Distributed Randomized Block Stochastic Gradient Tracking Method

Farzad Yousefian  Jayesh Yevale  Harshal D. Kaushik

Abstract—We consider distributed optimization over networks where each agent is associated with a smooth and strongly convex local objective function. We assume that the agents only have access to unbiased estimators of the gradient of their objective functions. Motivated by big data applications, our goal lies in addressing this problem when the dimensionality of the solution space is possibly large and consequently, the computation of the local gradient mappings may become expensive. We develop a randomized block-coordinate variant of the recently developed distributed stochastic gradient tracking (DSGT) method. We derive non-asymptotic convergence rates of the order 1/k and 1/k^2 in terms of an optimality metric and a consensus violation metric, respectively. Importantly, while block coordinate schemes have been studied for distributed optimization problems before, the proposed algorithm appears to be the first randomized block-coordinate gradient tracking method that is equipped with the aforementioned convergence rate statements. We validate the performance of the proposed method on the MNIST and a synthetic data set under different network settings.

I. INTRODUCTION

We consider the distributed optimization problem

$$\min_{x \in \mathbb{R}^n} \sum_{i=1}^{m} \mathbb{E}[f_i(x, \xi_i)],$$

where agents communicate over an undirected graph denoted by $G = (\mathcal{N}, \mathcal{E})$, where $\mathcal{N}$ is the node set and $\mathcal{E} \subseteq \mathcal{N} \times \mathcal{N}$. Here, $\xi_i \in \mathbb{R}^d$ denotes a local random variable. For the ease of exposition, we let $f_i(\bullet, \xi_i)$ and $f_i(\bullet) \triangleq \mathbb{E}[f_i(\bullet, \xi_i)]$ denote the stochastic and deterministic local objectives, respectively. We consider the following assumption.

Assumption 1: For all $i \in \{1, \ldots, m\}$, function $f_i(\bullet)$ is $\mu$-strongly convex and $L$-smooth.

Distributed optimization problems over networks and in particular problem 1, find a wide range of applications in statistical learning, wireless sensor networks, and control theory (see [14] and [9], [15], [1], [26], [27], [8], [28] as examples). In this paper, motivated by large-scale applications, we are interested in addressing problem 1 in stochastic and high-dimensional settings. To this end, we assume that agents only have access to noisy local gradient mappings denoted by $\nabla f_i(\bullet, \xi_i)$ satisfying the following standard assumption.

Assumption 2: For all $i \in \{1, \ldots, m\}$, $\xi_i$ are independent from each other. Also, for all $i$ and $x \in \mathbb{R}^n$,

$$\mathbb{E}[\nabla f_i(x, \xi_i) | x] = \nabla f_i(x),$$

and

$$\mathbb{E} \left[ \| \nabla f_i(x, \xi_i) - \nabla f_i(x) \| ^2 | x \right] \leq \nu^2$$

for some $\nu > 0$.

* This work is supported by the CAREER grant ECCS-1944500.

The authors are with the School of Industrial Engineering and Management, Oklahoma State University, Stillwater, OK 74078, USA. Emails: <farzad.yousefian, jayesh.yevale, harshal.kaushik>@okstate.edu.

To account for high-dimensionality, our goal in this work lies in the development of a randomized block-coordinate gradient scheme where at each iteration, agents evaluate only random blocks of their stochastic gradient mapping. To this end, throughout, we consider a block structure for $x$ given by $x = [x^{(1)}; \ldots; x^{(b)}]$ where $x^{(\ell)} \in \mathbb{R}^{n_\ell}$ denotes the $\ell$-th block-coordinate of $x \in \mathbb{R}^n$ and $\sum_{\ell=1}^{b} n_\ell = n$.

Existing methods and research gap. Among the recent advancements in distributed optimization algorithms, gradient tracking methods have been recently studied. In these schemes, agents track the average of the global gradient mapping through communicating their estimate of the gradient locally with their neighbors in convex [16], [20], [19], [29], and nonconvex regime [10], [21], [23], [24]. In [20], Push-Pull, G-Push-Pull algorithms and their variants are developed for addressing distributed optimization over directed graphs and a linear rate of convergence was established. Recently, a stochastic variant of gradient tracking methods has been developed in [19], namely the DSGT method, where non-asymptotic convergence rates of the order 1/k and 1/k^2 in terms of an optimality metric and a consensus violation metric were derived, respectively. Further, in [7], integrating the ideas from DIGing [16] and a fast incremental gradient method (SAGA) [3], S-DIGing algorithm is developed.

In the aforementioned schemes, agents have to evaluate full-dimensional gradient vectors at each iteration of the method. A popular avenue for addressing this issue is the class of block-coordinate schemes. Block-coordinate schemes, and specifically their randomized variants, have been widely studied in addressing optimization problems and games in deterministic [17], [22], [25], [4], [5] and stochastic regimes [2], [30], [13]. In randomized block schemes, at each iteration only a randomly selected block of the gradient mapping is evaluated, requiring significantly lower computational effort per iteration than the standard schemes. Although block-coordinate schemes have been studied for distributed optimization problems before [11], [18], the convergence rate statements of randomized block gradient tracking methods are not yet established. Inspired by the DSGT method [19], our goal in this paper lies in extending DSGT to a randomized block variant that is equipped with new non-asymptotic performance guarantees.

