A Distributed Algorithm For Convex Constrained Optimization Under Noise

Nikolaos Chatzipanagiotis, *Student Member, IEEE,* and Michael M. Zavlanos, *Member, IEEE*

***Abstract*—We present a novel distributed algorithm for convex constrained optimization problems that are subject to noise corruption and uncertainties. The proposed scheme can be classified as a distributed stochastic approximation method, where a unique feature here is that we allow for multiple noise terms to appear in both the computation and communication stages of the distributed iterative process. Specifically, we consider problems that involve multiple agents optimizing a separable convex objective function subject to convex local constraints and linear coupling constraints. This is a richer class of problems compared to those that can be handled by existing distributed stochastic approximation methods which consider only consensus constraints and fewer sources of noise. The proposed algorithm utilizes the augmented Lagrangian (AL) framework, which has been widely used recently to solve deterministic optimization problems in a distributed way. We show that the proposed method generates sequences of primal and dual variables that converge to their respective optimal sets almost surely.**

***Index Terms*—Distributed optimization, stochastic optimization, stochastic approximation, multi-agent systems, noisy communications.**

I. Introduction

Distributed optimization methods [1] have recently received significant attention due to the ever increasing size and complexity of modern day problems, and the ongoing advancements in the parallel processing capabilities of contemporary computers. By decomposing the original problem into smaller, more manageable subproblems that are solved in parallel, distributed methods scale much better than their centralized counterparts. For this reason, they are widely used to solve large-scale problems arising in areas as diverse as wireless communications, optimal control, machine learning, artificial intelligence, computational biology, finance and statistics, to name a few. Moreover, distributed algorithms avoid the cost and fragility associated with centralized coordination, and provide better privacy for the autonomous decision makers. These are desirable properties, especially in applications involving networked robotics, communication or sensor networks, and power distribution systems.

A classic method used for distributed optimization is that of *dual decomposition* and is based on Lagrangian duality theory [1, 2]. Dual methods are simple and popular, however, they suffer from exceedingly slow convergence rates and require strict convexity of the objective function. These drawbacks are alleviated by utilizing the augmented Lagrangian (AL) framework, which has recently received considerable attention as a most efficient approach for distributed optimization in determistic settings; see e.g. [3]–[7]. Alternative algorithms for distributed optimization include Newton methods [8, 9], projection-based approaches [10, 11], Nesterov-like algorithms [12, 13], online methods [14, 15], and even continuous-time approaches [16].

In this paper we propose a distributed algorithm for convex constrained optimization problems that are subject to noise and uncertainties. The focus of our investigation is the following convex problem

$$\begin{aligned} \min \ & \sum_{i=1}^{N} f_i(\mathbf{x}_i) \\ \text{subject to } & \sum_{i=1}^{N} \mathbf{A}_i \mathbf{x}_i = \mathbf{b}, \\ & \mathbf{x}_i \in \mathcal{X}_i, \quad i = 1, 2, \ldots, N, \end{aligned} \tag{1}$$

where, for every $i \in \mathcal{I} = \{1, 2, \ldots, N\}$, the $\mathcal{X}_i \subseteq \mathbb{R}^{n_i}$ denotes a nonempty closed, convex subset of n_i-dimensional Euclidean space, the $f_i : \mathbb{R}^{n_i} \to \mathbb{R}$ is a convex function, and the $\mathbf{A}_i$ is a matrix of dimension $m \times n_i$.

Problem (1) models situations where a set of N decision makers, henceforth referred to as agents, need to determine local decisions $\mathbf{x}_i \in \mathcal{X}_i$ that minimize a collection of local cost functions $f_i(\mathbf{x}_i)$, while respecting a set of affine constraints $\sum_{i=1}^{N} \mathbf{A}_i \mathbf{x}_i = \mathbf{b}$ that couple the local decisions between agents. In previous work [7], we presented the *Accelerated Distributed Augmented Lagrangians* (ADAL) method to solve such problems in a distributed fashion. ADAL is a primal-dual iterative scheme based on the augmented Lagrangian framework. Each iteration of ADAL consists of three steps. First, every agent solves a local convex optimization problem based on a separable approximation of the AL, that utilizes only locally available variables. Then, the agents update and communicate their primal variables to neighboring agents. Finally, they update their dual variables based on the values of the communicated primal variables. The computations at each step are performed in parallel. It was shown in [17] that ADAL has a worst-case $O(1/k)$ convergence rate, where k denotes the number of iterations.

In this paper, we extend ADAL to address the case where problem (1) needs to be solved distributedly in the presence of uncertainty and noise. In particular, we consider the scenario where: i) the agents have access only to noisy approximations of their objective functions at each iteration or, equivalently,

This work is supported by NSF CNS under grant #1261828, and by ONR under grant #N000141410479.

Nikolaos Chatzipanagiotis and Michael M. Zavlanos are with the Dept. of Mechanical Engineering and Materials Science, Duke University, Durham, NC, 27708, USA, {n.chatzip,michael.zavlanos}@duke.edu.

the agents can calculate only approximate subgradients of their objective functions, and ii) noise corrupts the primal and dual message exchanges between agents during the iterative execution of the method. To address this stochastic framework, ADAL needs to be modified; we refer to the new algorithm as the *Stochastic Accelerated Distributed Augmented Lagrangians* (SADAL) method. We show that SADAL generates sequences of primal and dual variables that converge to their optimal values almost surely (a.s.).

Our work is also directly related to the literature of *stochastic approximation* (SA) techniques. Generally speaking, the term SA characterizes those stochastic methods that attempt to iteratively solve convex optimization problems based on noisy (sub)gradient observations. SA has been an active area of research since the 1950s, beginning with the seminal work of [18], which introduced the *Robbins-Monro* algorithm for unconstrained optimization problems and proved that the method generates iterates that converge to the optimal solution in the mean square sense. Since then, a significant amount of SA literature has emerged, with some of the most representative works being [19]–[25]. Certain works follow the so-called "limiting mean ODE" (ordinary differential equations) technique to prove the convergence of SA schemes; see for example [26]–[28]. On the other hand, there exists a significantly smaller number of works that considers distributed stochastic approximation schemes. The existing literature on such distributed approaches is mostly concerned with consensus constrained optimization problems, wherein a set of agents with separable objective functions need to agree on a common decision vector; see, e.g., [29]–[38].

The contribution of this paper is threefold. First, we propose a distributed stochastic approximation algorithm that can address more general constraint sets compared to the relevant existing literature in SA which is concerned only with consensus constraints. In fact, problems with consensus constraints can be expressed in the form of (1), such that they are a special case of the scenario under consideration here. Second, we allow for multiple noise terms, namely four, to appear in both the computation and communication stages of the proposed distributed iterative method. Typically, distributed stochastic approximation algorithms contain a single source of noise, affecting either the computations or the communications, with only a few works considering two noise terms simultaneously [30]–[32].

Finally, the proposed method is based on the augmented Lagrangian framework, for which only a limited amount of works exist that consider it in a stochastic setting. For example, in [37, 38] the authors examine the stability properties of the *Alternating Direction Method of Multipliers* (ADMM) when applied to consensus problems that suffer from noise in the message exchanges. Moreover, in [39] a convergence result is derived for the ADMM applied to problems of the form (1) with $N = 2$ and noise appearing in one of the two objective functions. In this paper we study a more general framework than the aforementioned stochastic ADMM works, and also provide a stronger convergence result (a.s.). Specifically, we consider the general class of problems (1), where N is allowed to be larger than 2, and where noise appears in all the objective functions and all message exchanges at the same time. The main challenge here is that AL methods are primal-dual schemes, which means that the effects from multiple sources of noise corruption and uncertainty propagate in-between the two domains.

The rest of this paper is organized as follows. In section II we discuss some essential facts regarding duality and augmented Lagrangians. We also provide a description of the ADAL method and recall its convergence properties. In section III we describe the proposed SADAL algorithm and elaborate on the specific noise terms that can appear during its iterative execution. In section IV we establish the a.s. convergence of SADAL. Finally, in Section V we present numerical results to verify the validity of the proposed approach.

II. Preliminaries

In this section we introduce some basic facts about solving (1) using the augmented Lagrangian framework in a deterministic setting, i.e., when the agents have exact knowledge of the objective functions f_i and when there is no noise in the message exchanges between them (communication is necessary when these algorithms are implemented in a distributed fashion). Moreover, we briefly discuss the distributed ADAL method [7] to lay the ground for the development of its stochastic counterpart, the SADAL, which will be subsequently presented in Section III. We denote

$$f(\mathbf{x}) = \sum_{i=1}^{N} f_i(\mathbf{x}_i)$$

where $\mathbf{x} = [\mathbf{x}_1^\top, \ldots, \mathbf{x}_N^\top]^\top \in \mathbb{R}^n$ with $n = \sum_{i=1}^N n_i$. Furthermore, we denote $\mathbf{A} = [\mathbf{A}_1 \ldots \mathbf{A}_N] \in \mathbb{R}^{m \times n}$. The constraint $\sum_{i=1}^{N} \mathbf{A}_i \mathbf{x}_i = \mathbf{b}$ of problem (1) takes on the form $\mathbf{A}\mathbf{x} = \mathbf{b}$. Associating Lagrange multipliers $\boldsymbol{\lambda} \in \mathbb{R}^m$ with that constraint, the Lagrange function is defined as

$$\begin{aligned} L(\mathbf{x}, \boldsymbol{\lambda}) &= F(\mathbf{x}) + \langle \boldsymbol{\lambda}, \mathbf{A}\mathbf{x} - \mathbf{b} \rangle \\ &= \sum_{i=1}^{N} L_i(\mathbf{x}_i, \boldsymbol{\lambda}) - \langle \mathbf{b}, \boldsymbol{\lambda} \rangle, \end{aligned} \tag{2}$$

where $L_i(\mathbf{x}_i, \boldsymbol{\lambda}) = f_i(\mathbf{x}_i) + \langle \boldsymbol{\lambda}, \mathbf{A}_i \mathbf{x}_i \rangle$, and $\langle \cdot, \cdot \rangle$ denotes inner product. Then, the dual function is defined as

$$g(\boldsymbol{\lambda}) = \inf_{\mathbf{x} \in \mathcal{X}} \; L(\mathbf{x}, \boldsymbol{\lambda}) = \sum_{i=1}^{N} g_i(\boldsymbol{\lambda}) - \langle \mathbf{b}, \boldsymbol{\lambda} \rangle,$$

where $\mathcal{X} = \mathcal{X}_1 \times \mathcal{X}_2 \cdots \times \mathcal{X}_N$, and

$$g_i(\boldsymbol{\lambda}) = \inf_{\mathbf{x}_i \in \mathcal{X}_i} \Big[f_i(\mathbf{x}_i) + \langle \boldsymbol{\lambda}, \mathbf{A}_i \mathbf{x}_i \rangle \Big].$$

The dual function is decomposable and this gives rise to various decomposition methods addressing the *dual problem*, which is defined by

$$\max_{\boldsymbol{\lambda} \in \mathbb{R}^m} \sum_{i=1}^{N} g_i(\boldsymbol{\lambda}) - \langle \mathbf{b}, \boldsymbol{\lambda} \rangle. \tag{3}$$

Algorithm 1 Augmented Lagrangian Method (ALM)

Set $k = 1$ and define initial Lagrange multipliers $\boldsymbol{\lambda}^1$.

1. For a fixed vector λ^k, calculate $\hat{\mathbf{x}}^k$ as a solution of the problem:
$$\min_{\mathbf{x}\in\mathcal{X}} \ \Lambda_\rho(\mathbf{x}, \boldsymbol{\lambda}^k). \tag{4}$$
2. If the constraints $\sum_{i=1}^N \mathbf{A}_i\hat{\mathbf{x}}_i^k = \mathbf{b}$ are satisfied, then stop (optimal solution found). Otherwise, set :
$$\boldsymbol{\lambda}^{k+1} = \boldsymbol{\lambda}^k + \rho\left(\sum_{i=1}^N \mathbf{A}_i\hat{\mathbf{x}}_i^k - \mathbf{b}\right), \tag{5}$$
Increase k by one and return to Step 1.

Dual methods suffer from well-documented disadvantages, the most notable ones being their exceedingly slow convergence rates and the requirement for strictly convex objective functions. These drawbacks can be alleviated by the augmented Lagrangian framework [1, 40]. The augmented Lagrangian associated with problem (1) is given by

$$\Lambda_\rho(\mathbf{x}, \boldsymbol{\lambda}) \ = \ f(\mathbf{x}) \ + \ \big\langle \boldsymbol{\lambda}, \mathbf{A}\mathbf{x} - \mathbf{b}\big\rangle \ + \ \frac{\rho}{2}\|\mathbf{A}\mathbf{x} - \mathbf{b}\|^2,$$

where $\rho > 0$ is a penalty parameter. We recall the standard augmented Lagrangian method (ALM), also referred to as the "Method of Multipliers" in the literature [1, 40], in Alg. 1.

The convergence of the augmented Lagrangian method is ensured when problem (3) has an optimal solution independently of the starting point $\boldsymbol{\lambda}^1$. Under convexity assumptions and a constraint qualification condition, every accumulation point of the sequence $\{\mathbf{x}^k\}$ is an optimal solution of problem (1). It is also worth mentioning that the augmented Lagrangian method exhibits convergence properties in non-convex settings, cf. [40].

The convergence speed and the numerical advantages of augmented Lagrangian methods (see, e.g., [40, 41]) provide a strong motivation for creating decomposed versions of them. However, achieving such a decomposition is not as straightforward as in the simple dual method, since the quadratic term of the augmented Lagrangian introduces cross-terms between all variables. Early specialized techniques that allow for decomposition of the augmented Lagrangian can be traced back to the works [42, 43]. More recent literature involves the *Diagonal Quadratic Approximation* (DQA) algorithm [5, 6, 44], the *Alternating Direction Method of Multipliers* (ADMM) [1, 3, 4, 45], as well as the *Accelerated Distributed Augmented Lagrangians* (ADAL) method which was recently developed by the authors in [7, 17]. The DQA method replaces each minimization step in the augmented Lagrangian algorithm by a separable approximation of the augmented Lagrangian function. The ADMM methods are based on the relations between splitting methods for monotone operators, such as Douglas-Rachford splitting, and the proximal point algorithm [3, 43]. In what follows we focus our discussion on the ADAL method and briefly describe its algorithmic form and convergence properties. For a more detailed discussion on the differences and similarities between ADAL, DQA, and ADMM, the interested reader is directed to [7].

Algorithm 2 Accelerated Distributed Augmented Lagrangians (ADAL)

Set $k = 1$ and define initial Lagrange multipliers $\boldsymbol{\lambda}^1$ and initial primal variables $\mathbf{x}^1$.

1. For fixed Lagrange multipliers $\boldsymbol{\lambda}^k$, determine $\hat{\mathbf{x}}_i^k$ for every $i \in \mathcal{I}$ as the solution of the following problem:
$$\min_{\mathbf{x}_i\in\mathcal{X}_i} \ \bar{\Lambda}_\rho^i(\mathbf{x}_i, \mathbf{x}^k, \boldsymbol{\lambda}^k). \tag{7}$$
2. Set for every $i \in \mathcal{I}$
$$\mathbf{x}_i^{k+1} = \mathbf{x}_i^k + \tau(\hat{\mathbf{x}}_i^k - \mathbf{x}_i^k). \tag{8}$$
3. If the constraints $\sum_{i=1}^N \mathbf{A}_i\mathbf{x}_i^{k+1} = \mathbf{b}$ are satisfied and $\mathbf{A}_i\hat{\mathbf{x}}_i^k = \mathbf{A}_i\mathbf{x}_i^k$, then stop (optimal solution found). Otherwise, set:
$$\boldsymbol{\lambda}^{k+1} = \boldsymbol{\lambda}^k + \rho\tau\left(\sum_{i=1}^N \mathbf{A}_i\mathbf{x}_i^{k+1} - \mathbf{b}\right), \tag{9}$$
increase k by one and return to Step 1.

A. The ADAL algorithm

The ADAL method is based on defining the *local augmented Lagrangian* function $\bar{\Lambda}_\rho^i : \mathbb{R}^{n_i}\times\mathbb{R}^n\times\mathbb{R}^m \to \mathbb{R}$ for every agent $i = 1, \ldots, N$ at each iteration k, according to

$$\begin{aligned}\bar{\Lambda}_\rho^i(\mathbf{x}_i, \mathbf{x}^k, \boldsymbol{\lambda}) \ &= \ f_i(\mathbf{x}_i) \ + \ \big\langle \boldsymbol{\lambda}, \mathbf{A}_i\mathbf{x}_i\big\rangle \\ &+ \ \frac{\rho}{2}\|\mathbf{A}_i\mathbf{x}_i + \sum_{j\in\mathcal{I}}^{j\neq i}\mathbf{A}_j\mathbf{x}_j^k - \mathbf{b}\|^2.\end{aligned} \tag{6}$$

ADAL has two parameters: a positive penalty parameter ρ and a stepsize parameter $\tau \in (0, 1)$. Each iteration of ADAL is comprised of three steps: i) a minimization step of all the local augmented Lagrangians, ii) an update step for the primal variables, and iii) an update step for the dual variables. The computations at each step are performed in a parallel fashion, so that ADAL resembles a *Jacobi*-type algorithm; see [1] for more details on Jacobi and Gauss-Seidel type algorithms. The ADAL method is summarized in Alg. 2.

