Straggler-Robust Distributed Optimization with the Parameter Server Utilizing Coded Gradient

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

Optimization in distributed networks plays a central role in almost all distributed machine learning problems. In principle, the use of distributed task allocation has reduced the computational time, allowing better response rates and higher data reliability. However, for these computational algorithms to run effectively in complex distributed systems, the algorithms ought to compensate for communication asynchrony, and network node failures and delays known as stragglers. These issues can change the effective connection topology of the network, which may vary through time, thus hindering the optimization process. In this paper, we propose a new distributed unconstrained optimization algorithm for minimizing a strongly convex function which is adaptable to a parameter server network. In particular, the network worker nodes solve their local optimization problems, allowing the computation of their local coded gradients, and send them to different server nodes. Then each server node aggregates its communicated local gradients, allowing convergence to the desired optimizer. This algorithm is robust to network worker node failures or disconnection, or delays known as stragglers. One way to overcome the straggler problem is to allow coding over the network. We further extend this coding framework to enhance the convergence of the proposed algorithm under such varying network topologies. Finally, we implement the proposed scheme in MATLAB and provide comparative results demonstrating the effectiveness of the proposed framework.

Index Terms

distributed optimization, gradient coding, synchronous, centralized networks

I. INTRODUCTION

Many problems in distributed systems over the cloud, or in wireless ad hoc networks [6,13,14], are formulated as convex optimization programs in a parallel computing scheme. Depending on the structure of these networks, i.e., centralized, decentralized or fully distributed, the optimization techniques are adapted to accommodate such structures. However, the malfunctioning of processors directly impacts the overall performance of parallel computing. Dealing with this malfunctioning is referred to as the straggling problem. Many applications, whether over the cloud

* This material is based upon work supported by the National Science Foundation under grants CCF-1718195.
or in local distributed networks, have experienced considerable time delays, due in part to this straggling problem. Asynchronous \cite{10, 5} and synchronous algorithms \cite{17, 2} have been proposed to overcome this problem. While Lee et al. \cite{7} and Dutta et al. \cite{4} describe techniques for mitigating stragglers in different applications, a recent work by Tandon et al. \cite{16} focused on codes for recovering the batch gradient of a loss function (i.e., synchronous gradient descent). Specifically, a coding scheme in \cite{16} was proposed, enabling a distributed division of tasks into uncoded (naive) and coded parts. This partition alleviates the effect of straggling servers in a trade-off among computational, communication complexity and time delay. This novel coding scheme solves this problem by providing robustness to partial failure or delay of nodes.

II. PROBLEM SETUP

We consider a network of \(n\) server nodes indexed by \(V = \{1, 2, \ldots, n\}\) and \(m\) worker nodes on a parameter server platform using a multi-bus multiprocessor system with shared memory. The objective is to solve a minimization problem where the solution set \(X^*\) belongs to a convex set \(X\) where the gradient is bounded. To that end we require that the global function \(f\) is divided into \(p\) partitions with arbitrary number of replication for each. Thus, we require arbitrary interleaved connections according to availability. Meanwhile, in Figure 1 due to the complexity of the system schematic we show a restrictive setup where each server is connected to its own partition all the time where the redundancy of each partition is unity. We show in Figure 2 a more elaborate schematic for the general case when discussing the push/pull steps for one server.

The optimization problem is the unconstrained optimization given by

\[
\hat{x} = \arg \min_{x \in \mathbb{R}^n} f(x) = \sum_{i=1}^{n} f^{(i)}(x)
\]  

Due to the random behavior of this distributed system we adapt a similar approach to analyze the unconstrained Problem \ref{1} as in \cite{8}. Lee and Nedić. In their approach the authors solve a constrained distributed optimization problem based on stochastic gradient descent (i.e., partial gradients) and random projections

\[
\min_{x \in \bigcap_i X_i} f(x) = \sum_{i=1}^{n} f^{(i)}(x)
\]  

on \(n\) nodes, where the local optimization problems are carried on the nodes themselves and \(X_i\) are convex sets such that \(\bigcap_i X_i\) is the constraint set. We note that we use a similar approach as their random approach which employs the supermartingale theorem due to the structure of our distributed parameter servers and the randomness of connections along the multi-bus between servers and workers. This analysis also utilizes the supermartingale theorem, however the projection is on the set \(X\) of bounded gradients which has nothing in common with any projection step in the algorithm and is their only for the mathematical analysis.

In this work to solve the unconstrained distributed optimization problem \ref{1} we use a gradient descent method without projections through utilizing Straggler-Robust Distributed Optimization (SRDO) Algorithm. As its name infers our algorithm has the extra feature of being robust to stragglers. To this end, coded local gradients \(\nabla g^{(i)}_j\)
Fig. 1. Parameter server network in the special case with \( n \) server nodes and \( m \) workers nodes. The worker nodes are divided into \( p = n \) partitions where partition \( i \) has \( n_i \) workers and each server is connected to its unique partition at all iterations. (i.e., local coded gradient of worker \( j \) in partition \( i \)) are carried on the different partitions of worker nodes and decoded to the partition's gradient \( \nabla f^{(i)} \) (i.e., the gradient of the function of the partition connected to server \( i \), cf. Remark 1) on the connected server nodes (cf. Fig.1).

Meanwhile, after dividing the load into different partitions each partition \( i \in \{1, \ldots, p\} \) is distributed with an arbitrary redundancy among the workers. Thus, each partition replica utilizes a gradient coding similar to that of \([16]\) to enable robustness to an allowed number of stragglers per each partition. More specifically, under a global clock each server sends its estimate \( v_i(k) \) to possibly any worker under the interleaved shared memory multi-bus system. Each worker then calculates its coded gradient relative to the partition replica it belongs to and using the estimate it received from a server. Then under a global clock, the servers synchronously call for different partitions and decode their local functions gradients \( \nabla f^{(i)} \) to compute their estimates \( x_i(k) \).

Thus, different servers are working in a synchronous manner to compute their solution \( x_i(k) \) through a gradient descent step decoded from partial coded gradients of the connected partition replica. However, each partition worker calculates its coded gradient by evaluating the coded load on the weighted averages \( v_q(k - k'_q) \) which can be from different servers and different time step with bounded delay, (i.e., \( q \in \{1, \ldots, n\} \) and \( 0 \leq k'_q \leq H \).

Next, we briefly outline the steps of our algorithm implemented at the server and worker nodes; and elaborate on these steps in Section [III] After the distribution of the load in the distribution step accordingly.

On the servers side:

**Push step:** Under a global clock each server \( i \) sends a message containing the weighted average \( v_i(k) \) to an arbitrary number of workers (i.e., that could be in different partitions). Then each worker starts computing its local
Algorithm 1 Algorithm Updating at Each Server SRDO Algorithm

Given: $f(x) = \sum_{i=1}^p f^{(i)}(x)$

e.g. Least Squares Prototype: $A \in \mathbb{R}^{M \times N}$, $y \in \mathbb{R}^M$ where $f(x) = \frac{1}{2}\|y - Ax\|_2^2$.

Find: $x^* \in \mathbb{R}^N$.

Initialization: Each server $i$ sets $v_i(0)$ to an arbitrary random vector.

Set: $t_0, \epsilon_i = \text{large number}$.

while $\epsilon_i > t_0$

At each server $i$:

Push step of iteration $k$: &

Send $v_i(k)$ to & arbitrary workers

Pull step at iteration $k$: 

Decoding the partition gradient $\hat{\nabla} f^{(i)}(k)$ from the sent local coded gradients $\nabla g^{(i)}_j(k)$

Iteration $k + 1$:

$x_i(k+1) = v_i(k) - \alpha_k \hat{\nabla} f^{(i)}(k)$

$v_i(k+1) = \sum_{j=1}^n w_{ij}(k+1)x_j(k+1)$.

$\epsilon_i = \|v_i(k+1) - v_i(k)\|_2$

end

Output: $x^* = x_i(k)$.

coded gradient in the worker computation of coded gradient step.

Pull step: Under a global clock each server gets activated and calls for coded gradients from an arbitrary partition. The workers from that partition $\iota$ will send the coded gradients $\nabla g^{(i)}_j$ to the connected server $i$. (i.e., some workers are stragglers thus don’t send their coded gradients). Server $i$ after receiving the coded gradients from the connected partition $\iota$ decodes the partition gradient $\nabla f^{(i)}$.

Remark 1. Here, we identify the partition $\iota$ at this connection instant with server $i$.

Then the server calculates the estimate $x_i(k+1)$.

Consensus step: Under a global clock each server gets activated again and computes its weighted average $v_i(k+1)$ from its connected servers $x_j(k+1)$ according to (3). Then the algorithm state goes back to the push step and continues henceforth until convergence.

On the workers side:
Worker Computation of Coded Gradient: When a worker $i$ of a partition $\iota$ receives a weighted average $v_{j}(k)$ from a server $j$ it gets activated and starts calculating its coded gradient $\nabla g_{\iota}(i)$ relative to the coding scheme used on partition $\iota$. Workers can get delayed in their computation of coded gradients and need to send their computed gradients to a connected server to their partition at a subsequent time instant. Here, there is one aspect of asynchronous behavior in the algorithm that influence the computed partition gradient used at the connected server. That is, we don’t require that the coded gradients of which the partition’s (or connected server) partial gradient is decoded to be of the consecutive previous instant weighted averages evaluations but rather of possibly older weighted averages evaluations.

This approach is tolerant to the allowed number of stragglers, and it is also robust to more than the allowed number of stragglers.

We prove convergence in the general case where $n$, $m$ and $p$ are arbitrary.

A. Assumptions on the Convex Functions

**Assumption 1.** Let the following conditions be satisfied:

(a) Each function $f^{(i)} : \mathbb{R}^N \rightarrow \mathbb{R}$ is convex.

(b) The functions $f^{(i)}$, $i \in 1, 2, \ldots, n$, are differentiable and have Lipschitz gradients with a constant $L$ over $\mathbb{R}^N$,

$$\|\nabla f^{(i)}(x) - \nabla f^{(i)}(y)\| \leq L \|x - y\|$$

for all $x, y \in \mathbb{R}^N$.

(c) The gradients $\nabla f^{(i)}(x)$, where $i \in V$ are bounded over the set $\mathcal{X}$ where $\mathcal{X} = \{x | x = \text{arg min } f(x)\} \subset \mathcal{X}$; i.e., there exists a constant $G_f$ such that $\|\nabla f^{(i)}(x)\| \leq G_f$ for all $x \in \mathcal{X}$ and all $i \in V$.

The assumed structure on $f$ is typical for problems of this kind and enables a detailed convergence analysis. Next, we make the following assumptions about the server-server edge weights in the consensus step.

**Assumption 2.** [Row Stochastic] For all $k \geq 0$, we have:

- The matrices $w_{ij}(k)$ in (3c) are chosen such that $w_{ij} = W(k)$ depending on the network server-connection topology in a way that allows consensus.

- $\sum_{j=1}^{n}[W(k)]_{ij} = 1$ for all $i \in V$.

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

- $\sum_{i=1}^{n}[W(k)]_{ij} \leq 1 - \mu$ for all $j \in V$ and $0 < \mu < 1$.

- If server $i$ is disconnected from server $j$ at instant $k$, then $[W(k)]_{ij} = 0$.

**Remark 2.** Notice that for the matrices $W(k)$ we have for $b_i = \sum_{j=1}^{n}[W(k)]_{ij}a_j$ that

$$\sum_{i=1}^{n}\|b_i\| \leq \sum_{i=1}^{n}\sum_{j=1}^{n}[W(k)]_{ij}\|a_j\| \leq (1 - \mu)\sum_{j=1}^{n}\|a_j\| \leq \sum_{j=1}^{n}\|a_j\|$$

and we are going to use either inequality as needed in our analysis.

**Assumption 3.** Bounded Delayed Evaluation and Gradient Computation:
We assume the decoding of gradient $\nabla f^{(i)}(k)$ at time $k$ decoded from coded gradients evaluated of weighted averages $v_i(k - k')$ where $0 \leq k' \leq H$. We assume the use of stale gradients in gradient computation scenario 3 of weighted average evaluations of instants $k - k''$ where $0 \leq k'' \leq \kappa$ to be more explicit, i.e., that are of the global iteration $k - \bar{k}$ where $\bar{k} \leq k'' \leq \bar{k} + H$. Without a loss of generality, we assume $H = \kappa$.

**Assumption 4. Choice of Partition by Server in Pull Step**

Each server $i$ gets connected to a partition ($i$) out of the $p$ partitions with a probability $\gamma(i)$ and gets no connection with any partition with a probability $\gamma(0)$.

**Assumption 5. Diminishing Coordinated Synchronized Stepsizes**

The stepsizes $\alpha_{i,k}$ of server $i$ are coordinated and synchronized where $\alpha_{i,k} = \alpha_k > 0$ and $\alpha_k \to 0, \sum_{k=0}^{\infty} \alpha_k = \infty, \sum_{k=0}^{\infty} \alpha_k^2 < \infty$. For example, a unanimous stepsize $\alpha_{i,k} = \alpha_k = \frac{1}{k+1}$ among all servers $i$ per iteration $k$.

**III. MAIN ALGORITHM: STRAGGLERS ROBUST DISTRIBUTED OPTIMIZATION ALGORITHM (SRDO)**

In solving problem (1), we propose a synchronous iterative gradient descent method. This method is robust to an allowed number of stragglers, and is also valid for a varying network topology with more than the allowed number of stragglers as we are going to show.

An appropriate implementable platform for the algorithm is a multi-bus distributed parameter server shared memory network. The network is equipped with a universal clock that synchronizes the actions of its nodes.

**Initialization:** Each server node $i$ at global iteration $k = 0$ begins with random weighted average $v_i(0) \in \mathbb{R}^N$ for $i \in \{1,2,\ldots,n\}$, and sends $v_i(0)$ to arbitrary number of worker nodes in the push step, as shown in Fig. 2. Each worker node $j$ of a partition $u$ then solves for its partition gradient in a coded manner using its received $v_i(0)$, finding a coded local gradient $(\nabla g_j^{(u)})^T = B_j^{(u)} \nabla f^{(u)}$ evaluated at $v_i(0)$ where $f^{(u)}$ corresponds to the partition $u$ function such that $f^{(u)} = \sum_{i=1}^{n_i} f_i^{(u)}$ and $\nabla f^{(u)} = [\nabla f_1^{(u)}]^T (\nabla f_2^{(u)})^T \ldots (\nabla f_{n_i}^{(u)})^T]^T$.

Then in the pull step of iteration $k$, each server node $i$ according to the global clock gets activated and aggregates those local coded gradients received from its partition $i$ worker nodes and finds the partition gradient $\nabla f^{(i)}$ (i.e., notice that we identified partition $i$ with server $i$ in the pull step, as shown in Fig. 2, i.e., decoding the partition’s gradient in the respective network topology scenario, i.e., related to the number of stragglers, and according to the used gradient computation scenario). Afterward, the algorithm adapts the values $v_i(0)$ at the server nodes by that connected partition computed gradient if it exists, and finds the estimates $x_i(1)$ according to (3b). It then forms the weighted averages $v_i(1)$ according to (3b), and sends them to the worker nodes, as the cycle continues henceforth until the algorithm converges to the optimizer. We prove the convergence in Section VII.

**Remark 3.** Each worker node $j$ of partition $i$ uses the estimate $v_i(k)$ and finds the coded local gradient $\nabla g_j^{(i)}(k)(v_i(k))$ corresponding to its local optimization at $v_i(k)$; i.e., here, $g_j^{(i)}(k)$ is a function employed due to the coding scheme, and $\nabla g_j^{(i)}(k)$ corresponds to $g_j^{(i)}(k) = \sum_{q=1}^{n_i} B_{j,q}^{(i)} f_q^{(i)}(k)$. The function $f_q^{(i)}$ corresponds to $q \in \{1,2,\ldots,n_i\}$, where $\sum_{q=1}^{n_i} f_q^{(i)} = f^{(i)}$, and $n_i$ is the number of worker nodes in partition $i$. In the pull step, each server node uses the received coded local gradients that are employed at probably different estimates $v_i(k - k')$ of different
time instants $k' \leq \kappa$, and tries to decode the partition gradient of the function $f^{(i)}$ by using a different coded scheme according to a specific partition sub-partition, (i.e., which need not to be unanimous to all partitions). We define the set of connected nodes at iteration $k$ to server node $i$ as $\Gamma_i(k)$, and thus, the set of stragglers to node $i$ as $\Gamma_i^\ominus(k) \triangleq \{1, 2, \ldots, n_i\} \setminus \Gamma_i(k)$.

**Remark 4.** It is worth noting the following about the synchronous behavior of SRDO:

- If it happens that a server node $j$ still didn’t receive the coded gradients of any of the partitions and is unable to decode the partition’s partial gradient and calculate its estimate $x_j(k+1) = \mathbf{v}_j$.
- If at the push step of iteration $k$ a worker is unable to receive any of the weighted averages $\mathbf{v}_q(k - k'_q)$ for all $0 \leq k'_q \leq H$, where $q \in \{1, \ldots, n\}$, that allows it to compute its local coded gradient in time before the pull step for the same instant $k$, then that worker is considered a straggler.
- Each worker has to finish its computation before interacting with another server. In that respect, gradients evaluated at previous weighted averages can still be used by a server as long as the worker would send its coded gradient when it finishes computation at the time of the synchronous update. Moreover, a similar scenario is when these coded gradients from prior are kept in memory of a server and are used in the subsequent updates. The benefit of the latter scenario on the prior one is that it mitigates the effect of stragglers at the instant of update.

A. **Remark on the Computation of the Gradient under Different Scenarios**

We distinguish three different scenarios for different partition’s gradient computation scenarios:

**Scheme 1:**
In this scheme, the number of all partition’s worker nodes disconnected to their respective server node $i$ at the pull step is less than or equal to the maximum allowed number of stragglers (i.e., $|\Gamma_i^\ominus| \leq s$). And the server decodes the partition’s inexact gradient $\nabla f^{(i)}$ by the brute application of the described coding scheme.

**Scheme 2:**
In this scheme, the number of all partition’s worker nodes disconnected (i.e., fail or get delayed) from their respective server node $i$ at the pull step is greater than the maximum allowed number of stragglers, (i.e., $|\Gamma_i^\ominus| > s$). Server node $i$ uses only the received coded local gradients from its connection set $\Gamma_i(k)$ to compute the partition’s inexact gradient $\nabla f^{(i)}$.

**Scheme 3:**
In this scheme, the number of all worker nodes disconnected (i.e., fail or get delayed) from their respective server node $i$ at the pull step is greater than the maximum allowed number of stragglers, (i.e., $|\Gamma_i^\ominus| > s$). Server node $i$ uses the received local gradients at instant $k$ from its connection set $\Gamma_i(k)$ of the connected worker nodes of partition $i$ identified with server $i$ or stale delayed coded local gradients to compute the partition’s inexact gradient $\nabla f^{(i)}$, (see Remark 3 for different scenarios of this scheme).
Hence, in SRDO step (3b), the iterate $x_i(k+1)$ employing the partition’s decoded gradient is calculated by the server node. i.e.,

$$x_i(k+1) = v_i(k) - \alpha_k \nabla f^{(i)}(k)$$

where $\nabla f^{(i)}(k) = \sum_{j \in \Gamma(k)} A^{(i)}_{f_{it},j} \nabla g^{(i)}_j (v_q(k-k_q'))$ where $0 \leq k_q' \leq H$ and $q \in V = \{1, \ldots, n\}$.

The server node aggregates the weighted local coded gradients at different estimates $v_q(k)$ available on each connected partition’s worker node, to decode the gradient of the partition.

This decoded partition’s gradient utilized in the updating step of the partition’s connected server in its more general form compatible with the aforementioned three schemes can be further written in a more reduced form corresponding to each index $i \in \{1, \ldots, n\}$ as:

$$\hat{\nabla f^{(i)}}(k) = \nabla f^{(i)}(v_i(k)) + \sum_{j \in \Gamma_i(k)} A^{(i)}_{f_{it},j} \sum_{l=1}^{n_i} B^{(i)}_{j,l} (\nabla f_l^{(i)}(v_q(k-k_q')) - \nabla f_l^{(i)}(v_i(k))).$$

Where here the server node $i$ receives coded gradients in the pull step from a worker node $l$ which previously received an arbitrary estimate $v_q$ in the prior consecutive push step. Here, we seek arbitrary server to worker connections in both the pull and push steps. That is, servers are not only connected to the same workers partition in both the pull and the push steps as then in that case an exact but different partition’s gradients are decoded on different servers.