Main contributions. To highlight our contributions, we have prepared Table 1. Our main contributions are as follows:

(i) We develop a distributed randomized block stochastic gradient tracking algorithm (DRBSGT) for solving problem 1. At each iteration, each agent evaluates only random blocks of its local stochastic gradient mapping. Importantly, we assume that the agents choose both their stochastic gradient mapping

```
and their random block-coordinates independently from each other (see Algorithm 1).

(ii) In Theorem 1 we derive a rate of $O(1/k)$ on a sub-optimality and $O(1/k^2)$ on a consensus violation metric for Algorithm 1. Importantly, while DRBSGT generalizes DSGT to a randomized block variant, these rate statements match those of DSGT, indicating that there is no sacrifice in terms of the order of magnitude of the rate statements.

(iii) To validate the theoretical claims, we compare the performance of our scheme with that of other existing gradient tracking schemes and provide preliminary results on different data sets and under different network assumptions.

Outline. The rest of the paper is organized as follows. Section II includes the algorithm outline and some preliminaries. Section III provides the error bounds for the optimality and the consensus violation in Proposition 1. The rate results are presented in Theorem 1. Section IV includes the numerical experiments. We provide concluding remarks in Section V.

Notation. We let $x^*$ denote the unique global optimal solution of problem (1). We let $N(i)$ denote the set of neighbors of agent $i$, i.e., $N(i) = \{j \mid (i, j) \in E\}$. We use $[m]$ to denote $\{1, 2, \ldots, m\}$ for any integer $m \geq 1$.

Throughout, we let $\| \cdot \|$ denote the Euclidean norm and Frobenius norm of a vector and a matrix, respectively. We define $U_i \in \mathbb{R}^{n 	imes m}$ for $i \in [b]$ such that $\{U_1, \ldots, U_b\} = I_n$ where $I_n$ denotes the $n \times n$ identity matrix. Note that we can write $x = \sum_{i=1}^b U_i x^{(i)}$, $\|U_i x^{(i)}\|_2 = \|x^{(i)}\|_2$.

We consider the following notation throughout the paper. $x := [x_1, x_2, \ldots, x_m]^T$, $y := [y_1, y_2, \ldots, y_m]^T \in \mathbb{R}^{m \times n}$, $f(x) := \sum_{i=1}^m f_i(x)$, $f(x) := \sum_{i=1}^m f_i(x)$.

**Assumption 3:** For $k \geq 0$ and $i \in [m]$, let $\ell_{i,k}$ be generated from a discrete uniform distribution, i.e., $\text{Prob}(\ell_{i,k} = \ell) = b^{-1}$ for all $\ell \in [b]$.

**Remark 1:** Note that in Assumption 3 the block $\ell_{i,k}$ for each agent $i$ is independently selected. Also, note that this block selection is independent from the random variables $\xi_i$.

Algorithm 1 can compactly be written as

$$x_{k+1} = W(x_k - \gamma_k y_k),$$
$$y_{k+1} = W y_k + b^{-1} (G(x_{k+1}, \xi_{k+1}) - e_{k+1}) - b^{-1} (G(x_k, \xi_k) - e_k),$$

where $W$ is a doubly stochastic weight matrix and $\gamma_k$ is a nonincreasing strictly positive stepsize sequence. Similar to [19], we consider the following assumptions.

**Assumption 4:** The weight matrix $W$ is doubly stochastic and we have $w_{i,i} > 0$ for all $i \in [m]$.

**Assumption 5:** Let the graph $G$ corresponding to the communication network be undirected and connected.

Throughout, we show the history of the scheme for $k \geq 1$ as $f_k \triangleq \cup_{i=1}^m \{x_1, 0, \xi_{i,0}, \xi_{i,0}, \ldots, \xi_{i,k-1}, \xi_{i,k-1}\}$, with $f_0 \triangleq \cup_{i=1}^m \{x_1, 0\}$. We define the stochastic errors of the randomized block-coordinate scheme as follows.

$$e_{i,k} \triangleq \nabla f_i(x_{i,k}, \xi_{i,k}) - bU_{i,k} \nabla f_{\ell_{i,k}}(x_{i,k}, \xi_{i,k}),$$
$$e_k \triangleq e_{1,k}, e_{2,k}, \ldots, e_{m,k}^T \in \mathbb{R}^{m \times x},$$
$$\bar{e}_k \triangleq \frac{1}{m} \sum_{i=1}^m e_{i,k}.$$

We show some key properties of the randomized errors.

**Lemma 1:** We have for all $i \in [m]$ and $k \geq 0$

(a) $E[e_{i,k} \mid f_k] = E[e_k] - E[f_k] = 0.$

