Walkman: A Communication-Efficient Random-Walk Algorithm for Decentralized Optimization

Xianghui Mao, Kun Yuan, Yubin Hu, Yuantao Gu, *Senior Member, IEEE*, Ali H. Sayed, *Fellow, IEEE*, and Wotao Yin

***Abstract*—This paper introduces a new algorithm for consensus optimization in a multi-agent network, where all agents collaboratively find a minimizer for the sum of their private functions. All decentralized algorithms rely on communications between adjacent nodes. One class of algorithms use communications between *some or all pairs of adjacent agents* at each iteration. Another class of algorithms uses a *random walk incremental* strategy, which sequentially activates a succession of agents. Existing incremental algorithms require diminishing step sizes to converge to the solution, and their convergence is slow. In this work, we propose a random walk algorithm that uses a fixed step size and converges faster to the solution than the existing random walk incremental algorithms. Our algorithm uses only one link to communicate the latest information from an agent to another. Since this style of communication mimics a man walking in a network, we call our algorithm *Walkman*. We establish convergence for convex and nonconvex objectives. For decentralized least squares, we derive a linear rate of convergence and obtain a better communication complexity than those of other decentralized algorithms. Numerical experiments verify our analysis results.**

***Index Terms*—Consensus optimization, decentralized method, random walk.**

I. Introduction

CONSIDER a directed graph $G = (V, E)$, where $V = \{1, 2, \ldots, n\}$ is the set of agents and E is the set of m edges. We aim to solve the following optimization problem:

$$\min_{x\in\mathbb{R}^p} \quad r(x) + \frac{1}{n}\sum_{i=1}^{n} f_i(x), \tag{1}$$

Manuscript received November 30, 2018; revised June 14, 2019 and November 4, 2019; accepted March 4, 2020. Date of publication March 30, 2020; date of current version May 1, 2020. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Soummya Kar. The work of Xianghui Mao, Yubin Hu, and Yuantao Gu was supported in part by the National Key Research and Development Program of China Project 2017YFC0403600 and in part by the National Natural Science Foundation of China Project 61971266. The work of Ali H. Sayed was supported by NSF under Grant CCF-1524250. The work of Wotao Yin was supported in part by NSF under Grant DMS-1720237 and in part by ONR under Grant N000141712162. *(Corresponding author: Wotao Yin.)*

Xianghui Mao, Yubin Hu, and Yuantao Gu are with the Beijing National Research Center for Information Science and Technology (BNRist) and Department of Electronic Engineering, Tsinghua University, Beijing 100084, China (e-mail: sherrymao92@gmail.com; hu-yb16@mails.tsinghua.edu.cn; gyt@tsinghua.edu.cn).

Kun Yuan is with the Electrical and Computer Engineering Department, University of California, Los Angeles, CA 90095 USA (e-mail: kunyuan@ucla.edu).

Ali H. Sayed is with the Ecole Polytechnique Federale de Lausanne, School of Engineering, 1015 Lausanne, Switzerland (e-mail: ali.sayed@epfl.ch).

Wotao Yin is with the Mathematics Department, University of California, Los Angeles, CA 90095 USA (e-mail: wotaoyin@math.ucla.edu).

Digital Object Identifier 10.1109/TSP.2020.2983167

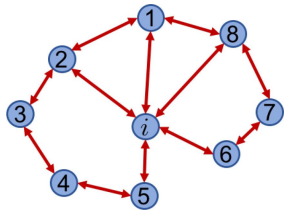


Fig. 1. Communications in the k-th iteration for gossip type methods.

where each f_i is locally held by agent i and r is a globally known regularizer. Both f_i and r can be non-convex. An algorithm is decentralized if it relies only on communications between neighbors (adjacent agents); there is no central node that collects or distributes information to the agents. Decentralize consensus optimization finds applications in various areas including wireless sensor networks, multi-vehicle and multi-robot control systems, smart grid implementations, distributed adaptation and estimation [1], [2], distributed statistical learning [3]–[5] and clustering [6].

A. The Literature

There are several decentralized numerical approaches to solve problem (1) or its special case without the regularizer r. One well-known approach lets *every* agent exchange information with all, or a random subset, of its direct neighbors per iteration. This is illustrated in Fig. 1, where agent i is collecting information from all its neighbors (to update its local variables). This approach includes well-known algorithms such as diffusion [1], [2] and consensus [7], [8], distributed ADMM (D-ADMM) [9]–[13], EXTRA [14], PG-EXTRA [15], gradient tracking [16]–[21], exact diffusion [22], NIDS [23] and beyond. Among them, Push-Sum [24], DEXTRA [25], EXTRAPUSH [26], subgradient-push [27], and Push-Pull methods [20], [21] are designed for directed graphs, while DIGing is for time-varying graphs. The aforementioned algorithms have good convergence rates in the number of iterations. D-ADMM, EXTRA, DIGing, exact diffusion, and NIDS all converge linearly to the exact solution assuming strong convexity and using constant step-sizes. Their communication per iteration is relatively high. Depending on the density of the network, the costs are $O(n)$ computation and $O(n)$–$O(n^2)$ communications per iteration.

To alleviate the communication burden of decentralized optimization methods, another line of works [28]–[30] study the communication pattern illustrated in Fig. 2, specifically, randomly activating one edge for bi-directional communication

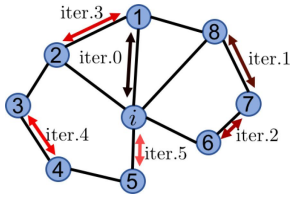


Fig. 2. Communications in 5 adjacent iterations for randomized gossip type methods.

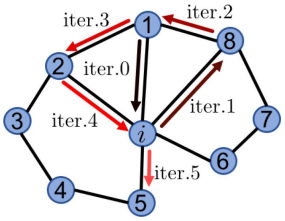


Fig. 3. Communications in 5 adjacent iterations for random walk based methods.

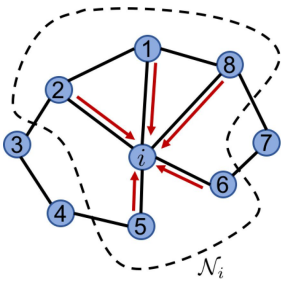


Fig. 4. Communications in the k-th iteration for RW-ADMM.

in each iteration. Among them, randomized gossip algorithms proposed in [28], [29] are designed to solve average consensus problem. More recently, ESDACD [30] implements such random activation to solve general smooth strongly-convex consensus problem. In general, the selected edges are not continuous, some global coordination is required to ensure non-overlapping of iterations.

Another approach is based on the (random) walk (sub) gradient method [31]–[35], where a variable x will move through a (random) succession of agents in the network. At each iteration, the agent i that receives x updates it using one of the subgradients of f_i, followed by sending x to a (random) neighbor. Fig. 3 illustrates the communications along a walk $(1, i, 8, 1, 2, i, 5, \ldots)$. Since only one node and one link are used at each iteration, this approach only costs $O(1)$ computation and $O(1)$ communication per iteration. Thanks to the natural continuity of random walk, it is easy for the involved agents to coordinate. The works [34], [35] apply random walks in the context of adaptive networks and relies on stochastic gradients. If these algorithms use a constant step-size, their iterates converge to a neighborhood of the solution. If the step-size is small, the neighborhood upper bound will be proportionally small but convergence becomes slow. For applications where convergence to the exact solution is required, decaying step-sizes must be used, which leads to slow convergence. The authors of recent work [36] study a mixture of each node exchanging information with all pattern and random walk pattern as shown in Fig. 4, and propose RW-ADMM, where each node in the random walk starts computing after collecting information from all its neighbors. RW-ADMM is proved to converge under constant stepsize on the sacrifice of more communication per iteration.

TABLE I
COMMUNICATION COMPLEXITIES OF VARIOUS ALGORITHMS WHEN SOLVING DECENTRALIZED LEAST SQUARES PROBLEM WITH $\lambda_2(\mathbf{P})$ IS CLOSE TO 1. THE NETWORK HAS n NODES AND m ARCS, $m \in [n, n(n-1)]$, WITH EACH NODE IS CONNECTED BY m/n ARCS. THE QUANTITY ϵ IS THE TARGET ACCURACY, $\mathbf{P}$ IS THE PROBABILITY TRANSITION MATRIX, AND $\lambda_2(\mathbf{P})$ IS THE SECOND LARGEST EIGENVALUE OF $\mathbf{P}$, ONE OF THE MEASURES OF THE CONNECTIVITY OF THE NETWORK

Algorithm	Communication Complexity
Walkman (proposed)	$O\left(\ln\left(\frac{1}{\epsilon}\right)\cdot\frac{n\ln^3(n)}{(1-\lambda_2(\mathbf{P}))^2}\right)$
D-ADMM [11]	$O\left(\ln\left(\frac{1}{\epsilon}\right)\cdot\left(\frac{m}{(1-\lambda_2(\mathbf{P}))^{1/2}}\right)\right)$
EXTRA [14]	$O\left(\ln\left(\frac{1}{\epsilon}\right)\cdot\left(\frac{m}{1-\lambda_2(\mathbf{P})}\right)\right)$
Exact diffusion [22]	$O\left(\ln\left(\frac{1}{\epsilon}\right)\cdot\left(\frac{m}{1-\lambda_2(\mathbf{P})}\right)\right)$
ESDACD [30]	$O\left(\ln(\frac{1}{\epsilon})\cdot\frac{\sqrt{mn}}{\sqrt{(1-\lambda_2(\mathbf{P}))}}\right)$
RW-ADMM [36]	$O\left(\ln\left(\frac{1}{\epsilon}\right)\cdot\frac{m^2}{n\sqrt{(1-\lambda_2(\mathbf{P}))}}\right)$

B. Contribution

In this paper we propose Walkman, a new random walk algorithm[1] for decentralized consensus optimization that uses a *fixed step-size* and converges to the *exact solution*. It is significantly faster than the existing random-walk (sub)gradient incremental methods.

When both r and f_i are possibly non-convex and f_i are Lipschitz differentiable, we show that the iterates x^k generated by Walkman will converge to the stationary point $x^\star$ almost surely. In addition, we establish a linear convergence rate for decentralized least squares.

Walkman is communication efficient. For decentralized least squares, the communication complexity of Walkman compares favorably with existing popular algorithms. In Walkman, the activation of agents follows a Markov chain – the probability to activate adjacent agent i_{k+1} depends only on the current agent i_k, which is defined as $p(i_{k+1}|i_k)$. Define the transition probability matrix $\mathbf{P} \in \mathbb{R}^{n\times n}$ where $[\mathbf{P}]_{ij} = p(i_{k+1} = j | i_k = i) \in [0, 1]$. Matrix $\mathbf{P}$ models the communication pattern of Walkman and affects its convergence performance. The communication complexity of various decentralized algorithms are summarized in Table I. We show that, if

$$\lambda_2(\mathbf{P}) \leq 1 - \frac{\ln^{4/3}(n)}{m^{2/3}} \approx 1 - \frac{1}{m^{2/3}}, \tag{2}$$

which implies the connectivity of the network is at least moderate, then Walkman uses less communication than *all* the state-of-the-art decentralized algorithms listed in the table.

C. Discussion

Walkman is a random-walk algorithm. Its efficiency depends on how long it takes the walk to visit all the agents. This time is known as the *cover time*. When Walkman only needs to visit every agent at least once (which is the case to compute the

[1]Communication pattern is shown in Fig. 3.

consensus average), the cover time is exactly the complexity of Walkman. For the cover times of random walks in various graphs, we refer the reader to [37, Chapter 11].

For more general problems, Walkman must visit each agent infinitely many times to converge. Its efficiency depends on how frequently all of the agents are revisited. For a random walk, this can be described by the *mixing time* of the underlying Markov chain. Next, we present relevant assumptions.

Assumption 1: The random walk $(i_k)_{k\geq 0}$, $i_k \in V$, forms an irreducible and aperiodic Markov chain with transition probability matrix $\mathbf{P} \in \mathbb{R}^{n\times n}$ where $[\mathbf{P}]_{ij} = p(i_{k+1} = j | i_k = i) \in [0, 1]$ and stationary distribution π satisfying $\pi^T \mathbf{P} = \pi^T$.

If the underlying network is a complete graph, we can choose P so that $P_{ij} = p(i_{k+1} = j | i_k = i) = \frac{1}{n}$ for all $i, j \in V$, a case analyzed in [38, §2.6.1] (barring asynchronicity therein). For a more general network that is connected, we need the *mixing time* (for given $\delta > 0$), which is defined as the smallest integer $\tau(\delta)$ such that, for all $i \in V$,

$$\left\| [\mathbf{P}^{\tau(\delta)}]_{i,:} - \pi^{\mathsf{T}} \right\| \leq \delta \pi_*, \tag{3}$$

where $\pi_* := \min_{i\in V} \pi_i$, and $[\mathbf{P}^{\tau(\delta)}]_{i,:}$ denotes the ith row of $\mathbf{P}^{\tau(\delta)}$. This inequality states: regardless of current state i and time k, the probability of visiting each state j after $\tau(\delta)$ more steps is $(\delta\pi_*)$-close to π_j, that is, for all $i, j \in V$,

$$\left| [\mathbf{P}^{\tau(\delta)}]_{ij} - \pi_j \right| \leq \delta \pi_*. \tag{4}$$

A good reference for mixing time is [37]. The mixing time requirement, inequality (3), is guaranteed to hold for[2]

$$\tau(\delta) := \left\lceil \frac{1}{1 - \sigma(\mathbf{P})} \ln \frac{\sqrt{2}}{\delta \pi_*} \right\rceil \tag{5}$$

for $\sigma(\mathbf{P}) := \sup\{\|f^{\mathsf{T}}\mathbf{P}\|/\|f\| : f^{\mathsf{T}}\mathbf{1} = 0, f \in \mathbb{R}^n\}$.

We will use inequality (4) to show the sufficient descent of a Lyapunov function L_β^k.[3] A similar Lyapunov function has been used in [39] and extended in [40]. However, the analyses in [39], [40] only help us show $L_\beta^k \geq L_\beta^{k+1}$ and the existence of a lower bound. Because a random walk $(i_k)_{k\geq 0}$ is neither essentially cyclic nor i.i.d. random, we must use a new analytic technique, which is motivated by the recent paper [41]. This new technique integrates mixing-time bounds with a conventional line of convergence analysis.

For decentralized least squares, we give the communication complexity bound of Walkman in term of $\sigma(\mathbf{P})$. This quantity also determines the communication complexity bounds of D-ADMM, EXTRA, and exact diffusion. Therefore, we can compare their communication complexities. For moderately well connected networks, we show in §V that the bound of Walkman is the lowest.

[2]Here is a trivial proof. For any $k \geq 1$, by definition, it holds $[\mathbf{P}^k]_{i,:} - \pi^{\mathsf{T}} = ([\mathbf{P}^{k-1}]_{i,:} - \pi^{\mathsf{T}})\mathbf{P}$, and $([\mathbf{P}^k]_{i,:} - \pi^{\mathsf{T}})\mathbf{1} = 0$. Hence, $\|[\mathbf{P}^k]_{i,:} - \pi^{\mathsf{T}}\| \leq \|[\mathbf{P}^{k-1}]_{i,:} - \pi^{\mathsf{T}}\|\sigma(\mathbf{P}) \leq \cdots \leq \|\mathbf{I}_{i,:} - \pi^{\mathsf{T}}\|\sigma^k(\mathbf{P})$. We can bound $\|\mathbf{I}_{i,:} - \pi^{\mathsf{T}}\|^2 \leq (1 - \pi_i)^2 + \sum_{j\neq i} \pi_j^2 \leq (1 - \pi_*)^2 + (1 - \pi_*)^2 = 2(1 - \pi_*)^2$. Therefore, by ensuring $\sqrt{2}(\sigma(\mathbf{P}))^{\tau(\delta)}(1 - \pi_*) \leq \delta\pi_*$, which simplifies to condition (5) by Taylor series, we guarantee (3) to hold.

