2022 American Control Conference (ACC)
Atlanta, USA, June 8-10, 2022

Distributed Optimization over Time-Varying Networks: Imperfect Information with Feedback is as Good as Perfect Information

Hadi Reisizadeh, Behrouz Touri, and Soheil Mohajer

***Abstract*— The convergence of an error-feedback algorithm is studied for decentralized stochastic gradient descent (DSGD) algorithm with compressed information sharing over time-varying graphs. It is shown that for both strongly-convex and convex cost functions, despite of imperfect information sharing, the convergence rates match those with perfect information sharing. To do so, we show that for strongly-convex loss functions, with a proper choice of a step-size, the state of *each* node converges to the global optimizer at the rate of $\mathcal{O}\left(T^{-1}\right)$. Similarly, for general convex cost functions, with a proper choice of step-size, we show that the value of loss function at a *temporal average* of *each* node's estimates converges to the optimal value at the rate of $\mathcal{O}(T^{-1/2+\epsilon})$ for any $\epsilon > 0$.**

I. Introduction

Due to the emergence of big data analytics over large-scale computing architectures, the study of multi-agent (time-varying) networks has received significant attention in various application domains, such as large-scale machine learning [1], power control [2], and sensor networks [3], [4]. In general and in the absence of a central server, we deal with decentralized computing nodes (agents) that are interested in collaboratively solving optimization problems. Moreover, due to privacy and data ownership concerns, each node performs local and on-device computation on its own available data. To ensure the convergence to an optimal solution of the original problem, nodes exchange information over a (time-varying) network. However, exchanging exact information can potentially introduce a massive communication overhead in the network. This communication overhead increases as the number of the decision variables grows. In this paper, we study distributed optimization problems and propose a gradient-based distributed algorithm that addresses the limitation of the communication load over time-varying graphs, by exchanging compressed information. Surprisingly, the algorithm achieves the same performance as one with exact information sharing.

Related Works. Various algorithms and methods have been proposed for convex [5]–[7] and non-convex [8], [9] distributed optimization problems, where various sub-gradient methods with diminishing step-sizes have been used to show convergence to a desired point. These works assume an exact information sharing among nodes, i.e., each node is allowed to communicate real-valued vectors with its neighbors. Various compression approaches have been used in the literature to mitigate the communication constraint [10]–[13].

H. Reisizadeh (email: hadir@umn.edu) and S. Mohajer (email: soheil@umn.edu) are with the University of Minnesota, and B. Touri (email: btouri@ucsd.edu) is with the University of California San Diego. The work of H. Reisizadeh and S. Mohajer is supported in part by the National Science Foundation under Grants CCF-1749981.

A DSGD algorithm with quantized information sharing over a fixed graph is presented in [11]. It is shown that, for strongly convex functions, the (time-invariant) step-sizes of the algorithm can be tuned in terms of a given *stopping time* (iteration) T, such that the algorithm reaches a point within $cT^{-1/2+\epsilon}$-neighborhood of the optimum point, for some $c > 0$ and any $\epsilon > 0$ [11]. However, the theoretical result only holds for $T \geq T_{\min}$, where $T_{\min}$ depends on ϵ, and grows unboundedly as ϵ vanishes. In our recent work [14], [15], a two time-scale DSGD algorithm, DIMIX, with two vanishing step-sizes is proposed for time-varying networks and imperfect information sharing. One time-scale suppresses the noise induced by incoming information from neighboring agents, and the other time-scale regulates the local loss functions' gradients. It is shown that for strongly convex setting with a proper choice for step-sizes, each node finds the optimal solution at the rate of $\mathcal{O}(T^{-1/2})$ [14]. A similar result is presented in a parallel work [16], for time-invariant networks. Although these proposed algorithms address noisy information sharing over time-varying graphs, they offer a reduced speed of convergence in comparison to the exact information sharing methods. In this regard, an error-feedback mechanism with diminishing step-size over fixed-network is introduced in [10], and it is shown to achieve the convergence rate of $\mathcal{O}(T^{-1})$ only for strongly convex functions. In another related work [12], error-feedback is used for the push-sum algorithm with fixed step-size (that depends on the stopping time T) over a fixed network. It is shown that the algorithm stops at a point close to a desired point of the loss function.

Contributions. In this work, the distributed optimization problem with quantized information sharing over *time-varying* networks is studied. Each node updates a local state by exploiting local computations and the quantized information received from its neighbors. We extend the convergence analysis of a novel decentralized stochastic gradient descent algorithm that utilizes diminishing step-sizes and a feedback mechanism to damp the quantization noise, proposed in [10]. We show that for strongly convex loss functions with a proper choice of step-size, each node's state converges to the optimal point at a rate of $\mathcal{O}(T^{-1})$, which matches with the convergence rate of the DSGD with perfect information sharing [5]. Also, for convex loss functions, we show that the loss associated with the temporal average of the states' for *each* node converges to the optimal loss at the rate of $\mathcal{O}(T^{-1/2+\epsilon})$ for any $\epsilon > 0$. To establish our results, we analyze the convergence rate of the close-loop system with a certain sum-product expression, then provide a novel analysis for the asymptotic behavior and the convergence rate of

978-1-6654-5196-3/$31.00 ©2022 AACC

the sum-product expression. Our simulations results strongly support the tightness of the provided analysis.

Notation and Basic Terminology. We use $[n]$ to denote $\{1, 2, \ldots, n\}$. Since we are dealing with minimizing a function in $\mathbb{R}^d$, we assume that the underlying functions are acting on **row** vectors, i.e., $\mathbf{x} \in \mathbb{R}^{1\times d} = \mathbb{R}^d$. The rest of the vectors, i.e., those in $\mathbb{R}^{n\times 1} = \mathbb{R}^n$, are assumed to be column vectors. The L_2-norm of a vector $\mathbf{x} \in \mathbb{R}^d$ is defined by $\|\mathbf{x}\|^2 = \sum_{j=1}^d |x_j|^2$, and the Frobenius norm of a matrix $A \in \mathbb{R}^{n\times d}$ is defined by $\|A\|_F^2 = \sum_{i=1}^n \|A_i\|^2 = \sum_{i=1}^n \sum_{j=1}^d |A_{ij}|^2$, where A_i denotes the ith row of A. A vector $\mathbf{r} \in \mathbb{R}^n$ is called stochastic if $r_i \geq 0$ and $\sum_{i=1}^n r_i = 1$. A non-negative matrix $A \in \mathbb{R}^{n\times d}$ is called (row) stochastic if $\sum_{j=1}^d A_{ij} = 1$ for every $i \in [n]$.