At the first step of each iteration, each agent minimizes its local AL subject to its local convex constraints. This computation step requires only local information. To see this, note that the variables $\mathbf{A}_j\mathbf{x}_j^k$, appearing in the penalty term of the local AL (6), correspond to the local primal variables of agent j that were communicated to agent i for the optimization of its local Lagrangian $\bar{\Lambda}_\rho^i$. With respect to agent i, these are considered fixed parameters. The penalty term of each $\bar{\Lambda}_\rho^i$ can be equivalently expressed as

$$\begin{aligned}&\|\mathbf{A}_i\mathbf{x}_i + \sum\nolimits_{j\in\mathcal{I}}^{j\neq i}\mathbf{A}_j\mathbf{x}_j^k - \mathbf{b}\|^2 \\ &\qquad = \ \sum\nolimits_{l=1}^m \Big(\big[\mathbf{A}_i\mathbf{x}_i\big]_l + \sum\nolimits_{j\in\mathcal{I}}^{j\neq i}\big[\mathbf{A}_j\mathbf{x}_j^k\big]_l - b_l\Big)^2,\end{aligned}$$

where $[\mathbf{A}_i]_j$ denotes the j-th row of matrix $\mathbf{A}_i$. The above penalty term is involved only in the minimization computation (7). Hence, for those l such that $[\mathbf{A}_i]_l = \mathbf{0}$, the terms $\sum_{j\in\mathcal{I}}^{j\neq i}\big[\mathbf{A}_j\mathbf{x}_j^k\big]_l - \mathbf{b}_l$ are just constant terms in the minimization

step, and can be excluded. Here, $[\mathbf{A}_i]_l$ denotes the l-th row of $\mathbf{A}_i$ and $\mathbf{0}$ stands for a zero vector of proper dimension. This implies that subproblem i needs access only to the decisions $\big[\mathbf{A}_j\mathbf{x}_j^k\big]_l$ from all subproblems $j \neq i$ that are involved in the same constraints l as i. Moreover, regarding the term $\langle \boldsymbol{\lambda}, \mathbf{A}_i\mathbf{x}_i\rangle$ in (6), we have that $\langle \boldsymbol{\lambda}, \mathbf{A}_i\mathbf{x}_i\rangle = \sum_{j=1}^m \lambda_j[\mathbf{A}_i\mathbf{x}_i]_j$. Hence, we see that, in order to compute (7), each subproblem i needs access only to those λ_j for which $[\mathbf{A}_i]_j \neq \mathbf{0}$. Intuitively speaking, each agent needs access only to the information that is relevant to the constraints that this agent is involved in.

After the local optimization steps have been carried out, the second step consists of each agent updating its primal variables by taking a convex combination with the corresponding values from the previous iteration. This update depends on a stepsize τ which must satisfy $\tau \in (0, \frac{1}{q})$ in order to ensure convergence of the algorithm [7]. Here, q is defined as the *maximum degree*, and is a measure of sparsity of the total constraint matrix $\mathbf{A}$. Specifically, for each constraint $j = 1, \ldots, m$, we introduce a measure of involvement. We denote the number of agents i associated with this constraint by q_j, that is, q_j is the number of all $i \in \mathcal{I} : [\mathbf{A}_i]_j \neq \mathbf{0}$. Here, $\mathbf{0}$ denotes a zero vector of proper dimension. We define q to be the maximum over all q_j, i.e.,

$$q = \max_{1\leq j\leq m} q_j. \tag{10}$$

Intuitively, q is the number of agents coupled in the "most populated" constraint of the problem.

The third and final step of each ADAL iteration consists of the dual update. This step is distributed by structure, since the Lagrange multiplier of the j-th constraint is updated according to $\lambda_j^{k+1} = \lambda_j^k + \rho\tau\big(\sum_{i=1}^N \big[\mathbf{A}_i\mathbf{x}_i^{k+1}\big]_j - b_j\big)$, which implies that the udpate of λ_j needs only information from those i for which $[\mathbf{A}_i]_j \neq \mathbf{0}$. We can define, without loss of generality, a set $\mathcal{M} \subseteq \{1, \ldots, m\}$ of agents that perform the dual updates, such that an agent $j \in \mathcal{M}$ is responsible for the update of the dual variables corresponding to a subset of the coupling constraint set $\mathbf{A}\mathbf{x} = \mathbf{b}$ (without overlapping agents). For example, if the cardinality of $\mathcal{M}$ is equal to the number of constraints m, then each agent $j \in \mathcal{M}$ is responsible for the update of the dual variable of the j-th constraint. In practical settings, $\mathcal{M}$ can be a subset of $\mathcal{I}$, or it can be a separate set of agents, depending on the application.

The convergence of ADAL, relies on the following three assumptions, which are mild, technical, and commonly required in the analysis of convex optimization methods:

(a1) The functions $f_i : \mathbb{R}^{n_i} \to \mathbb{R}$ for $i \in \mathcal{I} = \{1, 2, \ldots, N\}$ are convex, and the sets $\mathcal{X}_i \subseteq \mathbb{R}^{n_i}$ for $i \in \mathcal{I}$ are nonempty closed convex sets.

(a2) The Lagrange function $L(\mathbf{x}, \boldsymbol{\lambda})$, cf. (2), has a saddle point $(\mathbf{x}^*, \boldsymbol{\lambda}^*) \in \mathbb{R}^n \times \mathbb{R}^m$ so that

$$L(\mathbf{x}^*, \boldsymbol{\lambda}) \leq L(\mathbf{x}^*, \boldsymbol{\lambda}^*) \leq L(\mathbf{x}, \boldsymbol{\lambda}^*),\ \forall\, \mathbf{x} \in \mathcal{X},\ \forall\, \boldsymbol{\lambda} \in \mathbb{R}^m.$$

(a3) All subproblems (7) are solvable at every iteration.

Assumption (a2) implies that the point $\mathbf{x}^*$ is a solution of problem (1), the point $\boldsymbol{\lambda}^*$ is a solution of (3), and the strong duality relation holds, i.e., the optimal values of the primal and dual problems are equal. Assumption (a3) is satisfied if for every $i = 1, \ldots, N$, either the set $\mathcal{X}_i$ is compact, or the function $f_i(\mathbf{x}_i) + \frac{\rho}{2}\|\mathbf{A}_i\mathbf{x}_i - \tilde{\mathbf{b}}\|^2$ is inf-compact for any vector $\tilde{\mathbf{b}}$. The latter condition, means that the level sets of the function are compact sets, implying that the set $\{\mathbf{x}_i \in \mathcal{X}_i : f_i(\mathbf{x}_i) + \frac{\rho}{2}\|\mathbf{A}_i\mathbf{x}_i - \tilde{\mathbf{b}}\|^2 \leq \alpha\}$ is compact for any $\alpha \in \mathbb{R}$.

The convergence proof of ADAL hinges on showing that the Lyapunov/Merit function $\phi(\mathbf{x}^k, \boldsymbol{\lambda}^k)$ defined by

$$\begin{aligned}\phi(\mathbf{x}^k, \boldsymbol{\lambda}^k) \;&=\; \rho\sum_{i=1}^N \|\mathbf{A}_i(\mathbf{x}_i^k - \mathbf{x}_i^*)\|^2 \\ &\quad+\; \frac{1}{\rho}\|\boldsymbol{\lambda}^k + \rho(1-\tau)\mathbf{r}(\mathbf{x}^k) - \boldsymbol{\lambda}^*\|^2\end{aligned} \tag{11}$$

is strictly decreasing throughout the iterations k. Here, we define the *residual* $\mathbf{r}(\mathbf{x}) \in \mathbb{R}^m$ as the vector containing the amount of all constraint violations with respect to the primal variable $\mathbf{x}$, i.e.

$$\mathbf{r}(\mathbf{x}) = \sum_{i=1}^N \mathbf{A}_i\mathbf{x}_i - \mathbf{b}.$$

We state the main convergence result of ADAL from [7].

Theorem 1: Assume (a1)–(a3). If the stepsize satisfies $0 < \tau < \frac{1}{q}$, then, the sequence $\{\phi(\mathbf{x}^k, \boldsymbol{\lambda}^k)\}$, is strictly decreasing. Moreover, the ADAL method either stops at an optimal solution of problem (3), or generates a sequence of λ^k converging to an optimal solution of it. Any sequence $\{\mathbf{x}^k\}$ generated by the ADAL algorithm has an accumulation point and any such point is an optimal solution of (1).

III. Stochastic ADAL

In this section we develop the SADAL algorithm that allows for a distributed solution of (1) when: i) the local computation steps are inexact or are performed in the presence of uncertainty, and ii) the message exchanges between agents are corrupted by noise. The basic algorithmic structure of SADAL is essentially the same as that of ADAL. Nevertheless, to account for the presence of uncertainty and noise, appropriate adaptations need to be made. SADAL is summarized in Alg. 3.

In SADAL, each agent $i \in \mathcal{I}$ receives noise-corrupted versions of the actual primal and dual variables. We let $\tilde{\mathbf{x}}_{ij}^k$ and $\tilde{\boldsymbol{\lambda}}_i^k$ denote the noise-corrupted versions of the primal $\mathbf{x}_j^k$ and the dual $\boldsymbol{\lambda}^k$ variables, respectively, as received by agent i at iteration k. Consequently, the local augmented Lagrangian function $\Lambda_\rho^i : \mathbb{R}^{n_i} \times \mathbb{R}^n \times \mathbb{R}^m \to \mathbb{R}$ of each agent $i \in \mathcal{I}$ is now formed based on these noise-corrupted variables, cf. (12), i.e., it takes the form

$$\begin{aligned}\Lambda_i\big(\mathbf{x}_i, \tilde{\mathbf{x}}_i^k, \tilde{\boldsymbol{\lambda}}_i^k, \xi_i^k\big) \;&=\; F_i(\mathbf{x}_i, \xi_i^k) + \big\langle \tilde{\boldsymbol{\lambda}}_i^k, \mathbf{A}_i\mathbf{x}_i\big\rangle \\ &\quad+\; \frac{\rho}{2}\|\mathbf{A}_i\mathbf{x}_i + \sum_{j\in\mathcal{I}}^{j\neq i}\mathbf{A}_j\tilde{\mathbf{x}}_{ij}^k - \mathbf{b}\|^2,\end{aligned}$$

where $\tilde{\mathbf{x}}_i^k = \{\tilde{\mathbf{x}}_{i1}^k, \ldots, \tilde{\mathbf{x}}_{iN}^k\}$ denotes the collection of the noise corrupted variables $\tilde{\mathbf{x}}_{ij}^k$. Note that every local AL is now defined with respect to the function $F_i(\mathbf{x}_i, \xi_i^k)$, which is the

Algorithm 3 Stochastic Accelerated Distributed Augmented Lagrangians (SADAL)

Set $k = 1$ and define initial Lagrange multipliers $\boldsymbol{\lambda}^1$ and initial primal variables $\mathbf{x}^1$.

1. For fixed Lagrange multipliers $\boldsymbol{\lambda}^k$, determine $\hat{\mathbf{x}}_i^k$ for every $i \in \mathcal{I}$ as the solution of the following problem:

$$\begin{aligned} \min_{\mathbf{x}_i} \quad & \Lambda_\rho^i\big(\mathbf{x}_i, \tilde{\mathbf{x}}_i^k, \tilde{\boldsymbol{\lambda}}_i^k, \xi_i^k\big) \\ \text{s.t.} \quad & \mathbf{x}_i \in \mathcal{X}_i \end{aligned} \tag{12}$$

2. Set for every $i \in \mathcal{I}$

$$\mathbf{x}_i^{k+1} \;=\; \mathbf{x}_i^k \;+\; \tau_k\big(\hat{\mathbf{x}}_i^k - \mathbf{x}_i^k\big) \tag{13}$$

$$\mathbf{y}_i^{k+1} \;=\; \mathbf{x}_i^k \;+\; \frac{1}{q}\big(\hat{\mathbf{x}}_i^k - \mathbf{x}_i^k\big) \tag{14}$$

3. If the constraints $\sum_{i=1}^N \mathbf{A}_i \mathbf{x}_i^{k+1} = \mathbf{b}$ are satisfied and $\mathbf{A}_i \hat{\mathbf{x}}_i^k = \mathbf{A}_i \mathbf{x}_i^k$, then stop (optimal solution found). Otherwise, set:

$$\boldsymbol{\lambda}^{k+1} = \boldsymbol{\lambda}^k + \rho\tau_k \left(\sum_{i=1}^N \mathbf{A}_i \tilde{\mathbf{y}}_i^{k+1} - \mathbf{b} \right) \tag{15}$$

increase k by one and return to Step 1.

noise-corrupted version of the true objective function f_i. Here, the term ξ_i^k represents the uncertainty at iteration k.

Moreover, in SADAL the stepsize parameter τ has to be defined as τ_k, cf. (13) and (15), which must be a decreasing, nonnegative sequence that is square summable, but not summable. This decrease property of the stepsize is essential in the vast majority of works within the relevant stochastic approximation literature; see, e.g., [26]. Finally, SADAL introduces the additional auxiliary variables $\mathbf{y}_i^{k+1} = \mathbf{x}_i^k + \frac{1}{q}(\hat{\mathbf{x}}_i^k - \mathbf{x}_i^k)$ that are updated locally at each agent $i \in \mathcal{I}$, cf. (14). Note that the difference between the $\mathbf{y}_i^{k+1}$ and the $\mathbf{x}_i^{k+1}$ updates lies in the stepsize choice; we always use $\frac{1}{q}$ for the $\mathbf{y}_i^{k+1}$ and τ_k for the $\mathbf{x}_i^{k+1}$. The $\mathbf{y}_i^{k+1}$ variables are then used for the dual update step of SADAL, cf. (15).

In what follows, we elaborate on the specific noise terms that appear during the iterative execution of SADAL (12)-(15), such that we provide a specific definition for the notion of uncertainty and noise corruption in our particular setting. We assume that there is a probability space $\big(\Omega, \mathcal{F}, P\big)$, where the set Ω is arbitrary, $\mathcal{F}$ is a σ-algebra of subsets of Ω, and P is a probability measure defined on $\mathcal{F}$. All σ-algebras will be sub-σ-algebras of $\mathcal{F}$, and all random variables will be defined on this space.

Noise in the message exchanges for the formation of the local augmented Lagrangians: At each iteration k, agent i receives, via communication, the noise corrupted primal variables $\mathbf{A}_j \tilde{\mathbf{x}}_{ij}^k$ from agent j, and also the noise corrupted dual variables $\tilde{\boldsymbol{\lambda}}_i^k$ according to

$$\mathbf{A}_j \tilde{\mathbf{x}}_{ij}^k \;=\; \mathbf{A}_j \mathbf{x}_j^k \;+\; \mathbf{v}_{ij}^k \tag{16}$$

$$\tilde{\boldsymbol{\lambda}}_i^k \;=\; \boldsymbol{\lambda}^k \;+\; \mathbf{w}_i^k \tag{17}$$

where the $\mathbf{v}_{ij}^k : \Omega \to \mathbb{R}^m$, and $\mathbf{w}_i^k : \Omega \to \mathbb{R}^m$ are random vectors of appropriate size whose entries are assumed to be i.i.d. random variables with zero mean and bounded variance. Essentially, $\mathbf{v}_{ij}^k$ represents the noise corruption in the communication of the actual primal variables $\mathbf{A}_j \mathbf{x}_j^k$ of agent j towards agent i. Similarly, $\mathbf{w}_i^k$ denotes the noise corruption on the dual variables $\boldsymbol{\lambda}^k$ as perceived by agent i after the corresponding message exchanges.

Note that we formulate the message exchanges with respect to the products $\mathbf{A}_j \mathbf{x}_j^k$, despite the fact that the $\mathbf{x}_j^k$ are the actual variables and the matrices $\mathbf{A}_j$ are essentially problem parameters. This is because each agent i does not need to know the matrices $\mathbf{A}_j$ of the other agents; it is only interested in the products $\mathbf{A}_j \mathbf{x}_j^k$. In fact, it only needs those entries of the vector $\mathbf{A}_j \mathbf{x}_j^k$ which correspond to their common coupling constraints (cf. the pertinent discussion in section II-A).

Noise in the local computations: After receiving the communicated variables, each agent i determines the minimizers $\hat{\mathbf{x}}_i^k$ of its local augmented Lagrangian $\Lambda_\rho^i\big(\mathbf{x}_i, \tilde{\mathbf{x}}_i^k, \tilde{\boldsymbol{\lambda}}_i^k, \xi_i^k\big)$ according to (12). Each $\Lambda_\rho^i\big(\mathbf{x}_i, \tilde{\mathbf{x}}_i^k, \tilde{\boldsymbol{\lambda}}_i^k, \xi_i^k\big)$ contains the function $F_i : \mathbb{R}^{n_i} \times \Omega \to \mathbb{R}$, where ξ_i is an element of the probability space $\big(\Omega, \mathcal{F}, P\big)$. We assume that each $F_i(\cdot, \xi_i)$ is a convex function of $\mathbf{x}_i$ for each $\xi_i \in \Omega$, and that $F_i(\mathbf{x}_i, \cdot)$ is an integrable function of ξ_i for each $\mathbf{x}_i \in \mathbb{R}^{n_i}$, i.e., we assume that $\mathbb{E}\big[|F_i(\mathbf{x}_i, \xi_i)|\big] < \infty$ for each $\mathbf{x}_i \in \mathbb{R}^{n_i}$. We also assume that the functions F_i satisfy

$$f_i(\mathbf{x}_i) = \mathbb{E}\big[F_i(\mathbf{x}_i, \xi_i)\big].$$

The above relation implies that the $f_i(\mathbf{x}_i)$ are convex and that the following relation also holds (see, e.g., [46])

$$\partial f_i(\mathbf{x}_i) = \mathbb{E}\big[\partial F_i(\mathbf{x}_i, \xi_i)\big],$$

where $\partial f_i(\mathbf{x}_i)$ and $\partial_{\mathbf{x}_i} F_i(\mathbf{x}_i, \xi_i)$ denote the convex subdifferentials of the convex functions $f_i(\mathbf{x}_i)$ and $F_i(\mathbf{x}_i, \xi_i)$, respectively, at a point $\mathbf{x}_i$. If we let $\mathbf{s}_{f_i} \in \partial f_i(\mathbf{x}_i)$ denote a subgradient of f_i at the point $\mathbf{x}_i$, and $\mathbf{s}_{F_i} \in \partial F_i(\mathbf{x}_i, \xi_i)$ be a subgradient of F_i with respect to $\mathbf{x}_i$, then $\mathbf{s}_{f_i} = \mathbb{E}\big[\mathbf{s}_{F_i}\big]$ also holds, which can be equivalently expressed as

$$\mathbf{s}_{f_i} = \mathbf{s}_{F_i} + \mathbf{e}_i^k, \tag{18}$$

where the noise vector $\mathbf{e}_i^k : \Omega \to \mathbb{R}^{n_i}$ must satisfy $\mathbb{E}[\mathbf{e}_i^k] = \mathbf{0}$ for all iterations k. Since the functions F_i appear only in the local computation steps (12), the above arguments reveal that, for our particular method, the requirement $f_i(\mathbf{x}_i) = \mathbb{E}\big[F_i(\mathbf{x}_i, \xi_i)\big]$ is equivalent to $\mathbf{s}_{f_i} = \mathbf{s}_{F_i} + \mathbf{e}_i^k$.

Essentially, the aforementioned formulation for noise in the local computations of SADAL enables us to model cases where: i) at each iteration k the agents have access only to noisy observations $F_i(\mathbf{x}_i, \xi_i^k)$ of their true objective functions f_i, or ii) cases where the agents want to optimize the expected values of the F_i's, but have access only to sample values $F_i(\mathbf{x}_i, \xi_i^k)$, or even iii) cases where the subgradients of F_i can only be computed with an error $\mathbf{e}_i^k$ at each iteration k.