[3]See equation (16) for the definition of L_β^k.

Even though D-ADMM, EXTRA, and exact diffusion use more total communications, their communications are simultaneous over different edges, so they *may* take less *total time*. However, this time will increase and even overpass the Walkman time if different edges have different communication latencies and bandwidths, and if synchronization overhead is included. In an ideal situation where every communication takes the same amount of time and synchronization has no overhead, Walkman is found to be slower in time, unsurprisingly.

Although this paper does not discuss data privacy, Walkman protects privacy better than diffusion, consensus, D-ADMM, etc., since the communication path is random and hence instead of periodically communicating local information, only the current iterate x^k is sent out by the active agent. Updating history is neither sent nor traceable.

The limitation of this paper lies in that the linear convergence rate analysis applies only to least squares (though convergence and a sublinear convergence rate are established for more general problems) and that the transition matrix is stationary. They need more space to address in our future work. Another direction to generalize this work is to create multiple simultaneous random walks, which may reduce the total solution time. The information exchange across random walks will require careful design and analysis.

In the rest of this paper, §II derives Walkman, §III presents the main convergence result and the key lemmas, §IV focuses on least squares and obtains its linear convergence rate of Walkman, §V analyzes communication complexities and make comparisons between Walkman and other algorithms, §VI presents numerical simulation results, and finally §VII summarizes the findings of this paper.

II. Derivation of Walkman

Walkman can be derived by modifying existing algorithms to use a random walk, for example, ADMM [42], [43] or PPG [44]. By defining

$$y := \mathrm{col}\{y_1, y_2, \ldots, y_n\} \in \mathbb{R}^{np},\ F(y) := \sum_{i=1}^n f_i(y_i), \tag{6}$$

where the operation $\mathrm{col}(\cdot)$ refers to stacking vectors column by column, we can compactly rewrite problem (1) as

$$\begin{aligned} \min_{x,y} \quad & r(x) + \frac{1}{n} F(y), \\ \text{subject to} \quad & \mathbb{1} \otimes x - y = 0, \end{aligned} \tag{7}$$

where $\mathbb{1} = [1\ 1\ \ldots\ 1]^{\mathsf{T}} \in \mathbb{R}^n$ and $\otimes$ is the Kronecker product. The constraint is equivalent to $x - y_i = 0$ for $i = 1, \ldots, n$. The augmented Lagrangian for problem (7) is

$$\begin{aligned} L_\beta(x, y; z) := r(x) + \frac{1}{n}\Bigg(& F(y) + \langle z, \mathbb{1} \otimes x - y \rangle \\ & + \frac{\beta}{2} \|\mathbb{1} \otimes x - y\|^2 \Bigg), \end{aligned} \tag{8}$$

where $z := \mathrm{col}\{z_1, \ldots, z_n\} \in \mathbb{R}^{np}$ is the dual variable (Lagrange multipliers) and $\beta > 0$ is a constant parameter.

The standard ADMM algorithm is an iteration that minimizes $L_\beta(x,y;z)$ in x, then in y, and finally updates z. Applying ADMM to problem (7) yields (not our algorithm)

$$\bar{x}^{k+1} = \frac{1}{n}\sum_{i=1}^{n}\left(y_i^k - \frac{z_i^k}{\beta}\right), \tag{9a}$$

$$x^{k+1} = \mathbf{prox}_{\frac{1}{\beta}r}(\bar{x}^{k+1}), \tag{9b}$$

$$y_i^{k+1} = \mathbf{prox}_{\frac{1}{\beta}f_i}\left(x^{k+1} + \frac{z_i^k}{\beta}\right), \quad \forall i \in V \tag{9c}$$

$$z_i^{k+1} = z_i^k + \beta(x^{k+1} - y_i^{k+1}), \quad \forall i \in V \tag{9d}$$

where the proximal operator is defined as $\mathbf{prox}_{\gamma f}(x) := \arg\min_y f(y) + \frac{1}{2\gamma}\|y-x\|_2^2$. Since computing the sum in (9a) needs information from all the agents, it is too expensive to realize in a decentralized fashion. However, if each ADMM iteration updates only y_{i_k} and z_{i_k} in (9c) and (9d), keeping the remaining $\{y_i\}_{i\neq i_k}, \{z_i\}_{i\neq i_k}$ unchanged, the algorithm then changes to:

$$x^{k+1} = \mathbf{prox}_{\frac{1}{\beta}r}(\bar{x}^{k+1}), \tag{10a}$$

$$y_i^{k+1} = \begin{cases} \mathbf{prox}_{\frac{1}{\beta}f_i}\left(x^{k+1} + \frac{z_i^k}{\beta}\right), & i = i_k \\ y_i^k, & \text{otherwise} \end{cases} \tag{10b}$$

$$z_i^{k+1} = \begin{cases} z_i^k + \beta(x^{k+1} - y_i^{k+1}), & i = i_k \\ z_i^k, & \text{otherwise} \end{cases} \tag{10c}$$

$$\bar{x}^{k+2} = \bar{x}^{k+1} + \frac{1}{n}\left(y_{i_k}^{k+1} - \frac{z_{i_k}^{k+1}}{\beta}\right) - \frac{1}{n}\left(y_{i_k}^{k} - \frac{z_{i_k}^{k}}{\beta}\right). \tag{10d}$$

If we initialize $\{y_i^0\}_{i=1}^n$ and $\{z_i^0\}_{i=1}^n$ so that

$$\bar{x}^1 = \frac{1}{n}\sum_{i=1}^{n}\left(y_i^0 - \frac{z_i^0}{\beta}\right) = 0, \tag{11}$$

for example, by simply setting $y_i^0 = 0$ and $z_i^0 = 0, i = 1,\dots,n$, then with only the i_k-th part of variables y and z updated in each (10), mathematical induction implies that (10d) automatically maintains

$$\bar{x}^{k+1} = \frac{1}{n}\sum_{i=1}^{n}\left(y_i^k - \frac{z_i^k}{\beta}\right).$$

Note that, the second equation of the initialization condition in (11) can be conducted locally, whereas the constraint on $\bar{x}^1$ only involves the agent where the random walk starts. Therefore, a simple initialization satisfying (11) can be realized without any "consensus"-type preprocessing. We call (10) Walkman. Its decentralized implementation is presented in Algorithm 1. The variable $\bar{x}^k$ is updated by agent i_k and passed as a token to agent i_{k+1}.

Use ∇f_i instead of $\mathbf{prox}_{f_i}$. If the regularizer r is proximable, i.e., $\mathbf{prox}_{\gamma r}$ can be computed in $O(n)$ or $O(n\text{polylog}(n))$ time, the computational resources are mainly consumed on solving the minimization problem in step (10b). We can avoid it by using the cheaper gradient descent, like in diffusion, consensus, EXTRA, DIGing, exact diffusion, and NIDS. If f_i is differentiable, we replace (10b) with the update:

$$y_i^{k+1} = \begin{cases} x^{k+1} + \frac{1}{\beta}z_i^k - \frac{1}{\beta}\nabla f_i(y_i^k), & i = i_k \\ y_i^k, & \text{otherwise.} \end{cases} \tag{10b'}$$

Compare to (10b), update (10b') saves computations but can cause more iterations and thus more total communications. One can choose between (10b) and (10b') based on computation and communication tradeoffs in applications. In the next section, we are going to analyze their performance.

Algorithm 1: Walkman.

Initialization: initialize y_i^0 and z_i^0 so that (11) holds;
Repeat for $k = 0, 1, 2, \dots$ until convergence
  agent i_k do:
    update x^{k+1} according to (10a);
    update $y_{i_k}^{k+1}$ according to (10b) or (10b');
    update $z_{i_k}^{k+1}$ according to (10c);
    update $\bar{x}^{k+2}$ according to (10d);
    send $\bar{x}^{k+2}$ via edge (i_k, i_{k+1}) to agent i_{k+1};
End

III. Convergence

In this section we present convergence of Walkman based on the following assumptions.

Assumption 2: The objective function in original problem (1), $r(x) + \frac{1}{n}\sum_{i=1}^n f_i(x)$, is bounded from below by $\underline{f}$ and is coercive over $\mathbb{R}^p$, that is, $r(x) + \frac{1}{n}\sum_{i=1}^n f_i(x) \to \infty$ for any sequence $\{x^k\}_{k\geq 0} \subset \mathbb{R}^p$ with $\|x^k\| \overset{k\to\infty}{\to} \infty$.

Assumption 2 is *not* over $\mathbb{R}^{np}$ but $\mathbb{R}^p$, so it is easy to satisfy.

Assumption 3: Each $f_i(x)$ is L-Lipschitz differentiable, that is, for any $u, v \in \mathbb{R}^p$,

$$\|\nabla f_i(u) - \nabla f_i(v)\| \leq L\|u - v\|, \quad i = 1,\dots,n. \tag{12}$$

Assumption 4: The lower semi-continuous function $r(x)$ is γ-semiconvex, that is, $r(\cdot) + \frac{\gamma}{2}\|\cdot\|^2$ is convex or equivalently,

$$r(y) + \frac{\gamma}{2}\|y-x\|^2 \geq r(x) + \langle d, y-x\rangle, \forall x, y, \forall d \in \partial r(x). \tag{13}$$

We first introduce the notation used in our analysis. The first time that the Markov chain $(i_k)_{k\geq 0}$ hits agent i is denoted as $T_i := \min\{k : i_k = i\}$, and their max over i is

$$T := \max\{T_1, \dots, T_n\}. \tag{14}$$

By iteration T, every agent has been visited at least once. Based on Assumption 1, the Markov chain is positive recurrent and, therefore, $\Pr(T < \infty) = 1$. For $k > T$, let $\tau(k,i)$ denote the iteration of the last visit to agent i before k, that is,

$$\tau(k,i) := \max\{k' : i_{k'} = i, k' < k\}. \tag{15}$$

Next, we define two separate Lyapunov functions for Walkman updating y using (10b) (computing $\mathbf{prox}_{\frac{1}{\beta}f_i}$) and (10b')

(computing $\nabla f_i(y_i^k)$):

$$L_\beta^k := L_\beta(x^k, y^k; z^k), \tag{16}$$

$$M_\beta^k := L_\beta^k + \frac{L^2}{n}\sum_{i=1}^n \|y_i^{\tau(k,i)+1} - y_i^{\tau(k,i)}\|^2, \tag{17}$$

where $L_\beta(x, y; z)$ is defined in (8). We establish the descent of L_β^k (resp. M_β^k) for Walkman using (10b) (resp. (10b')).

Lemma 1: Under Assumptions 2, 3, and 4, the iterates $(x^k, y^k, z^k)_{k\geq 0}$ generated by Walkman (10), or Algorithm 1, satisfy the following properties:

1) for (10b) and $\beta \geq \max\{\gamma, 2L+2\}$, $(L_\beta^k)_{k\geq 0}$ is lower bounded and convergent;
1') for (10b') and $\beta > \max\{\gamma, 2L^2+L+2\}$, $(M_\beta^k)_{k\geq 0}$ is lower bounded and convergent;
2) for Walkman with either (10b) or (10b'), the sequence $(x^k, y^k, z^k)_{k\geq 0}$ is bounded.

See the Appendix for a proof. Based on Lemma 1, we establish the convergence of subgradients of L_β^k.

Lemma 2: Take Assumptions 1–4 and Walkman with β given in Lemma 1. For any given subsequence (including the whole sequence) with its index $(k_s)_{s\geq 0}$, there exists a sequence $\{g^k\}_{k\geq 0}$ with $g^k \in \partial L_\beta^{k+1}$ containing an almost surely convergent subsubsequence $(g^{k_{s_j}})_{j\geq 0}$, that is,

$$\Pr\left(\lim_{j\to\infty} \|g^{k_{s_j}}\| = 0\right) = 1.$$

Proof: The proof sketch is summarized as follows.

1) We construct $g^k \in \partial L_\beta^{k+1}$ and show that its subvector $q_i^k := (g_x^k, g_{y_i}^k, g_{z_i}^k)$ satisfies $\lim_{k\to\infty} \mathbb{E}\|q_{i_k}^{k-\tau(\delta)-1}\|^2 = 0$, where the mixing time $\tau(\delta)$ is defined in (5).
2) For $k \geq 0$, define the filtration of sigma algebras:

$$\chi^k := \sigma\left(x^0, \ldots, x^k, y^0, \ldots, y^k, z^0, \ldots, z^k, i_0, \ldots, i_k\right).$$

We show that

$$\mathbb{E}\left(\|q_{i_k}^{k-\tau(\delta)-1}\|^2 \middle| \chi^{k-\tau(\delta)}\right) \geq (1-\delta)\pi_*\|g^{k-\tau(\delta)-1}\|^2,$$

where π_* is the minimal value in the Markov chain's stationary distribution. From this bound and the result in step 1), we can get $\lim_{k\to\infty}\mathbb{E}\|g^k\| = 0$.
3) From the result in the last step, we use some inequalities and the Borel-Cantelli lemma to obtain an almost surely convergent subsubsequence of g^k.

The details of these steps are given in the Appendix. ■

Theorem 1: Under Assumptions 1–4, for $\beta > \max\{\gamma, 2L+2\}$ (resp. $\beta > \max\{\gamma, 2L^2+L+2\}$), it holds that any limit point (x^*, y^*, z^*) of the sequence (x^k, y^k, z^k) generated by Walkman with (10b) (resp. (10b')) satisfies: $x^* = y_i^*$, $i = 1, \ldots, n$, where x^* is a stationary point of (1), with probability 1, that is,

$$\Pr\left(0 \in \partial r(x^*) + \frac{1}{n}\sum_{i=1}^n \nabla f_i(x^*)\right) = 1. \tag{18}$$

If the objective of (1) is convex, then x^* is a minimizer.

Proof: By statement 2) of Lemma 1, the sequence (x^k, y^k, z^k) is bounded so there exists a convergent subsequence $(x^{k_s}, y^{k_s}, z^{k_s})$ converging to a limit point (x^*, y^*, z^*) as $s \to \infty$. By continuity, we have

$$L_\beta(x^*, y^*, z^*) = \lim_{s\to\infty} L_\beta(x^{k_s}, y^{k_s}, z^{k_s}). \tag{19}$$

Lemma 2 finds a subsubsequence $g^{k_{s_j}} \in \partial L_\beta^{k+1}$ such that $\Pr(\lim_{j\to\infty}\|g^{k_{s_j}}\| = 0) = 1$. By the definition of *general subgradient* (cf. [45, Def. 8.3]), we have $0 \in \partial L_\beta(x^*, y^*, z^*)$.

This completes the proof of Theorem 1. ■

Next, we derive the convergence rate for Walkman with a specific initialization, $z_i^0 = \nabla f_i(y_i^0)$. Specifically, to avoid consensus preprocessing, we need $\nabla f_i(y_i^0) = \beta y_i^0$. In other words, y_i^0 is a stationary point for the problem $\min_{y\in\mathbb{R}^p} f_i(y) - \frac{\beta}{2}\|y\|^2$. This preprocessing can be accomplished without communication.