For an $n \times d$ matrix A and a strictly positive stochastic vector $\mathbf{r} \in \mathbb{R}^n$, we define the $\mathbf{r}$-norm of A by $\|A\|_{\mathbf{r}}^2 = \sum_{i=1}^n r_i \|A_i\|^2$. It can be verified that $\|\cdot\|_{\mathbf{r}}$ is a norm on the space of $n \times d$ matrices.

II. Problem Setup and Main Results

In this section, we discuss the problem formulation and the main results of this work.

A. Problem Setup

Consider a time-varying network of $n \geq 2$ agents. Each agent i has the cost function $f_i : \mathbb{R}^{1\times d} \to \mathbb{R}$. The goal is to minimize the function $f(\mathbf{x}) := \sum_{i=1}^n r_i f_i(\mathbf{x})$, or equivalently solve the following consensus optimization problem

$$\min_{\mathbf{x}_1,\ldots,\mathbf{x}_n \in \mathbb{R}^d} \sum_{i=1}^n r_i f_i(\mathbf{x}_i) \quad \text{s.t.} \quad \mathbf{x}_1 = \mathbf{x}_2 = \cdots = \mathbf{x}_n, \quad (1)$$

where $\mathbf{r} = (r_1, \ldots, r_n)$ is a stochastic vector.

We represent the time-varying network at time $t \geq 1$ by a directed weighted graph $\mathcal{G}(t) = ([n], \mathcal{E}, W(t))$, where the vertex set $[n]$ is the set of agents, the set $\mathcal{E} \subseteq [n] \times [n]$ is the set of edges/links connecting them, and the $n \times n$ matrix $W(t)$ represents the weight of the edges on the graph. In our model, the entry $W_{ij}(t)$ is the weight that agent i assigns to the incoming information from agent j. Here, $W_{ij}(t) > 0$ only if $(j, i) \in \mathcal{E}$. Let $\mathcal{N}_{\text{in}}^i := \{j \mid (j, i) \in \mathcal{E}\}$ and $\mathcal{N}_{\text{out}}^i := \{j \mid (i, j) \in \mathcal{E}\}$ be the sets of in-neighbors and out-neighbors of node $i \in [n]$, respectively.

To discuss our Perturbed Compressed Decentralized Stochastic Gradient Descent (PCOM-DSGD) algorithm, let $\mathbf{x}_j(t)$ be the state of node j (of a global optimizer of (1) at time t. We also use $\hat{\mathbf{x}}_j(t)$ to denote the estimate of $\mathbf{x}_j(t)$, reconstructed at *every* out-neighbors of node j. This estimate can be reconstructed recursively. At time t, node j losslessly broadcasts the quantized vector $Q(\mathbf{x}_j(t) - \hat{\mathbf{x}}_j(t-1))$ to every $i \in \mathcal{N}_{\text{out}}^j$. Then, node i updates its estimate of $\mathbf{x}_j(t)$ as

$$\hat{\mathbf{x}}_j(t) = \hat{\mathbf{x}}_j(t-1) + Q(\mathbf{x}_j(t) - \hat{\mathbf{x}}_j(t-1)). \quad (2)$$

Next, each node i computes its local gradient ∇f_i at $\mathbf{x}_i(t)$, and updates its local decision variable as

$$\mathbf{x}_i(t+1) = W_{ii}(t)\mathbf{x}_i(t) + \sum_{j\in\mathcal{N}_{\text{in}}^i} W_{ij}(t)\hat{\mathbf{x}}_j(t) - \alpha(t)\nabla f_i(\mathbf{x}_i(t)),$$

where $\alpha(t) = \frac{\alpha_0}{(t+\tau)^\nu}$ is the diminishing gradient step-size. Here, we assume $\nu \in (0, 1]$ and $\alpha_0, \tau \in \mathbb{R}^+$ are fixed throughout the algorithm. For notational convenience, assume that $\mathbf{x}_i(t)$, $\hat{\mathbf{x}}_i(t)$, and $\nabla f_i(\mathbf{x}_i(t))$ are all row vectors and consider matrices $X(t), E(t), \nabla f(X(t))$, whose ith rows are $\mathbf{x}_i(t), \mathbf{e}_i(t)$, and $\nabla f_i(\mathbf{x}_i(t))$, respectively. Thus, our update algorithm

$$X(t+1) = W(t)X(t) + E(t) - \alpha(t)\nabla f(X(t)), \quad (3)$$

where $E(t) := (W(t) - W_D(t))(\hat{X}(t) - X(t))$ is the accumulative perturbation matrix, and $W_D(t)$ is a diagonal matrix with $W_{ii}(t)$ as its ith diagonal entry.

Remark 1: Note that the dynamics in (3) with $E(t) = 0$ subsumes the decentralized gradient descent methods with *exact* information sharing.

In the rest of this paper, we will show that under some conditions, $E(t)$ decays to zero, and consequently, the proposed algorithm convergences to a desired optimal point.

B. Assumptions

We make the following assumptions on the quantizer, the sequence of stochastic weight matrices $\{W(t)\}$, and the local objective functions $\{f_i : i \in [n]\}$.

*Assumption 1 (**Quantizer Assumption**):* The quantizer $Q(\cdot)$ is a random quantizer, i.e., for any $\mathbf{x} \in \mathbb{R}^d$, $Q(\mathbf{x})$ is a random vector in $\mathbb{R}^d$. Furthermore, it is an unbiased and has a bounded average distortion (in L_2-norm), i.e., for some fixed $\omega \in (0, 1]$ and all $x \in \mathbb{R}^d$,

$$\mathbb{E}\left[Q(\mathbf{x})\right] = \mathbf{x}, \quad \mathbb{E}\left[\|Q(\mathbf{x}) - \mathbf{x}\|^2\right] \leq \omega\|\mathbf{x}\|^2. \quad (4)$$

Finally, we assume that $Q(\cdot)$ is an independent quantizer, i.e., the collection of random vectors $\{Q(\mathbf{x}_i(t) - \hat{\mathbf{x}}_i(t-1)) : i \in [n], t \geq 1\}$ in (2) is an independent collection of random vectors.

In the following, we assume that Assumption 1 holds for some $\omega \leq \omega_0$, where ω_0 depends on the problem parameters. Let us discuss an (important) example of such a quantizer.

