SPARQ-SGD: Event-Triggered and Compressed Communication in Decentralized Optimization

Navjot Singh, *Student Member, IEEE*, Deepesh Data, Jemin George, *Member, IEEE*, and Suhas Diggavi, *Fellow, IEEE*

***Abstract*—In this article, we propose and analyze SParsified Action Regulated Quantized–Stochastic Gradient Descent (SPARQ-SGD), a communication-efficient algorithm for decentralized training of large-scale machine learning models over a graph with n nodes, where communication efficiency is achieved using *compressed* exchange of local model parameters among neighboring nodes, which is *triggered* only when an *event* (a locally computable condition) is satisfied. Specifically, in SPARQ-SGD, each node takes a fixed number of local gradient steps and then checks if the model parameters have significantly changed compared to its last update; only when the change is beyond a certain threshold (specified by a design criterion), it compresses its local model parameters using both quantization and sparsification and communicates them to its neighbors. We prove that SPARQ-SGD converges as $O(\frac{1}{nT})$ and $O(\frac{1}{\sqrt{nT}})$ in the strongly convex and nonconvex settings, respectively, matching the convergence rates of plain decentralized SGD. This demonstrates that we get communication efficiency achieved by aggressive compression, local iterations, and event-triggered communication essentially for *free*. We evaluate SPARQ-SGD over real datasets to demonstrate significant amount of savings in communication over the state-of-the-art while achieving similar performance.**

***Index Terms*—Consensus, convergence, decentralized algorithms, efficient-communication, event-triggered, multiagent systems, stochastic optimization.**

I. Introduction

THERE has been a recent interest in communication-efficient *decentralized* training of large-scale machine learning models [1]–[3]. In decentralized training, the nodes do not have a central coordinator, and are not directly connected to all other nodes, but are connected through a communication graph. This implies that the communication is inherently more efficient, as the local connection (degree) of such graphs could be a small constant, independent of the network size. In this article, we propose SPARQ-SGD,[1] an optimization algorithm to improve communication-efficiency in decentralized training through *event-driven* exchange of *quantized* and *sparsified* model parameters between the nodes.

Over the past few years, a number of different methods have been developed to achieve communication efficiency in *distributed* SGD, where there exists a central coordinator. These can be broadly divided into two categories. In the first category, to reduce communication, workers send *compressed* updates either with sparsification [4]–[8] or quantization [9]–[12] or a combination of both [13].[2] In another class of algorithms that are based on the idea of *infrequent communication*, workers do not communicate in each iteration; rather, they send the updates after performing a *fixed* number of local gradient steps [13]–[16]. The idea of compressed communication, using quantization or sparsification, has been extended to the setting of *decentralized* optimization [2], [3], [17].

In this article, we propose SPARQ-SGD, which combines compression with *event-triggered* communication, where a node initiates a (compressed communication) action regulated by a locally computable triggering condition (event), thereby further reducing the communication among nodes. In particular, the proposed triggering condition is such that at least a fixed number of local gradient steps or iterations (say, H local iterations) are first completed and after that the condition checks if there is a significant change (beyond a certain threshold) in its local model parameter vector since the last time communication occurred. Only if the change in model parameter exceeds the prescribed threshold, does a node trigger compressed communication. As far as we know, such an idea of event-triggered and compressed communication has not been proposed and analyzed in the context of decentralized (stochastic) training of large-scale machine learning models.

Manuscript received 4 February 2021; revised 11 February 2021, 26 August 2021, and 30 August 2021; accepted 8 January 2022. Date of publication 25 January 2022; date of current version 30 January 2023. This work was supported in part by NSF under Grant #2007714 and under Grant #1955632, in part by UC-NL under Grant LFR-18-548554, and in part by Army Research Laboratory under Cooperative Agreement W911NF-17-2-0196. Recommended by Associate Editor J. Lavaei. *(Corresponding author: Navjot Singh.)*

Navjot Singh, Deepesh Data, and Suhas Diggavi are with the Department of Electrical and Computer Engineering, University of California, Los Angeles, CA 90095 USA (e-mail: navjotsingh@ucla.edu; deepesh.data@gmail.com; suhas@ee.ucla.edu).

Jemin George is with the US Army Research Lab, Adelphi, MD 20783 USA (e-mail: jemin.george.civ@mail.mil).

Color versions of one or more figures in this article are available at https://doi.org/10.1109/TAC.2022.3145576.

Digital Object Identifier 10.1109/TAC.2022.3145576

[1] Acronym stands for SParsified Action Regulated Quantized–Stochastic Gradient Descent.

[2] In sparsification, the vector sparsification is done by selecting either its top k entries (in terms of the absolute value) or random k entries, where k is less than the dimension of the vector. Quantization consists of discretization of the vector by rounding off its entries either randomly or deterministically (in the extreme case, this can be just the sign operator).

As mentioned earlier, in addition to event-triggered communication, we also incorporate compression of the model parameters: When a node communicates with its neighbors, it sends a quantized and sparsified version of its local model parameters. We therefore combine the recent ideas applied to communication efficient training (quantization and sparsification) with our event-triggered communication to propose SPARQ-SGD;[3] see Algorithm 1. We analyze the performance of our algorithm for both convex and (smooth) nonconvex objective functions, in terms of its convergence rate as a function of the number of iterations T (and also the number of communication rounds) and the amount of communication bits exchanged to learn a model to a certain accuracy. We prove that SPARQ-SGD converges with rates $O(\frac{1}{nT})$ and $O(\frac{1}{\sqrt{nT}})$ in strongly convex and nonconvex settings, respectively, demonstrating that such aggressive compression, including event-triggered communication does not affect the overall convergence rate as compared to a uncompressed decentralized training [1]. Moreover, we show that SPARQ-SGD yields significant amount of saving in communication over the state-of-the-art; see Section VII for more details.

Related Work: In decentralized setting, Tang *et al.* [2] and Reisizadeh *et al.* [18] propose unbiased stochastic compression for gradient exchange. Assran *et al.* [19] and Tatarenko and Touri[20] analyze Stochastic Gradient Push algorithm for nonconvex objectives, which approximates distributed averaging instead of compressing the gradients. Our work most closely relates to [3], which proposed CHOCO-SGD that uses compressed (sparsified *or* quantized) updates; the distinction is that we propose an *event-triggered* communication where sparsified *and* quantized model parameters are transmitted only if they have changed significantly after performing some fixed number of local iterations, further reducing communication. The idea of event-triggered communication has been explored previously in the control community [21]–[25] and in optimization literature [26]–[32]. These papers focus on continuous-time, deterministic optimization algorithms for convex problems; in contrast, we propose event-driven, *compressed, stochastic* gradient descent algorithms for both convex and nonconvex problems. Chen *et al.*[33] propose an adaptive scheme to skip gradient computations in a *distributed* setting for *deterministic* gradients; moreover, their focus is on saving communication rounds, and do not have any compressed communication. Sub-gradient descent with quantization for *deterministic* decentralized optimization has been studied in [34] and [35] for convex objectives only, with the former showing convergence only within a neighborhood of the optimum and the latter employing an adaptive quantization scheme to recover rates attained by un-quantized schemes. Decentralized consensus with quantization over time varying topology has been analyzed in [36]. Pu *et al.* [37] consider inexact proximal gradient with quantization in decentralized optimization for strongly convex objectives, showing convergence to the global optimum. As far as we know, ours is the first article which uses event-triggered (incorporating infrequent communication) and compressed communication for decentralized *stochastic* optimization of both strongly convex and nonconvex objectives.

Contributions: We study optimization in a decentralized setup, where n different workers, each having a different dataset $\mathcal{D}_i$ (with an associated local objective function $f_i : \mathbb{R}^d \to \mathbb{R}$), are linked through a connected graph $\mathcal{G} = ([n], \mathcal{E})$, where $[n] := \{1, 2, \ldots, n\}$. Vertex i in $\mathcal{G}$ corresponds to the ith worker who can only communicate with its neighbors $\mathcal{N}_i = \{j \in [n] : \{i, j\} \in \mathcal{E}\}$. We consider the loss function

$$f(\mathbf{x}) = \frac{1}{n} \sum_{i=1}^{n} f_i(\mathbf{x}) \tag{1}$$

where $f_i(\mathbf{x}) = \mathbb{E}_{\xi_i \sim \mathcal{D}_i}[F_i(\mathbf{x}, \xi_i)]$, where $\xi_i \sim \mathcal{D}_i$ denotes a random data sample from $\mathcal{D}_i$ and $F_i(\mathbf{x}, \xi_i)$ denotes the risk associated with the data sample ξ_i w.r.t. $\mathbf{x}$ at the ith worker node. We solve the decentralized optimization in (1) using SPARQ-SGD. Our theoretical results are the convergence analyses for both strongly convex and nonconvex objectives in the synchronous setting; see Theorems 1 and 2, respectively. In the strongly convex setting, we show a convergence rate of $\mathcal{O}(\frac{1}{nT}) + \mathcal{O}(\frac{c_0}{\delta^2 T^{(1+\epsilon)}}) + \mathcal{O}(\frac{H^2}{\delta^4 \omega^2 T^2}) + \mathcal{O}(\frac{H^3}{\omega^3 \delta^6 T^3})$ for some $\epsilon \in (0, 1)$. Here c_0 is for triggering threshold, H is for number of local iterations, ω quantifies compression, and δ is the spectral-gap of the connectivity matrix of the graph $\mathcal{G}$. Note that all these factors for communication-efficiency appear in the higher order terms. Thus, for large enough T, they do not affect the dominating term $\mathcal{O}(\frac{1}{nT})$, which, in fact, is the convergence rate of centralized vanilla SGD with mini-batch size of n. Similar observation is also made in the nonconvex setting, where we get a convergence rate of $\mathcal{O}(\frac{1}{\sqrt{nT}})$; see Corollaries 1, 2, and the following remarks for more details. Hence, for both the objectives, we get essentially the same convergence rate as that of vanilla SGD, even after applying SPARQ-SGD to gain communication efficiency; and hence, we get communication efficiency essentially "for free". We compare our algorithm against CHOCO-SGD [17], which is the state-of-the-art in compressed decentralized training and provide theoretical justification for communication efficiency of SPARQ-SGD over CHOCO-SGD to achieve the same target accuracy. We corroborate our theoretical understanding with numerical results in Section VII where we demonstrate that SPARQ-SGD yields significant savings in communication bits. For a convex objective simulated on the MNIST dataset, SPARQ-SGD saves total communicated bits by a factor of $40\times$ compared to CHOCO-SGD [3] and by $1000\times$ compared to vanilla SGD to converge to the same target accuracy. Similarly, for a nonconvex objective simulated on the CIFAR-10 dataset [38], we save total bits by a factor of $40\times$ compared to CHOCO-SGD [17] and around $3\text{K}\times$ compared to vanilla SGD to reach the same target accuracy. We further provide experiments comparing total training time when communicating over bandlimited links, showing a speedup by a factor of $60\times$ and $700\times$ compared to CHOCO-SGD [3] and vanilla SGD, respectively, to achieve the same target accuracy.

[3]The idea of combining compression and *fixed* number of local iterations has been carried out in a *distributed* setting (the master-worker architecture) in [13]. In this work, in addition to *extending* this combination to the *decentralized* setting, we also propose and analyze event-triggered communication.

Article Organization: We describe SPARQ-SGD, our proposed algorithm, in Section II. Our main results are stated in Section III for strongly convex and nonconvex objectives and we prove them in Section V. Intermediate lemmas used for proving these theorems concerning consensus among nodes are proved in Section VI. We validate our theoretical findings with numerical experiments in Section VII. Omitted proof details and additional experiments are provided in our extended article [39].

II. Our Algorithm: SPARQ-SGD

In this section, we describe SPARQ-SGD, our decentralized SGD algorithm with compression and event-triggered communication. First, we need to define its main ingredients.

Definition 1 (Compression, [7]): A (possibly randomized) function $\mathcal{C} : \mathbb{R}^d \to \mathbb{R}^d$ is called a *compression* operator, if there exists a positive constant $\omega < 1$, such that the following holds for every $\mathbf{x} \in \mathbb{R}^d$:

$$\mathbb{E}_{\mathcal{C}}[\|\mathbf{x} - \mathcal{C}(\mathbf{x})\|_2^2] \leq (1-\omega)\|\mathbf{x}\|_2^2 \tag{2}$$

where expectation is taken over the randomness of $\mathcal{C}$. We assume $\mathcal{C}(\mathbf{0}) = \mathbf{0}$.

It is known that some important sparsifiers as well as quantizers are examples of compression operators: 1) Top_k and $Rand_k$ sparsifiers (in which we select k entries; see Footnote 2) with $\omega = k/d$ [7], 2) stochastic quantizer Q_s from [9][4] with $\omega = (1-\beta_{d,s})$ for $\beta_{d,s} < 1$, and 3) deterministic quantizer $\frac{\|\mathbf{x}\|_1}{d} Sign(\mathbf{x})$ from [12] with $\omega = \frac{\|\mathbf{x}\|_1^2}{d\|\mathbf{x}\|_2^2}$. It was shown in [13] that if we compose these sparsifiers and quantizers, the resulting operator also gives compression and outperforms their individual components. For example, for any $Comp_k \in \{Top_k, Rand_k\}$, the following are compression operators: (iv) $\frac{1}{(1+\beta_{k,s})} Q_s(Comp_k)$ with $\omega = (1 - \frac{k}{d(1+\beta_{k,s})})$ for any $\beta_{k,s} \geq 0$, and (v) $\frac{\|Comp_k(\mathbf{x})\|_1 SignComp_k(\mathbf{x})}{k}$ with $\omega = \max\{\frac{1}{d}, \frac{k}{d}(\frac{\|Comp_k(\mathbf{x})\|_1^2}{d\|Comp_k(\mathbf{x})\|_2^2})\}$.

Event-Triggered Communication: As mentioned in Section I, our proposed event-triggered communication consists of two phases: In the first phase, nodes perform a fixed number H of local iterations, and in the second phase, they check for the communication-triggering condition (event), if satisfied, then they send the (compressed) updates to their neighbors. Let $\mathcal{I}_T \subseteq [T]$ denote a set of indices at which workers check for the triggering condition. Since we are in the synchronous setting, we assume that $\mathcal{I}_T$ is same for all workers. Let $\mathcal{I}_T = \{I_{(1)}, I_{(2)}, \ldots, I_{(k)}, \ldots\}$. The gap of $\mathcal{I}_T$ is defined as $gap(\mathcal{I}_T) := \max_{i \in [k-1]}\{(I_{(i+1)} - I_{(i)})\}$, [14], which is equal to the maximum number of local iterations a worker performs before checking for the triggering condition. Note that $gap(\mathcal{I}_T) = 1$ is equivalent to the case when workers check for the communication triggering criterion in every iteration.

Our algorithm, SPARQ-SGD, for optimizing (1) in a decentralized setting is presented in Algorithm 1. For designing this, in addition to combining sparsification *and* quantization, we carefully incorporate local iterations and event-triggered[5] communication into the CHOCO-SGD algorithm from [3], which uses only sparsified *or* quantized updates. This poses several technical challenges in proving the convergence; see the proofs of Theorems 1, 2, and in particular, the proof of Lemma 2. The resulting algorithm is presented in Algorithm 1.

In SPARQ-SGD, each node $i \in [n]$ maintains a local parameter vector $\mathbf{x}_i^{(t)}$, and their goal is to achieve consensus among themselves on the value of $\mathbf{x}$ that minimizes (1), while allowing only for compressed and infrequent communication. Node i updates $\mathbf{x}_i^{(t)}$ in each iteration t by a stochastic gradient step (line 4). An estimate $\hat{\mathbf{x}}_i^{(t)}$ of $\mathbf{x}_i^{(t)}$ is also maintained at each neighbor $j \in \mathcal{N}_i$ and at i itself. Thus, each node maintains an estimate of all its neighbors' local parameter vectors and of itself. In our algorithm, $\mathcal{I}_T$ is the set of indices for which the workers check for the triggering condition and take a consensus step. We also allow the triggering threshold c_t to vary with t with the requirement that c_t is $o(t)$. At time-step t, if $(t+1) \in \mathcal{I}_T$, the nodes check for the triggering condition (line 6), if satisfied, then each node $i \in [n]$ sends to all its neighbors the compressed difference between its local parameter vector and its estimate that its neighbors have (line 12); and based on the messages received from its neighbors, the ith node updates $\hat{\mathbf{x}}_j^{(t)}$—the estimate of the jth node's local parameter vector (line 13), and then every node performs the consensus step (line 15).

III. Main Results

Our main results are under the following assumptions.

Assumptions: 1) *L-Smoothness:* Each local function f_i for $i \in [n]$ is L-smooth, i.e, $\forall \mathbf{x}, \mathbf{y} \in \mathbb{R}^d$, we have $f_i(\mathbf{y}) \leq f_i(\mathbf{x}) + \langle \nabla f_i(\mathbf{x}), \mathbf{y} - \mathbf{x}\rangle + \frac{L}{2}\|\mathbf{y} - \mathbf{x}\|^2$. 2) *Bounded variance:* For every $i \in [n]$, we have $\mathbb{E}_{\xi_i}\|\nabla F_i(\mathbf{x}, \xi_i) - \nabla f_i(\mathbf{x})\|^2 \leq \sigma_i^2$, for some finite σ_i, where $\nabla F_i(\mathbf{x}, \xi_i)$ is the unbiased gradient at worker i such that $\mathbb{E}_{\xi_i}[\nabla F_i(\mathbf{x}, \xi_i)] = \nabla f_i(\mathbf{x})$. We define the average variance across all workers as $\bar{\sigma}^2 := \frac{1}{n}\sum_{i=1}^n \sigma_i^2$. 3) *Bounded second moment*[6]*:* For every $i \in [n]$, we have $\mathbb{E}_{\xi_i}\|\nabla F_i(\mathbf{x}, \xi_i)\|^2 \leq G^2$, for some finite G.

Before stating the main results, we need some notations about the underlying communication graph $\mathcal{G}$ first. Let $W \in \mathbb{R}^{n \times n}$ denote the weighted connectivity matrix of $\mathcal{G}$, with w_{ij} for every $i, j \in [n]$ being its (i,j)th entry, which denotes the weight on the link between worker i and j. W is assumed to be symmetric and doubly stochastic, which implies that all its eigenvalues $\lambda_i(W), i = 1, 2, \ldots, n$, lie in $[-1, 1]$. Without loss of generality, assume that $|\lambda_1(W)| > |\lambda_2(W)| \geq \ldots \geq |\lambda_n(W)|$. Since W is doubly stochastic, we have $\lambda_1(W) = 1$, and since $\mathcal{G}$ is connected, we have $\lambda_2(W) < \lambda_1(W)$. Let the spectral gap of W be defined as $\delta := 1 - |\lambda_2(W)|$. Since $|\lambda_2(W)| \in [0, 1)$, we have that $\delta \in (0, 1]$. Simple matrices W with $\delta > 0$ are known to exist for every connected graph [3].