Theorem 2: [Gradient sublinear convergence] Consider Walkman using either (10b) or (10b') update. Under Assumptions 1–4, with β given in Lemma 1, and local variables initialized as $\nabla f_i(y_i^0) = \beta y_i^0 = z_i^0$, $\forall i \in \{1, \ldots, n\}$, there exists a sequence $\{g^k\}_{k\geq 0}$ with $g^k \in \partial L_\beta^{k+1}$ satisfying

$$\min_{k\leq K} \mathbb{E}\|g^k\|^2 \leq \frac{\bar{C}}{K}(L_\beta^0 - \underline{f}),\ \forall K > \tau(\delta) + 2, \tag{20}$$

where $\bar{C}$ is a constant merely depending on β, L, γ and $n, \tau(\delta)$. With β, L, γ independent from the network structure, one has $\bar{C} \sim O(\frac{\tau(\delta)^2+1}{(1-\delta)n\pi_*})$, where $\tau(\delta)$ is defined in (5).

Proof: The detailed proof can be found in Appendix C. ■

It is possible, though more cumbersome, to show a sublinear convergence rate under a more general initialization. We decided not to pursue it.

IV. Linear Convergence for Least Squares

In this section, we focus on the decentralize least-squares problem:

$$\begin{aligned} \min \quad & \frac{1}{n}\sum_{i=1}^n \frac{1}{2}\|\mathbf{A}_i y_i - b_i\|^2, \\ \text{subject to} \quad & y_1 = y_2 = \cdots = y_n = x, \end{aligned} \tag{21}$$

which is a special case (7) with regularizer $r = 0$, local objective $f_i(y_i) := \frac{1}{2}\|\mathbf{A}_i y_i - b_i\|^2$ and gradient $\nabla f_i(y_i) = \mathbf{A}_i^\intercal(\mathbf{A}_i y_i - b_i)$. The Lipschitz constant L in Assumption 3 equals $\sigma_{\max}^* := \max_i \sigma_{\max}(\mathbf{A}_i^\intercal \mathbf{A}_i)$, where $\sigma_{\max}(\cdot)$ takes largest eigenvalue. To assure that there exists a single optimum to problem (21), the following analysis is based on the assumption that the matrix $\sum_{i=1}^n \mathbf{A}_i^\intercal \mathbf{A}_i$ is reversible, which implies (21) is strongly convex.

We apply Walkman (or Algorithm 1) updating with $\mathbf{prox}_{f_i}$, i.e., utilizing (10b), and starting from

$$y_i^0 = (\mathbf{A}_i^\intercal \mathbf{A}_i - \beta\mathbf{I})^{-1}(\mathbf{A}_i^\intercal b_i),\ \forall i \in V, \tag{22}$$

$$z_i^0 = \nabla f_i(y_i^0) = \mathbf{A}_i^\intercal(\mathbf{A}_i y_i^0 - b_i),\ \forall i \in V, \tag{23}$$

where (22) is well defined for $\beta > \max_i \sigma_{\max}(\mathbf{A}_i^\intercal \mathbf{A}_i)$. This is to ensure $y_i^0 - z_i^0/\beta = 0$ and thus (11) for all $k \geq 0$.

We analyze the complexities of Walkman for problem (21) based on the Lyapunov function $h_\beta(\mathcal{Y}) : \mathbb{R}^{np} \to \mathbb{R}$,

$$h_\beta(\mathcal{Y}) := \frac{1}{n}\sum_{i=1}^{n}\left(\frac{\beta}{2}\|y_i^k\|^2 - \frac{1}{2}\|\mathbf{A}_i y_i^k\|^2 + \frac{1}{2}\|b_i\|^2\right) - \frac{\beta}{2}\|\mathbf{T}\mathcal{Y} + c\|^2, \tag{24}$$

where $\mathbf{T} := \frac{1}{n}[(\mathbf{I} - \frac{1}{\beta}\mathbf{A}_1^\mathsf{T}\mathbf{A}_1), \ldots, (\mathbf{I} - \frac{1}{\beta}\mathbf{A}_n^\mathsf{T}\mathbf{A}_n)] \in \mathbb{R}^{p\times np}$ and $c := \frac{1}{n\beta}\sum_{i=1}^{n}\mathbf{A}_i^\mathsf{T} b_i \in \mathbb{R}^p$. The following lemma relates $h_\beta(\mathcal{Y})$ and the augmented Lagrangian sequence.

Lemma 3: With initialization (22) and (23), it holds that

$$h_\beta(\mathcal{Y}^k) = L_\beta(x^{k+1}, \mathcal{Y}^k; \mathcal{Z}^k). \tag{25}$$

Proof: From the optimality condition of (10b), we can verify

$$\mathbf{A}_{i_k}^\mathsf{T}(\mathbf{A}_{i_k} y_{i_k}^{k+1} - b_{i_k}) = \beta x^{k+1} + z_{i_k}^k - \beta y_{i_k}^{k+1} \overset{(a)}{=} z_{i_k}^{k+1}, \tag{26}$$

for $k \geq 1$, where (a) follows from (10c). In Walkman, each pair of y_i and z_i is either updated together, or both not updated. Then by applying (23) and (26), we get

$$z_i^k = \mathbf{A}_i^\mathsf{T}(\mathbf{A}_i y_i^k - b_i),\ \forall i \in V,\ k \geq 0. \tag{27}$$

Substituting (27) into (10d) and (10a) yields $x^{k+1} = \mathbf{T}\mathcal{Y}^k + c, \quad \forall k \geq 0$. Eliminating z_i^k and x^{k+1} in $L_\beta(x^{k+1}, \mathcal{Y}^k; \mathcal{Z}^k)$ using the above formulas produces (25). ■

The following lemma establishes that $h_\beta(\mathcal{Y})$ is strongly convex and Lipschitz differentiable.

Lemma 4: For a network with $n \geq 2$ agents, and the parameter $\beta > \sigma_{\max}^*$, where $\sigma_{\max}^* := \max_i \sigma_{\max}(\mathbf{A}_i^\mathsf{T}\mathbf{A}_i)$, the function $h_\beta(\cdot)$ is

1) strongly convex with modulus $\nu = \frac{(n-1)(\beta - \sigma_{\max}^*)}{n^2}$, and
2) Lipschitz differentiable with Lipschitz constant $\bar{L} = \frac{\beta}{n}(1 - \frac{1}{n}(1 - \frac{\sigma_{\max}^*}{\beta})^2)$.

Proof: As a quadratic function, $h_\beta(\cdot)$ is ν-strongly convex with $\bar{L}$-Lipschitz gradients if, and only if, its Hessian (by (24)) $\mathbf{H}$ satisfies

$$\nu\mathbf{I} \preceq \mathbf{H} := \frac{\beta}{n}\mathbf{I}_{np} - \frac{1}{n}\mathbf{A} - \beta\mathbf{T}^\mathsf{T}\mathbf{T} \preceq \bar{L}\mathbf{I}, \tag{28}$$

where $\mathbf{A} := \mathrm{diag}(\mathbf{A}_1^\mathsf{T}\mathbf{A}_1, \mathbf{A}_2^\mathsf{T}\mathbf{A}_2, \ldots, \mathbf{A}_n^\mathsf{T}\mathbf{A}_n)$. With $\beta > \max_i \sigma_{\max}(\mathbf{A}_i^\mathsf{T}\mathbf{A}_i)$, we define the symmetric positive definite matrices $\mathbf{D}_i := (\mathbf{I} - \frac{1}{\beta}\mathbf{A}_i^\mathsf{T}\mathbf{A}_i)^{1/2}$ for $i \in V$. The spectral norm of $\mathbf{D}_i$ satisfies

$$\left(1 - \frac{\sigma_{\max}^*}{\beta}\right)^{\frac{1}{2}} \leq \left(1 - \frac{\sigma_{\max}(\mathbf{A}_i^\mathsf{T}\mathbf{A}_i)}{\beta}\right)^{\frac{1}{2}} \leq \|\mathbf{D}_i\| \leq 1. \tag{29}$$

Stacking $\mathbf{D}_i$'s into

$$\mathbf{D} := \begin{bmatrix}\mathbf{D}_1 \\ \vdots \\ \mathbf{D}_n\end{bmatrix}. \tag{30}$$

Then, for any vector $w := \mathrm{col}\{w_1, \ldots, w_n\} \in \mathbb{R}^{np}$ where $w_i \in \mathbb{R}^p$, we have the interval bounds for $\|\mathrm{diag}(\mathbf{D})w\|$:

$$\|\mathrm{diag}(\mathbf{D})w\| = \left\|\begin{bmatrix}\mathbf{D}_1 w_1 \\ \vdots \\ \mathbf{D}_n w_n\end{bmatrix}\right\| \tag{31}$$

$$\in \left[\left(1 - \frac{\sigma_{\max}^*}{\beta}\right)^{\frac{1}{2}}\|w\|, \|w\|\right]. \tag{32}$$

It is easy to check

$$w^\mathsf{T}\mathbf{H}w = \frac{\beta}{n}(\mathrm{diag}(\mathbf{D})w)^\mathsf{T}\left(\mathbf{I} - \frac{1}{n}\mathbf{D}^\mathsf{T}\mathbf{D}\right)(\mathrm{diag}(\mathbf{D})w). \tag{33}$$

Therefore, we get (28) from

$$w^\mathsf{T}\mathbf{H}w \geq \frac{\beta}{n}\left(1 - \frac{1}{n}\right)\|\mathrm{diag}(\mathbf{D})w\|^2 \tag{34}$$

$$\geq \underbrace{\frac{\beta}{n}\left(1 - \frac{1}{n}\right)\left(1 - \frac{\sigma_{\max}^*}{\beta}\right)}_{\nu}\|w\|^2 \tag{35}$$

and

$$w^\mathsf{T}\mathbf{H}w \leq \frac{\beta}{n}\left(\|\mathrm{diag}(\mathbf{D})w\|^2 - \frac{1}{n}\right) \tag{36}$$

$$\leq \frac{\beta}{n}\left(\|w\|^2 - \frac{1}{n}\left(1 - \frac{\sigma_{\max}^*}{\beta}\right)^2\|w\|^2\right) \tag{37}$$

$$= \underbrace{\frac{\beta}{n}\left(1 - \frac{1}{n}\left(1 - \frac{\sigma_{\max}^*}{\beta}\right)^2\right)}_{\bar{L}}\|w\|^2. \tag{38}$$

■

Lemma 5: With $\beta > \sigma_{\max}^*$, the unique minimizer of $h_\beta(\cdot)$ is $\mathcal{Y}^\star := \mathrm{col}\{y_1^\star, \ldots, y_n^\star\}$ with $y_i^\star \equiv x^\star = (\sum_{i=1}^n \mathbf{A}_i^\mathsf{T}\mathbf{A}_i)^{-1}(\sum_{i=1}^n \mathbf{A}_i^\mathsf{T} b_i)$. These components are also the unique solution to (21), as well as the unique minimizer of $\sum_{i=1}^n \frac{1}{2}\|\mathbf{A}_i x - b_i\|^2$.

Proof: Since $\mathcal{Y}^*$ must satisfy $\nabla h_\beta(\mathcal{Y}^*) = 0$, we have

$$\nabla_i h_\beta(\mathcal{Y}^*) = \frac{\beta}{n}\left(y_i^\star - \frac{1}{\beta}\mathbf{A}_i^\mathsf{T}\mathbf{A}_i y_i^\star\right) - \frac{\beta}{n}\left(\mathbf{I} - \frac{1}{\beta}\mathbf{A}_i^\mathsf{T}\mathbf{A}_i\right)(\mathbf{T}\mathcal{Y}^\star + c)$$

$$= \frac{\beta}{n}\left(\mathbf{I} - \frac{1}{\beta}\mathbf{A}_i^\mathsf{T}\mathbf{A}_i\right)(y_i^\star - \mathbf{T}\mathcal{Y}^\star - c) = 0. \tag{39}$$

Since $\mathbf{I} - \frac{1}{\beta}\mathbf{A}_i^\mathsf{T}\mathbf{A}_i \succ 0$ with $\beta > \sigma_{\max}^*$, we conclude

$$y_i^\star - \mathbf{T}\mathcal{Y}^\star - c = 0,\ \forall i = 1, \ldots, n, \tag{40}$$

which implies $y^\star$ given in the Lemma. It is easy to verify the rest of the Lemma using optimality conditions. ■

Define **one epoch** as $\tau(\delta)$ iterations, and let

$$h_\beta^\star := \min_{\mathcal{Y}}\{h_\beta(\mathcal{Y})\}, \quad F_t := \mathbb{E}h_\beta(\mathcal{Y}^{t\tau(\delta)}) - h_\beta^\star, \tag{41}$$

where we use t to index an epoch. The next lemma is fundamental to the remaining analysis.

Lemma 6: Under Assumption 1 and $\beta > 2\sigma^*_{\max} + 2$, for any $\delta > 0$, we have

$$F_t^2 \le \frac{2\beta^2\tau(\delta)}{n(1-\delta)\pi_*}\left(F_t - F_{t+1}\right)\cdot \mathbb{E}\|y^{t\tau(\delta)} - y^\star\|^2, \tag{42}$$

where $\tau(\delta)$ is defined in (5).

Proof: We first upper bound $\|\nabla h_\beta(y^k)\|^2$. Verify

$$\nabla_i h_\beta(y^k) = \frac{\beta}{n}\mathbf{D}_i^2\left(y_i^k - \mathbf{T}y^k - c\right). \tag{43}$$

Investigate step (10b) for $i = i_k$ as

$$\begin{aligned} y_i^{k+1} &= \arg\min_y \frac{1}{2}\|\mathbf{A}_i y - b_i\|^2 + \frac{\beta}{2}\left\|y - x^{k+1} - \frac{1}{\beta}z_i^k\right\|^2 \\ &= \left(\mathbf{A}_i^\intercal\mathbf{A}_i + \beta\mathbf{I}\right)^{-1}\left(\mathbf{A}_i^\intercal b_i + \beta x^{k+1} + z_i^k\right) \\ &\overset{(a)}{=} \left(\mathbf{A}_i^\intercal\mathbf{A}_i + \beta\mathbf{I}\right)^{-1}\left(\beta\mathbf{T}y^k + \beta c + \mathbf{A}_i^\intercal\mathbf{A}_i y_i^k\right) \\ &= y_i^k + \left(\mathbf{I} + \frac{1}{\beta}\mathbf{A}_i^\intercal\mathbf{A}_i\right)^{-1}\left(\mathbf{T}y^k + c - y_i^k\right) \\ &\overset{(43)}{=} y_i^k - \frac{n}{\beta}\left(\mathbf{I} + \frac{1}{\beta}\mathbf{A}_i^\intercal\mathbf{A}_i\right)^{-1}\mathbf{D}_i^{-2}\left(\nabla_i h_\beta(y^k)\right), \end{aligned} \tag{44}$$

where (a) follows from (27) and $\mathbf{T}$'s definition. Thence,

$$\begin{aligned} \|\nabla_{i_k} h_\beta(y^k)\| &= \frac{\beta}{n}\left\|\left(\mathbf{I} - \frac{1}{\beta^2}(\mathbf{A}_{i_k}^\intercal\mathbf{A}_{i_k})^2\right)(y_{i_k}^{k+1} - y_{i_k}^k)\right\| \\ &\le \frac{\beta}{n}\|y^{k+1} - y^k\|, \end{aligned} \tag{45}$$