IV. Background Material

A. Gradient Coding Scheme

As previously discussed, when solving Problem (1) using a distributed synchronous gradient descent worker nodes may be stragglers [3, 5, 10]; i.e., the nodes fail or get delayed significantly in computing or communicating the gradient vector to the server node. Tandon et al. [16] proposed replicating some data across machines in a defined coding scheme to address this problem. This scheme allows the recovery of the overall gradient by the aggregation of the computed local gradients from the connected nodes active in the network at iteration $k$.

Specifically, in [16], the authors find a lower bound on the structure of the coding partition scheme that allows the computation of the overall gradient in the presence of any $s$ or fewer stragglers, (i.e., if we have fewer stragglers than the maximum allowed, we can use any $n - s$ combination of the connected nodes). This bound is on the minimum number of replicas for each partition $J_i$ forming the local function $f^{(i)}$ such that $f = \sum_{i=1}^d f^{(i)}$ which should be at least replicated $s + 1$ times across all machines. A coding scheme robust to any $s$ stragglers or less, corresponding to $n$ nodes and $m$ data partitions. Without loss of generality, we assume the number of nodes $n$ is equal to the number of partitions $m$ in the algorithm.

Therefore, to employ coding and decoding of the overall gradient in the case of the allowed number of stragglers not to exceed $s$, we require a scheme in which [16]:

$$B \in \mathbb{R}^{n \times d}, \quad A \in \mathbb{R}^{(s) \times n} \quad \text{and} \quad AB = 1^{(s) \times d},$$

(7)
where the decoding matrix $A$ and the encoding matrix $B$ can be calculated from Algorithm 1 and Algorithm 2 in [16], respectively.

We exploit the above coded scheme to compute $\nabla f^{(i)}$, where $\nabla f = \sum_{i=1}^{n} \nabla f^{(i)}$. Thus, we apply this coding scheme to compute $\nabla f^{(i)} = \sum_{l=1}^{n_i} \eta_l^{(i)}$ (i.e., in the coding scheme, the number of data partitions $d$ is without loss of generality equal to the number of nodes $n$; therefore, in our partitions $d = n_i$, the number of nodes of partition $i$). See Remark 3 in Section III.

Fig. 2. Parameter server schematic with coding scheme 1 of allowed number of stragglers on partition 1 of 3 worker nodes connecting to sever 2. Notice that in this general scenario, the $\hat{\nabla} f_2$ that is calculated at server node 2 is identified with partition 1 gradient $\hat{\nabla} f^{(1)}$. 

\[ v_2(k) = \sum_{l=1}^{2} [W]_2x_1(k) \]
\[ x_2(k + 1) = v_2(k) - \alpha_\delta \nabla f^{(1)}(k) \]
\[ \nabla f^{(1)}(k) = 2\nabla g_1^{(1)}(v_2(k)) - \nabla g_2^{(1)}(v_1(k)) \]
V. Theoretical Analysis Background

A. Specific Relations and Theorems

1) Convexity of functions: We refer to [11] for the definition of convex functions and the properties concerning nonexpansiveness of projections and the definition of minimum distance between a point and a closed set.

2) Supermartingale Convergence Result: In our analysis as in paper [8], we also make use of the supermartingale convergence result due to Robbins and Siegmund (see Lemma 10-11, p. 49-50 [12] or original paper [15], p. 111-135)

Theorem 1. Let \(v_k, u_k, a_k\) and \(b_k\) be sequences of non-negative random variables such that

\[
\mathbb{E}[v_{k+1} | F_k] \leq (1 + a_k) v_k - u_k + b_k \quad \text{for all } k \geq 0 \text{ a.s.}
\]

where \(F_k\) denotes the collection \(v_0, ..., v_k, u_0, ..., u_k, a_0, ..., a_k\) and \(b_0, ..., b_k\). Also, \(\sum_{k=0}^{\infty} a_k < \infty\) and \(\sum_{k=0}^{\infty} b_k < \infty\) a.s.

Then, we have \(\lim_{k \to \infty} v_k = v\) for a random variable \(v \geq 0\) a.s. and \(\sum_{k=0}^{\infty} u_k < \infty\) a.s.

VI. Convergence Analysis for the Main Algorithm SRDO

Definition 1. Let

\[
R_i(k) = -\alpha_k \sum_{j \in \Gamma_i(k)} A_{ji}^{(i)} B_{ji}^{(i)} (\nabla f_j^{(i)} (v_l(k) - k_j') - \nabla f_j^{(i)} (v_l(k)))
\]

\[
= \alpha_k \epsilon_i(k)
\]

Hence, from (5) and (6) we have

\[
x_i(k + 1) = v_i(k) - \alpha_k \nabla f^{(i)} (v_i(k)) + R_i(k)
\]

Or

\[
x_i(k + 1) = v_i(k) - \alpha_k \nabla f^{(i)} (v_i(k)) + R_i(k),
\]

where \(x_i(k + 1) = v_i(k) - \alpha_k \nabla f^{(i)} (v_i(k))\)

Lemma 1. Let Assumptions 1, 2, 4 and 5 hold. Let the sequences \(\{x_i(k)\}\) and \(\{v_i(k)\}\) be generated by method [3]. Then we have

\[
\sum_{i=1}^{n} \|v_i(k + 1) - \mathbf{x}^*\|^2 \leq \sum_{i=1}^{n} \|v_i(k) - \mathbf{x}^*\|^2
\]

\[
+ 2\alpha_k \sum_{j=1}^{n} \gamma_{i(i)} (\nabla f^{(i)} (v_j(k)), \mathbf{x}^* - v_j(k))
\]

\[
+ 2\alpha_k \sum_{j=1}^{n} \gamma_{i(i)} (\epsilon_{j(i)}(k), v_j(k) - \mathbf{x}^*)
\]

\[
+ 2\alpha_k \sum_{j=1}^{n} \sum_{(i)=1}^{p} \|\nabla f^{(i)} (v_j(k))\|^2 + \|\epsilon_{j(i)}(k)\|^2
\]

(11)
Lemma 2. Let Assumptions 1, 2, 4 and 5 hold. Let the functions in Assumption 1 also satisfy inequality (122) of Lemma 7 for strongly convex for every \( \sigma \) for \( k \). Let Assumptions 1-5 hold. Let the functions in Assumption 1 be strongly convex and satisfy given in Assumption 1. Then, the sequences \( \{x_i(k)\} \) and \( \{v_i(k)\} \), \( i \in V \) be generated by method 3. Then we have

\[
(1 - \mu) \sum_{i=1}^{n} \|v_i(k + 1) - x^*\|^2 \leq (1 - \mu) \sum_{j=1}^{n} \|v_j(k) - x^*\|^2 + 2\alpha_k \sum_{j=1}^{n} \sum_{i=1}^{p} \gamma(i) \langle \epsilon_j(i)(k), v_j - x^* \rangle + 2\alpha_k^2 \sum_{j=1}^{n} \sum_{i=1}^{p} \gamma(i)^2 \|\epsilon_j(i)(k)\|^2
\]

(12)

- \( 2\alpha_k \sum_{j=1}^{n} \sum_{i=1}^{p} \gamma(i)(b - \frac{2\alpha_k a}{(\sum_{i=1}^{p} \gamma(i))^{2/3}})\|v_j(k) - x^*(i)\|^2 \)

Proposition 1. Let Assumptions 1-5 hold. Let the functions in Assumption 1 be strongly convex and satisfy \( f^{(i)}(x^*) = f^{(i)}(x^{(i)}) \) for all \( i \). Let the sequences \( \{x_i(k)\} \) and \( \{v_i(k)\} \), \( i \in V \) be generated by method 3 with stepsizes and errors as given in Assumptions 3 and 5. Assume that problem 1 has a non-empty optimal solution set \( X^* \) as given in Assumption 1. Then, the sequences \( \{x_i(k)\} \) and \( \{v_i(k)\} \), \( i \in V \) converge to the same random point in \( X^* \) with probability 1.

Proof. With errors as in Assumption 3 we have \( \|R_i(k)\| \) and \( \|\epsilon_j(i)(k)\| \) as given in Appendix. Then having \( f^{(i)}(x^*) = f^{(i)}(x^{(i)}) \) for all \( i \) also satisfied we have Lemma 2 satisfied. Then we can use the resulting inequality with the substitution of \( \|\epsilon_j(i)(k)\| \) from Appendix to get

\[
\sum_{i=1}^{n} \|v_i(k + 1) - x^*\|^2 \leq (1 - \mu) \sum_{j=1}^{n} \|v_j(k) - x^*\|^2 + 4L\alpha_k \sum_{j=1}^{n} \max_{i} \|A^{(i)}\|_\infty \|B^{(i)}\|_\infty \max_{k-H \leq l \leq k} \|v_q(k) - x^*\|^2
\]

(13)

\[
+ 8L^2\alpha_k^2 \sum_{j=1}^{n} \max_{i} \|A^{(i)}\|_\infty \|B^{(i)}\|_\infty \max_{k-H \leq l \leq k} \|v_q(k) - x^*\|^2 \frac{(\sum_{p} \gamma(i))^{2/3}}{(\sum_{i=1}^{p} \gamma(i))^{2/3}}
\]

But in order to be able to use Lemma 7 the last term in (13) should be negative. Which means \( b - \frac{2\alpha_k a}{(\sum_{i=1}^{p} \gamma(i))^{2/3}} \geq 0 \) where \( b = \langle \overrightarrow{u}, \overrightarrow{v} \rangle \). And \( \nabla f^{(i)}(v_j(k)) - \nabla f^{(i)}(x^{(i)}) \) = \( a\|v_j(k) - x^{(i)}\| \overrightarrow{u} \) where \( a \leq L \). However, \( f^{(i)} \) is strongly convex for every \( i \), then \( \langle \nabla f^{(i)}(v_j(k)) - \nabla f^{(i)}(x^{(i)}), v_j(k) - x^{(i)} \rangle = a\|v_j(k) - x^{(i)}\|^2 \langle \overrightarrow{u}, \overrightarrow{v} \rangle \geq \sigma_{(i)}^2 \|v_j(k) - x^{(i)}\|^2 \), that is \( \langle \overrightarrow{u}, \overrightarrow{v} \rangle \geq \frac{\sigma_{(i)}}{a} \). Therefore, a sufficient condition is

\[
\frac{2\alpha_k a}{(\sum_{i=1}^{p} \gamma(i))^{2/3}} \leq \frac{2\alpha_k L}{(\sum_{i=1}^{p} \gamma(i))^{2/3}} \leq \frac{\sigma_{(i)}}{a} = \langle \overrightarrow{u}, \overrightarrow{v} \rangle
\]

(14)

The sufficient condition in (14) is satisfied for \( k \geq k_0 \) since \( \alpha_k \rightarrow 0 \). Then (13) is similar to the martingale inequality 122 of Lemma 7 for \( k \geq k_0 \). By the result of Lemma 7 we have for \( v_k = \sum_{i=1}^{n} \|v_i(k) - x^*\|^2 \) that

\[
\sum_{i=1}^{n} \|v_i(k) - x^*\|^2 \leq \rho^k V_0
\]

(15)

for \( k \geq k_1 = \tilde{k} = k_0 + B \) where \( \rho, V_0 \) and \( \tilde{k} \) are as in the lemma. Therefore, as \( k \rightarrow \infty \) we have \( \sum_{i=1}^{n} \|v_i(k) - x^*\|^2 \rightarrow 0 \). That is, \( \|v_i(k) - x^*\| \rightarrow 0 \) for all \( i \in V \). Then in view of (3b) where \( x_i(k + 1) = v_i(k) - \)
Proposition 2. Let Assumptions 1-5 hold. Let the functions in Assumption 1 satisfy $f(i(x^*)) = f(i(x^{(i)}))$ for all (i). Let the sequences $\{x_i(k)\}$ and $\{v_i(k)\}$, $i \in V$ be generated by method \([3]\) with stepsizes and errors as given in Assumptions 3 and 5. Assume that problem \([1]\) has a non-empty optimal solution set $X^*$ as given in Assumption 1. Then, the sequences $\{x_i(k)\}$ and $\{v_i(k)\}$, $i \in V$ converge to the same random point in $X^*$ with probability 1.

Proof. With errors as in Assumption 3 we have $\| R_i(k) \|$ and $\|e_{j(i)}(k)\|$ as given in Appendix. Then having $f(i(x^*)) = f(i(x^{(i)}))$ for all (i) also satisfied we have Lemma \([2]\) satisfied. Then we can use the resulting inequality \([12]\) with the substitution of $\|e_{j(i)}(k)\|$ from Appendix to get

$$\sum_{i=1}^{n} \|v_i(k+1) - x^*\|^2 \leq (1 - \mu) \sum_{j=1}^{n} \|v_j(k) - x^*\|^2 + 4L\alpha_k \sum_{j=1}^{n} \max_{(i)} \|A^{(i)}\|_\infty \|B^{(i)}\|_{2,\infty} \max_{k-H \leq k < q \in V} \|v_q(k) - x^*\|^2$$

$$+ 8L^2 \alpha_k^2 \sum_{j=1}^{n} \max_{(i)} \|A^{(i)}\|_\infty^2 \|B^{(i)}\|_{2,\infty}^2 \max_{k-H \leq k < q \in V} \|v_q(k) - x^*\|^2$$

$$- 2a\alpha_k \sum_{j=1}^{n} \sum_{(i)=1}^{p} \gamma_{(i)}(b - \frac{2a\alpha_k}{(\sum_{(i)=1}^{p} \gamma_{(i)})^2}) \|v_j(k) - x^{(i)}\|^2$$

But in order to be able to use Lemma \([7]\) on \([16]\) we must have

$$- \mu \sum_{j=1}^{n} \|v_j(k) - x^*\|^2 - 2a\alpha_k \sum_{j=1}^{n} \sum_{(i)=1}^{p} \gamma_{(i)}(b - \frac{2a\alpha_k}{(\sum_{(i)=1}^{p} \gamma_{(i)})^2}) \|v_j(k) - x^{(i)}\|^2 \leq 0$$

That is

$$- \mu \sum_{j=1}^{n} \|v_j(k) - x^*\|^2 + \sum_{j=1}^{n} \sum_{(i)=1}^{p} \gamma_{(i)} \left(\frac{4a^2 \alpha_k^2}{(\sum_{(i)=1}^{p} \gamma_{(i)})^2}\right) \|v_j(k) - x^{(i)}\|^2$$

$$\leq 2a\alpha_k \sum_{j=1}^{n} \sum_{(i)=1}^{p} \gamma_{(i)} b \|v_j(k) - x^{(i)}\|^2$$

which reduces to the sufficient condition

$$- \mu \sum_{j=1}^{n} \|v_j(k) - x^*\|^2 + \sum_{j=1}^{n} \sum_{(i)=1}^{p} \gamma_{(i)} \left(\frac{8a^2 \alpha_k^2}{(\sum_{(i)=1}^{p} \gamma_{(i)})^2}\right) \|v_j(k) - x^*\|^2$$

$$+ \sum_{j=1}^{n} \sum_{(i)=1}^{p} \gamma_{(i)} \left(\frac{8a^2 \alpha_k^2}{(\sum_{(i)=1}^{p} \gamma_{(i)})^2}\right) \|x^* - x^{(i)}\|^2$$

$$\leq 2a\alpha_k \sum_{j=1}^{n} \sum_{(i)=1}^{p} \gamma_{(i)} b \|v_j(k) - x^{(i)}\|^2$$
But $0 \geq b = \langle \overline{w}, \overline{v} \rangle \leq 1$. And $\nabla f^{(i)}(v_j(k)) - \nabla f^{(i)}(x^{(i)}) = a \|v_j(k) - x^{(i)}\| \overline{w}$ where $a \leq L$. Thus, the right hand side of (19) is nonnegative. Therefore, the sufficient condition in (19) reduces to

$$\sum_{j=1}^{n} \sum_{(i)=1}^{p} \gamma(i) \langle \epsilon_j, (i)(k), v_j - x^{(i)} \rangle \leq (\mu - \sum_{(i)=1}^{p} \gamma(i)) \sum_{j=1}^{n} \|v_j(k) - x^{(i)}\|^2$$

But the sufficient condition in (20) is satisfied for $0 < \mu < 1$ and $0 < a < L$ since $\alpha_k \rightarrow 0$. Thus,

$$\sum_{j=1}^{n} \sum_{(i)=1}^{p} \gamma(i) \langle \epsilon_j, (i)(k), v_j - x^{(i)} \rangle \leq (\mu - \sum_{(i)=1}^{p} \gamma(i)) \sum_{j=1}^{n} \|v_j(k) - x^{(i)}\|^2$$

Therefore, for $k \geq k_0$ we have the right hand side of (21) to be greater than $E$ and the left hand side to be less than $E$ since $\alpha_k \rightarrow 0$. Then (16) is similar to the martingale inequality [22] of Lemma [7] for $k \geq k_0$. By the result of Lemma [7] we have for $v_k = \sum_{i=1}^{n} \|v_i(k) - x^{*}\|^2$

$$\sum_{i=1}^{n} \|v_i(k) - x^{*}\|^2 \leq \rho^k V_0$$

for $k \geq k_1 = \overline{k} = k_0 + B$ where $\rho$, $V_0$ and $\overline{k}$ are as in the lemma. Therefore, as $k \rightarrow \infty$ we have $\sum_{i=1}^{n} \|v_i(k) - x^{*}\|^2 \rightarrow 0$. That is, $\|v_i(k) - x^{*}\| \rightarrow 0$ for all $i \in V$. Then in view of (3b) where $x_i(k+1) = v_i(k) - \alpha_k \nabla f^{(i)}(v_i(k)) + R_i(k)$ and since $R_i(k) \rightarrow 0$ because $\|v_i(k) - x^{*}\| \rightarrow 0$ or $\alpha_k \rightarrow 0$ and $\alpha_k \nabla f^{(i)}(v_i(k)) \rightarrow 0$ since $\alpha_k \rightarrow 0$ and $v_i(k) \rightarrow x^{*}$, thus $v_i(k) \in X$ where $\nabla f^{(i)}(v_i(k)) \leq Gf_i$. Therefore, $x_i(k+1) = v_i(k) = x^{*}$ as $k \rightarrow \infty$.

Lemma 3. Let Assumptions 1, 2, 4 and 5 hold. Let the functions in Assumption 1 also satisfy $f^{(i)}(x^{*}) > f^{(i)}(x^{(i)})$ for at least one $(i)$. Let the sequences $\{x_i(k)\}$ and $\{v_i(k)\}$, $i \in V$ be generated by method (3). Then we have

$$1 - \mu \sum_{i=1}^{n} \|v_i(k+1) - x^{*}\|^2 \leq (1 - \mu) \sum_{j=1}^{n} \|v_j(k) - x^{*}\|^2$$

$$+ 2\alpha_k \sum_{j=1}^{n} \sum_{(i)=1}^{p} \gamma(i) \langle \epsilon_j, (i)(k), v_j - x^{*} \rangle + 2\alpha_k \sum_{j=1}^{n} \sum_{(i)=1}^{p} \gamma(i) \|\epsilon_j, (i)(k)\|^2$$

$$+ 2\alpha_k \sum_{j=1}^{n} \sum_{(i) \in I} \gamma(i) \langle \nabla f^{(i)}(v_j(k)), x^{*} - x^{(i)} \rangle$$

$$- 2\alpha_k \sum_{j=1}^{n} \sum_{(i)=1}^{p} \gamma(i) \langle b - \frac{2\alpha_k a}{\sum_{(i)=1}^{p} \gamma(i)} \rangle \|v_j(k) - x^{(i)}\|^2$$

(23)
Proposition 3. Let Assumptions 1-5 hold. Let the functions in Assumption 1 be strongly convex and satisfy \( f^{(i)}(x^*) > f^{(i)}(x^{(i)}) \) for at least one \( (i) \). Let the sequences \( \{x_i(k)\} \) and \( \{v_i(k)\} \), \( i \in V \) be generated by method (3) with stepsizes and errors as given in Assumptions 3 and 5. Assume that problem (1) has a non-empty optimal solution set \( X^* \) as given in Assumption 1. Then, the sequence \( \{\sum_{i=1}^{n} \|v_i(k) - x^*\|^2\} \) converge to a nonnegative value \( D \).