[4] $Q_s : \mathbb{R}^d \to \mathbb{R}^d$ is a stochastic quantizer, if for every $\mathbf{x} \in \mathbb{R}^d$, we have (i) $\mathbb{E}[Q_s(\mathbf{x})] = \mathbf{x}$ and (ii) $\mathbb{E}[\|\mathbf{x} - Q_s(\mathbf{x})\|_2^2] \leq \beta_{d,s}\|\mathbf{x}\|_2^2$. Q_s from [9] satisfies this definition with $\beta_{d,s} = \min\{\frac{d}{s^2}, \frac{\sqrt{d}}{s}\}$.

[5] The Zeno phenomenon [21] does not occur in our setup as we have a discrete sampling period as well as a fixed number of local iterations, giving a lower bound to the event intervals of at least H times the sampling period.

[6] Bounded second moment is a standard assumption in stochastic optimization with *compressed* communication [7], [8].

Algorithm 1: SPARQ-SGD: SParsified Action Regulated Quantized SGD.

Initial values $\mathbf{x}_i^{(0)} \in \mathbb{R}^d$ on each node $i \in [n]$, consensus stepsize γ, SGD stepsizes $\{\eta_t\}_{t\geq 0}$, threshold sequence $\{c_t\}_{t\geq 0}$, compression operator $\mathcal{C}$ having parameter ω, communication graph $G = ([n], E)$ and mixing matrix W, set of synchronization indices $\mathcal{I}_T$, initialize $\hat{\mathbf{x}}_i^{(0)} := \mathbf{0}$ for all $i \in [n]$
for $t = 0$ **to** $T-1$ in parallel for all workers $i \in [n]$ **do**
  Sample $\xi_i^{(t)}$, stochastic gradient $\mathbf{g}_i^{(t)} := \nabla F_i(\mathbf{x}_i^{(t)}, \xi_i^{(t)})$
  $\mathbf{y}_i^{(t+1)} := \mathbf{x}_i^{(t)} - \eta_t \mathbf{g}_i^{(t)}$
  If $(t+1) \in \mathcal{I}_T$ **then**
    If $\|\mathbf{y}_i^{(t+1)} - \hat{\mathbf{x}}_i^{(t)}\|_2^2 > c_t \eta_t^2$ **then**
      Set $\mathbf{q}_i^{(t)} := \mathcal{C}(\mathbf{y}_i^{(t+1)} - \hat{\mathbf{x}}_i^{(t)})$
    else
      Set $\mathbf{q}_i^{(t)} := \mathbf{0}$
    end if
    for neighbors $j \in \mathcal{N}_i \cup i$ **do**
      Send $\mathbf{q}_i^{(t)}$ and receive $\mathbf{q}_j^{(t)}$
      $\hat{\mathbf{x}}_j^{(t+1)} := \mathbf{q}_j^{(t)} + \hat{\mathbf{x}}_j^{(t)}$
    end for
    $\mathbf{x}_i^{(t+1)} = \mathbf{y}_i^{(t+1)} + \gamma \sum_{j\in\mathcal{N}_i} w_{ij}(\hat{\mathbf{x}}_j^{(t+1)} - \hat{\mathbf{x}}_i^{(t+1)})$
  else
    $\hat{\mathbf{x}}_i^{(t+1)} = \hat{\mathbf{x}}_i^{(t)}$ and $\mathbf{x}_i^{(t+1)} = \mathbf{y}_i^{(t+1)}$
  end if
end for

Now we state the main results of this article both for strongly convex and nonconvex objectives. As mentioned in Section I, even after using compression and infrequent communication, we prove a convergence rate matching with that of vanilla SGD in both strongly convex and nonconvex settings.

Theorem 1 (Smooth and strongly convex objective with decaying learning rate): Suppose f_i, for all $i \in [n]$ is L-smooth and μ-strongly convex. Let $\mathcal{C}$ be a compression operator with parameter equal to $\omega \in (0, 1]$. Let $gap(\mathcal{I}_T) \leq H$. If we run SPARQ-SGD with consensus step-size $\gamma = \frac{2\delta\omega}{64\delta+\delta^2+16\beta^2+8\delta\beta^2-16\delta\omega}$ (where $\beta = \max_i\{1-\lambda_i(W)\}$), an increasing threshold function $c_t \leq c_0 t^{1-\epsilon}$ for all t where constant $c_0 \geq 0$ and $\epsilon \in (0, 1)$ and decaying learning rate $\eta_t = \frac{8}{\mu(a+t)}$, where $a \geq \max\{\frac{5H}{p}, \frac{32L}{\mu}\}$ for $p = \frac{\gamma\delta}{8}$, and let the algorithm generate $\{\mathbf{x}_i^{(t)}\}_{t=0}^{T-1}$ for $i \in [n]$, then the following holds:

$$\mathbb{E}f(\mathbf{x}_{avg}^{(T)}) - f^* \leq \frac{\mu a^3}{8S_T}\|\mathbf{x}^{(0)} - \mathbf{x}^*\|^2 + \frac{4T(T+2a)}{\mu S_T}\frac{\bar{\sigma}^2}{n} + \frac{Z_1 T G^2 H^2}{\mu^2 S_T p^2}(2L+\mu) + \frac{Z_2 c_0 \omega T^{(2-\epsilon)}}{\mu^2(2-\epsilon)S_T}\left(\frac{2L+\mu}{p}\right)$$

where $\bar{\mathbf{x}}_{avg}^{(T)} = \frac{1}{S_T}\sum_{t=0}^{T-1} w_t \bar{\mathbf{x}}^{(t)}$, where $\bar{\mathbf{x}}^{(t)} = \frac{1}{n}\sum_{i=1}^n \mathbf{x}_i^{(t)}$, weights $w_t = (a+t)^2$, $S_T = \sum_{t=0}^{T-1} w_t \geq \frac{1}{3}T^3$ and Z_1, Z_2 are universal constants.

We provide a proof of Theorem 1 in Section V-A. The analysis provided also works for any $c_t = o(t)$; however, we provide it for $c_t \leq c_0 t^{1-\epsilon}$ to highlight the main idea. Observe that the consensus step-size γ does not appear explicitly in the above rate expression, but it does affect the convergence indirectly through $p = \gamma\delta/8$. Note that $\delta \in (0, 1]$, $\beta \leq 2$, and $\omega \geq 0$. Substituting these in the expression of γ and p gives $\gamma \geq \frac{2\delta\omega}{161}$ and $p \geq \frac{\delta^2\omega}{644}$; see also the proof of Lemma 2. Now we simplify the above expression to gain further insights as to how our techniques for reducing communication is affecting the convergence rate.

Corollary 1: Using $\mathbb{E}\|\mathbf{x}^{(0)} - \mathbf{x}^*\|_2^2 \leq \frac{4G^2}{\mu^2}$ (from [40, Lemma 2]) and $p \geq \frac{\delta^2\omega}{644}$, hiding constants (including L) in the $\mathcal{O}$ notation, the rate expression in Theorem 1 is simplified as

$$\mathbb{E}[f(\bar{\mathbf{x}}_{avg}^{(T)})] - f^* \leq \mathcal{O}\left(\frac{\bar{\sigma}^2}{\mu n T}\right) + \mathcal{O}\left(\frac{c_0}{\mu^2\delta^2 T^{(1+\epsilon)}}\right) + \mathcal{O}\left(\frac{G^2H^2}{\mu^2\delta^4\omega^2 T^2}\right) + \mathcal{O}\left(\frac{G^2H^3}{\mu\omega^3\delta^6 T^3}\right).$$

Remark 1: Observe that the dominating term $\mathcal{O}(\frac{\bar{\sigma}^2}{\mu n T})$ is not affected by the compression factor ω, the number of local iterations H, the factor c_0 in the triggering condition, and the topology of the underlying communication graph (which is controlled by the spectral gap δ)—they all appear in the higher order terms. In order to ensure that they do not affect the dominating term while converging at a rate of $\mathcal{O}(\frac{\bar{\sigma}^2}{\mu n T})$, we would require $T \geq T_0 := C \times \max\{(\frac{nc_0}{\mu\delta^2\bar{\sigma}^2})^{\frac{1}{\epsilon}}, (\frac{nH^2G^2}{\mu\bar{\sigma}^2\delta^4\omega^2})\}$ for sufficiently large constant C. This implies that for large enough T, we get benefits of all these techniques in saving communication bits, without affecting the convergence rate.

Now we analyze the effect of ω, H, c_0, δ on T_0: 1) If we compress the communication more, i.e., smaller ω, then T_0 increases, as expected; 2) if we take more number of local iterations H, T_0 would again increase, as expected, because increasing H means communicating less frequently; 3) if we increase c_0, which means that the triggering threshold has become bigger, we expect less frequent communication, thus T_0 increases, as expected; 4) if the spectral gap $\delta \in (0, 1]$ is closer to 1, which implies that the graph is well-connected, then the threshold T_0 decreases, which is also expected, as good connectivity means faster spreading of information, resulting in faster consensus.[7] Additional experiments to support the arguments made in this remark can be found in the extended version [39].

Remark 2: Observe that after a large enough $T \geq T_0$, we get the same rate as that of distributed vanilla SGD and also a distributed gain of n with the number of nodes. Thus, we essentially converge at the same rate as that of vanilla SGD, while significantly saving in terms of communication bits among all the workers; this can be seen in our numerical results in Section VII.

Now we state our result for nonconvex objectives.

Theorem 2 (Smooth and nonconvex objective with fixed learning rate): Suppose f_i, for all $i \in [n]$ be L-smooth. Let $\mathcal{C}$ be a

[7] If we are to design the underlying communication graph, one possible choice is to consider the *expander graphs* [41], that will simultaneously give low communication and faster convergence, as they have constant degree and large spectral gap [42].

compression operator with parameter equal to $\omega \in (0, 1]$. Let $gap(\mathcal{I}_T) \leq H$. If we run SPARQ-SGD for $T \geq 64nL^2$ iterations with fixed learning rate $\eta = \sqrt{\frac{n}{T}}$, a fixed threshold function c_t such that $c_t < \frac{c_0}{\eta^{1-\epsilon}}$ for all t where $\epsilon \in (0, 1)$, some constant c_0, and consensus step-size $\gamma = \frac{2\delta\omega}{64\delta+\delta^2+16\beta^2+8\delta\beta^2-16\delta\omega}$ (where $\beta = \max_i\{1 - \lambda_i(W)\}$), and let the algorithm generate $\{\mathbf{x}_i^{(t)}\}_{t=0}^{T-1}$ for $i \in [n]$, then the averaged iterates $\bar{\mathbf{x}}^{(t)} := \frac{1}{n}\sum_{i=0}^{n}\mathbf{x}_i^{(t)}$ satisfy

$$\frac{\sum_{t=0}^{T-1}\mathbb{E}\|\nabla f(\bar{\mathbf{x}}^{(t)})\|_2^2}{T} \leq \frac{4\left(f(\bar{\mathbf{x}}_0) - f^* + L\bar{\sigma}^2\right)}{\sqrt{nT}} + \frac{\tilde{Z}_1 G^2H^2L^2n}{Tp\delta^2\omega} + \frac{\tilde{Z}_2 G^2H^2L^3n^{3/2}}{T^{3/2}p\delta^2\omega} + \frac{\tilde{Z}_3 L^2\omega c_0\sqrt{n^{(1+\epsilon)}}}{p\sqrt{T^{(1+\epsilon)}}} + \frac{\tilde{Z}_4 L^3\omega c_0\sqrt{n^{(2+\epsilon)}}}{p\sqrt{T^{(2+\epsilon)}}}$$

Here $p = \frac{\gamma\delta}{8}$ and $\tilde{Z}_1, \tilde{Z}_2, \tilde{Z}_3, \tilde{Z}_4$, are universal constants.

We prove Theorem 2 in Section V-B. As mentioned after Theorem 1, though the consensus step-size γ does not appear in the rate expression, it does affect the convergence through the parameter p. As argued after Theorem 1, we can show similarly show that $p \geq \frac{\delta^2\omega}{644}$. Now we simplify the above expression in the following corollary.

Corollary 2: Let $f(\bar{\mathbf{x}}^{(0)}) - f^* \leq J^2$, where $J^2 < \infty$ is a constant. Using $p \geq \frac{\delta^2\omega}{644}$, and hiding constants (including L) in the $\mathcal{O}$ notation, we can simplify the rate expression in Theorem 2 to the following:

$$\frac{\sum_{t=0}^{T-1}\mathbb{E}\|\nabla f(\bar{\mathbf{x}}^{(t)})\|_2^2}{T} \leq \mathcal{O}\left(\frac{J^2+\bar{\sigma}^2}{\sqrt{nT}}\right) + \mathcal{O}\left(\left(\frac{n}{T} + \frac{n^{3/2}}{T^{3/2}}\right)\left(\frac{G^2H^2}{\omega^2\delta^4}\right) + \frac{n^{\frac{1+\epsilon}{2}}}{T^{\frac{1+\epsilon}{2}}}\left(1 + \sqrt{\frac{n}{T}}\right)\left(\frac{c_0}{\delta^2}\right)\right).$$

Remark 3: Observe that ω, H, δ do not affect the dominating term $\mathcal{O}(\frac{J^2+\bar{\sigma}^2}{\sqrt{nT}})$. Since Theorem 2 provides nonasymptotic guarantee, we need to decide the horizon T before running the algorithm; so, to ensure that the dominating term does not get affected by these different factors, while converging at a rate of $\mathcal{O}(\frac{J^2+\bar{\sigma}^2}{\sqrt{nT}})$, we would be required to fix $T \geq T_1 := C_1 \times \max\{(\frac{c_0^2 n^{(2+\epsilon)}}{(J^2+\bar{\sigma}^2)^2\delta^4})^{1/\epsilon}, \frac{n^3G^4H^4}{(J^2+\bar{\sigma}^2)^2\omega^4\delta^4}\}$ for sufficiently large constant C_1. This implies that for large enough T, we get the benefits of all these techniques in saving the communication bits, essentially for "free," without affecting the convergence rate by too much. The rest of Remark 1 and Remark 2 are also applicable here.

Theoretical Justification for Communication Gain:

The convergence result for SPARQ-SGD highlights savings in communication compared to CHOCO-SGD [3]. For the sake of argument, consider the case when SPARQ-SGD only performs local iterations and no threshold based triggering ($c_t = 0, \forall t$). For the same compression operator ω used for both SPARQ and CHOCO, to transmit the same number of bits (i.e., having same number of communication rounds), T iterations of CHOCO would correspond to $T \times H$ iterations of SPARQ (due to H local SGD steps). Thus, for the same number of bits transmitted, the bound on suboptimality for convex objective for CHOCO is $\mathcal{O}(1/\mu nT) + \mathcal{O}(G^2/\omega^2\delta^4\mu^2T^2)$, while for SPARQ, it is $\mathcal{O}(1/\mu nHT) + \mathcal{O}(G^2/\omega^2\delta^4\mu^2T^2)$. Thus, for the same amount of communication (same number of communication rounds), SPARQ-SGD has a better performance compared to CHOCO-SGD (the first dominant term is affected by H). Similarly, for the same number of communication rounds, the bound on suboptimality for CHOCO-SGD for nonconvex objectives is $\mathcal{O}(1/\sqrt{T}) + \mathcal{O}(1/T)$, while for SPARQ-SGD, it is $\mathcal{O}(1/\sqrt{HT}) + \mathcal{O}(H/T)$. Thus, it can be seen that for large values of T, the performance of SPARQ-SGD is better than that of CHOCO-SGD for the number of communicated bits. Thus, there is theoretical justification for our algorithm to have a better performance while using less bits for communication and this claim is also supported through our experiments.

IV. Preliminaries

In this section, we define the matrix notation which would be used throughout the proofs of Theorems 1 and 2 given in Section V.

Consider the set of parameters $\{\mathbf{x}_i^{(t)}\}_{i=1}^n$ at the nodes at timestep t and estimates of the parameter $\{\hat{\mathbf{x}}_i^{(t)}\}_{i=1}^n$. The matrix notation is given by

$$\mathbf{X}^{(t)} := [\mathbf{x}_1^{(t)}, \ldots, \mathbf{x}_n^{(t)}] \in \mathbb{R}^{d\times n}$$

$$\hat{\mathbf{X}}^{(t)} := [\hat{\mathbf{x}}_1^{(t)}, \ldots, \hat{\mathbf{x}}_n^{(t)}] \in \mathbb{R}^{d\times n}$$

$$\bar{\mathbf{X}}^{(t)} := [\bar{\mathbf{x}}^{(t)}, \ldots, \bar{\mathbf{x}}^{(t)}] \in \mathbb{R}^{d\times n}$$

$$\nabla\mathbf{F}(\mathbf{X}^{(t)}, \boldsymbol{\xi}^{(t)}) := [\nabla F_1(\mathbf{x}_1^{(t)}, \xi_1^{(t)}), \ldots, \nabla F_n(\mathbf{x}_n^{(t)}, \xi_n^{(t)})] \in \mathbb{R}^{d\times n}$$

where $\nabla F_i(\mathbf{x}_i^{(t)}, \xi_i^{(t)})$ denotes the stochastic gradient at node i at timestep t and the vector $\bar{\mathbf{x}}^{(t)} := \frac{1}{n}\sum_{i=1}^n \mathbf{x}_i^{(t)}$ denotes the average of node parameters at time t.

Let $\Gamma^{(t)} \subseteq [n]$ be the set of nodes that do not communicate at time t. We define $\mathbf{P}^{(t)} \in \mathbb{R}^{n\times n}$, a diagonal matrix with $\mathbf{P}_{ii}^{(t)} = 0$ for $i \in \Gamma^{(t)}$ and $\mathbf{P}_{ii}^{(t)} = 1$ otherwise.

SPARQ-SGD in Matrix Notation: Consider Algorithm 1 with synchronization indices given by the set $\{I_{(1)}, I_{(2)}, \ldots, I_{(k)}, \ldots\}$. Using the above notation, the sequence of parameters updates from synchronization index $I_{(t)}$ to $I_{(t+1)}$ is given by

$$\mathbf{Y}^{I_{(t+1)}} = \mathbf{X}^{I_{(t)}} - \sum_{t'=I_{(t)}}^{I_{(t+1)}-1}\eta_{t'}\nabla\mathbf{F}(\mathbf{X}^{(t')}, \boldsymbol{\xi}^{(t')})$$

$$\hat{\mathbf{X}}^{I_{(t+1)}} = \hat{\mathbf{X}}^{I_{(t)}} + \mathcal{C}((\mathbf{Y}^{I_{(t+1)}} - \hat{\mathbf{X}}^{I_{(t)}})\mathbf{P}^{(I_{(t+1)}-1)})$$

$$\mathbf{X}^{I_{(t+1)}} = \mathbf{Y}^{I_{(t+1)}} + \gamma\hat{\mathbf{X}}^{I_{(t+1)}}(\mathbf{W} - \mathbf{I})$$

where $\mathcal{C}(.)$ denotes the compression operator applied columnwise to the argument matrix and $\mathbf{I}$ is the identity matrix. We now note some useful properties of the iterates in matrix notation, which would be used throughout the article.