For any $k \ge \tau(\delta) - 1$, we further have

$$\begin{aligned} &\|\nabla_{i_k} h_\beta(y^{k-\tau(\delta)+1})\|^2 \\ &= \|\nabla_{i_k} h_\beta(y^{k-\tau(\delta)+1}) - \nabla_{i_k} h_\beta(y^k) + \nabla_{i_k} h_\beta(y^k)\|^2 \\ &\le 2\|\nabla_{i_k} h_\beta(y^{k-\tau(\delta)+1}) - \nabla_{i_k} h_\beta(y^k)\|^2 + 2\|\nabla_{i_k} h_\beta(y^k)\|^2 \\ &\overset{(45)}{\le} \frac{2\beta^2\tau'}{n^2}\sum_{d=k-\tau(\delta)+1}^{k-1}\|y^{d+1} - y^d\|^2 + \frac{2\beta^2}{n^2}\|y^{k+1} - y^k\|^2 \\ &\le \max\left\{\frac{2\beta^2\tau'}{n^2}, \frac{2\beta^2}{n^2}\right\}\sum_{d=k-\tau(\delta)+1}^{k}\|y^{d+1} - y^d\|^2 \\ &\le \frac{2\beta^2\tau(\delta)}{n^2}\sum_{d=k-\tau(\delta)+1}^{k}\|y^{d+1} - y^d\|^2, \end{aligned} \tag{46}$$

where $\tau' = \tau(\delta) - 1$ and the last inequality holds because $\beta > \sigma^*_{\max}$. With the filtration $\mathcal{X}^k = \sigma\{y^0, \ldots, y^k, i_0, \ldots, i_{k-1}\}$,

$$\begin{aligned} &\mathbb{E}\left(\|\nabla_{i_k} h_\beta(y^{k-\tau(\delta)+1})\|^2 | \chi^{k-\tau(\delta)+1}\right) \\ &\quad= \mathbb{E}\left(\|\nabla_{i_k} h_\beta(y^{k-\tau(\delta)+1})\|^2 | y^{k-\tau(\delta)+1}, i_{k-\tau(\delta)}\right) \\ &\quad= \sum_{j=1}^N [\mathbf{P}^{\tau(\delta)}]_{i_{k-\tau(\delta)},j}\|\nabla_j h_\beta(y^{k-\tau(\delta)+1})\|^2 \\ &\quad\overset{(4)}{\ge} (1-\delta)\pi_*\|\nabla h_\beta(y^{k-\tau(\delta)+1})\|^2. \end{aligned} \tag{47}$$

Reverting the sides of (47) and taking expectation over $\mathcal{X}^{k-\tau(\delta)+1}$, followed by applying (46), we have for $k \ge \tau(\delta) - 1$

$$\begin{aligned} &\mathbb{E}\|\nabla h_\beta(y^{k-\tau(\delta)+1})\|^2 \\ &\le \frac{2\beta^2\tau(\delta)}{n^2(1-\delta)\pi_*}\sum_{d=k-\tau(\delta)+1}^{k}\mathbb{E}\left(\|y^{d+1} - y^d\|^2\right). \end{aligned} \tag{48}$$

Notice that

$$\begin{aligned} &h_\beta(y^k) - h_\beta(y^{k+1}) \\ &\overset{(25)}{=} L_\beta(x^{k+1}, y^k; z^k) - L_\beta(x^{k+2}, y^{k+1}; z^{k+1}) \\ &= L_\beta(x^{k+1}, y^k; z^k) - L_\beta^{k+1} + L_\beta^{k+1} - L_\beta(x^{k+2}, y^{k+1}; z^{k+1}) \\ &\ge \frac{1}{n}\|y^k - y^{k+1}\|^2, \end{aligned} \tag{49}$$

where the last line follows from parts 1 and 2 of Lemma 1. Combining (49) and (48), we get

$$\begin{aligned} &\mathbb{E}\|\nabla h_\beta(y^{k-\tau(\delta)+1})\|^2 \\ &\le \frac{2\beta^2\tau(\delta)}{n(1-\delta)\pi_*}\mathbb{E}\left(h_\beta(y^{k-\tau(\delta)+1}) - h_\beta(y^{k+1})\right). \end{aligned} \tag{50}$$

Now with $k = (t+1)\tau(\delta) - 1$, (50) reduces to

$$\begin{aligned} &\mathbb{E}\|\nabla h_\beta(y^{t\tau(\delta)})\|^2 \\ &\le \frac{2\beta^2\tau(\delta)}{n(1-\delta)\pi_*}\mathbb{E}\left(h_\beta(y^{t\tau(\delta)}) - h_\beta(y^{(t+1)\tau(\delta)})\right) \\ &\overset{(41)}{=} \frac{2\beta^2\tau(\delta)}{n(1-\delta)\pi_*}\left(F_t - F_{t+1}\right) \end{aligned} \tag{51}$$

By the convexity of $h_\beta(\cdot)$,

$$\mathbb{E}h_\beta(y^{t\tau(\delta)}) - h_\beta^\star \le \mathbb{E}\langle\nabla h_\beta(y^{t\tau(\delta)}), y^{t\tau(\delta)} - y^\star\rangle. \tag{52}$$

Since both sides of (52) are nonnegative, we square them and use the Cauchy-Schwarz inequality to get

$$F_t^2 \le \mathbb{E}\|\nabla h_\beta(y^{t\tau(\delta)})\|^2\cdot\mathbb{E}\|y^{t\tau(\delta)} - y^\star\|^2. \tag{53}$$

Substituting (50) into (53) completes the proof. ■

Now we are ready to establish the linear convergence rate of the sequence $(F_t)_{t\ge0}$.

Theorem 3: Under Assumption 1, for $\beta > 2\sigma^*_{\max} + 2$, we have linear convergence (with ν given in Lemma 4):

$$F_{t+1} \le \left(1 + \frac{n(1-\delta)\pi_*\nu}{4\beta^2\tau(\delta)}\right)^{-1}F_t, \quad \forall\, t \ge 0. \tag{54}$$

Proof: By the strong convexity of $h_\beta(\cdot)$ and $y^\star = \arg\min h_\beta(y)$, it holds for any $y \in \mathbb{R}^{np}$ that,

$$\frac{\nu}{2}\|y - y^\star\|^2 \le h_\beta(y) - h_\beta(y^\star). \tag{55}$$

Hence,

$$\mathbb{E}\|y^{t\tau(\delta)} - y^\star\|^2 \le \frac{2(\mathbb{E}h_\beta(y^{t\tau(\delta)}) - h_\beta^\star)}{\nu} = \frac{2F_t}{\nu}. \tag{56}$$

Substituting (56) into (42), we have

$$F_t^2 \le \frac{C}{\nu}\left(F_t - F_{t+1}\right)F_t, \quad \text{where} \quad C = \frac{4\beta^2\tau(\delta)}{n(1-\delta)\pi_*}. \tag{57}$$

By (49), the sequence $\{h_\beta(y^k)\}$ is non-increasing, implying $0 \le F_{t+1} \le F_t$. This together with (57) yields

$$F_t F_{t+1} \le \frac{C}{\nu}\left(F_t - F_{t+1}\right) F_t, \tag{58}$$

which is equivalent to (54). ■

Theorem 3 states that Walkman for decentralized least squares converges linearly by epoch (every $\tau(\delta)$ iterations).

V. Communication Analysis

This section derives and compares communication complexities with some state-of-the-art methods to solve problem (68) in solving two different types of problems: the decentralized least squares problem and the general nonconvex nonsmooth problem. In the following analysis, communication of p-dimensional variables between a pair of agents is taken as 1 unit of communication, while ν and L are taken as constants independent from network scale n.

A. Solving Least Squares Problem

First, we establish the communication complexity of Walkman. From (55) and (41), we have

$$\mathbb{E}\|y^{t\tau(\delta)} - y^\star\|^2 \le \frac{2}{\nu}F_t \overset{(54)}{\le} \left(\frac{2}{\nu}\right)\left(1 + \frac{n(1-\delta)\pi_*\nu}{4\beta^2\tau(\delta)}\right)^{-t} F_0. \tag{59}$$

To achieve mean-square deviation $G_t := \mathbb{E}\|y^{t\tau(\delta)} - y^\star\|^2 \le \epsilon$, it is enough to have

$$\left(\frac{2F_0}{\nu}\right)\left(1 + \frac{n(1-\delta)\pi_*\nu}{4\beta^2\tau(\delta)}\right)^{-t} \le \epsilon, \tag{60}$$

which is implied by

$$t = \ln\left(\frac{2F_0}{\nu\epsilon}\right) \Big/ \ln\left(1 + \frac{n(1-\delta)\pi_*\nu}{4\beta^2\tau(\delta)}\right). \tag{61}$$

Since β can be regarded as constants that are independent of network size n, and ν is $O(\frac{1}{n})$, we can write:

$$t \sim O\left(\ln\left(\frac{n}{\epsilon}\right) / \ln\left(1 + \frac{(1-\delta)\pi_*}{\tau(\delta)}\right)\right) \tag{62}$$

For each epoch t, there are $\tau(\delta)$ iterations, which use $O(\tau(\delta))$ communication. Hence, to guarantee $G_t \le \epsilon$, the total communication complexity is

$$O\left(\left(\ln\left(\frac{n}{\epsilon}\right) / \ln\left(1 + \frac{(1-\delta)\pi_*}{\tau(\delta)}\right)\right) \cdot \tau(\delta)\right) \tag{63}$$

Recall the definition of $\tau(\delta)$ in (5), by setting δ as $1/2$, the communication complexity is

$$O\left(\left(\underbrace{\ln\left(\frac{n}{\epsilon}\right) / \ln\left(1 + \frac{(1-\sigma(\mathbf{P}))\pi_*}{2\ln\frac{2}{\pi_*}}\right)}_{\text{epoch number}}\right) \cdot \underbrace{\frac{1}{1-\sigma(\mathbf{P})}\ln\frac{2}{\pi_*}}_{\text{comm. per epoch}}\right), \tag{64}$$

where we remember $\sigma(\mathbf{P}) := \sup_{f\in\mathbb{R}^n : f^\mathsf{T}\mathbf{1}=0} \|f^\mathsf{T}\mathbf{P}\|/\|f\|$.

For simplicity of expression and comparison, in the succeeding parts we assume that the Markov chain is reversible with $\mathbf{P} = \mathbf{P}^\mathsf{T}$ and it admits a uniform stationary distribution $\pi^T = \pi^T\mathbf{P}$ with:

$$\pi = [1/n, \ldots, 1/n]^\mathsf{T} \in \mathbb{R}^n,$$

which implies $\pi_* = \min_i \pi_i = 1/n$. With $\mathbf{P}$ being a symmetric real matrix, we also have $\sigma(\mathbf{P}) = \lambda_2(\mathbf{P}) = \max\{|\lambda_i(\mathbf{P})| : \lambda_i(\mathbf{P}) \ne 1\}$. We get the total communication complexity of Walkman as

$$O\Bigg(\underbrace{\left(\ln\left(\frac{n}{\epsilon}\right) \Big/ \ln\left(1 + \frac{1-\lambda_2(\mathbf{P})}{2n\ln(2n)}\right)\right)}_{\text{epoch numbers}} \underbrace{\left(\frac{\ln(n)}{1-\lambda_2(\mathbf{P})}\right)}_{\text{comm. per epoch}}\Bigg) \tag{65}$$

1) Communication Comparisons: For comparison, we list the communication complexities of some existing algorithms.

Firstly, we study the communication complexity of ESDACD [30], which is an accelerated generalization to the randomized gossip method originally designed to solve the average consensus problem. When applying ESDACD, the agents should be synchronized to keep track of which iteration the network is going through. And in each iteration, only one edge in the network is activated to communicate bi-directionally. Based on Theorem 1 and following the derivation of Section B2 of [30], with algorithmic parameter $\mu_{i,j} = 1$ and all the edges uniformly selected with probability $1/m$,[4] the communication complexity of ESDACD to achieve the deviation $G_t \le \epsilon$ is

$$O\left(\ln\left(\frac{1}{\epsilon}\right) \cdot \frac{m}{\sqrt{d_{\min}(1-\lambda_2(\mathbf{P}))}}\right), \tag{66}$$

where $d_{\min}$ denotes the smallest degree among $d_1, \ldots, d_n$.

As for RW-ADMM [36], in each iteration, it evokes the communications between the activated agent and all its neighbors, and thus consumes d_{ave} communications per iteration where $d_{\text{ave}} = \frac{1}{n}\sum_{i=1}^n d_i$ is the averaged degree in the network. This implies that RW-ADMM requires more communications than Walkman per iteration. To calculate the communication complexity of RW-ADMM, we consider a simple d-regular graph in which $m = nd/2$. We also assume the state transition matrix $\mathbf{P}$ is symmetric and doubly stochastic so that $\pi_{\max} = \pi_{\min} = 1/n$. In addition, we assume the current agent will activate one of its d neighbors with a uniform probability $p = 1/d$ and thus it holds that $p_{\max} = p_{\min} = 1/d$. Under these conditions, the communication complexity for RW-ADMM is verified as

$$O\left(\ln\left(\frac{1}{\epsilon}\right) \cdot \frac{md}{\sqrt{(1-\lambda_2(\mathbf{P}))}}\right). \tag{67}$$

Next, we consider gossip based methods. D-ADMM [11] has:

$$O\left(\left(\underbrace{\ln\left(\frac{1}{\epsilon}\right) \Big/ \ln\left(1 + \sqrt{1-\lambda_2(\mathbf{P})}\right)}_{\text{iteration numbers}}\right) \cdot \underbrace{m}_{\text{comm. per iter.}}\right) \tag{68}$$

[4]Nonuniform selection of edges is not practical in real applications. Since each agent should generate the randomly selected edge in each iteration with the same seed, nonuniform selection of edges implies each agent should cache the diverse sampling probabilities for each edge.

where m is the number of edges. The communication complexity of EXTRA [14] is

$$O\left(\left(\ln\left(\frac{1}{\epsilon}\right)\Big/\ln\left(2-\lambda_2(\mathbf{P})\right)\right)\cdot m\right) \tag{69}$$

As to exact diffusion [22], the communication complexity is

$$O\left(\left(\ln\left(\frac{1}{\epsilon}\right)\Big/\ln\left(1+\frac{1-\lambda_2(\mathbf{P})}{\lambda_2(\mathbf{P})+C}\right)\right)\cdot m\right) \tag{70}$$

where C only depends on the condition number of the objective function, independent of $\lambda_2(\mathbf{P})$ and n.

Considering the case $\epsilon \le 1/\mathrm{e}$, it holds $\ln(n/\epsilon) \le \ln n \cdot \ln(1/\epsilon)$. Since $\ln(1+x) \approx x$ for x close to 0, Walkman in (65) can be simplified to:

$$O\left(\ln\left(\frac{1}{\epsilon}\right)\cdot\frac{n\ln^3(n)}{(1-\lambda_2(\mathbf{P}))^2}\right). \tag{71}$$

We similarly simplify the communication complexities in (68), (69), and (70). They are listed in Table I in §I-A. Clearly, ESDACD has a better communication complexity than all the compared methods but may still be worse than Walkman.

Walkman is more communication efficient than ESDACD when

$$\frac{n\ln^3(n)}{(1-\lambda_2(\mathbf{P}))^2} \le \frac{m}{(d_{\min}(1-\lambda_2(\mathbf{P})))^{1/2}}. \tag{72}$$

With $d_{\min} \le m/n$, a sufficient condition for (72) is

$$\lambda_2(\mathbf{P}) \le 1-\frac{n^{1/3}[\ln(n)]^2}{m^{1/3}} \approx 1-\left(\frac{n}{m}\right)^{1/3}, \tag{73}$$

where the approximation holds for $\ln(n) \ll n$ and with $\ln(n)$ ignored. Condition (73) indicates the network has moderately good connectivity. When this holds, Walkman exhibits superior communication efficiency than every compared algorithm.

