of class $\mathcal{K}$ and unbounded. A function $\chi : \mathbb{R}_{\geq 0} \to \mathbb{R}_{\geq 0}$ is *of class* $\mathcal{L}$ if it is non-increasing and $\lim_{t\to\infty}\chi(t) = 0$. A function $\beta : \mathbb{R}_{\geq 0}\times\mathbb{R}_{\geq 0} \to \mathbb{R}_{\geq 0}$ is *of class* $\mathcal{KL}$ if $\beta(\cdot,t) \in \mathcal{K}$ for each $t \in \mathbb{R}_{\geq 0}$ and $\beta(s,\cdot) \in \mathcal{L}$ for each $s \in \mathbb{R}_{\geq 0}$ Let $V^\circ(x;v) := \limsup_{h\to 0^+,y\to x}\frac{V(y+hv)-V(y)}{h}$ be the generalized directional derivative of Clarke of a locally Lipschitz function V at x in the direction of v. This is the standard directional derivative $\langle\nabla V(x),v\rangle$ for continuously differentiable V [12].

B. Proof of Lemma 3

Since $\bar{\alpha} = \max_{i\in\mathcal{I}_a}\alpha_{(i)}$, all elements of the sequence $\{\bar{\alpha}_{[k]}\}$ must belong to the set $\mathcal{M} := \{a \in \mathbb{R} : a = \alpha_{(i),[k]},\ i \in \mathbb{Z}_{|\bar{\mathcal{E}}|}\backslash\{0\}, k \in \mathbb{Z}_{\geq 0}\}$, which is the set of all elements of local stepsize sequences. Thus, the sum of squares of $\{\bar{\alpha}_{[k]}\}$ must be upper bounded by the sum of squares of all elements in $\mathcal{M}$. Note that, for any i in the finite set $\mathbb{Z}_{|\bar{\mathcal{E}}|}\backslash\{0\}$, the sequence $\{\alpha_{(i),[k]}\}$ is square summable. Thus, all elements in $\mathcal{M}$ must also be square summable. That is $\sum_{k=0}^\infty \bar{\alpha}_{[k]}^2 < \sum_{i=1}^{|\bar{\mathcal{E}}|}\sum_{k=0}^\infty \alpha_{(i),[k]}^2 < \infty$ for $\{\alpha_{(i),[k]}\}$ generated by $\alpha_{(i)}^+ = A_{(i)}(\alpha_{(i)})$. By Assumption 2, $\mathcal{I}_a \neq \emptyset$ infinitely often, almost surely. Therefore, $\bar{\alpha}_{[k]} > 0$ for infinitely many k, almost surely. Since $|\bar{\mathcal{E}}|$ is finite, $\bar{\alpha}_{[k]} > 0$ for infinitely many k implies that $\{\bar{\alpha}_{[k]}\}$ must contain infinitely many terms from at least one sequence $\{\alpha_{(i),[k]}\}$ for some $i \in \mathbb{Z}_{|\bar{\mathcal{E}}|}\backslash\{0\}$. Otherwise $\{\alpha_{(i),[k]}\}$ will only have a finite number of elements which contradicts the fact that $\bar{\alpha}_{[k]} > 0$ for infinitely many k. Therefore $\sum_{k=0}^\infty \bar{\alpha}_{[k]} = \infty$ almost surely.