(b) $E[\|e_{i,k}\|_2^2 \mid f_k] \leq (b - 1) \left(\sigma^2 + \|\nabla f_i(x_{i,k})\|_2^2\right).$

(c) $E[\|e_k\|_2^2 \mid f_k] \leq (b - 1)\sigma^2 + b \frac{1}{m} \|G(x_k)\|_2^2.$

**Proof:** (a) We can write

$$E[e_{i,k} \mid f_k \cup \{\xi_{i,k}\}] = \nabla f_i(x_{i,k}, \xi_{i,k}) - bE[U_{i,k} \nabla f_{\ell_{i,k}}(x_{i,k}, \xi_{i,k}) \mid f_k \cup \{\xi_{i,k}\}] = \nabla f_i(x_{i,k}, \xi_{i,k}) - b \sum_{i=1}^b - b \nabla f_i(x_{i,k}, \xi_{i,k}).$$

The desired result follows by taking expectations from the preceding relation with respect to $\xi_{i,k}$.

(b) Throughout the proof, we use the compact notation $\nabla f_{\ell} \triangleq \nabla f_i(x_{i,k}, \xi_{i,k})$. Taking conditional expectations, we
Algorithm 1 Distributed Randomized Block Stochastic Gradient Tracking (DRBSGT)

1: **Input:** Agents choose $\gamma_0 > 0$ the weight matrix $W$. For all $i \in [m]$, agent $i$ chooses a random initial point $x_{i,0} \in \mathbb{R}^n$.

2: For all $i \in [m]$, agent $i$ generates samples $\xi_{i,0}$ and $\ell_{i,0}$ and $y_{i,0}^{(\ell_{i,0})} := \nabla f_i(x_{i,0}, \xi_{i,0})$ and $y_{i,0}^{(\ell)} := 0$ for all $\ell \neq \ell_{i,0}$.

3: for $k = 0, 1, \ldots,$ do

4: For all $i \in [m]$, agent $i$ does the following update: $x_{i,k+1} := \sum_{j=1}^{m} W_{ij} (x_{j,k} - \gamma_k y_{j,k})$.

5: For all $i \in [m]$, agent $i$ generates realizations of the random variables $\xi_{i,k+1}$ and $\ell_{i,k+1}$.

6: For all $i \in [m]$, agent $i$ does the following update:

$$y_{i,k+1}^{(\ell)} := \begin{cases} \sum_{j=1}^{m} W_{ij} y_{j,k}^{(\ell)} + \nabla f_i(x_{i,k+1}, \xi_{i,k+1}) - \nabla f_i(x_{i,k}, \xi_{i,k}), & \text{if } \ell = \ell_{i,k+1} = \ell_{i,k} \\ \sum_{j=1}^{m} W_{ij} y_{j,k}^{(\ell)} + \nabla f_i(x_{i,k+1}, \xi_{i,k+1}), & \text{if } \ell = \ell_{i,k+1} \neq \ell_{i,k} \\ \sum_{j=1}^{m} W_{ij} y_{j,k}^{(\ell)} - \nabla f_i(x_{i,k}, \xi_{i,k}), & \text{if } \ell = \ell_{i,k} \neq \ell_{i,k+1} \\ \sum_{j=1}^{m} W_{ij} y_{j,k}, & \text{if } \ell \neq \ell_{i,k+1}, \ell \neq \ell_{i,k}. \end{cases}$$

7: end for

have

$$\mathbb{E} \left[ \|e_{i,k}\|^2 \mid F_k \cup \left( \bigcup_{j=1}^{m} \{\xi_{j,k}\} \right) \right] = \left\| \nabla_{i,k} \right\|^2 + b \sum_{\ell=1}^{b} \left\| U_{\ell} \nabla_{i,k} \right\|^2 - 2 \nabla_{i,k}^T \sum_{\ell=1}^{b} U_{\ell} \nabla_{i,k}.$$

We have $\sum_{\ell=1}^{b} \left\| U_{\ell} \nabla_{i,k} \right\|^2 \leq \left\| \nabla_{i,k} \right\|^2$. From the two preceding relations, we obtain

$$\mathbb{E} \left[ \|e_{i,k}\|^2 \mid F_k \cup \left( \bigcup_{j=1}^{m} \{\xi_{j,k}\} \right) \right] = (b - 1) \left\| \nabla_{i,k} \right\|^2.$$

The desired relation holds by taking expectations with respect to $\bigcup_{j=1}^{m} \{\xi_{j,k}\}$ from both sides and invoking Assumption 3.

(c) This relation follows from part (b) and by noting that we have $\|e_{i,k}\|^2 \leq \frac{1}{m} \sum_{j=1}^{m} \|e_{i,k}\|^2$.

The following two lemmas will be applied in the analysis and can be found in [19].

**Lemma 2:** Let Assumptions 1 and 5 hold. Let $\rho_W$, denote the spectral norm of $W - \frac{1}{m} I$ and $\tilde{u} \triangleq \frac{1}{m} I u$. Then, $\rho_W < 1$, and $\|W u - \tilde{u}\| \leq \rho_W \|u - \tilde{u}\|$ for all $u \in \mathbb{R}^{m \times n}$.

**Lemma 3:** Let Assumption 1 hold. For any $\alpha \leq \frac{\sqrt{m}}{\mu + 2}$, we have $\|x_k - \alpha g(x_k) - x^*\| \leq \|x_k - x^*\|$.

We also make use of the following result later in the analysis.

**Lemma 4:** Let $u, v \in \mathbb{R}^{m \times n}$. Then, follows the holds.

(a) $u v = \sum_{i=1}^{m} u_i v_i^T = \sum_{i=1}^{m} u_i^T v_i^T$.

(b) $\|u + v\|^2 = \|u\|^2 + 2 \langle u, v \rangle + \|v\|^2$.

(c) $\langle u, v \rangle \leq \|u\| \|v\| \leq \frac{1}{\lambda} \left( \|u\|^2 + \frac{1}{\lambda} \|v\|^2 \right)$ for any $\lambda > 0$.

III. CONVERGENCE AND RATE ANALYSIS