2) Communication for Different Graphs: Let us consider three classes of graphs for concrete communication complexities.

Example 1 (Complete graph) In a complete graph, every agent connects with all the other nodes. The number of edges $m = O(n^2)$ and $\lambda_2(\mathbf{P}) = 0$, $d_{\min} = n$. Consequently, the communication complexity of Walkman is $O(\ln(1/\epsilon)n\ln^3(n))$ while that of ESDACD is $O(\ln(1/\epsilon)n^{3/2})$, and those of the other algorithms are $O(\ln(1/\epsilon)n^2)$. Noticing $\ln^3(n) \ll n^{1/2}$, Walkman is more communication efficient.

Example 2 (Random graph) Consider the random graphs by Edgar Gilbert [46], $G(n,p)$, in which an n-node graph is generated with each edge populating independently with probability $p \in (0,1)$. Let $\mathbf{A} \in \mathbb{R}^{n\times n}$ denote the adjacency matrix of the generated graph, with $A_{i,j} = 1$ if nodes i and j are connected, and 0 otherwise. The (i,j)-th $(i \ne j)$ entry of the transition probability matrix $\mathbf{P}$ is $P_{i,j} = \frac{A_{i,j}}{d_{\max}}$, where $d_{\max} = \max_i \sum_{j=1}^n A_{i,j}$ is the maximal degree of all the n nodes. By union bound and Bernstein's inequality and Theorem 1 of [47], one has

$$1-\lambda_2(\mathbf{P}) = O(1). \tag{74}$$

With such setting, Walkman exhibits a communication complexity of roughly $O(\ln(1/\epsilon)n\ln^3 n)$ while that of ESDACD is $O(\ln(1/\epsilon)n^{3/2})$, and the other algorithms have $O(\ln(1/\epsilon)n^2)$. Hence, Walkman is more communication-efficient when n is sufficiently large.

Example 3 (Cycle graph) Consider a cycle, where each agent connects with its previous and next neighbors. One can verify that

$$1-\lambda_2(\mathbf{P}) = O\left(1-\cos(2\pi/n)\right) = O\left(1/n^2\right), \tag{75}$$

and $m = O(n)$. Hence, Walkman has a communication complexity of roughly $O(\ln(1/\epsilon)n^5\ln^3 n)$ while, in (68) and (66), D-ADMM and ESDACD have $O(n^2\ln(1/\epsilon))$, and in (70)–(70), EXTRA and exact diffusion have $O(n^3\ln(1/\epsilon))$, so Walkman is less communication-efficient.

B. Solving General Nonconvex Nonsmooth Problems

According to Theorem 2, we first derive the communication complexity of Walkman. To achieve the ergodic gradient deviation $E_t := \min_{k\le t}\mathbb{E}\|g^k\|^2 \le \epsilon$ for any $t > \tau(\delta)+2$, it is sufficient to have

$$\frac{\bar{C}}{t}\left(L_\beta^0 - \underline{f}\right) \le \epsilon. \tag{76}$$

Taking L_β^0 and $\underline{f}$ as constant independent from n and the network structure, one has

$$t \sim O\left(\frac{1}{\epsilon}\cdot\frac{\tau(\delta)^2+1}{(1-\delta)n\pi_*}\right) \tag{77}$$

Recall the definition of $\tau(\delta)$ in (5), by setting δ as $1/2$, the communication complexity is

$$O\left(\frac{1}{\epsilon}\cdot\frac{\ln^2\left(\frac{1}{\pi_*}\right)}{n\pi_*(1-\sigma(\mathbf{P}))^2}\right). \tag{78}$$

We consider a reversible Markov chain with $\mathbf{P}^\mathsf{T} = \mathbf{P}$ embedded on an undirected graph, and have the communication complexity of Walkman is

$$O\left(\frac{1}{\epsilon}\cdot\frac{\ln^2 n}{(1-\lambda_2(\mathbf{P}))^2}\right). \tag{79}$$

1) Communication Comparisons on Different Graphs: Next, we compare the communication complexity of Walkman with existing algorithms, D-GPDA [48] and xFILTER [48] on two specific types of graph structures. On a complete graph, the communication complexity of Walkman is $O(\frac{\ln^2 n}{\epsilon})$, whereas, according to [48], the better communication complexity between D-GPDA and xFILTER is $O(\frac{n^2}{\epsilon})$. Next, we consider the cycle graph, which is sparsely connected. Walkman consumes $O(n^4\frac{\ln^2 n}{\epsilon})$ amount of communication on it, whereas the better communication complexity between D-GPDA and xFILTER is $O(\frac{n^2}{\epsilon})$. Hence, we can draw a similar conclusion as in Section V-A, that is, Walkman is more communication efficient on a more densely connected graph.

VI. Numerical Experiments

In this section, we compare Walkman with existing state-of-the-art decentralized methods through numerical experiments. Consider a network of 50 nodes that are randomly placed in a 30×30 square. Any two nodes within a distance of 15 are

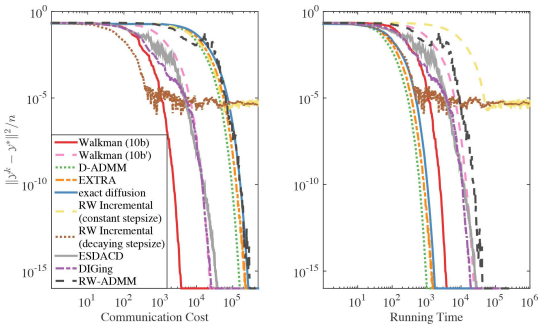


Fig. 5. Performance of decentralized algorithms on least squares.

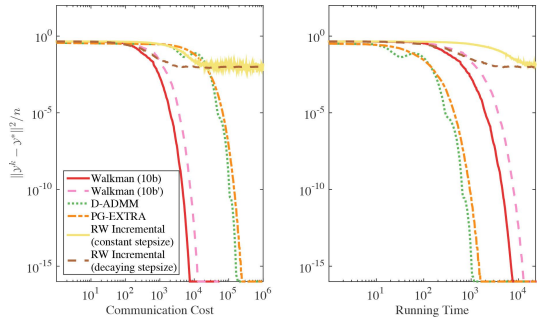


Fig. 6. Performance of decentralized algorithms on logistic regression.

connected; others are not. We set the probability transition matrix $\mathbf{P}$ as $[\mathbf{P}]_{ij} = 1/d_i$. Algorithmic parameters in the following experiments are set as follows. For the random-walk (RW) incremental algorithm, we have used both a fixed step-size of 0.001 and a sequence of decaying step-sizes $\min\{0.01, 5/k\}$. For other algorithms, we have hand-optimized their parameters by grid-search.

A. Decentralized Least Squares

The first experiment uses least squares in (21) with $\mathbf{A}_i \in \mathbb{R}^{5\times 10}$, $x \in \mathbb{R}^{10}$ and $b_i \in \mathbb{R}^5$. Each entry in $\mathbf{A}_i$ is generated from the standard Gaussian distribution, and $b_i := \mathbf{A}_i x_0 + v_i$, where $x_0 \sim \mathcal{N}(0, I_{10})$ and $v_i \sim \mathcal{N}(0, 0.1 \times I_5)$. Fig. 5 compares different algorithms. In this experiment, the comparison methods include the randomized-gossip type method (ESDACD), those with dense communications (D-ADMM, EXTRA, exact diffusion), DIGing over time-varying graph, RW Incremental method and RW-ADMM with mixed communication pattern. To be noted, we implement all these methods in the synchronous fashion, i.e., an iteration would not start before a priori iteration completes. As for a method over time-varying graph, DIGing is conducted with merely one edge uniformly randomly chosen in each time instance. For ESDACD, the activated edge in each iteration is also drawn independently from a uniform distribution.

In the left plot of Fig. 5, we count one communication for each transmission of a p-length vector ($p = 10$ is the dimension of x). It is observed that Walkman with (10b) is much more communication efficient than the other algorithms, while Walkman with (10b') is comparable to ESDACD and DIGing. In the right plot of Fig. 5, we illustrate the running times of these methods.

While a running time should in general include the times of computing, communication, and other overheads, we only include communication time and allows simultaneous communication over multiple edges for non-incremental algorithms. However, we assume each communication follows an i.i.d. exponential distribution with parameter 1. Each iteration of D-ADMM, EXTRA, and exact diffusion waits for the completion of the slowest communication (out of $2m$ communications), which determines the communication time of that iteration. In contrast, ESDACD, DIGing, random-walk incremental algorithms and Walkman only use one communication per iteration. The communication time per iteration of RW-ADMM is in between, as it waits for the slowest communication in the neighborhood to complete. Under our setting, Walkman takes longer to converge than D-ADMM, EXTRA, and exact diffusion. It is observed that Walkman with (10b) outperforms RW-ADMM and ESDACD in both communication cost and running time. In addition, D-ADMM is also observed more efficient than RW-ADMM, which is consistent with the communication complexity we derived in (67).

B. Decentralized Sparse Logistic Regression

The second experiment solves the logistic regression problem

$$\min_{x\in\mathbb{R}^p} \lambda\|x\|_1 + \frac{1}{nb}\sum_{i=1}^{n}\sum_{j=1}^{b}\log\left(1+\exp(-y_{ij}v_{ij}^{\mathsf{T}}x)\right), \qquad (80)$$

where $y_{ij} \in \{-1, 1\}$ denotes the label of the jth sample kept by the ith agent, and $v_{ij} \in \mathbb{R}^p$ represents its feature vector, and there are b samples kept by each agent. In this experiment, we set $b = 10, p = 5$. Each sample feature $v_{ij} \sim \mathcal{N}(0, 1)$. To generate y_{ij}, we first generate a random vector $x^0 \in \mathbb{R}^5 \sim \mathcal{N}(0, I)$. Then we generate a uniformly distributed variable $z_{ij} \sim \mathcal{U}(0, 1)$, and if $z_{ij} \le 1/[1 + \exp(-v_{ij}^{\mathsf{T}}x^0)]$, y_{ij} is taken as 1; otherwise y_{ij} is set as -1. We run the simulation over the same network as the above least-square problem. Due to the nonsmooth term in (80), EXTRA and exact diffusion is not applicable in this problem. Instead, we compare Walkman with PG-EXTRA [15], D-ADMM and random walk proximal gradient method, which conducts one-step proximal gradient operation when an agent receives the variable x.

The communication efficiency of Walkman is also observed in Fig. 6.

C. Decentralized Non-Negative Principal Component Analysis

To test the performance on solving nonconvex, nonsmooth problem, the third experiment solves the Non-Negative Principal Component Analysis (NN-PCA) problem

$$\min_{x\in\mathbb{R}^p} \frac{1}{n}\sum_{i=1}^{n} -x^{\mathsf{T}}\left(\frac{1}{b}\sum_{j=1}^{b} y_{ij}y_{ij}^{\mathsf{T}}\right)x + \mathbf{1}_C(x), \qquad (81)$$

where $\mathbf{1}_C$ denotes the indicator function of the feasible space $C := \{x \in \mathbb{R}^p : \|x\| \le 1, x_i \ge 0, \forall i \in \{1, \ldots, p\}\}$, $y_{ij} \in \mathbb{R}^p$ denotes the j-th sample kept by the i-th agent, and there are b samples kept by each agent. In this experiment, we utilize the training set of the MNIST [49] dataset to form the samples, and set $b = 1000$. Each agent only keeps samples with a same label. Noticing that the NN-PCA problem is nonconvex, we use

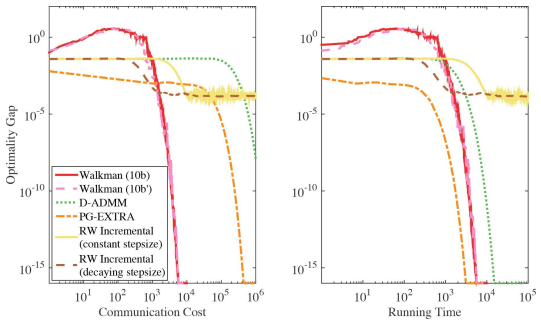


Fig. 7. Performance of decentralized algorithms on NN-PCA.

optimality gap to measure the distance between the algorithmic variables to problem's saddle points, which is defined as

$$\|\text{proj}_{\partial_r(x^k)}(-\nabla f(x^k)) + \nabla f(x^k)\|^2 + \|sy^k - \mathbb{1}\otimes x^k\|^2,$$

where the first term measures how close is $\partial_r(x^k) + \nabla f(x^k)$ to 0, and the second term measures the consensus violation of the copies kept by agents. For PG-EXTRA and D-ADMM, since there is only y^k, we take x^k as the mean of $\{y_1^k, \ldots, y_n^k\}$. For RW Incremental methods, since there is only x^k, the second term of optimality gap is naturally 0. We run the simulation over the same network as the above two problems. Under either optimality criterion, the communication efficiency of Walkman is also observed in Fig. 7.

VII. Conclusion

We have proposed a (random) walk algorithm, called Walkman, for decentralized consensus optimization. The (random) walk carries the current solution x and lets it updated by every visited agent. Any limit point of the sequence of x is almost surely a stationary point. Under convexity assumption, the sequence converges to the optimal solution with a fixed step-size, which makes Walkman more efficient than the existing random-walk algorithms. We have found Walkman uses less total communication than popular algorithms such as D-ADMM, EXTRA, exact diffusion, and PG-EXTRA though taking longer wall-clock time to converge. Random walks also add another layer of privacy protection.

Appendix A
Proof of Lemma 1

The proof of Lemma 1 takes a few steps, Lemmas 7–9.

Lemma 7 shows that the update on the dual variable can be bounded by that of the primal variable.

Lemma 7: Under Assumption 3, $(x^k, y^k, z^k)_{k>T}$, the sequence generated by Walkman iteration (10), satisfies

1) if Walkman uses (10b), it holds

$$\|z^{k+1} - z^k\| = \|z_{i_k}^{k+1} - z_{i_k}^k\| \le L\|y^{k+1} - y^k\|; \quad (82)$$

2) if Walkman uses (10b'), it hold

$$\|z^{k+1} - z^k\| = \|z_{i_k}^{k+1} - z_{i_k}^k\| \le L\|y^{\tau(k,i_k)+1} - y^{\tau(k,i_k)}\|. \quad (83)$$

Proof: Part 1) Remember agent i_k is activated at iteration k. The optimality condition of (10b) for $i = i_k$ implies

$$\nabla f_i(y_{i_k}^{k+1}) - \left(z_{i_k}^k + \beta(x^{k+1} - y_{i_k}^{k+1})\right) = 0. \quad (84)$$

Substituting the above into (10c) yields

$$\nabla f_i(y_i^{k+1}) = z_i^{k+1}, \quad \text{for } i = i_k. \quad (85)$$

Hence, for $i = i_k$, we have:

$$\begin{aligned}\|z_i^{k+1} - z_i^k\| &\overset{(a)}{=} \|z_i^{k+1} - z_i^{\tau(k,i)+1}\| \overset{(85)}{=} \|\nabla f_i(y_i^{k+1}) \\ &\quad - \nabla f_i(y_i^{\tau(k,i)+1})\| \\ &\overset{(12)}{\le} L\|y_i^{k+1} - y_i^{\tau(k,i)+1}\| \overset{(b)}{=} L\|y_i^{k+1} - y_i^k\|,\end{aligned} \quad (86)$$

where $\tau(k,i)$ is defined in (15). Equality (a) holds because $z_i^k = z_i^{\tau(k,i)+1}$ and (b) holds because $y_i^k = y_i^{\tau(k,i)+1}$. On the other hand, when $i \neq i_k$, agent i is not activated at k, so $\|z_i^{k+1} - z_i^k\| = L\|y_i^{k+1} - y_i^k\| = 0$, and we have (82).