Example 1: **Stochastic Quantizer.** For a number of quantization levels s (with $\log_2(s)$ bits of communication per dimension), and a vector $\mathbf{x} \in \mathbb{R}^d$, the stochastic quantizer $Q_s^S(\mathbf{x})$ is a random vector in $\mathbb{R}^d$, where

$$\left[Q_s^S(\mathbf{x})\right]_j = \|\mathbf{x}\| \cdot \text{sgn}(x_j) \cdot \zeta\left(\frac{|x_j|}{\|\mathbf{x}\|}, s\right), \quad j \in [d], \quad (5)$$

and $\zeta_j(x, s)$ is a random variable taking values in $\{0, \frac{1}{s}, \frac{2}{s}, \ldots, \frac{s-1}{s}\}$, and we have[1]

$$\zeta(x, s) = \begin{cases} \lceil sx \rceil / s & \text{w.p. } sx - \lfloor sx \rfloor \\ \lfloor sx \rfloor / s & \text{w.p. } \lceil sx \rceil - sx. \end{cases} \quad (6)$$

It is shown in [17] that the stochastic quantizer is unbiased (i.e., $\mathbb{E}\left[Q_s^S(\mathbf{x})\right] = \mathbf{x}$) and satisfies

$$\mathbb{E}\left[\|Q_s^S(\mathbf{x}) - \mathbf{x}\|^2\right] \leq \min\left(d/s, \sqrt{d}/s\right)\|\mathbf{x}\|^2.$$

[1]We define $\zeta(x, s) = x$ whenever sx is an integer.

Therefore, for $d < s^2$, the stochastic quantizer would satisfy Assumption 1 with $\omega = d/s^2$. Note that we can always tune the parameter s to guarantee $\omega \leq \omega_0$.

We need certain connectivity conditions for the underlying network to guarantee the convergence of the algorithm.

*Assumption 2 (**Connectivity Assumption**):* We assume that the weight matrix sequence $\{W(t)\}$ in (3) satisfies the following properties

(a) *Stochastic with Common Stationary Distribution*: $W(t)$ is row-stochastic and $\mathbf{r}^T W(t) = \mathbf{r}^T$ for all $t \geq 1$, where $\mathbf{r} > 0$ is the given weight vector.

(b) *Bounded Nonzero Elements*: There exists some $\eta > 0$ such that if for some $i, j \in [n]$ and $t \geq 1$ we have $W_{ij}(t) > 0$, then $W_{ij}(t) \geq \eta$.

(c) *B-Connected*: For a fixed integer $B \geq 1$, the graph $\left([n], \bigcup_{k=t+1}^{t+B} \mathcal{E}(k)\right)$ is strongly connected for all $t \geq 1$, where $\mathcal{E}(k) = \{(j,i) \mid W_{ij}(k) > 0\}$.

The next assumption describes the properties of the loss functions we study in this work.

*Assumption 3 (**Objective Function Assumptions**):* We assume the following properties on the function f_i for all i

(a) The function ∇f_i is L-Lipschitz, i.e., for any $\mathbf{x}, \mathbf{y} \in \mathbb{R}^d$ we have that $\|\nabla f_i(\mathbf{x}) - \nabla f_i(\mathbf{y})\| \leq L\|\mathbf{x} - \mathbf{y}\|$.

(b) The function f_i is μ-strongly convex, i.e., for any $\mathbf{x}, \mathbf{y} \in \mathbb{R}^d$, we have $\langle \nabla f_i(\mathbf{x}) - \nabla f_i(\mathbf{y}), \mathbf{x} - \mathbf{y} \rangle \geq \mu \|\mathbf{x} - \mathbf{y}\|^2$.

(c) f_is have bounded gradients, i.e., there exists a scalar $K > 0$ such that $\|\nabla f_i(\mathbf{x})\|^2 \leq K$ for all $\mathbf{x} \in \mathbb{R}^d$.

Remark 2: Using Assumptions 3-(a), 3-(c), and the mean-value theorem for $g_i : [0,1] \to \mathbb{R}$ given by $g_i(t) = f_i(\mathbf{x} + t(\mathbf{y} - \mathbf{x}))$, we can conclude that $f_i(\mathbf{x})$ is a $\sqrt{K}$-Lipschitz function for all $i \in [n]$.

C. Main Results and Discussion

With these preliminary discussions, we are ready to present the main results of the paper. In the rest of this paper, we refer to any quantity that does not depend on iteration t simply as a *constant*. Such a constant can possibly depend on the parameters of the problem, i.e., the network parameters $(n, B, \eta, \mathbf{r})$, function parameters (L, μ, K), dynamic parameters (α_0, ν, τ), and quantizer parameter ω.

Strongly Convex Objectives. For strongly convex loss functions, we have the following result regarding the convergence rate of each node's estimate to the optimal point defined by $\mathbf{x}^\star := \arg\min f(\mathbf{x})$.

Theorem 1: Suppose that Assumptions 1–3 hold and let $\alpha(t) = \frac{\alpha_0}{(t+\tau)^\nu}$. Then, if $\nu \in (0,1)$ the dynamics in (3) satisfies

$$\mathbb{E}\left[\|X(T) - \mathbf{1}\mathbf{x}^\star\|_{\mathbf{r}}^2\right] = \mathcal{O}\left(T^{-\nu}\right), \tag{7}$$

provided that $\tau \geq \tau_1$. Moreover, if $\nu = 1$, $\alpha_0 \geq \frac{\mu + L}{\mu L}$, and $\tau \geq \tau_2$ we have

$$\mathbb{E}\left[\|X(T) - \mathbf{1}\mathbf{x}^\star\|_{\mathbf{r}}^2\right] = \mathcal{O}\left(T^{-1}\right). \tag{8}$$

Here, τ_1 and τ_2 are constant shift parameters.

This result follows from Proposition 1 and Proposition 2, and its sketch is presented in Section III. We refer to the long version of the paper [18] for its detailed proof.

Remark 3: Theorem 1 guarantees the exact convergence (in L_2 sense) of each local state to the global optimal with diminishing step-size. When $\nu = 1$ we get the maximum exponent for the convergence rate as in (8), which recovers the convergence rate of decentralized gradient descent method with exact information sharing [19]. This shows that the feedback mechanism suppresses the noise generated by the random quantizer. Our algorithm is inspired by DIMIX [14] (and a similar work in [16]), where a second vanishing time-scale is used to damp the quantization noise without having a feedback mechanism. The additional time-scale leads to a slower convergence of $\mathcal{O}(T^{-1/2})$ as reported in [14].

Remark 4: In a related work [10], for strongly convex loss functions, under the limited setting of *time-invariant* networks, *uniform* weights $\mathbf{r} = \mathbf{1}$, and a *particular choice* of parameter $\nu = 1$, an algorithm is proposed where achieves the same convergence rate estimate $\mathcal{O}(T^{-1})$.