Proof. With errors as in Assumption 3 we have \( \|R_i(k)\| \) and \( \|e_i(k)\| \) as given in Appendix. Then having \( f^{(i)}(x^*) > f^{(i)}(x^{(i)}) \) for at least one \( (i) \) also satisfied we have Lemma 3 satisfied. Then we can use the resulting inequality (23) with the substitution of \( \|e_{j,(i)}(k)\| \) to get

\[
\sum_{l=1}^{n} \|v_l(k+1) - x^*\|^2 \leq (1 - \mu) \sum_{j=1}^{n} \|v_j(k) - x^*\|^2 \\
+ 2\alpha_k n I/L \max_{(i)} \|x^* - x^{(i)}\|^2 \\
+ 4L\alpha_k \sum_{j=1}^{n} \|A^{(i)}\|_{\infty} \|B^{(i)}\|_{2,\infty} \max_{k-H \leq \hat{k} \leq k; q \in V} \|v_q(\hat{k}) - x^*\|^2 \\
+ 8L^2 \alpha_k \sum_{j=1}^{n} \|A^{(i)}\|_{2,\infty} \|B^{(i)}\|_{2,\infty} \max_{k-H \leq \hat{k} \leq k; q \in V} \|v_q(\hat{k}) - x^*\|^2 \\
- 2\alpha_k \sum_{j=1}^{n} \sum_{p=1}^{p} \gamma_{(i)}(b - \frac{2\alpha_k a}{(\sum_{(i)=1}^{\gamma_{(i)}})^2}) \|v_j(k) - x^{(i)}\|^2 \tag{24}
\]

But in order to be able to use Lemma 3 the last term in (24) should be negative. Which means \( b - \frac{\gamma_{(i)}}{a} \geq 0 \) where \( b = \langle \bar{u}, \bar{v} \rangle \). And \( \nabla f^{(i)}(v_j(k)) - \nabla f^{(i)}(x^{(i)}) = a \|v_j(k) - x^{(i)}\| \bar{u} \) where \( a \leq L \). However, \( f^{(i)} \) is strongly convex for every \( (i) \), then \( \langle \nabla f^{(i)}(v_j(k)) - \nabla f^{(i)}(x^{(i)}), v_j(k) - x^{(i)} \rangle = a \|v_j(k) - x^{(i)}\|^2 \langle \bar{u}, \bar{v} \rangle \geq \sigma_{(i)}^2 \|v_j(k) - x^{(i)}\|^2 \), that is \( \langle \bar{u}, \bar{v} \rangle \geq \frac{\sigma_{(i)}^2}{a} \). Therefore, a sufficient condition is

\[
\frac{2\alpha_k a}{(\sum_{(i)=1}^{\gamma_{(i)}})^2} \leq \frac{2\alpha_k L}{(\sum_{(i)=1}^{\gamma_{(i)}})^2} \leq \frac{\sigma_{(i)}^2}{a} = \langle \bar{u}, \bar{v} \rangle \tag{25}
\]

The sufficient condition in (25) is satisfied for \( k \geq k_0 \) since \( \alpha_k \to 0 \). Since \( \max_{(i)} \|x^* - x^{(i)}\|^2 \) is fixed independent of \( k \) then (24) is similar to the martingale inequality (32) of Lemma 8. By the result of Lemma 8 we have for \( v_k = \sum_{i=1}^{n} \|v_i(k) - x^*\|^2 \) that

\[
\sum_{i=1}^{n} \|v_i(k) - x^*\|^2 \leq \rho^k V_0 + \eta \tag{26}
\]

for \( k \geq \bar{k}_2 = \bar{k} = k_0 + B \) where \( \rho, V_0, \bar{k} \) are as in the lemma and where \( \eta > 0 \) as substituted from the inequality (24) by using the lemma. Therefore, as \( k \to \infty \) we have \( \sum_{i=1}^{n} \|v_i(k) - x^*\|^2 \to \eta \). That is, \( \sum_{i=1}^{n} \|v_i(k) - x^*\|^2 < D \) for all \( i \in V \) as \( k \geq \bar{k}_2 = \bar{k} \).

\( \square \)

Proposition 4. Let Assumptions 1-5 hold. Let the functions in Assumption 1 satisfy \( f^{(i)}(x^*) > f^{(i)}(x^{(i)}) \) for at least one \( (i) \). Let the sequences \( \{x_i(k)\} \) and \( \{v_i(k)\} \), \( i \in V \) be generated by method (3) with stepsizes and errors as given in Assumptions 3 and 5. Assume that problem (1) has a non-empty optimal solution set \( X^* \) as given in Assumption 1. Then, the sequence \( \{\sum_{i=1}^{n} \|v_i(k) - x^*\|^2\} \) converge to a nonnegative value \( D \).
Proof. With errors as in Assumption 3 we have \( \|R_i(k)\| \) and \( \|e_i(k)\| \) as given in Appendix. Then having \( f^{(i)}(x^*) > f^{(i)}(x^{(i)}) \) for at least one \( (i) \) also satisfied we have Lemma 3 satisfied. Then we can use the resulting inequality \( (23) \) with the substitution of \( \|e_{j,(i)}(k)\| \) to get

\[
\sum_{j=1}^{n} \|v_j(k+1) - x^*\|^2 \\
\leq (1 - \mu) \sum_{j=1}^{n} \|v_j(k) - x^*\|^2 \\
+ 2\alpha_k n |L| L \max_{(i)} \|x^* - x^{(i)}\|^2 \\
+ 4L\alpha_k \sum_{j=1}^{n} \|A^{(i)}\|_{\infty} \|B^{(i)}\|_{2,\infty} \max_{k-H \leq k \leq k; q \in V} \|v_q(\hat{k}) - x^*\|^2 \\
+ 8L^2\alpha_k^2 \sum_{j=1}^{n} \|A^{(i)}\|_{\infty} \|B^{(i)}\|_{2,\infty} \max_{k-H \leq k \leq k; q \in V} \|v_q(\hat{k}) - x^*\|^2 \\
- 2\alpha_k \sum_{j=1}^{n} \sum_{j=1}^{p} \gamma_{(i)} b - \frac{2\alpha_k a}{(\sum_{j=1}^{p} \gamma_{(i)})^2} \|v_j(k) - x^{(i)}\|^2.
\] (27)

But in order to be able to use Lemma 8 on (16) we must have

\[
-\mu \sum_{j=1}^{n} \|v_j(k) - x^*\|^2 - 2\alpha_k \sum_{j=1}^{n} \sum_{j=1}^{p} \gamma_{(i)} b - \frac{2\alpha_k a}{(\sum_{j=1}^{p} \gamma_{(i)})^2} \|v_j(k) - x^{(i)}\|^2 \leq 0
\] (28)

That is

\[
-\mu \sum_{j=1}^{n} \|v_j(k) - x^*\|^2 + \sum_{j=1}^{n} \sum_{j=1}^{p} \gamma_{(i)} b - \frac{4\alpha_k^2 a^2}{(\sum_{j=1}^{p} \gamma_{(i)})^2} \|v_j(k) - x^{(i)}\|^2 \\
\leq 2\alpha_k \sum_{j=1}^{n} \sum_{j=1}^{p} \gamma_{(i)} b \|v_j(k) - x^{(i)}\|^2
\] (29)

which reduces to the sufficient condition

\[
-\mu \sum_{j=1}^{n} \|v_j(k) - x^*\|^2 + \sum_{j=1}^{n} \sum_{j=1}^{p} \gamma_{(i)} \frac{8\alpha_k^2 a^2}{(\sum_{j=1}^{p} \gamma_{(i)})^2} \|v_j(k) - x^*\|^2 \\
+ \sum_{j=1}^{n} \sum_{j=1}^{p} \gamma_{(i)} \frac{8\alpha_k^2 a^2}{(\sum_{j=1}^{p} \gamma_{(i)})^2} \|x^* - x^{(i)}\|^2 \\
\leq 2\alpha_k \sum_{j=1}^{n} \sum_{j=1}^{p} \gamma_{(i)} b \|v_j(k) - x^{(i)}\|^2
\] (30)

But \( 0 \geq b = (\overline{u}, \overline{v}) \leq 1 \). And \( \nabla f^{(i)}(v_j(k)) - \nabla f^{(i)}(x^{(i)}) = a \|v_j(k) - x^{(i)}\| \overline{u} \) where \( a \leq L \). Thus, the right hand side of (30) is nonnegative. Therefore, the sufficient condition in (30) reduces to

\[
+ \sum_{j=1}^{n} \sum_{j=1}^{p} \gamma_{(i)} \frac{8\alpha_k^2 a^2}{(\sum_{j=1}^{p} \gamma_{(i)})^2} \|x^* - x^{(i)}\|^2 \\
\leq (\mu - \sum_{j=1}^{n} \gamma_{(i)} \frac{8\alpha_k^2 a^2}{(\sum_{j=1}^{p} \gamma_{(i)})^2} \sum_{j=1}^{n} \|v_j(k) - x^*\|^2
\] (31)
But the sufficient condition in (31) is satisfied for $0 < \mu < 1$ and $0 < a < L$ since $\alpha_k \to 0$. Thus,

$$+ \sum_{j=1}^{n} \sum_{(i)=1}^{P} \gamma(i) \frac{8\alpha_k^2 a^2}{(\sum_{(i)=1}^{P} \gamma(i))^2} \|x^* - x^{(i)}\|^2$$

$$+ \sum_{j=1}^{n} \sum_{(i)=1}^{P} \gamma(i) \frac{8\alpha_k^2 L^2}{(\sum_{(i)=1}^{P} \gamma(i))^2} \|x^* - x^{(i)}\|^2$$

$$\leq (\mu - \sum_{(i)=1}^{P} \gamma(i) \frac{8\alpha_k^2 L^2}{(\sum_{(i)=1}^{P} \gamma(i))^2} \sum_{j=1}^{n} \|v_j(k) - x^*\|^2$$

(32)

$$\leq (\mu - \sum_{(i)=1}^{P} \gamma(i) \frac{8\alpha_k^2 a^2}{(\sum_{(i)=1}^{P} \gamma(i))^2} \sum_{j=1}^{n} \|v_j(k) - x^*\|^2$$

Therefore, for $k \geq k_0$ we have the right hand side of (32) to be greater than $E$ and the left hand side to be less than $E$ since $\alpha_k \to 0$. Since $\max_{(i)} \|x^* - x^{(i)}\|^2$ is fixed independent of $k$ then (27) is similar to the martingale inequality (132) of Lemma 8. By the result of Lemma 8 we have for $v_k = \sum_{i=1}^{n} \|v_i(k) - x^*\|^2$ that

$$\sum_{i=1}^{n} \|v_i(k) - x^*\|^2 \leq \rho^k V_0 + \eta$$

(33)

for $k \geq \tilde{k}_2 = \tilde{k} = k_0 + B$ where $\rho$, $V_0$, $\tilde{k}$ are as in the lemma and where $\eta > 0$ as substituted from the inequality (27) by using the lemma. Therefore, as $k \to \infty$ we have $\sum_{i=1}^{n} \|v_i(k) - x^*\|^2 \to \eta$. That is, $\sum_{i=1}^{n} \|v_i(k) - x^*\|^2 < D$ for all $i \in V$ as $k \geq \tilde{k}_2 = \tilde{k}$.

Proposition 5. Let Assumptions 1-5 hold. Let the functions in Assumption 1 be strongly convex and satisfy $f^{(i)}(x^*) > f^{(i)}(x^{(i)})$ for at least one (i) and let $p < \frac{1}{\gamma_{min}}$. Let the sequences $\{x_i(k)\}$ and $\{v_i(k)\}$, $i \in V$ be generated by method (3) with stepsizes and errors as given in Assumptions 3 and 5. Assume that problem (1) has a non-empty optimal solution set $X^*$ as given in Assumption 1. Then, the sequences $\{x_i(k)\}$ and $\{v_i(k)\}$, $i \in V$ converge to the same random point in $X^*$ with probability 1.

Proof. Having the assumptions of Proposition 5 holding then Proposition 3 is satisfied. Then Lemma 4 premises are satisfied and the lemma follows. Then Lemma 5 premises are satisfied and the lemma follows. Subsequently, Lemma 6 premises are satisfied and the lemma follows proving the proposition.

 Proposition 6. Let Assumptions 1-5 hold. Let the functions in Assumption 1 satisfy $f^{(i)}(x^*) > f^{(i)}(x^{(i)})$ for at least one (i) and let $p < \frac{1}{\gamma_{min}}$. Let the sequences $\{x_i(k)\}$ and $\{v_i(k)\}$, $i \in V$ be generated by method (3) with stepsizes and errors as given in Assumptions 3 and 5. Assume that problem (1) has a non-empty optimal solution set $X^*$ as given in Assumption 1. Then, the sequences $\{x_i(k)\}$ and $\{v_i(k)\}$, $i \in V$ converge to the same random point in $X^*$ with probability 1.

Proof. Having the assumptions of Proposition 6 holding then Proposition 4 is satisfied. Then Lemma 4 premises are satisfied and the lemma follows. Then Lemma 5 premises are satisfied and the lemma follows. Subsequently, Lemma 6 premises are satisfied and the lemma follows proving the proposition.
Lemma 4. Let Assumption I hold, and \([W(k)]\) be row stochastic as in Assumption 2, let Proposition 3 or 4 be satisfied and \(p < \frac{1}{\gamma_{\text{min}}}\) and let \(\sum_{k=0}^{\infty} \alpha_k^2 < \infty\). Then, \(\sum_{k=0}^{\infty} \text{dist}^2(v_i(k), X) < \infty\) for all \(i \in V\) a.s.

Proof. We are going to prove Lemma 4 by using the following lemmas 9 and 10 and then exploiting the supermartingale Theorem 1.

Remark 5. Note that the case of Least squares optimization for the set \(X\) to acquire bounded gradients it must be a bounded set and thus from the representation theorem of closed bounded sets it must be that \(X \equiv \bigcap_{i=1}^{t} X_i\) then let \(\hat{x} \in X\). This implies that \(\hat{x} \in X_i\) for all \(i \in \{1, 2, \ldots, t\}\). And hence we only require \(f_i\) to be Lipschitz on \(X_i\) for all \(i \in \{1, 2, \ldots, t\}\) for Lemmas 9 and 10 to be applied, then \(f_j\) is Lipschitz on \(X\). However, in the analysis we require the estimates and weighted averages gradients to be also Lipschitz without insisting that they must lie in any \(X_i\) (i.e., they can be anywhere in \(\mathbb{R}^N\)). So to avoid any obstacle that can hinder the flow of the proof we assume Assumption 1(b) and require Lipschitz gradients on all of \(\mathbb{R}^N\) for any scenario of Problem (1). (In fact in our simulation where we investigate the Least squares problem, we take \(f_j\) is Lipschitz on the whole \(\mathbb{R}^N\), this for ease of implementation since we pick the matrix \(A\) randomly.

Take \(x_i(k+1) = x - \alpha \nabla \Phi(x) \in \mathbb{R}^N\), then being in \(\mathbb{R}^N\) the following inequalities hold,

\[
\text{dist}(x_i(k+1), X) = \|x_i(k+1) - \Pi_X(x_i(k+1))\| \\
\leq \|x_i(k+1) - \hat{x}\|
\]

In fact, we can do better. For any \(\hat{x} \in X\)

\[
\|\Pi_X(x_i(k+1)) - \hat{x}\| \leq \|x_i(k+1) - \hat{x}\|^2 - \|\Pi_X(x_i(k+1)) - x_i(k+1)\|^2
\]

\[\Rightarrow \text{dist}^2(x_i(k+1), X) = \|\Pi_X(x_i(k+1)) - x_i(k+1)\|^2 \leq \|x_i(k+1) - \hat{x}\|^2 - \|\Pi_X(x_i(k+1)) - \hat{x}\|^2.
\]

This implies that

\[
\text{dist}^2(x_i(k+1), X) = \|\Pi_X(x_i(k+1)) - x_i(k+1)\|^2 \leq \|x_i(k+1) - \hat{x}\|^2 - a\|\Pi_X(x_i(k+1)) - \hat{x}\|^2, \tag{34}
\]

where \(a \leq 1\).

We used this form of the inequality with a constant \(a\) that we can adjust in order for the supermartingale theorem hypothesis to hold.

Remark 6. Notice from the procedure of the proof that \(\tau, \eta > 0\) are arbitrary and independent from any quantity, i.e., they are not related to any bound on any quantity. In other words, the lemmas inequalities can adapt to any values \(\tau, \eta > 0\).

Using equation (34) and substituting the results of Lemma 9 and Lemma 10 on the above substitutions, we have for each use of \(f^{(i)}\) the following applies on node \(i\)
\[ \text{dist}^2(\mathbf{x}_i(k+1), \mathcal{X}) \leq \] 
\[ \|\mathbf{x}_i(k+1) - \hat{x}\|^2 - a\|\Pi_X(\mathbf{x}_i(k+1)) - \hat{x}\|^2 \] 
\[ \leq (1 - a + (A_\eta - aA_\tau)\alpha^2)\|\mathbf{x} - \mathbf{x}\|^2 \] 
\[ - (2 - 2a)\alpha(\Phi(\mathbf{z}) - \Phi(\hat{x})) + \frac{3a}{4}\|\mathbf{y} - \mathbf{x}\|^2 \] 
\[ + \left(\frac{3}{8\eta} - \frac{3a}{8\tau} + (2 - 2a)\alpha L\right)\|\mathbf{x} - \mathbf{z}\|^2 \] 
\[ + (B_\eta - aB_\tau)\alpha^2\|\nabla\Phi(\hat{x})\|^2, \tag{35} \]

where \( A_\tau = 8L^2 + 16\tau L^2, B_\tau = 8\tau + 8, \tau > 0 \) is arbitrary, \( A_\eta = 2L^2 + 16\eta L^2, B_\eta = 2\eta + 8 \) and \( \eta > 0 \) is arbitrary. Here, we note that \( \hat{x} \in \mathcal{X} \).

Now, we use the above inequality, (35) for the purpose of making the hypothesis of the supermartingale theorem hold. We thus use the following substitutions: \( \Phi = f^{(i)}, \alpha = \alpha_k, \hat{x} \in \mathcal{X} = \bigcap_{i=1}^n \mathcal{X}_i, \mathbf{y} = \Pi_X(\mathbf{x}_i(k+1), \mathbf{x}_i(k+1) = \mathbf{x} - \alpha_k \nabla f(\mathbf{x}) \) and \( \mathbf{x} = \mathbf{v}_i(k) \).

In particular, if we take \( \hat{x} = \Pi_X(\mathbf{v}_i(k)) \in \mathcal{X} \) in the feasibility region and \( \mathbf{z} = \Pi_X(\mathbf{v}_i(k)) = \hat{x} \) then the above conclusion (35) becomes

\[ \text{dist}^2(\mathbf{x}_i(k+1), \mathcal{X}) \leq (1 - a + (A_\eta - aA_\tau)\alpha_k^2)\text{dist}^2(\mathbf{v}_i(k), \mathcal{X}) \] 
\[ + \frac{3a}{4}\|\Pi_X(\mathbf{v}_i(k) - \alpha_k \nabla f^{(i)}(\mathbf{v}_i(k))) - \mathbf{v}_i(k)\|^2 \] 
\[ + \left(\frac{3}{8\eta} - \frac{3a}{8\tau} + (2 - 2a)\alpha_k L\right)\text{dist}^2(\mathbf{v}_i(k), \mathcal{X}) \] 
\[ + (B_\eta - aB_\tau)\alpha_k^2\|\nabla f^{(i)}(\hat{x})\|^2, \tag{36} \]

But by using the nonexpansiveness property of a projection we have

\[ \|\Pi_X(\mathbf{v}_i(k) - \alpha_k \nabla f^{(i)}(\mathbf{v}_i(k))) - \mathbf{v}_i(k)\|^2 \leq \|\alpha_k \nabla f^{(i)}(\mathbf{v}_i(k))\|^2 \] 
\[ - \|\Pi_X(\mathbf{v}_i(k) - \alpha_k \nabla f^{(i)}(\mathbf{v}_i(k))) - \mathbf{v}_i(k) - \alpha_k \nabla f^{(i)}(\mathbf{v}_i(k))\|^2 \]

Then for \( 0 \leq a \leq 1 \). Then from \( \hat{x} = \Pi_X(\mathbf{v}_i(k)) \), \( (a + b)^2 \leq 2a^2 + 2b^2 \) and the Lipschitz property of \( f \) from Assumption 1(b), we get

\[ \frac{3a}{4}\|\Pi_X(\mathbf{v}_i(k) - \alpha_k \nabla f^{(i)}(\mathbf{v}_i(k))) - \mathbf{v}_i(k)\|^2 \leq \] 
\[ \frac{3a}{4}\|\alpha_k \nabla f^{(i)}(\mathbf{v}_i(k))\|^2 - \frac{3a}{4}\text{dist}^2(\mathbf{v}_i(k) - \alpha_k \nabla f^{(i)}(\mathbf{v}_i(k)), \mathcal{X}) \] 
\[ \leq \frac{3a}{4}\|\alpha_k \nabla f^{(i)}(\mathbf{v}_i(k))\|^2 \]

Then for \( \hat{x} = \Pi_X(\mathbf{v}_i(k)) \), \( (a + b)^2 \leq 2a^2 + 2b^2 \) and the Lipschitz property of \( f \) from Assumption 1(b), we get
\[ \frac{3\alpha}{4} \| \Pi_X(v_i(k) - \alpha_k \nabla f(i)(v_i))(k)) - v_i(k) \|^2 \leq \]
\[ \| \alpha_k \nabla f(i)(v_i(k)) - \nabla f(\hat{x}) + \nabla f(i)(\hat{x}) \|^2 \leq 2\alpha_k^2 L^2 \| v_i(k) - \hat{x} \|^2 + 2\alpha_k^2 \| \nabla f(i)(\hat{x}) \|^2 \leq 2\alpha_k^2 L^2 \text{dist}^2(v_i(k), X) + 2\alpha_k^2 \| \nabla f(i)(\hat{x}) \|^2. \]

Then the result (36) becomes

\[ \text{dist}^2(x_i(k+1), X) \leq \]
\[ (1 - a + (A_\eta - aA_\tau)\alpha_k^2) \text{dist}^2(v_i(k), X) \]
\[ + \left( \frac{3 - 3a}{8\eta} + 2\alpha_k^2 L^2 + (2 - 2a)\alpha_k L \right) \text{dist}^2(v_i(k), X) \]
\[ + (B_\eta - aB_\tau + 2)\alpha_k^2 \| \nabla f(i)(\hat{x}) \|^2 \]
(37)

From (10) and (5) substitution we can write

\[ x_i(k+1) = \tilde{x}_i(k+1) + R_i(k), \]
(38)

where \( \tilde{x}_i(k+1) = v_i(k) - \alpha_k \nabla f(i)(v_i(k)) \)

(cf. Definition [1] for the definition of \( R_i(k) \)).