Noise in the message exchanges for the dual updates: Similar to the discussion in Section II-A, we assume that there exists a set $\mathcal{M} \subseteq \{1, \ldots, m\}$ of agents that perform the dual updates. After the local minimization and primal update steps have been performed, each agent $i \in \mathcal{I}$ communicates

the updated $\mathbf{A}_i\mathbf{y}_i^{k+1}$ variables to the agents responsible for the dual updates. This message exchange is also corrupted by noise, such that the received messages $\mathbf{A}_i\mathbf{y}_i^{k+1}$ take the form

$$\mathbf{A}_i\tilde{\mathbf{y}}_i^{k+1} = \mathbf{A}_i\mathbf{y}_i^{k+1} + \mathbf{u}_i^{k+1} \tag{19}$$

where $\mathbf{u}_i^{k+1} : \Omega \to \mathbb{R}^m$ is a random vector whose entries are assumed to be i.i.d. random variables with zero mean and bounded variance. Here, the entry $[\mathbf{u}_i^{k+1}]_j$ corresponds to the noise corruption on the respective actual variable $[\mathbf{A}_i]_j\mathbf{y}_i^{k+1}$ as received by agent $j \in \mathcal{M}$ after the corresponding message exchanges.

Note that we formulate the message exchanges (19) with respect to the products $\mathbf{A}_i\tilde{\mathbf{y}}_i$, following the same reasoning as discussed above regarding the message exchanges for the formation of the local ALs (16)-(17). Furthermore, note that in practice if a row j of $\mathbf{A}_i$ is zero, then the corresponding j-th entry of $\mathbf{u}_i$ should also be identically zero (i.e., it has zero mean and variance), since agent i does not need to communicate anything for the update of the dual variable of constraint j.

IV. Convergence Analysis.

In this section we establish the almost sure convergence of SADAL to the optimal solution of (1). We need the following assumptions:

(A1) For every $i \in \mathcal{I}$ consider the function $F_i(\mathbf{x}_i, \xi_i)$, where $F_i : \mathbb{R}^{n_i} \times \Omega \to \mathbb{R}$. We assume that $F_i(\cdot, \xi_i)$ is a convex function of $\mathbf{x}_i$ for each $\xi_i \in \Omega$, and that $F_i(\mathbf{x}_i, \cdot)$ is an integrable function of ξ_i for each $\mathbf{x}_i \in \mathbb{R}^{n_i}$, i.e., $\mathbb{E}\big[|F_i(\mathbf{x}_i, \xi_i)|\big] < \infty$ for each $\mathbf{x}_i \in \mathbb{R}^{n_i}$. Moreover, each F_i satisfies the relation $f_i(\mathbf{x}_i) = \mathbb{E}\big[F_i(\mathbf{x}_i, \xi_i)\big]$, where $f_i : \mathbb{R}^{n_i} \to \mathbb{R}$. It follows that f_i is also convex; see, e.g., [46]. We also assume that the sets $\mathcal{X}_i \subseteq \mathbb{R}^{n_i}$ are nonempty closed, convex sets for all $i \in \mathcal{I}$.

(A2) The Lagrange function $L(\mathbf{x}, \boldsymbol{\lambda})$, cf. (2), has a saddle point $(\mathbf{x}^*, \boldsymbol{\lambda}^*) \in \mathbb{R}^n \times \mathbb{R}^m$:

$$L(\mathbf{x}^*, \boldsymbol{\lambda}) \;\le\; L(\mathbf{x}^*, \boldsymbol{\lambda}^*) \;\le\; L(\mathbf{x}, \boldsymbol{\lambda}^*), \tag{20}$$

for every $\mathbf{x} \in \mathcal{X}$, and $\boldsymbol{\lambda} \in \mathbb{R}^m$.

(A3) All subproblems (12) are solvable at any iteration $k = 1, 2, \ldots$.

(A4) The stepsize sequence τ_k satisfies

$$\tau_k \in (0, \frac{1}{q}) \quad , \quad \sum_{k=1}^{\infty} \tau_k = \infty \quad \text{and} \quad \sum_{k=1}^{\infty} \tau_k^2 < \infty. \tag{21}$$

(A5) The penalty parameter ρ is strictly positive.

(A6) The noise corruption vectors $\mathbf{e}_i^k, \mathbf{v}_{ij}^k, \mathbf{w}_i^k, \mathbf{u}_i^k$, for every $i, j \in \mathcal{I}$, have entries which are i.i.d random variables with zero mean. Moreover, the entries of the noise terms $\mathbf{v}_{ij}^k, \mathbf{w}_i^k, \mathbf{u}_i^k$ have bounded variance.

(A7) The noise terms

$$\boldsymbol{\epsilon}_i^k = \mathbf{w}_i^k + \rho \sum_{j \neq i} \mathbf{v}_{ij}^k \tag{22}$$

satisfy a.s.

$$\sum_{k=1}^{\infty} \tau_k \mathbb{E}_k\big[\|\boldsymbol{\epsilon}_i^k\|^2\big] < \infty, \tag{23}$$

where $\mathbb{E}_k$ denotes conditional expectation with respect to the σ-algebra pertaining to iteration k.

(A8) The noise terms in the local computation steps satisfy

$$\mathbf{e}_i^k \xrightarrow{a.s.} \mathbf{0}.$$

Assumptions (A2) and (A3) are the same as in the deterministic ADAL method. For the sake of clarity, we recall that assumption (A3) is satisfied if for every $i = 1, \ldots, N$, either the set $\mathcal{X}_i$ is compact, or the function $F_i(\mathbf{x}_i, \tilde{\xi}) + \frac{\rho}{2}\|\mathbf{A}_i\mathbf{x}_i - \tilde{\mathbf{b}}\|^2$ is inf-compact for any given $\tilde{\xi}$ and $\tilde{\mathbf{b}}$. The latter condition, means that the level sets of the function are compact sets, implying that the set $\{\mathbf{x}_i \in \mathcal{X}_i : F_i(\mathbf{x}_i, \tilde{\xi}) + \frac{\rho}{2}\|\mathbf{A}_i\mathbf{x}_i - \tilde{\mathbf{b}}\|^2 \le \alpha\}$ is compact for any $\alpha \in \mathbb{R}$.

Assumption (A4) includes the stepsize condition from ADAL, i.e., the fact that $\tau_k < \frac{1}{q}$. Moreover, the conditions that the stepsize sequence should be square-summable, but not summable, are typical in relevant stochastic approximation literature; see, e.g., [26] for a comprehensive overview. Assumption (A5) is the typical assumption that $\rho > 0$, which is necessary in all augmented Lagrangian methods.

Assumptions (A6)-(A8) are necessary to prove the a.s. convergence of SADAL. The zero mean assumption is a common assumption ensuring that the presence of noise does not introduce bias in the computations in the long run, while the bounded variance assumption is a mild technical condition that is needed to guarantee the convergence of the iterative procedure. Assumption (A7) is necessary to establish the a.s. convergence of a supermartingale sequence that we construct to show the a.s. convergence of SADAL; similar assumptions have been used in the existing literature, e.g., [33, 47, 48]. Assumption (A8) is used to guarantee that SADAL indeed converges to the optimal set of the original problem (1); see, e.g., [49]–[51] for a range of applications where this assumption can be valid.

Essentially, assumptions (A6)-(A8) require that the noise corruption terms appearing in the local AL computations vanish in the limit; relation (23) also requires that the noise terms $\boldsymbol{\epsilon}_i^k$ vanish "quickly enough". Note that we do not impose any decrease conditions on the noise terms $\mathbf{u}_i^k$ that appear in the dual update step of SADAL, cf. (15) and (19). An intuitive explanation about the different assumptions on the noise terms can be reached if we recall that in the AL framework we perform gradient ascent in the dual domain, where the gradient of the dual function at each iteration is given by the residual of the constraints. For instance, the AL method (cf. Alg. 1) can be viewed as the proximal point method applied to the dual problem [1]. In classical stochastic gradient descent methods it is not essential that the gradient noise terms vanish in the limit, just that they are unbiased. This is exactly the case here; for the noise terms $\mathbf{u}_i^k$ that directly affect the residual calculation (the gradient of the dual function) we only require that they are unbiased, cf. assumption (A6). However, we cannot guarantee the same unbiased behavior for the noise terms $\mathbf{v}_{ij}^k$, $\mathbf{w}_i^k$, and $\mathbf{e}_i^k$ that appear in the local AL computations, since the effects of noise corruption can propagate and affect the dual gradient in more complicated ways that may cause bias. While this work constitutes a first effort to address the presence of noise within ADAL, it is

certainly an interesting topic to characterize the error bounds, in terms of optimality and feasibility, for scenarios where the aforementioned noise terms do not necessarily vanish in the limit, or even propose alternative algorithms that allow us to relax the decrease conditions.

To avoid cluttering the notation, in what follows we will use the simplified notation $\sum_i$ to denote summation over all $i \in \mathcal{I}$, i.e. $\sum_i = \sum_{i=1}^N$, unless explicitly noted otherwise. We use $\mathcal{N}_{\mathcal{X}}(\mathbf{x})$ to denote the normal cone to the set $\mathcal{X}$ at point $\mathbf{x}$ [40], i.e.,

$$\mathcal{N}_{\mathcal{X}}(\mathbf{x}) = \{\mathbf{h} \in \mathbb{R}^n : \langle \mathbf{h}, \mathbf{y} - \mathbf{x} \rangle \leq 0, \quad \forall\ \mathbf{y} \in \mathcal{X}\}.$$

We also define the auxiliary variables:

$$\hat{\boldsymbol{\lambda}}^k = \boldsymbol{\lambda}^k + \rho \mathbf{r}(\hat{\mathbf{x}}^k), \tag{24}$$

available at iteration k.

Note that, in the following analysis, the various primal and dual variables at each iteration k are essentially functions of the entire history of the generated random process up to that iteration and, hence, are random variables. With that in mind, recall that $(\Omega, \mathcal{F}, P)$ is our probability space, and let $\mathcal{F}_1 \subset \mathcal{F}_2 \subset \ldots$ denote an increasing sequence of sub-σ-algebras of $\mathcal{F}$, with $\mathcal{F}_k$ denoting the sub-σ-algebra pertaining to iteration k

$$\mathcal{F}_k = \sigma\Big(\{\ \mathbf{x}_i^s,\ \mathbf{y}_i^s,\ \boldsymbol{\lambda}^s,\ \bar{\boldsymbol{\lambda}}^s,\ \tilde{\mathbf{y}}_i^s,\ \mathbf{u}_i^s,\ \hat{\mathbf{x}}_i^{s-1},\ \tilde{\mathbf{x}}_i^{s-1},\ \tilde{\boldsymbol{\lambda}}_i^{s-1},\ \xi_i^{s-1}, \mathbf{w}_i^{s-1},\ \mathbf{v}_{ij}^{s-1},\ \boldsymbol{\epsilon}_i^{s-1},\ \mathbf{e}_i^{s-1} :\ i,j \in \mathcal{I},\ 1 \leq s \leq k\ \}\Big). \tag{25}$$

The main idea behind the convergence proof is to show that the sequence $\{\phi(\mathbf{x}^k, \boldsymbol{\lambda}^k)\}$, defined as

$$\phi(\mathbf{x}^k, \boldsymbol{\lambda}^k) \ = \ \rho \sum_i \|\mathbf{A}_i(\mathbf{x}_i^k - \mathbf{x}_i^*)\|^2 + \ \frac{1}{\rho}\|\boldsymbol{\lambda}^k + \rho(1 - \frac{1}{q})\mathbf{r}(\mathbf{x}^k) - \boldsymbol{\lambda}^*\|^2, \tag{26}$$

converges a.s. to some finite random variable $\bar{\phi}$. Note that (26) is almost the same as (11), with the only difference being that we have now replaced τ with its upper bound $\frac{1}{q}$ in the term involving the dual variables. To establish the a.s. convergence of $\{\phi(\mathbf{x}^k, \boldsymbol{\lambda}^k)\}$ in the stochastic setting, we will make use of the following theorem from [52], which is called the *non-negative almost-supermartingale convergence theorem*.

Theorem 2: Let (Ω, F, P) be a probability space and $F_1 \subset F_2 \subset \ldots$ be a sequence of σ-subfields of F. For each $k = 1, 2, \ldots$, let $\zeta_k, \chi_k, \psi_k, \eta_k$ be nonnegative, F_k-measurable random variables such that

$$\mathbb{E}(\zeta_{k+1}|F_k) \ \leq \ (1 + \eta_k)\zeta_k + \chi_k - \psi_k. \tag{27}$$

If $\sum_1^\infty \eta_k < \infty$ and $\sum_1^\infty \chi_k < \infty$ hold, then $\lim_{k\to\infty} \zeta_k$ exists and is finite almost surely, and $\sum_1^\infty \psi_k < \infty$.

Essentially, in what follows we prove that $\{\phi(\mathbf{x}^k, \boldsymbol{\lambda}^k)\}$ is such a non-negative almost-supermartingale. Then, we use this convergence result to infer that, for almost all $\omega \in \Omega$, the sequence of dual variables $\{\lambda^k(\omega)\}$ converges to an optimal solution of problem (3), and that any sequence of primal variables $\{\mathbf{x}^k(\omega)\}$ generated by SADAL has an accumulation point, and any such point is an optimal solution of problem (1). In the following lemma, we utilize the first order optimality conditions for each local subproblem (12) to obtain a first result towards proving the a.s. convergence of the sequence $\{\phi(\mathbf{x}^k, \boldsymbol{\lambda}^k)\}$.

Lemma 1: Assume (A1)-(A3). Then, the following inequality holds:

$$\begin{aligned}\frac{1}{\rho}\Big\langle \hat{\boldsymbol{\lambda}}^k - \boldsymbol{\lambda}^*, \boldsymbol{\lambda}^k - \hat{\boldsymbol{\lambda}}^k \Big\rangle \ \geq \ & \rho \sum_i \Big\langle \mathbf{A}_i(\hat{\mathbf{x}}_i^k - \mathbf{x}_i^*), \sum_{j\neq i} \mathbf{A}_j(\mathbf{x}_j^k - \hat{\mathbf{x}}_j^k) \Big\rangle \\ & + \ \sum_i \Big[\Big\langle \mathbf{A}_i(\hat{\mathbf{x}}_i^k - \mathbf{x}_i^*), \boldsymbol{\epsilon}_i^k \Big\rangle \\ & + f_i(\mathbf{x}_i^*) - f_i(\hat{\mathbf{x}}_i^k) + F_i(\hat{\mathbf{x}}_i^k, \xi_i^k) - F_i(\mathbf{x}_i^*, \xi_i^k) \Big],\end{aligned} \tag{28}$$

where $(\mathbf{x}^*, \boldsymbol{\lambda}^*)$ is a saddle point of the Lagrangian L and $\hat{\boldsymbol{\lambda}}^k$, $\hat{\mathbf{x}}_i^k$ are calculated at iteration k.

Proof: The first order optimality conditions for problem (12) imply the following inclusion for the minimizer $\hat{\mathbf{x}}_i^k$

$$\begin{aligned}0 \ \in \ & \partial_{\mathbf{x}_i} F_i(\hat{\mathbf{x}}_i^k, \xi_i^k) \ + \ \mathbf{A}_i^T \tilde{\boldsymbol{\lambda}}_i^k \\ & + \ \rho \mathbf{A}_i^T \Big(\mathbf{A}_i \hat{\mathbf{x}}_i^k + \sum_{j\neq i} \mathbf{A}_j \tilde{\mathbf{x}}_{ij}^k - \mathbf{b} \Big) \ + \ \mathcal{N}_{\mathcal{X}_i}(\hat{\mathbf{x}}_i^k).\end{aligned} \tag{29}$$

We infer that subgradients $\mathbf{s}_{F_i}^k \in \partial_{\mathbf{x}_i} F_i(\hat{\mathbf{x}}_i^k, \xi_i^k)$ and normal elements $\mathbf{z}_i^k \in \mathcal{N}_{\mathcal{X}_i}(\hat{\mathbf{x}}_i^k)$ exist such that we can express (29) as follows:

$$0 \ = \ \mathbf{s}_{F_i}^k \ + \ \mathbf{A}_i^\top \tilde{\boldsymbol{\lambda}}_i^k \ + \ \rho \mathbf{A}_i^\top \Big(\mathbf{A}_i \hat{\mathbf{x}}_i^k + \sum_{j\neq i} \mathbf{A}_j \tilde{\mathbf{x}}_{ij}^k - \mathbf{b} \Big) \ + \ \mathbf{z}_i^k.$$

Taking inner product with $\mathbf{x}_i^* - \hat{\mathbf{x}}_i^k$ on both sides of this equation and using the definition of a normal cone, we obtain

$$\begin{aligned}\Big\langle \mathbf{s}_{F_i}^k + \mathbf{A}_i^\top \tilde{\boldsymbol{\lambda}}_i^k + \rho \mathbf{A}_i^\top \Big(\mathbf{A}_i \hat{\mathbf{x}}_i^k + \sum_{j\neq i} \mathbf{A}_j \tilde{\mathbf{x}}_{ij}^k - \mathbf{b} \Big), \mathbf{x}_i^* - \hat{\mathbf{x}}_i^k \Big\rangle \\ = \ \Big\langle -\mathbf{z}_i^k, \mathbf{x}_i^* - \hat{\mathbf{x}}_i^k \Big\rangle \ \geq \ 0.\end{aligned} \tag{30}$$

Using the variables $\hat{\boldsymbol{\lambda}}^k$ defined in (24) and also substituting $\tilde{\mathbf{x}}_{ij}^k, \tilde{\boldsymbol{\lambda}}_i^k$ from (16)-(17) in (30), we obtain

$$\begin{aligned}0 \ \leq \ & \Big\langle \mathbf{s}_{F_i}^k + \mathbf{A}_i^T \Big[\hat{\boldsymbol{\lambda}}^k - \rho \sum_j \mathbf{A}_j \hat{\mathbf{x}}_j^k \\ & + \mathbf{w}_i^k + \rho \Big(\mathbf{A}_i \hat{\mathbf{x}}_i^k + \sum_{j\neq i} \mathbf{A}_j \tilde{\mathbf{x}}_{ij}^k \Big) \Big], \mathbf{x}_i^* - \hat{\mathbf{x}}_i^k \Big\rangle \\ = \ & \Big\langle \mathbf{s}_{F_i}^k + \mathbf{A}_i^T \big(\mathbf{w}_i^k + \rho \sum_{j\neq i} \mathbf{v}_{ij}^k \big) \\ & + \mathbf{A}_i^T \Big[\hat{\boldsymbol{\lambda}}^k + \rho \Big(\sum_{j\neq i} \mathbf{A}_j \mathbf{x}_j^k - \sum_{j\neq i} \mathbf{A}_j \hat{\mathbf{x}}_j^k \Big) \Big], \mathbf{x}_i^* - \hat{\mathbf{x}}_i^k \Big\rangle.\end{aligned}$$