Convex Objectives. The next theorem shows the convergence of the algorithm for general convex cost functions.

Theorem 2: Under Assumptions 1, 2, and Assuming 3-(a), 3-(c), and convexity for the local cost functions $f_i(\cdot)$, the dynamics (3) satisfies

$$M_\theta(\nu) := \left[\frac{1}{T}\sum_{t=1}^{T}\left[\mathbb{E}\left[f(\mathbf{x}_i(t))\right] - f(\mathbf{x}^\star)\right]^\theta\right]^{\frac{1}{\theta}} = \mathcal{O}\left(T^{-\min\{\nu, 1-\nu\}}\right),$$

for $\alpha(t) = \frac{\alpha_0}{(t+\tau)^\nu}$ with $\nu \in (0,1]$, $\nu \neq 1/2$, and all $\theta \in (0,1)$, provided that $\tau \geq \tau_0$. Moreover, for $\nu^\star = 1/2$, we have the optimal convergence rate of

$$M_\theta(\nu^\star) = \mathcal{O}\left(T^{-1/2}\ln T\right).$$

Corollary 1: Under the conditions of Theorem 2, for the optimum choice of $\nu^\star = 1/2$, we get

$$\mathbb{E}\left[f\left(\frac{1}{T}\sum_{t=1}^{T}\mathbf{x}_i(t)\right)\right] - f(\mathbf{x}^\star) = \mathcal{O}\left(T^{-1/2+\epsilon}\right),$$

for any $\epsilon > 0$.

Remark 5: In a related work [12], a quantized push-sum algorithm is presented for convex objective loss functions under a limited setting of *time-invariant* networks, *uniform* weights $\mathbf{r}$, and a *fixed* step-size, that depends on a the *stopping time* T, and for some constant $c > 0$ it is shown that

$$\mathbb{E}\left[f\left(\frac{1}{T}\sum_{t=1}^{T}\mathbf{x}_i(t)\right)\right] - f(\mathbf{x}^\star) \leq cT^{-1/2}.$$

III. Proofs and Discussions

Here, we provide the proof sketches of the main results presented in this paper. The proof of the main theorems are built on four major lemmas. The details of these proofs are far beyond the page limits of the current publication but they are provided in [18]. First, we present a technical lemma that plays an important role in the proof of the main results. We refer to [18] for the proof of the lemma.

Lemma 1: For any $0 \leq \delta < \min(1,\sigma)$, and every $t > t_0 := (\frac{2(\sigma-\delta)}{a})^{\frac{1}{1-\delta}}$, we have

$$\sum_{s=1}^{t-1}\left[\frac{1}{s^\sigma}\prod_{k=s+1}^{t-1}\left(1-\frac{a}{k^\delta}\right)\right] \leq A(a,\sigma,\delta)t^{-(\sigma-\delta)},$$

where $A(a,\sigma,\delta)$ is a constant that only depends on a, σ, and δ. Moreover, for $\delta = 1$ and $a-\sigma+1 \neq 0$, we have

$$\sum_{s=1}^{t-1}\left[\frac{1}{s^\sigma}\prod_{k=s+1}^{t-1}\left(1-\frac{a}{k}\right)\right] \leq A(a,\sigma,1)t^{-\min(\sigma-1,a)},$$

where $A(a,\sigma,1) = 2^\sigma\left(1+\frac{1}{|a-\sigma+1|}\right)$.

The following proposition bounds the deviation of each node's estimate from the average state. Let us denote the average state by $\bar{\mathbf{x}}(t) := \mathbf{r}^T X(t) = \sum_{i=1}^n r_i \mathbf{x}_i(t)$ for $t \geq 1$.

Proposition 1: Under Assumptions 1, 2, and 3-(c) for $\alpha(t) = \frac{\alpha_0}{(t+\tau)^\nu}$ when $\nu \in (0,1]$, the dynamics (3) satisfies

$$\mathbb{E}\left[\|X(T) - \mathbf{1}\bar{\mathbf{x}}(T)\|_{\mathbf{r}}^2\right] \leq \xi_1\alpha^2(T), \tag{9}$$

for any $T \geq 1$, provided that $\tau \geq \tau_0$. Here ξ_1 is a constant.

Proof Sketch of Proposition 1: The proof is based on establishing two recursive inequalities between $\phi(t) := \mathbb{E}\left[\|X(t)-\mathbf{1}\bar{\mathbf{x}}(t)\|_{\mathbf{r}}^2\right]$ and $\psi(t) := \mathbb{E}\left[\|\hat{X}(t)-X(t)\|^2\right]$. Using the linearity of (3), the fact that $\{E(t)\}$ is a zero-mean independent process, and $\{W(t)\}$ admits a common stationary distribution $\mathbf{r}$, we can show

$$\mathbb{E}\left[\|X(t)-\mathbf{1}\bar{\mathbf{x}}(t)\|_{\mathbf{r}}^2\right] \leq 2\mathbb{E}\left[\left\|\sum_{s=1}^{t-1}\alpha(s)P(t:s)\nabla f(X(s))\right\|_{\mathbf{r}}^2\right] + 2\sum_{s=1}^{t-1}\mathbb{E}\left[\|P(t:s)E(s)\|_{\mathbf{r}}^2\right], \tag{10}$$

where $P(t:s) := W(t-1)\cdots W(s+1) - \mathbf{1}\mathbf{r}^T$. Furthermore, the B-connectivity condition in Assumption 2 implies that $\|P(t:s)U\|_{\mathbf{r}}^2 \leq \kappa(1-\lambda)^{t-s-1}\|U\|_{\mathbf{r}}^2$ for any matrix U, where $\lambda := \frac{\eta\mathbf{r}_{\min}}{2Bn^2}$, $\kappa := (1-B\lambda)^{-1}$. Using the above contraction, and the bounded gradient condition in Assumption 3-(c), for the first term in (10), we have

$$2\mathbb{E}\left[\left\|\sum_{s=1}^{t-1}\alpha(s)P(t:s)\nabla f(X(s))\right\|_{\mathbf{r}}^2\right] \leq \xi_2\alpha^2(t), \tag{11}$$

where ξ_2 is constant. Moreover, using the facts that $\|AB\|_{\mathbf{r}}^2 \leq \|A\|_{\mathbf{r}}^2\|B\|_F^2$ and $\|W(s)-W_D(s)\|_{\mathbf{r}}^2 \leq 1$, we can write $\|E(s)\|_{\mathbf{r}}^2 \leq \|\hat{X}(s)-X(s)\|^2$. Hence, (10) leads to

$$\phi(t) \leq \xi_2\alpha^2(t) + 2\kappa\sum_{s=1}^{t-1}(1-\lambda)^{t-s-1}\psi(s). \tag{12}$$

