2019 57th Annual Allerton Conference on Communication, Control, and Computing (Allerton)
Allerton Park and Retreat Center
Monticello, IL, USA, September 24-27, 2019

Distributed Asynchronous Random Projection Algorithm (DARPA) with Arbitrary Uniformly Bounded Delay

Elie Atallah, Nazanin Rahnavard *Senior Member, IEEE*, and Chinwendu Enyioha
Department of Electrical and Computer Engineering
University of Central Florida, Orlando, FL
Emails: {elieatallah@knights., nazanin@eecs., cenyioha@}ucf.edu

***Abstract*—In this paper, an asynchronous random projection algorithm is introduced to solve a distributed constrained convex optimization problem over a time-varying multi-agent network. In this asynchronous case, each agent computes its estimate by exchanging information with its neighbors within a bounded delay lapse. For diminishing uncoordinated stepsizes and some standard conditions on the gradient errors, we provide a convergence analysis of Distributed Asynchronous Random Projection Algorithm (DARPA) to the same optimal point under an arbitrary uniformly bounded delay.**

***Index Terms*—random projections, asynchronous, gradient error, delays**

I. Introduction

The focus of this paper is the convergence analysis of a communication-efficient distributed algorithm whereby agents exchange local information and update in an asynchronous manner. We propose a gradient descent algorithm with random projections which is implementable in a fully asynchronous communication framework. The random projection algorithm is of interest for constrained optimization when the constraint set is not known in advance or the projection operation on the whole constraint set is computationally prohibitive. A synchronous randomized algorithm for distributed optimization problems [5] and centralized problems [8] were presented. However, asynchronous algorithms based on a gossip scheme have been proposed and analyzed for a scalar objective function and a diminishing stepsize [11], and a vector objective function and a constant stepsize [12].

To the best of our knowledge, the case of (fully) asynchronous distributed random projection algorithm was left untreated, while partly asynchronous cases such as gossiping [6] or broadcast [7] were investigated as mentioned earlier. Motivated by these considerations, this paper proposes a Distributed Asynchronous Random Projection Algorithm (DARPA) over a time-varying network in which agents are activated with some probability, receive (possibly delayed) information from their neighbors, with which they estimate their local gradient and project to the feasible set at each activation time instance. We prove that the DARPA converges to an optimal solution assuming that the next update is performed by a random agent and asynchronous communication is subject to an arbitrary uniformly bounded delay. With reasonable assumptions on gradient estimation errors, we prove that the iterates of all agents converge to the same point in the optimal set almost surely. Additionally, the case of exact evaluation of the gradients follows directly from this proof. We use the common nomenclature utilized in the literature.

The remainder of the paper is organized as follows: In Section II, we present the problem setup with the model assumptions. In Section III, we describe the proposed algorithm and present the theoretical background needed for our analysis. In Section IV, we state the main convergence theorem with its detailed proof. In Section V, we conclude the paper.

II. Problem Setup

We consider a distributed constrained convex optimization problem over a network of n nodes indexed by $V = \{1, .., n\}$. We assume that the agents communicate over a network with time-varying network topology represented by an undirected graph $(V, \mathcal{E}(k))$, where $\mathcal{E}(k)$ is the set of undirected edges at time k. There are no self loops in this graph and we have $\{i, j\} \in \mathcal{E}(k)$ only if agents i and j can communicate with each other at time k. Our aim is to solve the constrained distributed optimization problem

$$\min f(x) = \sum_{i=1}^{n} f_i(x) \quad \text{where } x \in \mathcal{X} = \cap_{i=1}^{n} \mathcal{X}_i \ , \qquad (1)$$

where $f_i : \mathbb{R}^N \to \mathbb{R}$ is a convex function, representing the local objective of agent i, and $\mathcal{X}_i \subset \mathbb{R}^N$ is a closed convex set, representing the local constraint set of agent i. The function f_i and the set $\mathcal{X}_i$ are known to agent i only. We assume that Problem (1) is feasible.

Moreover, each constraint set $\mathcal{X}_i$ is the intersection of finitely many closed convex sets, i.e., $\mathcal{X}_i = \cap \mathcal{X}_i^j$ for $j \in I_i$ where $I_i = \{1, .., d_i\}$ and $I = \cup_{i=1}^{n} I_i$ and $I_i \cap I_j = \phi$ for $i \neq j$.

We present DARPA, a distributed optimization algorithm for problem (1), that is based on the random projections and the complete asynchronous communication protocol DARPA is described in Algorithm 1 and the updating equations are given by (3) where $\mathbf{x}_i(k)$ is the estimate of the solution at node i and iteration k for $i = \{1, 2, \ldots, n\}$, and $\mathbf{v}_i(k)$ is the weighted average calculated from the connection of node i to neighboring nodes. $\Pi_{\mathcal{X}_i^{\Omega_i(k)}}$ is a random projection on a convex feasible set of the constraints and ∇f_i is the gradient of the local function f_i of node i where $f = \sum_{i=1}^{n} f_i$.

This material is based upon work supported by the National Science Foundation under Grant No. CCF-1718195.

978-1-7281-3151-1/19/$31.00 ©2019 IEEE

To allow agents to communicate and carry out local computations at different time instances and for different duration, the proposed algorithm introduces delays into the consensus updates of (3) where the weighted average $\mathbf{v}_i(k)$ can use delayed information received from its neighbors, i.e., $0 \leq t_{ij} \leq B$. The information may be several iterations out of date. Under uniformly bounded (but arbitrary) delay and the condition that the next update is done by a random agent, we show that the estimates converge with consensus to the optimizer of the constrained problem (1) almost surely.

Assumption 1. *[Assumptions on the Local Functions f_i] We make the following assumptions on the local objective functions and constraint sets.*

- **(a)** *Each function $f_i : \mathbb{R}^N \to \mathbb{R}$ is convex.*
- **(b)** *The functions f_i, $i \in 1, 2, \ldots, n$, are differentiable and have Lipschitz gradients with a constant L over $\mathbb{R}^N$,*

$$\|\nabla f_i(\mathbf{x}) - \nabla f_i(\mathbf{y})\| \leq L\|\mathbf{x} - \mathbf{y}\|$$

  for all $\mathbf{x}$, $\mathbf{y} \in \mathbb{R}^N$.
- **(c)** *The gradients $\nabla f_i(\mathbf{x})$, where $i \in V$ are bounded over the set $\mathcal{X}$ where $\mathcal{X}^* \subset \mathcal{X}$ and $\mathcal{X}^* = \{\mathbf{x} | \mathbf{x} = \arg\min f(\mathbf{x})\}$; i.e., there exists a constant G_f such that $\|\nabla f_i(\mathbf{x})\| \leq G_f$ for all $\mathbf{x} \in \mathcal{X}$ and all $i \in V$.*
- **(d)** *Each projection set $\mathcal{X}_i^j$ is not necessarily bounded, where $\mathcal{X} \triangleq \cap_{i=1, j \in I_i} \mathcal{X}_i^j$ and I_i as defined earlier.*

Remark 1. *It is worth mentioning that we do not require that the described projection sets be bounded. The proof for this case easily follows from the supermartingale theorem and there is no need for the analysis adapted in Procedure A.*

Assumption 2. *[Network Topology and Weight Matrices] For all $k \geq 0$, we have:*

(a) *The matrices $[\mathbf{W}(k)]_{ij}$ are equal to $w_{ij}(k)$ in (3c) which are chosen locally depending on the network connection topology at each activated node.*

(b) *$\sum_{j=1}^{n} [\mathbf{W}(k)]_{ij} = 1$ for all $i \in V$. This is a local behavior which can be adjusted locally at each activated node that receives estimates. That is $\mathbf{W}$ is row stochastic.*

(c) *There exists a scalar $\nu \in (0, 1)$ such $[\mathbf{W}(k)]_{ij} \geq \nu$ if $[\mathbf{W}(k)]_{ij} > 0$.*

(d) *$\sum_{i=1}^{n} [\mathbf{W}(k)]_{ij} \leq n$ for all $j \in V$. Which is satisfied since $[\mathbf{W}(k)]_{ij} \leq 1$.* ***(e)*** *If server i is disconnected from server j at instant k, then $[\mathbf{W}(k)]_{ij} = 0$.*

Assumption 3 (Bounded Delay). *We assume that our asynchronous algorithm DARPA has a uniformly but arbitrary bounded delay of B. That is, the updating equations (3) has $0 \leq t_{ij}(k) \leq B$ In other words, each activated node i at global instant k can receive estimates from neighboring nodes of instants $k^{'}$ where $0 \leq k - k^{'} \leq B$*

Assumption 4 (Gradient Estimation Error). *We assume that there is a scalar σ such that $\mathbb{E}[\|\epsilon_i(k)\|^2 | \mathcal{F}_{k-1}, I_k] \leq \sigma^2$ with probability 1 for all i and $k \geq k_*$. See [7], [1] and [2].*

Assumption 5 (Uncoordinated Diminishing Stepsizes). *For a diminishing stepsize, we use $\alpha_{k,i} = \frac{1}{\Gamma_i(k)}$ where $\Gamma_i(k)$ denotes the number of updates that agent i has performed until time k as in [7], (i.e., the activation probability of node i is $\gamma_i = \frac{1}{n} \sum_{j \in \mathcal{N}_i} p_{ij}$, where now the connection links allows to the activation of node i in receiving information). We define $\gamma_{min} \triangleq \min_i \gamma_i$.*

Using Lemma 3 in [7] we get that for $p_{min} = \min_{i,j \in \mathcal{E}(k)} p_{ij}$ where $p_{ij} > 0$ is the probability that the edge $\{i, j\}$ is functioning. And let $0 < q < \frac{1}{2}$. Then, there exists a large enough $\tilde{k}^* = \tilde{k}^*(q, n)$ such that with probability 1 for all $k \geq \tilde{k}^*$ and $i \in V$,

$$\alpha_{i,k} \leq \frac{2}{k\gamma_i}, \quad \alpha_{i,k}^2 \leq \frac{4n^2}{k^2 p_{min}^2}, \quad f_k = |\alpha_{i,k} - \frac{1}{k\gamma_i}| \leq \frac{2}{k^{\frac{3}{2}-q} p_{min}^2}.$$

Then $\sum_{k=0}^{\infty} \alpha_{i,k}^2 < \infty$ and $\sum_{k=0}^{\infty} f_k < \infty$.