But by using an equivalent of (46) for instant \( k \) instead of \( k+1 \), we have

\[ \text{dist}(x_j(k), X) = \| \tilde{x}_j(k) + R_j(k-1) - \Pi_X[\tilde{x}_j(k) + R_j(k-1)] \|
\leq \| \tilde{x}_j(k) - \Pi_X[\tilde{x}_j(k)] + R_j(k-1) \|
\leq \text{dist}(\tilde{x}_j(k), X) + \| R_j(k-1) \|. \]
(39)

But in our algorithm we have,

\[ v_i(k) = \sum_{j=1}^n [W(k)]_{ij} x_j(k), \]
(40)

Then from the convexity of the norm squared, we have

\[ \text{dist}^2(v_i(k), X) \leq \sum_{j=1}^n [W(k)]_{ij} \text{dist}^2(x_j(k), X). \]
(41)

But

\[ (1 - a + (A_\eta - aA_\tau)\alpha_k^2) \text{dist}^2(v_i(k), X) = \]
\[ (1 + (A_\eta - aA_\tau)\alpha_k^2) \text{dist}^2(v_i(k), X) \]
\[ + (1 - a) \text{dist}^2(v_i(k), X), \]
(42)

Then by using (42) and (41), (37) becomes
\( \text{dist}^2(\mathbf{x}_i(k+1), \mathcal{X}) \leq \)

\[
(1 + (A_\eta - aA_\tau)\alpha_k^2) \sum_{i=1}^{n} [\mathbf{W}(k)]_{ij} \text{dist}^2(\mathbf{x}_j(k), \mathcal{X})
\]

\[
+ (1 - a) \text{dist}^2(\mathbf{v}_i(k), \mathcal{X}) + (B_\eta + aB_\tau + 2)\alpha_k^2 G_f^2
\]

\[
+ \left( \frac{3}{8\eta} - \frac{3a}{8\tau} + 2\alpha_k^2 L^2 + (2 - 2a)\alpha_k L \right) \text{dist}^2(\mathbf{v}_i(k), \mathcal{X})
\]

(43)

where we used \( \| \nabla f^{(i)}(\bar{x}) \| \leq G_f \) (i.e., gradient is bounded on set \( \mathcal{X} \)).

Let \( \mathcal{F}_k \) be the \( \sigma \)-algebra generated by the entire history of the algorithm up to time \( k \) inclusively, that is \( \mathcal{F}_k = \{ \mathbf{x}_i(0), i \in V \} \cup \{ \Omega_\tau(l) : 0 \leq l \leq k, i \in V \} \). Therefore, given \( \mathcal{F}_k \), the collection \( \mathbf{x}_i(0), \ldots, \mathbf{x}_i(k+1) \) and \( \mathbf{v}_i(0), \ldots, \mathbf{v}_i(k+1) \) generated by the algorithm is fully determined. So by applying expectation on (43), and taking consideration the probability \( 0 \leq \gamma_{(i)} \leq 1 \) of accessing a partition per each server, we have for \( k > 0 \), a.s. that

\[
\mathbb{E}[\text{dist}^2(\mathbf{x}_i(k+1), \mathcal{X})/\mathcal{F}_k] \leq \]

\[
\sum_{(i)=0}^{p} \gamma_{(i)}(1 + (A_\eta - aA_\tau)\alpha_k^2) \sum_{j=1}^{n} [\mathbf{W}(k)]_{ij} \text{dist}^2(\mathbf{x}_j(k), \mathcal{X})
\]

\[
+ \sum_{(i)=0}^{p} \gamma_{(i)}(1 - a) \text{dist}^2(\mathbf{v}_i(k), \mathcal{X}) + \sum_{(i)=1}^{p} \gamma_{(i)}(B_\eta - aB_\tau + 2)\alpha_k^2 G_f^2
\]

\[
+ \sum_{(i)=1}^{p} \gamma_{(i)}\left( \frac{3}{8\eta} - \frac{3a}{8\tau} + 2\alpha_k^2 L^2 + (2 - 2a)\alpha_k L \right) \text{dist}^2(\mathbf{v}_i(k), \mathcal{X})
\]

(44)

Then by using \( 0 \leq \gamma_{\text{min}} \leq \gamma_{(i)} \leq 1 \) for partitions \( (i) \in \{1, \ldots, p\} \) and choosing \( 0 < a < 1 \) and the fact that \( \sum_{(i)=1}^{p} \gamma_{(i)} \leq \sum_{(i)=0}^{p} \gamma_{(i)} = 1 \), the above reduces to

\[
\mathbb{E}[\text{dist}^2(\mathbf{x}_i(k+1), \mathcal{X})/\mathcal{F}_k] \leq \]

\[
(1 + (A_\eta - aA_\tau)\alpha_k^2) \sum_{j=1}^{n} [\mathbf{W}(k)]_{ij} \text{dist}^2(\mathbf{x}_j(k), \mathcal{X})
\]

\[
+ (1 - a) \text{dist}^2(\mathbf{v}_i(k), \mathcal{X})
\]

\[
+ \left( \frac{3}{8\eta} - \sum_{(i)=1}^{p} \frac{3\gamma_{(i)}a}{8\tau} + 2\alpha_k^2 L^2 + (2 - 2a)\alpha_k L \right) \text{dist}^2(\mathbf{v}_i(k), \mathcal{X})
\]

\[
(B_\eta - aB_\tau + 2)\alpha_k^2 G_f^2.
\]

(45)

Then summing overall \( n \), and having that \( [\mathbf{W}(k)]_{ij} \) doubly stochastic, we have a.s. that
$$\mathbb{E}\left[\sum_{i=1}^{n} \text{dist}^2(\mathbf{x}_i(k+1), \mathcal{X}) / \mathcal{F}_k \right] \leq$$

$$(1 + (A_\eta - aA_\tau)\alpha_k^2) \sum_{i=1}^{n} \text{dist}^2(\mathbf{x}_i(k), \mathcal{X})$$

$$+ \left( \frac{3}{8}\eta - \sum_{(i=1)}^{n} \frac{3\gamma^{(i)}a}{8\tau} + 2\alpha_k^2L^2 + (2 - 2a)\alpha_k L \right) \sum_{i=1}^{n} \text{dist}^2(\mathbf{v}_i(k), \mathcal{X})$$

$$+ n(B_\eta - aB_\tau + 2)\alpha_k^2G_f^2 + (1 - a) \text{dist}^2 \sum_{i=1}^{n} (\mathbf{v}_i(k), \mathcal{X})$$

From (10) and (5) substitution we can write

$$\mathbf{x}_i(k+1) = \mathbf{x}_i(k+1) + R_i(k),$$

where

$$\mathbf{x}_i(k+1) = \mathbf{v}_i(k) - \alpha_k \nabla f^{(i)}(\mathbf{v}_i(k))$$

(cf. Definition 1 for the definition of $R_i(k)$).

But by using an equivalent of (46) for instant $k$ instead of $k + 1$, we have

$$\text{dist}(\mathbf{x}_j(k), \mathcal{X}) = \| \mathbf{x}_j(k) + R_j(k-1) - \Pi_X[\mathbf{x}_j(k) + R_j(k-1)] \|

\leq \| \mathbf{x}_j(k) - \Pi_X[\mathbf{x}_j(k)] + R_j(k-1) \|

\leq \text{dist}(\mathbf{x}_j(k), \mathcal{X}) + \| R_j(k-1) \|.$$  

Using $2ab \leq a^2 + b^2$ and squaring both sides, the above becomes

$$\text{dist}^2(\mathbf{x}_j(k), \mathcal{X}) \leq 2\text{dist}^2(\mathbf{x}_j(k), \mathcal{X}) + 2\| R_j(k-1) \|^2. \quad (48)$$

**Applying the Supermartingale Theorem in Lemma 4** Then using the result of (48) and writing the above in the format of the supermartingale theorem, we see that

$$\mathbb{E}\left[\sum_{i=1}^{n} \text{dist}^2(\mathbf{x}_i(k+1), \mathcal{X}) / \mathcal{F}_k \right] \leq$$

$$(1 + (A_\eta - aA_\tau)\alpha_k^2) \sum_{i=1}^{n} \text{dist}^2(\mathbf{x}_i(k), \mathcal{X})$$

$$+ (1 - a + \frac{3}{8}\eta - \frac{3\gamma_{\min}a\eta}{8\tau} + 2\alpha_k^2L^2 + (2 - 2a)\alpha_k L) \sum_{i=1}^{n} \text{dist}^2(\mathbf{v}_i(k), \mathcal{X})$$

$$+ n(B_\eta - aB_\tau + 2)\alpha_k^2G_f^2$$

$$+ (1 + (A_\eta - aA_\tau)\alpha_k^2) \sum_{i=1}^{n} \| R_i(k-1) \|^2 \text{ a.s.} \quad (49)$$
Then by using the result of (146) which implies the boundedness of $\sum_{i=1}^{n} \|R_i(k-1)\|^2$ for $k \geq \bar{k}$, we can apply the supermartingale convergence theorem on (49) with the following substitutions: $v_{k+1} = \sum_{i=1}^{n} \text{dist}^2(\kappa_i(k) + 1, \mathcal{X})$, i.e., $v_k = \sum_{i=1}^{n} \text{dist}^2(\kappa_i(k), \mathcal{X})$, $a_k = (A_\eta - aA_\tau) \alpha_k^2$ and $A_\eta - aA_\tau > 0$, $b_k = n(B_\eta - aB_\tau + 2) \alpha_k^2 G + (1 + (A_\eta - aA_\tau) \alpha_k^2) \sum_{i=1}^{n} \|R_i(k-1)\|^2$ where $\sum_{k=0}^{\infty} a_k < \infty$ and $\sum_{k=n}^{\infty} b_k < \infty$ since $\sum_{k=0}^{\infty} \alpha_k^2 < \infty$ and $u_k = \sum_{i=1}^{n} \text{dist}^2(\nu_i, \mathcal{X})$, i.e., for $b_k$ notice that $\sum_{i=1}^{n} \|R_i(k-1)\|^2$ is bounded since $\sum_{i=1}^{n} \|v_i(k-1) - x^*\|^2 < D < \infty$ for every $k > \bar{k}$ from the Appendix (C) (i.e., Proposition [3] or [4]).

But $u_k = \sum_{i=1}^{n} \text{dist}^2(\nu_i(k), \mathcal{X}) \leq \sum_{i=1}^{n} \|v_i(k) - x^*\|^2 < D < \infty$ for $k \geq \bar{k}$, (i.e., Proposition 3 or 4), and its coefficient is $(1 - a + \frac{3}{8} - \frac{3\gamma_{\min}p}{8\tau} + 2\alpha_k^2 L^2 + (2 - 2a)\alpha_k L) < -1$ for $k \geq k_1$ in order to apply the supermartingale theorem.

Remark 7. But $\alpha_k \to 0$ (since $\alpha_k$ is chosen such that $\sum_{k=0}^{\infty} \alpha_k^2 < \infty$). Thus, we can bound $2\alpha_k^2 L^2 + (2 - 2a)\alpha_k L < \epsilon$ as $\alpha_k \to 0$. By 2. $2 + \frac{3}{8} + \epsilon < a + \frac{3\gamma_{\min}p}{8\tau}$ Thus, taking $\epsilon < 1$ and $\eta > \tau$ where $\eta = l\tau$ and $l > 1$ such that what preceded applies. Then $a > \frac{1}{\gamma_{\min}p + 1}$ where we can choose $l$ such that $a < 1$. Therefore, by reducing this inequality with the conditions on the values of the above variables we have $8l\tau < 3(\gamma_{\min}p - 1)$ where we get a sufficient condition for convergence which is that $\gamma_{\min}p l > 1$, but $l > \frac{1}{\gamma_{\min}p} > 1$, then a sufficient condition is for $p < \frac{1}{\gamma_{\min}}$.

E.g. for $l = 3$, $\gamma_{\min}p = 2$ we have $\tau < \frac{1}{3}$, so we can choose $\tau = \frac{1}{4}$, $\eta = \frac{3}{4}$ and $1 > a > \frac{3}{8}$. Thus, choosing an $a = \frac{3}{4}$ is sufficient for this Lemma 1 to follow.

Thus, the coefficient of $u_k$ is negative for $k \geq k_1$.

Then from the supermartingale theorem holding for the tail of the sequences (i.e., $k > \max(k_1, \bar{k})$) we have $\sum_{k=0}^{\infty} u_k < \infty$. That is, $\sum_{k=0}^{\infty} \sum_{i=1}^{n} \text{dist}^2(\nu_i(k), \mathcal{X}) < \infty$. We can interchange infinite and finite sums, as an implicit consequence of the linearity of these sums. Thus, we have $\sum_{i=1}^{n} \sum_{k=0}^{\infty} \text{dist}^2(\nu_i(k), \mathcal{X}) < \infty \implies$ the argument inside the finite sum is bounded, i.e.,

$$\sum_{k=0}^{\infty} \text{dist}^2(\nu_i(k), \mathcal{X}) < \infty,$$

the result we require. And $\lim_{k \to \infty} \text{dist}^2(\nu_i(k), \mathcal{X}) = 0$. \hfill $\square$

Lemma 5. Let Assumptions [7] hold and Proposition [3] or [4] satisfied. Also, assume that the stepsize sequence $\{\alpha_k\}$ is non-increasing such that $\sum_{k=0}^{\infty} \alpha_k^2 < \infty$, and define $\epsilon_i = x_i(k+1) - v_i(k)$ for all $i \in V$ and $k \geq 0$. Then, we have a.s.

$$\sum_{k=0}^{\infty} ||\epsilon_i(k)||^2 < \infty \text{ for all } i \in V,$$

$$\sum_{k=0}^{\infty} \alpha_k ||v_i(k) - \nabla(k)|| < \infty \text{ for all } i \in V,$$

where $\nabla(k) = \frac{1}{n} \sum_{l=1}^{n} v_i(l)$.

Proof. Let
\[ \epsilon_i(k) = x_i(k + 1) - v_i(k) \]
\[ = x_i(k + 1) + R_i(k) - v_i(k) \]
\[ = v_i(k) - \alpha_k \nabla f^{(i)}(v_i(k)) + R_i(k) - v_i(k) \]
\[ = -\alpha_k \nabla f^{(i)}(v_i(k)) + R_i(k). \]

(52)

Notice that in our case, \( f^{(i)} \) is not fixed and can vary on the same node \( i \) and this also complies with \( \epsilon_i \) in lemma 6 in [8] which is an arbitrary error at node \( i \).

And, let \( z_i(k) = \Pi_X[v_i(k)] \) then we have

\[
\| \epsilon_i(k) \| \leq ||x_i(k + 1) - z_i(k)|| + ||z_i(k) - v_i(k)|| \\
\leq ||v_i(k) - \alpha_k \nabla f^{(i)}(v_i(k)) + R_i(k) - z_i(k)|| + ||z_i(k) - v_i(k)|| \\
\leq 2 ||v_i(k) - z_i(k)|| + \alpha_k ||\nabla f^{(i)}(v_i(k))|| + ||R_i(k)|| \\
\leq 2 ||v_i(k) - z_i(k)|| + \alpha_k ||\nabla f^{(i)}(z_i(k))|| \\
+ \alpha_k ||\nabla f^{(i)}(z_i(k)) - \nabla f^{(i)}(v_i(k))|| + ||R_i(k)||.
\]

But \( ||\nabla f^{(i)}(z_i(k))|| \leq G_f \) since \( z = \Pi_X(v_i(k)) \in \mathcal{X}, \)

and \( ||\nabla f^{(i)}(z_i(k)) - \nabla f^{(i)}(v_i(k))|| \leq L ||v_i(k) - z_i(k)||. \)

Therefore, (53) is reduced to

\[
\| \epsilon_i(k) \| \leq (2 + \alpha_k L)\|v_i(k) - z_i(k)\| + \alpha_k G_f \\
\alpha_k^2 \|A^{(i)}\|_\infty^2 \|B^{(i)}\|_\infty^2 \sum_{k \geq 1} 4n L^2 \max_{k - H \leq k \leq k - 1} \|v_q(\hat{k}) - x^*\|^2
\]

From \((a + b + c)^2 \leq 3a^2 + 3b^2 + 3c^2\), we get

\[
\| \epsilon_i(k) \|^2 \leq 3(2 + \alpha_k L)^2 \|v_i(k) - z_i(k)\|^2 + 3\alpha_k^2 G_f^2 \\
+ 48 \alpha_k^4 \|A^{(i)}\|_\infty^4 \|B^{(i)}\|_\infty^4 \sum_{k \geq 1} n^2 L^4 \max_{k - H \leq k \leq k - 1} \|v_q(\hat{k}) - x^*\|^4
\]

\[
\sum_{k=0}^{\infty} \| \epsilon_i(k) \|^2 \leq 3 \sum_{k=0}^{\infty} (4 + \alpha_k^2 L^2) \|v_i(k) - z_i(k)\|^2 \\
+ 12 \sum_{k=0}^{\infty} \alpha_k L \|v_i(k) - z_i(k)\|^2 + 3 \sum_{k=0}^{\infty} \alpha_k^2 G_f^2 \\
+ 48 \sum_{k=0}^{\infty} \alpha_k^4 \|A^{(i)}\|_\infty^4 \|B^{(i)}\|_\infty^4 L^4 \max_{k - H \leq k \leq k - 1} \|v_q(\hat{k}) - x^*\|^4.
\]

But, we also have from Lemma 2

\[
\sum_{k=0}^{\infty} \|v_i(k) - z_i(k)\|^2 = \sum_{k=0}^{\infty} \|v_i(k) - \Pi_X(v_i(k))\|^2 \\
= \sum_{k=0}^{\infty} dist^2(v_i(k), \mathcal{X}) < \infty,
\]

(55)
and
\[ \sum_{k=0}^{\infty} \alpha_k \|v_i(k) - z_i(k)\|^2 = \sum_{k=0}^{\infty} \alpha_k \text{dist}^2(v_i(k), \mathcal{X}) \]
\[ = \frac{1}{2} \sum_{k=0}^{\infty} \alpha_k^2 + \frac{1}{2} \sum_{k=0}^{\infty} \text{dist}^4(v_i(k), \mathcal{X}) < \infty, \]  
(56)
and since \( \sum_{k=0}^{\infty} \alpha_k^2 < \infty \) which also implies that \( \sum_{k=0}^{\infty} \alpha_k^4 < \infty \). And \( \sum_{i=1}^{n} \|x_i(k)\|^2 \) is bounded since \( \|v_i(k) - x_i\|^2 \) is bounded (i.e., Proposition 3 or 4).