On the other hand, we have

$$\begin{aligned}\psi(t+1) &= \mathbb{E}\left[\mathbb{E}\left[\|\hat{X}(t+1)-X(t+1)\|^2\middle|\mathcal{F}_t\right]\right]\\ &\overset{(a)}{=} \mathbb{E}\left[\mathbb{E}\left[\|\hat{X}(t)+Q(X(t+1)-\hat{X}(t))-X(t+1)\|^2\middle|\mathcal{F}_t\right]\right]\\ &\overset{(b)}{\leq} \omega\mathbb{E}\left[\|X(t+1)-\hat{X}(t)\|^2\right],\end{aligned} \tag{13}$$

where (a) follows from (2) and (b) follows from Assumption 1. Rewriting $X(t+1)$ from (3) we arrive at

$$\begin{aligned}X(t+1)-\hat{X}(t) &= W_D(t)X(t) + (W(t)-W_D(t))\hat{X}(t)\\ &\quad -\alpha(t)\nabla f(X(t)) - \hat{X}(t)\\ &\overset{(c)}{=} (W(t)-W_D(t)-I)(\hat{X}(t)-X(t))\\ &\quad + (W(t)-I)(X(t)-\mathbf{1}\bar{\mathbf{x}}(t)) - \alpha(t)\nabla f(X(t)),\end{aligned}$$

where in (c) we used the fact that $W(s)\mathbf{1} = \mathbf{1}$ for every $s \geq 1$. Then, convexity of the norm implies that

$$\begin{aligned}&\|X(t+1)-\hat{X}(t)\|^2\\ &\leq 3\|W(t)-W_D(t)-I\|^2\|\hat{X}(t)-X(t)\|^2\\ &\quad + 3\|W(t)-I\|^2\|X(t)-\mathbf{1}\bar{\mathbf{x}}(t)\|^2 + 3\alpha^2(t)\|\nabla f(X(t))\|^2\\ &\overset{(d)}{\leq} 6n\|\hat{X}(t)-X(t)\|^2 + 3n\|X(t)-\mathbf{1}\bar{\mathbf{x}}(t)\|^2 + 3nK\alpha^2(t),\end{aligned}$$

where (d) follows from Assumption 3-(c). Taking expectation from both sides and plugging that into (13) we get

$$\psi(t+1) \leq 3n\omega\left(2\psi(t) + \mathbf{r}_{\min}^{-1}\phi(t) + K\alpha^2(t)\right), \tag{14}$$

in which we used the fact that $\|U\|^2 \leq \mathbf{r}_{\min}^{-1}\|U\|_{\mathbf{r}}^2$ for any matrix U. Next, having (12) and (14), we can use induction on t to prove that $\phi(t) \leq \xi_1\alpha^2(t)$ and $\psi(t) \leq \xi_3\alpha^2(t)$ for every $t \geq 1$. Due to the page limit, the induction details are provided in [18]. It is worth noting that the condition $\omega \leq \omega_0$ is required to complete the induction. ■

The next proposition bounds bounds the gap between the average state and the optimal point of a strongly-convex function, i.e., $\mathbf{x}^\star := \arg\min f(\mathbf{x})$.

Proposition 2: Let $\alpha(t) := \frac{\alpha_0}{(t+\tau)^\nu}$. Under Assumptions 1–3, the dynamics (3) satisfies

$$\mathbb{E}\left[\|\bar{\mathbf{x}}(T)-\mathbf{x}^\star\|^2\right] \leq \xi_4 e^{-\xi_5(T+\tau)^{1-\nu}} + \xi_6(T+\tau)^{-\nu},$$

for any iteration $T \geq 1$, provided that $\nu \in (0,1)$ and $\tau \geq \tau_1$. Furthermore, if $\nu = 1$ and $\alpha_0 \geq \frac{\mu+L}{\mu L}$, we get

$$\mathbb{E}\left[\|\bar{\mathbf{x}}(T)-\mathbf{x}^\star\|^2\right] \leq \xi_7(T+\tau)^{-1},$$

for any $T \geq 1$, assuming that $\tau \geq \tau_2$. Note that $\xi_4, \xi_5, \xi_6, \xi_7$ and as well as τ_1 and τ_2 are constant.

Proof Sketch of Proposition 2: Define $\bar{\mathbf{x}}(t) = \mathbf{r}^T\mathbf{x}(t)$, $g(t) := \mathbf{r}^T\nabla f(X(t)) = \sum_i r_i\nabla f_i(\mathbf{x}_i(t))$ and $V(t) := \|\bar{\mathbf{x}}(t)-\mathbf{x}^\star\|^2$. Then, multiplying both sides of (3) by $\mathbf{r}^T$, subtracting $\mathbf{x}^\star$, and using the fact that $\mathbf{r}^T W(t) = \mathbf{r}^T$, we get

$$\bar{\mathbf{x}}(t+1) - \mathbf{x}^\star = \bar{\mathbf{x}}(t) - \mathbf{x}^\star - \alpha(t)g(t) + \mathbf{r}^T E(t). \tag{15}$$

Since, $\mathbb{E}[E(t)] = 0$ (see Assumption 1), from (15) we have

$$\begin{aligned}&\mathbb{E}[V(t+1)\mid\mathcal{F}_t] = \mathbb{E}\left[\|\bar{\mathbf{x}}(t+1)-\mathbf{x}^\star\|^2\middle|\mathcal{F}_t\right]\\ &= \|\bar{\mathbf{x}}(t)-\mathbf{x}^\star-\alpha(t)g(t)\|^2 + \mathbb{E}\left[\|\mathbf{r}^T E(t)\|^2\middle|\mathcal{F}_t\right].\end{aligned} \tag{16}$$