Here, we present some properties of the gradient maps.

**Lemma 5:** Consider Algorithm 1. Let Assumptions 1 and 2 hold. Then, for all $k \geq 0$, the following results hold.

(a) $b y_k = G(x_k, \xi_k) - \tilde{e}_k$.

(b) $E [b y_k | F_k] = G(x_k)$.

(c) $E \|b y_k - G(x_k)\|^2 | F_k \leq \left( \frac{1}{m} + b - 1 \right) \nu^2 + \frac{b^2}{m} \|G(x_k)\|^2$.

(d) For any $u, v \in \mathbb{R}^{m \times n}$, $\|G(u) - G(v)\| \leq \nu \sqrt{m} |u - v|$.

(e) $\|G(x_k) - G(\bar{x}_k)\| \leq \frac{\nu}{\sqrt{m}} \|x_k - \bar{x}_k\|$.

(f) $\|G(\bar{x}_k)\| \leq \frac{\nu}{\sqrt{m}} \|x_k - \bar{x}_k\|$.

(g) $\|G(x_k)\|^2 \leq 2 \|x_k - \bar{x}_k\|^2 + 2m \|x_k - x^*\|^2$.

**Proof:** (a) We use induction on $k$. For $k = 0$ we have $b y_0 = \frac{1}{m} \sum_{i=1}^{m} b^{-1} (\nabla f_i(x_{i,0}, \xi_{i,0}) - e_{i,0}) = G(x_0, \xi_0) - \tilde{e}_0$.

Now suppose (a) holds for some $k$. We have

$$b y_{k+1} = \frac{1}{m} \sum_{i=1}^{m} b^{-1} (\nabla f_i(x_{i,k+1}, \xi_{i,k+1}) - e_{i,k}) = \frac{1}{m} \sum_{i=1}^{m} (\nabla f_i(x_{i,k+1}, \xi_{i,k+1}) - e_{i,k+1})$$

Therefore, the induction hypothesis statement holds for $k + 1$ and hence, the desired relation holds for all $k \geq 0$.

(b) Taking conditional expectations from the equation in part (a) and utilizing Lemma 1(a), we have

$$E [b y_k | F_k] = E [G(x_k, \xi_k) - \tilde{e}_k | F_k] = G(x_k) - E [\tilde{e}_k | F_k] = G(x_k).$$

(c) From part (a), we have

$$E \|b y_k - G(x_k)\|^2 | F_k = E \|G(x_k, \xi_k) - \tilde{e}_k - G(x_k)\|^2$$

$$\|F_k \leq \frac{\nu^2}{m} + (b - 1) \nu^2 + \frac{b^2}{m} \|G(x_k)\|^2,$$

where the last relation is obtained from Lemmas 1 and 2.

(d) For any $u, v \in \mathbb{R}^{m \times n}$, with $u_i, v_i \in \mathbb{R}^{n}$ denoting the $i$th row of $u, v$, respectively, we have

$$\|G(u) - G(v)\| = \frac{1}{m} \sum_{i=1}^{m} \nabla f_i(u_i) - \sum_{i=1}^{m} \nabla f_i(v_i) \|u - v\|$$

$$= \frac{1}{m} \|G(x_k) - G(\bar{x}_k)\| \leq \frac{\nu}{\sqrt{m}} \|x_k - \bar{x}_k\|.$$

(e) We have $\|G(x_k) - G(\bar{x}_k)\| = \|G(x_k) - G(\bar{x}_k)\| \leq \frac{\nu}{\sqrt{m}} \|x_k - \bar{x}_k\|.$

(f) Invoking $G(1x^*) = 0$, we have $\|G(\bar{x}_k)\| = \|G(\bar{x}_k)\| = 0 \|x_k - \bar{x}_k\| \leq \frac{\nu}{\sqrt{m}} \|x_k - \bar{x}_k\|.$

(g) We can write

$$\|G(x_k)\|^2 \leq 2 \|x_k - \bar{x}_k\|^2 + 2m \|x_k - x^*\|^2 \leq 2 \|x_k - \bar{x}_k\|^2 + 2m \|x_k - x^*\|^2.$$
In the following, we derive three recursive error bounds that will be used later to derive the convergence rate statements.

**Proposition 1 (Recursive error bounds):** Consider Algorithm 1 Let Assumptions 1-2-3-4 and 5 hold. Then, if \( \gamma_k \leq \min \left\{ \frac{2b}{\mu + L}, \frac{b\mu}{(b-1)L^2} \right\} \), for any \( \eta > 0 \) we have

(a) \( \mathbb{E} \left[ \| x_{k+1} - x^* \|^2 \right] \leq (1 - \frac{\mu - \gamma_k}{2}) \mathbb{E} \left[ \| x_k - x^* \|^2 \right] + b^{-1} \gamma_k L^2 \left( \frac{1}{\mu} + b^{-1}(2b-1)\gamma_k \right) \mathbb{E} \left[ \| x_k - 1 \bar{x}_k \|^2 \right] + b^{-2} \gamma_k^2 \left( \frac{1}{\mu} + b - 1 \right) \nu^2. \)

(b) \( \mathbb{E} \left[ \| y_{k+1} - 1 \bar{y}_k \|^2 \right] \leq \left( 1 + \frac{b^{-1} \eta}{2} \right) \rho^2_W + \gamma_k \left( \frac{1}{\mu} + \frac{b}{m} \right) \mathbb{E} \left[ \| y_k - 1 \bar{y}_k \|^2 \right] + \frac{2b^{-1} \gamma_k L^2}{m} \left( \frac{1}{\mu} + b^{-2}(2b-1) \gamma_k \right) \mathbb{E} \left[ \| x_k - x^* \|^2 \right] + \left( \frac{1}{\mu} + b^{-2} \gamma_k \left( \frac{1}{\mu} + b - 1 \right) \nu^2 \right) \mathbb{E} \left[ \| y_k - 1 \bar{y}_k \|^2 \right]. \)

(c) \( \mathbb{E} \left[ \| x_{k+1} - x^* \|^2 \right] \leq \left( 1 - \frac{\mu - \gamma_k}{2} \right) \mathbb{E} \left[ \| x_k - x^* \|^2 \right] + b^{-1} \gamma_k L^2 \left( \frac{1}{\mu} + b^{-1}(2b-1)\gamma_k \right) \mathbb{E} \left[ \| x_k - 1 \bar{x}_k \|^2 \right] + b^{-2} \gamma_k^2 \left( \frac{1}{\mu} + b - 1 \right) \nu^2. \)