For the random projections on subsets of the local sets which correspond to the random sequences $\{\Omega_i(k)\}$, $i \in V$, we assume the following:

Assumption 6 (Random Projections Process). *As in [5], the sequences $\{\Omega_i(k)\}$ for $i \in V$, are independent and identically distributed, and independent of the initial points $v_i(0)$ for $i \in \{1, ..., n\}$. We have $\pi_i^j \triangleq Pr\{\Omega_i(k) = j\} > 0$ for all $j \in I_i$ and $i \in V$. the variable $\Omega_i(k)$ is a random sample at time k of a random variable Ω_i that takes values $j \in I_i$ with probability π_i^j.*

We assume the following condition holds, and refer readers to [3] and [4] for details on how it is satisfied given the problem model and assumptions above.

Condition 1. *For all $i \in V$, there exists a constant $c > 0$ such that for all $x \in \mathbb{R}^d$,*

$$dist^2(x, \mathcal{X}) \leq c\, \mathbb{E}[dist^2(x, \mathcal{X}_i^{\Omega_i(k)})]. \tag{2}$$

III. Main Algorithm: Distributed Asynchronous Random Projection Algorithm (DARPA)

We propose DARPA that uses the asynchronous time model described in using a single virtual clock [6, 7]. We assume that agents wakes up in an arbitrary and asynchronous manner that satisfies the Bounded Delay Assumption (i.e., Assumption 3, i.e., $0 \leq t_{ij} \leq B$) is satisfied.

In our analysis, as in paper [5], we also make use of the supermartingale convergence result due to Robbins and Siegmund (see Lemma 10-11, p. 49-50 [10] or original paper [13], p. 111-135) stated in Theorem 1.

Theorem 1. *Suppose $\mathcal{F}_k$ denotes the collection $v_0, ..., v_k$, $u_0, ..., u_k$, $e_0, ..., e_k$ and $c_0, ..., c_k$; and let v_k, u_k, e_k and c_k be sequences of non-negative random variables such that*

$$\mathbb{E}[v_{k+1} | \mathcal{F}_k] \leq (1 + e_k) v_k - u_k + c_k \quad \text{for all } k \geq 0 \text{ a.s.} \tag{4}$$

Also, let $\sum_{k=0}^{\infty} e_k < \infty$ and $\sum_{k=0}^{\infty} c_k < \infty$ a.s; then $\lim_{k \to \infty} v_k = v$ for a random variable $v \geq 0$ a.s. and $\sum_{k=0}^{\infty} u_k < \infty$ a.s.

IV. Convergence Analysis for the Main Algorithm DARPA

In what follows we prove the convergence of DARPA. It is worth mentioning that the more common distributed asynchronous gradient descent can be viewed as a special case of the asynchronous distributed random projection algorithm

Algorithm 1 DARPA Algorithm

Input: Initialization: Initialize estimates $\mathbf{x}_i(0)$; set $\mathbf{x}_i(0)$ to its initial value or to an arbitrary random value. Define Tolerance value tol for halting the algorithm and set $e_i(0) = tol$. $k = 0$.

1: **while** $e_i(k) \geq tol$ **do** {Halting is done at each node independently with no coordination}

2: $k = k + 1$

3: **At each node** i **Update Estimates**

$$\mathbf{v}_i(k) = \sum_{j=1}^{n} w_{ij}(k)\mathbf{x}_j(k - t_{ij}(k)). \tag{3a}$$

$$\begin{aligned} &\textbf{Estimating the local gradient } \widehat{\nabla f_i}(\mathbf{v}_i(k)) \\ &\textbf{where } \widehat{\nabla f_i}(\mathbf{v}_i(k)) = \nabla f_i(\mathbf{v}_i(k)) + \epsilon_i(k) \end{aligned} \tag{3b}$$

$$\mathbf{x}_i(k+1) = \Pi_{\mathcal{X}_i^{\Omega_i(k)}}(\mathbf{v}_i(k) - \alpha_{i,k}\widehat{\nabla f_i}(\mathbf{v}_i(k))) \tag{3c}$$

4: **Find error** $e_i(k) = \|\mathbf{x}_i(k) - \mathbf{x}_i(k-1)\|$ **for all** $1 \leq i \leq n$

5: **end while**

Output: $\mathbf{x}_i(k)$ **for the corresponding** k **for each node**

where the constraint set $\mathcal{X} = \mathbb{R}^N$. That is, the projection is over $\mathcal{X}_i^{\Omega_i(k)} = \mathbb{R}^N$ where $\mathcal{X} = \cap_i \mathcal{X}_i^{\Omega_i(k)} = \cap_i \mathbb{R}^N = \mathbb{R}^N$. In other words, the asynchronous distributed gradient descent algorithm solves for the unconstrained problem

$$\min_{x \in \mathbb{R}^N} f(x). \tag{5}$$

We thus have that the convergence proof of this algorithm as well. The proof with the gradient estimation error follows if Assumption 4 is satisfied. While the exact gradient evaluation case follows directly by assuming the variance is zero.

Theorem 2. *Consider Algorithm 1, and suppose Assumptions 1-7 hold. Let* $f^* = min_{x \in \mathcal{X}} f(\mathbf{x})$ *and* $\mathcal{X}^* = \{x \in \mathcal{X} | f(\mathbf{x}) = f^*\}$. *Assume then that* $\mathcal{X}^* \neq \Phi$. *Then, the iterates* $\{x(k)\}$ *generated by our algorithm (3a)-(3c) converge almost surely to the solution* $\mathbf{x}^* \in \mathcal{X}^*$, *i.e.,*

$$\lim_{k \to \infty} x(k) = \mathbf{x}^* \ \textit{for all } i \in V \ \textit{a.s.} \tag{6}$$

Proof: The theorem follows by applying Lemma 1, 3 and 4 in the described order. ■

A. Convergence Proof Main Parts

Assumption 7.

$$\frac{1}{2}\alpha_{i,k}^2(\sum_{i=1}^{nB}\|R_i(k-1)\|^2 + \sum_{i=1}^{nB}\|R_i(k-B-1)\|^2) + nBB_\tau\alpha_{i,k}^2G_f^2$$

$$\leq (-1 + \frac{3}{8\tau} - A_\tau\alpha_{i,k}^2 - 2\alpha_{i,k}L)\sum_{i=1}^{nB} dist^2(\tilde{\mathbf{v}}_i(k), \mathcal{X}) \ \ a.s. \textit{ and}$$

$$4\alpha_{i,k}G_f\sum_{i=1}^{nB}\|\tilde{\mathbf{v}}_i(k) - \bar{\mathbf{v}}(k)\| + \frac{1}{2}\alpha_{i,k}^2\sum_{i=1}^{nB}\|R_i(k-1)\|^2$$

$$+ \frac{1}{2}\alpha_{i,k}^2\sum_{i=1}^{nB}\|R_i(k-B-1)\|^2 + nB_\tau\alpha_{i,k}^2G_f^2 \leq 2\alpha_{i,k}B(f(\bar{\mathbf{z}}(k)) - f(\mathbf{x}^*))$$

a.s. for $k \geq \tilde{k}$.

Remark 2. *Under the assumptions on the uncoordinated stepsizes presented in Assumption 5 and having for* $k \geq \tilde{k_*}$ *that Assumption 4 is satisfied then Assumption 7 is satisfied. However, we refrain from analyzing that here due to the limited space but we refer you to a future extended version.*

B. Reduction to a Consensus Problem without Delay

Here, we reduce the original agent system with delays to a system without delays, under the Bounded Delays assumption [cf. Assumption 3]. This idea has also been used in the distributed computation model of Nedich et al. [9], and it motivates our development here.

With some modifications adapted to form mathematical structures that suit our case we get:

The relation in (3) for the evolution of estimates of computing agents is given by: for all $i \in \{1, \ldots, , nB\}$,

$$\tilde{\mathbf{v}}_i(k+1) = \sum_{j=1}^{2nB}[\bar{\mathbf{W}}(k+1)]_{ij}\tilde{\mathbf{x}}_j(k+1). \tag{7}$$

Notice that for weighted averages $\tilde{\mathbf{v}}_i$ we have $i \in \{1, \ldots, , nB\}$. And that for estimates $\tilde{\mathbf{x}}_i$ we have $j \in \{1, \ldots, , 2nB\}$. This is because the weighted averages of the computing and noncomputing nodes total to nB for instant k. And these weighted averages are each dependent on estimates of delay from 0 up to B of that instant. Thus, we have nB estimates for $\tilde{\mathbf{v}}_i(k+1)$ of computing and noncomputing nodes beginning from $\tilde{\mathbf{x}}_p(k)$ and ending in $\tilde{\mathbf{x}}_{p+nB-1}(k)$ where $p = ((i-1)\ div\ n)n + 1$ and $i \in \{1, \ldots, , nB\}$. Thus, for $\tilde{\mathbf{x}}_j(k)$ we have $j \in \{1, \ldots, , 2nB\}$.

And for all $i \in \{1, \ldots, nB\}$ and $h \in \{1, \ldots, 2nB\}$.

$$[\bar{\mathbf{W}}(k+1)]_{lh} = \begin{cases} [\mathbf{W}(k+1-s)]_{ij} & \text{if } h = j + tn,\ t = t_{ij}(k+1) \\ & i = ((l-1)\ mod\ n) + 1 \\ & s = (l-1)\ div\ n \\ 0 & \text{otherwise} \quad \text{for all } k \geq 0 \end{cases} \tag{8}$$

and $[\mathbf{W}(k)]_{ij}$ are the weights used by the agents in the original network.