Letting $\bar{g}(t) := \nabla f(\bar{\mathbf{x}}(t)) = \sum_{i=1}^n r_i\nabla f_i(\bar{\mathbf{x}}(t))$, we have

$$\begin{aligned}&\|\bar{\mathbf{x}}(t) - \mathbf{x}^\star - \alpha(t)g(t)\|^2\\ &\quad = \|\bar{\mathbf{x}}(t) - \mathbf{x}^\star - \alpha(t)\bar{g}(t) + \alpha(t)\bar{g}(t) - \alpha(t)g(t)\|^2\\ &\quad \leq (1+\rho(t))\|\bar{\mathbf{x}}(t) - \mathbf{x}^\star - \alpha(t)\bar{g}(t)\|^2\\ &\quad\quad + \alpha^2(t)(1+1/\rho(t))\|g(t)-\bar{g}(t)\|^2\end{aligned} \tag{17}$$

for any $\rho(t) > 0$. Now, we can use [14, Lemma 7] for *strongly convex* function $f(\cdot)$ and write

$$\langle \bar{\mathbf{x}}(t) - \mathbf{x}^\star, \bar{g}(t) - 0 \rangle \geq c_1 \|\bar{g}(t)\|^2 + c_2 \|\bar{\mathbf{x}}(t) - \mathbf{x}^\star\|^2, \quad (18)$$

where $c_1 = \frac{1}{\mu+L}$ and $c_2 = \frac{\mu L}{\mu+L}$. Therefore, for the first term in (17), we can write

$$\begin{aligned}
&\|\bar{\mathbf{x}}(t) - \mathbf{x}^\star - \alpha(t)\bar{g}(t)\|^2 \\
&= \|\bar{\mathbf{x}}(t) - \mathbf{x}^\star\|^2 + \alpha^2(t)\|\bar{g}(t)\|^2 - 2\alpha(t)(\bar{\mathbf{x}}(t) - \mathbf{x}^\star)^T \bar{g}(t) \\
&\leq (1 - 2c_2\alpha(t))\|\bar{\mathbf{x}}(t) - \mathbf{x}^\star\|^2 + \alpha(t)(\alpha(t) - 2c_1)\|\bar{g}(t)\|^2 \\
&= (1-2c_2\alpha(t))V(t) + \alpha(t)(\alpha(t)-2c_1)\|\bar{g}(t)\|^2. \quad (19)
\end{aligned}$$

Hence, for sufficiently large t with $\alpha(t) \leq 2c_1$ we have

$$\|\bar{\mathbf{x}}(t) - \mathbf{x}^\star - \alpha(t)\bar{g}(t)\|^2 \leq (1-2c_2\alpha(t))V(t). \quad (20)$$

Moreover, convexity of $\|\cdot\|^2$ and Assumption 3-(a) imply

$$\begin{aligned}
\mathbb{E}\left[\|g(t) - \bar{g}(t)\|^2\right] &\leq \mathbb{E}\left[\sum_{i=1}^n r_i \|\nabla f_i(\mathbf{x}_i(t)) - \nabla f_i(\bar{\mathbf{x}}(t))\|^2\right] \\
&\leq L \sum_{i=1}^n r_i \mathbb{E}\left[\|\bar{\mathbf{x}}(t) - \mathbf{x}_i(t)\|^2\right] \leq L\xi_1 \alpha^2(t), \quad (21)
\end{aligned}$$

where the last inequality follows from Lemma 1. Taking expectation from both sides of (17) and using (20) and (21) with $\rho(t) = c_2\alpha(t)$, we arrive at

$$\begin{aligned}
&\mathbb{E}\left[\|\bar{\mathbf{x}}(t) - \mathbf{x}^\star - \alpha(t)g(t)\|^2\right] \quad (22) \\
&\leq (1 - c_2\alpha(t))\mathbb{E}\left[V(t)\right] + L\xi_1\alpha(t)(\alpha(t) + 1/c_2)\alpha^2(t) \\
&\leq (1 - c_2\alpha(t))\mathbb{E}\left[V(t)\right] + L\xi_1\alpha_0(\alpha_0 + 1/c_2)\alpha^2(t).
\end{aligned}$$

The second term in (16) can be also bounded as

$$\begin{aligned}
&\mathbb{E}\left[\mathbb{E}\left[\|\mathbf{r}^T E(t)\|^2 \big| \mathcal{F}_t\right]\right] \leq \mathbb{E}\left[\|\mathbf{r}^T E(t)\|^2\right] \\
&\leq \|\mathbf{r}\|^2 \|W(s) - W_D(s)\|^2 \mathbb{E}\left[\|\hat{X}(t) - X(t)\|^2\right] \\
&\leq \psi(t) \leq \xi_2 \alpha^2(t), \quad (23)
\end{aligned}$$

where the last inequality is discussed in the proof of Lemma 1. Taking expectation from both sides of (16) and using (22) and (23), we arrive at the following recursive inequality

$$\begin{aligned}
\mathbb{E}\left[V(t+1)\right] &= \mathbb{E}\left[\mathbb{E}\left[V(t+1) \mid \mathcal{F}_t\right]\right] \quad (24) \\
&\leq (1 - c_2\alpha(t))\mathbb{E}\left[V(t)\right] + \xi_8 \alpha^2(t),
\end{aligned}$$

where $\xi_8 := L\xi_1\alpha_0(\alpha_0 + 1/c_2) + \xi_2$. This leads to

$$\begin{aligned}
\mathbb{E}\left[V(t)\right] \leq &\left(\prod_{s=T_0}^{t-1}(1 - c_2\alpha(s))\right)\mathbb{E}\left[V(T_0)\right] \\
&+ \xi_8 \sum_{s=T_0}^{t-1} \alpha^2(s) \left[\prod_{\ell=s+1}^{t-1}(1 - c_2\alpha(\ell))\right].
\end{aligned}$$

Simplifying the latter expression using Lemma 1 we arrive at the claim of the lemma. ■

Next, we present a lemma that bounds the gap between $f(\bar{\mathbf{x}})$ and $f(\mathbf{x}^\star)$ for a convex function.

Proposition 3: Let Assumptions 1, 2, 3-(a), and 3-(c) hold. Then, assuming the convexity of local cost functions $f_i(\cdot)$, $\tau \geq \tau_0$, and for a constant $\xi_9 > 0$, the dynamics (3) satisfies

$$\sum_{t=1}^T \alpha(t)\left(\mathbb{E}\left[f(\bar{\mathbf{x}}(t))\right] - f(\mathbf{x}^\star)\right) \leq \frac{1}{2}\|\mathbf{x}^\star\|^2 + \xi_9 \sum_{t=1}^T \alpha^2(t).$$