**Proof:** (a) Multiplying both sides of first equation in (3) by \( 1^T \) and invoking \( 1^T W = 1^T \), we obtain \( \bar{x}_{k+1} = \bar{x}_k - \gamma_k \bar{y}_k \). Using Lemma 5(b) and (c), we can write

\[
\mathbb{E} \left[ \| x_{k+1} - x^* \|^2 \ | \mathcal{F}_k \right] = \mathbb{E} \left[ \| x_k - x^* \|^2 \ | \mathcal{F}_k \right] - 2 \gamma_k \langle x_k - x^*, x_k - x^* \rangle + \gamma_k^2 \mathbb{E} \left[ \| x_k - 1 \bar{x}_k \|^2 \right] + b^{-2} \gamma_k^2 \left( \frac{1}{\mu} + b - 1 \right) \nu^2 + \frac{b^{-2} \gamma_k^2}{m} \mathbb{E} \left[ \| G(x_k) \|^2 \right].
\]

Taking expectations from the last relation, we obtain (b).

(c) Next we obtain the third recursive relation. For the ease of presentation, we will use the following compact notation.

\( G_k = G(x_k), \ G_k = G(x_k, x_k), \ G_k^T = \hat{G}_k - e_k, \ \nabla_{i,k} = \nabla f_i(x_{i,k}), \ \nabla_{i,k} = \nabla f_i(x_{i,k}, x_{i,k}), \ \nabla_{i,k}^e = \nabla f_i(x_{i,k}, x_{i,k}) - e_{i,k}. \)
Note that $E \left[ \tilde{G}_k^c | \mathcal{F}_k \right] = G_k$. Also, we have, $2E \left[ (G_{k+1} - \tilde{G}_k^c + G_k) | \mathcal{F}_k \right] \leq E \left[ \|G_{k+1} - G_k\|^2 | \mathcal{F}_k \right] + E \left[ \|\tilde{G}_k^c - G_k\|^2 | \mathcal{F}_k \right].$

From the last three relations, we obtain Claim 1.

Claim 2: We have, $E \left[ \|G_{k+1} - G_k\|^2 | \mathcal{F}_k \right] \leq m b \nu^2 + (b - 1) E \left[ \|G_{k+1} - G_k\|^2 | \mathcal{F}_k \right].$

Proof: From Assumption 2 and Lemma 1 we have

$$E \left[ \|G_k^c - G_k\|^2 | \mathcal{F}_k \right] = E \left[ \|G_k - e_k - G_k^c\|^2 | \mathcal{F}_k \right] = m b \nu^2 + (b - 1) E \left[ \|G_k\|^2 | \mathcal{F}_k \right].$$

Using this relation, we can also write

$$E \left[ \|\tilde{G}_k^c - G_k\|^2 | \mathcal{F}_k \right] = E \left[ \|\tilde{G}_k^c - G_k + G_k^c - G_k + G_k\|^2 | \mathcal{F}_k \right]$$

$$\leq m b \nu^2 + (b - 1) E \left[ \|G_k + G_k^c\|^2 | \mathcal{F}_k \right].$$

Claim 3: The following inequality holds.

$$E \left[ \|G_{k+1} - G_k\|^2 | \mathcal{F}_k \right] \leq 2 L^2 \left( b^2 L^2 \gamma_k^2 + \|W - I\|^2 \right) \|x_k - \bar{x}_k\|^2 + 2 L^2 \rho_W \gamma_k^2 \|x_k - 1\|_2^2 | \mathcal{F}_k$$

$$+ 2 b^2 m L^4 \gamma_k^2 \|x_k - x^*\|^2 + L^2 \gamma_k^2 b^{-2} \|x_k - 1\|_2^2 | \mathcal{F}_k$$

$$+ m L^2 b^2 - 2 \|b y_k\|^2 \|b y_k\|^2 | \mathcal{F}_k \right].$$

Proof: From the Lipschitzian property of the local objectives we have $\|G_{k+1} - G_k\|^2 \leq 2 L^2 \|x_{k+1} - x_k\|^2$. Next, we estimate the term $\|x_{k+1} - x_k\|^2$. We have

$$\|x_{k+1} - x_k\|^2 = \|W x_k - \gamma_k W y_k - x_k\|^2$$

$$= \|(W - I)(x_k - \bar{x}_k) - \gamma_k W y_k\|^2$$

$$= \|(W - I)^2 \|x_k - \bar{x}_k\|^2 + \gamma_k^2 \|W y_k - 1\|_2^2$$

$$+ m \gamma_k^2 \|y_k\|^2 - 2 \gamma_k \langle (W - I)(x_k - \bar{x}_k), W y_k - 1\rangle$$

$$\leq \|(W - I)^2 \|x_k - \bar{x}_k\|^2 + \rho_W \gamma_k^2 \|y_k - 1\|_2^2$$

$$+ m \gamma_k^2 \|y_k\|^2 + 2 \rho_W \gamma_k^2 \|x_k - 1\|_2^2$$

$$+ m \gamma_k^2 \|y_k\|^2 + 2 \rho_W \gamma_k^2 \|y_k - 1\|_2^2 | \mathcal{F}_k$$

$$\leq 2 \|(W - I)^2 \|x_k - x^*\|^2 + 2 \rho_W \gamma_k^2 \|y_k - 1\|_2^2$$

$$+ m \gamma_k^2 \|y_k\|^2 | \mathcal{F}_k \right]\.$$
Claim 4: We have
\[\mathbb{E} \left[ \left\| G_{k+1} \right\|_{F}^{2} \right] \leq 2L^{2}m \left\| \bar{x}_{k} - x^{*} \right\|^{2} + \frac{2L^{2}\gamma_{b}^{2}(1+\rho_{t})}{1-\rho_{t}^{2}} \mathbb{E} \left[ \left\| y_{k} - \bar{y}_{k} \right\|^{2} \right] + 2mL^{2}b^{2}\rho_{t}^{2}(\frac{1}{m} + b)^{2} + 2mL^{2}b^{2}\rho_{t}^{2}(\frac{1}{m} + b) \gamma_{k}^{2}.\]

\(\text{Proof:}\) From Lemma 5(g) we can write
\[\mathbb{E} \left[ \left\| G_{k+1} \right\|_{F}^{2} \right] \leq 2L^{2}m \left\| \bar{x}_{k+1} - \bar{x}_{k+1} \right\|^{2} \mathbb{E} \left[ \left\| y_{k} - \bar{y}_{k+1} \right\|^{2} \right],\]
from the first two recursive bounds, substituting \(\mathbb{E} \left[ \left\| x_{k+1} - 1\bar{x}_{k+1} \right\|^{2} \right]\) and \(\mathbb{E} \left[ \left\| \bar{x}_{k+1} - x^{*} \right\|^{2} \right]\), we can conclude Claim 4.