From the convexity of F_i, we have that $F_i(\mathbf{x}_i^*, \xi_i^k) - F_i(\hat{\mathbf{x}}_i^k, \xi_i^k) \geq \mathbf{s}_{F_i}(\hat{\mathbf{x}}_i)^T \big(\mathbf{x}_i^* - \hat{\mathbf{x}}_i^k \big)$, so the above inequality can

be expressed as

$$F_i(\mathbf{x}_i^*,\xi_i^k) - F_i(\hat{\mathbf{x}}_i^k,\xi_i^k) + \Big\langle \mathbf{w}_i^k + \rho\sum_{j\neq i}\mathbf{v}_{ij}^k \qquad (31)$$
$$+\hat{\boldsymbol{\lambda}}^k + \rho\Big(\sum_{j\neq i}\mathbf{A}_j\mathbf{x}_j^k - \sum_{j\neq i}\mathbf{A}_j\hat{\mathbf{x}}_j^k\Big), \mathbf{A}_i\big(\mathbf{x}_i^* - \hat{\mathbf{x}}_i^k\big)\Big\rangle \;\geq\; 0.$$

The assumptions (A1) and (A2) entail that the following optimality conditions are satisfied at the point $(\mathbf{x}^*,\boldsymbol{\lambda}^*)$:

$$0 \in \partial f_i(\mathbf{x}_i^*) + \mathbf{A}_i^\top\boldsymbol{\lambda}^* + \mathcal{N}_{\mathcal{X}_i}(\mathbf{x}_i^*), \quad \forall\; i=1,\ldots,N. \quad (32)$$

Inclusion (32) implies that subgradients $\mathbf{s}_{f_i}^* \in \partial f_i(\mathbf{x}_i^*)$ and normal vectors $\mathbf{z}_i^* \in \mathcal{N}_{\mathcal{X}_i}(\mathbf{x}_i^*)$ exist, such that we can express (32) as:

$$0 = \mathbf{s}_{f_i}^* + \mathbf{A}_i^\top\boldsymbol{\lambda}^* + \mathbf{z}_i^*$$

Taking inner product with $\hat{\mathbf{x}}_i^k - \mathbf{x}_i^*$ on both sides of this equation and using the definition of a normal cone, we infer

$$\Big\langle \mathbf{s}_{f_i}^* + \mathbf{A}_i^\top\boldsymbol{\lambda}^*, \hat{\mathbf{x}}_i^k - \mathbf{x}_i^*\Big\rangle \;\geq\; 0, \quad \forall\; i=1,\ldots,N,$$

or, using the convexity of f_i,

$$f_i(\hat{\mathbf{x}}_i^k) - f_i(\mathbf{x}_i^*) + \Big\langle \mathbf{A}_i^\top\boldsymbol{\lambda}^*, \hat{\mathbf{x}}_i^k - \mathbf{x}_i^*\Big\rangle \;\geq\; 0, \qquad (33)$$

for all $i=1,\ldots,N$. Denoting $\boldsymbol{\epsilon}_i^k = \mathbf{w}_i^k + \rho\sum_{j\neq i}\mathbf{v}_{ij}^k$, cf. (22), and combining (31) and (33), we obtain the following inequalities for all $i=1,\ldots,N$:

$$\Big\langle \boldsymbol{\lambda}^* - \hat{\boldsymbol{\lambda}}^k - \boldsymbol{\epsilon}_i^k - \rho\Big(\sum_{j\neq i}\mathbf{A}_j\mathbf{x}_j^k - \sum_{j\neq i}\mathbf{A}_j\hat{\mathbf{x}}_j^k\Big), \mathbf{A}_i\big(\hat{\mathbf{x}}_i^k - \mathbf{x}_i^*\big)\Big\rangle$$
$$\geq\; f_i(\mathbf{x}_i^*) - f_i(\hat{\mathbf{x}}_i^k) + F_i(\hat{\mathbf{x}}_i^k,\xi_i^k) - F_i(\mathbf{x}_i^*,\xi_i^k).$$

Adding the inequalities for all $i=1,\ldots,N$ and rearranging terms, we get:

$$\Big\langle \boldsymbol{\lambda}^* - \hat{\boldsymbol{\lambda}}^k, \sum_i \mathbf{A}_i(\hat{\mathbf{x}}_i^k - \mathbf{x}_i^*)\Big\rangle \;\geq$$
$$\rho\sum_i\Big\langle \mathbf{A}_i\hat{\mathbf{x}}_i^k - \mathbf{A}_i\mathbf{x}_i^*, \sum_{j\neq i}\mathbf{A}_j(\mathbf{x}_j^k - \hat{\mathbf{x}}_j^k)\Big\rangle \qquad (34)$$
$$+\;\sum_i\Big[\Big\langle \mathbf{A}_i(\hat{\mathbf{x}}_i^k - \mathbf{x}_i^*), \boldsymbol{\epsilon}_i^k\Big\rangle$$
$$+ f_i(\mathbf{x}_i^*) - f_i(\hat{\mathbf{x}}_i^k) + F_i(\hat{\mathbf{x}}_i^k,\xi_i^k) - F_i(\mathbf{x}_i^*,\xi_i^k)\Big].$$

Substituting $\sum_{i=1}^N\mathbf{A}_i\mathbf{x}_i^* = \mathbf{b}$ and $\sum_{i=1}^N\mathbf{A}_i\hat{\mathbf{x}}_i^k - \mathbf{b} = \frac{1}{\rho}(\hat{\boldsymbol{\lambda}}^k - \boldsymbol{\lambda}^k)$ in the left hand side of (34), we conclude that

$$\frac{1}{\rho}\Big\langle \hat{\boldsymbol{\lambda}}^k - \boldsymbol{\lambda}^*, \boldsymbol{\lambda}^k - \hat{\boldsymbol{\lambda}}^k\Big\rangle \;\geq$$
$$\rho\sum_i\Big\langle \mathbf{A}_i(\hat{\mathbf{x}}_i^k - \mathbf{x}_i^*), \sum_{j\neq i}\mathbf{A}_j(\mathbf{x}_j^k - \hat{\mathbf{x}}_j^k)\Big\rangle$$
$$+\;\sum_i\Big[\Big\langle \mathbf{A}_i(\hat{\mathbf{x}}_i^k - \mathbf{x}_i^*), \boldsymbol{\epsilon}_i^k\Big\rangle$$
$$+ f_i(\mathbf{x}_i^*) - f_i(\hat{\mathbf{x}}_i^k) + F_i(\hat{\mathbf{x}}_i^k,\xi_i^k) - F_i(\mathbf{x}_i^*,\xi_i^k)\Big],$$

as required. ■

In the next lemma, we further manipulate the result from Lemma 1 to bring us one step closer to proving the a.s. convergence of the sequence $\{\phi(\mathbf{x}^k,\boldsymbol{\lambda}^k)\}$. To avoid cluttering the notation, in what follows we denote the rightmost term in (28) as

$$\alpha^k \;=\; \sum_i\Big[\Big\langle \mathbf{A}_i(\hat{\mathbf{x}}_i^k - \mathbf{x}_i^*), \boldsymbol{\epsilon}_i^k\Big\rangle \qquad (35)$$
$$+ f_i(\mathbf{x}_i^*) - f_i(\hat{\mathbf{x}}_i^k) + F_i(\hat{\mathbf{x}}_i^k,\xi_i^k) - F_i(\mathbf{x}_i^*,\xi_i^k)\Big].$$

Lemma 2: Under assumptions (A1)-(A3), the following estimate holds:

$$\sum_i\rho\Big\langle \mathbf{A}_i(\mathbf{x}_i^k - \mathbf{x}_i^*), \mathbf{A}_i(\mathbf{x}_i^k - \hat{\mathbf{x}}_i^k)\Big\rangle + \frac{1}{\rho}\Big\langle \boldsymbol{\lambda}^k - \boldsymbol{\lambda}^*, \boldsymbol{\lambda}^k - \hat{\boldsymbol{\lambda}}^k\Big\rangle$$
$$\geq\; \sum_i\rho\|\mathbf{A}_i(\mathbf{x}_i^k - \hat{\mathbf{x}}_i^k)\|^2 \;+\; \frac{1}{\rho}\|\hat{\boldsymbol{\lambda}}^k - \boldsymbol{\lambda}^k\|^2 \qquad (36)$$
$$+\;\Big\langle \hat{\boldsymbol{\lambda}}^k - \boldsymbol{\lambda}^k, \mathbf{r}(\mathbf{x}^k) - \mathbf{r}(\hat{\mathbf{x}}^k)\Big\rangle \;+\; \alpha^k.$$

Proof: Add the term $\rho\sum_i\Big\langle \mathbf{A}_i(\hat{\mathbf{x}}_i^k - \mathbf{x}_i^*), \mathbf{A}_i(\mathbf{x}_i^k - \hat{\mathbf{x}}_i^k)\Big\rangle$ to both sides of inequality (28) from Lemma 1, to get

$$\rho\sum_i\Big\langle \mathbf{A}_i(\hat{\mathbf{x}}_i^k - \mathbf{x}_i^*), \mathbf{A}_i(\mathbf{x}_i^k - \hat{\mathbf{x}}_i^k)\Big\rangle + \frac{1}{\rho}\Big\langle \boldsymbol{\lambda}^k - \boldsymbol{\lambda}^*, \boldsymbol{\lambda}^k - \hat{\boldsymbol{\lambda}}^k\Big\rangle$$
$$\geq\; \rho\sum_i\Big\langle \mathbf{A}_i(\hat{\mathbf{x}}_i^k - \mathbf{x}_i^*), \mathbf{A}_i(\mathbf{x}_i^k - \hat{\mathbf{x}}_i^k)\Big\rangle$$
$$+\rho\sum_i\Big\langle \mathbf{A}_i(\hat{\mathbf{x}}_i^k - \mathbf{x}_i^*), \sum_{j\neq i}\mathbf{A}_j(\mathbf{x}_j^k - \hat{\mathbf{x}}_j^k)\Big\rangle \;+\; \alpha^k.$$

Grouping the terms in the right-hand side of the above inequality by their common factor, we transform it as follows:

$$\rho\sum_i\Big\langle \mathbf{A}_i(\hat{\mathbf{x}}_i^k - \mathbf{x}_i^*), \mathbf{A}_i(\mathbf{x}_i^k - \hat{\mathbf{x}}_i^k)\Big\rangle + \frac{1}{\rho}\Big\langle \boldsymbol{\lambda}^k - \boldsymbol{\lambda}^*, \boldsymbol{\lambda}^k - \hat{\boldsymbol{\lambda}}^k\Big\rangle$$
$$\geq\; \rho\sum_i\Big\langle \mathbf{A}_i(\hat{\mathbf{x}}_i^k - \mathbf{x}_i^*), \sum_i\mathbf{A}_i(\mathbf{x}_i^k - \hat{\mathbf{x}}_i^k)\Big\rangle \;+\; \alpha^k. \qquad (37)$$

Recall that $\sum_i\mathbf{A}_i(\mathbf{x}_i^k - \hat{\mathbf{x}}_i^k) = \mathbf{r}(\mathbf{x}^k) - \mathbf{r}(\hat{\mathbf{x}}^k)$. This term does not depend on the summation over i in the right hand side of (37). Moreover, $\sum_i\mathbf{A}_i\mathbf{x}_i^* = \mathbf{b}$. Substituting these terms at the right-hand side of (37), yields

$$\rho\sum_i\Big\langle \mathbf{A}_i(\hat{\mathbf{x}}_i^k - \mathbf{x}_i^*), \mathbf{A}_i(\mathbf{x}_i^k - \hat{\mathbf{x}}_i^k)\Big\rangle + \frac{1}{\rho}\Big\langle \boldsymbol{\lambda}^k - \boldsymbol{\lambda}^*, \boldsymbol{\lambda}^k - \hat{\boldsymbol{\lambda}}^k\Big\rangle$$
$$\geq\; \rho\Big\langle \sum_i\mathbf{A}_i(\hat{\mathbf{x}}_i^k - \mathbf{x}_i^*), \mathbf{r}(\mathbf{x}^k) - \mathbf{r}(\hat{\mathbf{x}}^k)\Big\rangle \;+\; \alpha^k$$
$$=\; \rho\Big\langle \sum_i\mathbf{A}_i\hat{\mathbf{x}}_i^k - \mathbf{b}, \mathbf{r}(\mathbf{x}^k) - \mathbf{r}(\hat{\mathbf{x}}^k)\Big\rangle \;+\; \alpha^k$$
$$=\; \Big\langle \hat{\boldsymbol{\lambda}}^k - \boldsymbol{\lambda}^k, \mathbf{r}(\mathbf{x}^k) - \mathbf{r}(\hat{\mathbf{x}}^k)\Big\rangle \;+\; \alpha^k. \qquad (38)$$

In a last step, we represent

$$(\mathbf{A}_i\hat{\mathbf{x}}_i^k - \mathbf{A}_i\mathbf{x}_i^*) = (\mathbf{A}_i\mathbf{x}_i^k - \mathbf{A}_i\mathbf{x}_i^*) + (\mathbf{A}_i\hat{\mathbf{x}}_i^k - \mathbf{A}_i\mathbf{x}_i^k)$$
$$\text{and} \quad \hat{\boldsymbol{\lambda}}^k - \boldsymbol{\lambda}^* = (\boldsymbol{\lambda}^k - \boldsymbol{\lambda}^*) + (\hat{\boldsymbol{\lambda}}^k - \boldsymbol{\lambda}^k)$$

in the left-hand side of (38). We obtain

$$\begin{aligned}\sum_i \rho\Big\langle \mathbf{A}_i(\mathbf{x}_i^k - \mathbf{x}_i^*), \mathbf{A}_i(\mathbf{x}_i^k - \hat{\mathbf{x}}_i^k)\Big\rangle + \frac{1}{\rho}\Big\langle \boldsymbol{\lambda}^k - \boldsymbol{\lambda}^*, \boldsymbol{\lambda}^k - \hat{\boldsymbol{\lambda}}^k\Big\rangle \\ \geq \sum_i \rho\|\mathbf{A}_i(\mathbf{x}_i^k - \hat{\mathbf{x}}_i^k)\|^2 + \frac{1}{\rho}\|\hat{\boldsymbol{\lambda}}^k - \boldsymbol{\lambda}^k\|^2 \\ + \Big\langle \hat{\boldsymbol{\lambda}}^k - \boldsymbol{\lambda}^k, \mathbf{r}(\mathbf{x}^k) - \mathbf{r}(\hat{\mathbf{x}}^k)\Big\rangle + \alpha^k,\end{aligned}$$

as required. ■

To avoid cluttering the notation, in what follows we use the following variable

$$\bar{\boldsymbol{\lambda}}^k = \boldsymbol{\lambda}^k + \rho(1 - \frac{1}{q})\mathbf{r}(\mathbf{x}^k). \tag{39}$$

Note that $\bar{\boldsymbol{\lambda}}^k$ appears in the term containing the dual variables of our stochastic Lyapunov/Merit function $\phi(\mathbf{x}^k, \boldsymbol{\lambda}^k)$, cf. (26).

Lemma 3: Under the assumptions (A1)-(A3), the following estimate holds

$$\begin{aligned}\sum_i \rho\Big\langle \mathbf{A}_i(\mathbf{x}_i^k - \mathbf{x}_i^*), \mathbf{A}_i(\mathbf{x}_i^k - \hat{\mathbf{x}}_i^k)\Big\rangle + \frac{1}{\rho}\Big\langle \bar{\boldsymbol{\lambda}}^k - \boldsymbol{\lambda}^*, \boldsymbol{\lambda}^k - \hat{\boldsymbol{\lambda}}^k\Big\rangle \\ \geq \sum_i \frac{\rho}{2}\|\mathbf{A}_i(\mathbf{x}_i^k - \hat{\mathbf{x}}^k)\|^2 + \frac{1}{2q\rho}\|\boldsymbol{\lambda}^k - \hat{\boldsymbol{\lambda}}^k\|^2 + \alpha^k,\end{aligned}$$

where the $\bar{\boldsymbol{\lambda}}^k$ are defined in (39).