Par 2) Substituting (10b') into (10c) yields

$$\nabla f_i(y_i^k) = z_i^{k+1}, \quad \text{for } i = i_k. \quad (87)$$

Comparing (85) and (87) and using $z_i^k = z_i^{\tau(k,i)+1}$ and $y_i^k = y_i^{\tau(k,i)+1}$, we get (83) using a similar derivation for (86). ■

Lemma 8 shows that the x-update in Walkman, i.e., (10a), provides sufficient descent of the augmented Lagrangian.

Lemma 8: Recall L_β^k defined in (16). Under Assumption 4, for $\beta > \gamma, k \ge 0$, the Walkman iterates satisfy

$$L_\beta^k - L_\beta(x^{k+1}, y^k; z^k) \ge \frac{\beta - \gamma}{2}\|x^k - x^{k+1}\|^2. \quad (88)$$

Proof: We rewrite the augmented Lagrangian in (8) as

$$L_\beta(x, y; z) = r(x) + \frac{1}{n}\left(F(y) + \frac{\beta}{2}\left\|\mathbb{1}\otimes x - y + \frac{z}{\beta}\right\|^2 - \frac{\|z\|^2}{2\beta}\right). \quad (89)$$

Applying the cosine identity $\|b+c\|^2 - \|a+c\|^2 = \|b-a\|^2 + 2\langle a+c, b-a\rangle$, we have

$$\begin{aligned}&L_\beta^k - L_\beta(x^{k+1}, y^k; z^k) - r(x^k) + r(x^{k+1}) \\ &= \frac{\beta}{2n}\left\|\mathbb{1}\otimes x^k - y^k + \frac{z^k}{\beta}\right\|^2 - \frac{\beta}{2n}\left\|\mathbb{1}\otimes x^{k+1} - y^k + \frac{z^k}{\beta}\right\|^2 \\ &= \frac{\beta}{2n}\sum_{i=1}^n\left(\|x^k - x^{k+1}\|^2 + 2\left\langle x^{k+1} - y_i^k + \frac{z_i^k}{\beta}, x^k - x^{k+1}\right\rangle\right) \\ &\ge \frac{\beta}{2}\|x^k - x^{k+1}\|^2 - \langle d^k, x^k - x^{k+1}\rangle,\end{aligned} \quad (90)$$

where d^k is defined as

$$d^k := -\frac{\beta}{n}\sum_{i=1}^n\left(x^{k+1} - y_i^k + \frac{z_i^k}{\beta}\right) \overset{(a)}{\in} \partial r(x^{k+1}), \quad (91)$$

where (a) comes from the optimality condition of (10a). Assumption 4 states $r(x^k) + \frac{\gamma}{2}\|x^k - x^{k+1}\|^2 \ge r(x^{k+1}) + \langle d^k, x^k - x^{k+1}\rangle$, substituting which into (90) gives us (88). ■

In Lemma 9, we derive the lower bound of descent in the augmented Lagrangian over the updates of y and z.

Lemma 9: Recall L_β^k defined in (16). Under Assumption 3, for any $k > T$,

1) if $\beta \geq 2L+2$, Walkman using (10b) satisfies

$$L_\beta(x^{k+1}, y^k; z^k) - L_\beta^{k+1} \geq \frac{1}{n}\|y^k - y^{k+1}\|^2. \quad (92)$$

2) if $\beta > L$, Walkman using (10b') satisfies

$$L_\beta(x^{k+1}, y^k; z^k) - L_\beta^{k+1} \geq \frac{\beta - L}{2n}\|y^k - y^{k+1}\|^2 - \frac{L^2}{n\beta}\|y^{\tau(k,i_k)+1} - y^{\tau(k,i_k)}\|^2. \quad (93)$$

Proof: From the Lagrangian (8), we derive

$$\begin{aligned}
&L_\beta(x^{k+1}, y^k; z^k) - L_\beta^{k+1} \\
&= \frac{1}{n}\Bigg(f_{i_k}(y_{i_k}^k) + \langle z_{i_k}^k, x^{k+1} - y_{i_k}^k\rangle + \frac{\beta}{2}\|x^{k+1} - y_{i_k}^k\|^2 \\
&\quad - f_{i_k}(y_{i_k}^{k+1}) - \langle z_{i_k}^{k+1}, x^{k+1} - y_{i_k}^{k+1}\rangle - \frac{\beta}{2}\|x^{k+1} - y_{i_k}^{k+1}\|^2\Bigg) \\
&\overset{(a)}{=} \frac{1}{n}\Bigg(f_{i_k}(y_{i_k}^k) - f_{i_k}(y_{i_k}^{k+1}) + \frac{\beta}{2}\|y_{i_k}^k - y_{i_k}^{k+1}\|^2 \\
&\quad - \langle y_{i_k}^k - y_{i_k}^{k+1}, z_{i_k}^{k+1}\rangle - \frac{1}{\beta}\|z_{i_k}^{k+1} - z_{i_k}^k\|^2\Bigg) \quad (94) \\
&\overset{(b)}{=} \frac{1}{n}\Bigg(f_{i_k}(y_{i_k}^k) - f_{i_k}(y_{i_k}^{k+1}) + \frac{\beta}{2}\|y_{i_k}^k - y_{i_k}^{k+1}\|^2 \\
&\quad - \langle y_{i_k}^k - y_{i_k}^{k+1}, \nabla f_{i_k}(y_{i_k}^{k+1})\rangle - \frac{1}{\beta}\|z_{i_k}^{k+1} - z_{i_k}^k\|^2\Bigg) \quad (95) \\
&\overset{(c)}{\geq} \frac{1}{n}\left(-\frac{L}{2}\|y_{i_k}^k - y_{i_k}^{k+1}\|^2 + \frac{\beta}{2}\|y_{i_k}^k - y_{i_k}^{k+1}\|^2 - \frac{1}{\beta}\|z_{i_k}^{k+1} - z_{i_k}^k\|^2\right) \quad (96) \\
&\overset{(d)}{\geq} \frac{1}{n}\Bigg(-\frac{L}{2}\|y_{i_k}^k - y_{i_k}^{k+1}\|^2 + \frac{\beta}{2}\|y_{i_k}^k - y_{i_k}^{k+1}\|^2 \\
&\quad - \frac{L^2}{\beta}\|y_{i_k}^k - y_{i_k}^{k+1}\|^2\Bigg) \\
&\overset{(e)}{\geq} \frac{1}{n}\|y_{i_k}^k - y_{i_k}^{k+1}\|^2 = \frac{1}{n}\|y^k - y^{k+1}\|^2. \quad (97)
\end{aligned}$$

where equality (a) holds due to $\|b+c\|^2 - \|a+c\|^2 = \|b-a\|^2 + 2\langle a+c, b-a\rangle$ and recursion (10c), equality (b) holds because of (85), inequality (c) holds because $f_i(\cdot)$ is L-Lipschitz differentiable, inequality (d) holds because of (86), and inequality (e) follows from the assumption $\beta \geq 2L+2$.

Next, we study Walkman using (10b'). The above equation array holds to (94). By substituting (87) into (94), we get

$$\begin{aligned}
&L_\beta(x^{k+1}, y^k; z^k) - L_\beta^{k+1} \\
&= \frac{1}{n}\Bigg(f_{i_k}(y_{i_k}^k) - f_{i_k}(y_{i_k}^{k+1}) + \frac{\beta}{2}\|y_{i_k}^k - y_{i_k}^{k+1}\|^2 \\
&\quad - \langle y_{i_k}^k - y_{i_k}^{k+1}, \nabla f_{i_k}(y_{i_k}^k)\rangle - \frac{1}{\beta}\|z_{i_k}^{k+1} - z_{i_k}^k\|^2\Bigg). \quad (98)
\end{aligned}$$

While (95) has $\nabla f_{i_k}(y_{i_k}^{k+1})$, (98) involves $\nabla f_{i_k}(y_{i_k}^k)$. However, from (98), using $\nabla f_{i_k}(\cdot)$ being L-Lipschitz, we still get (96), to which we can apply Lemma 7 2) to get (93). ■

In Lemma 10, we establish the sufficient descent in Lyapunov functions of Walkman.

Lemma 10: Recall L_β^k and M_β^k defined in (16) and (17). Under Assumptions 3 and 4, for any $k > T$,

1) if $\beta > \max\{\gamma, 2L+2\}$, Walkman using (10b) satisfies

$$L_\beta^k - L_\beta^{k+1} \geq \frac{\beta - \gamma}{2}\|x^k - x^{k+1}\|^2 + \frac{1}{n}\|y^k - y^{k+1}\|^2; \quad (99)$$

2) if $\beta > \max\{\gamma, 2L^2 + L + 2\}$ the Walkman using (10b') satisfies

$$M_\beta^k - M_\beta^{k+1} \geq \frac{\beta - \gamma}{2}\|x^k - x^{k+1}\|^2 + \frac{1}{n}\|y^k - y^{k+1}\|^2 + \frac{L^2}{2n}\|y^{\tau(k,i_k)+1} - y^{\tau(k,i_k)}\|^2. \quad (100)$$

Proof: Statement 1) is a direct result of adding (88) and (92). To prove statement 2), noticing

$$y_i^{\tau(k+1,i)+1} - y_i^{\tau(k+1,i)} = \begin{cases} y_i^{k+1} - y_i^k, & i = i_k \\ y_i^{\tau(k,i)+1} - y_i^{\tau(k,i)}, & \text{otherwise,} \end{cases} \quad (101)$$

we derive

$$\begin{aligned}
&M_\beta^k - M_\beta^{k+1} \\
&= L_\beta^k - L_\beta^{k+1} + \frac{L^2}{n}\left(\|y_{i_k}^{\tau(k,i_k)+1} - y_{i_k}^{\tau(k,i_k)}\|^2 - \|y_{i_k}^{k+1} - y_{i_k}^k\|^2\right) \\
&= L_\beta^k - L_\beta^{k+1} + \frac{L^2}{n}\left(\|y^{\tau(k,i_k)+1} - y^{\tau(k,i_k)}\|^2 - \|y^{k+1} - y^k\|^2\right). \quad (102)
\end{aligned}$$

Substituting (88) and (93) into (102) and using $\frac{\beta}{2} - \frac{L}{2} - L^2 \geq 1$ and $1 - \frac{1}{\beta} > \frac{1}{2}$, we complete the proof of statement 2). ■

Lemma 11 states that both Lyapunov functions are lower bounded.

Lemma 11: For $\beta > \max\{\gamma, 2L+2\}$ (resp. $\beta > \max\{\gamma, 2L^2 + L + 2\}$), Walkman using (10b) (resp. (10b')) ensures a lower bounded sequence $(L_\beta^k)_{k\geq 0}$ (resp. $(M_\beta^k)_{k\geq 0}$).

Proof: For Walkman using (10b) and $k > T$, we have

$$\begin{aligned}
L_\beta^k &= r(x^k) + \frac{1}{n}\sum_{j=1}^n \left(f_j(y_j^k) + \langle z_j^k, x^k - y_j^k\rangle\right) \\
&\quad + \frac{\beta}{2n}\|\mathbb{1}\otimes x^k - y^k\|^2 \\
&\overset{(85)}{=} r(x^k) + \frac{1}{n}\sum_{j=1}^n \left(f_j(y_j^k) + \langle \nabla f_j(y_j^k), x^k - y_j^k\rangle\right) \\
&\quad + \frac{\beta}{2n}\|\mathbf{1}\otimes x^k - y^k\|^2 \\
&\overset{(a)}{\geq} r(x^k) + \frac{1}{n}\sum_{j=1}^n f_j(x^k) + \frac{\beta - L}{2n}\|\mathbf{1}\otimes x^k - y^k\|^2
\end{aligned}$$

$$\geq \min_x \left\{ r(x) + \frac{1}{n}\sum_{j=1}^n f_j(x) \right\} + \frac{\beta - L}{2n}\|\mathbb{1}\otimes x^k - y^k\|^2$$
$$\overset{(b)}{>} -\infty, \tag{103}$$

where (a) holds as each f_j is Lipschitz differentiable and (b) from Assumption 2 and $\beta > L$. So, L_β^k is lower bounded.

Next, for Walkman using (10b') and $k > T$, we derive

$$\begin{aligned}
M_\beta^k &\overset{(87)}{=} r(x^k) + \frac{1}{n}\sum_{j=1}^n \left(f_j(y_j^k) + \langle \nabla f_j(y_j^{\tau(k,j)}), x^k - y_j^k\rangle \right) \\
&\quad + \frac{\beta}{2n}\|\mathbb{1}\otimes x^k - y^k\|^2 + \frac{L^2}{n}\sum_{i=1}^n \|y_i^{\tau(k,i)+1} - y_i^{\tau(k,i)}\|^2 \\
&\overset{(a)}{\geq} r(x^k) + \frac{1}{n}\sum_{j=1}^n \left(f_j(x^k) + \langle \nabla f_j(y_j^{\tau(k,j)}) - \nabla f_j(y_j^k), x^k - y_j^k\rangle \right) \\
&\quad + \frac{\beta - L}{2n}\|\mathbb{1}\otimes x^k - y^k\|^2 + \frac{L^2}{n}\sum_{i=1}^n \|y_i^k - y_i^{\tau(k,i)}\|^2 \\
&\overset{(b)}{\geq} r(x^k) + \frac{1}{n}\sum_{j=1}^n \left(f_j(x^k) - \|\nabla f_j(y_j^{\tau(k,j)}) - \nabla f_j(y_j^k)\|^2 \right) \\
&\quad + \frac{\beta - L - 2}{2n}\|\mathbb{1}\otimes x^k - y^k\|^2 + \frac{L^2}{n}\sum_{i=1}^n \|y_i^k - y_i^{\tau(k,i)}\|^2 \\
&\overset{(c)}{\geq} \min_x \left\{ r(x^k) + \frac{1}{n}\sum_{j=1}^n f_j(x^k) \right\} + \frac{2L^2}{2n}\|\mathbb{1}\otimes x^k - y^k\|^2 \\
&\overset{(d)}{>} -\infty,
\end{aligned} \tag{104}$$

where (a) holds because each f_j is Lipschitz differentiable, (b) holds due Young's inequality, (c) follows from the assumption $\beta > 2L^2 + L + 2$ and the Lipschitz smoothness of each f_j, and (d) holds due to Assumption 2. Therefore, M_β^k is bounded from below. ■

With above lemmas, we are ready to prove Lemma 1.

Proof of Lemma 1: Recall that the maximal hitting time T is almost surely finite. The monotonicity of $(L_\beta^k)_{k>T}$ (resp. $(M_\beta^k)_{k>T}$) in Lemma 10 and their lower boundedness in Lemma 11 ensure convergence of $(L_\beta^k)_{k\geq 0}$ (resp. $(M_\beta^k)_{k\geq 0}$).

For statement 2), We first consider Walkman with (10b). By statement 1) and (103), $r(x^k) + \frac{1}{n}F(\mathbf{x}^k)$ is upper bounded by $\max\{\max_{t\in\{0,\ldots,T\}}\{r(x^t) + \frac{1}{n}F(\mathbf{x}^t)\}, L_\beta^{T+1}\}$, and $\|\mathbb{1}\otimes x^k - y^k\|^2$ is upper bounded by $\max\{\max_{t\in\{0,\ldots,T\}}\{\|\mathbb{1}\otimes x^t - y^t\|^2\}, L_\beta^{T+1}\}$. By Assumption 2, the sequence (x^k) is bounded. The boundedness of $\|\mathbb{1}\otimes x^k - y^k\|^2$ further leads to that of (y^k). Finally, (85) and Assumption 3 ensure (z^k) is bounded, too. Altogether, (x^k, y^k, z^k) is bounded. Starting from statement 1)' and (104), a similar argument leads to boundedness of (x^k, y^k, z^k) for Walkman using (10b'). ■

Appendix B
Proof of Lemma 2

Following the aforementioned proof idea, we provide the detailed proof of Lemma 2 in this Section.