C. Proof of Proposition 1

By Lemma 3, $\{\bar{\alpha}_{[k]}\}_{k=1}^\infty$ is square summable almost surely and the largest element of $\theta(t)$ is 1 for all t. Hence, $\sum_{k=1}^\infty \mathbb{E}[|\zeta_{[k+1]}-\zeta_{[k]}|^2|\mathcal{F}_k] \leq \sum_{k=1}^\infty \bar{\alpha}_{[k]}^2\mathbb{E}[|e_{[k+1]}|^2|\mathcal{F}_k] \leq \sum_{k=1}^\infty \bar{\alpha}_{[k]}^2 K < \infty$ almost surely, due to Assumption 4 and Lemma 3. Since $\beta_{[k]} = 0$ for all k, $\mathbb{E}[\langle\bar{\alpha}_{[m_1]}\theta(t_{[m_1]})e_{[m_1+1]},\bar{\alpha}_{[m_2]}\theta(t_{[m_2]})e_{[m_2+1]}\rangle] = 0$ holds for any $m_1 \neq m_2 \in \mathbb{Z}_{\geq 0}$. Thus, for any $k \geq 1$, $\mathbb{E}[|\zeta_{[k]}|^2] = \sum_{k=1}^\infty \mathbb{E}[|\bar{\alpha}_{[k-1]}e_{[k]}|^2] < \infty$ follows from the fact that $|\theta(t)|$ is bounded. Consequently, all assumptions of Lemma 7 are satisfied and $\zeta_{[k]}$ converges almost surely to a random variable. The definition of $r(t)$ and Lemma 3 yield $r(t_{[k]}) \to \infty$ as $k \to \infty$. Therefore, the sequence $\{\zeta_{[r(t_{[k]}+\cdot)]}-\zeta_{[r(t_{[k]})]}\}_{k=1}^\infty$ converges to 0 uniformly on $[0,T]$ for any $T > 0$ almost surely when k is sufficiently large. Thus the proof is complete.

D. Proof of Proposition 2

From Proposition 1, the sequence $\{\zeta_{[r(t_{[k]}+u)]}-\zeta_{[r(t_{[k]})]}\}_{k=1}^\infty$ almost surely converges to 0 on $[0,\infty)$. Moreover, due to Lemma 3, $t_{[k+1]}-t_{[k]} = \bar{\alpha}_{[k]} \to 0$ as $k \to \infty$, thus for $t \in [t_{[k]},t_{[k+1]})$, $\int_0^u \theta(t+s)g(\lambda_{[r(t+s)]})ds - \int_0^u \theta(t+s)g(\bar{\lambda}(t+s))ds \to 0$ as $k \to \infty$ and the result follows since $k \to \infty$ implies $t_{[k]} \to \infty$ by Lemma 3.

E. Proof of Proposition 3

The proof is based on [7, Chapter 7, Theorem 2]. First, let $\tilde{\theta}(t) := (\tau_{(1),[k]},\ldots,\tau_{(|\bar{\mathcal{E}}|),[k]})$, for $t \in [t_{[k]},t_{[k+1]})$ which is the vector containing all diagonal elements of $\theta(t)$. Then, we view $\tilde{\theta}(\cdot)$ as an element of $\mathcal{V}$, where $\mathcal{V}$ is the space of measurable maps $y(\cdot) : [0,\infty) \to [0,1]^{\bar{n}}$ with the coarsest topology that renders the maps $y(\cdot) \to \int_0^T\langle g(s),y(s)\rangle ds$ continuous for all $T > 0, g(\cdot) \in \mathcal{L}_2([0,T];\mathbb{R}^{\bar{n}})$. This defines the weak topology on $\mathcal{L}_2([0,T];\mathbb{R}^{\bar{n}})$ (see Lemma 6). Since $g(\cdot) \in \mathcal{L}_2([0,T];\mathbb{R}^{\bar{n}})$ and all elements of $\tilde{\theta}(t)$ are uniformly bounded for all $t \geq 0$, $\mathcal{V}$ is a bounded subset of $\mathcal{L}_2([0,T];\mathbb{R}^{\bar{n}})$. Then by the Banach-Alaoglu Theorem (see Lemma 6), $\mathcal{V}$ is relatively compact. Thus, by choosing an appropriate common subsequence of $\{\tilde{\theta}(t_{[k]}+\cdot)\}$ and $\{\bar{\lambda}(t_{[k]}+\cdot)\}$ indexed by $\{k'\}$, we can show that any limit point of $\{\bar{\lambda}(t_{[k]}+\cdot)\}$ in $C([0,T];\mathbb{R}^{\bar{n}})$ denoted by $\tilde{\lambda}(\cdot)$ is of the form

$$\lim_{k'\to\infty}\bar{\lambda}(t_{[k']}+t) := \tilde{\lambda}(t) = \tilde{\lambda}(0) - \int_0^t \Theta(s)g(\tilde{\lambda}(s))ds,$$