By combining (5), Claims 1 to 4, and Lemma 5(g), we obtain
\[\mathbb{E} \left[ \left\| y_{k+1} - 1\bar{y}_{k+1} \right\|^{2} \right] \leq \left( \left(1 + b^{-1}\eta\right)\rho_{t}^{2} + \gamma_{k}^{2}(1 + \frac{1}{m}) \right) (2L^{2}\rho_{t}^{2}) \leq \mathbb{E} \left[ \left\| y_{k} - \bar{y}_{k+1} \right\|^{2} \right] + 2mL^{2}(\frac{1}{m} + b) \gamma_{k}^{2} \left(1 + \frac{1}{m} \right) \gamma_{k}^{2} + 2L^{2} \gamma_{b}^{2}(1 + \frac{1}{m}) (2L^{2}\rho_{t}^{2}) \left(1 + \frac{1}{m} \right) \gamma_{k}^{2} + 2b^{2}\rho_{t}^{2}(\frac{1}{m} + b)^{2} + 2mL^{2}b^{2}\rho_{t}^{2}(\frac{1}{m} + b) \gamma_{k}^{2}.\]

We are now ready to present the main result of the paper.

**Theorem 1 (Rate statements):** Consider Algorithm 1. Let Assumptions 2, 3, and 5 hold. Let us define \(\text{err}_{1,k} := \mathbb{E} \left[ \left\| \bar{x}_{k} - x^{*} \right\|^{2} \right] + \text{err}_{2,k} \equiv \mathbb{E} \left[ \left\| y_{k} - \bar{y}_{k+1} \right\|^{2} \right]\) for \(k \geq 0\). Suppose \(\gamma_{k} := \frac{k}{k+1}\) with \(\gamma > 0\),
\[\Gamma \geq \gamma \min \left\{ \frac{2b}{\mu L}, \frac{b\rho_{t}}{3(1-bL)} \right\} - 1.\]

(a) Then, there exist positive scalars \(\theta_{1} > 0\) for \(t = 1, \ldots, 9\) with \(\theta_{4} < 1\) and \(\theta_{0} < 1\) such that for all \(k \geq 0\) we have
\[\text{err}_{1,k+1} \leq (1 - \theta_{1}\gamma_{k})\text{err}_{1,k} + \theta_{2}\gamma_{k}\text{err}_{2,k} + \theta_{3}\gamma_{k}^{2},\]
\[\text{err}_{2,k+1} \leq (1 - \theta_{1}\gamma_{k})\text{err}_{2,k} + \theta_{4}\gamma_{k}\text{err}_{3,k},\]
\[\text{err}_{3,k+1} \leq (1 - \theta_{1}\gamma_{k})\text{err}_{3,k} + \theta_{5}\gamma_{k}\text{err}_{1,k} + \theta_{6}\text{err}_{2,k} + \theta_{9}.\]

(b) Let \(\gamma > \frac{1}{\theta_{1}}\). Let us define \(\text{err}_{1} := \Gamma \text{err}_{1,0} N_{1}, \text{err}_{2} := \text{err}_{2,0} N_{2},\) and \(\text{err}_{3} := \max \left\{ \Gamma \text{err}_{1,0} N_{3}, \text{err}_{3,0} \right\},\) where \(N_{1}, N_{2}, N_{3} \geq 0\) are given as \(N_{1} := \frac{2C_{1}}{\theta_{1} C_{2}, C_{3}, C_{4}, C_{5}}\), \(N_{2} := \frac{2C_{1}}{\theta_{1} C_{2}, C_{3}, C_{4}, C_{5}}\), and \(N_{3} := \frac{2C_{1}}{\theta_{1} C_{2}, C_{3}, C_{4}, C_{5}}\), where \(C_{1} := \gamma_{b}(1 + \frac{1}{m}) (2L^{2}\rho_{t}^{2}) \left(1 + \frac{1}{m} \right) \gamma_{k}^{2} + 2b^{2}\rho_{t}^{2}(\frac{1}{m} + b)^{2} + 2mL^{2}b^{2}\rho_{t}^{2}(\frac{1}{m} + b) \gamma_{k}^{2} + \theta_{2}\gamma_{k}\text{err}_{2,k} + \theta_{3}\gamma_{k}^{2}
\]

\[\text{err}_{2,k} \leq \text{err}_{2,0} + \frac{\gamma_{k}^{2}}{(k+1)^{2}}.\]

Therefore, we have
\[\frac{\gamma_{k}^{2} b^{2}}{(k+1)^{2}} + \frac{2b^{2}\rho_{t}^{2}(\frac{1}{m} + b)^{2} + 2mL^{2}b^{2}\rho_{t}^{2}(\frac{1}{m} + b) \gamma_{k}^{2}}{(k+1)^{2}} < 1.\]

(b) Without loss of generality, assume that \(\theta_{3}\) is arbitrarily large such that \(N_{1} \geq 1\). Consecutively, we can state that \(N_{2} \geq 1\) and \(N_{3} \geq 1\). We have \(\text{err}_{1,0} \leq \text{err}_{2,0} \gamma_{t} \leq \text{err}_{2,0} \gamma_{t}^{2} \lambda_{t}^{2}\). The first inequality in (11) holds true for \(k = 0\). Now from the definition of \(\text{err}_{2,0}\), we have \(\text{err}_{2,0} N_{2} \leq \text{err}_{2,0} \gamma_{t} \leq \text{err}_{2,0} \gamma_{t}^{2}\). This implies the second inequality in (11) holds true for \(k = 0\). Further, for \(\text{err}_{3,0}\), we have \(\text{err}_{3,0} \leq \text{err}_{3,0} \gamma_{t} \leq \text{err}_{3,0} \gamma_{t}^{2}\). Therefore, the third inequality in (11) holds true for \(k = 0\). Now let the induction hypothesis holds true for some \(k \geq 0\). From the definition of \(N_{1}\), we have \(C_{2} \leq (C_{3} - C_{2} C_{4} C_{5}) N_{1}\). Therefore, we have \(C_{2} N_{1} + C_{2} \leq C_{3} N_{1}\). Next, substituting the values of \(C_{1}, C_{2},\) and \(C_{3}\), in the above, we have