Proof: Add the term $\rho(1-\frac{1}{q})\Big\langle \mathbf{r}(\mathbf{x}^k), \frac{1}{\rho}(\boldsymbol{\lambda}^k - \hat{\boldsymbol{\lambda}}^k)\Big\rangle = \rho(1-\frac{1}{q})\Big\langle \mathbf{r}(\mathbf{x}^k), -\mathbf{r}(\hat{\mathbf{x}}^k)\Big\rangle$ to inequality (36) to get:

$$\begin{aligned}\sum_i \rho\Big\langle \mathbf{A}_i(\mathbf{x}_i^k - \mathbf{x}_i^*), \mathbf{A}_i(\mathbf{x}_i^k - \hat{\mathbf{x}}_i^k)\Big\rangle + \frac{1}{\rho}\Big\langle \bar{\boldsymbol{\lambda}}^k - \boldsymbol{\lambda}^*, \boldsymbol{\lambda}^k - \hat{\boldsymbol{\lambda}}^k\Big\rangle \\ \geq \sum_i \rho\|\mathbf{A}_i(\mathbf{x}_i^k - \hat{\mathbf{x}}_i^k)\|^2 + \frac{1}{\rho}\|\hat{\boldsymbol{\lambda}}^k - \boldsymbol{\lambda}^k\|^2 + \alpha^k \\ + \Big\langle \hat{\boldsymbol{\lambda}}^k - \boldsymbol{\lambda}^k, \mathbf{r}(\mathbf{x}^k) - \mathbf{r}(\hat{\mathbf{x}}^k)\Big\rangle - \rho(1 - \frac{1}{q})\Big\langle \mathbf{r}(\mathbf{x}^k), \mathbf{r}(\hat{\mathbf{x}}^k)\Big\rangle.\end{aligned} \tag{40}$$

Isolate the term $\Big\langle \hat{\boldsymbol{\lambda}}^k - \boldsymbol{\lambda}^k, \mathbf{r}(\mathbf{x}^k) - \mathbf{r}(\hat{\mathbf{x}}^k)\Big\rangle - \rho(1 - \frac{1}{q})\Big\langle \mathbf{r}(\mathbf{x}^k), \mathbf{r}(\hat{\mathbf{x}}^k)\Big\rangle$ at the right hand side for a bit. We manipulate it to yield:

$$\begin{aligned}&\Big\langle \hat{\boldsymbol{\lambda}}^k - \boldsymbol{\lambda}^k, \mathbf{r}(\mathbf{x}^k) - \mathbf{r}(\hat{\mathbf{x}}^k)\Big\rangle - \rho(1 - \frac{1}{q})\Big\langle \mathbf{r}(\mathbf{x}^k), \mathbf{r}(\hat{\mathbf{x}}^k)\Big\rangle \\ &= \rho\Big\langle \mathbf{r}(\hat{\mathbf{x}}^k), \mathbf{r}(\mathbf{x}^k) - \mathbf{r}(\hat{\mathbf{x}}^k)\Big\rangle - \rho(1 - \frac{1}{q})\Big\langle \mathbf{r}(\mathbf{x}^k), \mathbf{r}(\hat{\mathbf{x}}^k)\Big\rangle \\ &= \rho\Big\langle \mathbf{r}(\hat{\mathbf{x}}^k), \mathbf{r}(\mathbf{x}^k) - \mathbf{r}(\hat{\mathbf{x}}^k)\Big\rangle \\ &\qquad - \rho(1 - \frac{1}{q})\Big\langle \mathbf{r}(\mathbf{x}^k) - \mathbf{r}(\hat{\mathbf{x}}^k) + \mathbf{r}(\hat{\mathbf{x}}^k), \mathbf{r}(\hat{\mathbf{x}}^k)\Big\rangle \\ &= \frac{1}{q}\rho\Big\langle \mathbf{r}(\hat{\mathbf{x}}^k), \mathbf{r}(\mathbf{x}^k) - \mathbf{r}(\hat{\mathbf{x}}^k)\Big\rangle - (1 - \frac{1}{q})\rho\|\mathbf{r}(\hat{\mathbf{x}}^k)\|^2 \\ &= \frac{1}{q}\Big\langle \hat{\boldsymbol{\lambda}}^k - \boldsymbol{\lambda}^k, \sum_i \mathbf{A}_i(\mathbf{x}_i^k - \hat{\mathbf{x}}_i^k)\Big\rangle - (1 - \frac{1}{q})\frac{1}{\rho}\|\hat{\boldsymbol{\lambda}}^k - \boldsymbol{\lambda}^k\|^2.\end{aligned}$$

Then, (40) becomes:

$$\begin{aligned}\sum_i \rho\Big\langle \mathbf{A}_i(\mathbf{x}_i^k - \mathbf{x}_i^*), \mathbf{A}_i(\mathbf{x}_i^k - \hat{\mathbf{x}}_i^k)\Big\rangle + \frac{1}{\rho}\Big\langle \bar{\boldsymbol{\lambda}}^k - \boldsymbol{\lambda}^*, \boldsymbol{\lambda}^k - \hat{\boldsymbol{\lambda}}^k\Big\rangle \\ \geq \sum_i \rho\|\mathbf{A}_i(\mathbf{x}_i^k - \hat{\mathbf{x}}_i^k)\|^2 + \frac{1}{q\rho}\|\hat{\boldsymbol{\lambda}}^k - \boldsymbol{\lambda}^k\|^2 \\ + \frac{1}{q}\Big\langle \hat{\boldsymbol{\lambda}}^k - \boldsymbol{\lambda}^k, \sum_i \mathbf{A}_i(\mathbf{x}_i^k - \hat{\mathbf{x}}_i^k)\Big\rangle + \alpha^k.\end{aligned} \tag{41}$$

The terms $\frac{1}{q}\Big\langle \hat{\boldsymbol{\lambda}}^k - \boldsymbol{\lambda}^k, \mathbf{A}_i(\mathbf{x}_i^k - \hat{\mathbf{x}}_i^k)\Big\rangle$ can be bounded below by considering

$$\begin{aligned}&\frac{1}{q}\Big\langle \hat{\boldsymbol{\lambda}}^k - \boldsymbol{\lambda}^k, \mathbf{A}_i(\mathbf{x}_i^k - \hat{\mathbf{x}}_i^k)\Big\rangle \\ &\qquad \geq -\frac{1}{2}\Big(\rho||\mathbf{A}_i(\mathbf{x}_i^k - \hat{\mathbf{x}}_i^k)||^2 + \frac{1}{q^2\rho}||\hat{\boldsymbol{\lambda}}^k - \boldsymbol{\lambda}^k||^2\Big).\end{aligned}$$

Summing the inequality over all i, we observe that the quantity $|\hat{\boldsymbol{\lambda}}_j^k - \boldsymbol{\lambda}_j^k|^2$, where λ_j indicates the Lagrange multiplier of the j-th constraint, appears at most q times. This is because

$$\begin{aligned}&\sum_{i=1}^N \frac{1}{q}\Big\langle \hat{\boldsymbol{\lambda}}^k - \boldsymbol{\lambda}^k, \mathbf{A}_i(\mathbf{x}_i^k - \hat{\mathbf{x}}_i^k)\Big\rangle \\ &\qquad = \frac{1}{q}\sum_{i=1}^N\sum_{j=1}^m (\hat{\boldsymbol{\lambda}}_j^k - \boldsymbol{\lambda}_j^k)\big[\mathbf{A}_i(\mathbf{x}_i^k - \hat{\mathbf{x}}_i^k)\big]_j \\ &\qquad = \frac{1}{q}\sum_{j=1}^m (\hat{\boldsymbol{\lambda}}_j^k - \boldsymbol{\lambda}_j^k)\sum_{i=1}^N \big[\mathbf{A}_i(\mathbf{x}_i^k - \hat{\mathbf{x}}_i^k)\big]_j.\end{aligned}$$

Thus, recalling that q denotes the maximum number of non-zero blocks $[\mathbf{A}_i]_j$ over all j, we can conclude that each term $\|\hat{\boldsymbol{\lambda}}_j^k - \boldsymbol{\lambda}_j^k\|^2,\ j = 1, \ldots, m$ appears at most q times in the summation. This observation leads us to

$$\begin{aligned}&\sum_{i=1}^N \frac{1}{q}\Big\langle \hat{\boldsymbol{\lambda}}^k - \boldsymbol{\lambda}^k, \mathbf{A}_i(\mathbf{x}_i^k - \hat{\mathbf{x}}_i^k)\Big\rangle \\ &\qquad \geq -\frac{1}{2}\Big(\sum_i \rho\|\mathbf{A}_i(\mathbf{x}_i^k - \hat{\mathbf{x}}_i^k)\|^2 + \frac{1}{q\rho}\|\hat{\boldsymbol{\lambda}}^k - \boldsymbol{\lambda}^k\|^2\Big).\end{aligned} \tag{42}$$

Finally, substituting (42) into (41) we get

$$\begin{aligned}\sum_i \rho\Big\langle \mathbf{A}_i(\mathbf{x}_i^k - \mathbf{x}_i^*), \mathbf{A}_i(\mathbf{x}_i^k - \hat{\mathbf{x}}_i^k)\Big\rangle + \frac{1}{\rho}\Big\langle \bar{\boldsymbol{\lambda}}^k - \boldsymbol{\lambda}^*, \boldsymbol{\lambda}^k - \hat{\boldsymbol{\lambda}}^k\Big\rangle \\ \geq \sum_i \frac{\rho}{2}\|\mathbf{A}_i(\mathbf{x}_i^k - \hat{\mathbf{x}}_i^k)\|^2 + \frac{1}{2q\rho}\|\boldsymbol{\lambda}^k - \hat{\boldsymbol{\lambda}}^k\|^2 + \alpha^k,\end{aligned}$$

which completes the proof. ■

Now we are ready to prove the a.s. convergence of our Lyapunov/Merit function $\phi(\mathbf{x}^k, \boldsymbol{\lambda}^k)$ by utilizing the almost-supermartingale convergence result from Theorem 2.

Lemma 4: Assume (A1)-(A7). Then, the sequence

$$\phi(\mathbf{x}^k, \boldsymbol{\lambda}^k) = \sum_{i=1}^N \rho\|\mathbf{A}_i(\mathbf{x}_i^k - \mathbf{x}_i^*)\|^2 + \frac{1}{\rho}\|\bar{\boldsymbol{\lambda}}^k - \boldsymbol{\lambda}^*\|^2 \tag{43}$$

generated by SADAL converges a.s. to some finite random variable $\bar{\phi}$. Moreover, for all $i = 1, \ldots, N$ we have

$$\mathbf{r}(\hat{\mathbf{x}}^k) \xrightarrow{L^2} \mathbf{0}$$
$$\text{and } \mathbf{A}_i\hat{\mathbf{x}}_i^k \xrightarrow{L^2} \mathbf{A}_i\mathbf{x}_i^k,$$

where $\xrightarrow{L^2}$ denotes mean-square convergence.

Proof: First, we show that the dual update step (15) in the SADAL method results in the following update rule for the variables $\bar{\boldsymbol{\lambda}}^k$, which are defined in (39):

$$\bar{\boldsymbol{\lambda}}^{k+1} = \bar{\boldsymbol{\lambda}}^k + \tau_k\rho\mathbf{r}(\hat{\mathbf{x}}^k) + \tau_k\rho\sum_i \mathbf{u}_i^{k+1} \tag{44}$$

Indeed,

$$\begin{aligned}
\boldsymbol{\lambda}^{k+1} &= \boldsymbol{\lambda}^k + \tau_k\rho\mathbf{r}(\tilde{\mathbf{y}}^{k+1}) \\
&= \boldsymbol{\lambda}^k + \tau_k\rho\Big[\mathbf{r}(\mathbf{y}^{k+1}) + \rho\sum_i \mathbf{u}_i^{k+1}\Big] \\
&= \boldsymbol{\lambda}^k + \tau_k\rho\Big[(1-\frac{1}{q})\mathbf{r}(\mathbf{x}^k) + \frac{1}{q}\mathbf{r}(\hat{\mathbf{x}}^k) + \rho\sum_i \mathbf{u}_i^{k+1}\Big] \\
&= \boldsymbol{\lambda}^k + \tau_k\Big[-(1-\frac{1}{q})\rho\Big(\mathbf{r}(\hat{\mathbf{x}}^k) - \mathbf{r}(\mathbf{x}^k)\Big) \\
&\qquad + \rho\mathbf{r}(\hat{\mathbf{x}}^k) + \rho\sum_i \mathbf{u}_i^{k+1}\Big] \\
&= \boldsymbol{\lambda}^k - (1-\frac{1}{q})\rho\tau_k\Big(\mathbf{r}(\hat{\mathbf{x}}^k) - \mathbf{r}(\mathbf{x}^k)\Big) \\
&\qquad + \tau_k\rho\mathbf{r}(\hat{\mathbf{x}}^k) + \tau_k\rho\sum_i \mathbf{u}_i^{k+1},
\end{aligned}$$

where the third equality follows from the definition of the $\mathbf{y}^{k+1}$ variables in the primal update step of SADAL, cf. (13). Adding $(1-\frac{1}{q})\rho\mathbf{r}(\mathbf{x}^k)$ on both sides of the above relation and rearranging terms, we obtain

$$\begin{aligned}
&\boldsymbol{\lambda}^{k+1} + (1-\frac{1}{q})\rho\Big[\mathbf{r}(\mathbf{x}^k) + \tau_k\Big(\mathbf{r}(\hat{\mathbf{x}}^k) - \mathbf{r}(\mathbf{x}^k)\Big)\Big] \\
&\quad = \boldsymbol{\lambda}^k + (1-\frac{1}{q})\rho\mathbf{r}(\mathbf{x}^k) + \tau_k\rho\mathbf{r}(\hat{\mathbf{x}}^k) + \tau_k\rho\sum_i \mathbf{u}_i^{k+1}.
\end{aligned}$$

The left hand side of the above is equal to $\bar{\boldsymbol{\lambda}}^{k+1}$ by definition (39) and the fact that $\mathbf{r}(\mathbf{x}^{k+1}) = \mathbf{r}(\mathbf{x}^k) + \tau_k\big(\mathbf{r}(\hat{\mathbf{x}}^k) - \mathbf{r}(\mathbf{x}^k)\big)$. Hence, we arrive at

$$\begin{aligned}
\bar{\boldsymbol{\lambda}}^{k+1} &= \boldsymbol{\lambda}^k + (1-\frac{1}{q})\rho\mathbf{r}(\mathbf{x}^k) + \tau_k\rho\mathbf{r}(\hat{\mathbf{x}}^k) + \tau_k\rho\sum_i \mathbf{u}_i^{k+1} \\
&= \bar{\boldsymbol{\lambda}}^k + \tau_k\rho\mathbf{r}(\hat{\mathbf{x}}^k) + \tau_k\rho\sum_i \mathbf{u}_i^{k+1},
\end{aligned}$$

as required.

Using (44), we can now evaluate $\phi(\mathbf{x}^{k+1}, \bar{\boldsymbol{\lambda}}^{k+1})$ as

$$\begin{aligned}
&\phi\big(\mathbf{x}^{k+1}, \bar{\boldsymbol{\lambda}}^{k+1}\big) = \\
&= \sum_{i=1}^N \rho||\mathbf{A}_i(\mathbf{x}_i^{k+1} - \mathbf{x}_i^*)||^2 + \frac{1}{\rho}||\bar{\boldsymbol{\lambda}}^{k+1} - \boldsymbol{\lambda}^*||^2 \\
&= \sum_{i=1}^N \rho||\mathbf{A}_i(\mathbf{x}_i^k - \mathbf{x}_i^*) + \tau_k\mathbf{A}_i(\hat{\mathbf{x}}_i^k - \mathbf{x}_i^k)||^2 \\
&\qquad + \frac{1}{\rho}||\bar{\boldsymbol{\lambda}}^k - \boldsymbol{\lambda}^* + \tau_k\rho\mathbf{r}(\hat{\mathbf{x}}^k) + \tau_k\rho\sum_i \mathbf{u}_i^{k+1}||^2.
\end{aligned}$$

Expanding the right hand side of the above relation, we get

$$\begin{aligned}
&\phi\big(\mathbf{x}^{k+1}, \bar{\boldsymbol{\lambda}}^{k+1}\big) = \sum_{i=1}^N \rho||\mathbf{A}_i(\mathbf{x}_i^k - \mathbf{x}_i^*)||^2 + \frac{1}{\rho}||\bar{\boldsymbol{\lambda}}^k - \boldsymbol{\lambda}^*||^2 \\
&- 2\tau_k\Bigg[\rho\sum_{i=1}^N \Big\langle \mathbf{A}_i(\mathbf{x}_i^k - \mathbf{x}_i^*), \mathbf{A}_i(\mathbf{x}_i^k - \hat{\mathbf{x}}_i^k)\Big\rangle \\
&\qquad - \frac{1}{\rho}\Big\langle \bar{\boldsymbol{\lambda}}^k - \boldsymbol{\lambda}^*, \rho\mathbf{r}(\hat{\mathbf{x}}^k) + \rho\sum_i \mathbf{u}_i^{k+1}\Big\rangle\Bigg] \\
&+ \tau_k^2\Bigg[\sum_{i=1}^N \rho||\mathbf{A}_i(\hat{\mathbf{x}}_i^k - \mathbf{x}_i^k)||^2 + \frac{1}{\rho}||\rho\mathbf{r}(\hat{\mathbf{x}}^k) + \rho\sum_i \mathbf{u}_i^{k+1}||^2\Bigg].
\end{aligned}$$

After expanding the very last term on the above and recalling the definition $\hat{\boldsymbol{\lambda}}^k = \boldsymbol{\lambda}^k + \rho\mathbf{r}(\hat{\mathbf{x}}^k)$, cf. (24), we arrive at

$$\begin{aligned}
&\phi\big(\mathbf{x}^{k+1}, \bar{\boldsymbol{\lambda}}^{k+1}\big) = \phi\big(\mathbf{x}^k, \bar{\boldsymbol{\lambda}}^k\big) \\
&- 2\tau_k\Bigg[\rho\sum_{i=1}^N \Big\langle \mathbf{A}_i(\mathbf{x}_i^k - \mathbf{x}_i^*), \mathbf{A}_i(\mathbf{x}_i^k - \hat{\mathbf{x}}_i^k)\Big\rangle \\
&\qquad - \frac{1}{\rho}\Big\langle \bar{\boldsymbol{\lambda}}^k - \boldsymbol{\lambda}^*, \rho\mathbf{r}(\hat{\mathbf{x}}^k) + \rho\sum_i \mathbf{u}_i^{k+1}\Big\rangle\Bigg] \\
&+ \tau_k^2\Bigg[\sum_i \rho||\mathbf{A}_i(\hat{\mathbf{x}}_i^k - \mathbf{x}_i^k)||^2 + \frac{1}{\rho}||\hat{\boldsymbol{\lambda}}^k - \boldsymbol{\lambda}^k||^2 \\
&\qquad + \frac{2}{\rho}\Big\langle \hat{\boldsymbol{\lambda}}^k - \boldsymbol{\lambda}^k, \rho\sum_i \mathbf{u}_i^{k+1}\Big\rangle + \frac{1}{\rho}\|\rho\sum_i \mathbf{u}_i^{k+1}\|^2\Bigg].
\end{aligned}$$

We use Lemma 3 to substitute the term $-2\tau_k\rho\sum_{i=1}^N \big\langle \mathbf{A}_i(\mathbf{x}_i^k - \mathbf{x}_i^*), \mathbf{A}_i(\mathbf{x}_i^k - \hat{\mathbf{x}}_i^k)\big\rangle - \frac{2\tau_k}{\rho}\big\langle \bar{\boldsymbol{\lambda}}^k - \boldsymbol{\lambda}^*, \rho\mathbf{r}(\hat{\mathbf{x}}^k)\big\rangle$ with its lower bound in the above relation, to arrive at

$$\begin{aligned}
&\phi\big(\mathbf{x}^{k+1}, \boldsymbol{\lambda}^{k+1}\big) \le \phi\big(\mathbf{x}^k, \boldsymbol{\lambda}^k\big) - \\
&\sum_i \rho(\tau_k - \tau_k^2)||\mathbf{A}_i(\mathbf{x}_i^k - \hat{\mathbf{x}}_i^k)||^2 - (\frac{\tau_k}{q} - \tau_k^2)\frac{1}{\rho}||\boldsymbol{\lambda}^k - \hat{\boldsymbol{\lambda}}^k||^2 \\
&+ 2\tau_k\Big\langle \bar{\boldsymbol{\lambda}}^k - \boldsymbol{\lambda}^*, \sum_i \mathbf{u}_i^{k+1}\Big\rangle + 2\tau_k^2\Big\langle \hat{\boldsymbol{\lambda}}^k - \boldsymbol{\lambda}^k, \sum_i \mathbf{u}_i^{k+1}\Big\rangle \\
&+ \rho\tau_k^2\|\rho\sum_i \mathbf{u}_i^{k+1}\|^2 - 2\tau_k\alpha^k. \qquad (45)
\end{aligned}$$

Now, take the conditional expectation $\mathbb{E}_k$ (with respect to $\mathcal{F}_k$) on the above relation. First, by the definition of $\mathcal{F}_k$ in (25), we get that

$$\mathbb{E}_k\Big\langle \bar{\boldsymbol{\lambda}}^k - \boldsymbol{\lambda}^*, \sum_i \mathbf{u}_i^{k+1}\Big\rangle = \Big\langle \bar{\boldsymbol{\lambda}}^k - \boldsymbol{\lambda}^*, \mathbb{E}_k\sum_i \mathbf{u}_i^{k+1}\Big\rangle = 0, \tag{46}$$

where in the last equality we used the zero-mean assumption (A6) for $\mathbf{u}^{k+1}$. Moreover, it is true that

$$\mathbb{E}_k\Big\langle \hat{\boldsymbol{\lambda}}^k - \boldsymbol{\lambda}^k, \sum_i \mathbf{u}_i^{k+1}\Big\rangle \;=\; \mathbb{E}_k \rho \mathbf{r}(\hat{\mathbf{x}}^k)^T \;\mathbb{E}_k \sum_i \mathbf{u}_i^{k+1} \;=\; 0. \tag{47}$$

This follows from the following facts. The terms $\mathbf{r}(\hat{\mathbf{x}}^k)$, and $\sum_i \mathbf{u}_i^{k+1}$ are conditionally independent (recall from (19) that $\mathbf{u}^{k+1}$ denotes the noise in the message exchanges for the dual updates). Moreover, due to assumption (A3) the minimizers $\hat{\mathbf{x}}_i^k$ belong to compact sets for all $i = 1, \ldots, N$, hence we have that $\mathbb{E}_k \mathbf{r}(\hat{\mathbf{x}}^k) < \infty$. Finally, assumption (A6) implies that $\mathbb{E}_k \sum_i \mathbf{u}_i^{k+1} = \mathbf{0}$.