From the preceding four assumptions we have that in (54),
\[ \sum_{k=0}^{\infty} \|\epsilon_i(k)\|^2 < \infty. \]  
(57)

By applying \( 2ab \leq a^2 + b^2 \) on \( \sum_{k=0}^{\infty} \alpha_k^2 < \infty \) and \( \sum_{k=0}^{\infty} \|\epsilon_i(k)\|^2 < \infty \), we have
\[ \sum_{k=0}^{\infty} \alpha_k \|\epsilon_i(k)\| \leq \frac{1}{2} \sum_{k=0}^{\infty} \alpha_k^2 + \frac{1}{2} \sum_{k=0}^{\infty} \|\epsilon_i(k)\|^2 < \infty. \]  
(58)

But \( x_i(k+1) = v_i(k) + \epsilon_i(k) \) and \( v_i(k) = \sum_{j=1}^{n} [W(k)]_{ij} x_j(k) \) where \( \sum_{j=1}^{n} [W(k)]_{ij} = 1 \) and \( \sum_{k=0}^{\infty} \alpha_k \|\epsilon_i(k)\| < \infty \) a.s.

Therefore, by Lemma 11 this implies that \( \sum_{k=0}^{\infty} \alpha_k \|x_i(k) - x_j(k)\| < \infty \)

Now, let \( \bar{v}(k) = \frac{1}{n} \sum_{i=1}^{n} v_i(k) \).

Since \( v_i(k) = \sum_{j=1}^{n} [W(k)]_{ij} x_j(k) \) where \( \sum_{j=1}^{n} [W(k)]_{ij} = 1 \) implies that
\[ \|v_i(k) - \bar{v}(k)\| = \| \sum_{j=1}^{n} [W(k)]_{ij} x_j(k) - \sum_{j=1}^{n} [W(k)]_{ij} \bar{v}(k) \| \]
\[ \leq \sum_{j=1}^{n} [W(k)]_{ij} \|x_j(k) - \bar{v}(k)\|, \]  
(59)
where the inequality follows by the convexity of the norm. But, \([W(k)] \) is doubly stochastic, so we have \( \bar{v}(k) = \frac{1}{n} \sum_{i=1}^{n} v_i(k) = \frac{1}{n} \sum_{i=1}^{n} \sum_{j=1}^{n} [W(k)]_{ij} x_j(k) = \frac{1}{n} \sum_{i=1}^{n} x_i(k) \).

\[ \|v_i(k) - \bar{v}(k)\| \leq \sum_{j=1}^{n} [W(k)]_{ij} \|x_j(k) - \bar{v}(k)\| \]
\[ \leq \sum_{j=1}^{n} \|x_j(k) - \bar{v}(k)\| \]
\[ \leq \sum_{j=1}^{n} \|x_j(k) - \frac{1}{n} \sum_{i=1}^{n} x_i(k)\|, \]  
(60)
where in the first inequality we used convexity of the norm, in the second 0 ≤ [W(k)]_{ij} ≤ 1 and in the third we substituted the preceding result on \( \bar{v}(k) \). Therefore,
\[
\|v_i(k) - \bar{v}(k)\| \leq \frac{n}{\gamma} \sum_{j=1}^{\infty} \|x_j(k) - \frac{1}{n} \sum_{i=1}^{n} x_i(k)\| \leq \frac{1}{n} \sum_{j=1}^{\infty} \sum_{i=1}^{n} \|x_j(k) - x_i(k)\| ,
\]
where by the convexity of norm we have the last inequality. Thus, we have
\[
\alpha_k \|v_i(k) - \bar{v}(k)\| \leq \frac{\alpha_k}{\gamma} \sum_{j=1}^{\infty} \sum_{i=1}^{n} \|x_j(k) - x_i(k)\| .
\]

Then
\[
\sum_{k=0}^{\infty} \alpha_k \|v_i(k) - \bar{v}(k)\| \leq \sum_{k=0}^{\infty} \frac{\alpha_k}{\gamma} \sum_{j=1}^{\infty} \sum_{i=1}^{n} \|x_j(k) - x_i(k)\| \leq \frac{1}{n} \sum_{j=1}^{\infty} \sum_{i=1}^{n} \sum_{k=0}^{\infty} \alpha_k \|x_j(k) - x_i(k)\|. 
\]
The second inequality is valid since we can interchange infinite sum with finite sum. But, from a previous result (Lemma[11], we have \( \sum_{k=0}^{\infty} \alpha_k \|x_i(k) - x_j(k)\| < \infty \) which through the preceding, \( \sum_{k=0}^{\infty} \alpha_k \|v_i(k) - \bar{v}(k)\| < \infty \).

But \( \sum_{k=0}^{\infty} \alpha_k = \infty \) by choice. \( \Rightarrow \|v_i(k) - \bar{v}(k)\| \to 0 \), i.e., \( \lim_{k \to \infty} v_i(k) = \lim_{k \to \infty} \bar{v}(k) \) if they exist.

Therefore, Lemma[5] follows. 

Lemma 6. Let Assumptions[7] hold and Proposition[3] or[4] satisfied. Let \( p < \frac{1}{\gamma_{\min}} \) and let the step-size be such that \( \sum_{k=0}^{\infty} \alpha_k = \infty \) and \( \sum_{k=0}^{\infty} \alpha_k^2 < \infty \). Let \( f^* = \min_{x \in \mathcal{X}} f(x) \) and \( \mathcal{X}^* = \{ x \in \mathcal{X} | f(x) = f^* \} \). Assume then that \( \mathcal{X}^* \neq \Phi \). Then, the iterates \( \{x(k)\} \) generated by SRDO algorithm (3a)-(3c) converge almost surely to the solution \( x^* \in \mathcal{X}^* \), i.e.,
\[
\lim_{k \to \infty} x_i(k) = x^* \quad \text{for all} \quad i \in V \ a.s.
\]

Proof. We begin the proof of Lemma[6] with the following.

For any \( x^* \in \mathcal{X} \), we have
\[
\Rightarrow \|x_i(k+1) - x^*\|^2 \leq \|x_i(k+1) - x^*\|^2 + a\|\Pi_{\mathcal{X}}(x_i(k+1)) - x^*\|^2 .
\]
where \( a > 0 \).

And we have a similar inequality as (54)
\[
\Rightarrow \|\Pi_{\mathcal{X}}(x_i(k+1) - x_i(k+1))\|^2 \leq \|x_i(k+1) - x^*\|^2 - a\|\Pi_{\mathcal{X}}(x_i(k+1)) - x^*\|^2 ,
\]

(64)
where $a \leq 1$.

Using Lemma 9 and 10 with the following substitutions $y = x_i(k + 1), x = v_i(k)$, and $x^* = argmin f(x) \in X$ for $Y = X$, $\alpha = \alpha_k$ and $\Phi = f$ and $z_i(k) = \Pi_X(v_i(k)) \in X$, equation (64) where $Y = X$, we get the following equivalent of (35).

\[
\|x_i(k + 1) - x^*\|^2 \leq \\
(1 - a + (A_\eta - aA_\tau)\alpha_k^2)\|v_i(k) - x^*\|^2 \\
- 2(1 - a)\alpha_k(f^{(i)}(z_i(k)) - f^{(i)}(x^*)) \\
+ \frac{3a}{4}\|\Pi_X[v_i(k) - \alpha_k \nabla f(v_i(k))] - v_i(k)\|^2 \\
+ \left(\frac{3}{8\eta} - \frac{3a}{8\tau} + (2 - 2a)\alpha_k L\right)\|v_i(k) - z_i(k)\|^2 \\
+ (B_\eta - aB_\tau)\alpha_k^2\|\nabla f^{(i)}(x^*)\|^2,
\]

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

But we know that $v_i(k) = \sum_{j=1}^{n}[W(k)]_{ij}x_j(k)$ then by the convexity of the norm squared and double stochasticity of $[W(k)]_{ij}$, we have by summing from $i = 1$ to $n$,

\[
\sum_{i=1}^{n}\|v_i(k) - x^*\|^2 \leq \sum_{i=1}^{n}\sum_{j=1}^{n}[W(k)]_{ij}\|x_j(k) - x^*\|^2 \\
\leq \sum_{i=1}^{n}\|x_j(k) - x^*\|^2.
\]

Following a similar analysis as in the Appendix and taking $x^i = \min f^{(i)}(x)$ for each $i \in \{1, \ldots, p\}$. Having for $i \in I$ that $f^{(i)}(x^*) > f^{(i)}(x^i)$ and for $i \in I^C$ that $f^{(i)}(x^*) \leq f^{(i)}(x^i)$ and for $\bar{z}(k) = \frac{1}{n}\sum_{i=1}^{n}z_i(k)$, where $z_i(k) = \Pi_X(v_i(k)) \in X$, we have

\[
\sum_{(i=1)}^{p}\gamma(i)(f^{(i)}(z_i(k)) - f^{(i)}(x^*)) \\
\geq \sum_{(i=1)}^{p}\gamma(i)(f(z_i(k)) - f^{(i)}(\bar{z}(k))) + \gamma_{\min}(f(\bar{z}(k)) - f(x^*)) \\
+ \sum_{(i) \in I}\gamma(i)(\nabla f^{(i)}(x^*), x^{(i)} - x^*).
\]

But

\[
\sum_{(i=1)}^{p}\gamma(i)(f^{(i)}(z_i(k)) - f^{(i)}(\bar{z}(k))) \\
\geq \sum_{i=1}^{p}\gamma(i)\langle \nabla f^{(i)}(\bar{z}(k)), z_i(k) - \bar{z}(k)\rangle \\
\geq -\sum_{(i=1)}^{p}\gamma(i)\|\nabla f^{(i)}(\bar{z}(k))\|\|z_i(k) - \bar{z}(k)\|.
\]
Let \( z_i(k) = \Pi_{\mathcal{X}}(v_i(k)) \in \mathcal{X} \). Since \( \bar{z}(k) \) is a convex combination of \( z_i(k) \in \mathcal{X} \implies \bar{z}(k) \in \mathcal{X} \) and thus \( \| \nabla f^{(i)}(\bar{z}(k)) \| \leq G_f \).

\[
\implies \sum_{(i)=1}^{p} \gamma(i)(f(z_i(k)) - f(\bar{z}(k))) \geq -G_f \sum_{(i)=1}^{p} \gamma(i)\|z_i(k) - \bar{z}(k)\|. \tag{69}
\]

But

\[
\|z_i(k) - \bar{z}(k)\| = \frac{1}{n} \sum_{j=1}^{n} (z_i(k) - z_j(k))
\]

\[
\leq \frac{1}{n} \sum_{j=1}^{n} \|z_i(k) - z_j(k)\| \leq \frac{1}{n} \sum_{j=1}^{n} \|\Pi_{\mathcal{X}}(v_i(k)) - \Pi_{\mathcal{X}}(v_j(k))\|
\]

\[
\leq \frac{1}{n} \sum_{j=1}^{n} \|v_i(k) - v_j(k)\|
\]

where the first inequality follows from the convexity of the norm and the last inequality follows from the non-expansiveness of the projection \( \Pi \).

But by the triangle inequality, we have \( \|v_i(k) - v_l(k)\| \leq \|v_i(k) - \bar{v}(k)\| + \|v_l(k) - \bar{v}(k)\| \),

Thus,

\[
\|z_i(k) - \bar{z}(k)\| \leq \frac{1}{n} \sum_{i=1}^{n} \|v_i(k) - v_l(k)\|
\]

\[
\leq \frac{1}{n} \sum_{i=1}^{n} \|v_i(k) - \bar{v}(k)\| + \frac{1}{n} \sum_{i=1}^{n} \|v_l(k) - \bar{v}(k)\|
\]

\[
\leq \frac{n}{n} \|v_i(k) - \bar{v}(k)\| + \frac{1}{n} \sum_{i=1}^{n} \|v_l(k) - \bar{v}(k)\|
\]

\[
\implies \|z_i(k) - \bar{z}(k)\| \leq \|v_i(k) - \bar{v}(k)\| + \frac{1}{n} \sum_{i=1}^{n} \|v_l(k) - \bar{v}(k)\|. \tag{70}
\]

Then by summing over \( i \), we get

\[
\sum_{i=1}^{n} \|z_i(k) - \bar{z}(k)\| \leq \sum_{i=1}^{n} \|v_i(k) - \bar{v}(k)\| + \frac{1}{n} \sum_{i=1}^{n} \sum_{l=1}^{n} \|v_l(k) - \bar{v}(k)\|
\]

\[
\leq \sum_{i=1}^{n} \|v_i(k) - \bar{v}(k)\| + \frac{n}{n} \sum_{l=1}^{n} \|v_l(k) - \bar{v}(k)\|
\]

\[
\leq 2 \sum_{i=1}^{n} \|v_i(k) - \bar{v}(k)\|	ext{,}
\]

which follows since indices \( i \) and \( l \) are arbitrary indexes.

Then, by using \( (67) \), substituting \( (69) \) and \( (70) \) and the Lipschitz continuity of the gradients accordingly, we have
\[ \sum_{(i)=1}^{p} \gamma(i)(f^{(i)}(z(k)) - f^{(i)}(x^*)) \geq \sum_{(i)=1}^{p} \gamma(i)(f^{(i)}(z_i(k)) - f^{(i)}(z(k))) + \gamma_{\min}(f(z(k)) - f(x^*)) + \sum_{(i) \in I} \gamma(i)\langle \nabla f^{(i)}(x^*), x^{(i)} - x^* \rangle \]

\[ \geq -G_f \sum_{(i)=1}^{p} \gamma(i)\|z_i(k) - z(k)\| + \gamma_{\min}(f(z(k)) - f(x^*)) \]

\[ - \sum_{(i) \in I} \gamma(i)\|\nabla f^{(i)}(x^*)\||x^{(i)} - x^*\| \]

\[ \geq -2G_f \sum_{(i)=1}^{p} \gamma(i)\|v_i(k) - \bar{v}(k)\| + \gamma_{\min}(f(\bar{z}(k)) - f(x^*)) \]

\[ - \sum_{(i) \in I} \gamma(i)L\|x^{(i)} - x^*\|^2. \]

Then using (65) and taking the expectation on the history \( \mathcal{F}_k \) up to \( k \), we have after the consideration the probability \( 0 \leq \gamma_{\min} \leq \gamma(i) \leq 1 \) of accessing a partition \( (i) \in \{1, \ldots, p\} \) per each server, \( 0 < a < 1 \) and \( \sum_{(i)=1}^{p} \leq \sum_{(i)=0}^{p} = 1 \) that

\[ \mathbb{E}[\|x_i(k + 1) - x^*\|^2 / \mathcal{F}_k] \leq \]

\[ \sum_{(i)=0}^{p} \gamma(i)(1 + (A_{\eta} - aA_r)\alpha_k^2 \sum_{j=1}^{n}[W(k)]_{ij}\|x_i(k) - x^*\|^2 \]

\[ + \sum_{(i)=0}^{p} \gamma(i)(1 - a)\|v_i(k) - x^*\|^2 \]

\[ - 2\sum_{(i)=1}^{p} \gamma(i)(1 - a)\alpha_k(f^{(i)}(z_i(k)) - f^{(i)}(x^*)) \]

\[ + \sum_{(i)=1}^{p} \gamma(i)\frac{3a}{4}\|\Pi_x[v_i(k) - \alpha_k \nabla f(v_i(k))] - v_i(k)\|^2 \]

\[ + \sum_{(i)=1}^{p} \gamma(i)\left( \frac{3}{8R} - \frac{3a}{8T} + (2 - 2a)\alpha_kL\|v_i(k) - z\| \right) \]

\[ + \sum_{(i)=1}^{p} \gamma(i)(B_{\eta} - aB_r)\alpha_k^2\|\nabla f^{(i)}(x^*)\|^2 \ a.s. \]

But

\[ \|x_i(k) - x^*\| = \|x_i(k) + R_i(k - 1) - x^*\|. \]

And by squaring the norm and applying \( 2ab \leq a^2 + b^2 \), we have

\[ \|x_i(k) - x^*\|^2 \leq 2\|x_i(k) - x^*\|^2 + 2\|R_i(k - 1)\|^2. \]
Then through summing over all $i$ and having $[W]_{ij}$ doubly stochastic, the substitution of (71) and having $0 \leq \gamma_{min} \leq \gamma_{(i)} \leq 1$ for $(i) \in \{1, \ldots, p\}$ and $0 < a < 1$, (72) becomes

$$
E\left[\sum_{i=1}^{n} \|x^{(k+1)} - x^*\|^2 / F_k\right] \leq
$$

$$(1 + (\eta - aA_{\tau})\alpha_k^2) \sum_{i=1}^{n} \|x^{(k)} - x^*\|^2$$

$$+ (1 - a) \sum_{i=1}^{n} \|v_i(k) - x^*\|^2$$

$$+ 2(1 - a) \sum_{i=1}^{n} \|x^{(k)} - x^*\|^2$$

$$+ 4(1 - a)\alpha_k G_f \sum_{i=1}^{n} \|v_i(k) - x^*\|^2$$

$$- 2\gamma_{min} n(1 - a)\alpha_k \left(f(\bar{z}(k)) - f(x^*)\right) + \left(B_\eta - aB_\tau + 2\right)\alpha_k^2 G_f^2$$

$$+ \left(\frac{3}{8\eta} - \frac{3\gamma_{min} ap}{8\tau} + 2\alpha_k^2 L^2 + (2 - 2a)\alpha_k L\right) \sum_{i=1}^{n} \text{dist}^2(v_i(k), \mathcal{X})$$

$$+ (1 + (\eta - aA_{\tau})\alpha_k^2) \sum_{i=1}^{n} \|x^{(k)} - x^*\|^2$$

But we have

$$
\left(\frac{3}{8\eta} - \frac{3\gamma_{min} ap}{8\tau} + 2\alpha_k^2 L^2 + (2 - 2a)\alpha_k L\right) \text{dist}^2(v_i(k), \mathcal{X})$$

$$+ (1 - a) \sum_{i=1}^{n} \|v_i(k) - x^*\|^2 \leq$$

$$\left(1 - a + \frac{3}{8\eta} - \frac{3\gamma_{min} ap}{8\tau} + 2\alpha_k^2 L^2 + (2 - 2a)\alpha_k L\right) \sum_{i=1}^{n} \|v_i(k) - x^*\|^2 \leq$$

$$2(1 - a + \frac{3}{8\eta} - \frac{3\gamma_{min} ap}{8\tau} + 2\alpha_k^2 L^2 + (2 - 2a)\alpha_k L) \sum_{i=1}^{n} \|v_i(k) - x^{(i)}\|^2$$

$$2(1 - a + \frac{3}{8\eta} - \frac{3\gamma_{min} ap}{8\tau} + 2\alpha_k^2 L^2 + (2 - 2a)\alpha_k L) n \|x^{(i)} - x^*\|^2.$$
Introducing the same analysis as in Lemma 4 we have a sufficient condition for 
the terms containing $i \gamma$ that 
sum_{i=1}^{n} \|x_i(k) - x^*\|^2 + 4(1-a)\alpha_k G_f \sum_{i=1}^{n} \|v_i(k) - \bar{v}(k)\| 
- 2\gamma_{min}(1-a)\alpha_k (f(\bar{z}(k)) - f(x^*)) + n (B_\eta - aB_r + 2)\alpha_k^2 G_f^2 
+ 2(1-a + \frac{3}{8\eta} - \frac{3\gamma_{min}ap}{8\tau} + 2\alpha_k^2 L^2 + (2 - 2a)\alpha_k L) \sum_{i=1}^{n} \min_{(i)} \|v_i(k) - x^{(i)}\|^2 
+ 2n(1-a + \frac{3}{8\eta} - \frac{3\gamma_{min}ap}{8\tau} + 2\alpha_k^2 L^2 
+ (1-a)(2 + \min(|I|\gamma_{max},1))\alpha_k L) \max_{(i)} \|x^{(i)} - x^*\|^2 
(1 + (A_\eta - aA_r)\alpha_k^2)\|R_f(k-1)\|^2 \ a.s.