[4]A set-valued mapping $M : \mathbb{R}^n \rightrightarrows \mathbb{R}^n$ is upper semicontinuous if for each $x \in \mathbb{R}^n$ and $\varepsilon > 0$ there exists $\delta > 0$ such that $M(x+\delta\mathbb{B}) \subseteq M(x)+\varepsilon\mathbb{B}$.

where $\Theta(\cdot) = \mathrm{diag}(\tilde{\theta}^*_{(1)}(\cdot), \ldots, \tilde{\theta}^*_{(|\bar{\mathcal{E}}|)}(\cdot))$ and $\tilde{\theta}^*(\cdot) = (\tilde{\theta}^*_{(1)}(\cdot), \ldots, \tilde{\theta}^*_{(|\bar{\mathcal{E}}|)}(\cdot))$ is the subsequential limits of $\{\tilde{\theta}(t_{[k']}+\cdot)\}$ in $\mathcal{V}$ corresponding to $\{k'\}$. The limit point of $\{g(\tilde{\lambda}(t_{[k']}+\cdot))\}$ in $\mathcal{L}_2([0,T];\mathbb{R}^{\bar{n}})$ belongs to $\partial f(\tilde{\lambda}(\cdot))$ since ∂f is outer semicontinuous with closed and convex values which is guaranteed by Lemma 8.

For any given $n \geq 0$ and $s > 0$, let $N(n,s)$ be the smallest integer m such that $t_{[m]} \geq t_{[n]} + s$. Since $\tilde{\theta}(\cdot)$ is a piecewise constant function by definition and $\bar{\alpha}_{[k]} = t_{[k+1]} - t_{[k]}$, for any given i and k, $\int_{t_{[k]}}^{t_{[k+1]}} \tilde{\theta}_{(i)}(t)dt = \frac{\alpha_{(i),[\gamma_{(i),[k]}]} v_{(i),[k+1]}}{\bar{\alpha}_{[k]}} \bar{\alpha}_{[k]}$. As a result, the integral of $\tilde{\theta}(\cdot)$ can be replaced by appropriate sums. Then for any $i \in \mathbb{Z}_{|\bar{\mathcal{E}}|} \backslash \{0\}$, $\int_t^{t+s} \tilde{\theta}^*_{(i)}(y)dy = \lim_{k'\to\infty} \sum_{m=k'}^{N(k',s)} \frac{\alpha_{(i),[\gamma_{(i),[m]}]} v_{(i),[m+1]}}{\bar{\alpha}_{[m]}} \bar{\alpha}_{[m]}$ which is equal to $\lim_{k'\to\infty} \sum_{m=\gamma_{(i),[k']}}^{\gamma_{(i),[N(k',s)]}} \alpha_{(i),[m]}$. The last equality follows from the fact that $\alpha_{(i)}$ and $\gamma_{(i)}$ get updated if and only if $v^+_{(i)} = 1$ which means at time k, the stepsize $\alpha_{(1),[k]}$ generated by (10) is the same as $\alpha_{(i),[\gamma_{(i),[k]}]}$ generated from the pre-selected update rule $\alpha^+_{(i)} = A_{(i)}(\alpha_{(i)})$. By Assumptions 2, 3, $\frac{\int_t^{t+s} \tilde{\theta}^*_{(i)}(y)dy}{\int_t^{t+s} \tilde{\theta}^*_{(j)}(y)dy} = \lim_{k'\to\infty} \frac{\sum_{m=\gamma_{(i),[k']}}^{\gamma_{(i),[N(k',s)]}} \alpha_{(i),[m]}}{\sum_{m=\gamma_{(j),[k']}}^{\gamma_{(j),[N(k',s)]}} \alpha_{(j),[m]}} \geq \liminf_{k'\to\infty} \frac{\Delta \alpha_{(i),[\gamma_{(i),[N(k',s)]}]}}{\alpha_{(j),[\gamma_{(j),[k']}]}} \geq \min\{\Delta\bar{\delta}, \frac{\Delta}{\bar{\Delta}}\} := \varepsilon$ for any $i, j \in \mathbb{Z}_{|\bar{\mathcal{E}}|} \backslash \{0\}$. The first inequality follows from replacing the sum in the numerator (respectively denominator) by the sum of its smallest (resp. largest) element (assuming without loss of generality all stepsize sequences are non-increasing) and Assumption 2 that the ratio of the number of updates is lower bounded by Δ. Since s is arbitrary and all diagonal elements of $\Theta(t)$ are upper bounded by 1 by construction, $\Theta(t)$ has diagonal entries in $[\varepsilon, 1]$.