1) Since $\mathbf{W} \in [0, 1]^{n\times n}$ is a doubly stochastic matrix

$$\bar{\mathbf{X}}^{(t)} = \mathbf{X}^{(t)}\frac{1}{n}\mathbf{1}\mathbf{1}^T, \qquad \bar{\mathbf{X}}^{(t)}\mathbf{W} = \bar{\mathbf{X}}^{(t)}. \quad (3)$$

Here $\mathbf{1} \in \mathbb{R}^n$ is all ones vector. The first expression follows from the definition of $\bar{\mathbf{X}}^{(t)}$. The second expression follows as $\bar{\mathbf{X}}^{(t)}\frac{1}{n}\mathbf{1}\mathbf{1}^T = \bar{\mathbf{X}}^{(t)}$ and $\mathbf{1}^T\mathbf{W} = \mathbf{1}^T$, which imply $\bar{\mathbf{X}}^{(t)}\mathbf{W} = \bar{\mathbf{X}}^{(t)}\frac{1}{n}\mathbf{1}\mathbf{1}^T\mathbf{W} = \bar{\mathbf{X}}^{(t)}\frac{1}{n}\mathbf{1}\mathbf{1}^T = \bar{\mathbf{X}}^{(t)}$.

2) The average of the iterates in Algorithm 1 follows:

$$\bar{\mathbf{X}}^{(t+1)} = \bar{\mathbf{Y}}^{(t+1)} + \mathbb{1}_{(t+1)\in\mathcal{I}_T}\left[\gamma\hat{\mathbf{X}}^{(t+1)}(\mathbf{W}-\mathbf{I})\frac{1}{n}\mathbf{1}\mathbf{1}^T\right] = \bar{\mathbf{Y}}^{(t+1)} \tag{4}$$

where $\mathcal{I}_T$ denotes the set of synchronization indices of Algorithm 1 and $\mathbb{1}_{(t+1)\in\mathcal{I}_T}$ is the indicator variable taking value 1 if time instant $(t+1) \in \mathcal{I}_T$ or 0 otherwise. The above follows from the observation that $\mathbf{W1} = \mathbf{1}$ as $\mathbf{W}$ is a doubly stochastic matrix.

Fact 1: Consider any two matrices $\mathbf{A} \in \mathbb{R}^{d\times n}$, $\mathbf{B} \in \mathbb{R}^{n\times n}$. Then, the following holds:

$$\|\mathbf{AB}\|_F \leq \|\mathbf{A}\|_F\|\mathbf{B}\|_2 \tag{5}$$

where $\|.\|_F$ denotes the Frobenius norm.

Fact 2: (Triggering rule) Consider the set of nodes $\Gamma^{(t)}$ which do not communicate at time t. For a threshold sequence $\{c_t\}_{t=0}^{T-1}$, the triggering rule in Algorithm 1 dictates

$$\|\mathbf{y}_i^{(t+1)} - \hat{\mathbf{x}}_i^{(t)}\|_2^2 \leq c_t\eta_t^2 \qquad \forall i \in \Gamma^{(t)}.$$

Using the matrix notation, since $|\Gamma^{(t)}| \leq n$, this implies

$$\left\|(\mathbf{Y}^{(t+1)} - \hat{\mathbf{X}}^{(t)})(\mathbf{I} - \mathbf{P}^{(t)})\right\|_F^2 \leq nc_t\eta_t^2. \tag{6}$$

Fact 3 (Lemma 16 in [3]): For a doubly stochastic matrix $\mathbf{W}$ with second largest eigenvalue $1-\delta = |\lambda_2(\mathbf{W})| < 1$ and for any non-negative integer k, we have

$$\left\|\mathbf{W}^k - \frac{1}{n}\mathbf{1}\mathbf{1}^T\right\|_2 = (1-\delta)^k. \tag{7}$$

Fact 4: Consider the set of synchronization indices $\{I_{(1)}, I_{(2)}, \ldots, I_{(k)}, \ldots\} \in \mathcal{I}_T$ with the maximum gap between any two consecutive elements bounded by H. Then, for $I_{(k)}, I_{(k+1)} \in \mathcal{I}$ and $\xi := \{\xi^{(t')} : I_{(k)} \leq t' \leq I_{(k+1)}\}$

$$\mathbb{E}_\xi\left[\left\|\sum_{t'=I_{(k)}}^{I_{(k+1)}-1}\eta_{t'}\nabla\mathbf{F}(\mathbf{X}^{(t')}, \xi^{(t')})\right\|_F^2\right] \leq \eta_{I_{(k)}}^2 H^2 nG^2. \tag{8}$$

Fact 5 (Variance bound for workers): Consider the variance bound on the stochastic gradient for nodes $i \in [n]$: $\mathbb{E}_{\xi_i}\|\nabla F_i(\mathbf{x}, \xi_i) - \nabla f_i(\mathbf{x})\|^2 \leq \sigma_i^2$, where $\mathbb{E}_{\xi_i}[\nabla F_i(\mathbf{x}, \xi_i)] = \nabla f_i(\mathbf{x})$. Then, we have

$$\mathbb{E}_{\boldsymbol{\xi}^{(t)}}\left\|\frac{1}{n}\sum_{j=1}^n\left(\nabla f_j(\mathbf{x}_j^{(t)}) - \nabla F_j(\mathbf{x}_j^{(t)}, \xi_j^{(t)})\right)\right\|^2 \leq \frac{\bar{\sigma}^2}{n} \tag{9}$$

where $\xi^{(t)} = \{\xi_1^{(t)}, \xi_2^{(t)}, \ldots, \xi_n^{(t)}\}$ denotes the stochastic samples for the nodes at any timestep t and $\frac{\sum_{j=1}^n\sigma_j^2}{n} = \bar{\sigma}^2$.

V. Proofs of Theorems 1 and 2

In this section, we give proofs of Theorems 1 and 2. Our proof outlines take inspiration from [3], [17], with significant changes in the proof details arising due to the use of local iterations and event-triggered communication. The main idea of our proof involves using the perturbed iterate analysis [43] for the update of the global parameter vector

$$\bar{\mathbf{x}}^{(t+1)} = \bar{\mathbf{x}}^{(t)} - \frac{\eta_t}{n}\sum_{j=1}^n\nabla F_j(\mathbf{x}_j^{(t)}, \xi_j^{(t)}). \tag{10}$$

This requires carefully bounding the term $\sum_{i=1}^n\|\bar{\mathbf{x}}^{(t)} - \mathbf{x}_i^{(t)}\|$ even when performing compression, local iterations, and event-based triggering, and forms the main ingredient in our convergence analyses. More importantly, since we are in a decentralized setting, this result also establishes that the worker nodes reach a *consensus* when optimizing the global objective (1). We provide proof of this result in Lemmas 2 and 3.

A. Proof of Theorem 1

Omitted details for this proof can be found in the extended version of our article [39]. To proceed with the proof, we first note the following lemma from [3, Lemma 20].

Lemma 1: Let $\{\mathbf{x}_t^{(i)}\}_{t=0}^{T-1}$ be generated according to Algorithm 1 with stepsize η_t and define $\bar{\mathbf{x}}_t = \frac{1}{n}\sum_{i=1}^n\mathbf{x}_t^{(i)}$. Then, we have the following result for $\bar{\mathbf{x}}^{(t)}$:

$$\mathbb{E}_{\boldsymbol{\xi}^{(t)}}\|\bar{\mathbf{x}}^{(t+1)} - \mathbf{x}^*\|^2 \leq \left(1 - \frac{\eta_t\mu}{2}\right)\|\bar{\mathbf{x}}^{(t)} - \mathbf{x}^*\|^2 + \frac{\eta_t^2\bar{\sigma}^2}{n} + \eta_t\left(\frac{2\eta_tL^2 + L + \mu}{n}\right)\sum_{j=1}^n\|\bar{\mathbf{x}}^{(t)} - \mathbf{x}_j^{(t)}\|^2 - 2\eta_t(1 - 2L\eta_t)(f(\bar{\mathbf{x}}^{(t)}) - f^*)$$

where $\boldsymbol{\xi}^{(t)} := \{\xi_1^{(t)}, \xi_2^{(t)}, \ldots, \xi_n^{(t)}\}$ is the set of random samples at each worker at time step t and $\bar{\sigma}^2 = \frac{1}{n}\sum_{i=1}^n\sigma_i^2$

Using result of Lemma 1 and taking expectation with respect to the whole process, we have

$$\mathbb{E}\|\bar{\mathbf{x}}^{(t+1)} - \mathbf{x}^*\|^2 \leq \left(1 - \frac{\eta_t\mu}{2}\right)\mathbb{E}\|\bar{\mathbf{x}}^{(t)} - \mathbf{x}^*\|^2 + \frac{\eta_t^2\bar{\sigma}^2}{n} - 2\eta_t(1 - 2L\eta_t)(\mathbb{E}f(\bar{\mathbf{x}}^{(t)}) - f^*) + \eta_t\left(\frac{2\eta_tL^2 + L + \mu}{n}\right)\sum_{j=1}^n\mathbb{E}\|\bar{\mathbf{x}}^{(t)} - \mathbf{x}_j^{(t)}\|^2. \tag{11}$$

As our algorithm uses multiple iterations of local gradient steps, it is useful to include it in the analysis. Let $I_{(t)_0}$ denote the latest synchronization step before or equal to t. Then, we have

$$\mathbf{X}^{(t)} = \mathbf{X}^{I_{(t)_0}} - \sum_{t'=I_{(t)_0}}^{t-1}\eta_{t'}\nabla\mathbf{F}(\mathbf{X}^{(t')}, \boldsymbol{\xi}^{(t')})$$

$$\bar{\mathbf{X}}^{(t)} = \bar{\mathbf{X}}^{I_{(t)_0}} - \sum_{t'=I_{(t)_0}}^{t-1}\eta_{t'}\nabla\mathbf{F}(\mathbf{X}^{(t')}, \boldsymbol{\xi}^{(t')})\tfrac{1}{n}\mathbf{1}\mathbf{1}^T.$$

Thus, the following holds:

$$\mathbb{E}\|\mathbf{X}^{(t)} - \bar{\mathbf{X}}^{(t)}\|_F^2 = \mathbb{E}\left\|\mathbf{X}^{I_{(t)_0}} - \bar{\mathbf{X}}^{I_{(t)_0}} - \sum_{t'=I_{(t)_0}}^{t-1}\eta_{t'}\nabla\mathbf{F}(\mathbf{X}^{(t')}, \boldsymbol{\xi}^{(t')})\left(\mathbf{I} - \tfrac{1}{n}\mathbf{1}\mathbf{1}^T\right)\right\|_F^2$$

$$\leq 2\mathbb{E}\|\mathbf{X}^{I_{(t)_0}} - \bar{\mathbf{X}}^{I_{(t)}}\|_F^2 + 2\mathbb{E}\left\|\sum_{t'=I_{(t)_0}}^{t-1} \eta_{t'}\nabla\mathbf{F}(\mathbf{X}^{(t')}, \boldsymbol{\xi}^{(t')})\left(\mathbf{I} - \frac{1}{n}\mathbf{1}\mathbf{1}^T\right)\right\|_F^2.$$

Using (5) for the second term in above and noting that $\mathbb{E}\|\sum_{t'=I_{(t)_0}}^{t-1}\eta_{t'}\nabla\mathbf{F}(\mathbf{X}^{(t')}, \boldsymbol{\xi}^{(t')})\|_F^2 \leq \eta_{I_{(t)_0}}^2 H^2 nG^2$ from (8) and $\|\frac{1}{n}\mathbf{1}\mathbf{1}^T - \mathbf{I}\|_2^2 = 1$, we get

$$\mathbb{E}\|\mathbf{X}^{(t)} - \bar{\mathbf{X}}^{(t)}\|_F^2 \leq 2\mathbb{E}\|\mathbf{X}^{I_{(t)_0}} - \bar{\mathbf{X}}^{I_{(t)_0}}\|_F^2 + 2\eta_{I_{(t)_0}}^2 H^2 nG^2. \tag{12}$$

We now note the following lemma to bound the first term in RHS of (12), which is proved later in Section VI-A.

Lemma 2: (Consensus—contracting deviation of local iterates and the averaged iterates). Under the assumptions of Theorem 1, for any $I_{(t)} \in \mathcal{I}_T$, we have

$$\mathbb{E}\|\mathbf{X}^{I_{(t)}} - \bar{\mathbf{X}}^{I_{(t)}}\|_F^2 \leq \frac{20nA_{I_{(t)}}\eta_{I_{(t)}}^2}{p^2}$$

where $A_{I_{(t)}} = 4p(8H^2G^2(\frac{1}{p} + \frac{4}{\omega}) + \omega c_{I_{(t)}})$ with $c_{I_{(t)}}$ denoting the triggering threshold at time $I_{(t)}$.

Note that $\{A_t\}_{t=0}^{T-1}$ is an increasing sequence, which follows because $\{c_t\}_{t=0}^{T-1}$ is increasing sequence. Now, since $I_{(t)_0}$ is the last synchronization index before t, we have $A_{I_{(t)_0}} \leq A_t$. We also note the following relation for the learning rate:

$$\frac{\eta_{I_{(t)_0}}}{\eta_t} = \frac{8/\mu(a+I_{(t)_0})}{8/\mu(a+t)} \leq \frac{a + I_{(t)_0} + H}{a + I_{(t)_0}} \leq \left(1 + \frac{H}{a}\right) \overset{(a\geq H)}{\leq} 2.$$

Using these and the bound in Lemma 2 in (12) gives

$$\mathbb{E}\|\mathbf{X}^{(t)} - \bar{\mathbf{X}}^{(t)}\|_F^2 \leq \frac{40nA_t(4\eta_t^2)}{p^2} + 2(4\eta_t^2)H^2nG^2.$$

Using the above bound for the last term in (11) gives

$$\mathbb{E}\|\bar{\mathbf{x}}^{(t+1)} - \mathbf{x}^*\|^2 \leq \left(1 - \frac{\eta_t\mu}{2}\right)\mathbb{E}\|\bar{\mathbf{x}}^{(t)} - \mathbf{x}^*\|^2 + \frac{\eta_t^2\bar{\sigma}^2}{n} - 2\eta_t(1 - 2L\eta_t)(\mathbb{E}f(\bar{\mathbf{x}}^{(t)}) - f^*) + 4\eta_t\left(\frac{2\eta_t L^2 + L + \mu}{n}\right)\left(\frac{40nA_t}{p^2} + 2nH^2G^2\right)\eta_t^2.$$

For $\eta_t = \frac{8}{\mu(a+t)}$ and $a \geq \max\{\frac{32L}{\mu}, \frac{5H}{p}\}$, we have $\eta_t \leq \frac{1}{4L}$. This implies $2L\eta_t - 1 \leq -\frac{1}{2}$ and $(2\eta_t L^2 + L + \mu) \leq (2L + \mu)$. Using these in the above equation gives

$$\mathbb{E}\|\bar{\mathbf{x}}^{(t+1)} - \mathbf{x}^*\|^2 \leq \left(1 - \frac{\eta_t\mu}{2}\right)\mathbb{E}\|\bar{\mathbf{x}}^{(t)} - \mathbf{x}^*\|^2 - \eta_t\mathbb{E}f(\bar{\mathbf{x}}^{(t)}) + \eta_t f^* + \frac{\eta_t^2\bar{\sigma}^2}{n} + 4\eta_t^3(2L + \mu)\left(\frac{40A_t}{p^2} + 2H^2G^2\right).$$

Substituting value of $A_t = 4p(8H^2G^2(\frac{1}{p} + \frac{4}{\omega}) + \omega c_t)$:

$$\mathbb{E}\|\bar{\mathbf{x}}^{(t+1)} - \mathbf{x}^*\|^2 \leq \left(1 - \frac{\eta_t\mu}{2}\right)\mathbb{E}\|\bar{\mathbf{x}}^{(t)} - \mathbf{x}^*\|^2 - \eta_t(\mathbb{E}f(\bar{\mathbf{x}}^{(t)}) - f^*) + \frac{\eta_t^2\bar{\sigma}^2}{n} + Y_1\eta_t^3(2L + \mu)\left(\frac{1}{p^2} + \frac{1}{p\omega} + \frac{\omega c_t}{pH^2G^2} + 2\right)G^2H^2$$

where Y_1 is a universal constant. This gives a recursive relation for the error $\mathbb{E}\|\bar{\mathbf{x}}^{(t)} - \mathbf{x}^*\|^2$. We now use define the following for the sequence relation above:

$$a_t = \mathbb{E}\|\bar{\mathbf{x}}^{(t)} - \mathbf{x}^*\|^2, e_t = \mathbb{E}f(\bar{\mathbf{x}}^{(t)}) - f^*, P = 1, Q = \frac{\bar{\sigma}^2}{n}$$

$$R = Y_1(2L + \mu)\left(\frac{1}{p^2} + \frac{1}{p\omega} + 2\right)G^2H^2$$

$$U_t = Y_1\left(\frac{2L + \mu}{p}\right)\omega c_t$$

using the variant of [7, Lemma 3.3] (proved in [39]), for $w_t = (a + t)^2$ and $e_t = \mathbb{E}f(\bar{\mathbf{x}}^{(t)}) - f^*$, we have

$$\frac{1}{S_T}\sum_{t=0}^{T-1} w_t e_t \leq \frac{\mu a^3}{8S_T}a_0^2 + \frac{4T(T + 2a)}{\mu S_T}\frac{\bar{\sigma}^2}{n} + \frac{Y_1'T}{\mu^2 S_T}(2L + \mu)\left(\frac{1}{p^2} + \frac{1}{p\omega} + 2\right)G^2H^2 + \frac{Y_2'c_0\omega T^{(2-\epsilon)}}{\mu^2(2 - \epsilon)S_T}\left(\frac{2L + \mu}{p}\right)$$

where $\epsilon \in (0, 1)$ and Y_1' and Y_2' are constants. From the convexity of f, we finally have

$$\mathbb{E}f(\mathbf{x}_{avg}^{(T)}) - f^* \leq \frac{1}{S_T}\sum_{t=0}^{T-1} w_t e_t$$

where $\bar{\mathbf{x}}_{avg}^{(T)} = \frac{1}{S_T}\sum_{t=0}^{T-1} w_t\bar{\mathbf{x}}^{(t)}$. We finally use the fact that $p \leq \omega$ (as $\delta \leq 1$ and $p := \frac{\gamma^*\delta}{8}$ with $\gamma^* \leq \omega$). This implies that the above expression can be simplified as

$$\mathbb{E}f(\mathbf{x}_{avg}^{(T)}) - f^* \leq \frac{\mu a^3}{8S_T}a_0^2 + \frac{4T(T + 2a)}{\mu S_T}\frac{\bar{\sigma}^2}{n} + \frac{Z_1T}{\mu^2 S_T}\frac{(2L + \mu)}{p^2}G^2H^2 + \frac{Z_2c_0\omega T^{(2-\epsilon)}}{\mu^2(2 - \epsilon)S_T}\left(\frac{2L + \mu}{p}\right)$$

where Z_1, Z_2 are constants, completing proof of Theorem 1.

B. Proof of Theorem 2

We start the proof with learning rate set to η_t. We do not use any algebraic property of the learning rate until (18), thus the analysis remains the same till then for both constant learning rate $\eta_t = \eta$ and for decaying η_t. We do this to reuse the analysis till (18) in the proof for nonconvex objective with varying step size provided in extended version [39].