Thus, after taking the conditional expectation of $\phi\big(\mathbf{x}^{k+1}, \boldsymbol{\lambda}^{k+1}\big)$ with respect to $\mathcal{F}_k$, and using (45)-(47), we get that

$$\begin{aligned}
\mathbb{E}_k \phi\big(\mathbf{x}^{k+1}, \boldsymbol{\lambda}^{k+1}\big) \;&\leq\; \phi\big(\mathbf{x}^k, \boldsymbol{\lambda}^k\big) \\
-\; \mathbb{E}_k \Bigg[&\sum_i \rho(\tau_k - \tau_k^2) ||\mathbf{A}_i(\mathbf{x}_i^k - \hat{\mathbf{x}}_i^k)||^2 \\
+\; \big(\frac{\tau_k}{q} - \tau_k^2\big)\frac{1}{\rho} &||\boldsymbol{\lambda}^k - \hat{\boldsymbol{\lambda}}^k||^2 - \tau_k^2 \rho \| \sum_i \mathbf{u}_i^{k+1}\|^2 + 2\tau_k \alpha^k \Bigg].
\end{aligned}$$

Consider the term $-2\tau_k \mathbb{E}_k \alpha^k$, and recall from (35) that $\alpha^k = \sum_i \Big[\big\langle \mathbf{A}_i(\hat{\mathbf{x}}_i^k - \mathbf{x}_i^*), \boldsymbol{\epsilon}_i^k\big\rangle + f_i(\mathbf{x}_i^*) - f_i(\hat{\mathbf{x}}_i^k) + F_i(\hat{\mathbf{x}}_i^k, \xi_i) - F_i(\mathbf{x}_i^*, \xi_i)\Big]$. By the definition of the functions F_i and f_i, cf. assumption (A1), we have that

$$\mathbb{E}_k\Big[f_i(\mathbf{x}_i^*) - f_i(\hat{\mathbf{x}}_i^k) + F_i(\hat{\mathbf{x}}_i^k, \xi_i^k) - F_i(\mathbf{x}_i^*, \xi_i^k)\Big] \;=\; 0, \tag{48}$$

for all $i = 1, \ldots, N$. Note that, according to the definition of the sub-σ-algebra $\mathcal{F}_k$ in (25), the conditional expectation in (48) is taken with respect to both the random variables $\hat{\mathbf{x}}_i^k$ and ξ_i^k. Hence, to see why (48) is true, we need to consider the tower property of conditional expectation, which states that for some random variable X and some sub-σ-algebras $\mathcal{S}_1 \subset \mathcal{S}_2 \subset \mathcal{F}$ we have $\mathbb{E}\big(X|\mathcal{S}_1\big) = \mathbb{E}\big(\mathbb{E}(X|\mathcal{S}_2)|\mathcal{S}_1\big)$. Now, recall that assumption (A1) essentialy says that $\mathbb{E}\big(F_i(\mathbf{x}_i, \xi_i) - f_i(\mathbf{x}_i)|\mathbf{x}_i\big) = 0$. Then, (48) holds true from the tower property for $\mathcal{S}_1 = \mathcal{F}_k$ and $\mathcal{S}_2 = \mathcal{F}_k \cup \sigma(\hat{\mathbf{x}}^k)$.

Hence, we have that

$$-2\tau_k \mathbb{E}_k \alpha^k \;=\; -2\tau_k \mathbb{E}_k \sum\nolimits_i \Big\langle \mathbf{A}_i(\hat{\mathbf{x}}_i^k - \mathbf{x}_i^*), \boldsymbol{\epsilon}_i^k\Big\rangle$$

Now, by assumption (A6) we have that $\mathbb{E}_k(\mathbf{A}_i \mathbf{x}_i^*)^T \boldsymbol{\epsilon}_i^k = 0 = \mathbb{E}_k(\mathbf{A}_i \mathbf{x}_i^k)^T \boldsymbol{\epsilon}_i^k$, since $\mathbb{E}_k \boldsymbol{\epsilon}_i^k = 0$ and the fact that $\boldsymbol{\epsilon}_i^k$ and $\mathbf{x}_i^k$ are conditionally independent given the definition of $\mathcal{F}_k$ in (25). Thus, we can substitute $\mathbf{A}_i \mathbf{x}_i^*$ with $\mathbf{A}_i \mathbf{x}_i^k$ in the term involving $\boldsymbol{\epsilon}_i^k$ in the above relation, and then use the fact that $-2\big\langle \mathbf{A}_i(\hat{\mathbf{x}}_i^k - \mathbf{x}_i^k), \boldsymbol{\epsilon}_i^k\big\rangle \;\leq\; \frac{1}{C}\|\mathbf{A}_i(\hat{\mathbf{x}}_i^k - \mathbf{x}_i^k)\|^2 + C\|\boldsymbol{\epsilon}_i^k\|^2$, for any $C < \infty$, to get

$$\begin{aligned}
\mathbb{E}_k \phi\big(\mathbf{x}^{k+1}, \boldsymbol{\lambda}^{k+1}\big) \;&\leq\; \phi\big(\mathbf{x}^k, \boldsymbol{\lambda}^k\big) \\
+ \mathbb{E}_k \Bigg[&\tau_k \sum_{i=1}^N \Big(\frac{1}{C}\|\mathbf{A}_i(\hat{\mathbf{x}}_i^k - \mathbf{x}_i^k)\|^2 + C\|\boldsymbol{\epsilon}_i^k\|^2\Big) \\
-\; &\sum_i \rho(\tau_k - \tau_k^2)||\mathbf{A}_i(\mathbf{x}_i^k - \hat{\mathbf{x}}_i^k)||^2 \\
-\; \big(\frac{\tau_k}{q} - \tau_k^2\big)&\frac{1}{\rho}||\boldsymbol{\lambda}^k - \hat{\boldsymbol{\lambda}}^k||^2 \;+\; \tau_k^2 \rho \| \sum_i \mathbf{u}_i^{k+1}\|^2 \Bigg],
\end{aligned}$$

which, after rearranging terms, can be expressed as

$$\begin{aligned}
\mathbb{E}_k \phi\big(\mathbf{x}^{k+1}, \boldsymbol{\lambda}^{k+1}\big) \;&\leq\; \phi\big(\mathbf{x}^k, \boldsymbol{\lambda}^k\big) \\
&+\; \frac{1}{\rho}\big(\tau_k^2 - \frac{\tau_k}{q}\big)\mathbb{E}_k||\boldsymbol{\lambda}^k - \hat{\boldsymbol{\lambda}}^k||^2 \\
&+\; \Big[\rho\tau_k^2 - (\rho - \frac{1}{C})\tau_k\Big]\mathbb{E}_k \sum_{i=1}^N ||\mathbf{A}_i(\mathbf{x}_i^k - \hat{\mathbf{x}}_i^k)||^2 \\
&+\; C\sum_{i=1}^N \tau_k \mathbb{E}_k \|\boldsymbol{\epsilon}_i^k\|^2 \;+\; \rho\tau_k^2 \;\mathbb{E}_k \| \sum_i \mathbf{u}_i^{k+1}\|^2.
\end{aligned} \tag{49}$$

We can now recall Theorem 2 and observe that relation (49) is of the form (27), with

$$\begin{aligned}
\eta_k \;&=\; 0, \\
\chi_k \;&=\; C\sum_{i=1}^N \tau_k \mathbb{E}_k \|\boldsymbol{\epsilon}_i^k\|^2 \;+\; \rho\tau_k^2 \;\mathbb{E}_k \| \sum_i \mathbf{u}_i^{k+1}\|^2 \\
&\quad+\; \rho\tau_k^2 \mathbb{E}_k \sum_{i=1}^N ||\mathbf{A}_i(\mathbf{x}_i^k - \hat{\mathbf{x}}_i^k)||^2 \;+\; \frac{\tau_k^2}{\rho}\mathbb{E}_k||\boldsymbol{\lambda}^k - \hat{\boldsymbol{\lambda}}^k||^2, \\
\psi_k \;&=\; \big(\rho - \frac{1}{C}\big)\tau_k \mathbb{E}_k \sum_{i=1}^N ||\mathbf{A}_i(\mathbf{x}_i^k - \hat{\mathbf{x}}_i^k)||^2 \\
&\qquad\qquad +\; \frac{\tau_k}{q\rho}\mathbb{E}_k||\boldsymbol{\lambda}^k - \hat{\boldsymbol{\lambda}}^k||^2.
\end{aligned}$$

By assumption (A5), we have that $(\rho - \frac{1}{C}) > 0$, and by assumption (A4), we have that $\frac{\tau_k}{q} > 0$. Hence, the variable ψ_k is nonnegative at all times, as required for the application of Theorem 2. Moreover, by assumption (A6), we have that $\mathbb{E}_k \| \sum_i \mathbf{u}_i^{k+1}\|^2 \leq M_1 < \infty$ for all iterations $k = 1, 2, \ldots$, and by assumption (A4) we have that $\sum_{k=1}^\infty \tau_k^2 < \infty$, from which we infer that $\sum_{k=1}^\infty \rho\tau_k^2 \;\mathbb{E}_k \| \sum_i \mathbf{u}_i^{k+1}\|^2 < \infty$.

Furthermore, note that the random variables $\|\mathbf{A}_i(\mathbf{x}_i^k - \hat{\mathbf{x}}_i^k)\|^2$ for all $i \in \mathcal{I}$, and $\|\boldsymbol{\lambda}^k - \hat{\boldsymbol{\lambda}}^k\|^2 = \|\rho \mathbf{r}(\hat{\mathbf{x}}^k)\|^2$ are bounded for every k. This is because the iterates $\hat{\mathbf{x}}_i^k$ belong to compact sets for every $i \in \mathcal{I}$ and all k, due to assumption (A3). From the fact that $\mathbf{x}_i^{k+1}$ is a convex combination between $\hat{\mathbf{x}}_i^k$ and $\mathbf{x}_i^k$, cf. (13), and given that the initial value $\mathbf{x}_i^1$ is bounded, it is straightforward to show by induction that the sequences $\mathbf{x}_i^k$ remain bounded for every $i \in \mathcal{I}$. Hence, we have that $\mathbb{E}_k \sum_{i=1}^N ||\mathbf{A}_i(\mathbf{x}_i^k - \hat{\mathbf{x}}_i^k)||^2 \leq M_2 < \infty$, and $\mathbb{E}_k||\boldsymbol{\lambda}^k - \hat{\boldsymbol{\lambda}}^k||^2 \leq M_3 < \infty$ for all iterations $k = 1, 2, \ldots$. We infer that $\sum_{k=1}^\infty \Big[\rho\tau_k^2 \mathbb{E}_k \sum_{i=1}^N ||\mathbf{A}_i(\mathbf{x}_i^k - \hat{\mathbf{x}}_i^k)||^2 + \frac{\tau_k^2}{\rho}\mathbb{E}_k||\boldsymbol{\lambda}^k - \hat{\boldsymbol{\lambda}}^k||^2\Big] < \infty$. In addition, by assumption (A7) we have that $C\sum_{i=1}^N \sum_{k=1}^\infty \tau_k \mathbb{E}_k \|\boldsymbol{\epsilon}_i^k\|^2 < \infty$. These facts combined lead to $\sum_{k=1}^\infty \chi_k < \infty$.

Thus, the conditions of Theorem 2 are satisfied. We conclude that the sequence $\{\phi(\mathbf{x}^k,\boldsymbol{\lambda}^k)\}$ converges almost surely to some finite random variable $\bar{\phi}$. By Theorem 2, we also have that

$$\sum_{k=1}^{\infty}\Big[(\rho-\frac{1}{C})\tau_k\mathbb{E}_k\sum_{i=1}^{N}||\mathbf{A}_i(\mathbf{x}_i^k-\hat{\mathbf{x}}_i^k)||^2 + \frac{\tau_k}{q\rho}\mathbb{E}_k||\boldsymbol{\lambda}^k-\hat{\boldsymbol{\lambda}}^k||^2\Big] \;<\; \infty.$$

The random variables $\|\mathbf{A}_i(\mathbf{x}_i^k-\hat{\mathbf{x}}_i^k)\|^2$ for all $i\in\mathcal{I}$, and $\|\boldsymbol{\lambda}^k-\hat{\boldsymbol{\lambda}}^k\|^2=\|\rho\mathbf{r}(\hat{\mathbf{x}}^k)\|^2$ are integrable, due to the aforementioned arguments about their boundedness. Hence, we can use the law of iterated expectation on the above relation which, combined with assumption (A4) that states $\sum_{k=1}^{\infty}\tau_k=\infty$, finally gives us that $\mathbf{r}(\hat{\mathbf{x}}^k)\xrightarrow{L^2}\mathbf{0}$ and $\mathbf{A}_i\hat{\mathbf{x}}_i^k\xrightarrow{L^2}\mathbf{A}_i\mathbf{x}_i^k$ for all $i=1,\ldots,N$. ■

We are now ready to prove the main result of this paper. The central idea behind the following proof is to show that there exists a subsequence over which $\{\phi(\mathbf{x}^k,\boldsymbol{\lambda}^k)\}$ converges almost surely to zero. Then, we use the result of lemma 4, which states that $\{\phi\big(\mathbf{x}^k,\boldsymbol{\lambda}^k\big)\}$ converges a.s. over all k to a finite limit, to infer that the generated sequences of primal and dual variables converge to their respective optimal sets almost surely over all k.

Theorem 3: Assume (A1)-(A8). Then, the SADAL method generates sequences of dual variables $\{\lambda^k(\omega)\}$ that converge to an optimal solution of problem (3) for almost all $\omega\in\Omega$. Moreover, any sequence $\{\mathbf{x}^k(\omega)\}$ generated by SADAL has an accumulation point and any such point is an optimal solution of problem (1) for almost all $\omega\in\Omega$.

Proof: In Lemma 4, we proved that $\mathbf{r}(\hat{\mathbf{x}}^k)\xrightarrow{L^2}\mathbf{0}$ and $\mathbf{A}_i\hat{\mathbf{x}}_i^k\xrightarrow{L^2}\mathbf{A}_i\mathbf{x}_i^k$ for all $i=1,\ldots,N$. It is known that the mean square convergence of a sequence of random variables implies convergence in probability, which in turn implies that there exists a subsequence such that a.s. convergence holds. Hence, there exists a subsequence $\mathcal{K}_1\subset\mathcal{K}$ such that $\mathbf{A}_i\hat{\mathbf{x}}_i^k\xrightarrow{a.s.}\mathbf{A}_i\mathbf{x}_i^k$ for $k\in\mathcal{K}_1$ holds. Similarly, $\mathbf{r}(\hat{\mathbf{x}}^k)\xrightarrow{L^2}\mathbf{0}$ over $\mathcal{K}_1$, which in turn means that there exists a sub-subsequence $\mathcal{K}_2\subset\mathcal{K}_1$ such that $\mathbf{r}(\hat{\mathbf{x}}^k)\xrightarrow{a.s.}\mathbf{0}$ for $k\in\mathcal{K}_2$. Hence, we have that $\{\mathbf{r}\big(\hat{\mathbf{x}}^k(\omega)\big)\}_{k\in\mathcal{K}_2}$ and $\{\mathbf{A}_i\hat{\mathbf{x}}_i^k(\omega)-\mathbf{A}_i\mathbf{x}_i^k(\omega)\}_{k\in\mathcal{K}_2}$ converge to zero for almost all ω. Combining these two results, we infer that $\mathbf{r}(\mathbf{x}^k)\xrightarrow{a.s.}\mathbf{0}$ over $\mathcal{K}_2$, also.