F. Proof of Proposition 4

By Proposition 3, $\Theta(t) \in \mathbb{R}^{\bar{n}\times\bar{n}}$ is a diagonal matrix with diagonal elements in the interval $[\varepsilon, 1]$ for all t. Since $f(\tilde{\lambda})$ is locally Lipschitz, the radially unbounded Lyapunov function $V(\tilde{\lambda}) := f(\tilde{\lambda}) + Q^*$ is also locally Lipschitz. Moreover, by convexity of f, both V and ∇f vanish on Λ_{opt} and nowhere else. Define $\phi_1(s) := \inf_{|\tilde{\lambda}|_{\Lambda_{\mathrm{opt}}} \geq s} V(\tilde{\lambda})$ and $\phi_2(s) := \sup_{|\tilde{\lambda}|_{\Lambda_{\mathrm{opt}}} \leq s} V(\tilde{\lambda})$. It can be seen that ϕ_1 and ϕ_2 are positive definite and non-decreasing. Since $|\tilde{\lambda}|_{\Lambda_{\mathrm{opt}}}$ is a continuous function of $\tilde{\lambda}$ for any given Λ_{opt}, the functions ϕ_1 and ϕ_2 are continuous. Because $V(\tilde{\lambda})$ is radially unbounded and Λ_{opt} is compact, $\phi_1(s)$ and $\phi_2(s)$ tend to infinity as $s \to \infty$. Hence we can choose $\psi_1(s) \leq k_1\phi_1(s)$ with $k_1 \in (0,1)$ and $\psi_2(s) \geq k_2\phi_2(s)$ with $k_2 \in (1,\infty)$ to belong to class $\mathcal{K}_\infty$. Therefore, $\psi_1(|\tilde{\lambda}|_{\Lambda_{\mathrm{opt}}}) \leq \phi_1(|\tilde{\lambda}|_{\Lambda_{\mathrm{opt}}}) \leq V(\tilde{\lambda}) \leq \phi_2(|\tilde{\lambda}|_{\Lambda_{\mathrm{opt}}}) \leq \psi_2(|\tilde{\lambda}|_{\Lambda_{\mathrm{opt}}})$ holds. By Rademacher's theorem, $V(\tilde{\lambda})$ is differentiable almost everywhere and $\langle \frac{\partial V}{\partial \tilde{\lambda}}, -\Theta(t)\nabla f(\tilde{\lambda})\rangle = -\nabla^T f(\tilde{\lambda})\Theta(t)\nabla f(\tilde{\lambda})$ holds for almost all $\tilde{\lambda}$. Thus, from [29, item 5, p. 100], for almost all t and each $v \in -\Theta(t)\partial f(\tilde{\lambda})$, $\dot{V}(\tilde{\lambda}) \leq V^\circ(\tilde{\lambda}; v) \leq -\langle \nabla f(\tilde{\lambda}), \Theta(t)\nabla f(\tilde{\lambda})\rangle$. Since $\Theta(t)$ is diagonal and all its diagonal elements are in $[\varepsilon, 1]$, $\nabla^T f(\tilde{\lambda})\Theta(t)\nabla f(\tilde{\lambda})$ is lower bounded by the continuous function $\varepsilon\nabla^T f(\tilde{\lambda})\nabla f(\tilde{\lambda})$ which vanishes only on Λ_{opt} and is positive everywhere else. Thus there exists a positive definite function $\psi_3(s) := \inf_{|\tilde{\lambda}|_{\Lambda_{\mathrm{opt}}} \geq s} \varepsilon\nabla^T f(\tilde{\lambda})\nabla f(\tilde{\lambda})$, such that $\dot{V} \leq -\psi_3(|\tilde{\lambda}|_{\Lambda_{\mathrm{opt}}})$ for almost all t. By [2, Lemma IV.1], there exist $\rho_1 \in \mathcal{K}_\infty$ and $\rho_2 \in \mathcal{L}$ such that $\psi_3(s) \geq \rho_1(s)\rho_2(s)$. Therefore, for almost all t, $\dot{V} \leq -\psi_3(|\tilde{\lambda}|_{\Lambda_{\mathrm{opt}}}) \leq -\rho_1(|\tilde{\lambda}|_{\Lambda_{\mathrm{opt}}})\rho_2(|\tilde{\lambda}|_{\Lambda_{\mathrm{opt}}}) \leq -\rho_1(\psi_2^{-1}(V))\rho_2(\psi_1^{-1}(V)) := -\rho(V)$. Since ρ is positive definite, by [2, Lemma IV.2], there exists a $\beta \in \mathcal{KL}$ such that $V(\tilde{\lambda}(t)) \leq \beta(V(\tilde{\lambda}(t_0)), t - t_0)$. Therefore, all solutions to (16) starting at $\tilde{\lambda}(t_0)$ satisfy $|\tilde{\lambda}(t)|_{\Lambda_{\mathrm{opt}}} \leq \psi_1^{-1}(\beta(\psi_2(|\tilde{\lambda}(t_0)|_{\Lambda_{\mathrm{opt}}}), t - t_0)) := \bar{\beta}(|\tilde{\lambda}(t_0)|_{\Lambda_{\mathrm{opt}}}, t - t_0)$ with $\bar{\beta} \in \mathcal{KL}$. Thus, Λ_{opt} is UGAS for (16).