Initial part of the proof uses techniques from [17, Th. A.2]. Consider expectation taken over sampling at time instant t: $\boldsymbol{\xi}^{(t)} = \{\xi_1^{(t)}, \xi_2^{(t)}, \ldots, \xi_n^{(t)}\}$ and using $\bar{\mathbf{X}}^{(t+1)} = \bar{\mathbf{Y}}^{(t+1)}$ [from (4)] which gives $\bar{\mathbf{x}}^{(t+1)} = \bar{\mathbf{x}}^{(t)} - \frac{\eta_t}{n}\sum_{j=1}^n \nabla F_j(\mathbf{x}_j^{(t)}, \xi_j^{(t)})$. Thus, we have

$$\mathbb{E}_{\boldsymbol{\xi}^{(t)}}f(\bar{\mathbf{x}}^{(t+1)}) = \mathbb{E}_{\boldsymbol{\xi}^{(t)}}f\left(\bar{\mathbf{x}}^{(t)} - \frac{\eta_t}{n}\sum_{j=1}^n \nabla F_j(\mathbf{x}_j^{(t)}, \boldsymbol{\xi}_j^{(t)})\right).$$

Using L-smoothness of f, we get

$$\mathbb{E}_{\boldsymbol{\xi}^{(t)}}f(\bar{\mathbf{x}}^{(t+1)}) \leq f(\bar{\mathbf{x}}^{(t)}) + \mathbb{E}_{\boldsymbol{\xi}^{(t)}}\frac{L}{2}\eta_t^2\left\|\sum_{j=1}^n \frac{\nabla F_j(\mathbf{x}_j^{(t)}, \boldsymbol{\xi}_j^{(t)})}{n}\right\|_2^2$$

$$- \mathbb{E}_{\boldsymbol{\xi}^{(t)}} \left\langle \nabla f(\bar{\mathbf{x}}^{(t)}), \frac{\eta_t}{n} \sum_{j=1}^{n} \nabla F_j(\mathbf{x}_j^{(t)}, \boldsymbol{\xi}_j^{(t)}) \right\rangle. \quad (13)$$

To estimate the second term in (13), we note that

$$\begin{aligned}
&- \eta_t \mathbb{E}_{\boldsymbol{\xi}^{(t)}} \left\langle \nabla f(\bar{\mathbf{x}}^{(t)}), \frac{1}{n} \sum_{j=1}^{n} \nabla F_j(\mathbf{x}_j^{(t)}, \boldsymbol{\xi}_j^{(t)}) \right\rangle \\
&= - \eta_t \left\langle \nabla f(\bar{\mathbf{x}}^{(t)}), \frac{1}{n} \sum_{j=1}^{n} \nabla f_j(\mathbf{x}_j^{(t)}) \right\rangle \\
&\overset{(a)}{=} \eta_t \langle \nabla f(\bar{\mathbf{x}}^{(t)}), \nabla f(\bar{\mathbf{x}}^{(t)}) - \frac{1}{n} \sum_{j=1}^{n} \nabla f_j(\mathbf{x}_j^{(t)}) \rangle \\
&\quad - \eta_t \|\nabla f(\bar{\mathbf{x}}^{(t)})\|_2^2 \\
&= \eta_t \langle \nabla f(\bar{\mathbf{x}}^{(t)}), \frac{1}{n} \sum_{j=1}^{n} (\nabla f_j(\bar{\mathbf{x}}^{(t)}) - \nabla f_j(\mathbf{x}_j^{(t)})) \rangle \\
&\quad - \eta_t \|\nabla f(\bar{\mathbf{x}}^{(t)})\|_2^2 \\
&\overset{(b)}{\leq} - \frac{\eta_t}{2} \|\nabla f(\bar{\mathbf{x}}^{(t)})\|_2^2 + \frac{\eta_t}{2n} \sum_{j=1}^{n} \|\nabla f_j(\bar{\mathbf{x}}^{(t)}) - \nabla f_j(\mathbf{x}_j^{(t)})\|^2
\end{aligned}$$

where in (a) we add and subtract $\nabla f(\bar{\mathbf{x}}^{(t)})$ and (b) follows by noting that $\langle \mathbf{p}, \mathbf{q} \rangle \leq \frac{\|\mathbf{p}\|^2 + \|\mathbf{q}\|^2}{2}$ for any $\mathbf{p}, \mathbf{q} \in \mathbb{R}^d$. Using L-Lipschitz continuity of gradient of f_j for $j \in [n]$, we have

$$\begin{aligned}
&- \eta_t \mathbb{E}_{\boldsymbol{\xi}^{(t)}} \left\langle \nabla f(\bar{\mathbf{x}}^{(t)}), \frac{1}{n} \sum_{j=1}^{n} \nabla F_j(\mathbf{x}_j^{(t)}, \boldsymbol{\xi}_j^{(t)}) \right\rangle \leq \\
&- \frac{\eta_t}{2} \|\nabla f(\bar{\mathbf{x}}^{(t)})\|_2^2 + \frac{\eta_t L^2}{2n} \sum_{j=1}^{n} \|\bar{\mathbf{x}}^{(t)} - \mathbf{x}_j^{(t)}\|^2. \quad (14)
\end{aligned}$$

To estimate the last term in (13), we add and subtract $\nabla f(\bar{\mathbf{x}}^{(t)}) = \frac{1}{n} \sum_{j=1}^{n} \nabla f_i(\bar{\mathbf{x}}_t)$ and $\frac{1}{n} \sum_{j=1}^{n} \nabla f_j(\mathbf{x}_t^{(j)})$

$$\begin{aligned}
&\frac{L}{2} \eta_t^2 \mathbb{E}_{\boldsymbol{\xi}^{(t)}} \left\| \frac{1}{n} \sum_{j=1}^{n} \nabla F_j(\mathbf{x}_j^{(t)}, \boldsymbol{\xi}_j^{(t)}) \right\|_2^2 \\
&= \mathbb{E}_{\boldsymbol{\xi}^{(t)}} \left[\frac{L}{2} \eta_t^2 \left\| \frac{1}{n} \sum_{j=1}^{n} (\nabla F_j(\mathbf{x}_j^{(t)}, \boldsymbol{\xi}_j^{(t)}) - \nabla f_j(\mathbf{x}_j^{(t)})) \right.\right. \\
&\qquad \left.\left. + \frac{1}{n} \sum_{j=1}^{n} (\nabla f_j(\mathbf{x}_j^{(t)}) - \nabla f_j(\bar{\mathbf{x}}^{(t)})) + \nabla f(\bar{\mathbf{x}}^{(t)}) \right\|_2^2 \right] \\
&\leq L \eta_t^2 \mathbb{E}_{\boldsymbol{\xi}^{(t)}} \left\| \frac{1}{n} \sum_{j=1}^{n} (\nabla F_j(\mathbf{x}_j^{(t)}, \boldsymbol{\xi}_j^{(t)} - \nabla f_j(\mathbf{x}_j^{(t)})) \right\|_2^2 \\
&\quad + \frac{2L\eta_t^2}{n} \sum_{j=1}^{n} \left\| (\nabla f_j(\mathbf{x}_j^{(t)}) - \nabla f_j(\bar{\mathbf{x}}^{(t)})) \right\|_2^2 \\
&\quad + 2L\eta_t^2 \left\| \nabla f(\bar{\mathbf{x}}^{(t)}) \right\|_2^2.
\end{aligned}$$

Using the variance bound (5) for the first term and L-Lipschitz continuity of gradients of f_j for $j \in [n]$ for the second term in above, we get

$$\begin{aligned}
&\frac{L}{2} \eta_t^2 \mathbb{E}_{\boldsymbol{\xi}^{(t)}} \left\| \frac{1}{n} \sum_{j=1}^{n} \nabla F_j(\mathbf{x}_j^{(t)}, \boldsymbol{\xi}_j^{(t)}) \right\|_2^2 \leq \frac{L\eta_t^2 \bar{\sigma}^2}{n} \\
&+ \frac{2L^3 \eta_t^2}{n} \sum_{j=1}^{n} \left\| \mathbf{x}_j^{(t)} - \bar{\mathbf{x}}^{(t)} \right\|_2^2 + 2L\eta_t^2 \left\| \nabla f(\bar{\mathbf{x}}^{(t)}) \right\|_2^2. \quad (15)
\end{aligned}$$

Substituting (14) and (15) to (13) and taking expectation with respect to the entire process gives

$$\begin{aligned}
&\mathbb{E}[f(\bar{\mathbf{x}}^{(t+1)})] \leq \mathbb{E} f(\bar{\mathbf{x}}^{(t)}) - \eta_t \left(\frac{1}{2} - 2L\eta_t \right) \mathbb{E} \|\nabla f(\bar{\mathbf{x}}^{(t)})\|_2^2 \\
&+ \frac{L\eta_t^2 \bar{\sigma}^2}{n} + \left(\frac{\eta_t L^2}{2n} + \frac{2L^3 \eta_t^2}{n} \right) \textstyle\sum_{j=1}^{n} \mathbb{E} \|\bar{\mathbf{x}}^{(t)} - \mathbf{x}_j^{(t)}\|^2. \quad (16)
\end{aligned}$$

Let $I_{(t+1)_0}$ denote the latest synchronization step before or equal to $(t+1)$. Then, we have

$$\begin{aligned}
\mathbf{X}^{(t+1)} &= \mathbf{X}^{I_{(t+1)_0}} - \sum_{t'=I_{(t+1)_0}}^{t} \eta_{t'} \nabla \mathbf{F}(\mathbf{X}^{(t')}, \boldsymbol{\xi}^{(t')}) \\
\bar{\mathbf{X}}^{(t+1)} &= \bar{\mathbf{X}}^{I_{(t+1)_0}} - \sum_{t'=I_{(t+1)_0}}^{t} \eta_{t'} \nabla \mathbf{F}(\mathbf{X}^{(t')}, \boldsymbol{\xi}^{(t')}) \tfrac{1}{n} \mathbf{1}\mathbf{1}^T.
\end{aligned}$$

Thus, the following holds:

$$\begin{aligned}
&\mathbb{E} \|\mathbf{X}^{(t+1)} - \bar{\mathbf{X}}^{(t+1)}\|_F^2 = \mathbb{E} \left\| \mathbf{X}^{I_{(t+1)_0}} - \bar{\mathbf{X}}^{I_{(t+1)_0}} \right. \\
&\qquad \left. - \sum_{t'=I_{(t+1)_0}}^{t} \eta_{t'} \nabla \mathbf{F}(\mathbf{X}^{(t')}, \boldsymbol{\xi}^{(t')}) \left(\mathbf{I} - \tfrac{1}{n} \mathbf{1}\mathbf{1}^T \right) \right\|_F^2 \\
&\leq 2\mathbb{E} \|\mathbf{X}^{I_{(t+1)_0}} - \bar{\mathbf{X}}^{I_{(t+1)_0}}\|_F^2 \\
&\qquad + 2\mathbb{E} \left\| \sum_{t'=I_{(t+1)_0}}^{t} \eta_{t'} \nabla \mathbf{F}(\mathbf{X}^{(t')}, \boldsymbol{\xi}^{(t')}) \left(\mathbf{I} - \tfrac{1}{n} \mathbf{1}\mathbf{1}^T \right) \right\|_F^2.
\end{aligned}$$

Using (5) for the second term in above and noting that $\mathbb{E} \| \sum_{t'=I_{(t+1)_0}}^{t} \eta_{t'} \nabla \mathbf{F}(\mathbf{X}^{(t')}, \boldsymbol{\xi}^{(t')}) \|_F^2 \leq \eta_{I_{(t+1)_0}} n H^2 G^2$ and $\| \frac{1}{n} \mathbf{1}\mathbf{1}^T - \mathbf{I} \|_2^2 = 1$ from (8) and (7) (with $k = 0$) respectively, we have

$$\begin{aligned}
\mathbb{E} \|\mathbf{X}^{(t+1)} - \bar{\mathbf{X}}^{(t+1)}\|_F^2 &\leq 2\mathbb{E} \|\mathbf{X}^{I_{(t+1)_0}} - \bar{\mathbf{X}}^{I_{(t+1)_0}}\|_F^2 \\
&\quad + 2H^2 n \eta_{I_{(t+1)_0}}^2 G^2. \quad (17)
\end{aligned}$$

By noting that $\sum_{j=1}^{n} \mathbb{E} \|\bar{\mathbf{x}}^{(t)} - \mathbf{x}_j^{(t)}\|^2 = \mathbb{E} \|\mathbf{X}^{(t)} - \bar{\mathbf{X}}^{(t)}\|_F^2$, we use (17) to bound the last term in (16), which gives

$$\begin{aligned}
&\mathbb{E}[f(\bar{\mathbf{x}}^{(t+1)})] \leq \mathbb{E} f(\bar{\mathbf{x}}^{(t)}) - \left(\frac{\eta_t}{2} - 2L\eta_t^2 \right) \mathbb{E} \|\nabla f(\bar{\mathbf{x}}^{(t)})\|_2^2 + \frac{L\eta_t^2 \bar{\sigma}^2}{n} \\
&+ \left(\frac{\eta_t L^2}{n} + \frac{4L^3 \eta_t^2}{n} \right) [\mathbb{E} \|\mathbf{X}^{I_{(t)_0}} - \bar{\mathbf{X}}^{I_{(t)_0}}\|_F^2 + H^2 n \eta_{I_{(t)_0}}^2 G^2]. \quad (18)
\end{aligned}$$

We now replace η_t with a fixed learning rate η

$$\begin{aligned}
&\mathbb{E}[f(\bar{\mathbf{x}}^{(t+1)})] \leq \mathbb{E} f(\bar{\mathbf{x}}^{(t)}) - \left(\frac{\eta}{2} - 2L\eta^2 \right) \mathbb{E} \|\nabla f(\bar{\mathbf{x}}^{(t)})\|_2^2 + \frac{L\eta^2 \bar{\sigma}^2}{n} \\
&+ \left(\frac{\eta L^2}{n} + \frac{4L^3 \eta^2}{n} \right) \left[\mathbb{E} \|\mathbf{X}^{I_{(t)_0}} - \bar{\mathbf{X}}^{I_{(t)_0}}\|_F^2 + H^2 n \eta^2 G^2 \right]. \quad (19)
\end{aligned}$$

We now note the following lemma, which is proved in Section VI-B, to bound the second term in RHS of (19).

Lemma 3: (Bounded deviation of local iterates and the averaged iterates). Under the assumptions of Theorem 2, for any $I_{(t)} \in \mathcal{I}_T$, we have

$$\mathbb{E}\|\mathbf{X}^{I_{(t)}} - \bar{\mathbf{X}}^{I_{(t)}}\|_F^2 \leq \frac{4nA\eta^2}{p^2}$$

where constant $A = 4p(8H^2G^2(\frac{4}{\omega} + \frac{1}{p}) + \frac{\omega c_0}{\eta^{(1-\epsilon)}})$.

Remark 4: Lemma 3 is about consensus with bounded error, i.e., the nodes do not reach a consensus, but within an error proportional to the learning rate η. Note that if we take a decaying learning rate η_t (as in the strongly convex case), then, as shown in Lemma 2, different nodes will asymptotically reach to a consensus; however, the convergence rate of our algorithm will no longer be $\mathcal{O}(\frac{1}{\sqrt{T}})$, but we will only get a rate of $\mathcal{O}(\frac{1}{\log T})$, which, though matches the convergence rate of running vanilla SGD with decaying learning rate on nonconvex objectives, is much slower than what we can get with a fixed learning rate. We provide convergence result of SPARQ-SGD with a decaying learning rate on smooth nonconvex objectives in our extended article [39].

Using Lemma 3, for $A = 4p(8H^2G^2(\frac{4}{\omega} + \frac{1}{p}) + \frac{\omega c_0}{\eta^{(1-\epsilon)}})$, we get $\mathbb{E}\|\mathbf{X}^{I_{(t)_0}} - \bar{\mathbf{X}}^{I_{(t)_0}}\|_F^2 \leq \frac{4nA\eta^2}{p^2}$. Substituting in (19)

$$\mathbb{E}f(\bar{\mathbf{x}}^{(t+1)}) \leq \mathbb{E}f(\bar{\mathbf{x}}^{(t)}) - \eta\left(\frac{1}{2} - 2L\eta\right)\mathbb{E}\|\nabla f(\bar{\mathbf{x}}^{(t)})\|_2^2 + \frac{L\bar{\sigma}^2\eta^2}{n} + \left(\frac{\eta L^2}{2n} + \frac{2L^3\eta^2}{n}\right)\left(\frac{8An}{p^2} + 2nH^2G^2\right)\eta^2$$

For $\eta = \sqrt{\frac{n}{T}}$ and $T \geq 64nL^2$, we have $\eta \leq \frac{1}{8L}$, giving

$$\mathbb{E}f(\bar{\mathbf{x}}^{(t+1)}) \leq \mathbb{E}f(\bar{\mathbf{x}}^{(t)}) - \frac{\eta}{4}\mathbb{E}\|\nabla f(\bar{\mathbf{x}}^{(t)})\|_2^2 + \frac{L\bar{\sigma}^2\eta^2}{n} + \frac{L^2}{2}\left(\frac{8A}{p^2} + 2H^2G^2\right)\eta^3 + 2L^3\left(\frac{8A}{p^2} + 2H^2G^2\right)\eta^4.$$

Rearranging terms and summing from 0 to $T-1$, we get

$$\sum_{t=0}^{T-1}\eta\mathbb{E}\|\nabla f(\bar{\mathbf{x}}^{(t)})\|_2^2 \leq 4\left(f(\bar{\mathbf{x}}_0) - \mathbb{E}f(\bar{\mathbf{x}}^{(t)})\right) + 2L^2\left(\frac{8A}{p^2} + 2H^2G^2\right)\sum_{t=0}^{T-1}\eta^3 + \frac{4L\bar{\sigma}^2}{n}\sum_{t=0}^{T-1}\eta^2 + 8L^3\left(\frac{8A}{p^2} + 2H^2G^2\right)\sum_{t=0}^{T-1}\eta^4.$$

Dividing both sides by ηT and using $\mathbb{E}f(\bar{\mathbf{x}}^{(t)}) \geq f^*$

$$\frac{\sum_{t=0}^{T-1}\mathbb{E}\|\nabla f(\bar{\mathbf{x}}^{(t)})\|_2^2}{T} \leq \frac{4\left(f(\bar{\mathbf{x}}_0) - f^*\right)}{\eta T} + \frac{4L\bar{\sigma}^2}{n}\eta + C_1L^2\left(\frac{A}{p^2} + H^2G^2\right)\eta^2 + C_2L^3\left(\frac{A}{p^2} + H^2G^2\right)\eta^3$$

where C_1, C_2 are constants. Noting that $\frac{A}{p^2} \geq H^2G^2$, we get

$$\frac{\sum_{t=0}^{T-1}\mathbb{E}\|\nabla f(\bar{\mathbf{x}}^{(t)})\|_2^2}{T} \leq \frac{4\left(f(\bar{\mathbf{x}}_0) - f^*\right)}{\eta T} + \frac{D_1L^2A}{p^2}\eta^2 + \frac{D_2L^3A}{p^2}\eta^3 + \frac{4L\bar{\sigma}^2}{n}\eta$$

for universal constant D_1, D_2. Substituting the value of $A = 4p\left(8H^2G^2(\frac{4}{\omega} + \frac{1}{p}) + \frac{\omega c_0}{\eta^{(1-\epsilon)}}\right)$, we have

$$\frac{\sum_{t=0}^{T-1}\mathbb{E}\|\nabla f(\bar{\mathbf{x}}^{(t)})\|_2^2}{T} \leq \frac{4\left(f(\bar{\mathbf{x}}_0) - f^*\right)}{\eta T} + \frac{D_1'G^2H^2L^2}{p}\eta^2(1+L\eta)\left(\frac{4}{\omega} + \frac{1}{p}\right) + \frac{D_2'L^2\omega c_0}{p}\eta^{(1+\epsilon)}(1+L\eta) + \frac{4L\bar{\sigma}^2}{n}\eta$$

for some constants D_1', D_2'. Substituting $\eta = \sqrt{\frac{n}{T}}$, we get the convergence rate as

$$\frac{\sum_{t=0}^{T-1}\mathbb{E}\|\nabla f(\bar{\mathbf{x}}^{(t)})\|_2^2}{T} \leq \frac{4\left(f(\bar{\mathbf{x}}_0) - f^* + L\bar{\sigma}^2\right)}{\sqrt{nT}} + \frac{D_1'G^2H^2L^2n}{Tp}\left(\frac{4}{\omega} + \frac{1}{p}\right) + \frac{4D_1'G^2H^2L^3n^{3/2}}{T^{3/2}p}\left(\frac{4}{\omega} + \frac{1}{p}\right) + \frac{D_2'L^2\omega c_0\sqrt{n^{(1+\epsilon)}}}{p\sqrt{T^{(1+\epsilon)}}} + \frac{4D_2'L^3\omega c_0\sqrt{n^{(2+\epsilon)}}}{p\sqrt{T^{(2+\epsilon)}}}$$

for some $\epsilon \in (0,1)$. We finally use the fact that $p \leq \omega$ (as $\delta \leq 1$ and $p := \frac{\gamma^*\delta}{8}$ with $\gamma^* \leq \omega$). This completes the proof of Theorem 2.