Recall from (15) that the update law for the dual sequence is $\boldsymbol{\lambda}^{k+1}=\boldsymbol{\lambda}^k+\rho\tau_k\big(\sum_{i=1}^{N}\mathbf{A}_i\tilde{\mathbf{y}}_i^{k+1}-\mathbf{b}\big)$, where $\mathbf{A}_i\tilde{\mathbf{y}}_i^{k+1}=\mathbf{A}_i\mathbf{y}_i^{k+1}+\mathbf{u}_i^{k+1}$, cf. (19). Combining these two, we have that

$$\boldsymbol{\lambda}^{k+1}=\boldsymbol{\lambda}^k+\rho\tau_k\mathbf{r}(\mathbf{y}^{k+1})+\rho\tau_k\sum_{i=1}^{N}\mathbf{u}_i^{k+1}. \tag{50}$$

By definition (14), it holds that $\mathbf{A}_i\mathbf{y}_i^{k+1}=\mathbf{A}_i\mathbf{x}_i^k+\frac{1}{q}\big(\mathbf{A}_i\hat{\mathbf{x}}_i^k-\mathbf{A}_i\mathbf{x}_i^k\big)$. Thus, using the results that $\mathbf{r}(\mathbf{x}^k)$ and $\mathbf{r}(\hat{\mathbf{x}}^k)$ converge a.s. to zero over $\mathcal{K}_2$, we infer that $\mathbf{r}(\mathbf{y}^k)\xrightarrow{a.s.}\mathbf{0}$ over $\mathcal{K}_2$. Moreover, from Chebyshev's inequality we have that for every $\delta>0$ the following holds

$$P\Big(|\rho\tau_k\mathbf{u}_i^{k+1}|\geq\delta\Big)\;\leq\;\frac{\rho^2\tau_k^2\mathbb{E}\|\mathbf{u}_i^{k+1}\|^2}{\delta^2}, \tag{51}$$

From assumption (A6) we have that $\mathbb{E}\|\mathbf{u}_i^{k+1}\|^2\leq M_4<\infty$ for all k which, combined with assumption (A4), gives us that $\sum_{k=1}^{\infty}\rho^2\tau_k^2\mathbb{E}\|\mathbf{u}_i^{k+1}\|^2<\infty$. This implies that $\sum_{k=1}^{\infty}P\Big(|\rho\tau_k\mathbf{u}_i^{k+1}|\geq\delta\Big)<\infty$, and by the Borel-Cantelli lemma this means that $\rho\tau_k\mathbf{u}_i^k\xrightarrow{a.s.}\mathbf{0}$. This, combined with the previous result that $\mathbf{r}(\mathbf{y}^k)\xrightarrow{a.s.}\mathbf{0}$ over $\mathcal{K}_2$ and the a.s. convergence of $\{\phi(\mathbf{x}^k,\boldsymbol{\lambda}^k)\}$ from Lemma 4, gives us that the dual sequence $\{\boldsymbol{\lambda}^k\}_{k\in\mathcal{K}_2}$ (50) converges a.s. to some finite limit $\bar{\boldsymbol{\mu}}$, i.e., $\boldsymbol{\lambda}^k\xrightarrow{a.s.}\bar{\boldsymbol{\mu}}$ over $\mathcal{K}_2$.

From assumption (A3), all sequences $\{\hat{\mathbf{x}}_i^k(\omega)\}$, $i=1,\ldots N$, are bounded. This, combined with the fact that the $\mathbf{x}_i^{k+1}$ is a convex combination between $\hat{\mathbf{x}}_i^k$ and $\mathbf{x}_i^k$, cf. (13), and given that the initial value $\mathbf{x}^1$ is bounded, means that the sequences $\{\mathbf{x}_i^k(\omega)\}$, $i=1,\ldots N$, are bounded also. This in turn implies that the sequences $\{\mathbf{x}_i^k(\omega)\}$ have accumulation points $\bar{\mathbf{x}}_i(\omega)$, which are also accumulation points of $\{\hat{\mathbf{x}}_i^k(\omega)\}$ due to the update step (13) of SADAL. We can choose a subsequence $\mathcal{K}_3\subset\mathcal{K}_2$ so that $\{\mathbf{x}_i^k(\omega)\}_{k\in\mathcal{K}_3}$ and $\{\hat{\mathbf{x}}_i^k(\omega)\}_{k\in\mathcal{K}_3}$ converge to $\bar{\mathbf{x}}_i(\omega)$ for all $i=1,\ldots,N$. Denoting $\bar{\mathbf{x}}(\omega)=\big[\bar{\mathbf{x}}_1(\omega),\ldots,\bar{\mathbf{x}}_N(\omega)\big]^T$, we observe that the point $\bar{\mathbf{x}}(\omega)$ is feasible due to the closedness of the sets $\mathcal{X}_i$ and the continuity of $\mathbf{r}(\cdot)$.

For any $i=1,\ldots,N$, consider the sequence $\{\mathbf{s}_{f_i}^k(\omega)\}_{k\in\mathcal{K}_3}$. The subdifferential mapping $\mathbf{x}\rightrightarrows\partial f(\mathbf{x})$ of any finite-valued convex function defined on $\mathbb{R}^n$ is upper semi-continuous and has compact images. Therefore, the sequences $\{\mathbf{s}_{f_i}^k(\omega)\}_{k\in\mathcal{K}_3}$ have convergent subsequences due to a fundamental result that goes back to [53]. We can choose $\mathcal{K}_4\subset\mathcal{K}_3$ such that $\{\mathbf{s}_{f_i}^k(\omega)\}_{k\in\mathcal{K}_4}$ converge to some $\bar{\mathbf{s}}_{f_i}(\omega)\subset\partial f_i\big(\bar{\mathbf{x}}_i(\omega)\big)$ for all $i=1,\ldots,N$ and almost all ω.

We recall that the optimality conditions for each subproblem $i=1,\ldots,N$, cf. (29), take the form

$$0\;=\;\mathbf{s}_{F_i}^k\;+\;\mathbf{A}_i^\top\tilde{\boldsymbol{\lambda}}_i^k\;+\;\rho\mathbf{A}_i^\top\Big(\mathbf{A}_i\hat{\mathbf{x}}_i^k+\sum_{j\neq i}\mathbf{A}_j\tilde{\mathbf{x}}_j^k-\mathbf{b}\Big)\;+\;\mathbf{z}_i^k.$$

Gathering all the noise terms, the above equation can be equivalently expressed as

$$0\;=\;\mathbf{s}_{f_i}^k\;+\;\mathbf{A}_i^\top\boldsymbol{\lambda}^k\;+\;\rho\mathbf{A}_i^\top\Big(\mathbf{A}_i\hat{\mathbf{x}}_i^k+\sum_{j\neq i}\mathbf{A}_j\mathbf{x}_j^k-\mathbf{b}\Big) + \underbrace{\mathbf{e}_i^k+\mathbf{w}_i^k+\rho\sum_{j\neq i}\mathbf{v}_{ij}^k}_{\text{noise terms}}\;+\;\mathbf{z}_i^k. \tag{52}$$

From assumptions (A6)–(A8), we have that the noise terms $\mathbf{e}_i^k+\mathbf{w}_i^k+\rho\sum_{j\neq i}\mathbf{v}_{ij}^k$ converge to $\mathbf{0}$ a.s. as $k\to\infty$. Hence, passing to the limit over $\mathcal{K}_4$ in (52), we infer that each sequence $\{\mathbf{z}_i^k(\omega)\}_{k\in\mathcal{K}_4}$ converges to a point $\bar{\mathbf{z}}_i(\omega)$, for all $i=1,\ldots,N$ and almost all ω. The mapping $\mathbf{x}_i\rightrightarrows\mathcal{N}_{\mathcal{X}_i}(\mathbf{x}_i)$ has closed graph and, hence, $\bar{\mathbf{z}}_i(\omega)\in\mathcal{N}_{\mathcal{X}_i}\big(\bar{\mathbf{x}}_i(\omega)\big)$ [40, Lemma 2.42]. After the limit pass in (52) over $k\in\mathcal{K}_4$, we conclude that

$$0=\bar{\mathbf{s}}_{f_i}(\omega)+\mathbf{A}_i^\top\bar{\boldsymbol{\mu}}(\omega)+\bar{\mathbf{z}}_i(\omega),\quad\forall\,i=1\ldots,N.$$

for almost all ω. These relations are exactly the optimality conditions of each $i \in \mathcal{I}$ for the saddle point of the original problem (1), cf. (32). This result, together with the feasibility of $\bar{\mathbf{x}}(\omega)$, implies that $\bar{\mathbf{x}}(\omega)$ is a solution of the primal problem (1) and $\bar{\boldsymbol{\mu}}(\omega)$ is a solution of the dual problem (3).

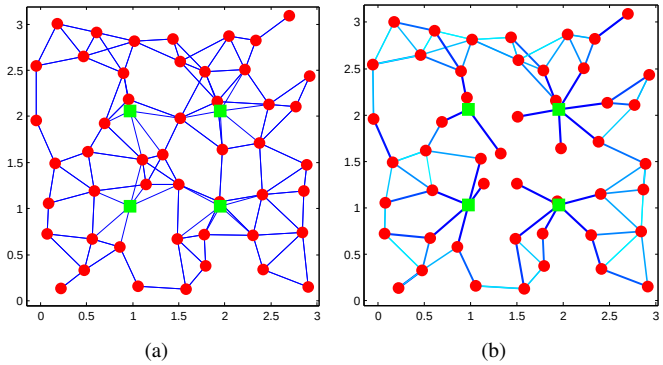
Due to the continuity of $\phi(\cdot)$, it follows that $\{\phi(\mathbf{x}^k, \boldsymbol{\lambda}^k)\}$ converges a.s. to zero over $\mathcal{K}_4$. Combining this with the result that $\{\phi(\mathbf{x}^k, \boldsymbol{\lambda}^k)\}$ converges a.s. to some finite limit for all $k = 1, 2, \dots$ from Lemma 4, we infer that $\{\phi(\mathbf{x}^k, \boldsymbol{\lambda}^k)\}$ converges a.s. to zero for all $k = 1, 2, \dots$. This further implies that the terms $\|\mathbf{A}_i\mathbf{x}_i^k - \mathbf{A}_i\mathbf{x}_i^*\|^2$ for all $i \in \mathcal{I}$ and $\|\bar{\boldsymbol{\lambda}}^k - \boldsymbol{\lambda}^*\|^2$ converge a.s. to zero for all $k = 1, 2, \dots$, due to the nonnegativity of all these terms in $\phi(\cdot)$. Hence, we infer that $\mathbf{A}_i\mathbf{x}_i^k - \mathbf{A}_i\mathbf{x}_i^* \xrightarrow{a.s.} \mathbf{0}$ for all $i \in \mathcal{I}$, and $\bar{\boldsymbol{\lambda}}^k \xrightarrow{a.s.} \boldsymbol{\lambda}^*$. Combining the result that $\mathbf{A}_i\mathbf{x}_i^k - \mathbf{A}_i\mathbf{x}_i^* \xrightarrow{a.s.} \mathbf{0}$ for all $i \in \mathcal{I}$ with the definition $\mathbf{r}(\mathbf{x}^k) = \sum_i \mathbf{A}_i\mathbf{x}_i^k - \mathbf{b}$ and the fact that the optimal solution is feasible $\sum_i \mathbf{A}_i\mathbf{x}_i^* = \mathbf{b}$, we get that $\mathbf{r}(\mathbf{x}^k) \xrightarrow{a.s.} \mathbf{0}$. Hence, after recalling that $\bar{\boldsymbol{\lambda}}^k = \boldsymbol{\lambda}^k + \rho(1 - \frac{1}{q})\mathbf{r}(\mathbf{x}^k)$ by the definition (39), we can use the results $\bar{\boldsymbol{\lambda}}^k \xrightarrow{a.s.} \boldsymbol{\lambda}^*$ and $\mathbf{r}(\mathbf{x}^k) \xrightarrow{a.s.} \mathbf{0}$ to infer that $\boldsymbol{\lambda}^k$ converges a.s. to $\boldsymbol{\lambda}^*$ for all $k = 1, 2, \dots$, as required. ■

V. Numerical Experiments

In order to verify the validity of the proposed method, in this section we present numerical results of SADAL applied on a network optimization problem. Consider an undirected graph $G = (N, A)$ with a set of nodes N and a set of arcs A. The set of nodes consists of two subsets $N = \{S, D\}$, where S is the set of source nodes and D is the set of destination nodes. Let s_i denote the flow generated at source node $i \in S$ and $c_i \geq 0$ denote the reward coefficient for s_i. Each s_i is subject to a minimum threshold constraint $s_i \geq s_i^{min}$, where the s_i^{min} is a problem parameter. Also, let t_{ij} denote the flow through arc (i, j). Each arc (i, j) has a feasible range of flows $a_{ij} \leq t_{ij} \leq b_{ij}$, where a_{ij}, b_{ij} are given numbers. Denote the neighborhood of node i as $\mathcal{C}_i = \{j : (i, j) \in A\}$. The conservation of flow at each node $i \in S$ is expressed as $\sum_{\{j \in \mathcal{C}_i\}} t_{ij} - \sum_{\{j | i \in \mathcal{C}_j\}} t_{ji} = s_i$. The problem we consider is a network utility maximization (NUM) problem, where we seek to find routing decisions t_{ij} that maximize the amount of flow generated at all source nodes, subject to flow conservation, arc capacity, and minimum flow generation constraints. The NUM takes the form

$$\begin{aligned} \max \quad & \sum_{i \in S} c_i s_i \\ \text{subject to} \quad & \sum_{\{j \in \mathcal{C}_i\}} t_{ij} - \sum_{\{j | i \in \mathcal{C}_j\}} t_{ji} = s_i, \quad \forall\, i \in S \\ & a_{ij} \leq t_{ij} \leq b_{ij}, \quad \forall\, (i, j) \in A, \\ & s_i \geq s_i^{min}, \quad \forall\, i \in S. \end{aligned}$$

In our consideration, the destination nodes are modeled as sinks and can absorb any amount of incoming rates. Hence, no flow conservation constraints are necessary for these nodes. Note that, if some nodes are neither sources or destinations then we set the c_i and s_i^{min} equal to zero. Note also that for this problem the distributed agents are all the source nodes $i \in S$. Moreover, the local constraint sets are $\mathcal{X}_i = \{s_i, t_{ij} \in \mathbb{R},\ \forall j \in N :\ s_i \geq s_i^{min},\ a_{ij} \leq t_{ij} \leq b_{ij}\}$, for all source nodes $i \in S$, while the coupling linear constraint set consists of the flow conservation constraints $\sum_{\{j \in \mathcal{C}_i\}} t_{ij} - \sum_{\{j | i \in \mathcal{C}_j\}} t_{ji} = s_i$ for all $i \in S$.

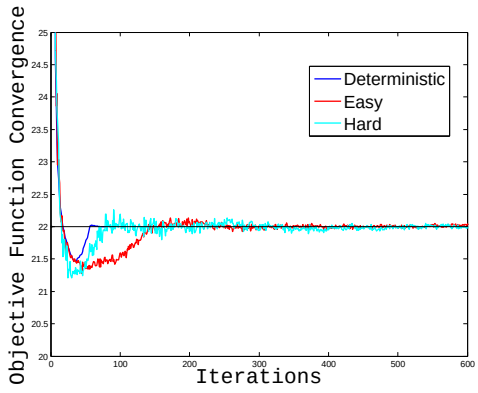


(a) (b)

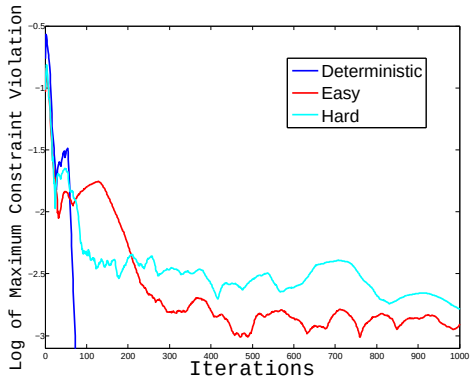
Fig. 1. A randomly generated network for the NUM problem with 50 sources (red dots) and 4 sinks (green squares): a) All available arcs $(i, j) \in A$ are depicted as blue lines, b) Flow routing decisions after solving the problem with SADAL. The rate of flow t_{ij} through arc (i, j) defines the thickness and color of the corresponding drawn line. Thicker, darker blue lines indicate larger flows.

In what follows, we consider normalized rates $s_i, t_{ij} \in [0, 1]$, without any loss of generality. The parameters c_i and s_i^{min} are randomly generated by sampling uniformly in the intervals $[0.1, 1]$ and $[0, 0.3]$, respectively. Unless otherwise noted, the penalty parameter is set to $\rho = 1$. Moreover, observe that $q = \max_i |\mathcal{C}_i|$ and, according to the presented convergence analysis, the stepsize τ of ADAL and the initial stepsize τ_1 of SADAL must be less than $1/q$. In all simulations, the objective function value $\sum_{i \in S} c_i s_i^k$ and the maximum residual $\max_i \mathbf{r}_i^k = \max_i \left(\sum_{\{j \in \mathcal{C}_i\}} t_{ij}^k - \sum_{\{j | i \in \mathcal{C}_j\}} t_{ji}^k - s_i^k \right)$, i.e., the maximum constraint violation among all the flow conservation constraints $j = 1, \dots, m$ at each iteration k, were monitored as the criteria of convergence. The examined networks were randomly generated with the agents uniformly distributed in rectangle boxes. All the simulation results that are presented in what follows correspond to the network configuration depicted in Fig. 1(a). For reference, the typical routing decisions obtained after solving this problem with SADAL are depicted in Fig. 1(b).

In Fig. 2 we present simulation results for different levels of noise corruption and compare them with the deterministic solution from ADAL. The simulations We consider two cases with different variances of the noise terms, labeled "easy" and "hard"; in the hard case all the noise terms have larger variance, compared to the easy case. In both cases the noise terms are modeled as uniform random variables. For the "easy" case, the noise corruption terms are modeled as follows: $\mathbf{v}_{ij}^k \sim \frac{1}{\mu^k} U(-0.1, 0.1)$, and $\mathbf{w}_i^k \sim \frac{1}{\mu^k} U(-0.1, 0.1)$, where $U(u, v)$ denotes the uniform distribution with support $[u, v]$. Here, μ^k is a coefficient that we introduce to make the noise terms decreasing; we initially set $\mu^1 = 1$ and then every 5 iterations we increase μ by one, i.e., $\{\mu\}_{k=1}^{\infty} = \{1, 1, 1, 1, 1, 2, 2, 2, 2, 2, 3, \dots\}$. The non-decreasing noise terms are generated as $\mathbf{u}_i^k \sim U(-0.03, 0.03)$,



(a)



(b)

Fig. 2. Comparative simulation results for different noise levels. We consider the noise-free case (deterministic) where ADAL is applied, and two noise scenarios, labeled "easy" and "hard" (larger noise), where SADAL is applied. a) Objective function convergence, b) Maximum constraint violation convergence.

and the noise in the linear objective functions is modeled as $\tilde{c}_i^k = c_i p_i^k$, where $p_i^k \sim \frac{1}{\mu^k} U(-0.3, 0.3)$. For the "hard" case the random variables are generated according to $\mathbf{v}_{ij}^k \sim \frac{1}{\mu^k} U(-0.2, 0.2)$, $\mathbf{w}_i^k \sim \frac{1}{\mu^k} U(-0.2, 0.2)$, $\mathbf{u}_i^k \sim U(-0.05, 0.05)$, and $p_i^k \sim \frac{1}{\mu^k} U(-0.7, 0.7)$. In both cases, the stepsize sequence τ_k is defined as $\tau_k = 1/(q\nu^k)$. The sequence ν^k is generated similar to μ^k, with the difference that we increase it by one every 30 iterations.