\[\text{err}_{1,0} \leq \frac{\gamma_{k}^{2} b^{2}}{(k+1)^{2}} + \frac{2b^{2}\rho_{t}^{2}(\frac{1}{m} + b)^{2} + 2mL^{2}b^{2}\rho_{t}^{2}(\frac{1}{m} + b) \gamma_{k}^{2}}{(k+1)^{2}}.\]
We have $\gamma \geq \gamma_k$. Substituting in the above leads to
\[
\frac{\text{err}_k}{(k+\Gamma)^2} \leq \frac{\gamma_k \text{err}_k}{(k+\Gamma)^2} - \frac{\gamma_k \text{err}_k}{(k+1)^2}.
\]
Further, by bounding the left-hand term above, we obtain
\[
\frac{\text{err}_k}{(k+\Gamma+1)^2} \leq \frac{\gamma_k \text{err}_k}{(k+\Gamma+1)^2} - \frac{\gamma_k \text{err}_k}{(k+1)^2}.
\]
By induction hypothesis, and Theorem 1(a), the first inequality of (11) holds for $k+1$. Next, from the definition of $N_3$, we have $N_3 \geq C_3 N_1 \geq C_4 N_3$. Substituting for $C_4$ and rearranging the terms, we obtain $\text{err}_3 \leq \frac{\theta_3}{2\gamma_3} \text{err}_2 \leq \frac{1}{2\gamma_3} (\theta_4 - \theta_3) \text{err}_2$. From $\frac{2k+1}{(k+1)^2} \leq \frac{\theta_4}{\gamma_4}$, the preceding inequality becomes $\frac{2k+1}{(k+1)^2} \text{err}_2 \leq \theta_4 \text{err}_2 - 2\theta_3 \gamma_2 \text{err}_3$. Further, from $\frac{2k+2\Gamma+1}{(k+\Gamma+1)^2} \leq \frac{2k+1}{(k+1)^2}$, we have $\frac{\text{err}_k}{(k+1)^2} \leq \frac{\text{err}_k}{(k+\Gamma+1)^2}$. By induction hypothesis and Theorem 1(a), we show the second inequality of (11) holds for $k+1$.

Next, from the definition of $N_3$, we have $N_3 \Gamma \geq C_3 N_1$. Substituting value of $C_5$ and rearranging the terms, we obtain $\frac{\text{err}_1}{(k+\Gamma)^2} \leq \theta_5 \text{err}_3$. Now, from the upper bound of $\Gamma$, we have $\Gamma^2 \geq 2C_6$. From this, we have $\Gamma^2 \left( \frac{\Gamma}{N_1} \right) \geq 2C_6 \left( \frac{\Gamma}{N_1} \right) \geq C_6 \left( \frac{2C_6}{C_4} N_1 \right)$. Substituting for $C_6$ and rearranging terms, we have $\frac{\text{err}_3}{(k+1)^2} \leq \theta_6 \text{err}_1$. From the preceding two results and $3\theta_6 \leq \theta_6 \text{err}_1$, we have $1 - \theta_6 \text{err}_3 + \frac{\theta_6}{k+1} \text{err}_1 + \frac{\theta_6}{(k+\Gamma)^2} \text{err}_2 + \theta_6 \text{err}_3$. By induction hypothesis and Theorem 1(a), we conclude that the third inequality of (11) holds for $k+1$.

IV. NUMERICAL EXPERIMENTS

We provide numerical experiments for comparing the performance of Algorithm 1 with other gradient tracking schemes. For the experiments, we consider the distributed regularized logistic regression loss minimization problem.

Let the data be denoted by $D \triangleq \{(u_j, v_j) \in \mathbb{R}^n \times \{-1, +1\} | j \in S\}$ where $S \triangleq \{1, \ldots, s\}$ denotes the index set. Let $S_i$ denote the data locally known to agent $i$ where $\bigcup_{i=1}^{m} S_i = S_{\text{train}}$. The problem above can be formulated as $\min \sum_{i=1}^{m} f_i(x)$ where we define local functions $f_i$ as
\[
f_i(x) \triangleq \frac{1}{\bar{s}} \sum_{j \in S_i} \ln (1 + \exp (-v_j u_j^T x)) + \frac{\mu}{2n} \|x\|^2.
\]

Here, $u_j \in \mathbb{R}^n$ denotes attributes, and $v_j \in \{-1, 1\}$ for $j \in S_i$ denotes the binary label of the $j^{th}$ data point.

We simulate the proposed distributed randomized block stochastic gradient tracking method (DRBSGT) algorithm on a system consisting of $m$ agents. We provide a comparison of suboptimality and consensus metrics with the two other existing methods namely, distributed stochastic gradient tracking (DSGT) [19] and adapt then combine (ATC), a variant of block distributed Successive CONvex Approximation algorithm over Time-varying digRAPHs (block SONATA) for convex regimes [18].