VI. Proofs of Lemmas 2 and 3

In this section, we provide proofs for Lemmas 2 and 3 and bound $\sum_{j=1}^n \mathbb{E}\|\bar{\mathbf{x}}^{I_{(t)}} - \mathbf{x}_j^{I_{(t)}}\|^2$, thereby, establishing consensus among the nodes.

A. Proof Sketch for Lemma 2

We first provide a proof sketch of Lemma 2, which states that $e_{I_{(t)}}^{(1)} := \sum_{j=1}^n \mathbb{E}\|\bar{\mathbf{x}}^{I_{(t)}} - \mathbf{x}_j^{I_{(t)}}\|^2$—the expected difference between local and the average iterates at the synchronization indices – decays asymptotically to zero for decaying learning rate η_t. We first define another quantity $e_{I_{(t)}}^{(2)} := \sum_{j=1}^n \mathbb{E}\|\hat{\mathbf{x}}_j^{I_{(t+1)}} - \mathbf{x}_j^{I_{(t)}}\|^2$ which is the sum of expected difference between the local iterates and their copies.

We now note the following lemmas, which recursively bound $e_{I_{(t+1)}}^{(1)}$ and $e_{I_{(t+1)}}^{(2)}$ in terms of $e_{I_{(t)}}^{(1)}$ and $e_{I_{(t)}}^{(2)}$.

Lemma 4: Under the setting of Theorem 1, $e_{I_{(t+1)}}^{(1)} := \sum_{j=1}^n \mathbb{E}\|\bar{\mathbf{x}}^{I_{(t+1)}} - \mathbf{x}_j^{I_{(t+1)}}\|^2$ satisfies

$$e_{I_{(t+1)}}^{(1)} \leq (1+\alpha_5^{-1})R_1e_{I_{(t)}}^{(1)} + (1+\alpha_5^{-1})R_2e_{I_{(t)}}^{(2)} + Q_1\eta_{I_{(t+1)}}^2$$

where $R_1 = (1+\alpha_1)(1-\gamma\delta)^2, R_2 = (1+\alpha_1^{-1})\gamma^2\lambda^2$ and $Q_1 = 4H^2nG^2(1+\alpha_5)(R_1+R_2)$. Here α_1, α_5 are constants strictly greater than 0.

Lemma 5: Under the setting of Theorem 2, $e_{I_{(t+1)}}^{(2)} := \sum_{j=1}^n \mathbb{E}\|\hat{\mathbf{x}}_j^{I_{(t+2)}} - \mathbf{x}_j^{I_{(t+1)}}\|^2$ satisfies

$$e_{I_{(t+1)}}^{(2)} \leq (1+\alpha_5^{-1})R_3e_{I_{(t)}}^{(2)} + (1+\alpha_5^{-1})R_4e_{I_{(t)}}^{(1)} + \eta_{I_{(t+1)}}^2Q_2$$

where $R_3 = (1+\gamma\lambda)^2(1+\alpha_4)(1+\alpha_3)(1+\alpha_2)(1-\omega)$, $R_4 = \gamma^2\lambda^2(1+\alpha_4^{-1})(1+\alpha_3)(1+\alpha_2)(1-\omega)$ and $Q_2 = 4H^2nG^2((1+\alpha_5)(R_3+R_4)+(1+\alpha_2^{-1})+(1+\alpha_3^{-1})(1+\alpha_2)(1-\omega))+(1+\alpha_2)\omega nc_{I_{(t+1)}}$. Here $\alpha_2,\alpha_3,\alpha_4>0$ and $\alpha_5>0$ is the same as used in Lemma 4.

We prove both Lemmas 4 and 5 in Appendix A.

Proof of Lemma 2: To proceed with the proof, we first define the total error at synchronization index $I_{(t+1)}$ as

$$e_{I_{(t+1)}} := e^{(1)}_{I_{(t+1)}} + e^{(2)}_{I_{(t+1)}}. \tag{20}$$

Note that Lemmas 4 and 5 provide bounds for the first and the second term in the RHS of (20). Substituting them in (20) gives

$$e_{I_{(t+1)}} \le (R_1+R_4)(1+\alpha_5^{-1})e^{(1)}_{I_{(t)}} + (R_2+R_3)(1+\alpha_5^{-1})e^{(2)}_{I_{(t)}} + (Q_1+Q_2)\eta^2_{I_{(t+1)}}. \tag{21}$$

Define the following:

$$\pi_1(\gamma) := R_2+R_3 = \gamma^2\lambda^2(1+\alpha_1^{-1}) + (1+\gamma\lambda)^2(1+\alpha_4)(1+\alpha_3)(1+\alpha_2)(1-\omega)$$

$$\pi_2(\gamma) := R_1+R_4 = (1-\delta\gamma)^2(1+\alpha_1) + \gamma^2\lambda^2(1+\alpha_4^{-1})(1+\alpha_3)(1+\alpha_2)(1-\omega)$$

$$\pi_t := Q_1+Q_2 \le 4H^2nG^2(1+\alpha_5)(R_1+R_2+R_3+R_4) + 4H^2nG^2((1+\alpha_2^{-1})+(1-\omega)(1+\alpha_3^{-1})(1+\alpha_2)) + (1+\alpha_2)\omega nc_t.$$

With these substitutions, $e_{I_{(t+1)}}$ in (21) can be bounded as

$$e_{I_{(t+1)}} \le (1+\alpha_5^{-1})max\{\pi_1(\gamma),\pi_2(\gamma)\}\left[e^{(1)}_{I_{(t)}} + e^{(2)}_{I_{(t)}}\right] + \pi_{I_{(t+1)}}\eta^2_{I_{(t+1)}}. \tag{22}$$

Technical details for calculation of $\max\{\pi_1(\gamma),\pi_2(\gamma)\}$ and $\pi_{I_{(t)}}$ can be found in [39], where we show that $\max\{\pi_1(\gamma),\pi_2(\gamma)\} \le (1-\frac{\gamma^*\delta}{8}) \le (1-\frac{\delta^2\omega}{644})$ and $\pi_t \le (16H^2nG^2(\frac{1}{p}+\frac{4}{\omega})+2\omega nc_t)$, where $p := \frac{\gamma^*\delta}{8}$ and $\gamma^* = \frac{2\delta\omega}{64\delta+\delta^2+16\lambda^2+8\delta\lambda^2-16\delta\omega}$ is the consensus step size. Substituting these and $\alpha_5 = \frac{2}{p}$ (used to bound π_t) for $p = \frac{\delta\gamma^*}{8}$ in (22) gives

$$e_{I_{(t+1)}} \le \left(1+\frac{p}{2}\right)\left(1-\frac{\delta\gamma^*}{8}\right)\left[e^{(1)}_{I_{(t)}} + e^{(2)}_{I_{(t)}}\right] + \left(16H^2nG^2\left(\frac{1}{p}+\frac{4}{\omega}\right)+2\omega nc_{I_{(t+1)}}\right)\eta^2_{I_{(t+1)}}.$$

Define $z_t := \pi_t = (16H^2nG^2(\frac{1}{p}+\frac{4}{\omega})+2\omega nc_t)$. Since $c_{I_{(t+1)}}\eta^2_{I_{(t+1)}} \le 4c_{I_{(t)}}\eta^2_{I_{(t)}}$ (see Claim 2 in [39]) and $\eta_{I_{(t+1)}} \le \eta_{I_{(t)}}$, we have that $z_{I_{(t+1)}}\eta^2_{I_{(t+1)}} \le 4z_{I_{(t)}}\eta^2_{I_{(t)}}$. Thus, we have the following recurrence relation for $e_{I_{(t)}}$:

$$e_{I_{(t+1)}} \le \left(1+\frac{p}{2}\right)(1-p)e_{I_{(t)}} + 4z_{I_{(t)}}\eta^2_{I_{(t)}}.$$

Define $A_t := \frac{2pz_t}{n}$. Thus, we have the following relation:

$$e_{I_{(t+1)}} \le \left(1-\frac{p}{2}\right)e_{I_{(t)}} + \frac{2nA_{I_{(t)}}}{p}\eta^2_{I_{(t)}}. \tag{23}$$

The recursion for $e_{I_{(t)}}$ in (23) can be bounded as

$$e_{I_{(t)}} \le \frac{20}{p^2}nA_{I_{(t)}}\eta^2_{I_{(t)}}, \qquad \forall I_{(t)} \in \mathcal{I}_T.$$

Note that we also have $e^{(1)}_{I_{(t)}} \le e_{I_{(t)}}$. Thus, we get the following result for any synchronization index $I_{(t)} \in \mathcal{I}_T$:

$$e^{(1)}_{I_{(t)}} = \mathbb{E}\|\bar{\mathbf{X}}^{I_{(t)}} - \mathbf{X}^{I_{(t)}}\|^2_F \le \frac{20}{p^2}nA_{I_{(t)}}\eta^2_{I_{(t)}}$$

where $A_{I_{(t)}} = 4p(8H^2G^2(\frac{1}{p}+\frac{4}{\omega})+\omega c_{I_{(t)}}), p = \frac{\delta\gamma^*}{8} \le 1, \delta \in (0,1],\ \beta \in [0,1),\ \omega \in (0,1],\ \gamma^* = \frac{2\delta\omega}{64\delta+\delta^2+16\lambda^2+8\delta\lambda^2-16\delta\omega}$ is the chosen consensus step size, and $\{c_t\}_{t\ge0}$ is the triggering threshold sequence. This completes the proof for Lemma 2. □

Remark 5: Note that [3] also proved analogous inequalities to Lemmas 4 and 5 with $Q_1 = Q_2 = 0$. Here $Q_1, Q_2(t)$ are non-zero (with $Q_2(t)$ possibly varying with t) and arise due to the use of local iterations and event-triggered communication, which makes the proof of these inequalities significantly more involved than the corresponding inequalities in [3].

B. Proof Sketch for Lemma 3

Lemma 3 is essentially about consensus with bounded error, i.e., the nodes do not reach to a consensus, but within an error that is proportional to the learning rate η.

Proof of Lemma 3: The proof follows similar steps to proof of Lemma 2. Note that in proof of either Lemma 4 or Lemma 5 (provided in Appendix A), we do not use any properties of the learning rate sequence, and thus the bounds derived in these lemmas also hold with constant[8] step-size η. Following along the same steps as proof of Lemma 2 till (23), we can define the recurrence relation like (23), but with constant η as

$$e_{I_{(t+1)}} \le \left(1-\frac{p}{2}\right)e_{I_{(t)}} + \frac{2nA_{I_{(t)}}}{p}\eta^2$$

where $A_t := \frac{2p}{n}(16H^2nG^2(\frac{4}{\omega}+\frac{1}{p})+2\omega nc_t)$. As we restrict our triggering sequence $c_t \le \frac{c_0}{\eta^{(1-\epsilon)}}\ \forall t$, for some $\epsilon \in (0,1)$

$$e_{I_{(t+1)}} \le \left(1-\frac{p}{2}\right)e_{I_{(t)}} + \frac{2nA}{p}\eta^2 \tag{24}$$

where constant $A = 4p\left(8H^2G^2(\frac{4}{\omega}+\frac{1}{p})+\frac{\omega c_0}{\eta^{(1-\epsilon)}}\right)$. Using the recursive definition of $e_{I_{(t)}}$ in (24), it can be shown that for all $I_{(t)} \in \mathcal{I}_T$, we have

$$e_{I_{(t)}} \le \frac{4nA\eta^2}{p^2}.$$

Note that we also have $e^{(1)}_{I_{(t)}} \le e_{I_{(t)}}$. Thus, we get the following result for any synchronization index $I_{(t)} \in \mathcal{I}_T$:

$$e^{(1)}_{I_{(t)}} = \mathbb{E}\|\bar{\mathbf{X}}^{I_{(t)}} - \mathbf{X}^{I_{(t)}}\|^2_F \le \frac{4nA\eta^2}{p^2}$$

where $A = 4p\left(8H^2G^2(\frac{4}{\omega}+\frac{1}{p})+\frac{\omega c_0}{\eta^{(1-\epsilon)}}\right)$ for $p = \frac{\delta\gamma^*}{8},\ \epsilon > 0$ and $\gamma^* = \frac{2\delta\omega}{64\delta+\delta^2+16\beta^2+8\delta\beta^2-16\delta\omega}$ is the chosen consensus step size. This completes the proof for Lemma 3 □

[8]As a small note, we use $\eta_{I_{(t)}} \le 2\eta_{I_{(t+1)}}$ in proof of Lemma 4, however, when η is constant, it is easy to see that the bound of Lemma 4 still holds.

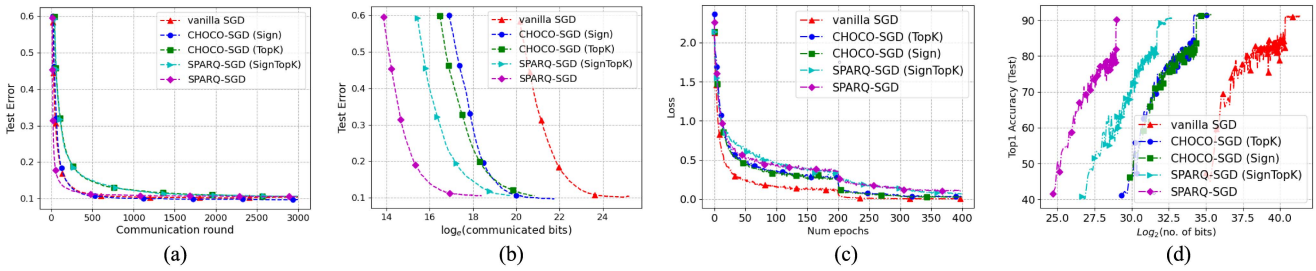

Fig. 1. (a) and (b) For convex objective showing test error vs. number of communication rounds and test error vs. total number of bits communicated, respectively, for different algorithms. (c) and (d) For nonconvex objective showing loss vs. epochs and Top-1 accuracy vs. total number of bits communicated, respectively.

VII. Experiments

In this section, we provide experimental results for our algorithm SPARQ-SGD and comparison with CHOCO-SGD [3], [17], which is the current state-of-the-art in efficient decentralized training, and vanilla decentralized training [1]. We compare these algorithms in terms of communication efficiency and training time over bandlimited links, and further perform an ablation study for the individual techniques comprising SPARQ-SGD. Additional experiments to support our argument in Remark 1 can be found in the extended article [39].

A. Communication Efficiency

We first show advantage of our algorithm in communication efficiency to reach a target test accuracy for both convex and nonconvex objectives. We compare SPARQ-SGD with CHOCO-SGD [3], [17] and vanilla decentralized SGD [1].

1) Convex Objectives: To simulate a convex objective, we consider the multiclass cross-entropy loss to model the local objectives $f_i, i \in [n]$ on the MNIST dataset [44]. We consider $n = 60$ nodes connected in a ring topology, each processing a mini-batch size of 50 per iteration and having heterogeneous distribution of data across classes. We work with learning rate $\eta_t = 1/(t + 10)$ and local iterations $H = 5$ for SPARQ-SGD. For compression, we use the composed operator $SignTopK$ [13] with top $k = 1\%$ (70 out of 7840 length parameter vector for MNIST dataset). For our experiments, we initially set the triggering constant $c_0 = 130$ in SPARQ-SGD (line 6) and keep it unchanged until a certain number of iterations and then increase it periodically under assumptions of Theorem 1; this is to prevent all the workers satisfying the triggering criterion in later iterations, as η_t eventually becomes very small. We also provide a plot for using the composed $SignTopK$ operator without local iterations or threshold triggering titled "SPARQ-SGD (Sign-TopK)" for comparison.

a) Results: We use $SignTopK$ compression in SPARQ-SGD and compare its performance against CHOCO-SGD. In Fig. 1(a), we observe that on account of using multiple SGD iterations, SPARQ-SGD can reach a target test error in fewer communication rounds while converging at a similar rate to vanilla SGD. The advantage to SPARQ-SGD comes from the significant savings in the number of bits communicated to achieve a desired test error, as seen in Fig. 1(b): To achieve a test error of around 0.1, SPARQ-SGD gets $100\times$ savings as compared to CHOCO-SGD with $Sign$ quantizer, around $40\times$ savings than CHOCO-SGD with $TopK$ sparsifier, and around $1000\times$ savings than vanilla decentralized SGD.