Since $\|x^* - x^*\|^2$ is bounded, (i.e., Proposition 5 or 4), and $\|v_i(k) - x^*\|^2$ is bounded then $\|v_i(k) - x^*\|^2$ is bounded, we need $1 + \frac{3}{8\eta} + \epsilon < \frac{3\gamma_{min}ap}{8\tau} + a$. Then for $\eta > \tau$ where $\eta = l\tau$ and $l > 1$ we get $a > \frac{8\eta + 3}{l(3\gamma_{min}p + 8\tau)}$. Introducing the same analysis as in Lemma 4 we have a sufficient condition for $p < \frac{1}{\gamma_{min}}$. Thus, for e.g. for $l = 8$, $\gamma_{min}p = 2$, $\eta = 4$ and $\tau = \frac{1}{2}$, we can take $1 > a > \frac{7}{18}$. Thus, by choosing $a = \frac{1}{2}$ then we have for $k > k_2$ that the terms containing $\|v_i(k) - x^*\|^2$ and $\|x^* - x^*\|^2$ to be negative.

And using the boundedness of $R_i(k-1)$ from the Appendix (77) (i.e., Proposition 3 or 4), becomes

$$
\mathbb{E}\left(\sum_{i=1}^{n} \|x_i(k) - x^*\|^2 / F_k\right) \leq (1 + (A_\eta - aA_r)\alpha_k^2) \sum_{i=1}^{n} \|x_i(k) - x^*\|^2 
+ (1 + (A_\eta - aA_r)\alpha_k^2)\alpha_k^{-1} nF + 4(1-a)\alpha_k G_f \sum_{i=1}^{n} \|v_i(k) - \bar{v}(k)\| 
- 2\gamma_{min}(1-a)\alpha_k (f(\bar{z}(k)) - f(x^*)) 
+ n (B_\eta - aB_r + 2)\alpha_k^2 G_f^2 \ a.s.
$$

But $\sum_{k=0}^{n} (A_\eta - aA_r)\alpha_k^2 < \infty$ since $\sum_{k=0}^{\infty} \alpha_k^2 < \infty$.

Also, $\sum_{k=0}^{\infty} n(B_\eta - aB_r)\alpha_k^2 G_f^2 < \infty$ similarly.

And, $\sum_{k=0}^{\infty} 4(1-a)\alpha_k G_f \sum_{i=1}^{n} \|v_i(k) - \bar{v}(k)\| < \infty$ from Lemma 6 holding. Therefore, the supermartingale theorem applies. Hence, the sequence $\{\|x_i(k) - x^*\|^2\}$ is convergent a.s. to a nonnegative random variable for any $i \in V$ and $x^* \in \mathcal{X}$ where $\mathcal{X}^* = \{x \in \mathcal{X} | f(x) = \min_{x \in \mathcal{X}} f(x)\}$. But $u_k = (1-a)\gamma_{min}\alpha_k (f(\bar{z}(k)) - f(x^*)) > 0$ with negative coefficient $-2 < -1$ since $0 < a = \frac{1}{2\gamma_{min}} < 1$ and $f(x^*) = \min f(x)$. And the theorem also implies that $\sum_{k=0}^{\infty} u_k = \sum_{k=0}^{\infty} (1-a)\gamma_{min}\alpha_k (f(\bar{z}(k)) - f(x^*)) \leq \infty$. This with the condition that $\sum_{k=0}^{\infty} \alpha_k = \infty$ and $f(\bar{z}(k)) - f(x^*)) \geq 0$ imply

$$
\lim_{k \to \infty} \inf_{k \to \infty} (f(\bar{z}(k)) - f(x^*)) = 0 \ a.s.
$$

(79)
And since \( f(\bar{z}(k)) - f(x^*) \geq 0 \) for all \( k \) since \( f(x^*) = \min f(x) \) then

\[
\lim_{k \to \infty} f(\bar{z}(k)) = f(x^*) \quad a.s.
\]  

(80)

By lemma 4, we have \( \sum_{k=0}^{\infty} \text{dist}^2(v_i(k), \mathcal{X}) < \infty \), i.e., \( \sum_{k=0}^{\infty} \|v_i(k) - z_i(k)\|^2 < \infty \) where \( z_i(k) = \Pi_{\mathcal{X}}(v_i(k)) \).

\[
\Rightarrow \lim_{k \to \infty} \|v_i(k) - z_i(k)\| \to 0.
\]  

(81)

But we have the sequence \( \{\|v_i(k) - x^*\|\} \) is also convergent a.s. for all \( i \in V \) and \( x^* \in \mathcal{X}^* \). By (81) it follows that \( \{\|z_i(k) - x^*\|\} \) is also convergent a.s for all \( i \in V \) and \( x^* \in \mathcal{X}^* \). But since \( \|\bar{v}(k) - x^*\| \leq \frac{1}{n} \sum_{i=1}^{n} \|v_i(k) - x^*\| \) and the sequence \( \{\|v_i(k) - x^*\|\} \) is convergent a.s for all \( i \in V \) and \( x^* \in \mathcal{X}^* \), it follows that \( \{\|\bar{v}(k) - x^*\|\} \) is convergent a.s for all \( x^* \in \mathcal{X}^* \). Using a similar argument, we can conclude that \( \{\|\bar{z}(k) - x^*\|\} \) is convergent a.s for all \( x^* \in \mathcal{X}^* \). As a particular consequence, it follows that the sequences \( \{\bar{v}(k)\} \) and \( \{\bar{z}(k)\} \) are a.s. bounded and, hence they have accumulation points.

From (80) and the continuity of \( f \), it follows that the sequence \( \bar{z}(k) \) must have one accumulation point in the set \( \mathcal{X}^* \) a.s. This and the fact that \( \{\|\bar{z}(k) - x^*\|\} \) is convergent a.s for every \( x^* \in \mathcal{X}^* \) imply that for a random point \( x^* \in \mathcal{X}^* \) (from (79))

\[
\lim_{k \to \infty} \bar{z}(k) = x^* \quad a.s.
\]  

(82)

Now, from \( \bar{z}(k) = \frac{1}{n} \sum_{i=1}^{n} z_i(k) \) and \( \bar{v}(k) = \frac{1}{n} \sum_{i=1}^{n} v_i(k) \) and using (81) \( \lim_{k \to \infty} \|v_i(k) - z_i(k)\| = 0 \) for all \( i \in V \) and the convexity of the norm, we obtain that

\[
\lim_{k \to \infty} \|\bar{v}(k) - \bar{z}(k)\| = \frac{1}{n} \sum_{i=1}^{n} \|v_i(k) - z_i(k)\| = 0 \quad a.s.
\]  

(83)

In view of (82), it follows that

\[
\lim_{k \to \infty} \bar{v}(k) = x^* \quad a.s.
\]  

(84)

By \( \sum_{k=0}^{\infty} \alpha_k \|v_i(k) - v(k)\| < \infty \) for all \( i \in V \) in Lemma 5 since \( \sum_{k=0}^{\infty} \alpha_k = \infty \), we have \( \lim_{k \to \infty} \inf \|v_i(k) - v(k)\| = 0 \) for all \( i \in V \) a.s.

This fact or the fact that \( \{\|v_i(k) - x^*\|\} \) is convergent a.s. for all \( i \in V \) together with the above limit equality (84), a consequence of (82), imply that

\[
\lim_{k \to \infty} \|v_i(k) - x^*\| = 0,
\]  

(85)

for all \( i \in V \) a.s.

Finally, from \( \sum_{k=0}^{\infty} \|e_i(k)\|^2 < \infty \) for all \( i \in V \) in Lemma 5 (i.e., Lemma 7 in [8]), we thus have

\[
\lim_{k \to \infty} \|x_i(k + 1) - v_i(k)\| = 0.
\]  

(86)
for all $i \in V$ a.s. where $\epsilon_i(k) = x_i(k + 1) - v_i(k)$, i.e., $\sum_{k=0}^{\infty} \|\epsilon_i(k)\|^2 = \sum_{k=0}^{\infty} \|x_i(k + 1) - v_i(k)\| < \infty$. This implies that $\lim_{k \to \infty} \|x_i(k + 1) - v_i(k)\| = 0$ which together with (85) (i.e., (86) and (85)) imply that

$$\lim_{k \to \infty} x_i(k) = x^*,$$

(87)

for all $i \in V$ a.s.

Hence, the result of Lemma 6 follows.

VII. CONVERGENCE RATE

We are going to show more explicitly with details how inequality (11) in the case where at least for one $(i)$ we have $f^{(i)}(x^{(i)}) < f^{(i)}(x^*)$ in the case of strongly convex $f^{(i)}$ for all $(i)$. In this case which corresponds to Lemma 3 and before reaching the explicit result of (24) we are going to use (121). Now,

$$2\alpha_k \sum_{j=1}^{n} \sum_{(i) \in I} \gamma^{(i)}(f^{(i)}(x^*) - f^{(i)}(x^{(i)}))$$

$$\leq 2\alpha_k \sum_{j=1}^{n} \sum_{(i) \in I} \gamma^{(i)} \nabla f^{(i)}(x^*) - \nabla f^{(i)}(x^{(i)}), x^* - x^{(i)}$$

$$\leq 2\alpha_k \sum_{j=1}^{n} \sum_{(i) \in I} \gamma^{(i)} \|x^{(i)} - x^*\|^2$$

$$\leq 2n\alpha_k L \min(|I|, \gamma_{\max}) \max_{(i)} \|x^{(i)} - x^*\|^2$$

And we have from the strong convexity of $f^{(i)}$ that

$$f^{(i)}(v_j(k))$$

$$\geq f^{(i)}(x^{(i)}) + \langle \nabla f^{(i)}(x^{(i)}), v_j(k) - x^{(i)} \rangle + \frac{\sigma^{(i)}}{2} \|x^{(i)} - v_j(k)\|^2$$

(89)

Thus, (96) becomes

$$f^{(i)}(x^{(i)}) - f^{(i)}(v_j(k))$$

$$\leq -\frac{\sigma^{(i)}}{2} \|x^{(i)} - v_j(k)\|^2$$

(90)

And since

$$\|x^{(i)} - v_j(k)\| \geq \|x^{(i)} - x^*\| - \|x^* - v_j(k)\|$$

(91)

then

$$\|x^{(i)} - v_j(k)\|^2 \geq \|x^{(i)} - x^*\|^2 + \|x^* - v_j(k)\|^2$$

$$- 2\|x^{(i)} - x^*\| \|x^* - v_j(k)\|$$

(92)
Thus,
\[ f(i)(x(i)) - f(i)(v_j(k)) \]
\[ \leq -\frac{\sigma(i)}{2} \Vert x(i) - v_j(k) \Vert^2 \]
\[ \leq -\frac{\sigma(i)}{2} \Vert x(i) - x^* \Vert^2 - \frac{\sigma(i)}{2} \Vert x^* - v_j(k) \Vert^2 + \sigma(i) \Vert x(i) - x^* \Vert \Vert x^* - v_j(k) \Vert \]

Then from convergence lemma (i.e., Lemma 6) since \( \Vert x^* - v_j(k) \Vert^2 \leq \Vert x(i) - x^* \Vert^2 \) for \( k \geq k_c \), we have
\[ f(i)(x(i)) - f(i)(v_j(k)) \]
\[ \leq -\frac{\sigma(i)}{2} \Vert x(i) - x^* \Vert^2 - \frac{\sigma(i)}{2} \Vert x^* - v_j(k) \Vert^2 \]

Therefore,
\[ \sum_{j=1}^{n} \sum_{i=1}^{P} \gamma(i)(f(i)(x(i)) - f(i)(v_j(k))) \]
\[ \leq p\gamma_{\min}(\frac{\sigma_{\max}}{2} \max_{(i)} \Vert x(i) - x^* \Vert^2 - \frac{\sigma_{\min}}{2} \Vert x^* - v_j(k) \Vert^2) \]

But we have from quadratic inequality due to convexity of \( f(i) \) that
\[ f(i)(v_j(k)) \leq f(i)(x(i)) + \langle \nabla f(i)(x(i)), v_j(k) - x(i) \rangle + \frac{L}{2} \Vert x(i) - v_j(k) \Vert^2 \]

Then
\[ f(i)(v_j(k)) - f(i)(x(i)) \]
\[ \leq \frac{L}{2} \Vert v_j(k) - x(i) \Vert^2 \]
\[ \leq L \Vert x(i) - x^* \Vert^2 + L \Vert x^* - v_j(k) \Vert^2 \]

Therefore,
\[ \sum_{j=1}^{n} \sum_{i=1}^{P} \gamma(i)(f(i)(v_j(k)) - f(i)(x(i))) \]
\[ \leq Ln \min(p\gamma_{\max}, 1)(\Vert x(i) - x^* \Vert^2 + \Vert x^* - v_j(k) \Vert^2) \]

Using \( (100) \) and evaluating the fourth term by \( (88) \) and using \( (95) \) to evaluate the first part of the fifth term and \( (98) \) to evaluate the second part we get
Thus, one way to tackle this problem is to adjust a lower bound on $\sum_{(i)=1}^{I(i)} \gamma_{(i)}$ using Definition 1 and (144) to get

$$\sum_{l=1}^{n} \|v_l(k+1) - x^*\|^2 \leq (1 - \mu - \alpha_k p \gamma_{\min} n \sigma_{\min})$$

$$+ 2 Ln \min(p \gamma_{\max}, 1) \alpha_k^2 \sum_{j=1}^{n} \|v_j(k) - x^*\|^2$$

$$+ 2 \alpha_k \sum_{j=1}^{n} \sum_{(i)=1}^{p} \gamma_{(i)} \langle e_{j,(i)}(k), v_j(k) - x^* \rangle$$

$$+ 2 \alpha_k \sum_{j=1}^{n} \sum_{(i)=1}^{p} \gamma_{(i)} \|e_{j,(i)}(k)\|^2$$

$$+ (2n \alpha_k L \min(I|\gamma_{\max}, 1) + \alpha_k p \gamma_{\min} n \sigma_{\max})$$

$$+ 2 Ln \min(p \gamma_{\max}, 1) \alpha_k^2 \max_{(i)} \|x^{(i)} - x^*\|^2$$

(99)

Then in order to use Lemma Martingale 2 at least one of these three inequalities must be satisfied:

- $\frac{2 Ln \min(p \gamma_{\max}, 1) \alpha_k^2}{\sum_{(i)=1}^{I(i)} \gamma_{(i)}^2} \leq \alpha_k p \gamma_{\min} n \sigma_{\min}$
- $\frac{2 Ln \min(p \gamma_{\max}, 1) \alpha_k^2}{\sum_{(i)=1}^{I(i)} \gamma_{(i)}^2} \leq \mu$
- $\frac{2 Ln \min(p \gamma_{\max}, 1) \alpha_k^2}{\sum_{(i)=1}^{I(i)} \gamma_{(i)}^2} \leq \mu + \alpha_k p \gamma_{\min} n \sigma_{\min}$

Thus, one way to tackle this problem is to adjust a lower bound on $\mu$ by adequately choosing the described row stochastic matrices $W_{(k)}$. But fortunately enough due to the strong convexity of the function and the existence of a lower bound on the monotonicity of the gradient of $f^{(i)}$ which is $\sigma_{(i)}$ the first inequality is satisfied for any decreasing sequence $\alpha_k$. Thus, eventually for $k \geq k_0$ an inequality in the form (132) is valid. And having it satisfied until the at least the first $B$ iterations after $k_0$ will result in applying the Martingale 2 result and thus bounding $\sum_{j=1}^{n} \|v_j(k) - x^*\|^2 \leq \eta$ where

$$\eta = \frac{(2n \alpha_k L \min(I|\gamma_{\max}, 1) + \alpha_k p \gamma_{\min} n \sigma_{\max} + \frac{2 Ln \min(p \gamma_{\max}, 1) \alpha_k^2}{\sum_{(i)=1}^{I(i)} \gamma_{(i)}^2} \max_{(i)} \|x^{(i)} - x^*\|^2}{\mu - 4 \alpha_k \|A^{(i)}\|_{\infty} \|B^{(i)}\|_{2,\infty} (1 + 2 \alpha_k \|A^{(i)}\|_{\infty} \|B^{(i)}\|_{2,\infty})}$$

(100)
Then using (78) from the third part of the convergence analysis and the strong convexity of $f$ resulting from the strong convexity of $f^{(i)}$ we get

$$
E\left[\sum_{i=1}^{n} \frac{\|\mathbf{x}_i(k+1) - \mathbf{x}^*\|^2}{F_k}\right] \leq (1 + (A_\eta - aA_\tau)\alpha_k^2) \sum_{i=1}^{n} \|\mathbf{x}_i(k) - \mathbf{x}^*\|^2 \\
+ (1 + (A_\eta - aA_\tau)\alpha_k^2)\alpha_k nF + 4(1 - a)\alpha_k G_f \sum_{i=1}^{n} \|\mathbf{v}_i(k) - \mathbf{v}(k)\| - \sigma \gamma_{\min}(1 - a)\alpha_k \|z(k) - \mathbf{x}^*\|^2 \\
+ n (B_\eta - aB_\tau + 2)\alpha_k^2 G_f^2
$$

(102)

But since $\|\Pi_X(\mathbf{v}_i(k)) - \mathbf{x}^*\| \geq \|\mathbf{v}_i(k) - \mathbf{x}^*\| - \|\Pi_X(\mathbf{v}_i(k)) - \mathbf{v}_i(k)\|, \|\Pi_X(\mathbf{v}_i(k)) - \mathbf{v}_i(k)\| \leq \|\mathbf{v}_i(k) - \mathbf{x}^*\|$ and the convexity of the norm we get

$$
E\left[\sum_{i=1}^{n} \frac{\|\mathbf{x}_i(k+1) - \mathbf{x}^*\|^2}{F_k}\right] \leq (1 + (A_\eta - aA_\tau)\alpha_k^2) \sum_{i=1}^{n} \|\mathbf{x}_i(k) - \mathbf{x}^*\|^2 \\
+ (1 + (A_\eta - aA_\tau)\alpha_k^2)\alpha_k nF + 16(1 - a)\alpha_k G_f \sum_{i=1}^{n} \|\mathbf{v}_i(k) - \mathbf{x}^*\|^2 \\
- \frac{\sigma}{n} \gamma_{\min}(1 - a)\alpha_k \sum_{i=1}^{n} \|\Pi_X(\mathbf{v}_i(k)) - \mathbf{v}_i(k)\|^2 \\
+ \frac{\sigma}{n} \gamma_{\min}(1 - a)\alpha_k \sum_{i=1}^{n} \|\mathbf{v}_i(k) - \mathbf{x}^*\|^2 \\
+ n (B_\eta - aB_\tau + 2)\alpha_k^2 G_f^2
$$

(103)

Using $\|\mathbf{x}_i(k) - \mathbf{x}^*\|^2 = \|\mathbf{v}_i(k-1) - \mathbf{x}^*\|^2 + 2\alpha_k \langle \nabla f^{(i)}(\mathbf{v}_i(k-1)), \mathbf{x}^* - \mathbf{v}_i(k-1) \rangle$ and $f(\mathbf{x}^*) - f(\mathbf{v}_i(k-1)) \leq -\sigma \alpha_k \|\mathbf{v}_i(k-1) - \mathbf{x}^*\|^2$, we get

$$
E\left[\sum_{i=1}^{n} \frac{\|\mathbf{x}_i(k+1) - \mathbf{x}^*\|^2}{F_k}\right] \leq \\
(1 + (2 - \sigma)\alpha_k)(1 + (A_\eta - aA_\tau)\alpha_k^2) \sum_{i=1}^{n} \|\mathbf{v}_i(k - 1) - \mathbf{x}^*\|^2 \\
+ (1 + (A_\eta - aA_\tau)\alpha_k^2)\alpha_k nF + 16(1 - a)\alpha_k G_f \sum_{i=1}^{n} \|\mathbf{v}_i(k) - \mathbf{x}^*\|^2 \\
- \frac{\sigma}{n} \gamma_{\min}(1 - a)\alpha_k \sum_{i=1}^{n} \|\Pi_X(\mathbf{v}_i(k)) - \mathbf{v}_i(k)\|^2 \\
+ \frac{\sigma}{n} \gamma_{\min}(1 - a)\alpha_k \sum_{i=1}^{n} \|\mathbf{v}_i(k) - \mathbf{x}^*\|^2 \\
+ n (B_\eta - aB_\tau + 2)\alpha_k^2 G_f^2
$$

(104)
But having $\sum_{i=1}^{n} \|v_i(k) - x^*\|^2 \leq \eta$, we get

$$
\mathbb{E} \left[ \sum_{i=1}^{n} \|x_i(k) - x^*\|^2 \right] \leq \left( 1 + (A_\eta - aA_\tau)\alpha_k^2 \right) \left( 1 + (2 - \sigma)\alpha_k + 4\sigma^2 nL^2\max_{(i)} \|A^{(i)}\|_\infty^2 \|B^{(i)}\|_2^2,\infty \right) + 16(1 - a)\alpha_k G_f + \sigma n \gamma_{\min} (1 - a)\alpha_k |\eta| + n (B_\eta - aB_\tau + 2)\alpha_k^2 G_f^2
$$

(105)

which is the expected convergence rate of a strongly convex function formed of $p$ strongly convex functions with $f^{(i)}(x^*) > f^{(i)}(x^i)$ for at least one $(i)$.