Proof of Lemma 2: First, recall Lemma 10 and $T < \infty$, for Walkman using (10b), we have

$$\sum_{k=0}^\infty \left(\mathbb{E}\|x^k - x^{k+1}\|^2 + \mathbb{E}\|y^k - y^{k+1}\|^2 \right) < +\infty; \tag{105}$$

and for Walkman using (10b'),

$$\begin{aligned}
\sum_{k=0}^\infty (&\mathbb{E}\|x^k - x^{k+1}\|^2 + \mathbb{E}\|y^k - y^{k+1}\|^2 \\
&+ \mathbb{E}\|y^{\tau(k,i_k)} - y^{\tau(k,i_k)+1}\|_2^2) < +\infty,
\end{aligned} \tag{106}$$

Hence, by Lemma 10,

$$\begin{aligned}
\sum_{k=0}^\infty (&\mathbb{E}\|x^k - x^{k+1}\|^2 + \mathbb{E}\|y^k - y^{k+1}\|^2 \\
&+ \mathbb{E}\|z^k - z^{k+1}\|_2^2) < +\infty,
\end{aligned} \tag{107}$$

holds for Walkman using either (10b) or (10b').

The proof starts with computing the subdifferentials of the augmented Lagrangian (89) with the updates in (10):

$$\begin{aligned}
\partial_x L_\beta^{k+1} &\ni d^k - \frac{\beta}{n}(y_{i_k}^{k+1} - y_{i_k}^k) + \frac{1}{n}(z_{i_k}^{k+1} - z_{i_k}^k), \\
&\overset{(91)}{=} -\frac{\beta}{n}(y_{i_k}^{k+1} - y_{i_k}^k) + \frac{1}{n}(z_{i_k}^{k+1} - z_{i_k}^k) =: w^k,
\end{aligned} \tag{108}$$

$$\nabla_{y_j} L_\beta^{k+1} = \frac{1}{n}\left(\nabla f_j(y_j^{k+1}) - z_j^{k+1} + \beta(y_j^{k+1} - x^{k+1}) \right), \tag{109}$$

$$\nabla_{z_j} L_\beta^{k+1} = \frac{1}{n}\left(x^{k+1} - y_j^{k+1} \right). \tag{110}$$

For notational brevity, we define g^k and q_i^k as

$$g^k := \begin{bmatrix} w^k \\ \nabla_y L_\beta^{k+1} \\ \nabla_z L_\beta^{k+1} \end{bmatrix}, \; q_i^k := \begin{bmatrix} w^k \\ \nabla_{y_i} L_\beta^{k+1} \\ \nabla_{z_i} L_\beta^{k+1} \end{bmatrix}, \tag{111}$$

where $i \in V$ is an agent index, and g^k is the gradient of L_β^{k+1}. For $\delta \in (0,1)$ and $k \geq \tau(\delta) + 1$, by the triangle inequality:

$$\begin{aligned}
\|q_{i_k}^{k-\tau(\delta)-1}\|^2 &= \|q_{i_k}^{k-\tau(\delta)-1} - q_{i_k}^k + q_{i_k}^k\|^2 \\
&\leq 2\underbrace{\|q_{i_k}^{k-\tau(\delta)-1} - q_{i_k}^k\|^2}_{A} + 2\underbrace{\|q_{i_k}^k\|^2}_{B}
\end{aligned} \tag{112}$$

Below, we upper bound A and B separately. A has three parts corresponding to the three components of g. Its first part is

$$\begin{aligned}
\|w^{k-\tau(\delta)-1} - w^k\|^2 &\leq 2\|w^{k-\tau(\delta)-1}\|^2 + 2\|w^k\|^2 \\
&\leq \frac{4}{n^2}\big(\beta^2\|y^{k+1} - y^k\|^2 + \beta^2\|y^{k-\tau(\delta)} - y^{k-\tau(\delta)-1}\|^2, \\
&\quad + \|z^{k+1} - z^k\|^2 + \|z^{k-\tau(\delta)} - z^{k-\tau(\delta)-1}\|^2\big)
\end{aligned} \tag{113}$$

where the 2nd inequality follows from (108). Then by (109), we bound the 2nd part of A

$$\begin{aligned}
&\|\nabla_{y_{i_k}} L_\beta^{k-\tau(\delta)-1} - \nabla_{y_{i_k}} L_\beta^{k+1}\|^2 \\
&\overset{(a)}{\leq} \frac{4L^2 + 4\beta^2}{n^2}\|y_{i_k}^{k-\tau(\delta)-1} - y_{i_k}^{k+1}\|^2 + \frac{4}{n^2}\|z_{i_k}^{k-\tau(\delta)-1} - z_{i_k}^{k+1}\|^2 \\
&\quad + \frac{4\beta^2}{n^2}\|x^{k-\tau(\delta)-1} - x^{k+1}\|^2
\end{aligned}$$

$$\leq \frac{(\tau(\delta)+2)(4+4\beta^2+4L^2)}{n^2}\sum_{t=k-\tau(\delta)-1}^{k}\big(\|x^t-x^{t+1}\|^2 + \|y^t-y^{t+1}\|^2+\|z^t-z^{t+1}\|^2\big), \quad (114)$$

where (a) uses the inequality of arithmetic and geometric means and the Lipschitz differentiability of f_j in Assumption 3. From (110), the 3rd part of A can be bounded as

$$\begin{aligned}&\|\nabla_{z_{i_k}}L_\beta^{k-\tau(\delta)-1}-\nabla_{z_{i_k}}L_\beta^{k+1}\|^2\\ &\leq \frac{2}{n^2}(\|x^{k-\tau(\delta)-1}-x^{k+1}\|^2+\|y_{i_k}^{k-\tau(\delta)-1}-y_{i_k}^{k+1}\|^2)\\ &\leq \frac{2(\tau(\delta)+2)}{n^2}\sum_{t=k-\tau(\delta)-1}^{k}\big(\|x^t-x^{t+1}\|^2+\|y^t-y^{t+1}\|^2\big). \quad (115)\end{aligned}$$

Substituting (113), (114) and (115) into term A, we get a constant $C_1\sim O(\frac{\tau(\delta)+1}{n^2})$, depending on $\tau(\delta),\beta,L$ and n, such that

$$A\leq C_1\sum_{t=k-\tau(\delta)-1}^{k}\big(\|x^t-x^{t+1}\|^2+\|y^t-y^{t+1}\|^2+\|z^t-z^{t+1}\|^2\big). \quad (116)$$

To bound the term B, using (110) and (10c), we have

$$\nabla_{z_{i_k}}L_\beta^{k+1}=\frac{1}{n\beta}(z_{i_k}^{k+1}-z_{i_k}^k), \quad (117)$$

Applying (109), (85) and (87), we derive $\nabla_{y_{i_k}}L_\beta^{k+1}$ for Walkman using (10b) or (10b'):

$$\text{(10b): } \nabla_{y_{i_k}}L_\beta^{k+1}=\frac{1}{n}\left(z_{i_k}^k-z_{i_k}^{k+1}\right), \quad (118)$$

$$\text{(10b'): } \nabla_{y_{i_k}}L_\beta^{k+1}=\frac{1}{n}\left(\nabla f_{i_k}(y_{i_k}^{k+1})-\nabla f_{i_k}(y_{i_k}^k)+z_{i_k}^k-z_{i_k}^{k+1}\right). \quad (119)$$

For both, we have

$$B\leq C_2\left(\|y^{k+1}-y^k\|^2+\|z^{k+1}-z^k\|^2\right), \quad (120)$$

for a constant C_2 depending on L,β and n, in the order of $\sim O(\frac{1}{n^2})$. Then substituting (116) and (120) into (112) and taking expectations yield

$$\mathbb{E}\|q_{i_k}^{k-\tau(\delta)-1}\|^2\leq C\sum_{t=k-\tau(\delta)-1}^{k}\big(\mathbb{E}\|x^t-x^{t+1}\|^2+\mathbb{E}\|y^t-y^{t+1}\|^2+\mathbb{E}\|z^t-z^{t+1}\|^2\big), \quad (121)$$

where $C=C_1+C_2$, and one has $C\sim O(\frac{\tau(\delta)+1}{n^2})$. Recalling (107), we get the convergence

$$\lim_{k\to\infty}\mathbb{E}\|q_{i_k}^{k-\tau(\delta)-1}\|^2=0, \quad (122)$$

which completes the proof of step 1).

In step 2), we compute the conditional expectation:

$$\begin{aligned}&\mathbb{E}\left(\|q_{i_k}^{k-\tau(\delta)-1}\|^2\mid\chi^{k-\tau(\delta)}\right)\\ &=\sum_{j=1}^{n}[\mathbf{P}^{\tau(\delta)}]_{i_{k-\tau(\delta)},j}\big(\|\nabla_x L_\beta^{k-\tau(\delta)}\|^2+\|\nabla_{y_j}L_\beta^{k-\tau(\delta)}\|^2\\ &\quad+\|\nabla_{z_j}L_\beta^{k-\tau(\delta)}\|^2\big)\\ &\overset{(a)}{\geq}(1-\delta)\pi_*\|g^{k-\tau(\delta)-1}\|^2, \quad (123)\end{aligned}$$

where (a) follows from (4) and the definition of g^k in (111). Then, with (122), it holds

$$\lim_{k\to\infty}\mathbb{E}\|g^k\|^2=\lim_{k\to\infty}\mathbb{E}\|g^{k-\tau(\delta)-1}\|^2=0. \quad (124)$$

By the Schwarz inequality $(\mathbb{E}\|g^k\|)^2\leq\mathbb{E}\|g^k\|^2$, we have

$$\lim_{k\to\infty}\mathbb{E}\|g^k\|=0. \quad (125)$$

Next, we prove step 3). By Markov's inequality, for each $\epsilon>0$, it holds that

$$\Pr(\|g^k\|\geq\epsilon)\leq\frac{\mathbb{E}\|g^k\|}{\epsilon}\overset{(125)}{\Rightarrow}\lim_{k\to\infty}\Pr(\|g^k\|\geq\epsilon)=0. \quad (126)$$

When a subsequence $(k_s)_{s\geq0}$ is provided, (126) implies.

$$\lim_{s\to\infty}\Pr(\|g^{k_s}\|\geq\epsilon)=0 \quad (127)$$

Then, for $j\in\mathbb{N}$, select $\epsilon=2^{-j}$ and we can find a nondecreasing subsubsequence (k_{s_j}), such that

$$\Pr(\|g^{k_s}\|\geq2^{-j})\leq2^{-j},\quad\forall k_s\geq k_{s_j}. \quad (128)$$

Since,

$$\sum_{j=1}^{\infty}\Pr(\|g^{k_{s_j}}\|\geq2^{-j})\leq\sum_{j=1}^{\infty}2^{-j}=1, \quad (129)$$

the Borel-Cantelli lemma yields

$$\Pr\left(\limsup_j\{\|g^{k_{s_j}}\|\geq2^{-j}\}\right)=0, \quad (130)$$

and thus

$$\Pr\left(\lim_j\|g^{k_{s_j}}\|=0\right)=1. \quad (131)$$

This completes step 3) and thus the entire Lemma 2. ■

APPENDIX C
PROOF OF THEOREM 2

We provide the detailed proof of Theorem 2 in this section.

Proof of Theorem 2: It can be simply verified that under the specific initialization, (85) and (87) hold for all $k\geq0$, and consequently ensure Lemmas 7–10 hold for all $k\geq0$. For g^k defined in (111), (121) and (123) hold. Jointly applying (121) and (123), for any $k\geq\tau(\delta)+1$, one has

$$\mathbb{E}\|g^{k-\tau(\delta)-1}\|^2\leq\frac{C}{(1-\delta)\pi_*}\sum_{t=k-\tau(\delta)-1}^{k}\big(\mathbb{E}\|x^t-x^{t+1}\|^2+\mathbb{E}\|y^t-y^{t+1}\|^2+\mathbb{E}\|z^t-z^{t+1}\|^2\big) \quad (132)$$

According to Lemmas 7 and 10, for Walkman using (10b) and $k \geq \tau(\delta)+1$, it holds,

$$\sum_{t=k-\tau(\delta)-1}^{k} \left(\mathbb{E}\|x^t - x^{t+1}\|^2 + \mathbb{E}\|y^t - y^{t+1}\|^2 + \mathbb{E}\|z^t - z^{t+1}\|^2\right) \leq \max\left\{\frac{2}{\beta-\gamma}, (1+L^2)n\right\}\left(\mathbb{E}L_\beta^{k-\tau(\delta)-1} - \mathbb{E}L_\beta^{k+1}\right) \tag{133}$$

It implies that for any $k \geq 0$, it holds

$$\mathbb{E}\|g^k\|^2 \leq C'\left(\mathbb{E}L_\beta^k - \mathbb{E}L_\beta^{k+\tau(\delta)+2}\right), \tag{134}$$

where $C' := \max\{\frac{2}{\beta-\gamma}, (1+L^2)n\}\frac{C}{(1-\delta)\pi_*}$. It can be simply verified that $C' = O(\frac{\tau(\delta)+1}{(1-\delta)n\pi_*})$ Let $\tau' := \tau(\delta)+2$. Then for any $K > \tau'$, summing (134) over $k \in \{K-\tau', \ldots, \mathrm{mod}_{\tau'}K\}$ gives

$$\sum_{l=1}^{\lfloor \frac{K}{\tau'} \rfloor} \mathbb{E}\|g^{K-l\tau'}\|^2 \leq C'\left(\mathbb{E}L_\beta^{\mathrm{mod}_{\tau'}K} - \mathbb{E}L_\beta^K\right) \leq C'(L_\beta^0 - \underline{f}), \tag{135}$$

where the last inequality follows from the nondecreasing property of the sequence $(L_\beta^k)_{k\geq 0}$ and the fact that $(L_\beta^k)_{k\geq 0}$ is lower bounded by $\underline{f}$.

According to (135), one has

$$\begin{aligned}\min_{k\leq K}\mathbb{E}\|g^k\|^2 &\leq \min_{1\leq l\leq \lfloor \frac{K}{\tau'}\rfloor}\mathbb{E}\|g^{K-l\tau'}\|^2 \\ &\leq \frac{1}{\lfloor \frac{K}{\tau'}\rfloor}\sum_{l=1}^{\lfloor \frac{K}{\tau'}\rfloor}\mathbb{E}\|g^{K-l\tau'}\|^2 \\ &\leq \frac{\tau' C'}{K-\tau'}(L_\beta^0 - \underline{f}) \\ &\leq \frac{C'(\tau'+1)}{K}(L_\beta^0 - \underline{f}),\end{aligned} \tag{136}$$

where the constant $C'(\tau'+1) = O(\frac{\tau(\delta)^2+1}{(1-\delta)n\pi_*})$. ■

We consider a reversible Markov chain on an undirected graph. Recalling the definition of $\tau(\delta)$ in (5) and taking δ as $1/2$, one has $\tau(\delta) \sim \frac{\ln n}{1-\lambda_2(\mathbf{P})}$. That is, to guarantee that $\min_{k\leq K}\mathbb{E}\|g^k\|^2 \leq \epsilon$,

$$O\left(\frac{1}{\epsilon}\cdot\left(\frac{\ln^2 n}{(1-\lambda_2(\mathbf{P}))^2}+1\right)\right) \tag{137}$$

iterations would be sufficient for Walkman using either (10b) or (10b').