But in the original system we have for $i \in \{1, \ldots, n\}$,

$$\mathbf{x}_i(k+1) = \Pi_{\mathcal{X}_i^{\Omega_i(k)}}(\mathbf{v}_i(k) - \alpha_{i,k}\widehat{\nabla f_i}(\mathbf{v}_i(k))), \tag{9}$$

$$\text{where } \widehat{\nabla f_i}(\mathbf{v}_i(k)) = \nabla f_i(\mathbf{v}_i(k)) + \epsilon_i(k). \tag{10}$$

$$\text{And } \mathbf{v}_i(k) = \sum_{j=1}^{n}[\mathbf{W}(k)]_{ij}\mathbf{x}_j(k - t_{ij}(k)), \tag{11}$$

Then for the extended system we have for $i \in \{1, \ldots, 2nB\}$

$$\tilde{\mathbf{x}}_i(k+1) = P_i(\sum_{j=1}^{2nB}[\tilde{\mathbf{W}}(k)]_{ij}\tilde{\mathbf{x}}_j(k) - \sum_{j=1}^{2nB}[\overline{\mathbf{W}}(k)]_{ij}\widehat{\overline{\nabla f}_j}(\mathbf{v}_i(k))), \tag{12}$$

where

$$P_i = \begin{cases} \Pi_{\mathcal{X}_l^{\Omega_{l(r)}}} & \text{for } i \in \{1, \ldots, nB\} \\ & l = ((i-1)\ mod\ n) + 1 \\ & r = k - ((i-1)\ div\ n) \\ 1 & \text{otherwise} \quad \text{for all } k \geq 0 \end{cases} \tag{13}$$

And $[\overline{\mathbf{W}}(k)]$ is such that

$$[\overline{\mathbf{W}}(k+1)]_{ij} = \begin{cases} \alpha_{l,r} & \text{for } j = i \text{ and } i \in \{1, \ldots, nB\} \\ & l = ((i-1)\ mod\ n) + 1 \\ & r = k - ((i-1)\ div\ n) \\ 0 & \text{otherwise} \quad \text{for all } k \geq 0 \end{cases} \tag{14}$$

and $\widehat{\overline{\nabla f}_q}(\mathbf{v}_i(k)) = \widehat{\nabla f}_j(\mathbf{v}_i(k))$ for $q = j + ln$ for $0 \leq l \leq 2B - 1$.

And the weights $[\tilde{\mathbf{W}}(k)]_{ij}$ for $i \in \{1,\dots,nB\}$ are given by $[\bar{\mathbf{W}}(k)]_{ij}$ of (8) where $j \in \{1,\dots,2nB\}$.

And the evolution of estimates for $i = nB+1,\dots,2nB$,

$$\tilde{\mathbf{x}}_i(k+1) = \tilde{\mathbf{x}}_{i-n}(k) \text{ for all } k \geq 0, \tag{15}$$

Therefore, for $i = nB+1,\dots,2nB$, we have

$$[\tilde{\mathbf{W}}(k+1)]_{ih} = \begin{cases} 1 & \text{for } h = i-n \\ 0 & \text{otherwise} \quad \text{for all } k \geq 0 \end{cases} \tag{16}$$

And the initial estimates are given by

$$\begin{cases} \tilde{\mathbf{x}}_i(0) = \mathbf{x}_i(0) & \text{for } i \in \{1,\dots,n\} \\ \tilde{\mathbf{x}}_i(q) = 0 & \text{otherwise} \quad \text{for all } i \in \{n+1,\dots,2nB\} \end{cases}$$

It is worth mentioning that (9) becomes

$$\mathbf{x}_i(k+1) = \Pi_{\mathcal{X}_i^{\Omega_i(k)}}(\mathbf{v}_i(k) - \alpha_{i,k}\nabla f_i(\mathbf{v}_i(k)) - \alpha_{i,k}\epsilon_i(k)), \tag{17}$$

where $\epsilon_i(k)$ is the error added by estimating the gradient.

Definition 1. *Let's denote* $R_i(k) = -\alpha_{i,k}\epsilon_i(k)$.

Then (12) for $i \in \{1,\dots,nB\}$ can be written

$$\begin{aligned} \tilde{\mathbf{x}}_i(k+1) &= \Pi_{\mathcal{X}_l^{\Omega_l(r)}}(\tilde{\mathbf{v}}_i(k) - \alpha_{i,k}\nabla f_i(\tilde{\mathbf{v}}_i(k)) + R_i(k-1)) \\ \textit{and} \quad \overline{\tilde{\mathbf{x}}_i}(k+1) &= \Pi_{\mathcal{X}_l^{\Omega_l(r)}}(\tilde{\mathbf{v}}_i(k) - \alpha_{i,k}\nabla f_i(\tilde{\mathbf{v}}_i(k))), \end{aligned} \tag{18}$$

with l and r as in (14).

C. Main Proof

1) **Part 1 of the Convergence Proof (Bound $\sum_{\mathbf{k=0}}^{\infty} \mathbf{dist^2}(\tilde{\mathbf{v}}_\mathbf{i}(\mathbf{k}), \mathcal{X}) < \infty$) ::**

Lemma 1. *Let Assumption 1-7 hold. Then* $\sum_{k=0}^{\infty} dist^2(\tilde{\mathbf{v}}_i(k), \mathcal{X}) < \infty$ *for all* $i \in V$ *a.s.*

Proof: We are going to prove Lemma 1 by using the following Lemma 2 and then exploiting the supermartingale Theorem 1.

Lemma 2. *Let* $\mathcal{Y} \subset \mathbb{R}^N$ *be a closed convex set. Let the function* $\Phi : \mathbb{R}^N \to \mathbb{R}$ *be convex and differentiable over* $\mathbb{R}^N$ *with Lipschitz continuous gradients with a constant* L.
Let y *be given by* $y = \Pi_{\mathcal{Y}}(\mathbf{x} - \alpha\nabla\Phi(\mathbf{x}))$ *for some* $x \in \mathbb{R}^N$, $\alpha > 0$.
Then, we have for any $\hat{\mathbf{x}} \in \mathcal{Y}$ *and* $z \in \mathbb{R}^N$,

$$\begin{aligned} \|\mathbf{y} - \hat{\mathbf{x}}\|^2 \leq (1 + A_\tau\alpha^2)\|\mathbf{x} - \hat{\mathbf{x}}\|^2 \\ - 2\alpha(\Phi(\mathbf{z}) - \Phi(\hat{\mathbf{x}})) - \frac{3}{4}\|\mathbf{y} - \mathbf{x}\|^2 \\ + (\frac{3}{8\tau} + 2\alpha L)\|\mathbf{x} - \mathbf{z}\|^2 + B_\tau\alpha^2\|\nabla\Phi(\hat{\mathbf{x}})\|^2, \end{aligned} \tag{19}$$

where $A_\tau = 8L^2 + 16\tau L^2$, $B_\tau = 8\tau + 8$ *and* $\tau > 0$ *is arbitrary.*

See Lemma 4 in [5] or Lemma 2 in [8].

Since $\mathcal{X} \triangleq \cap_{i=1}^n \mathcal{X}_i$ then let $\hat{\mathbf{x}} \in \mathcal{X}$. This implies that $\hat{\mathbf{x}} \in \mathcal{X}_i$ for all $i \in \{1,2,\dots,n\}$. But f_i is Lipschitz on $\mathcal{X}_i$ for all $i \in \{1,2,\dots,n\}$, then f_i is Lipschitz on $\mathcal{X}$ (in fact in our case, f is Lipschitz on the whole $\mathbb{R}^N$, this for ease of implementation since we pick the matrix A randomly.)

By requiring $i \in \{1,2,\dots,nB\}$, we take $\overline{\tilde{\mathbf{x}}_i}(k+1) = \Pi_{\mathcal{X}_l^{\Omega_l(r)}}(\tilde{\mathbf{v}}_i(k) - \alpha_{i,k}\nabla f_i(\tilde{\mathbf{v}}_i(k))) \in \mathbb{R}^N$ where l and r are as in (14), then being in $\mathbb{R}^N$ the following inequalities hold,

$$\begin{aligned} dist(\overline{\tilde{\mathbf{x}}_i}(k+1), \mathcal{X}) &= \|\overline{\tilde{\mathbf{x}}_i}(k+1) - \Pi_{\mathcal{X}}(\overline{\tilde{\mathbf{x}}_i}(k+1))\| \\ &\leq \|\overline{\tilde{\mathbf{x}}_i}(k+1) - \hat{\mathbf{x}}\|. \end{aligned} \tag{20}$$

To use Lemma 2, we thus use the following substitutions: $\Phi = f_i$, $\alpha = \alpha_{i,k}$, $\hat{\mathbf{x}} \in \mathcal{X} = \cap_{i=1}^n \mathcal{X}_i$, $\mathbf{y} = \overline{\tilde{\mathbf{x}}_i}(k+1) = \Pi_{\mathcal{X}_l^{\Omega_l(r)}}(\tilde{\mathbf{v}}_i(k) - \alpha_{i,k}\nabla f_i(\tilde{\mathbf{v}}_i(k)))$ and $\mathbf{x} = \tilde{\mathbf{v}}_i(k)$ where l and r are as in (14). In particular, if we take $\hat{\mathbf{x}} = \Pi_{\mathcal{X}}[\tilde{\mathbf{v}}_i(k)] \in \mathcal{X}$ in the feasibility region and $\mathbf{z} = \Pi_{\mathcal{X}}(\tilde{\mathbf{v}}_i(k)) = \hat{\mathbf{x}} \in \mathcal{X}$, but $\mathcal{X} = \cap_{i=1}^n \mathcal{X}_i \subset \cap_{i=1}^n \mathcal{X}_i^{\Omega_i(k)}$ since $\mathcal{X}_i \subset \mathcal{X}_i^{\Omega_i(k)}$ for all $i \in V$ then $\hat{\mathbf{x}} \in \mathcal{Y} = \mathcal{X}_l^{\Omega_l(r)} = \mathcal{X}_i^{\Omega_i(k)}$ for some i where l and r are as in (14), and $\mathbf{z} \in \mathbb{R}^N$ so we can apply Lemma 2 then

$$\begin{aligned} dist^2(\tilde{\mathbf{x}}_i(k+1), \mathcal{X}) \leq (1 + A_\tau\alpha_{i,k}^2)\|\tilde{\mathbf{v}}_i(k) - \Pi_{\mathcal{X}}(\tilde{\mathbf{v}}_i(k))\|^2 \\ - 2\alpha_{i,k}(f_i(\Pi_{\mathcal{X}}(\tilde{\mathbf{v}}_i(k))) - f_i(\Pi_{\mathcal{X}}(\tilde{\mathbf{v}}_i(k)))) \\ - \frac{3}{4}\|\Pi_{\mathcal{X}_l^{\Omega_l(r)}}(\tilde{\mathbf{v}}_i(k) - \alpha_{i,k}\nabla f_i(\tilde{\mathbf{v}}_i(k))) - \tilde{\mathbf{v}}_i(k)\|^2 \\ + (\frac{3}{8\tau} + 2\alpha_{i,k}L)\|\tilde{\mathbf{v}}_i(k) - \Pi_{\mathcal{X}}(\tilde{\mathbf{v}}_i(k))\|^2 \\ + B_\tau\alpha_{i,k}^2\|\nabla f_i(\hat{\mathbf{x}})\|^2, \end{aligned} \tag{21}$$