Setup. The simulations are performed on two data sets with $m$ agents. We use the complete and the ring graph structure to represent the communication among the agents. We implement the simulations on MNIST and Synthetic data set for $m = 5$ and $m = 5, 10$, respectively. The MNIST data set consists of 50,000 labels and 784 attributes, whereas the Synthetic data set has 10,000 labels and 10,000 attributes with a Gaussian distribution with mean as 5 and standard deviation as 0.5. We consider different parameters for different data sets mentioned in Table 2. Further, we use $\gamma = 1 + 1$, $\Gamma = 1 + 4$, $\mu = 1 - 1$, and the batch size for computing gradient from each agent $\epsilon = 1 + 2$ for both data sets. Taking into account the stochasticity involved in DRBSGT and DSGT schemes, we have obtained different sample paths in our implementations. The highlighted areas in the plots in Figure 1 represent the confidence intervals. We provide 90% confidence intervals on the errors of the sample paths for each setting in Table 3. We choose the total sample paths for MNIST and Synthetic data sets as 5 and 10, respectively.

Insights. In Figure 1 the proposed DRBSGT scheme converges in both MNIST and Synthetic data sets, considering both the suboptimality and consensus metrics. We observe that DRBSGT performs well compared to its counterparts. Note that Figure 2 presents the performance of the schemes with respect to the number of local gradient evaluations. This perhaps explains the slow convergence of ATC as the
scheme is deterministic and cycles through all the blocks at each iteration. From Table 3 and Figure 1, we observe that when either the number of attributes or the number of agents increases, the performance of DRB SGT improves. We do not observe any significant difference in the performance in terms of the network connectivity structure.

V. CONCLUDING REMARKS

We address a networked distributed optimization problem in stochastic regimes where each agent can only access unbiased estimators of its local gradient mapping. Motivated by big data applications, we address this problem considering a possibility of large-dimensionality of the solution space where the computation of the local gradient mappings may become expensive. We develop a distributed randomized block stochastic gradient tracking scheme and provide the non-asymptotic convergence rates of the order $1/k$ and $1/k^2$ in terms of an optimality metric and a consensus violation metric, respectively. We validate the performance of the proposed algorithm on the MNIST and a synthetic data set under different settings of the communication graph.

REFERENCES

[1] N. S. Aybat, Z. Wang, T. Lin, and S. Ma. Distributed linearized alternating direction method of multipliers for composite convex consensus optimization. IEEE Transactions on Automatic Control, 63(1):5–20, 2018.

[2] C. D. Dang and G. Lan. Stochastic block mirror descent methods for nonsmooth and stochastic optimization. SIAM Journal on Optimization, 25(2):856–881, 2015.

[3] A. Defazio, F. Bach, and S. Lacoste-Julien. SAGA: A fast incremental gradient method with support for non-strongly convex composite objectives. Advances in Neural Information Processing Systems, pages 1646–1654, 2014.

[4] H. D. Kaushik and F. Yousefian. A randomized block coordinate iterative regularized subgradient method for high-dimensional ill-posed convex optimization. In Proceedings of the American Control Conference, pages 3420–3425, IEEE, July 2019.

[5] H. D. Kaushik and F. Yousefian. A method with convergence rates for optimization problems with variational inequality constraints. SIAM Journal on Optimization, 31(3):2171–2198, 2021.

[6] B. Li, S. Cen, Y. Chen, and Yu Chi. Communication-efficient distributed optimization in networks with gradient tracking and variance reduction. Proceedings of the Twenty Third International Conference on Artificial Intelligence and Statistics, pages 1662–1672, 2020.

[7] H. Li, L. Zheng, Z. Wang, Y. Yan, L. Feng, and J. Guo. S-DIGing: A stochastic gradient tracking algorithm for distributed optimization. IEEE Transactions on Emerging Topics in Computational Intelligence, pages 1–30, 2020.