Proof Sketch of Proposition 3: Similar to the proof of Proposition 2, we can arrive at (16). However, the inequality in (18) does not hold when the function is only convex. Alternatively, we can use the convexity of the function, and write

$$\begin{aligned}
&\sum_{i=1}^n r_i \langle \nabla f_i(\mathbf{x}_i(t)), \mathbf{x}_i(t) - \mathbf{x}^\star \rangle \geq \sum_{i=1}^n r_i \left(f_i(\mathbf{x}_i(t)) - f_i(\mathbf{x}^\star)\right) \\
&= \sum_{i=1}^n r_i \left(f_i(\mathbf{x}_i(t)) - f_i(\bar{\mathbf{x}}(t))\right) + f(\bar{\mathbf{x}}(t)) - f(\mathbf{x}^\star) \\
&\geq \sum_{i=1}^n r_i \langle \nabla f_i(\bar{\mathbf{x}}(t)), \mathbf{x}_i(t) - \bar{\mathbf{x}}(t) \rangle + f(\bar{\mathbf{x}}(t)) - f(\mathbf{x}^\star).
\end{aligned}$$

Then, using the Cauchy-Schwarz inequality, we have

$$\begin{aligned}
&\langle g(t), \bar{\mathbf{x}}(t) - \mathbf{x}^\star \rangle = \left\langle \mathbf{r}^T \nabla f(X(t)), \bar{\mathbf{x}}(t) - \mathbf{x}^\star \right\rangle \\
&= \sum_{i=1}^n r_i [\langle \nabla f_i(\mathbf{x}_i(t)), \bar{\mathbf{x}}(t) - \mathbf{x}_i(t) \rangle + \langle \nabla f_i(\mathbf{x}_i(t)), \mathbf{x}_i(t) - \mathbf{x}^\star \rangle] \\
&\geq \sum_{i=1}^n r_i \langle \nabla f_i(\mathbf{x}_i(t)) - \nabla f_i(\bar{\mathbf{x}}(t)), \bar{\mathbf{x}}(t) - \mathbf{x}_i(t) \rangle \\
&\quad + f(\bar{\mathbf{x}}(t)) - f(\mathbf{x}^\star) \\
&\geq -\sum_{i=1}^n r_i \|\nabla f_i(\mathbf{x}_i(t)) - f_i(\bar{\mathbf{x}}(t))\| \|\mathbf{x}_i(t) - \bar{\mathbf{x}}(t)\| \\
&\quad + f(\bar{\mathbf{x}}(t)) - f(\mathbf{x}^\star) \\
&\geq -2K^{\frac{1}{2}} \sum_{i=1}^n r_i \|\mathbf{x}_i(t) - \bar{\mathbf{x}}(t)\| + f(\bar{\mathbf{x}}(t)) - f(\mathbf{x}^\star). \quad (25)
\end{aligned}$$

Moreover, from Assumption 3-(c) we get

$$\|g(t)\|^2 = \left\|\sum_{i=1}^n r_i \nabla f_i(\mathbf{x}_i(t))\right\|^2 \leq \sum_{i=1}^n r_i \|\nabla f_i(\mathbf{x}_i(t))\|^2 \leq K. \quad (26)$$

Continuing from (16), and using (23), (25) and (26), we have

$$\begin{aligned}
\mathbb{E}[V(t+1)\mathcal{F}_t] &= \|\bar{\mathbf{x}}(t) - \mathbf{x}^\star\|^2 - \alpha(t)\left\langle g(t), \bar{\mathbf{x}}(t) - \mathbf{x}^\star \right\rangle \\
&\quad + \alpha^2(t)\|g(t)\|^2 + \mathbb{E}\left[\|\mathbf{r}^T E(t)\|^2 \big| \mathcal{F}_t\right] \\
&\leq \|\bar{\mathbf{x}}(t) - \mathbf{x}^\star\|^2 + 2K^{\frac{1}{2}}\alpha(t) \sum_{i=1}^n r_i \|\mathbf{x}_i(t) - \bar{\mathbf{x}}(t)\| \\
&\quad - \alpha(t)\left(f(\bar{\mathbf{x}}(t)) - f(\mathbf{x}^\star)\right) + \alpha^2(t)K + \xi_2\alpha^2(t). \quad (27)
\end{aligned}$$

Using Jensen's inequality, we can show that

$$\mathbb{E}\left[\sum_{i=1}^n r_i \|\mathbf{x}_i(t) - \bar{\mathbf{x}}(t)\|\right] \leq \left(\mathbb{E}\left[\|X(t) - \mathbf{1}\bar{\mathbf{x}}(t)\|_{\mathbf{r}}^2\right]\right)^{\frac{1}{2}} = \phi^{\frac{1}{2}}(t),$$

which is upper bounded by $\xi_1^{\frac{1}{2}}\alpha(t)$. Hence, taking expectations from both sides of (27), we arrive at

$$\begin{aligned}
\mathbb{E}[V(t+1)] \leq &\mathbb{E}[V(t)] + \xi_9\alpha^2(t) \\
&- \alpha(t)\left(\mathbb{E}\left[f(\bar{\mathbf{x}}(t))\right] - f(\mathbf{x}^\star)\right), \quad (28)
\end{aligned}$$

where $\xi_9 := 2K^{\frac{1}{2}}\xi_1^{\frac{1}{2}} + K + \xi_2$. Note that unlike (24) where we have a multiplicative contraction, the recursive relationship in (28) can only possibly show an additive reduction. However, a telescopic summation of (28) over $t \in [T]$ and noting that $V(1) = \mathbf{x}^\star$ lead to the proposition's claim. ■

Proof Sketch of Theorem 1: Combining Proposition 1 and Proposition 2 using the triangle inequality, and the fact that $\nu \leq 1$ we readily arrive at the theorem's claim. ■

Proof Sketch of Theorem 2: First note that Proposition 1 guarantees that $\mathbb{E}\left[\|\mathbf{x}_i(t) - \bar{\mathbf{x}}(t)\|^2\right] \leq \mathbf{r}_{\min}^{-1}\xi_1\alpha^2(t)$. This together with the fact that f is a $\sqrt{K}$-Lipschitz function, implies that

$$\begin{aligned}\mathbb{E}\left[f(\mathbf{x}_i(t))\right] - \mathbb{E}\left[f(\bar{\mathbf{x}}(t))\right] &\leq \sqrt{K}\mathbb{E}\left[\|\mathbf{x}_i(t) - \bar{\mathbf{x}}(t)\|\right] \\ &\leq (K\mathbf{r}_{\min}^{-1}\xi_1)^{\frac{1}{2}}\alpha(t). \qquad (29)\end{aligned}$$