2) Nonconvex Objectives: We match the setting in CHOCO-SGD and perform our experiments on the CIFAR-10 [38] dataset and train a ResNet-20 [45] model with $n = 8$ nodes connected in a ring topology. Learning rate is initialized to 0.1, following a schedule consisting of a warmup period of 5 epochs followed by piecewise decay of 5 at epoch 200 and 300 and we stop training at epoch 400. The SGD algorithm is implemented with momentum with a factor of 0.9 and minibatch size of 256. SPARQ-SGD consists of $H = 5$ local iterations followed by checking for a triggering condition, and then communicating with the composed SignTopK operator, where we take top 1% elements of each tensor and only transmit the sign and norm of the result. The triggering threshold follows a schedule piecewise constant: initialized to 2.5 and increases by 1.5 after every 20 epochs till 350 epochs are complete; while maintaining that $c_t < {}^1/_\eta$ for all t. We compare performance of SPARQ-SGD against CHOCO-SGD with $Sign$, $TopK$ compression (taking top 1% of elements of the tensor) and decentralized vanilla SGD [1].

b) Results: We plot the global loss function evaluated at the average parameter vector across nodes in Fig. 1(c), where we observe SPARQ-SGD converging at a similar rate as CHOCO-SGD and vanilla decentralized SGD. Fig. 1(d) shows the performance for a given bit-budget, where we show the Top-1 test accuracy as a function of the total number of bits communicated. For target Top-1 test-accuracy of around 90%, SPARQ-SGD requires about $40\times$ less bits than CHOCO-SGD with $Sign$ or $TopK$ compression, and around $3K\times$ less bits than vanilla decentralized SGD.

B. Training Over Bandlimited Links

We now demonstrate that SPARQ-SGD is particularly suited for training over bandlimited links, where the communication time between nodes can be a bottleneck. For the following results, we consider the same setup as the convex setting described in Section VII-A1.

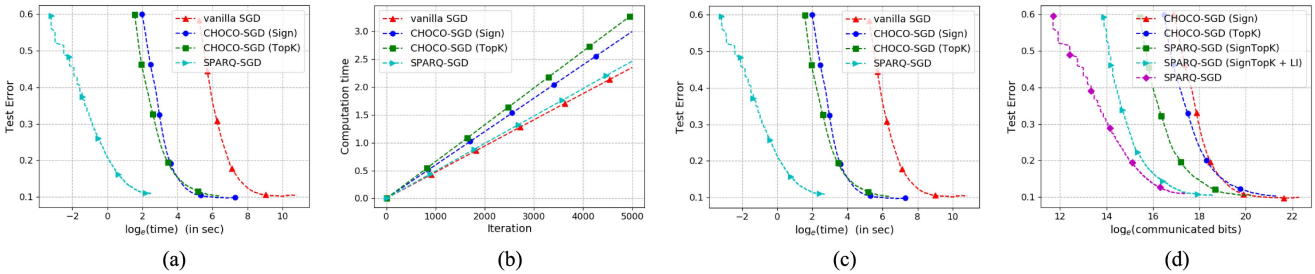

Fig. 2. (a)–(c) Demonstration of the time taken for each scheme when training over bandlimited links. (a) Comparison of test error vs. time taken when communicating over a rate pipe limited to 100 Kbps. (b) Computation time (in seconds) for a single worker as a function of iterations in the algorithms. (c) Total time (computation and communication) for all the schemes for the same number of iterations. (d) An ablation study for the different techniques comprising SPARQ-SGD, where we compare test error as a function of communication bits.

Fig. 1(b) shows the total communicated bits to achieve desired test accuracy. We remark that communicating large amount of data over rate limited channels can pose a significant bottleneck. To demonstrate this, we assume that the communication between nodes is rate-limited to 100 Kbps. This is typical average rate of wireless edge devices sharing a common bandwidth with other devices, therefore devices cannot have sustained high rates. Fig. 2(a) shows the time taken for a single worker (as communication with neighbors happens in parallel) to communicate over 100-Kbps rate links. This provides a comparison for communication time of each scheme over bandlimited links to achieve desired test accuracy. We will now include the computation time for each scheme. Fig. 2(b) shows the time taken for computation (performing gradient evaluation, compression and setting up triggering if required) for each node during training, and is lowest for SGD as expected. We note that SPARQ-SGD skips local iterations and thus does not require performing compression in every round, which makes it computationally efficient than CHOCO-SGD ($Sign/TopK$) [17]. We finally add the communication time over rate limited links [Fig. 2(a)] and the computation time [Fig. 2(b)] to demonstrate the test error performance with total time taken in Fig. 2(c). We observe that to achieve a test accuracy of around 90% in Fig. 2(c), SPARQ-SGD is about $60\times$ faster in total time than the closest competitor CHOCO-SGD (TopK). Compared to vanilla SGD training, SPARQ-SGD is about $700\times$ faster to achieve the same test accuracy when training over bandlimited networks.

C. Ablation Study

We now demonstrate the individual savings from each technique: compression, local iterations, and threshold-based communication which comprise our algorithm SPARQ-SGD. We consider the same setup as the convex setting described in Section VII-A1. In Fig. 2(d), we plot the test error as a function of total communicated bits on the log scale. All schemes are trained for a total of 5000 iterations (i.e., $T = 5000$). We observe that using only $SignTopK$ compression (without local iterations or triggering) in SPARQ-SGD (SignTopK) saves about $5\times$ total bits compared to the closest competitor [CHOCO-SGD (TopK)]. Including local iterations (without triggering) in SPARQ-SGD (SignTopK+LI) saves about $20\times$ bits compared to CHOCO-SGD (TopK), while further utilizing event-triggering, to comprise our proposed algorithm SPARQ-SGD, saves a factor of $50\times$ in total communicated bits compared to CHOCO-SGD (TopK). This also shows that SPARQ-SGD (employing event-triggered communication) can save about a factor of two in a number of communication rounds to achieve target test accuracy compared to SPARQ-SGD (SignTopK + LI) on account of suppressing, on average, half the total number of nodes in each communication round.

APPENDIX A
SUPPORTING LEMMAS FOR PROOF OF LEMMA 2 AND LEMMA 3

In the main article, we define $e^{(1)}_{I_{(t)}} := \sum_{j=1}^{n} \mathbb{E}\|\bar{\mathbf{x}}^{I_{(t)}} - \mathbf{x}_j^{I_{(t)}}\|^2$, which is the sum of expected difference between local and the average iterates at the synchronization index $I_{(t)}$ and another quantity $e^{(2)}_{I_{(t)}} := \sum_{j=1}^{n} \mathbb{E}\|\hat{\mathbf{x}}_j^{I_{(t+1)}} - \mathbf{x}_j^{I_{(t)}}\|^2$, which is the sum of expected difference between the local iterates and their copies.

In this section, we bound $e^{(1)}_{I_{(t+1)}}$ and $e^{(2)}_{I_{(t+1)}}$ in terms of $e^{(1)}_{I_{(t)}}$ and $e^{(2)}_{I_{(t)}}$, and establish a recurrence relations between them.

Using the matrix notation established in Section IV, we have

$$e^{(1)}_{I_{(t+1)}} = \mathbb{E}\|\mathbf{X}^{I_{(t+1)}} - \bar{\mathbf{X}}^{I_{(t+1)}}\|_F^2$$

$$e^{(2)}_{I_{(t+1)}} = \mathbb{E}\|\mathbf{X}^{I_{(t+1)}} - \hat{\mathbf{X}}^{I_{(t+2)}}\|_F^2.$$

We now state and prove these recurrence results.

Lemma (Restating Lemma 4): Consider the sequence of updates in Algorithm 1 in matrix form (refer IV for the matrix form of Algorithm 1). The expected deviation between the local node parameters $\mathbf{X}^{I_{(t+1)}}$ and the global average parameters $\bar{\mathbf{X}}^{I_{(t+1)}}$ evaluated at some $I_{(t+1)} \in \mathcal{I}_T$ satisfies

$$\mathbb{E}\|\mathbf{X}^{I_{(t+1)}} - \bar{\mathbf{X}}^{I_{(t+1)}}\|_F^2 \leq (1+\alpha_5^{-1})R_1\mathbb{E}\|\mathbf{X}^{I_{(t)}} - \bar{\mathbf{X}}^{I_{(t)}}\|_F^2 + (1+\alpha_5^{-1})R_2\mathbb{E}\|\mathbf{X}^{I_{(t)}} - \hat{\mathbf{X}}^{I_{(t+1)}}\|_F^2 + Q_1\eta^2_{I_{(t+1)}}$$

where $R_1 = (1+\alpha_1)(1-\gamma\delta)^2$, $R_2 = (1+\alpha_1^{-1})\gamma^2\lambda^2$ and $Q_1 = 4H^2nG^2(1+\alpha_5)(R_1+R_2)$. Here $\alpha_1, \alpha_5 > 0$, δ is the spectral gap, H is the synchronization gap, γ is the consensus stepsize, and $\lambda := \|\mathbf{W} - \mathbf{I}\|_2$, where $\mathbf{W}$ is a doubly stochastic mixing matrix.

Proof: Using the definition of $\mathbf{X}^{I_{(t+1)}}$ from Section IV, we have

$$\|\mathbf{X}^{I_{(t+1)}}-\bar{\mathbf{X}}^{I_{(t+1)}}\|_F^2$$
$$=\|\mathbf{Y}^{I_{(t+1)}}-\bar{\mathbf{X}}^{I_{(t+1)}}+\gamma\hat{\mathbf{X}}^{I_{(t+1)}}(\mathbf{W}-\mathbf{I})\|_F^2.$$

Noting that $\bar{\mathbf{X}}^{I_{(t+1)}}=\bar{\mathbf{Y}}^{I_{(t+1)}}$ from (4) and $\bar{\mathbf{Y}}^{I_{(t+1)}}(\mathbf{W}-\mathbf{I})=0$ from (3), we get

$$\|\mathbf{X}^{I_{(t+1)}}-\bar{\mathbf{X}}^{I_{(t+1)}}\|_F^2$$
$$=\|(\mathbf{Y}^{I_{(t+1)}}-\bar{\mathbf{Y}}^{I_{(t+1)}})((1-\gamma)\mathbf{I}+\gamma\mathbf{W})$$
$$+\gamma(\hat{\mathbf{X}}^{I_{(t+1)}}-\mathbf{Y}^{I_{(t+1)}})(\mathbf{W}-\mathbf{I})\|_F^2.$$

Using the fact $\|\mathbf{A}+\mathbf{B}\|_F^2\le(1+\alpha_1)\|\mathbf{A}\|_F^2+(1+\alpha_1^{-1})\|\mathbf{B}\|_F^2$ for any $\alpha_1>0$,

$$\|\mathbf{X}^{I_{(t+1)}}-\bar{\mathbf{X}}^{I_{(t+1)}}\|_F^2\le$$
$$(1+\alpha_1)\|(\mathbf{Y}^{I_{(t+1)}}-\bar{\mathbf{Y}}^{I_{(t+1)}})((1-\gamma)\mathbf{I}+\gamma\mathbf{W})\|_F^2$$
$$+(1+\alpha_1^{-1})\|\gamma(\hat{\mathbf{X}}^{I_{(t+1)}}-\mathbf{Y}^{I_{(t+1)}})(\mathbf{W}-\mathbf{I})\|_F^2.$$

Using $\|\mathbf{AB}\|_F\le\|\mathbf{A}\|_F\|\mathbf{B}\|_2$ as per (5), we have

$$\|\mathbf{X}^{I_{(t+1)}}-\bar{\mathbf{X}}^{I_{(t+1)}}\|_F^2\le$$
$$(1+\alpha_1)\|(\mathbf{Y}^{I_{(t+1)}}-\bar{\mathbf{Y}}^{I_{(t+1)}})((1-\gamma)\mathbf{I}+\gamma\mathbf{W})\|_F^2$$
$$+(1+\alpha_1^{-1})\gamma^2\|(\hat{\mathbf{X}}^{I_{(t+1)}}-\mathbf{Y}^{I_{(t+1)}})\|_F^2.\|(\mathbf{W}-\mathbf{I})\|_2^2. \quad (25)$$

To bound the first term in (25), we use the triangle inequality for Frobenius norm, giving us

$$\|(\mathbf{Y}^{I_{(t+1)}}-\bar{\mathbf{Y}}^{I_{(t+1)}})((1-\gamma)\mathbf{I}+\gamma\mathbf{W})\|_F\le$$
$$(1-\gamma)\|\mathbf{Y}^{I_{(t+1)}}-\bar{\mathbf{Y}}^{I_{(t+1)}}\|_F+\gamma\|(\mathbf{Y}^{I_{(t+1)}}-\bar{\mathbf{Y}}^{I_{(t+1)}})W\|_F.$$

It follows from (3) that $(\mathbf{Y}^{I_{(t+1)}}-\bar{\mathbf{Y}}^{I_{(t+1)}})\frac{1}{n}\mathbf{1}\mathbf{1}^T=\mathbf{0}$. Adding this inside the last term above, we get

$$\|(\mathbf{Y}^{I_{(t+1)}}-\bar{\mathbf{Y}}^{I_{(t+1)}})((1-\gamma)\mathbf{I}+\gamma\mathbf{W})\|_F\le$$
$$(1-\gamma)\|\mathbf{Y}^{I_{(t+1)}}-\bar{\mathbf{Y}}^{I_{(t+1)}}\|_F$$
$$+\gamma\left\|(\mathbf{Y}^{I_{(t+1)}}-\bar{\mathbf{Y}}^{I_{(t+1)}})\left(\mathbf{W}-\frac{1}{n}\mathbf{1}\mathbf{1}^T\right)\right\|_F.$$

Using $\|\mathbf{AB}\|_F\le\|\mathbf{A}\|_F\|\mathbf{B}\|_2$ from (5) and then using (7) with $k=1$, we can simplify the above to

$$\|(\mathbf{Y}^{I_{(t+1)}}-\bar{\mathbf{Y}}^{I_{(t+1)}})((1-\gamma)\mathbf{I}+\gamma\mathbf{W})\|_F\le$$
$$(1-\gamma\delta)\|\mathbf{Y}^{I_{(t+1)}}-\bar{\mathbf{Y}}^{I_{(t+1)}}\|_F.$$

Substituting the above in (25) and using $\lambda=\max_i\{1-\lambda_i(\mathbf{W})\}\Rightarrow\|\mathbf{W}-\mathbf{I}\|_2^2\le\lambda^2$, we get

$$\|\mathbf{X}^{I_{(t+1)}}-\bar{\mathbf{X}}^{I_{(t+1)}}\|_F^2\le$$
$$(1+\alpha_1)(1-\gamma\delta)^2\|\mathbf{Y}^{I_{(t+1)}}-\bar{\mathbf{Y}}^{I_{(t+1)}}\|_F^2$$
$$+(1+\alpha_1^{-1})\gamma^2\lambda^2\|\mathbf{Y}^{I_{(t+1)}}-\hat{\mathbf{X}}^{I_{(t+1)}}\|_F^2.$$

Taking expectation with respect to the entire process, we have

$$\mathbb{E}\|\mathbf{X}^{I_{(t+1)}}-\bar{\mathbf{X}}^{I_{(t+1)}}\|_F^2\le$$
$$(1+\alpha_1)(1-\gamma\delta)^2\mathbb{E}\|\mathbf{Y}^{I_{(t+1)}}-\bar{\mathbf{Y}}^{I_{(t+1)}}\|_F^2$$
$$+(1+\alpha_1^{-1})\gamma^2\lambda^2\mathbb{E}\|\mathbf{Y}^{I_{(t+1)}}-\hat{\mathbf{X}}^{I_{(t+1)}}\|_F^2.$$

Define $R_1=(1+\alpha_1)(1-\gamma\delta)^2, R_2=(1+\alpha_1^{-1})\gamma^2\lambda^2$. Using the update algorithm, we have

$$\mathbb{E}\|\mathbf{X}^{I_{(t+1)}}-\bar{\mathbf{X}}^{I_{(t+1)}}\|_F^2\le$$
$$R_1\mathbb{E}\left\|\bar{\mathbf{X}}^{I_{(t)}}-\mathbf{X}^{I_{(t)}}-\sum_{t'=I_{(t)}}^{I_{(t+1)}-1}\eta_{t'}\nabla\mathbf{F}(\mathbf{X}^{(t')},\xi^{(t')})\left(\frac{1}{n}\mathbf{1}\mathbf{1}^T-\mathbf{I}\right)\right\|_F^2$$
$$+R_2\mathbb{E}\left\|\hat{\mathbf{X}}^{I_{(t+1)}}-\mathbf{X}^{I_{(t)}}+\sum_{t'=I_{(t)}}^{I_{(t+1)}-1}\eta_{t'}\nabla\mathbf{F}(\mathbf{X}^{(t')},\xi^{(t')})\right\|_F^2.$$

Thus, for any $\alpha_5>0$, we have

$$\mathbb{E}\|\mathbf{X}^{I_{(t+1)}}-\bar{\mathbf{X}}^{I_{(t+1)}}\|_F^2\le R_1(1+\alpha_5^{-1})\mathbb{E}\left\|\bar{\mathbf{X}}^{I_{(t)}}-\mathbf{X}^{I_{(t)}}\right\|^2$$
$$+R_1(1+\alpha_5)\mathbb{E}\left\|\sum_{t'=I_{(t)}}^{I_{(t+1)}-1}\eta_{t'}\nabla\mathbf{F}(\mathbf{X}^{(t')},\xi^{(t')})\left(\frac{1}{n}\mathbf{1}\mathbf{1}^T-\mathbf{I}\right)\right\|_F^2$$
$$+R_2(1+\alpha_5)\mathbb{E}\left\|\sum_{t'=I_{(t)}}^{I_{(t+1)}-1}\eta_{t'}\nabla\mathbf{F}(\mathbf{X}^{(t')},\xi^{(t')})\right\|_F^2$$
$$+R_2(1+\alpha_5^{-1})\mathbb{E}\left\|\hat{\mathbf{X}}^{I_{(t+1)}}-\mathbf{X}^{I_{(t)}}\right\|^2.$$

Using $\|AB\|_F\le\|A\|_F\|B\|_2$ from (5) to split the third term, and $\|\frac{1}{n}\mathbf{1}\mathbf{1}^T-\mathbf{I}\|_2=1$, and further using the bound in (8) for the third and fourth terms, the above can be rewritten as

$$\mathbb{E}\|\mathbf{X}^{I_{(t+1)}}-\bar{\mathbf{X}}^{I_{(t+1)}}\|_F^2\le R_1(1+\alpha_5^{-1})\mathbb{E}\left\|\bar{\mathbf{X}}^{I_{(t)}}-\mathbf{X}^{I_{(t)}}\right\|^2$$
$$+R_2(1+\alpha_5^{-1})\mathbb{E}\left\|\hat{\mathbf{X}}^{I_{(t+1)}}-\mathbf{X}^{I_{(t)}}\right\|^2$$
$$+\eta_{I_{(t)}}^2H^2nG^2(1+\alpha_5)(R_1+R_2).$$

Noting that $\eta_{I_{(t)}}\le2\eta_{I_{(t+1)}}$[9] and defining $Q_1=4H^2nG^2(1+\alpha_5)(R_1+R_2)$ completes the proof of Lemma 4. □

Lemma (Restating Lemma 5): Consider the sequence of updates in Algorithm 1 in matrix form (refer IV) with the threshold sequence $\{c_t\}_{t=0}^{T-1}$ such that $c_t=o(t)$, and decaying learning rate $\eta_t=\frac{b}{(a+t)}$, for some $b>0$. The expected deviation between the local node parameters $\mathbf{X}^{I_{(t+1)}}$ and their estimates $\hat{\mathbf{X}}^{I_{(t+2)}}$ at a synchronization time step $I_{(t+1)}$ satisfies

$$\mathbb{E}\|\mathbf{X}^{I_{(t+1)}}-\hat{\mathbf{X}}^{I_{(t+2)}}\|_F^2\le(1+\alpha_5^{-1})R_3\mathbb{E}\|\mathbf{X}^{I_{(t)}}-\hat{\mathbf{X}}^{I_{(t+1)}}\|_F^2$$
$$+(1+\alpha_5^{-1})R_4\mathbb{E}\|\mathbf{X}^{I_{(t)}}-\bar{\mathbf{X}}^{I_{(t)}}\|_F^2+\eta_{I_{(t+1)}}^2Q_2$$

where $R_3=(1+\gamma\lambda)^2(1+\alpha_4)(1+\alpha_3)(1+\alpha_2)(1-\omega)$, $R_4=\gamma^2\lambda^2(1+\alpha_4^{-1})(1+\alpha_3)(1+\alpha_2)(1-\omega)$ and $Q_2=4H^2nG^2((1+\alpha_5)(R_3+R_4)+(1+\alpha_2^{-1})+(1+\alpha_3^{-1})(1+\alpha_2)(1-\omega))+(1+\alpha_2)\omega nc_{I_{(t+1)}}$. Here $\alpha_2,\alpha_3,\alpha_4>0,\alpha_5>0$ are the same as used in Lemma 4, δ is the spectral gap, H is the synchronization gap, γ is the consensus stepsize, and $\lambda=\|\mathbf{W}-\mathbf{I}\|_2$, where $\mathbf{W}$ is a doubly stochastic mixing matrix.