Not surprisingly, the results of Fig. 2 indicate that the stochastic problem converges slower than the noise-free case. On the other hand, the difference in the noise levels does not appear to affect the convergence speed significantly, at least for the two different noise scenarios studied here. We can also see that the residual (feasibility) convergence of SADAL slows down significanlty after reaching accuracy levels of about 10^{-3}. To put these results into perspective, note that, at iteration $k = 1000$, the noise term $\mathbf{v}_{ij}^k$ in the message exchange from agent i to j is distributed according to $\mathbf{v}_{ij}^k \sim 10^{-3} U(-0.5, 0.5)$. Now, consider that this is the noise just from one neighbor, however, each agent i has multiple neighbors, which means that the noise corruptions add up; the generated networks typically have a neighborhood size of about 5 or 6 for each node. Thus, there exists a (relatively) substantial amount of noise corruption even at iteration 1000 (recall also that the $\mathbf{u}_i^k$ noise terms do not decrease), which should be taken into consideration when evaluating the convergence results of Fig. 2.

In order to test how the choice of stepsize sequences $\{\tau_k\}$ affects the convergence, we have performed simulations where the ν^k is increased every 15, 60, or 100 iterations. In all cases we do not let the stepsize τ_k decrease below 0.01, i.e., $\tau_k = \max\{1/(q\nu^k), 0.01\}$. For completeness, we also test SADAL for a constant stepsize $\tau = \frac{1}{q}$, although this is not consistent with the assumptions of the algorithm. The results, corresponding to the "hard" formulation, are depicted in Fig. 3. We observe that the convergence of SADAL is not significantly affected by the choice of stepsize sequences, albeit stepsizes that decrease faster seem to exhibit a slightly better behavior (keep in mind that we do not let the stepsize decrease below 0.01). Moreover, the constant stepsize choice produces an oscillatory behavior and does not lead to convergence, which is in accordance with the theoretical analysis.

Finally, we examine how sensitive SADAL is to the choice of the user-defined penalty coefficient ρ, at least for the problem under consideration here. Fig. 4 depicts simulation results of the "hard" noise scenario for ρ taking the values 0.3, 1, 3, and 10. The μ^k is increased every 5 iterations, and ν^k every 30 iterations. We observe that convergence is not significantly affected by the choice of ρ, apart from the smallest value case $\rho = 0.3$ which lead to a more "oscillatory" behavior.

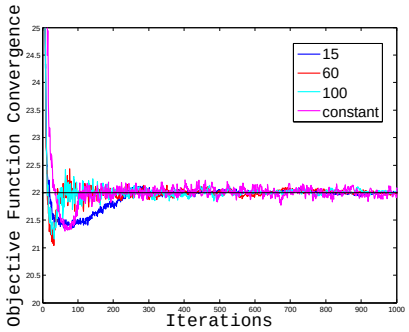
VI. Conclusions

In this paper we have investigated distributed solutions for a certain class of convex constrained optimization problems that are subject to noise and uncertainty. In particular, we have considered the problem of minimizing the sum of local objective functions whose arguments are local variables that are constrained to lie in closed, convex sets. The local variables are also globally coupled via a set of affine constraints. We have proposed an iterative distributed algorithm that is able to withstand the presence of multiple noise terms that affect both the computations and the communications. To the best of our knowledge, this is the first attempt to consider this class of problems in the context of distributed stochastic approximation techniques. Moreover, the proposed method is based on the augmented Lagrangian framework, which is well-known to be a very efficient approach for optimization in deterministic settings. Compared to existing distributed stochastic AL approaches, our method studies a more general framework, by allowing $N > 2$ and considering multiple noise terms that appear in all computations and all message exchanges at the same time. We have established conditions under which our method is guaranteed to converge a.s. to the optimal sets in both the primal and dual domains.

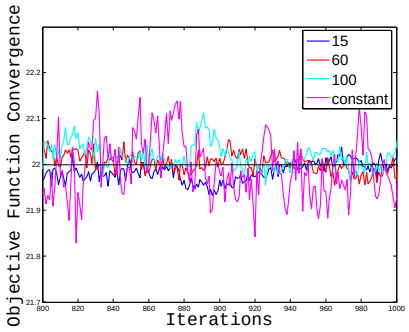
References

[1] D. P. Bertsekas and J. N. Tsitsiklis, *Parallel and Distributed Computation: Numerical Methods*. Athena Scientific, 1997.

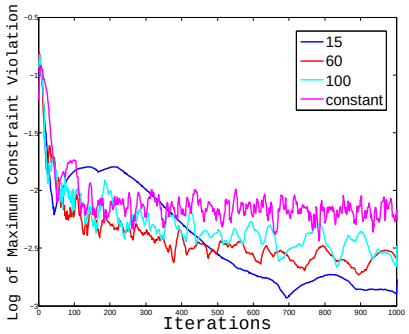
[2] A. Nedic and A. Ozdaglar, "Approximate primal solutions and rate analysis for dual subgradient methods," *SIAM Journal on Optimization*, vol. 19, no. 4, pp. 1757–1780, 2009.



(a)

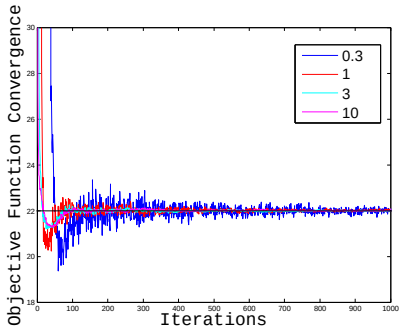


(b)

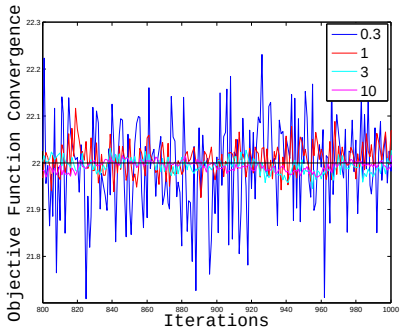


(c)

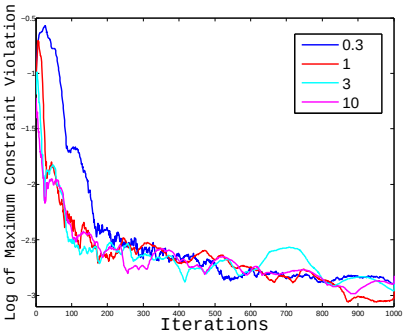
Fig. 3. Simulation results for different choices of the stepsize sequence $\tau_k = 1/(q\nu^k)$. We consider scenarios where ν^k is increased by one every 15, 60, and 100 iterations, and also a constant stepsize choice $\tau_k = 1/q$. The results correspond to the "hard" noise scenario. a) Objective function convergence, b) Magnified view of the objective function convergence, c) Maximum constraint violation convergence.



(a)



(b)



(c)

Fig. 4. Simulation results for different values of ρ. a) Objective function convergence, b) Magnified view of the objective function convergence, c) Maximum constraint violation convergence.

[3] J. Eckstein and D. P. Bertsekas, "On the Douglas-Rachford splitting method and the proximal point algorithm for maximal monotone operators," *Mathematical Programming,*, vol. 55, pp. 293–318, 1992.

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

[5] A. Ruszczyński, "On convergence of an Augmented Lagrangian decomposition method for sparse convex optimization," *Mathematics of Operations Research*, vol. 20, pp. 634–656, 1995.

[6] J. Mulvey and A. Ruszczyński, "A diagonal quadratic approximation method for large scale linear programs," *Operations Research Letters*, vol. 12, pp. 205–215, 1992.

[7] N. Chatzipanagiotis, D. Dentcheva, and M. M. Zavlanos, "An augmented Lagrangian method for distributed optimization," *Mathematical Programming*, 2014.

[8] E. Wei, A. Ozdaglar, and A. Jadbabaie, "A distributed newton method for network utility maximization," in *Proceedings of the 49th IEEE Conference on Decision and Control*, 2010.

[9] M. Zargham, A. Ribeiro, A. Jadbabaie, and A. Ozdaglar, "Accelerated dual descent for network optimization." in *Proceedings of the American Control Conference*, San Francisco, CA, 2011.

[10] S. Lee and A. Nedic, "Distributed random projection algorithm for convex optimization," *Selected Topics in Signal Processing, IEEE Journal of*, vol. 7, no. 2, pp. 221–229, April 2013.

[11] ——, "Gossip-based random projection algorithm for distributed optimization: Error bound," in *Decision and Control (CDC), 2013 IEEE 52nd Annual Conference on*, Dec 2013, pp. 6874–6879.

[12] Y. Nesterov, "Subgradient methods for huge-scale optimization problems," *Mathematical Programming*, vol. 146, no. 1-2, pp. 275–297, 2014.

[13] D. Jakovetic, J. Freitas Xavier, and J. Moura, "Convergence rates of distributed nesterov-like gradient methods on random networks," *Signal Processing, IEEE Transactions on*, vol. 62, no. 4, pp. 868–882, Feb 2014.

[14] S. Hosseini, A. Chapman, and M. Mesbahi, "Online distributed admm via dual averaging," in *Decision and Control (CDC), 2014 IEEE 53rd Annual Conference on*, Dec 2014, pp. 904–909.

[15] A. Koppel, F. Jakubiec, and A. Ribeiro, "A saddle point algorithm for networked online convex optimization," in *Acoustics, Speech and Signal Processing (ICASSP), 2014 IEEE International Conference on*, May 2014, pp. 8292–8296.

[16] S. S. Kia, J. Corts, and S. Martnez, "Distributed convex optimization

via continuous-time coordination algorithms with discrete-time communication," *Automatica*, vol. 55, pp. 254 – 264, 2015.

[17] N. Chatzipanagiotis and M. Zavlanos, "On the convergence rate of a distributed augmented lagrangian optimization algorithm," in *American Control Conference (ACC), 2015*, July 2015.

[18] H. Robbins and S. Monro, "A stochastic approximation method," *The Annals of Mathem. Statist.*, vol. 22, no. 3, pp. 400–407, Sep. 1951.

[19] J. Kiefer and J. Wolfowitz, "Stochastic estimation of the maximum of a regression function," *The Annals of Mathematical Statistics*, vol. 23, no. 3, pp. 462–466, Sep. 1952.

[20] B. Polyak and A. Juditsky, "Acceleration of stochastic approximation by averaging," *SIAM Journal on Control and Optimization*, vol. 30, no. 4, pp. 838–855, 1992.

[21] Y. M. Ermoliev, "Stochastic quasigradient methods," in *Numerical techniques for stochastic optimization*, 1983, pp. 141–185.

[22] A. Ruszczynski and W. Syski, "Stochastic approximation method with gradient averaging for unconstrained problems," *Automatic Control, IEEE Transactions on*, vol. 28, no. 12, pp. 1097–1105, Dec 1983.

[23] J. Spall, "Adaptive stochastic approximation by the simultaneous perturbation method," *Automatic Control, IEEE Transactions on*, vol. 45, no. 10, pp. 1839–1853, Oct 2000.

[24] A. Nemirovski, A. Juditsky, G. Lan, and A. Shapiro, "Robust stochastic approximation approach to stochastic programming," *SIAM Journal on Optimization*, vol. 19, no. 4, pp. 1574–1609, 2009.

[25] F. Yousefian, A. Nedić, and U. V. Shanbhag, "On stochastic gradient and subgradient methods with adaptive steplength sequences," *Automatica*, vol. 48, no. 1, pp. 56–67, Jan. 2012.

[26] H. Kushner and G. Yin, *Stochastic Approximation and Recursive Algorithms and Applications*. Springer, 2003.

[27] V. Borkar, *Stochastic approximation: a dynamical systems viewpoint*. Cambridge University Press, 2008.

[28] G. Roth and W. Sandholm, "Stochastic approximations with constant step size and differential inclusions," *SIAM Journal on Control and Optimization*, vol. 51, no. 1, pp. 525–555, 2013.

[29] J. Tsitsiklis, D. Bertsekas, and M. Athans, "Distributed asynchronous deterministic and stochastic gradient optimization algorithms," *Automatic Control, IEEE Transactions on*, vol. 31, no. 9, pp. 803–812, Sep 1986.

[30] H. Kushner and G. Yin, "Asymptotic properties of distributed and communicating stochastic approximation algorithms," *SIAM Journal on Control and Optimization*, vol. 25, no. 5, pp. 1266–1290, 1987.

[31] S. Stankovic, M. Stankovic, and D. Stipanovic, "Decentralized parameter estimation by consensus based stochastic approximation," *Automatic Control, IEEE Trans. on*, vol. 56, no. 3, pp. 531–543, March 2011.

[32] K. Srivastava and A. Nedic, "Distributed asynchronous constrained stochastic optimization," *Selected Topics in Signal Processing, IEEE Journal of*, vol. 5, no. 4, pp. 772–790, Aug 2011.

[33] A. Nedic, "Asynchronous broadcast-based convex optimization over a network," *Automatic Control, IEEE Transactions on*, vol. 56, no. 6, pp. 1337–1351, June 2011.

[34] S. Sundhar Ram, A. Nedic, and V. Veeravalli, "Distributed stochastic subgradient projection algorithms for convex optimization," *Journal of Optim. Theory and Applications*, vol. 147, no. 3, pp. 516–545, 2010.

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

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

[37] T. Erseghe, D. Zennaro, E. Dall'Anese, and L. Vangelista, "Fast consensus by the alternating direction multipliers method," *Signal Processing, IEEE Transactions on*, vol. 59, no. 11, pp. 5523–5537, Nov 2011.

[38] I. Schizas, G. Mateos, and G. Giannakis, "Distributed LMS for consensus-based in-network adaptive processing," *Signal Processing, IEEE Transactions on*, vol. 57, no. 6, pp. 2365–2382, June 2009.

[39] H. Ouyang, N. He, L. Tran, and A. G. Gray, "Stochastic alternating direction method of multipliers," in *Proc. of the 30th Internat. Conf. on Machine Learning (ICML-13)*, vol. 28, no. 1, 2013, pp. 80–88.

[40] A. Ruszczyński, *Nonlinear Optimization*. Princeton, NJ, USA: Princeton University Press, 2006.

[41] R. Rockafellar, "Augmented Lagrange multiplier functions and duality in nonconvex programming," *SIAM Journal on Control*, vol. 12, pp. 268–285, 1973.

[42] D. Gabay and B. Mercier, "A dual algorithm for the solution of nonlinear variational problems via finite element approximations," *Computers and Mathematics with Applications*, vol. 2, pp. 17–40, 1976.

[43] M. Fortin and R. Glowinski, *Augmented Lagrangian Methods: Applications to the Numerical Solution of Boundary-Value Problems*. Amsterdam: North-Holland, 1983.

[44] A. J. Berger, J. M. Mulvey, and A. Ruszczyński, "An extension of the DQA algorithm to convex stochastic programs," *SIAM J. Optim.*, vol. 4, no. 4, pp. 735–753, 1994.

[45] J. Eckstein and D. P. Bertsekas, "An alternating direction method for linear programming." LIDS, MIT, 1990.

[46] D. Bertsekas, "Stochastic optimization problems with nondifferentiable cost functionals," *Journal of Optimization Theory and Applications*, vol. 12, no. 2, pp. 218–231, 1973.

[47] I.-J. Wang and J. Spall, "A constrained simultaneous perturbation stochastic approximation algorithm based on penalty functions," in *American Control Conference, 1999. Proceedings of the 1999*, vol. 1, 1999, pp. 393–399 vol.1.

[48] ——, "Stochastic optimization with inequality constraints using simultaneous perturbations and penalty functions," in *Decision and Control, 2003. Proceedings. 42nd IEEE Conference on*, vol. 4, Dec 2003, pp. 3808–3813 vol.4.

[49] S. S. Singh, N. Kantas, B.-N. Vo, A. Doucet, and R. J. Evans, "Simulation-based optimal sensor scheduling with application to observer trajectory planning," *Automatica*, vol. 43, pp. 817 – 830, 2007.

[50] T. Salimans and D. A. Knowles, "On using control variates with stochastic approximation for variational bayes and its connection to stochastic linear regression," arXiv:1401.1022v3.

[51] R. Johnson and T. Zhang, "Accelerating stochastic gradient descent using predictive variance reduction," in *Advances in Neural Information Processing Systems*, 2013, pp. 315–323.

[52] H. Robbins and D. Siegmund, "A convergence theorem for non negative almost supermartingales and some applications," in *Proceedings of the Symposium on Optimizing methods in Statistics*, Columbus, OH, 1971, pp. 233–257.

[53] C. Berge, *Espaces Topologiques Functions Multivoques Dunod*. Paris, 1959.

Nikolaos Chatzipanagiotis received the Diploma in Mechanical Engineering, and the M.Sc. degree in Microsystems and Nanodevices from the National Technical University of Athens, Athens, Greece, in 2006 and 2008, respectively. He also received a Ph.D. degree in Mechanical Engineering from Duke University, Durham, NC, in 2015.

His research interests include optimization theory and algorithms with applications on networked control systems, wired and wireless communications, and multi-agent mobile robotic networks.

Michael M. Zavlanos received the Diploma in mechanical engineering from the National Technical University of Athens (NTUA), Athens, Greece, in 2002, and the M.S.E. and Ph.D. degrees in electrical and systems engineering from the University of Pennsylvania, Philadelphia, PA, in 2005 and 2008, respectively.

From 2008 to 2009 he was a post-doctoral researcher in the department of electrical and systems engineering at the University of Pennsylvania, Philadelphia. He then joined the Stevens Institute of Technology, Hoboken, NJ, as an assistant professor of mechanical engineering, where he remained until 2012. Currently, he is an assistant professor of mechanical engineering and materials science at Duke University, Durham, NC. He also holds a secondary appointment in the department of electrical and computer engineering. His research interests include a wide range of topics in the emerging discipline of networked systems, with applications in robotic, sensor, communication, and biomolecular networks. He is particularly interested in hybrid solution techniques, on the interface of control theory, distributed optimization, estimation, and networking.

Dr. Zavlanos is a recipient of the 2014 Office of Naval Research Young Investigator Program (YIP) Award and the 2011 National Science Foundation Faculty Early Career Development (CAREER) Award. He was also a finalist for the best student paper award at CDC 2006.