$$\begin{aligned} \text{But} \quad \|\Pi_{\mathcal{X}_i^{\Omega_i(k)}}(\tilde{\mathbf{v}}_i(k) - \alpha_{i,k}\nabla f_i(\tilde{\mathbf{v}}_i(k))) - \tilde{\mathbf{v}}_i(k)\|^2 \\ \geq \|\Pi_{\mathcal{X}_i^{\Omega_i(k)}}(\tilde{\mathbf{v}}_i(k)) - \tilde{\mathbf{v}}_i(k)\|^2 = dist^2(v_i(k), \mathcal{X}_i^{\Omega_i(k)}), \end{aligned} \tag{22}$$

$$\begin{aligned} \text{Then} \quad -\frac{3}{4}\|\Pi_{\mathcal{X}_i^{\Omega_i(k)}}(\tilde{\mathbf{v}}_i(k) - \alpha_{i,k}\nabla f_i(\tilde{\mathbf{v}}_i(k))) - \tilde{\mathbf{v}}_i(k)\|^2 \\ \leq -\frac{3}{4}dist^2(v_i(k), \mathcal{X}_i^{\Omega_i(k)}) \end{aligned} \tag{23}$$

$$\text{But} \quad \mathbb{E}[dist^2(v_i(k), \mathcal{X}_i^{\Omega_i(k)})/\mathcal{F}_k] \geq \frac{1}{\tau}dist^2(v_i(k), \mathcal{X}) \tag{24}$$

Let $\mathcal{F}_k$ be the σ-algebra generated by the entire history of the algorithm up to time k inclusively, that is $\mathcal{F}_k = \{\tilde{\mathbf{x}}_i(0), i \in V\} \cup \{\Omega_i(l) : 0 \leq l \leq k, i \in V\}$. Therefore, given $\mathcal{F}_k$, the collection $\tilde{\mathbf{x}}_i(0), ..., \tilde{\mathbf{x}}_i(k+1)$ and $\tilde{\mathbf{v}}_i(0), ..., \tilde{\mathbf{v}}_i(k+1)$ generated by the algorithm is fully determined.

But $\mathcal{Y} = \mathcal{X}_l^{\Omega_l(r)} = \mathcal{X}_i^{\Omega_i(k)}$ for some $i \in V$ where l and r are as in (14), then we have

$$\mathbb{E}[dist^2(v_i(k), \mathcal{X}_l^{\Omega_l(r)})/\mathcal{F}_k] \geq \frac{1}{\tau}dist^2(v_i(k), \mathcal{X}) \tag{25}$$

where l and r are as in (14). Then taking expectation over $\mathcal{F}_k$, (21) becomes

$$\begin{aligned} \mathbb{E}[dist^2(\overline{\tilde{\mathbf{x}}_i}(k+1), \mathcal{X})/\mathcal{F}_k] \leq (1 + A_\tau\alpha_{i,k}^2)dist^2(v_i(k), \mathcal{X}) \\ + (-\frac{3}{8\tau} + 2\alpha_{i,k}L)dist^2(v_i(k), \mathcal{X}) + B_\tau\alpha_{i,k}^2\|\nabla f_i(\hat{\mathbf{x}})\|^2 \quad a.s. \end{aligned} \tag{26}$$

Then from the convexity of the norm squared, we have

$$dist^2(\tilde{\mathbf{v}}_i(k), \mathcal{X}) \leq \sum_{j=1}^{2nB} [\tilde{\mathbf{W}}(k)]_{ij} dist^2(\tilde{\mathbf{x}}_j(k), \mathcal{X}). \tag{27}$$

But we can deduce using updating equation (3a) and projection Lemmas that

$$dist^2(\tilde{\mathbf{x}}_j(k), \mathcal{X}) \leq 2dist^2(\overline{\tilde{\mathbf{x}}_j}(k), \mathcal{X}) + 2\|R_i(k-1)\|^2. \tag{28}$$

By observing the first term in RHS of the inequality (26), we get

$$\begin{aligned} (1 + A_\tau\alpha_{i,k}^2)\ dist^2(\tilde{\mathbf{v}}_i(k), \mathcal{X}) \leq \\ \frac{1}{4nB}\alpha_{i,k}^2\ dist^2(\tilde{\mathbf{v}}_i(k), \mathcal{X}) + (1 + A_\tau\alpha_{i,k}^2)\ dist^2(\tilde{\mathbf{v}}_i(k), \mathcal{X}), \end{aligned} \tag{29}$$

and by using (27) on the first term, we get

$$\begin{aligned} dist^2(\overline{\tilde{\mathbf{x}}_i}(k+1), \mathcal{X})/\mathcal{F}_k \leq \frac{1}{4nB}\alpha_{i,k}^2 \sum_{j=1}^{2nB} [\tilde{\mathbf{W}}(k)]_{ij} dist^2(\tilde{\mathbf{x}}_j(k), \mathcal{X}) \\ + (1 + A_\tau\alpha_{i,k}^2)\ dist^2(\tilde{\mathbf{v}}_i(k), \mathcal{X}) \\ + (-\frac{3}{8\tau} + 2\alpha_{i,k}L)dist^2(\tilde{\mathbf{v}}_i(k), \mathcal{X} + B_\tau\alpha_{i,k}^2 G_f^2, \end{aligned} \tag{30}$$

where $\|\nabla f_i(\hat{\mathbf{x}})\| \leq G_f$ (i.e., gradient is bounded on set $\mathcal{X}$).

Then summing from $i = 1$ to nB, having (30) and taking into consideration activation with probability γ_i for each i and $\overline{\tilde{\mathbf{x}}_i}(k+1) = \tilde{\tilde{v}}_i(k)$ for no activation (N.B. notice that under suitable manipulation the inequality will reduce to the same as full activation except care should be taken for the third term where the coefficient is negative and have γ_i in $\alpha_{i,k}$) i.e., use $1 - \gamma_i \leq (1 + A_\tau\alpha_{i,k}^2)(1 - \gamma_i)$ and $0 \leq \gamma_i \leq 1$.

But having the matrix $[\tilde{\mathbf{W}}(k)]_{ij}$ for $i = \{1, \ldots, nB\}$ and $j = \{1, \ldots, 2nB\}$ constructed of B row stochastic block matrices (i.e., $\sum_{i=1}^{nB}[\tilde{\mathbf{W}}(k)]_{ij} \leq nB$), we have for $k \geq \tilde{k}^*$ by

Then the result of (28) for $i = \{1, \ldots, nB\}$

and having $x_j(k) = x_{j-n}(k-1)$ for $j = \{nB+1, \ldots, 2nB\}$ and cascading until $k - B$, we have $\sum_{i=nB+1}^{2nB} dist^2(\tilde{\mathbf{x}}_i(k), \mathcal{X}) = \sum_{i=1}^{nB} dist^2(\tilde{\mathbf{x}}_i(k-B), \mathcal{X})$ then we have for $k \geq \tilde{k}^*$

$$\begin{aligned}
&\mathbb{E}[\sum_{i=1}^{nB} dist^2(\overline{\tilde{\mathbf{x}}_i}(k+1), \mathcal{X})/\mathcal{F}_k] \leq \\
&\frac{1}{2}\alpha_{i,k}^2 \sum_{i=1}^{nB} dist^2(\overline{\tilde{\mathbf{x}}_i}(k), \mathcal{X}) + \frac{1}{2}\alpha_{i,k}^2 \sum_{i=1}^{nB} dist^2(\overline{\tilde{\mathbf{x}}_i}(k-B), \mathcal{X}) \\
&+ (1 - \frac{3\gamma_{min}}{8\tau} + A_\tau \alpha_{i,k}^2 + \frac{4}{k}L) \sum_{i=1}^{nB} dist^2(\tilde{\mathbf{v}}_i(k), \mathcal{X}) \\
&+ \frac{1}{2}\alpha_{i,k}^2 (\sum_{i=1}^{nB} \|R_i(k-1)\|^2 + \sum_{i=1}^{nB} \|R_i(k-B-1)\|^2) + nBB_\tau \alpha_{i,k}^2 G_f^2 \ a.s.
\end{aligned} \tag{31}$$

We now apply **Procedure A** with $a_k = 1 + C\alpha_{i,k}^2$ where $C > 0$ to give some degree of freedom of choosing $\{a_k\}$ with the required properties, then (31) implies

$$\begin{aligned}
&\mathbb{E}[\sum_{i=1}^{nB} dist^2(\overline{\tilde{\mathbf{x}}_i}(k+1), \mathcal{X})/\mathcal{F}_k] \leq (1 + C\alpha_{i,k}^2) \sum_{i=1}^{nB} dist^2(\overline{\tilde{\mathbf{x}}_i}(k), \mathcal{X}) \\
&+ (1 - \frac{3}{8\tau} + A_\tau \alpha_{i,k}^2 + 2\alpha_{i,k}L) \sum_{i=1}^{nB} dist^2(\tilde{\mathbf{v}}_i(k), \mathcal{X}) \\
&+ \frac{1}{2}\alpha_{i,k}^2 (\sum_{i=1}^{nB} \|R_i(k-1)\|^2 + \sum_{i=1}^{nB} \|R_i(k-B-1)\|^2) + nBB_\tau \alpha_{i,k}^2 G_f^2 \ a.s.
\end{aligned} \tag{32}$$

for $k > \hat{k} > \bar{k} > \bar{k} - B > \max(k^*, \tilde{k})$ where $k^* = \max(k_1, \tilde{k}^*)$.