Note that in the above expression, Q_2 depends on t (as captured by $c_{I_{(t)}}$ in the expression) as we allow for our triggering threshold to change with time.

[9] $\frac{\eta_{I_{(t)}}}{\eta_{I_{(t+1)}}}=\frac{I_{(t+1)}+a}{I_{(t)}+a}\le\frac{I_{(t)}+H+a}{I_{(t)}+a}=1+\frac{H}{I_{(t)}+a}\le1+\frac{H}{a}\le2$. The last inequality follows from that $a\ge\frac{5H}{p}\ge H$.

Proof: Note that $\hat{\mathbf{X}}^{I_{(t+2)}} = \hat{\mathbf{X}}^{I_{(t+1)}} + \mathcal{C}((\mathbf{Y}^{I_{(t+2)}} - \hat{\mathbf{X}}^{I_{(t+1)}})\mathbf{P}^{(I_{(t+2)}-1)})$; see Section IV in main article. Substituting this in $\mathbb{E}\|\mathbf{X}^{I_{(t+1)}} - \hat{\mathbf{X}}^{I_{(t+2)}}\|_F^2$, we get

$$\mathbb{E}\|\mathbf{X}^{I_{(t+1)}} - \hat{\mathbf{X}}^{I_{(t+2)}}\|_F^2 = \mathbb{E}\|\mathbf{X}^{I_{(t+1)}} - \hat{\mathbf{X}}^{I_{(t+1)}} - \mathcal{C}((\mathbf{Y}^{I_{(t+2)}} - \hat{\mathbf{X}}^{I_{(t+1)}})\mathbf{P}^{(I_{(t+2)}-1)})\|_F^2.$$

Adding and subtracting $\mathbf{Y}^{I_{(t+2)}}$ and rearranging terms

$$\mathbb{E}\|\mathbf{X}^{I_{(t+1)}} - \hat{\mathbf{X}}^{I_{(t+2)}}\|_F^2 = \mathbb{E}\|\mathbf{Y}^{I_{(t+2)}} - \hat{\mathbf{X}}^{I_{(t+1)}} + \mathbf{X}^{I_{(t+1)}} - \mathbf{Y}^{I_{(t+2)}} - \mathcal{C}((\mathbf{Y}^{I_{(t+2)}} - \hat{\mathbf{X}}^{I_{(t+1)}})\mathbf{P}^{(I_{(t+2)}-1)})\|_F^2.$$

Using $\|\mathbf{A}+\mathbf{B}\|_F^2 \le (1+\alpha_2)\|\mathbf{A}\|_F^2 + (1+\alpha_2^{-1})\|\mathbf{B}\|_F^2$ for any $\alpha_2 > 0$,

$$\begin{aligned}\mathbb{E}\|\mathbf{X}^{I_{(t+1)}} - \hat{\mathbf{X}}^{I_{(t+2)}}\|_F^2 &\le (1+\alpha_2)\mathbb{E}\|\mathbf{Y}^{I_{(t+2)}} - \hat{\mathbf{X}}^{I_{(t+1)}} - \mathcal{C}((\mathbf{Y}^{I_{(t+2)}} - \hat{\mathbf{X}}^{I_{(t+1)}})\mathbf{P}^{(I_{(t+2)}-1)})\|_F^2 \\ &\quad + (1+\alpha_2^{-1})\mathbb{E}\|\mathbf{X}^{I_{(t+1)}} - \mathbf{Y}^{I_{(t+2)}}\|_F^2 \\ &= (1+\alpha_2)\mathbb{E}\|\mathbf{Y}^{I_{(t+2)}} - \hat{\mathbf{X}}^{I_{(t+1)}} - \mathcal{C}((\mathbf{Y}^{I_{(t+2)}} - \hat{\mathbf{X}}^{I_{(t+1)}})\mathbf{P}^{(I_{(t+2)}-1)})\|_F^2 \\ &\quad + (1+\alpha_2^{-1})\mathbb{E}\left\|\sum_{t'=I_{(t+1)}}^{I_{(t+2)}-1}\eta_{t'}\nabla\mathbf{F}(\mathbf{X}^{t'},\xi^{t'})\right\|_F^2. \quad (26)\end{aligned}$$

Bounding the last term in (26) using (8), we get

$$\begin{aligned}\mathbb{E}\|\mathbf{X}^{I_{(t+1)}} - \hat{\mathbf{X}}^{I_{(t+2)}}\|_F^2 &\le (1+\alpha_2)\mathbb{E}\|\mathbf{Y}^{I_{(t+2)}} - \hat{\mathbf{X}}^{I_{(t+1)}} \\ &\quad -\mathcal{C}((\mathbf{Y}^{I_{(t+2)}} - \hat{\mathbf{X}}^{I_{(t+1)}})\mathbf{P}^{(I_{(t+2)}-1)})\|_F^2 \\ &\quad +(1+\alpha_2^{-1})\eta_{I_{(t+1)}}^2 H^2 nG^2.\end{aligned}$$

Note that both $\mathbf{P}^{(I_{(t+2)}-1)}$ and $\mathbf{I} - \mathbf{P}^{(I_{(t+2)}-1)}$ are diagonal matrices, with disjoint support on the diagonal entries, which implies that $\mathbb{E}\|\mathbf{Y}^{I_{(t+2)}} - \hat{\mathbf{X}}^{I_{(t+1)}}\|_F^2 = \mathbb{E}\|(\mathbf{Y}^{I_{(t+2)}} - \hat{\mathbf{X}}^{I_{(t+1)}})\mathbf{P}^{(I_{(t+2)}-1)}\|_F^2 + \mathbb{E}\|(\mathbf{Y}^{I_{(t+2)}} - \hat{\mathbf{X}}^{I_{(t+1)}})(\mathbf{I} - \mathbf{P}^{(I_{(t+2)}-1)})\|_F^2$. We get

$$\begin{aligned}&\mathbb{E}\|\mathbf{X}^{I_{(t+1)}} - \hat{\mathbf{X}}^{I_{(t+2)}}\|_F^2 \\ &\le (1+\alpha_2)\mathbb{E}\|(\mathbf{Y}^{I_{(t+2)}} - \hat{\mathbf{X}}^{I_{(t+1)}})\mathbf{P}^{(I_{(t+2)}-1)} - \mathcal{C}((\mathbf{Y}^{I_{(t+2)}} - \hat{\mathbf{X}}^{I_{(t+1)}})\mathbf{P}^{(I_{(t+2)}-1)})\|_F^2 \\ &\quad + (1+\alpha_2)\mathbb{E}\|(\mathbf{Y}^{I_{(t+2)}} - \hat{\mathbf{X}}^{I_{(t+1)}})(\mathbf{I} - \mathbf{P}^{(I_{(t+2)}-1)})\|_F^2 \\ &\quad + (1+\alpha_2^{-1})\eta_{I_{(t+1)}}^2 H^2 nG^2.\end{aligned}$$

Using the compression property of operator $\mathcal{C}$ as per (2), we have

$$\begin{aligned}\mathbb{E}\|\mathbf{X}^{I_{(t+1)}} - \hat{\mathbf{X}}^{I_{(t+2)}}\|_F^2 &\le (1+\alpha_2^{-1})\eta_{I_{(t+1)}}^2 H^2 nG^2 \\ &\quad + (1+\alpha_2)(1-\omega)\mathbb{E}\|(\mathbf{Y}^{I_{(t+2)}} - \hat{\mathbf{X}}^{I_{(t+1)}})\mathbf{P}^{(I_{(t+2)}-1)}\|_F^2 \\ &\quad + (1+\alpha_2)\mathbb{E}\|(\mathbf{Y}^{I_{(t+2)}} - \hat{\mathbf{X}}^{I_{(t+1)}})(\mathbf{I} - \mathbf{P}^{(I_{(t+2)}-1)})\|_F^2.\end{aligned}$$

Adding and subtracting $(1+\alpha_2)(1-\omega)\mathbb{E}\|(\mathbf{Y}^{I_{(t+2)}} - \hat{\mathbf{X}}^{I_{(t+1)}})(\mathbf{I} - \mathbf{P}^{(I_{(t+2)}-1)})\|_F^2$, we get

$$\begin{aligned}\mathbb{E}\|\mathbf{X}^{I_{(t+1)}} - \hat{\mathbf{X}}^{I_{(t+2)}}\|_F^2 &\le (1+\alpha_2^{-1})\eta_{I_{(t+1)}}^2 H^2 nG^2 \\ &\quad + (1+\alpha_2)(1-\omega)\mathbb{E}\|\mathbf{Y}^{I_{(t+2)}} - \hat{\mathbf{X}}^{I_{(t+1)}}\|_F^2 \\ &\quad + (1+\alpha_2)\omega\mathbb{E}\|(\mathbf{Y}^{I_{(t+2)}} - \hat{\mathbf{X}}^{I_{(t+1)}})(\mathbf{I} - \mathbf{P}^{(I_{(t+2)}-1)})\|_F^2.\end{aligned}$$

To bound the third term in the RHS above, note that $\hat{\mathbf{X}}^{I_{(t+2)}-1} = \hat{\mathbf{X}}^{I_{(t+1)}}$, because $\hat{\mathbf{X}}$ does not change in between the synchronization indices, which implies that $\mathbb{E}\|(\mathbf{Y}^{I_{(t+2)}} - \hat{\mathbf{X}}^{I_{(t+1)}})(\mathbf{I} - \mathbf{P}^{(I_{(t+2)}-1)})\|_F^2 = \mathbb{E}\|(\mathbf{Y}^{I_{(t+2)}} - \hat{\mathbf{X}}^{I_{(t+2)}-1})(\mathbf{I} - \mathbf{P}^{(I_{(t+2)}-1)})\|_F^2$, which we can upper bound using (6) by $nc_{I_{(t+2)}-1}\eta_{I_{(t+2)}-1}^2$. From Claim 2 provided in [39], it follows that $c_{I_{(t+2)}-1}\eta_{I_{(t+2)}-1}^2 \le 4c_{I_{(t+1)}}\eta_{I_{(t+1)}}^2$. Substituting this in the above gives

$$\begin{aligned}&\mathbb{E}\|\mathbf{X}^{I_{(t+1)}} - \hat{\mathbf{X}}^{I_{(t+2)}}\|_F^2 \\ &\le (1+\alpha_2)(1-\omega)\mathbb{E}\|\mathbf{Y}^{I_{(t+2)}} - \hat{\mathbf{X}}^{I_{(t+1)}}\|_F^2 \\ &\quad + (1+\alpha_2)\omega n4c_{I_{(t+1)}}\eta_{I_{(t+1)}}^2 + (1+\alpha_2^{-1})\eta_{I_{(t+1)}}^2 H^2 nG^2 \\ &= (1+\alpha_2)(1-\omega)\mathbb{E}\|\mathbf{X}^{I_{(t+1)}} - \sum_{t'=I_{(t+1)}}^{I_{(t+2)}-1}\eta_{t'}\nabla\mathbf{F}(\mathbf{X}^{t'},\xi^{t'}) - \hat{\mathbf{X}}^{I_{(t+1)}}\|_F^2 \\ &\quad + (1+\alpha_2^{-1})\eta_{I_{(t+1)}}^2 H^2 nG^2 + (1+\alpha_2)\omega n4c_{I_{(t+1)}}\eta_{I_{(t+1)}}^2 \\ &\le (1+\alpha_3)(1+\alpha_2)(1-\omega)\mathbb{E}\|\mathbf{X}^{I_{(t+1)}} - \hat{\mathbf{X}}^{I_{(t+1)}}\|_F^2 \\ &\quad + (1+\alpha_3^{-1})(1+\alpha_2)(1-\omega)\mathbb{E}\|\sum_{t'=I_{(t+1)}}^{I_{(t+2)}-1}\eta_{t'}\nabla\mathbf{F}(\mathbf{X}^{t'},\xi^{t'})\|_F^2 \\ &\quad +(1+\alpha_2^{-1})\eta_{I_{(t+1)}}^2 H^2 nG^2 + (1+\alpha_2)\omega n4c_{I_{(t+1)}}\eta_{I_{(t+1)}}^2 \quad (27)\end{aligned}$$

where in the last inequality, we have used bound on the sum[10] and $\alpha_3 > 0$. Using (8) to bound the penultimate term in (27)

$$\begin{aligned}&\mathbb{E}\|\mathbf{X}^{I_{(t+1)}} - \hat{\mathbf{X}}^{I_{(t+2)}}\|_F^2 \\ &\le (1+\alpha_3)(1+\alpha_2)(1-\omega)\mathbb{E}\|\mathbf{X}^{I_{(t+1)}} - \hat{\mathbf{X}}^{I_{(t+1)}}\|_F^2 \\ &\quad + (1+\alpha_3^{-1})(1+\alpha_2)(1-\omega)\eta_{I_{(t+1)}}^2 H^2 nG^2 \\ &\quad + (1+\alpha_2)\omega n4c_{I_{(t+1)}}\eta_{I_{(t+1)}}^2 + (1+\alpha_2^{-1})\eta_{I_{(t+1)}}^2 H^2 nG^2 \\ &= (1+\alpha_3)(1+\alpha_2)(1-\omega)\mathbb{E}\|\mathbf{Y}^{I_{(t+1)}} + \gamma\hat{\mathbf{X}}^{I_{(t+1)}}(\mathbf{W}-\mathbf{I}) - \hat{\mathbf{X}}^{I_{(t+1)}}\|_F^2 \\ &\quad + (1+\alpha_3^{-1})(1+\alpha_2)(1-\omega)\eta_{I_{(t+1)}}^2 H^2 nG^2 \\ &\quad + (1+\alpha_2)\omega n4c_{I_{(t+1)}}\eta_{I_{(t+1)}}^2 + (1+\alpha_2^{-1})\eta_{I_{(t+1)}}^2 H^2 nG^2 \\ &= (1+\alpha_3)(1+\alpha_2)(1-\omega)\mathbb{E}\|(\mathbf{Y}^{I_{(t+1)}} - \hat{\mathbf{X}}^{I_{(t+1)}}) \\ &\quad \times((1+\gamma)\mathbf{I} - \gamma\mathbf{W}) + \gamma(\mathbf{Y}^{I_{(t+1)}} - \bar{\mathbf{Y}}^{I_{(t+1)}})(\mathbf{W}-\mathbf{I})\|_F^2 \\ &\quad + (1+\alpha_3^{-1})(1+\alpha_2)(1-\omega)\eta_{I_{(t+1)}}^2 H^2 nG^2 \\ &\quad + (1+\alpha_2)\omega n4c_{I_{(t+1)}}\eta_{I_{(t+1)}}^2 + (1+\alpha_2^{-1})\eta_{I_{(t+1)}}^2 H^2 nG^2\end{aligned}$$

where in the last equality, we have used $\bar{\mathbf{Y}}^{I_{(t+1)}}(\mathbf{W}-\mathbf{I}) = 0$. For $\alpha_4 > 0$, using result stated in Footnote 10 gives us

$$\mathbb{E}\|\mathbf{X}^{I_{(t+1)}} - \hat{\mathbf{X}}^{I_{(t+2)}}\|_F^2 \le (1+\alpha_4)(1+\alpha_3)(1+\alpha_2)$$

[10]For any two matrices $\mathbf{A}, \mathbf{B} \in \mathbb{R}^{p\times q}$ and for any $\alpha > 0$, we have

$$\|\mathbf{A}+\mathbf{B}\|_F^2 \le (1+\alpha)\|\mathbf{A}\|_F^2 + (1+\alpha^{-1})\|\mathbf{B}\|_F^2$$

$$\times (1-\omega)\mathbb{E}\|(\mathbf{Y}^{I_{(t+1)}} - \hat{\mathbf{X}}^{I_{(t+1)}})((1+\gamma)\mathbf{I} - \gamma\mathbf{W})\|_F^2$$
$$+ (1+\alpha_4^{-1})(1+\alpha_3)(1+\alpha_2)(1-\omega)$$
$$\times \mathbb{E}\|\gamma(\mathbf{Y}^{I_{(t+1)}} - \bar{\mathbf{Y}}^{I_{(t+1)}})(\mathbf{W}-\mathbf{I})\|_F^2$$
$$+ (1+\alpha_3^{-1})(1+\alpha_2)(1-\omega)\eta_{I_{(t+1)}}^2 H^2 nG^2$$
$$+ (1+\alpha_2)\omega n 4c_{I_{(t+1)}}\eta_{I_{(t+1)}}^2 + (1+\alpha_2^{-1})\eta_{I_{(t+1)}}^2 H^2 nG^2.$$

Using $\|(1+\gamma)\mathbf{I} - \gamma W\|_2 = \|I + \gamma(\mathbf{I}-\mathbf{W})\|_2 = 1+\gamma\|\mathbf{I}-\mathbf{W}\|_2 = 1+\gamma\lambda$ (by definition of $\lambda = \max_i\{1-\lambda_i(\mathbf{W})\}$) and $\|\mathbf{I}-\mathbf{W}\|_2 = \lambda$ along with $\|\mathbf{AB}\|_F \le \|\mathbf{A}\|_F\|\mathbf{B}\|_2$ from (5)

$$\mathbb{E}\|\mathbf{X}^{I_{(t+1)}} - \hat{\mathbf{X}}^{I_{(t+2)}}\|_F^2 \le (1+\gamma\lambda)^2(1+\alpha_4)$$
$$\times (1+\alpha_3)(1+\alpha_2)(1-\omega)\mathbb{E}\|\mathbf{Y}^{I_{(t+1)}} - \hat{\mathbf{X}}^{I_{(t+1)}}\|_F^2$$
$$+ \gamma^2\lambda^2(1+\alpha_4^{-1})(1+\alpha_3)(1+\alpha_2)$$
$$\times (1-\omega)\mathbb{E}\|\mathbf{Y}^{I_{(t+1)}} - \bar{\mathbf{Y}}^{I_{(t+1)}}\|_F^2$$
$$+ \left((1+\alpha_2^{-1}) + (1+\alpha_3^{-1})(1+\alpha_2)(1-\omega)\right)\eta_{I_{(t+1)}}^2 H^2 nG^2$$
$$+ (1+\alpha_2)\omega n 4c_{I_{(t+1)}}\eta_{I_{(t+1)}}^2.$$