A. Convergence Rate for Strongly Convex Function with $f^{(i)}(x^*) = f^{(i)}(x^i)$ for all $(i)$

The convergence rate of a strongly convex function formed of $p$ strongly convex functions with $f^{(i)}(x^*) = f^{(i)}(x^i)$ for all $(i)$ can be deduced by applying Lemma 2 and is

$$
\mathbb{E} \left[ \sum_{i=1}^{n} \|v_i(k) - x^*\|^2 \right] \leq \left( 1 - \mu + \frac{4L\alpha_k \|A^{(i)}\|_\infty \|B^{(i)}\|_2,\infty \left( 1 + 2L\alpha_k \|A^{(i)}\|_\infty \|B^{(i)}\|_2,\infty \right)}{\left( \sum_{(i)=1}^{p} \gamma(i) \right)^2} \right)^k V_0
$$

(106)

for $k \geq k_0$.

Numerical Simulation

Our aim in the simulation is to verify the convergence of the proposed algorithm, while showing its convergence rate for different network topologies.

In this section, we restrict the optimization problem to the following unconstrained convex optimization problem on a network

$$
\arg\min_{x \in \mathbb{R}^N} \|Gx - y\|_2^2,
$$

(107)

where the network contains $n$ nodes, $G$ is a random matrix of size $M \times N$ whose entries are independent and identically distributed standard normal random variables, and

$$
y = Gx_o \in \mathbb{R}^M
$$

(108)

has entries of $x_o$ that are identically independent random variables sampled from the uniform bounded random distribution between $-1$ and 1. The solution $x^*$ of the optimization problem above is the least squares solution of the overdetermined system $y = Gx_o$, $x_o \in \mathbb{R}^N$. We demonstrate the performance of SRDO to solve the convex optimization problem (1), and match it with the calculated convergence rates.

Assume that the network has $s$ as the maximum number of allowed stragglers. Then we can repartition the network around those $m$ worker nodes, and accordingly, the random measurement matrix $G$, the measurement data $y$, and the objective function $f(x) := \|Gx - y\|_2^2$ in (1) as follows:

$$
f(x) = \sum_{i=1}^{n} f^{(i)}(x) := \sum_{i=1}^{n} \|G_i x - y_i\|_2^2.
$$
In the simulations, without a loss of generality, we assume the number of worker partitions equals the number of server nodes; i.e., \( p = n \). We also assume that the repartitioned parts have the same size; i.e., the number of rows in \( G_i \) and the lengths of vectors \( y_i, 1 \leq i \leq n \) are the same.

Each worker node finds its local coded gradient through a combination of uncoded local gradients computed through local optimization problems of overdetermined linear systems of equations. The stepsizes \( \alpha_k \) are chosen such that
\[
\alpha_k = \frac{1}{(k + a)^\theta}.
\]

We use the absolute error
\[
AE := \max_{1 \leq i \leq n} \frac{\|x_i(k) - x_o\|_2}{\|x_o\|_2}
\]
and consensus error
\[
CE := \max_{1 \leq i \leq n} \frac{\|x_i(k) - \bar{x}(k)\|_2}{\|x_o\|_2}
\]
to measure the performance of SRDO, where \( n \) is the number of server nodes (i.e., servers) in the network.

We simulate the algorithm for the fixed network topologies of the allowed number of stragglers or varying connection topologies according to gradient computation scenarios 1 through 3.

We present the simulation for 100 samples of parameter server networks of \( p = 5 \) equal sized partitions \( i \), where \( n_i = 3, s_i = 1 \) and \( n_i = 5, s_i = 2 \) for \( \mu = 100, N = 100 \), i.e., \( m_i = 300 \) and \( 500 \), and \( M = 1500 \) and \( 2500 \), respectively. Here, \( \mu \) stands for the number of rows in a partition sub-partition, which is assumed equal all over the network. That is, \( \mu \) corresponds to the functions \( f^{(i)}_l \), where \( l \) corresponds to worker node \( l \) in partition \( i \). We simulate different samples from variant connectivity levels:

- **Strong Connectivity** Condition: i) the allowed-number-of-stragglers connection topology (i.e., strong connectivity condition under gradient computation scenario 1), Fig. 3 in comparison with the full connection with the no failures scenario of uncoded centralized gradient descent algorithm.
- **Weak Connectivity** Condition: ii) the more-than-the-allowed number of stragglers connection with some instants of allowed number of stragglers connectivity under the weak straggler tolerance condition compared with the uncoded centralized gradient descent. For the two used gradient computation scenarios, we have:
  - a) Scenario 2 where only iteration available instant gradients are used, Fig. 4
  - b) Scenario 3 with delayed gradients used, Fig. 5 and Fig. 6

### B. Discussion of the Figures

In Fig. 3 SRDO algorithm converges in a rate close to the uncoded centralized gradient descent algorithm with full connection. The convergence can be faster or slower depending on the condition number of the coded matrices in relation to the uncoded matrices at each node; i.e., this is related to the respective Lipschitz constants \( L \). For Figs. 4, 5, and 6 the behavior matches the described convergence rates in Section VII. The fluctuation of the average consensus error for the SRDO has a larger variation, which is dependent on \( T = \max_{i \in [1:n]} T_i \), while the proposed algorithm corresponding absolute error behaves more smoothly.
Fig. 3. Allowed number of stragglers connection for SRDO algorithm network (n=5, s=2) with $\theta = \frac{1}{(k+300)^{0.55}}$ and centralized uncoded gradient descent algorithm for a full connection with no stragglers (n=5, s=0), for $M = 2500, N = 100$.

We can conceive from the simulations that SRDO has faster convergence for a smaller exponent $\theta \in (0, 1]$, which confirms its convergence rate estimate in Section VII. However, the simulations also indicate that decreasing exponent $\theta$ moves SRDO into the divergence phase, which could directly be related to the complexity of the network. It is worth mentioning that we can adequately calibrate this divergence by increasing the value of $a$ in our illustrated examples for a fixed exponent $\theta$. We can see that for a fixed value of $T$ and a fixed condition number, i.e., fixed Lipschitz constant $L$, (more specifically for a fixed matrix $G_i$), the decrease in the exponent $\theta$ allows the algorithm to enter the divergence instability region. Then an increase in the stepsize will make it converge the fastest where then any increase will ultimately degrade the algorithm to a slower convergence. Similarly, if we fix $T$ and the stepsize, the behavior of the convergence of the algorithm relative to the change in the condition number is the same as that relative to the stepsize in the previous scenario. Moreover, we see that when $T$ increases, and the allowed number of stragglers connection becomes less frequent, then the convergence is replaced by an anticipated
Fig. 4. Varying number of stragglers connection for SRDO algorithm network \((n=3, s=1)\) and \(T = 5\) for \(\alpha_k = \frac{1}{(k+300)^{0.35}}\) using gradient computation scenario 2 and centralized uncoded gradient descent algorithm for a full connection with no stragglers \((n=3, s=0)\), for \(M = 1500, N = 100\).

divergence. Then, for that \(T\), we can reenter the convergence region of the algorithm by increasing the exponent \(\theta\) for a fixed optimization problem. Convergence is also achieved for problems with matrices of higher condition numbers when the stepsize is fixed. As for scenario 3, we see that its performance is better than that of scenario 2, because scheme 3 exploits stale gradients to form the overall gradient at each instant, although both schemes 2 and 3 have slower convergence rates for the fixed stepsize and fixed condition number.

Thus, we can anticipate in Fig. 4 that the value of \(\theta = 0.35\) allowed a comparable convergence rate of the SRDO for \(T = 5\) as that of the centralized gradient descent algorithm. In Figs. 5 and 6, we realize that the lower value of \(\theta = 0.35\) is not permissible, because the SRDO algorithm will considerably enter the instability region, while a higher value of \(\theta = 0.55\) favors a better convergence rate for \(T = 10\), and the highest value of \(\theta = 0.75\) a better
Fig. 5. Varying number of stragglers connection for SRDO algorithm network (n=3, s=1) and $T = 10$ for $\alpha_k = \frac{1}{(k+300)^{0.55}}$ using gradient computation scenario 3 and centralized uncoded gradient descent algorithm for a full connection with no stragglers (n=3, s=0), for $M = 1500$, $N = 100$.

convergence rate for $T = 20$. Moreover, in the simulation we have used a definite coding scheme introduced in [16]. We could have used other schemes. As it is proven in Section VII, we anticipate a behavior dependent on the coding scheme as shown in (??). Meanwhile, we can adequately propose a coding scheme that will better enhance the convergence rate.

VIII. CONCLUSION

We have considered in this paper a parameter server network algorithm, SRDO, for minimizing a convex function that consists of a number of component functions. We restricted the simulation for the case of a quadratic function which corresponds to solving an overdetermined system of linear equations. A convergence proof for this algorithm in its general form (not necessarily a quadratic function) was provided in the case of network topologies where the
Varying number of stragglers connection for SRDO algorithm network (n=3, s=1) and $T = 20$ for $\alpha_k = \frac{1}{(k+300)^{0.75}}$ using gradient computation scenario 3 and centralized uncoded gradient descent algorithm for a full connection with no stragglers (n=3, s=0), for $M = 1500, N = 100$.

The convergence of the algorithm variants, in the case of the number of stragglers exceeding the quantity allowed, is formalized in Section VII through the described convergence rates. Furthermore, the simulation showed optimal results for the algorithm convergence rate, not only for the allowed number of stragglers scenarios but also for the number of stragglers exceeding the allowed threshold. These metrics matched the centralized gradient descent method with the bonus of robustness to an allowed number of stragglers, and to the case of stragglers exceeding the allowed number. We further analytically showed that the convergence rate can be considerably enhanced through applying an adequate coding scheme.
APPENDIX

A. Proof of Lemma 7

Proof.

\[
\|v_i(k + 1) - x^*\|^2 = \| \sum_{j=1}^{n} [W(k)]_{i,j} x_j(k + 1) - \sum_{j=1}^{n} [W(k)]_{i,j} x_j^* \|^2 \\
\leq \sum_{j=1}^{n} [W(k)]_{i,j} \|x_j(k + 1) - x^*\|^2 \\
\leq \sum_{j=1}^{n} [W(k)]_{i,j} \|v_j(k) - x^*\|^2 \\
\leq \sum_{j=1}^{n} [W(k)]_{i,j} [\|v_j(k) - x^*\|^2 - \alpha_k \sum_{i=1}^{p} \nabla f^{(i)}(v_j(k)) \cdot (v_j(k) - x^*)] \\
\leq \sum_{j=1}^{n} [W(k)]_{i,j} \|v_j(k) - x^*\|^2 - 2\alpha_k \sum_{i=1}^{p} \gamma_i \langle \nabla f^{(i)}(v_j(k)), v_j(k) - x^* \rangle + \alpha_k^2 \sum_{i=1}^{p} \gamma_i \|\nabla f^{(i)}(v_j(k))\|^2 ] \\
\leq \sum_{j=1}^{n} [W(k)]_{i,j} \|v_j(k) - x^*\|^2 + 2\alpha_k \sum_{i=1}^{p} \gamma_i \langle \nabla f^{(i)}(v_j(k)), x^* - v_j(k) \rangle \\
+ 2\alpha_k \sum_{i=1}^{p} \gamma_i \langle v_j, (i), k, v_j(k) - x^* \rangle + \alpha_k^2 \sum_{i=1}^{p} \gamma_i \|\nabla f^{(i)}(v_j(k)) - \epsilon_{j,i}(k)\|^2 ] \\
\leq \sum_{j=1}^{n} [W(k)]_{i,j} \|v_j(k) - x^*\|^2 + 2\alpha_k \sum_{i=1}^{p} \gamma_i \langle \nabla f^{(i)}(v_j(k)), x^* - v_j(k) \rangle \\
+ 2\alpha_k \sum_{i=1}^{p} \gamma_i \langle v_j, (i), k, v_j(k) - x^* \rangle + \alpha_k^2 \sum_{i=1}^{p} \gamma_i \frac{\|\nabla f^{(i)}(v_j(k)) - \epsilon_{j,i}(k)\|^2}{(\sum_{i=1}^{p} \gamma_i)^2} ] 
\]

(110)

Then summing from \( i = 1 \) to \( n \) and knowing that the sum of each column is less than or equal to \( 1 - \mu \) then the lemma follows.

B. Proof of Lemma 3

Having Lemma 7 then we have

\[
(1 - \mu) \sum_{i=1}^{n} \|v_i(k + 1) - x^*\|^2 \leq (1 - \mu) \sum_{j=1}^{n} \|v_j(k) - x^*\|^2 \\
+ 2\alpha_k \sum_{j=1}^{n} \sum_{i=1}^{p} \gamma_i \langle \epsilon_{j,i}(k), v_j - x^* \rangle + 2\alpha_k^2 \sum_{j=1}^{n} \sum_{i=1}^{p} \gamma_i \|\epsilon_{j,i}(k)\|^2 \\
+ 2\alpha_k \sum_{j=1}^{n} \sum_{i=1}^{p} \gamma_i \langle \nabla f^{(i)}(v_j(k)), x^* - v_j(k) \rangle \\
+ \frac{2\alpha_k^2}{(\sum_{i=1}^{p} \gamma_i)^2} \sum_{j=1}^{n} \sum_{i=1}^{p} \gamma_i \|\nabla f^{(i)}(v_j(k))\|^2 
\]

(111)
Then substituting (116) and (117) in (118) we have

\[
\langle \nabla f^{(i)}(v_j(k)), x^* - v_j(k) \rangle = -\langle \nabla f^{(i)}(v_j(k)), v_j(k) - x^* \rangle = -\langle \nabla f^{(i)}(v_j(k)), v_j(k) - x^{(i)} \rangle = -\langle \nabla f^{(i)}(v_j(k)) - \nabla f^{(i)}(x^{(i)}), v_j(k) - x^{(i)} \rangle
\]

(112)

While

\[
\nabla f^{(i)}(v_j(k)) - \nabla f^{(i)}(x^{(i)}) = a\|v_j(k) - x^{(i)}\| \overrightarrow{u}
\]

(113)

where \( \| \overrightarrow{u} \| = 1 \) and \( 0 \leq a \leq L \), and

\[
v_j(k) - x^{(i)} = \|v_j(k) - x^{(i)}\| \overrightarrow{v}
\]

(114)

where \( \| \overrightarrow{v} \| = 1 \). Using what preceded we have the expression in (112) equal to

\[
\langle \nabla f^{(i)}(v_j(k)), x^* - v_j(k) \rangle = -a\|v_j(k) - x^{(i)}\|^2 \langle \overrightarrow{u}, \overrightarrow{v} \rangle
\]

(115)

But since \( \langle \overrightarrow{u}, \overrightarrow{v} \rangle \geq 0 \) due to the monotonicity of the gradient we have \( 0 \leq \langle \overrightarrow{u}, \overrightarrow{v} \rangle \leq 1 \). Then

\[
\langle \nabla f^{(i)}(v_j(k)), x^* - v_j(k) \rangle = -a\|v_j(k) - x^{(i)}\|^2 b
\]

(116)

where \( 0 \leq b \leq 1 \).

Similarly,

\[
\| \nabla f^{(i)}(v_j(k)) \|^2 = \langle \nabla f^{(i)}(v_j(k)) - \nabla f^{(i)}(x^{(i)}), \nabla f^{(i)}(v_j(k)) - \nabla f^{(i)}(x^{(i)}) \rangle
\]

\[
= a^2\|v_j(k) - x^{(i)}\|^2 \overrightarrow{u}, \overrightarrow{u}
\]

\[
= a^2\|v_j(k) - x^{(i)}\|^2
\]

(117)

Then substituting (116) and (117) in (118) we have

\[
(1 - \mu) \sum_{l=1}^{n} \| v_l(k+1) - x^* \|^2 \leq (1 - \mu) \sum_{j=1}^{n} \| v_j(k) - x^* \|^2 + 2\alpha_k \sum_{j=1}^{n} \sum_{(i)=1}^{p} \gamma_{(i)} \langle \epsilon_{(i)}(k), v_j - x^* \rangle + 2\alpha_k a \sum_{j=1}^{n} \sum_{(i)=1}^{p} \gamma_{(i)} \| \epsilon_{(i)}(k) \| ^2 \]

\[
- 2\alpha_k \sum_{j=1}^{n} \sum_{(i)=1}^{p} \gamma_{(i)} (b - \frac{2\alpha_k a}{\sum_{(i)=1}^{p} \gamma_{(i)}}) \| v_j(k) - x^{(i)} \|^2
\]

(118)
C. Proof of Lemma

From Lemma 11 we have we have

\[
(1 - \mu) \sum_{l=1}^{n} \|v_l(k + 1) - x^*\|^2 \leq (1 - \mu) \sum_{j=1}^{n} \|v_j(k) - x^*\|^2 \\
+ 2\alpha_k \sum_{j=1}^{n} \sum_{(i) \in I} \gamma(i) \langle \epsilon_j(i)(k), v_j - x^* \rangle + 2\alpha_k^2 \frac{\sum_{j=1}^{n} \sum_{(i) = 1}^{p} \gamma(i)}{(\sum_{(i) = 1}^{p} \gamma(i))^2} \|\epsilon_j(i)(k)\|^2 \\
+ 2\alpha_k \sum_{j=1}^{n} \sum_{(i) \in I^c} \gamma(i) \langle \nabla f^{(i)}(v_j(k)), x^* - v_j(k) \rangle \\
+ 2\alpha_k^2 \frac{\sum_{j=1}^{n} \sum_{(i) = 1}^{p} \gamma(i)}{(\sum_{(i) = 1}^{p} \gamma(i))^2} \sum_{j=1}^{n} \sum_{(i) = 1}^{p} \gamma(i) \|\nabla f^{(i)}(v_j(k))\|^2 \\
\]

But for \((i) \in I\) we have \(f^{(i)}(x^{(i)}) < f^{(i)}(x^*)\) and for \((i) \in I^c\) we have \(f^{(i)}(x^*) = f^{(i)}(x^{(i)})\), then the above inequality becomes

\[
(1 - \mu) \sum_{l=1}^{n} \|v_l(k + 1) - x^*\|^2 \leq (1 - \mu) \sum_{j=1}^{n} \|v_j(k) - x^*\|^2 \\
+ 2\alpha_k \sum_{j=1}^{n} \sum_{(i) \in I} \gamma(i) \langle \epsilon_j(i)(k), v_j - x^* \rangle + 2\alpha_k^2 \frac{\sum_{j=1}^{n} \sum_{(i) = 1}^{p} \gamma(i)}{(\sum_{(i) = 1}^{p} \gamma(i))^2} \|\epsilon_j(i)(k)\|^2 \\
+ 2\alpha_k \sum_{j=1}^{n} \sum_{(i) \in I} \gamma(i) \langle \nabla f^{(i)}(v_j(k)), x^* - x^{(i)} - x^{(i)} - v_j(k) \rangle \\
+ 2\alpha_k^2 \frac{\sum_{j=1}^{n} \sum_{(i) = 1}^{p} \gamma(i)}{(\sum_{(i) = 1}^{p} \gamma(i))^2} \sum_{j=1}^{n} \sum_{(i) = 1}^{p} \gamma(i) \|\nabla f^{(i)}(v_j(k))\|^2 \\
\]

which becomes

\[
(1 - \mu) \sum_{l=1}^{n} \|v_l(k + 1) - x^*\|^2 \leq (1 - \mu) \sum_{j=1}^{n} \|v_j(k) - x^*\|^2 \\
+ 2\alpha_k \sum_{j=1}^{n} \sum_{(i) \in I} \gamma(i) \langle \epsilon_j(i)(k), v_j - x^* \rangle + 2\alpha_k^2 \frac{\sum_{j=1}^{n} \sum_{(i) = 1}^{p} \gamma(i)}{(\sum_{(i) = 1}^{p} \gamma(i))^2} \|\epsilon_j(i)(k)\|^2 \\
+ 2\alpha_k \sum_{j=1}^{n} \sum_{(i) \in I} \gamma(i) \langle \nabla f^{(i)}(v_j(k)), x^* - x^{(i)} \rangle \\
+ 2\alpha_k \sum_{j=1}^{n} \sum_{(i) \in I} \gamma(i) \langle \nabla f^{(i)}(v_j(k)), x^{(i)} - v_j(k) \rangle \\
+ 2\alpha_k^2 \frac{\sum_{j=1}^{n} \sum_{(i) = 1}^{p} \gamma(i)}{(\sum_{(i) = 1}^{p} \gamma(i))^2} \sum_{j=1}^{n} \sum_{(i) = 1}^{p} \gamma(i) \|\nabla f^{(i)}(v_j(k))\|^2 \\
\]
which by using (116) and (117) we get the result.