REFERENCES

[1] A. H. Sayed, "Adaptive networks," *Proc. IEEE*, vol. 102, no. 4, pp. 460–497, Apr. 2014.

[2] A. H. Sayed, "Adaptation, learning, and optimization over networks," *Found. Trends Mach. Learn.*, vol. 7, no. 4–5, pp. 311–801, 2014.

[3] J. C. Duchi, A. Agarwal, and M. J. Wainwright, "Dual averaging for distributed optimization: Convergence analysis and network scaling," *IEEE Trans. Autom. Control*, vol. 57, no. 3, pp. 592–606, Mar. 2012.

[4] J. Chen, Z. J. Towfic, and A. H. Sayed, "Dictionary learning over distributed models," *IEEE Trans. Signal Process.*, vol. 63, no. 4, pp. 1001–1016, Feb. 2015.

[5] S. Chouvardas, K. Slavakis, Y. Kopsinis, and S. Theodoridis, "A sparsity promoting adaptive algorithm for distributed learning," *IEEE Trans. Signal Process.*, vol. 60, no. 10, pp. 5412–5425, Oct. 2012.

[6] X. Zhao and A. H. Sayed, "Distributed clustering and learning over networks," *IEEE Trans. Signal Process.*, vol. 63, no. 13, pp. 3285–3300, Jul. 2015.

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

[8] K. Yuan, Q. Ling, and W. Yin, "On the convergence of decentralized gradient descent," *SIAM J. Optim.*, vol. 26, no. 3, pp. 1835–1854, 2016.

[9] G. Mateos, J. A. Bazerque, and G. B. Giannakis, "Distributed sparse linear regression," *IEEE Trans. Signal Process.*, vol. 58, no. 10, pp. 5262–5276, Oct. 2010.

[10] J. F. Mota, J. M. Xavier, P. M. Aguiar, and M. Püschel, "D-ADMM: A communication-efficient distributed algorithm for separable optimization," *IEEE Trans. Signal Process.*, vol. 61, no. 10, pp. 2718–2723, May 2013.

[11] W. Shi, Q. Ling, K. Yuan, G. Wu, and W. Yin, "On the linear convergence of the ADMM in decentralized consensus optimization," *IEEE Trans. Signal Process.*, vol. 62, no. 7, pp. 1750–1761, Apr. 2014.

[12] T.-H. Chang, M. Hong, and X. Wang, "Multi-agent distributed optimization via inexact consensus ADMM," *IEEE Trans. Signal Process.*, vol. 63, no. 2, pp. 482–497, Jan. 2014.

[13] N. S. Aybat, Z. Wang, T. Lin, and S. Ma, "Distributed linearized alternating direction method of multipliers for composite convex consensus optimization," *IEEE Trans. Autom. Control*, vol. 63, no. 1, pp. 5–20, Jan. 2017.

[14] W. Shi, Q. Ling, G. Wu, and W. Yin, "EXTRA: An exact first-order algorithm for decentralized consensus optimization," *SIAM J. Optim.*, vol. 25, no. 2, pp. 944–966, 2015.

[15] W. Shi, Q. Ling, G. Wu, and W. Yin, "A proximal gradient algorithm for decentralized composite optimization," *IEEE Trans. Signal Process.*, vol. 63, no. 22, pp. 6013–6023, Nov. 2015.

[16] J. Xu, S. Zhu, Y. C. Soh, and L. Xie, "Augmented distributed gradient methods for multi-agent optimization under uncoordinated constant stepsizes," in *Proc. IEEE Conf. Decis. Control*, 2015, pp. 2055–2060.

[17] P. D. Lorenzo and G. Scutari, "NEXT: In-network nonconvex optimization," *IEEE Trans. Signal Inf. Process. Over Netw.*, vol. 2, no. 2, pp. 120–136, Jun. 2016.

[18] A. Nedic, A. Olshevsky, and W. Shi, "Achieving geometric convergence for distributed optimization over time-varying graphs," *SIAM J. Optim.*, vol. 27, no. 4, pp. 2597–2633, 2017.

[19] G. Qu and N. Li, "Harnessing smoothness to accelerate distributed optimization," *IEEE Trans. Control Netw. Syst.*, vol. 5, no. 3, pp. 1245–1260, Sep. 2018.

[20] R. Xin and U. A. Khan, "A linear algorithm for optimization over directed graphs with geometric convergence," *IEEE Control Syst. Lett.*, vol. 2, no. 3, pp. 315–320, Jul. 2018.

[21] S. Pu, W. Shi, J. Xu, and A. Nedić, "A push-pull gradient method for distributed optimization in networks," in *Proc. IEEE Conf. Decis. Control*, 2018, pp. 3385–3390.

[22] K. Yuan, B. Ying, X. Zhao, and A. H. Sayed, "Exact dffusion for distributed optimization and learning – Part I: Algorithm development," *IEEE Trans. Signal Process.*, vol. 67, no. 3, pp. 708–723, Feb. 2019.

[23] Z. Li, W. Shi, and M. Yan, "A decentralized proximal-gradient method with network independent step-sizes and separated convergence rates," *IEEE Trans. Signal Process.*, vol. 67, no. 17, pp. 4494–4506, 2019.

[24] K. I. Tsianos, S. Lawlor, and M. G. Rabbat, "Push-sum distributed dual averaging for convex optimization," in *Proc. IEEE Conf. Decision Control.*, Maui, Hawaii, 2012, pp. 5453–5458.

[25] Chenguang Xi and Usman A Khan, "Dextra: A fast algorithm for optimization over directed graphs," *IEEE Trans. Autom. Control*, vol. 62, no. 10, pp. 4980–4993, Oct. 2017.

[26] J. Zeng and W. Yin, "ExtraPush for convex smooth decentralized optimization over directed networks," *J. Comput. Math.*, vol. 35, no. 4, 2017.

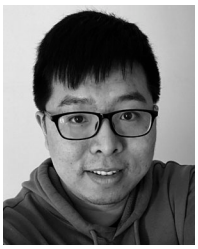
[27] A. Nedić and A. Olshevsky, "Stochastic gradient-push for strongly convex functions on time-varying directed graphs," *IEEE Trans. Autom. Control*, vol. 61, no. 12, pp. 3936–3947, Dec. 2016.

[28] S. Boyd, A. Ghosh, B. Prabhakar, and D. Shah, "Randomized gossip algorithms," *IEEE/ACM Trans. Netw.*, vol. 52, no. 6, pp. 2508–2530, Jun. 2006.

[29] M. Cao, D. A. Spielman, and E. M. Yeh, "Accelerated gossip algorithms for distributed computation," in *Proc. 44th Annu. Allerton Conf. Commun., Control, Comput.*, 2006, pp. 952–959.

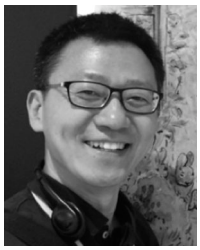
[30] H. Hendrikx, F. Bach, and L. Massoulié, "Accelerated decentralized optimization with local updates for smooth and strongly convex objectives," in *Proc. 22nd Int. Conf. Artif. Intell. Statist.*, 2019, pp. 897–906.
[31] D. P. Bertsekas, "A new class of incremental gradient methods for least squares problems," *SIAM J. Optim.*, vol. 7, no. 4, pp. 913–926, 1997.
[32] S. S. Ram, A Nedić, and V. V. Veeravalli, "Incremental stochastic subgradient algorithms for convex optimization," *SIAM J. Optim.*, vol. 20, no. 2, pp. 691–717, 2009.
[33] B. Johansson, M. Rabi, and M. Johansson, "A randomized incremental subgradient method for distributed optimization in networked systems," *SIAM J. Optim.*, vol. 20, no. 3, pp. 1157–1170, 2009.
[34] C. G. Lopes and A. H. Sayed, "Incremental adaptive strategies over distributed networks," *IEEE Trans. Signal Process.*, vol. 55, no. 8, pp. 4064–4077, Aug. 2007.
[35] C. G. Lopes and A. H. Sayed, "Randomized incremental protocols over adaptive networks," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, Dallas, TX, USA, 2010, pp. 3514–3517.
[36] S. M. Shah and K. E. Avrachenkov, "Linearly convergent asynchronous distributed ADMM via Markov sampling," 2018, *arXiv:1810.05067*.
[37] D. A. Levin and Y. Peres, *Markov Chains and Mixing Times*, vol. 107, 2nd ed. Providence, RI, USA: American Mathematical Soc., 2017.
[38] Z. Peng, Y. Xu, M. Yan, and W. Yin, "ARock: An algorithmic framework for asynchronous parallel coordinate updates," *SIAM J. Scientific Comput.*, vol. 38, no. 5, pp. A2851–A2879, 2016.
[39] M. Hong, Z.-Q. Luo, and M. Razaviyayn, "Convergence analysis of alternating direction method of multipliers for a family of nonconvex problems," *SIAM J. Optim.*, vol. 26, no. 1, pp. 337–364, 2016.
[40] Y. Wang, W. Yin, and J. Zeng, "Global convergence of ADMM in nonconvex nonsmooth optimization," *J. Scientific Comput.*, 2018, [Online]. Available: https://doi.org/10.1007/s10915-018-0757-z
[41] Tao Sun, Yuejiao Sun, and Wotao Yin, "On Markov chain gradient descent," in *Proc. Advances Neural Inf. Process. Syst.* 2018, pp. 9896–9905.
[42] R. Glowinski and A. Marroco, "Sur l'approximation, par éléments finis d'ordre un, et la résolution, par pénalisation-dualité d'une classe de problèmes de Dirichlet non linéaires," *ESAIM: Math. Modelling Numer. Anal.*, vol. 9, no. R2, pp. 41–76, 1975.
[43] D. Gabay and B. Mercier, "A dual algorithm for the solution of nonlinear variational problems via finite element approximation," *Comput. Math. Appl.*, vol. 2, no. 1, pp. 17–40, 1976.
[44] E. K. Ryu and W. Yin, "Proximal-proximal-gradient method," *J. Comput. Math.*, vol. 37, no. 6, pp. 778–812, 2019.
[45] R. Tyrrell Rockafellar and R. J-B Wets, *Variational Analysis*, vol. 317, Berlin Germany: Springer, 2009.
[46] E. N. Gilbert, "Random graphs," *Ann. Math. Statist.*, vol. 30, no. 4, pp. 1141–1144, 1959.
[47] F. Chung and M. Radcliffe, "On the spectra of general random graphs," *Electron. J. Combinatorics*, vol. 18, no. 1, pp. 215, 2011.
[48] H. Sun and M. Hong, "Distributed non-convex first-order optimization and information processing: Lower complexity bounds and rate optimal algorithms," in *Proc. 52nd Asilomar Conf. Signals, Syst., Comput.*, 2018, pp. 38–42.
[49] "Mnist dataset," 1998. [Online]. Available: http://yann.lecun.com/exdb/mnist/

Xianghui Mao received the B.E. and Ph.D. degrees from the Electronic Engineering Department, Tsinghua University, Beijing, China, in 2014 and 2019, respectively. Her Ph.D. advisor was Professor Yuantao Gu. She was a Visiting Graduate Researcher with the University of California, Los Angeles in 2017. Her recent research interests include decentralized optimization and machine learning.



Kun Yuan received the B.E. degree in telecommunication engineering from Xidian University, Xi'an, China, in 2011 and the M.S. degree in control theory and control engineering from the Department of Automation, University of Science and Technology of China, Hefei, China, 2014. He received the Ph.D. from the Electrical and Computer Engineering Department at University of California, Los Angeles, Los Angeles, CA, USA, in 2019. His research focus on decentralized optimization and machine learning.

Yubin Hu has been a student with the Department of Electronic Engineering, Tsinghua University, Beijing China, since 2016. He was a Visiting Undergraduate Research Intern of the Harvard John A. Paulson School of Engineering and Applied Sciences in 2019. His research interests include optimization related topics, 3D reconstruction and video processing.



Yuantao Gu (Senior Member, IEEE) received the B.E. degree from Xi'an Jiaotong University, Xi'an, China, in 1998, and the Ph.D. degree with honors from Tsinghua University, Beijing, China, in 2003, both in electronic engineering. He joined the faculty of Tsinghua University in 2003 and is currently a Professor with Department of Electronic Engineering. He was a Visiting Scientist with Microsoft Research Asia during 2005–2006, Research Laboratory of Electronics with the Massachusetts Institute of Technology during 2012–2013, and Department of Electrical Engineering and Computer Science with the University of Michigan in Ann Arbor during 2015. His research interests include high-dimensional statistics, sparse signal recovery, temporal-space and graph signal processing, related topics in wireless communications and information networks. He has been a Senior Area Editor for the IEEE TRANSACTIONS ON SIGNAL PROCESSING since 2019, an Elected Member of the IEEE Machine Learning for Signal Processing Technical Committee since 2019, and an Elected Member of the IEEE Signal Processing Theory and Methods (SPTM) Technical Committee since 2017. He was an Associate Editor for the IEEE TRANSACTIONS ON SIGNAL PROCESSING from 2015 to 2019 and a Handling Editor for *ELSEVIER Digital Signal Processing* from 2015 to 2017. He was the recipient of the Best Paper Award of IEEE GlobalSIP in 2015, the Award for Best Presentation of Journal Paper of ChinaSIP in 2015, and Zhang Si-Ying (CCDC) Outstanding Youth Paper Award (with his student) in 2017.

Ali H. Sayed (Fellow, IEEE) is the Dean of Engineering with EPFL, Lausanne, Switzerland. He was a Distinguished Professor and the Former Chairman of electrical engineering with the University of California, Los Angeles, where he established the UCLA Adaptive Systems Laboratory. He is the author of more than 530 scholarly publications and six books. His research involves several areas including adaptation and learning, data and network sciences, distributed optimization, and statistical inference. His work has been recognized with several awards including the 2014 Athanasios Papoulis Award from the European Association for Signal Processing, the 2015 Education Award, the 2013 Meritorious Service Award, and the 2012 Technical Achievement Award from the IEEE Signal Processing Society, the 2005 Terman Award from the American Society for Engineering Education, the 2003 Kuwait Prize, and the 1996 IEEE Donald G. Fink Prize. He was a Distinguished Lecturer for the IEEE Signal Processing Society in 2005 and an Editor-in Chief for the IEEE TRANSACTIONS ON SIGNAL PROCESSING (2003–2005). His articles received several best paper awards from the IEEE Signal Processing Society in 2002, 2005, 2012, and 2014. He is a fellow of the American Association for the Advancement of Science. He is recognized as a Highly Cited Researcher by Clarivate Analytics and is a member of the US National Academy of Engineering.

Wotao Yin received the master's and Ph.D. degrees in operations research from Columbia University, New York City, NY, USA, in 2003 and 2006, respectively. He is currently a Professor with the Department of Mathematics, University of California, Los Angeles, Los Angeles, CA, USA. His research interests include computational optimization and its applications in signal processing, machine learning, and other data science problems. He invented fast algorithms for sparse optimization, operator splitting, and large-scale distributed optimization problems. During 2006–2013, he was with Rice University. He received the NSF CAREER award in 2008, the Alfred P. Sloan Research Fellowship in 2009, and the Morningside Gold Medal in 2016, and has coauthored five papers receiving best paper-type awards. He is among the top 1% cited cross discipline and mathematical researchers by Clarivate Analytics in 2018 and 2019, respectively.