Define $R_3 = (1+\gamma\lambda)^2(1+\alpha_4)(1+\alpha_3)(1+\alpha_2)(1-\omega)$, $R_4 = \gamma^2\lambda^2(1+\alpha_4^{-1})(1+\alpha_3)(1+\alpha_2)(1-\omega)$ and $R_5 = ((1+\alpha_2^{-1}) + (1+\alpha_3^{-1})(1+\alpha_2)(1-\omega))H^2nG^2 + (1+\alpha_2)\omega n 4c_{I_{(t+1)}}$, then the above can be rewritten as

$$\mathbb{E}\|\mathbf{X}^{I_{(t+1)}} - \hat{\mathbf{X}}^{I_{(t+2)}}\|_F^2 \le R_3\mathbb{E}\|\mathbf{Y}^{I_{(t+1)}} - \hat{\mathbf{X}}^{I_{(t+1)}}\|_F^2$$
$$+ R_4\mathbb{E}\|\mathbf{Y}^{I_{(t+1)}} - \bar{\mathbf{Y}}^{I_{(t+1)}}\|_F^2 + R_5\eta_{I_{(t+1)}}^2.$$

Using the update rule, this can be expanded as

$$\mathbb{E}\|\mathbf{X}^{I_{(t+1)}} - \hat{\mathbf{X}}^{I_{(t+2)}}\|_F^2 \le$$
$$R_3\mathbb{E}\|\hat{\mathbf{X}}^{I_{(t+1)}} - \mathbf{X}^{I_{(t)}} + \sum_{t'=I_{(t)}}^{I_{(t+1)}-1}\eta_{t'}\nabla\mathbf{F}(\mathbf{X}^{(t')}, \xi^{(t')})\|_F^2 + R_5\eta_{I_{(t+1)}}^2$$
$$+ R_4\mathbb{E}\left\|\bar{\mathbf{X}}^{I_{(t)}} - \mathbf{X}^{I_{(t)}} - \sum_{t'=I_{(t)}}^{I_{(t+1)}-1}\eta_{t'}\nabla\mathbf{F}(\mathbf{X}^{(t')}, \xi^{(t')})\left(\frac{\mathbf{1}\mathbf{1}^T}{n} - I\right)\right\|_F^2.$$

For the same $\alpha_5 > 0$ as used in Lemma 4, we get

$$\mathbb{E}\|\mathbf{X}^{I_{(t+1)}} - \hat{\mathbf{X}}^{I_{(t+2)}}\|_F^2 \le R_3(1+\alpha_5^{-1})\mathbb{E}\left\|\hat{\mathbf{X}}^{I_{(t+1)}} - \mathbf{X}^{I_{(t)}}\right\|^2$$
$$+ R_4(1+\alpha_5^{-1})\mathbb{E}\left\|\bar{\mathbf{X}}^{I_{(t)}} - \mathbf{X}^{I_{(t)}}\right\|^2$$
$$+ R_4(1+\alpha_5)\mathbb{E}\left\|\sum_{t'=I_{(t)}}^{I_{(t+1)}-1}\eta_{t'}\nabla\mathbf{F}(\mathbf{X}^{(t')}, \xi^{(t')})\left(\frac{1}{n}\mathbf{1}\mathbf{1}^T - \mathbf{I}\right)\right\|_F^2$$
$$+ R_3(1+\alpha_5)\mathbb{E}\left\|\sum_{t'=I_{(t)}}^{I_{(t+1)}-1}\eta_{t'}\nabla\mathbf{F}(\mathbf{X}^{(t')}, \xi^{(t')})\right\|_F^2 + R_5\eta_{I_{(t+1)}}^2.$$

Using $\|\mathbf{AB}\|_F \le \|\mathbf{A}\|_F\|\mathbf{B}\|_2$ as per (5) to split the third term and then using $\|\frac{1}{n}\mathbf{1}\mathbf{1}^T - \mathbf{I}\| \le 1$, and further using the bound in (8) for the third and fourth terms, the above can be rewritten as

$$\mathbb{E}\|\mathbf{X}^{I_{(t+1)}} - \bar{\mathbf{X}}^{I_{(t+2)}}\|_F^2 \le R_3(1+\alpha_5^{-1})\mathbb{E}\left\|\hat{\mathbf{X}}^{I_{(t+1)}} - \mathbf{X}^{I_{(t)}}\right\|^2$$
$$+ R_4(1+\alpha_5^{-1})\mathbb{E}\left\|\bar{\mathbf{X}}^{I_{(t)}} - \mathbf{X}^{I_{(t)}}\right\|^2$$
$$+ \eta_{I_{(t)}}^2 H^2 nG^2(1+\alpha_5)(R_3+R_4) + R_5\eta_{I_{(t+1)}}^2.$$

Noting that $\eta_{I_{(t)}} \le 2\eta_{I_{(t+1)}}$ (see Footnote 9) and defining $Q_2 = 4H^2nG^2((1+\alpha_5)(R_3+R_4) + (1+\alpha_2^{-1}) + (1+\alpha_3^{-1})(1+\alpha_2)(1-\omega)) + 4(1+\alpha_2)\omega n c_{I_{(t+1)}} \ge 4H^2nG^2(1+\alpha_5)(R_3+R_4) + R_5$ completes the proof of Lemma 5. □

ACKNOWLEDGMENT

Views and conclusions contained in this document are those of the authors and should not be interpreted as representing the official policies, either expressed or implied, of Army Research Laboratory or the U.S. Government. The U.S. Government is authorized to reproduce and distribute reprints for Government purposes notwithstanding any copyright notation here on.

REFERENCES

[1] X. Lian, C. Zhang, H. Zhang, C.-J. Hsieh, W. Zhang, and J. Liu, "Can decentralized algorithms outperform centralized algorithms? A case study for decentralized parallel stochastic gradient descent," in *Proc. Neural Inf. Process. Syst.*, 2017, pp. 5330–5340.

[2] H. Tang, S. Gan, C. Zhang, T. Zhang, and J. Liu, "Communication compression for decentralized training," in *Proc. Neural Inf. Process. Syst.*, 2018, pp. 7663–7673.

[3] A. Koloskova, S. U. Stich, and M. Jaggi, "Decentralized stochastic optimization and gossip algorithms with compressed communication," in *Proc. Int. Conf. Mach. Learn.*, 2019, pp. 3478–3487.

[4] N. Strom, "Scalable distributed DNN training using commodity GPU cloud computing," in *Proc. Int. Speech Commun. Assoc. INTERSPEECH*, 2015, pp. 1488–1492.

[5] A. F. Aji and K. Heafield, "Sparse communication for distributed gradient descent," in *Proc. Empirical Methods Natural Lang. Process.*, 2017, pp. 440–445.

[6] Y. Lin, S. Han, H. Mao, Y. Wang, and W. J. Dally, "Deep gradient compression: Reducing the communication bandwidth for distributed training," in *Proc. Int. Conf. Learn. Representations*, 2018.

[7] S. U. Stich, J.-B. Cordonnier, and M. Jaggi, "Sparsified SGD with memory," in *Proc. Neural Inf. Process. Syst.*, 2018, pp. 4447–4458.

[8] D. Alistarh, T. Hoefler, M. Johansson, N. Konstantinov, S. Khirirat, and C. Renggli, "The convergence of sparsified gradient methods," in *Proc. Neural Inf. Process. Syst.*, 2018, pp. 5973–5983.

[9] D. Alistarh, D. Grubic, J. Li, R. Tomioka, and M. Vojnovic, "QSGD: Communication-efficient SGD via gradient quantization and encoding," in *Proc. Neural Inf. Process. Syst.*, 2017, pp. 1709–1720.

[10] W. Wen *et al.*, "Terngrad: Ternary gradients to reduce communication in distributed deep learning," in *Proc. Neural Inf. Process. Syst.*, 2017, pp. 1508–1518.

[11] A. T. Suresh, F. X. Yu, S. Kumar, and H. B. McMahan, "Distributed mean estimation with limited communication," in *Proc. Int. Conf. Mach. Learn.*, 2017, pp. 3329–3337.

[12] S. P. Karimireddy, Q. Rebjock, S. U. Stich, and M. Jaggi, "Error feedback fixes SignSGD and other gradient compression schemes," in *Proc. Int. Conf. Mach. Learn.*, 2019, pp. 3252–3261.

[13] D. Basu, D. Data, C. Karakus, and S. Diggavi, "Qsparse-local-SGD: Distributed SGD with quantization, sparsification and local computations," in *Proc. Neural Inf. Process. Syst.*, 2019, pp. 14 668–14679.

[14] S. U. Stich, "Local SGD converges fast and communicates little," in *Proc. Int. Conf. Learn. Representations*, 2019.

[15] H. Yu, S. Yang, and S. Zhu, "Parallel restarted SGD with faster convergence and less communication: Demystifying why model averaging works for deep learning," in *Proc. AAAI Conf. Artif. Intell.*, 2019, pp. 5693–5700.

[16] G. F. Coppola, "Iterative parameter mixing for distributed large-margin training of structured predictors for natural language processing," Ph.D. dissertation, School of Informatics, Univ. Edinburgh, Edinburgh, U.K., 2015.

[17] A. Koloskova, T. Lin, S. U. Stich, and M. Jaggi, "Decentralized deep learning with arbitrary communication compression," in *Proc. Int. Conf. Learn. Representations*, 2020.
[18] A. Reisizadeh, A. Mokhtari, H. Hassani, and R. Pedarsani, "Quantized decentralized consensus optimization," in *Proc. IEEE Conf. Decis. Control*, 2018, pp. 5838–5843.
[19] M. Assran, N. Loizou, N. Ballas, and M. Rabbat, "Stochastic gradient push for distributed deep learning," in *Proc. Int. Conf. Mach. Learn.*, 2019, pp. 344–353.
[20] T. Tatarenko and B. Touri, "Non-convex distributed optimization," *IEEE Trans. Autom. Control*, vol. 62, no. 8, pp. 3744–3757, Aug. 2017.
[21] W. P. M. H. Heemels, K. H. Johansson, and P. Tabuada, "An introduction to event-triggered and self-triggered control," in *Proc. IEEE Conf. Decis. Control*, 2012, pp. 3270–3285.
[22] D. V. Dimarogonas, E. Frazzoli, and K. H. Johansson, "Distributed event-triggered control for multi-agent systems," *IEEE Trans. Autom. Control*, vol. 57, no. 5, pp. 1291–1297, May 2012.
[23] G. S. Seyboth, D. V. Dimarogonas, and K. H. Johansson, "Event-based broadcasting for multi-agent average consensus," *Automatica*, vol. 49, no. 1, pp. 245–252, 2013.
[24] A. Girard, "Dynamic triggering mechanisms for event-triggered control," *IEEE Trans. Autom. Control*, vol. 60, no. 7, pp. 1992–1997, Jul. 2015.
[25] Y. Liu, C. Nowzari, Z. Tian, and Q. Ling, "Asynchronous periodic event-triggered coordination of multi-agent systems," in *Proc. IEEE Conf. Decis. Control*, 2017, pp. 6696–6701.
[26] S. S. Kia, J. Cortés, and S. Martínez, "Distributed convex optimization via continuous-time coordination algorithms with discrete-time communication," *Automatica*, vol. 55, pp. 254–264, 2015.
[27] W. Chen and W. Ren, "Event-triggered zero-gradient-sum distributed consensus optimization over directed networks," *Automatica*, vol. 65, pp. 90–97, 2016.
[28] W. Du, X. Yi, J. George, K. H. Johansson, and T. Yang, "Distributed optimization with dynamic event-triggered mechanisms," in *Proc. IEEE Conf. Decis. Control*, 2018, pp. 969–974.
[29] H. Li, S. Liu, Y. C. Soh, and L. Xie, "Event-triggered communication and data rate constraint for distributed optimization of multiagent systems," *IEEE Trans. Syst., Man, Cybern.: Syst.*, vol. 48, no. 11, pp. 1908–1919, Nov. 2018.
[30] X. Cao and T. Başar, "Decentralized online convex optimization with event-triggered communications," *IEEE Trans. Signal Process.*, vol. 69, pp. 284–299, Dec. 2020.
[31] Z. Wu, Z. Li, Z. Ding, and Z. Li, "Distributed continuous-time optimization with scalable adaptive event-based mechanisms," *IEEE Trans. Syst., Man, Cybern.: Syst.*, vol. 50, no. 9, pp. 3252–3257, Sep. 2020.
[32] C. Liu, H. Li, and Y. Shi, "Resource-aware exact decentralized optimization using event-triggered broadcasting," *IEEE Trans. Autom. Control*, vol. 66, no. 7, pp. 2961–2974, Jul. 2021.
[33] T. Chen, G. Giannakis, T. Sun, and W. Yin, "LAG: Lazily aggregated gradient for communication-efficient distributed learning," in *Proc. Neural Inf. Process. Syst.*, 2018, pp. 5050–5060.
[34] A. Nedic, A. Olshevsky, A. Ozdaglar, and J. N. Tsitsiklis, "Distributed subgradient methods and quantization effects," in *Proc. IEEE Conf. Decis. Control*, 2008, pp. 4177–4184.
[35] T. T. Doan, S. T. Maguluri, and J. Romberg, "Fast convergence rates of distributed subgradient methods with adaptive quantization," *IEEE Trans. Autom. Cont.*, vol. 66, no. 5, pp. 2191–2205, 2020.
[36] A. Nedich, A. Olshevsky, A. Ozdaglar, and J. N. Tsitsiklis, "On distributed averaging algorithms and quantization effects," *IEEE Trans. Autom. Control*, vol. 54, no. 11, pp. 2506–2517, Nov. 2009.
[37] Y. Pu, M. Zeilinger, and C. N. Jones, "Quantization design for distributed optimization," *IEEE Trans. Autom. Control*, vol. 62, no. 5, pp. 2107–2120, May 2017.
[38] A. Krizhevsky *et al.*, "Learning multiple layers of features from tiny images," Citeseer, 2009. [Online]. Available: http://www.cs.toronto.edu/kriz/cifar.html
[39] N. Singh, D. Data, J. George, and S. Diggavi, "SPARQ-SGD: Event-triggered and compressed communication in decentralized stochastic optimization," 2019, *arXiv:1910.14280*.
[40] A. Rakhlin, O. Shamir, and K. Sridharan, "Making gradient descent optimal for strongly convex stochastic optimization," in *Proc. Int. Conf. Mach. Learn.*, 2012, pp. 1571–1578.
[41] Y. T. Chow, W. Shi, T. Wu, and W. Yin, "Expander graph and communication-efficient decentralized optimization," in *Proc. 50th Asilomar Conf. Signals, Syst. Comput.*, 2016, pp. 1715–1720.
[42] S. Hoory, N. Linial, and A. Wigderson, "Expander graphs and their applications," *Bull. Amer. Math. Soc.*, vol. 43, no. 4, pp. 439–561, 2006.
[43] H. Mania, X. Pan, D. Papailiopoulos, B. Recht, K. Ramchandran, and M. I. Jordan, "Perturbed iterate analysis for asynchronous stochastic optimization," *SIAM J. Optim.*, vol. 27, no. 4, pp. 2202–2229, 2017.
[44] Y. LeCun, C. Cortes, and C. Burges, "MNIST handwritten digit database," *ATT Labs*, vol. 2, 2010.
[45] W. Wen, C. Wu, Y. Wang, Y. Chen, and H. Li, "Learning structured sparsity in deep neural networks," in *Proc. Neural Inf. Process. Syst.*, 2016, pp. 2074–2082.

Navjot Singh (Student Member, IEEE) received the B.Tech. and M.Tech. degrees in electrical engineering from the Indian Institute of Technology, Bombay, India, in 2018. He is currently working toward the Ph.D. degree with the Electrical and Computer Engineering Department, University of California, Los Angeles, CA, USA.

Deepesh Data received the B.Tech. degree in computer science and engineering from IIIT Hyderabad, Hyderabad, India, in 2011, and the M.Sc. and Ph.D. degrees from the School of Technology and Computer Science, TIFR Mumbai, Mumbai, India, in 2017.

He was a Postdoctoral Fellow with IIT Bombay, India, and with University of California, USA, in 2018. His research interests include federated machine learning, differential privacy, distributed optimization, cryptography, algorithms, and information and coding theory.

Dr. Data has received the Best Paper Award from ACM Conference on Computer and Communications Security (CCS) 2021, ACM India Doctoral Dissertation Award (Honorable Mention), TIFR-Sasken Best Ph.D. Thesis Award in Technology and Computer Sciences, and a Microsoft Research India Ph.D. Fellowship 2014–2017.

Jemin George (Member, IEEE) received the M.S. and Ph.D. degrees in aerospace engineering from the State University of New York at Buffalo, Buffalo, NY, USA, in 2007 and 2010, respectively.

He was a Summer Research Scholar with the U.S. Air Force Research Laboratory's Space Vehicles Directorate, in 2008, and with National Aeronautics and Space Administration Langley Aerospace Research, in 2009. From 2009 to 2010, he was a Research Fellow with the Department of Mathematics, Technische Universität Darmstadt, Darmstadt, Germany. He was with U.S. Army Research Laboratory in 2010. From 2014 to 2017, he was a Visiting Scholar with the Northwestern University, Evanston, IL, USA. His research interests include distributed learning, stochastic systems, control theory, nonlinear filtering, information fusion, and distributed sensing and estimation.

Suhas Diggavi (Fellow, IEEE) received M.Sc and Ph.D. degrees in electrical engineering from IIT, Delhi, India, and from Stanford University, Stanford, CA, USA.

He is currently a Professor of Electrical and Computer Engineering with University of California, Los Angeles, CA, USA, where he is the Director of Information Theory and Systems Laboratory. He was a Principal Member Research Staff with AT&T Shannon Laboratories and Director of the Laboratory for Information and Communication Systems (LICOS), EPFL, Lausanne, Switzerland. His research interests include information theory and its applications to several areas, including learning, security and privacy, data compression, wireless networks, cyber-physical systems, genomics, and neuroscience.

Dr. Diggavi has received several recognitions for his research from IEEE and ACM, including the 2013 IEEE Information Theory Society and Communications Society Joint Paper Award, 2021 ACM CCS Best Paper Award, 2013 ACM Mobihoc Best Paper Award, 2006 IEEE Donald Fink Prize Paper Award, 2019 Google Faculty Research Award, and the 2020 Amazon Faculty Research Award, 2021 Facebook Faculty Award. He was selected as a Guggenheim Fellow in 2021. He has also organized several IEEE and ACM conferences.