Define the following substitutions :

$$\begin{aligned}
&v_k = \sum_{i=1}^{nB} dist^2(\overline{\tilde{\mathbf{x}}_i}(k), \mathcal{X}), \ u_k = \sum_{i=1}^{nB} dist^2(\tilde{\mathbf{v}}_i(k), \mathcal{X}), \\
&-b_k = 1 - \frac{3\gamma_{min}}{8\tau} + A_\tau \alpha_{i,k}^2 + \frac{4}{k}L \\
&c_k = \frac{1}{2}\alpha_{i,k}^2 (\sum_{i=1}^{nB} \|R_i(k-1)\|^2 + \sum_{i=1}^{nB} \|R_i(k-B-1)\|^2) + nBB_\tau \alpha_{i,k}^2 G_f^2
\end{aligned} \tag{33}$$

We require initially that $-b_k \leq -1$ for $k > k_1$. But the terms $A_\tau \alpha_{i,k}^2 + \frac{4}{k}L$ in b_k although dependent on τ they can be controlled by the value of $\alpha_{i,k}$, thus k. Then, we can require that $A_\tau \alpha_{i,k}^2 + \frac{4}{k}L \leq A_\tau \frac{4n^2}{k^2 p_{min}^2} + \frac{4L}{k} < \epsilon$ for $k > k_1$ and $1 + \epsilon < \frac{3\gamma_{min}}{8\tau}$ where $1 - \frac{3\gamma_{min}}{8\tau} + \epsilon < -1$. This suffice to use $\tau < \frac{3\gamma_{min}}{16}$, e.g., $\tau = \frac{\gamma_{min}}{8}$. Thus,

$$\text{Taking } \tau < \frac{3\gamma_{min}}{16} \text{ , e.g., } \tau = \frac{\gamma_{min}}{8} \text{ we have for } k > k_1 \text{ that } b_k < -1 \tag{34}$$

To apply the supermartingale convergence theorem on (32) with the substitutions described in (33) along with $a_k = 1 + C\alpha_{i,k}^2$, We consider $k > \hat{k}$ which is an end part of the tail of the sequence where (34) is satisfied. Therefore, with the above condition on b_k satisfied for $k > \hat{k} > k_1$, (32) can be reduced to the supermartingale inequality. Therefore, for $k > \hat{k} > \max(k_1, \tilde{k}, \tilde{k}^*) \geq k_1$, (32) can be reduced to the supermartingale inequality. Then, having $a_k = 1 + e_k = 1 + C\alpha_{i,k}^2$ we get that $\sum_{k=0}^{\infty} e_k < \infty$ and $\sum_{k=0}^{\infty} c_k < \infty$ since $\sum_{k=0}^{\infty} \alpha_{i,k}^2 < \infty$. Then the supermartingale convergence theorem holds.

From the supermartingale theorem holding for the tail of the sequence $k > k'$ in $\mathbf{A}$, then we have $\sum_{k=0}^{\infty} u_k < \infty$, i.e., $\sum_{k=0}^{\infty} \sum_{i=1}^{n} dist^2(\tilde{\mathbf{v}}_i(k), \mathcal{X}) < \infty$. We can interchange infinite and finite sums, as an implicit consequence of the linearity of these sums. Thus, we have $\sum_{i=1}^{n}(\sum_{k=0}^{\infty} dist^2(\tilde{\mathbf{v}}_i(k), \mathcal{X})) < \infty \implies$ the argument inside the finite sum is bounded, i.e.,

$$\sum_{k=0}^{\infty} dist^2(\tilde{\mathbf{v}}_i(k), \mathcal{X}) < \infty, \qquad \blacksquare$$

2) **Part 2 of the Convergence Proof::**

Lemma 3. *Let Assumption 1-7 hold. Also, assume that the stepsize sequence $\{\alpha_{i,k}\}$ is non-increasing such that $\sum_{k=0}^{\infty} \alpha_{i,k}^2 < \infty$ (i.e., which is the case of Assumption 5), and define $\mathbf{e}_i(k) = \tilde{\mathbf{x}}_i(k+1) - \tilde{\mathbf{v}}_i(k)$ for all $i \in V$ and $k \geq 0$.*
Then, we have a.s.

$$\begin{aligned}
&\sum_{k=0}^{\infty} ||\mathbf{e}_i(k)||^2 < \infty \ for \ all \ i \in V, \\
&\sum_{k=0}^{\infty} \alpha_{i,k} ||\tilde{\mathbf{v}}_i(k) - \bar{\mathbf{v}}(k)|| < \infty \ for \ all \ i \in V, \\
&where \ \bar{\mathbf{v}}(k) = \frac{1}{2n^2B^2} \sum_{l=1}^{nB} \tilde{\mathbf{v}}_l(k).
\end{aligned} \tag{35}$$

Remark 3. *We refrain from listing the proof here due to the limited space. But the proof is similar to [5] with slight modifications. And we note that Assumption 4 on the gradient estimation error is needed for the proof to follow.*

3) **Part 3 of the Convergence Proof::**

Lemma 4. *Let Assumption 1-7 hold. Let $f^* = min_{x \in \mathcal{X}} f(\mathbf{x})$ and $\mathcal{X}^* = \{x \in \mathcal{X} | f(\mathbf{x}) = f^*\}$. Assume then that $\mathcal{X}^* \neq \Phi$. Then, the iterates $\{x(k)\}$ generated by our algorithm (3a)-(3c) converge almost surely to the solution $\mathbf{x}^* \in \mathcal{X}^*$, i.e.,*

$$\lim_{k \to \infty} x(k) = \mathbf{x}^* \ for \ all \ i \in V \ a.s. \tag{36}$$

Proof: We begin by substituting $\Phi = f_i$, $\alpha = \alpha_{i,k}$, $\hat{\mathbf{x}} \in \mathcal{X}^* \subset \mathcal{X} = \cap_{i=1}^n \mathcal{X}_i$, $\mathbf{y} = \overline{\tilde{\mathbf{x}}_i}(k+1) = \Pi_{\mathcal{X}_l^{\Omega_l(r)}}(\tilde{\mathbf{v}}_i(k) - \alpha_{i,k}\nabla f_i(\tilde{\mathbf{v}}_i(k)))$ and $\mathbf{x} = \tilde{\mathbf{v}}_i(k)$ where l and r are as in (14).

In particular, if we take $\hat{\mathbf{x}} = \mathbf{x}^* \in \mathcal{X}^* \subset \mathcal{X}$ in the solution set and $\mathbf{z} = \mathbf{z}_i(k) = \Pi_{\mathcal{X}}(\tilde{\mathbf{v}}_i(k))$. Then having $\hat{\mathbf{x}} \in \mathcal{X}$ and $\mathcal{X} = \cap_{i=1}^n \mathcal{X}_i \subset \cap_{i=1}^n \mathcal{X}_i^{\Omega_i(k)}$ since $\mathcal{X}_i \subset \mathcal{X}_i^{\Omega_i(k)}$ for all $i \in V$ we get that $\hat{\mathbf{x}} \in \mathcal{Y} = \mathcal{X}_l^{\Omega_l(r)} = \mathcal{X}_i^{\Omega_i(k)}$ for some i where l and r are as in (14), and $\mathbf{z} \in \mathbb{R}^N$ so we can apply Lemma 2 then

$$\begin{aligned}
\|\tilde{\mathbf{x}}_i(k+1) - \mathbf{x}^*\|^2 \leq (1 + A_\tau \alpha_{i,k}^2) \|\tilde{\mathbf{v}}_i(k) - \mathbf{x}^*\|^2 - 2\alpha_{i,k}(f_i(\mathbf{z}_i(k)) - f_i(\mathbf{x}^*)) \\
- \frac{3}{4}\|\Pi_{\mathcal{X}_l^{\Omega_l(r)}}(\tilde{\mathbf{v}}_i(k) - \alpha_{i,k}\nabla f^{(i)}(\tilde{\mathbf{v}}_i(k))) - \tilde{\mathbf{v}}_i(k)\|^2 \\
+ (\frac{3}{8\tau} + 2\alpha_{i,k}L)\|\tilde{\mathbf{v}}_i(k) - \mathbf{z}_i(k)\|^2 + B_\tau \alpha_{i,k}^2 \|\nabla f_i(\mathbf{x}^*)\|^2.
\end{aligned} \tag{37}$$

Then (37) becomes using the reduction (22)-(25), we get

$$\begin{aligned}
&\|\overline{\tilde{\mathbf{x}}_i}(k+1) - \mathbf{x}^*\|^2 \leq \\
&(1 + A_\tau \alpha_{i,k}^2)\|\tilde{\mathbf{v}}_i(k) - \mathbf{x}^*\|^2 - 2\alpha_{i,k}(f_i(\mathbf{z}_i(k)) - f_i(\mathbf{x}^*)) \\
&+ (-\frac{3}{8\tau} + 2\alpha_{i,k}L) dist^2(\tilde{\mathbf{v}}_i(k), \mathcal{X}) + B_\tau \alpha_{i,k}^2 \|\nabla f_i(\hat{\mathbf{x}})\|^2,
\end{aligned} \tag{38}$$

And for $\bar{\mathbf{z}}(k) = \frac{1}{n}\sum_{i=1}^n \mathbf{z}_i(k)$, and for the restricted available space, we follow the same analysis as in [5] but with $\bar{\mathbf{v}}(k) = \frac{1}{2n^2B^2}\sum_{l=1}^{nB} \tilde{\mathbf{v}}_l(k)$ and taking into consideration the probabilistic assumption on activation to arrive at

$$\begin{aligned}
&\sum_{i=1}^n \gamma_i \alpha_{i,k}(f_i(\mathbf{z}_i(k)) - f_i(\mathbf{x}^*)) = \sum_{i=1}^n \frac{2}{k}(f_i(\mathbf{z}_i(k)) - f_i(\bar{\mathbf{z}}(k))) + \frac{2}{k}(f(\bar{\mathbf{z}}(k)) - f(\mathbf{x}^*)) \\
&\geq -\frac{2}{k}G_f \sum_{i=1}^n \|\mathbf{z}_i(k) - \bar{\mathbf{z}}(k)\| + \frac{2}{k}(f(\bar{\mathbf{z}}(k)) - f(\mathbf{x}^*)) \\
&\geq -\frac{4}{k}G_f \sum_{i=1}^n \|\tilde{\mathbf{v}}_i(k) - \bar{\mathbf{v}}(k)\| + \frac{2}{k}(f(\bar{\mathbf{z}}(k)) - f(\mathbf{x}^*)).
\end{aligned} \tag{39}$$