D. Martingale 1

Lemma 7. Assume the following inequality holds a.s. for all \( k \geq k^* \),

\[
+1 \leq a_k + \max_{k-B \leq k} v_{k} \tag{122}
\]

where \( v_k, w_k, b_k, c_k, a_1 \) and \( a_{2,k} \) are non-negative random variables where \( a_1 + a_{2,k} \leq 1 \) and \( \{a_{2,k}\} \) is a decreasing sequence.

Then if for \( \rho = (a_1 + a_{2,1})^{\frac{1}{1+\beta}} \) and

\[
v_{k_0} \leq \rho \Phi(k_0) V_0' \quad \text{a.s.} \tag{123}
\]

for base case \( k = k_0 = \bar{k} - B \) (i.e., notice \( V_0 \) is not necessary the initial value \( v_0 \)). And \( \Phi \) 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 \( \rho \) is independent of \( k \)). i.e., \( k \in \{k_0 = \bar{k} - B, \ldots, \bar{k}\} \) and \( \bar{k} - B \geq \max(k^*, \bar{k}) \). That is

\[
v_k \leq \rho \Phi(k) V_0' \quad \text{a.s.} \tag{124}
\]

for \( k = \{k_0, \ldots, \bar{k}\} \).

Then we have

\[
v_k \leq \rho^k V_0 \quad \text{a.s.} \tag{125}
\]

for all \( k \geq \bar{k} \) where \( V_0 > 0 \) for all sequences patterns and \( \rho \) as before.

Proof: First since \( a_1 + a_{2,k} \leq 1 \) then

\[
1 \leq (a_1 + a_{2,k})^{\frac{1}{1+\beta}} \implies 1 \leq (a_1 + a_{2,1})^{\frac{1}{1+\beta}}
\]

\[
\implies (a_1 + a_{2,k}) \leq (a_1 + a_{2,1}) \tag{126}
\]

which implies that

\[
a_1 + a_{2,k} \rho^{-B} = a_1 + a_{2,k}(a_1 + a_{2,1})^{-\frac{1}{1+\beta}} \leq a_1(a_1 + a_{2,1})^{-\frac{1}{1+\beta}} + a_{2,k}(a_1 + a_{2,1})^{-\frac{1}{1+\beta}} = (a_1 + a_{2,k})(a_1 + a_{2,1})^{-\frac{1}{1+\beta}} \leq (a_1 + a_{2,k})^{\frac{1}{1+\beta}} = \rho
\]

That is

\[
a_1 + a_{2,k} \rho^{-B} \leq \rho \tag{127}
\]

Now, by induction we show that (125) for all \( k \geq k_0 \). Assume (123) 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
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 \( 0 \leq l \leq B \)

\[
v_k \leq \rho^{k-l}V_0' \\
v_{k-B} \leq \rho^{\Phi(k-B)}V_0'
\]

Then from (122) we have

\[
v_{k+1} \leq a_1 v_k + a_{2,k} \max_{k-B \leq l \leq k} v_l \\
\leq a_1 \rho^{k-l}V_0' + a_{2,k} \rho^{k-B}V_0' \\
\leq a_1 \rho^{k-l}V_0' + a_{2,k} \rho^{k-1-B}V_0' \\
= (a_1 + a_{2,k} \rho^{-B}) \rho^{k-1}V_0' \\
\leq \rho^{k+l+2}V_0' \ a.s.
\]

But without a loss of generality, we can find \( V_0 > 0 \) such that \( \rho^{k+l+2}V_0' \leq \rho^{k+1}V_0 \) to keep indexing tractable. And thus (129) is true for all \( k \geq \bar{k} + 1 \). i.e., notice that for \( k + 1 = \bar{k} + 2 \), we already have for \( k = \bar{k} + 1 \) that the power of \( \rho \) 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

\[
v_{k+2} \leq \rho^{k+2}V_0' \ a.s.
\]

Thus, (129) follows for all \( k \geq \bar{k} \). \( \square \)

**Remark 8.** i.e., notice that it is true for \( k = \bar{k} \) since

\[
v_{k+1} \leq \rho^{k-l}V_0' \text{ and } v_{k+1} \leq \rho^k V_0
\]

E. Martingale 2

**Lemma 8.** Assume the following inequality holds a.s. for all \( k \geq k^* \),

\[
v_{k+1} \leq a_1 v_k + a_{2,k} \max_{k-B \leq l \leq k} v_l + a_3
\]

\( v_k, \ u_k, \ a_3, \ a_1 \) and \( a_{2,k} \) are non-negative random variables where \( a_1 + a_{2,k} \leq 1 \) and \( \{a_{2,k}\} \) is a decreasing sequences. Then if for \( \rho = (a_1 + a_{2,1}) \frac{1}{1-a_1-a_2} \) and \( \eta = \frac{a_3}{1-a_1-a_2} \)

\[
v_{k_0} \leq \rho^{\Phi(k_0)}V_0' + \eta \ \ a.s.
\]

for base case \( k = k_0 = \bar{k} - B \). (i.e., notice \( V_0 \) is not necessary the initial value \( v_0 \)). And \( \Phi \) 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 \( \rho \) is independent of \( k \). i.e., \( k \in \{k_0 = \bar{k} - B, \ldots, \bar{k}\} \) and \( \bar{k} - B \geq \max(k^*, \bar{k}) \). That is

\[
v_k \leq \rho^{\Phi(k)}V_0' + \eta \ \ a.s.
\]
for $k = \{k_0, \ldots, \bar{k}\}$.

Then we have

$$v_k \leq \rho^k V_0 + \eta \quad \text{a.s.}$$  \hspace{1cm} (135)

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

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

$$1 \leq (a_1 + a_{2, k})^{-\frac{B}{\bar{k}+1}} \implies 1 \leq (a_1 + a_{2, 1})^{-\frac{B}{\bar{k}+1}}$$

$$\implies (a_1 + a_{2, k}) \leq (a_1 + a_{2, 1})$$

which implies that

$$a_1 + a_{2, k} \rho^{-B} = a_1 + a_{2, k}(a_1 + a_{2, 1})^{-\frac{B}{\bar{k}+1}}$$

$$\leq a_1(a_1 + a_{2, 1})^{-\frac{B}{\bar{k}+1}} + a_{2, k}(a_1 + a_{2, 1})^{-\frac{B}{\bar{k}+1}}$$

$$= (a_1 + a_{2, k})(a_1 + a_{2, 1})^{-\frac{B}{\bar{k}+1}}$$

$$\leq (a_1 + a_{2, k})^{-\frac{B}{\bar{k}+1}} = \rho$$

That is

$$a_1 + a_{2, k} \rho^{-B} \leq \rho$$  \hspace{1cm} (137)

Now, by induction we show that \((125)\) for all $k \geq k_0$. Assume \((133)\) 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 $0 \leq l \leq B$

$$v_k \leq \rho^{k-l} V_0' + \eta$$

$$v_{k-B} \leq \rho^{(k-B)} V_0' + \eta$$  \hspace{1cm} (138)

Then from \((132)\) we have

$$v_{k+1} \leq a_1 v_k + a_{2, k} \max_{k-B \leq l \leq k} v_k + a_3$$

$$\leq a_1 \rho^{k-l} V_0' + a_{2, k} \rho^{k-B} V_0' + a_1 \eta + a_{2, k} \eta + a_3$$

$$\leq a_1 \rho^{k-l} V_0' + a_{2, k} \rho^{k-B} V_0' + a_1 \eta + a_{2, k} \eta + a_3$$

$$= (a_1 + a_{2, k} \rho^{-B}) \rho^{k-l} V_0' + \eta$$

$$\leq \rho^{k-l+1} V_0' + \eta \quad \text{a.s.}$$  \hspace{1cm} (139)

But without a loss of generality, we can find $V_0 > 0$ such that $\rho^{k-l+1} V_0' \leq \rho^{k+1} V_0$ to keep indexing tractable. And thus \((139)\) is true for all $k \geq \bar{k} + 1$, i.e., notice that for $k + 1 = \bar{k} + 2$, we already have for $k = \bar{k} + 1$ that the power of $\rho$ in the recursive inequality after the coefficient $a_1$ is $\bar{k} + 1$. Thus, no matter what the arbitrary behavior for the prior $B$ terms is, we will have

$$v_{k+2} \leq \rho^{k+2} V_0' + \eta \quad \text{a.s.}$$  \hspace{1cm} (140)
Thus, (139) follows for all $k \geq \bar{k}$. □

**Remark 9.** i.e., notice that it is true for $k = \bar{k}$ since

$$v_{\bar{k}+1} \leq \rho^\bar{k}V_0 + \eta \text{ and } v_{\bar{k}+1} \leq \rho^kV_0 + \eta$$  (141)

**F. Evaluation of $\sum_{i=1}^{n} \|R_i(k - 1)\|^2$**

Then by Cauchy-Schwartz inequality, we have

$$\|R_i(k - 1)\|^2 \leq \alpha_{k-1}^2 \|A^{(i)}\|_\infty^2 \|B^{(i)}\|_\infty^2 \|f_i^{(i)}(v_M(k - 1)) - f_i^{(i)}(v_m(k - 1))\|^2$$

where $\|A^{(i)}\|_\infty$ is the norm of the row of $A^{(i)}$ with maximum $l_2$ norm which is bounded by the Euclidean norm of the vector formed by the support of a row of $A^{(i)}$ of length equals $n_i - s_i$ of maximum Euclidean norm, since not all coefficients are nonzero. And $v_M(k - 1)$ and $v_m(k - 1)$ are the instants at iteration $k'$ where $k - 1 - H \leq k' \leq k - 1$ with $\|f_i^{(i)}(v_M(k - 1)) - f_i^{(i)}(v_m(k - 1))\| = \max_{i,j} \|f_i^{(i)}(v_j(k')) - f_i^{(i)}(v_i(k'))\|$. Then

$$\|f_i^{(i)}(v_M(k - 1)) - f_i^{(i)}(v_m(k - 1))\|$$

$$\leq \|f_i^{(i)}(v_M(k - 1)) - f_i^{(i)}(x^*)\| + \|f_i^{(i)}(v_m(k - 1)) - f_i^{(i)}(x^*)\| + \|f_i^{(i)}(x^*)\| + \|f_i^{(i)}(x^*)\|.$$

Thus, squaring both sides and using $2ab \leq a^2 + b^2$ with the Lipschitz assumption on the gradients along with the nonexpansiveness property and the boundedness of gradients in the set $\mathcal{X}$ we have

$$\|f_i^{(i)}(v_M(k - 1)) - f_i^{(i)}(v_m(k - 1))\|^2$$

$$\leq 2L^2\|v_M(k - 1) - x^*\|^2 + 2L^2\|v_m(k - 1) - x^*\|^2$$

i.e., we used $f_i(x^*) = 0$. Therefore,

$$\|R_i(k - 1)\|^2$$

$$\leq \alpha_{k-1}^2 \|A^{(i)}\|_\infty^2 \|B^{(i)}\|_\infty^2 \|f_i^{(i)}(v_M(k - 1)) - f_i^{(i)}(v_m(k - 1))\|^2$$

$$\leq \alpha_{k-1}^2 \|A^{(i)}\|_\infty^2 \|B^{(i)}\|_\infty^2 \|f_i^{(i)}(v_M(k - 1)) - x^*\|^2$$

$$+ 2L^2\|v_m(k - 1) - x^*\|^2.$$  (142)

But we have

$$\|v_m(k - 1) - x^*\|^2 \leq \max_{k-1-H \leq k \leq k-1} \|v_q(\hat{k}) - x^*\|^2$$

where $q \in \{1, \ldots, q\}$. 

\[ \|\mathbf{v}_M(k-1) - x^*\|^2 \leq \max_{k-1-H \leq \hat{k} \leq k-1} \|\mathbf{v}_q(\hat{k}) - x^*\|^2 \]

where \( q \in \{1, \ldots, q\} \).

Then
\[ \|R_i(k-1)\| \leq \alpha_{k-1} \|\mathbf{A}^{(i)}\|_\infty \|\mathbf{B}^{(i)}\|_{2,\infty}(2L \max_{k-1-H \leq \hat{k} \leq k-1} \|\mathbf{v}_q(\hat{k}) - x^*\|) \] (143)

where \( q \in \{1, \ldots, n\} \).

and
\[ \|R_i(k-1)\|^2 \leq \alpha_{k-1}^2 \|\mathbf{A}^{(i)}\|_\infty^2 \|\mathbf{B}^{(i)}\|_{2,\infty}^2 (4L^2 \max_{k-1-H \leq \hat{k} \leq k-1, q \in \mathcal{V}} \|\mathbf{v}_q(\hat{k}) - x^*\|^2) \] (144)

and consequently
\[ \sum_{i=1}^{n} \|R_i(k-1)\|^2 \leq \alpha_{k-1}^2 \|\mathbf{A}^{(i)}\|_\infty^2 \|\mathbf{B}^{(i)}\|_{2,\infty}^2 \sum_{i=1}^{n} \max_{k-1-H \leq \hat{k} \leq k-1, q \in \mathcal{V}} \|\mathbf{v}_q(\hat{k}) - x^*\|^2 \] (145)

G. Bounding \( \sum_{i=1}^{n} \|R_i(k-1)\|^2 \)

From Proposition 3 or 4 we have
\[ \sum_{i=1}^{n} \|\mathbf{v}_i(\hat{k}) - x^*\|^2 \leq D \]

Then
\[ \sum_{i=1}^{n} \|R_i(k-1)\|^2 \leq \alpha_{k-1}^2 \|\mathbf{A}^{(i)}\|_\infty^2 \|\mathbf{B}^{(i)}\|_{2,\infty}^2 (4L^2D) \] (146)

\[ \leq \alpha_{k-1}^2 F \]

where
\[ F = 4L^2D \|\mathbf{A}^{(i)}\|_\infty^2 \|\mathbf{B}^{(i)}\|_{2,\infty}^2 < \infty \] (147)
H. Lemmas

**Lemma 9.** [A variant of Lemma 2 in (11)] Let \( Y \subset \mathbb{R}^N \) be a closed convex set. Let the function \( \Phi : \mathbb{R}^N \rightarrow \mathbb{R} \) be convex and differentiable over \( \mathbb{R}^N \) with Lipschitz continuous gradients with a constant \( L \).

Let \( y \) be given by \( y = x - \alpha \nabla \Phi(x) \) for some \( x \in \mathbb{R}^N \), \( \alpha > 0 \).

Then, we have for any \( \hat{x} \in Y \) and \( z \in \mathbb{R}^N \),
\[
\|y - \hat{x}\|^2 \leq (1 + A_\eta \alpha^2)\|x - \hat{x}\|^2 - 2\alpha(\Phi(z) - \Phi(\hat{x})) + \left(\frac{3}{8\eta} + 2\alpha L\right)\|x - z\|^2 + B_\eta \alpha^2\|
\]
where \( A_\eta = 2L^2 + 16\eta L^2 \), \( B_\eta = 2\eta + 8 \) and \( \eta > 0 \) is arbitrary.

A variant of the above lemma (Lemma 2 in (11) or Lemma 4 in [8]). Next we invoke Lemma 7(a) in [9]. which is given by

**Lemma 10.** Let \( Y \subset \mathbb{R}^N \) be a closed convex set. Let the function \( \Phi : \mathbb{R}^N \rightarrow \mathbb{R} \) be convex and differentiable over \( \mathbb{R}^N \) with Lipschitz continuous gradients with a constant \( L \).

Let \( y \) be given by, \( y = \Pi_Y(x - \alpha \nabla \Phi(x)) \) for some \( x \in \mathbb{R}^N \), \( \alpha > 0 \).

Then, we have for any \( \hat{x} \in Y \) and \( z \in \mathbb{R}^N \),
\[
\|y - \hat{x}\|^2 \leq (1 + A_\tau \alpha^2)\|x - \hat{x}\|^2 - 2\alpha(\Phi(z) - \Phi(\hat{x})) - \frac{3}{4}\|y - x\|^2 + \left(\frac{3}{8\tau} + 2\alpha L\right)\|x - z\|^2 + B_\tau \alpha^2\|
\]
where \( A_\tau = 8L^2 + 16\tau L^2 \), \( B_\tau = 8\tau + 8 \) and \( \tau > 0 \) is arbitrary. The proof is similar to Lemma 2 with the imposing of the non-expansive property to obtain the extra term.

We begin with the following result, Lemma 6 in [8].

**Lemma 11.** Let Assumptions [7] and [2] hold. Consider the iterates generated by
\[
\theta_i(k + 1) = \sum_{j=1}^n [W(k)]_{ij}\theta_j(k) + e_i(k) \quad \text{for all } i \in V.
\]

Suppose there exists a non-negative non-increasing sequence \( \{\alpha_k\} \) such that \( \sum_{k=0}^{\infty} \alpha_k \|e_i(k)\| < \infty \) for all \( i \in V \), then for all \( i, j \in V \) we have
\[
\sum_{k=0}^{\infty} \alpha_k \|\theta_i(k) - \theta_j(k)\| < \infty.
\]

REFERENCES

[1] Bertsekas, D. P., A. Nedic, and A. Ozdaglar (2003). Convex analysis and optimization. 2003. Athena Scientific.

[2] Chen, J., R. Monga, S. Bengio, and R. Jozefowicz (2016). Revisiting distributed synchronous sgd. In International Conference on Learning Representations Workshop Track.
[3] Dean, J., G. Corrado, R. Monga, K. Chen, M. Devin, M. Mao, A. Senior, P. Tucker, K. Yang, Q. V. Le, et al. (2012). Large scale distributed deep networks. In *Advances in neural information processing systems*, pp. 1223–1231.

[4] Dutta, S., V. Cadambe, and P. Grover (2016). Short-dot: Computing large linear transforms distributedly using coded short dot products. In *Advances In Neural Information Processing Systems*, pp. 2100–2108.

[5] Ho, Q., J. Cipar, H. Cui, S. Lee, J. K. Kim, P. B. Gibbons, G. A. Gibson, G. Ganger, and E. P. Xing (2013). More effective distributed ml via a stale synchronous parallel parameter server. In *Advances in neural information processing systems*, pp. 1223–1231.

[6] Johansson, B. (2008). *On distributed optimization in networked systems*. Ph. D. thesis, KTH.

[7] Lee, K., M. Lam, R. Pedarsani, D. Papailiopoulos, and K. Ramchandran (2016). Speeding up distributed machine learning using codes. In *Information Theory (ISIT), 2016 IEEE International Symposium on*, pp. 1143–1147. IEEE.

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

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

[10] Li, M., D. G. Andersen, A. J. Smola, and K. Yu (2014). Communication efficient distributed machine learning with the parameter server. In *Advances in Neural Information Processing Systems*, pp. 19–27.

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

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

[13] Rabbat, M. and R. Nowak (2004). Distributed optimization in sensor networks. In *Proceedings of the 3rd international symposium on Information processing in sensor networks*, pp. 20–27. ACM.

[14] Ram, S. S., V. V. Veeravalli, and A. Nedic (2009). Distributed non-autonomous power control through distributed convex optimization. In *INFOCOM 2009, IEEE*, pp. 3001–3005. IEEE.

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

[16] Tandon, R., Q. Lei, A. G. Dimakis, and N. Karampatziakis (2017, 06–11 Aug). Gradient coding: Avoiding stragglers in distributed learning. In D. Precup and Y. W. Teh (Eds.), *Proceedings of the 34th International Conference on Machine Learning*, Volume 70 of *Proceedings of Machine Learning Research*, International Convention Centre, Sydney, Australia, pp. 3368–3376. PMLR.

[17] Zaharia, M., A. Konwinski, A. D. Joseph, R. H. Katz, and I. Stoica (2008). Improving mapreduce performance in heterogeneous environments. In *Osdi*, Volume 8, pp. 7.