Yankai Lin (S'18) received the B.S. degree in electrical engineering and automation and the M.Eng. degree in electrical engineering from Huazhong University of Science and Technology, Wuhan, China, and the University of Melbourne, Parkville, VIC, Australia, in 2014 and 2016, respectively. He is currently pursuing the Ph.D. degree at the University of Melbourne.

His current research interests include optimization theory and networked control systems.

Iman Shames (M'11) received the B.Sc. degree in Electrical Engineering from Shiraz University in 2006, and the Ph.D. degree in engineering from the Australian National University, Canberra, Australia in 2011.

He is the Professor of Mechatronics Engineering at the School of Engineering, the Australian National University. Before joining the ANU, he was an Associate Professor at the Department of Electrical and Electronic Engineering, the University of Melbourne from 2020 to 2021, a senior lecturer from 2014 to 2019, and a McKenzie fellow at the same department from 2012 to 2014. Previously, he was an ACCESS Postdoctoral Researcher at the ACCESS Linnaeus Centre, the KTH Royal Institute of Technology, Stockholm, Sweden. His current research interests include, but are not limited to, decision making under uncertainty, optimisation theory and its application in control and estimation, mathematical systems theory of cyber-physical systems, and security and privacy.

Dragan Nešić (F'08) is a Professor at the Department of Electrical and Electronic Engineering at The University of Melbourne. He received his Bachelor of Mechanical Engineering Degree at the University of Belgrade (1990) and his Ph.D. at the Australian National University (1997). His research interests are in the broad area of control engineering including its mathematical foundations and its applications to various areas of engineering and science.

Prof. Nešić is a Fellow of IEEE and a Fellow of IFAC. He was a co-recipient of the George S. Axelby Outstanding Paper Award (2018). He was also awarded a Humboldt Research Fellowship (2003) by the Alexander von Humboldt Foundation, an Australian Professorial Fellowship (2004-2009), and a Future Fellowship (2010-2014) by the Australian Research Council.