Elaborating in a similar way as Part 1, (38) leads to the following inequality for $k \geq \tilde{k}^*$

$$\mathbb{E}[\sum_{i=1}^{nB} \|\overline{\tilde{\mathbf{x}}_i}(k+1) - \mathbf{x}^*\|^2/\mathcal{F}_k] \leq \frac{1}{2}\alpha_{i,k}^2 \sum_{i=1}^{nB} \|\overline{\tilde{\mathbf{x}}_i}(k) - \mathbf{x}^*\|^2$$
$$+ \frac{1}{2}\alpha_{i,k}^2 \sum_{i=1}^{nB} \|\overline{\tilde{\mathbf{x}}_i}(k-B) - \mathbf{x}^*\|^2 - \frac{4B}{k}(f(\bar{\mathbf{z}}(k)) - f(\mathbf{x}^*)) + \frac{8B}{k} G_f \sum_{i=1}^{nB} \|\tilde{\mathbf{v}}_i(k) - \bar{\mathbf{v}}(k)\|$$
$$+ (1 + A_\tau \alpha_{i,k}^2) \sum_{i=1}^{nB} \|\tilde{\mathbf{v}}_i(k) - \mathbf{x}^*\|^2 + (-\frac{3\gamma_{min}}{8\tau} + \frac{4}{k}L) \sum_{i=1}^{nB} dist^2(\tilde{\mathbf{v}}_i(k), \mathcal{X})$$
$$+ \frac{1}{2}\alpha_{i,k}^2 \sum_{i=1}^{nB} (\|R_i(k-1)\|^2 + \|R_i(k-B-1)\|^2) + nB_\tau \alpha_{i,k}^2 G_f^2 \quad a.s. \tag{40}$$

We can find k_2 such that for $k > k_2$ we have

$$(1 + A_\tau \alpha_{i,k}^2) \sum_{i=1}^{nB} \|\tilde{\mathbf{v}}_i(k) - \mathbf{x}^*\|^2 + (-\frac{3\gamma_{min}}{8\tau} + \frac{4}{k}L) \sum_{i=1}^{nB} dist^2(\tilde{\mathbf{v}}_i(k), \mathcal{X}) < 0 \tag{41}$$

That is, it suffices to have the following two inequalities

$$(1 + A_\tau \alpha_{i,k}^2) \sum_{i=1}^{nB} \|\tilde{\mathbf{v}}_i(k) - \mathbf{x}^*\|^2 + \frac{4}{k}L \sum_{i=1}^{nB} dist^2(\tilde{\mathbf{v}}_i(k), \mathcal{X})$$
$$< (1 + A_\tau \alpha_{i,k}^2 + \frac{4}{k}L) \sum_{i=1}^{nB} \|\tilde{\mathbf{v}}_i(k) - \mathbf{x}^*\|^2 < \frac{3\gamma_{min}}{8\tau} \sum_{i=1}^{nB} dist^2(\tilde{\mathbf{v}}_i(k), \mathcal{X}) \tag{42}$$

$$\sum_{i=1}^{nB} dist^2(\tilde{\mathbf{v}}_i(k), \mathcal{X}) = \sum_{i=1}^{nB} b\|\tilde{\mathbf{v}}_i(k) - \mathbf{x}^*\|^2 < \sum_{i=1}^{nB} \|\tilde{\mathbf{v}}_i(k) - \mathbf{x}^*\|^2 \tag{43}$$

where $0 < b < 1$. We can pick τ such that for $k > k_2$ we have $1 + \epsilon < \frac{3\gamma_{min} b}{8\tau}$. If we pick $\tau < \frac{3\gamma_{min} b}{8}$ we can have for $k > k_2$ where $0 < b < 1$ such that (41) is satisfied. Then by cancelling this negative term (41) we arrive at the reduced inequality for for $k > k^* = \max(k_2, \tilde{k}^*)$ and use the properties of $\alpha_{i,k}$ described in Assumption 5,

$$\mathbb{E}[\sum_{i=1}^{nB} \|\overline{\tilde{\mathbf{x}}_i}(k+1) - \mathbf{x}^*\|^2/\mathcal{F}_k] \leq$$
$$\frac{1}{2}\alpha_{i,k}^2 \sum_{i=1}^{nB} \|\overline{\tilde{\mathbf{x}}_i}(k) - \mathbf{x}^*\|^2 + \frac{1}{2}\alpha_{i,k}^2 \sum_{i=1}^{nB} \|\overline{\tilde{\mathbf{x}}_i}(k-B) - \mathbf{x}^*\|^2$$
$$- \frac{4B}{k}(f(\bar{\mathbf{z}}(k)) - f(\mathbf{x}^*)) + \frac{8B}{k} G_f \sum_{i=1}^{n} \|\tilde{\mathbf{v}}_i(k) - \bar{\mathbf{v}}(k)\| \tag{44}$$
$$+ \frac{1}{2}\alpha_{i,k}^2 \sum_{i=1}^{nB} (\|R_i(k-1)\|^2 + \|R_i(k-B-1)\|^2) + nB_\tau \alpha_{i,k}^2 G_f^2$$

We now apply **Procedure A** with $a_k = 1 + C\alpha_{i,k}^2$ where $C > 0$ to give some degree of freedom of choosing $\{a_k\}$ with the required properties, then (44) implies

$$\mathbb{E}[\sum_{i=1}^{nB} \|\overline{\tilde{\mathbf{x}}_i}(k+1) - \mathbf{x}^*\|^2/\mathcal{F}_k] \leq (1 + C\alpha_{i,k}^2) \sum_{i=1}^{nB} \|\overline{\tilde{\mathbf{x}}_i}(k) - \mathbf{x}^*\|^2$$
$$- \frac{4B}{k}(f(\bar{\mathbf{z}}(k)) - f(\mathbf{x}^*)) + \frac{8B}{k} G_f \sum_{i=1}^{nB} \|\tilde{\mathbf{v}}_i(k) - \bar{\mathbf{v}}(k)\| \tag{45}$$
$$+ \frac{1}{2}\alpha_{i,k}^2 \sum_{i=1}^{nB} (\|R_i(k-1)\|^2 + \|R_i(k-B-1)\|^2) + nB_\tau \alpha_{i,k}^2 G_f^2$$

We define the following substitutions:

$$v_k = \sum_{i=1}^{nB} \|\overline{\tilde{\mathbf{x}}_i}(k) - \mathbf{x}^*\|^2, \; u_k = \frac{4B}{k}(f(\bar{\mathbf{z}}(k)) - f(\mathbf{x}^*)), \; -b_k = -1,$$
$$c_k = \frac{8B}{k} G_f \sum_{i=1}^{nB} \|\tilde{\mathbf{v}}_i(k) - \bar{\mathbf{v}}(k)\|$$
$$+ \frac{1}{2}\alpha_{i,k}^2 \sum_{i=1}^{nB} (\|R_i(k-1)\|^2 + \|R_i(k-B-1)\|^2) + nB_\tau \alpha_{i,k}^2 G_f^2$$

Then, to apply the supermartingale convergence theorem on (45) we require the substitutions described above along with $a_k = 1 + C\alpha_{i,k}^2$. But (45) is satisfied for the tail where (41) is satisfied. Therefore, for $k > k_2$, (45) is equivalent to the supermartingale inequality. Then having $a_k = 1 + e_k = 1 + C\alpha_{i,k}^2$, we get that $\sum_{k=0}^{\infty} e_k < \infty$ and $\sum_{k=0}^{\infty} c_k < \infty$ since $\sum_{k=0}^{\infty} \alpha_{i,k}^2 < \infty$, $\sum_{k=0}^{\infty} \frac{2}{k\gamma_i} G_f \sum_{i=1}^{n} \|\tilde{\mathbf{v}}_i(k) - \bar{\mathbf{v}}(k)\| < \infty$ by Lemma 4. Then the supermartingale convergence theorem holds.

And then the supermartingale theorem implies that the sequence $\{\|\overline{\mathbf{x}}_i(k) - \mathbf{x}^*\|\}$ is convergent a.s for all $i \in V$ and every $\mathbf{x}^* \in \mathcal{X}^*$ and also implies that $\sum_{k=0}^{\infty} \frac{4}{k}(f(\bar{\mathbf{z}}(k)) - f(\mathbf{x}^*)) < \infty$. This with the condition that, $\sum_{k=0}^{\infty} \frac{4}{k} = \infty$.

$$\lim_{k\to\infty} \inf(f(\bar{\mathbf{z}}(k)) - f(\mathbf{x}^*)) = 0 \quad a.s. \tag{46}$$

And since $f(\bar{\mathbf{z}}(k)) - f(\mathbf{x}^*) \geq 0$ for all k since $f(\mathbf{x}^*) = \min f(\mathbf{x})$ then $\lim_{k\to\infty} f(\bar{\mathbf{z}}(k)) = f(\mathbf{x}^*) \quad a.s.$ (47)

After elaborating more we arrive at $\lim_{k\to\infty} \tilde{\mathbf{x}}_i(k) = \mathbf{x}^*,$ (48)

for all $i \in V$ a.s. See [5] for further details.

V. Conclusion

We have considered a Distributed Asynchronous Random Projection Algorithm (DARPA) for solving a distributed constrained optimization problem over a time-varying multi-agent network. The algorithm convergence was analyzed under standard assumptions on the functions, such as Lipschitz continuity and convexity along with the random behavior of projections onto the constraint sets. The gradient estimation is considered under a stochastic setting with a bounded gradient variance. With the use of uncoordinated diminishing stepsizes and under the above assumptions along with the boundedness of the information exchange delay, we establish almost sure convergence of the method to the same optimal point. The future analysis intends to investigate the convergence rates of the algorithm as well as to further relax the assumptions on the gradient errors.

Appendix

A. Proposition 1

Proposition 1. *Assume the following inequality holds a.s. for all $k \geq \hat{k}$,*

$$v_{k+1} \leq a_{1,k} v_k + a_{2,k} v_{k-B} - b_k u_k + c_k \tag{49}$$

Then for all $k \geq \hat{k}$, $v_{k+1} \leq a_{1,k} v_k + a_{2,k} \max_{k-B\leq\hat{k}\leq k} v_{\hat{k}} - b_k u_k + c_k$ (50)

holds a.s.

Follows easily by considering the maximum of a set is greater that that of a subset.