Multiplying both sides of (29) by $\alpha(t)$, summing up over t, and using Proposition 3, we arrive at

$$\sum_{t=1}^{T}\alpha(t)(\mathbb{E}\left[f(\mathbf{x}_i(t))\right] - f(\mathbf{x}^\star)) \leq \max\left(\|\mathbf{x}^\star\|^2, \xi_{10}\sum_{t=1}^{T}\alpha^2(t)\right),$$

where $\xi_{10} := \xi_9 + (K\mathbf{r}_{\min}^{-1}\xi_1)^{\frac{1}{2}}$. Finally, we use Hölder's inequality [20, Theorem 6.2],

$$\sum\nolimits_{t=1}^{T} a_t b_t \leq \left(\sum\nolimits_{t=1}^{T} a_t^p\right)^{\frac{1}{p}}\left(\sum\nolimits_{t=1}^{T} b_t^q\right)^{\frac{1}{q}},$$

with $a_t := (1/\alpha(t))^\theta$, $b_t := \alpha(t)(\mathbb{E}\left[f(\mathbf{x}_i(t))\right] - f(\mathbf{x}^\star))^\theta$, and $(p, q) = \left(\frac{1}{1-\theta}, \frac{1}{\theta}\right)$ to get the claim of the theorem. ■

Proof Sketch of Corollary 1: The claim of the corollary is a consequence of Theorem 2 and the convexity of the loss function, which yields to $f\left(\frac{1}{T}\sum_{t=1}^{T}\mathbf{x}_i(t)\right) \leq \frac{1}{T}\sum_{t=1}^{T} f(\mathbf{x}_i(t))$. It is worth noting that Theorem 2 cannot be directly applied for $\theta = 1$. We refer to [18] to address this subtle point. ■

IV. Experimental Results

Here, we provide a numerical experiment supporting our theoretical results. We apply PCom-DSGD (this work) to a linear regression problem over a time-varying network, and compare its performance against DSGD with perfect information sharing [5] and DIMIX [14] algorithms. For the sake of comparison, we are also providing the theoretical upper bounds for these algorithms. Due to the page limit, we refer to the extended version [18] for the details of the experiment as well as more experimental results.

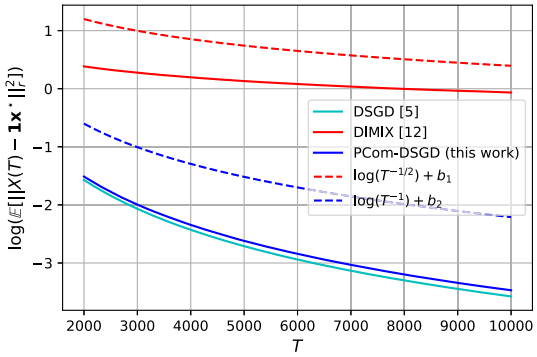

Fig. 1: Expected state distance to the optimal point vs. iterations: linear regression over a time-varying network

References

[1] K. I. Tsianos, S. Lawlor, and M. G. Rabbat, "Consensus-based distributed optimization: Practical issues and applications in large-scale machine learning," in Allerton, pp. 1543–1550, IEEE, 2012.

[2] S. S. Ram, V. V. Veeravalli, and A. Nedic, "Distributed non-autonomous power control through distributed convex optimization," in IEEE INFOCOM 2009, pp. 3001–3005, IEEE, 2009.

[3] M. Rabbat and R. Nowak, "Distributed optimization in sensor networks," in Int. Symp. on Inform. Proc. in sensor networks, pp. 20–27, 2004.

[4] S. Kar and J. M. Moura, "Distributed consensus algorithms in sensor networks with imperfect communication: Link failures and channel noise," IEEE Trans. Signal Process., vol. 57, no. 1, pp. 355–369, 2008.

[5] A. Nedic and A. Ozdaglar, "Distributed subgradient methods for multi-agent optimization," IEEE Trans. Automat. Contr., vol. 54, no. 1, pp. 48–61, 2009.

[6] A. Nedic, A. Ozdaglar, and P. A. Parrilo, "Constrained consensus and optimization in multi-agent networks," IEEE Trans. Automat. Contr., vol. 55, no. 4, pp. 922–938, 2010.

[7] K. Yuan, Q. Ling, and W. Yin, "On the convergence of decentralized gradient descent," SIAM Journal on Optimization, vol. 26, no. 3, pp. 1835–1854, 2016.

[8] T. Tatarenko and B. Touri, "Non-convex distributed optimization," IEEE Trans. Automat. Contr., vol. 62, no. 8, pp. 3744–3757, 2017.

[9] J. Zeng and W. Yin, "On nonconvex decentralized gradient descent," IEEE Trans. Signal Process., vol. 66, no. 11, pp. 2834–2848, 2018.

[10] A. Koloskova, S. Stich, and M. Jaggi, "Decentralized stochastic optimization and gossip algorithms with compressed communication," in ICML, pp. 3478–3487, PMLR, 2019.

[11] A. Reisizadeh, H. Taheri, A. Mokhtari, H. Hassani, and R. Pedarsani, "Robust and communication-efficient collaborative learning," in NeurIPS, pp. 8388–8399, 2019.

[12] H. Taheri, A. Mokhtari, H. Hassani, and R. Pedarsani, "Quantized decentralized stochastic learning over directed graphs," in ICML, pp. 9324–9333, PMLR, 2020.

[13] H. Reisizadeh, B. Touri, and S. Mohajer, "Adaptive bit allocation for communication-efficient distributed optimization," in 2021 60th IEEE Conference on Decision and Control (CDC), pp. 1994–2001, IEEE, 2021.

[14] H. Reisizadeh, B. Touri, and S. Mohajer, "Distributed optimization over time-varying graphs with imperfect sharing of information," arXiv preprint arXiv:2106.08469, 2021.

[15] H. Reisizadeh, B. Touri, and S. Mohajer, "Dimix: Diminishing mixing for sloppy agents." submitted, 2022.

[16] M. M. Vasconcelos, T. T. Doan, and U. Mitra, "Improved convergence rate for a distributed two-time-scale gradient method under random quantization," arXiv preprint arXiv:2105.14089, 2021.

[17] D. Alistarh, D. Grubic, J. Li, R. Tomioka, and M. Vojnovic, "Qsgd: Communication-efficient sgd via gradient quantization and encoding," in NeurIPS, pp. 1709–1720, 2017.

[18] H. Reisizadeh, B. Touri, and S. Mohajer, "Distributed optimization over time-varying networks: Imperfect information with feedback is as good as perfect information." https://tinyurl.com/yc259y9z.

[19] A. Agarwal, M. J. Wainwright, P. Bartlett, and P. Ravikumar, "Information-theoretic lower bounds on the oracle complexity of convex optimization," in NeurIPS, vol. 22, pp. 1–9, 2009.

[20] G. B. Folland, Real analysis: modern techniques and their applications, vol. 40. John Wiley & Sons, 1999.