B. Lemma 5

Lemma 5. *Assume the following inequality holds a.s. for all $k \geq k^*$,*

$$v_{k+1} \leq a_{1,k} v_k + a_{2,k} \max_{k-B\leq\hat{k}\leq k} v_{\hat{k}} - b_k u_k + c_k \tag{51}$$

v_k. u_k, b_k, c_k, $a_{1,k}$ and $a_{2,k}$ are non-negative random variables where $a_{1,k} + a_{2,k} \leq 1$. And $\{b_k\}$ and $\{c_k\}$ are increasing sequence and decreasing sequences, respectively, and $\{a_{1,k}\}$, $\{a_{2,k}\}$ are decreasing sequences and $c_k \leq b_k u_k$ for $k \geq \tilde{k}$. Then if for $\rho = (a_{1,1} + a_{2,1})^{\frac{1}{B+1}}$ and

$$v_{k_0} \leq \rho^{\Phi(k_0)} V_0' - b_0 u_0 + c_0 \quad a.s. \tag{52}$$

for base case $k = k_0 = \bar{k} - B$. (i.e., notice V_0 is not necessary the initial value v_0). And Φ is a random variable from $\mathbb{N}$ to $\mathbb{N}$ where $\Phi([n, m]) = [n, m]$.

And assume that this also holds for all $k \geq k_0$ up to $k = \bar{k}$ in an arbitrary manner (i.e., notice the power of ρ

is independent of k). i.e., $k \in \{k_0 = \bar{k} - B, \ldots, \bar{k}\}$ and $\bar{k} - B \geq \max(k^*, \tilde{k})$. That is

$$v_k \leq \rho^{\Phi(k)} V_0^{'} - b_{k-1} u_{k-1} + c_{k-1} \quad a.s. \tag{53}$$

for $k = \{k_0, \ldots, \bar{k}\}$.
Then we have

$$v_k \leq \rho^k V_0 - b_{k-1} u_{k-1} + c_{k-1} \quad a.s. \tag{54}$$

for all $k \geq \bar{k}$ where $V_0 > 0$ for all sequences patterns and ρ as before.

Proof: First since $a_{1,k} + a_{2,k} \leq 1$ then

$$\begin{aligned} & 1 \leq (a_{1,k} + a_{2,k})^{-\frac{B}{B+1}} \implies 1 \leq (a_{1,1} + a_{2,1})^{-\frac{B}{B+1}} \\ & \implies (a_{1,k} + a_{2,k}) \leq (a_{1,1} + a_{2,1}) \end{aligned} \tag{55}$$

which implies that

$$\begin{aligned} a_{1,k} + a_{2,k}\rho^{-B} &= a_{1,k} + a_{2,k}(a_{1,1} + a_{2,1})^{-\frac{B}{B+1}} \\ &\leq a_{1,k}(a_{1,1} + a_{2,1})^{-\frac{B}{B+1}} + a_{2,k}(a_{1,1} + a_{2,1})^{-\frac{B}{B+1}} \\ &= (a_{1,k} + a_{2,k})(a_{1,1} + a_{2,1})^{-\frac{B}{B+1}} \\ &\leq (a_{1,k} + a_{2,k})^{\frac{1}{B+1}} = \rho \end{aligned}$$

That is

$$a_{1,k} + a_{2,k}\rho^{-B} \leq \rho \tag{56}$$

Now, by induction we show that (54) for all $k \geq k_0$. Assume (52) is true for $k = k_0$ and that the induction hypothesis holds for all $k \geq k_0$ up to $\bar{k}$ where $k_0 = k - B \leq k \leq \bar{k}$. Then we have for any arbitrary behavior for k where $k_0 = k - B \leq k \leq \bar{k}$ that we can write the sequences v_k in a decreasing sequence. Without a loss of generality assume we will have for $0 \leq l \leq B$

$$\begin{aligned} v_{\bar{k}} &\leq \rho^{\bar{k}-l} V_0^{'} - b_{\bar{k}-1} u_{\bar{k}-1} + c_{\bar{k}-1} \\ v_{\bar{k}-B} &\leq \rho^{\Phi(\bar{k}-B)} V_0^{'} - b_{\bar{k}-B-1} u_{\bar{k}-B-1} + c_{\bar{k}-B-1} \end{aligned} \tag{57}$$

Then from (51) we have

$$\begin{aligned} v_{\bar{k}+1} &\leq a_{1,\bar{k}} v_{\bar{k}} + a_{2,\bar{k}} \max_{\bar{k}-B \leq \hat{k} \leq \bar{k}} v_{\hat{k}} - b_{\bar{k}} u_{\bar{k}} + c_{\bar{k}} \\ &\leq a_{1,\bar{k}} \rho^{\bar{k}-l} V_0^{'} + a_{2,\bar{k}} \rho^{\bar{k}-B} V_0^{'} - a_{1,\bar{k}} b_{\bar{k}-1} u_{\bar{k}-1} - a_{2,\bar{k}} b_{\bar{k}-B-1} u_{\bar{k}-B-1} \\ &\quad + a_{1,\bar{k}} c_{\bar{k}-1} + a_{2,\bar{k}} c_{\bar{k}-B-1} - b_{\bar{k}} u_{\bar{k}} + c_{\bar{k}} \end{aligned} \tag{58}$$

But $c_k \leq b_k u_k$ for all $k \geq \tilde{k}$ then

$$\begin{aligned} & -a_{1,\bar{k}} b_{\bar{k}-1} u_{\bar{k}-1} - a_{2,\bar{k}} b_{\bar{k}-B-1} u_{\bar{k}-B-1} + a_{1,\bar{k}} c_{\bar{k}-1} + a_{2,\bar{k}} c_{\bar{k}-B-1} = \\ & a_{1,\bar{k}}(c_{\bar{k}-1} - b_{\bar{k}-1} u_{\bar{k}-1}) + a_{2,\bar{k}}(c_{\bar{k}-B-1} - b_{\bar{k}-B-1} u_{\bar{k}-B-1}) \leq 0 \end{aligned} \tag{59}$$

Then

$$\begin{aligned} v_{\bar{k}+1} &\leq a_{1,\bar{k}} \rho^{\bar{k}-l} V_0 + a_{2,\bar{k}} \rho^{\bar{k}-B} V_0 - b_{\bar{k}} u_{\bar{k}} + c_{\bar{k}} \\ &\leq a_{1,\bar{k}} \rho^{\bar{k}-l} V_0^{'} + a_{2,\bar{k}} \rho^{\bar{k}-l-B} V_0^{'} - b_{\bar{k}} u_{\bar{k}} + c_{\bar{k}} \\ &= (a_{1,\bar{k}} + a_{2,\bar{k}} \rho^{-B}) \rho^{\bar{k}-l} V_0^{'} - b_{\bar{k}} u_{\bar{k}} + c_{\bar{k}} \\ &\leq \rho^{\bar{k}-l+1} V_0^{'} - b_{\bar{k}} u_{\bar{k}} + c_{\bar{k}} \quad a.s. \end{aligned} \tag{60}$$

But without a loss of generality, we can find $V_0 > 0$ such that $\rho^{\bar{k}-l+1} V_0^{'} \leq \rho^{\bar{k}+1} V_0$ to keep indexing tractable. And thus (60) is true for all $k \geq [k+1$. i.e., notice that for $k+1 = \bar{k}+2$, we already have for $k = \bar{k}+1$ that the power of ρ in the recursive inequality after the coefficient $a_{1,\bar{k}}$ is $\bar{k}+1$. Thus, no matter what the arbitrary behavior for the prior B terms is, we will have

$$v_{\bar{k}+2} \leq \rho^{\bar{k}+2} V_0^{'} - b_{\bar{k}+1} u_{\bar{k}+1} + c_{\bar{k}+1} \quad a.s. \tag{61}$$

Thus, (60) follows for all $k \geq \bar{k}$. ■

Remark 4. *i.e., notice that it is true for $k = \bar{k}$ since*

$$v_{\bar{k}+1} \leq \rho^{\bar{k}-l} V_0^{'} - b_{\bar{k}} u_{\bar{k}} + c_{\bar{k}} \text{ and } v_{\bar{k}+1} \leq \rho^{\bar{k}} V_0 - b_{\bar{k}} u_{\bar{k}} + c_{\bar{k}} \tag{62}$$

C. Proposition 2

Proposition 2. *If $v_k \leq \rho^k v_0 \quad a.s.$ for all $k > \bar{k}$ where $\rho < 1$ and v_k is non-negative, then $\{v_k\}$ is eventually a decreasing sequence a.s. That is there exists $\hat{k}$ such that for all $k > \hat{k} \geq \bar{k}$ we have $v_{k+1} < v_k \quad a.s.$*

Proof: We have $v_k \leq \rho^k v_0 \quad a.s.$ for all $k > \bar{k}$ then

$$\lim_{k \to \infty} v_k \leq \lim_{k \to \infty} \rho^k v_0 = 0 \tag{63}$$

But $v_k > 0$ then $\lim_{k \to \infty} v_k = 0$.

Suppose there exists $k > \bar{k}$ such that $v_{k+1} > v_k$. But $v_k \leq \rho^k v_0$ and $v_k > m > 0$ since $v_k > 0$ (i.e., a non-negative number). Then there esists $q \in \mathbb{N}$ $v_k < v_{k+q} \leq \rho^{k+q} v_0 < m$ since $\rho < 1$ and $\rho^k > \rho^{k'}$ if $k < k'$, then $v_k < v_{k+q} < m < v_k$ a contradiction. That is, for $k > \bar{k}$ where $v_{k+1} > v_k$ we have a limiting number $q \in \mathbb{N}$ where $v_{k+q} < v_k$. Thus, any increasing sequence pattern can last only finite terms. But $\lim_{k \to \infty} v_k = 0$, then its tail of infinite terms can not be increasing sequence because it is an infinite sequence. And it can not be constant since initial value $v_0 > 0$ and final value $\lim_{k \to \infty} v_k = 0$. Then it must be a decreasing subsequence. That is, there exists $\hat{k}$ such that for all $l_k > \hat{k}$ we have $v_{lk+1} < v_{lk} \quad a.s.$ Thus, without a loss of generality and assuming new indexing we have $\hat{k}$ such that for all $k > \hat{k}$ we have $v_{k+1} < v_k \quad a.s.$ where k is indexing the subsequence of the original sequence which can be seen as a sequence with this new indexing. ■

D. Proposition 3

Proposition 3. *If for all $k > \hat{k}$ we have $v_{k+1} < v_k$ and*

$$v_k \leq \rho^{l(k)} v_0 \tag{64}$$

where $v_k > 0$ and $\rho < 1$ and $l(k+1) > l(k)$. Then we can find a decreasing sequence $\{a_k\}$, that is, $a_{k+1} < a_k$ where $a_k > 1$ and

$$v_k \leq \rho^{l(k)} v_0 \leq a_{k-1} v_{k-1} \text{ for all } k > \hat{k} \tag{65}$$

Proof: We have for $k > \hat{k}$ that $v_{k+1} < v_k$, but

$$v_k \leq \rho^{l(k)} v_0 \text{ and } v_{k+1} \leq \rho^{l(k+1)} v_0$$

and $\rho^{l(k+1)} v_0 < \rho^{l(k)} v_0$ since $\rho < 1$ where $l(k+1) > l(k)$. Then we distinguish two cases:

$$\text{Either } v_k \leq \rho^{l(k)} v_0 \leq v_{k-1} \leq \rho^{l(k-1)} v_0 \tag{66}$$

$$\text{or } v_k \leq v_{k-1} \leq \rho^{l(k)} v_0 \leq \rho^{l(k-1)} v_0 \tag{67}$$

For case of (66) we have $v_k \leq \rho^{l(k)} v_0 \leq v_{k-1} \leq a_{k-1} v_{k-1}$ where $a_{k-1} > 1$. Then for this case we have

$$v_k \leq a_{k-1} v_{k-1}$$

is satisfied as long as $a_{k-1} > 1$. Then we can choose

$$a_{k-1} < a_{k-2} \text{ and } v_k \leq \rho^{l(k)} v_0 \leq a_{k-1} v_{k-1} \text{ as long as } a_{k-1} > 1 \tag{68}$$

For case of (67), we need to choose $a_{k-1} > 1$ such that

$$v_k \leq \rho^{l(k)} v_0 \leq a_{k-1} v_{k-1}.$$

But for this case we have $v_{k-1} \leq \rho^{l(k)} v_0$ then we can choose $a_{k-1} > 1$ such that

$$v_{k-1} \leq \rho^{l(k)} v_0 \leq a_{k-1} v_{k-1},$$

as long as $$\frac{\rho^{l(k)}}{a_{k-1}} \leq \frac{v_{k-1}}{v_0} \leq \frac{\rho^{l(k)} v_0}{v_0} \leq \rho^{l(k)} \tag{69}$$

But (69) is true for any $a_{k-1} > 1$. Then we can choose for this case $a_{k-1} < a_{k-2}$ such that

$$v_k \leq \rho^{l(k)} v_0 \leq a_{k-1} v_{k-1}.$$

is satisfied as long as $a_{k-1} > 1$. So for both complementary cases we can choose a decreasing sequence $\{a_k\}$, that is, $a_{k+1} < a_k$ such that for all $k > \hat{k}$ we have

$$v_k \leq \rho^{l(k)} v_0 \leq a_{k-1} v_{k-1}$$ ■

E. The bridge: Procedure A

1) Intuition and main theoretical contribution: We prove the convergence of DARPA using the same building lemmas as [6]. However, to accomodate for the delay we introduce in the proof an extended system without delay which is adapted from [9] with some modifications that makes it fit our problem. By performing that we introduce new variables that take into consideration all B instants and apply the analysis of [6] on these new variables. Meanwhile, due to the relaxed requirement on the projection sets, where we don't require that these sets be bounded, we are faced in Lemmas 1 and 4 by a form of inequalities with two priors that are not compatible with the supermartingale inequality. However, by applying Procedure A we are able to reduce these inequalities to a supermartingale inequality where we can immitate again [6] with slight modifications that fit our case to arrive at the required result.

We begin with the application of Procedure A on Lemma 1 in Part 1 of Section IV-C. We apply **Procedure A** where: Input Inequality: (31). Output Inequality: (32). Index: k_1

Procedure A: (where initial input inequality is (31))
We apply our analysis for $k > \tilde{k} \geq \max(2B-1, k_*)$ so that the extended variables $\tilde{\mathbf{v}}_i$ and $\tilde{\mathbf{x}}_i$ have all their entries taking a value through the algorithm process. Take $k^* = \max(k_1, \tilde{k}^*)$, then (31) is satisfied a.s. Then using Proposition 1 and having inequality (31) of the form in (49) then the inequality equivalent to the form (50) is satisfied for all $k > k^*$.

Then we can choose $k_0 = \bar{k} - B > \max(k^*, \tilde{k})$ as in Propositions 1-3 and Lemma 5 in Appendix. But we have for $k_0 = \bar{k} - B \leq k \leq \bar{k}$ that (31) satisfied. Then from (31), the following inequality follows

$$v_k \leq \rho \max(v_{k-1}, v_{k-B-1}) - b_{k-1} u_{k-1} + c_{k-1} \tag{70}$$

However, for arbitrary k_0 satisfying the above we have the corresponding term v_k of (31) behaving arbitrary for $k_0 = \bar{k} - B \leq k \leq \bar{k}$. Then we can choose a specific $V_0^{'} \in \mathbb{R}$ and $\rho = (a_{1,1} + a_{2,1})^{\frac{1}{B+1}} < 1$ for this arbitrary random behavior where for $k_0 = \bar{k} - B \leq k \leq \bar{k}$ we can have

$$\begin{aligned} \gamma_k = a_{1,k-1} v_{k-1} + a_{2,k-1} v_{k-B-1} &\leq \rho \max(v_{k-1}, v_{k-B-1}) \\ &\leq \rho^{\Phi(k)} V_0^{'} \end{aligned} \tag{71}$$

where Φ is a random variable from $\mathbb{N}$ to $\mathbb{N}$ such that $\Phi([n,m]) = [n,m]$. And (31) becomes

$$v_k \leq \gamma_k - b_{k-1} u_{k-1} + c_{k-1} \tag{72}$$

for $k_0 = \bar{k} - B \leq k \leq \bar{k}$.

Remark 5. *We can allow the above condition to be satisfied for any arbitrary behavior.*

Then applying Lemma 5 with the same identification of $\tilde{k}$, k^*, k_0 and $\bar{k}$, then we have for all $k > \bar{k} > \max(k^*, \tilde{k})$ that

$$v_k \leq \rho^k V_0 - b_{k-1} u_{k-1} + c_{k-1} \tag{73}$$

with the identifications assumed by (31) where $c_k < b_k u_k$. Then we have that

$$v_k \leq \rho^k V_0 \tag{74}$$

for $k > \bar{k}$. Then by applying Proposition 2, there exists (i.e., under the new subsequence indexing) $\hat{k} \geq \bar{k}$ such that

$$v_{k+1} \leq v_k \quad a.s. \tag{75}$$

Thus, we have for all $k > \hat{k} \geq \bar{k}$ that $v_{k+1} < v_k$ and $v_k \leq \rho^k V_0$ where $v_k > 0$ and $\rho < 1$

Then by applying Proposition 3 we can find a decreasing sequence $\{a_k\}$ where $a_k > 1$ such that

$$v_k \leq \rho^k V_0 \leq a_{k-1} v_{k-1} \tag{76}$$

for all $k > \hat{k} > \bar{k} > \max(k^*, \tilde{k})$. Then we arrive at the supermartingale inequality (32), that is for all $k > \hat{k} \geq \bar{k}$ we have (32) with the form equivalent to

$$v_k \leq a_{k-1} v_{k-1} - b_{k-1} u_{k-1} + c_{k-1} \tag{77}$$

and the corresponding identifications assumed in (31).

End of Procedure A (i.e. for Part 1)

Similarly, we apply Procedure A on Lemma 4 in Part 3 of Section IV-C. Then using **Procedure A** where: Input Inequality: (44). Output Inequality: (45). Index: k_2

That is, by following the analysis in part 3 up to (44) and using (44) in place of (31) in **Procedure A** and following the same steps we followed in Part 1 for (31) we arrive at an equivalent supermartingale inequality (45) instead of (32).

REFERENCES

[1] Bertsekas, D. P. and J. N. Tsitsiklis (2000). Gradient convergence in gradient methods with errors. *SIAM Journal on Optimization 10*(3), 627–642.

[2] Borkar, V. S. (2009). *Stochastic approximation: a dynamical systems viewpoint*, Volume 48. Springer.

[3] Burke, J. and M. C. Ferris (1993). Weak sharp minima in mathematical programming. *SIAM Journal on Control and Optimization 31*(5), 1340–1359.

[4] Gubin, L., B. Polyak, and E. Raik (1967). The method of projections for finding the common point of convex sets. *USSR Computational Mathematics and Mathematical Physics 7*(6), 1–24.

[5] Lee, S. and A. Nedic (2013). Distributed random projection algorithm for convex optimization. *IEEE Journal of Selected Topics in Signal Processing 7*(2), 221–229.

[6] Lee, S. and A. Nedić (2015). Asynchronous gossip-based random projection algorithms over networks. *IEEE Transactions on Automatic Control 61*(4), 953–968.

[7] Nedic, A. (2010). Asynchronous broadcast-based convex optimization over a network. *IEEE Transactions on Automatic Control 56*(6), 1337–1351.

[8] Nedić, A. (2011). Random algorithms for convex minimization problems. *Mathematical programming 129*(2), 225–253.

[9] Nedić, A. and A. Ozdaglar (2010). Convergence rate for consensus with delays. *Journal of Global Optimization 47*(3), 437–456.

[10] Polyak, B. T. (1987). Introduction to optimization. translations series in mathematics and engineering. *Optimization Software*.

[11] Ram, S. S., A. Nedić, and V. V. Veeravalli (2009). Asynchronous gossip algorithms for stochastic optimization. In *Proceedings of the 48h IEEE Conference on Decision and Control (CDC) held jointly with 2009 28th Chinese Control Conference*, pp. 3581–3586. IEEE.

[12] Ram, S. S., A. Nedić, and V. V. Veeravalli (2010). Distributed stochastic subgradient projection algorithms for convex optimization. *Journal of optimization theory and applications 147*(3), 516–545.

[13] Robbins, H. (2012). *Selected papers*. Springer.