2023 62nd IEEE Conference on Decision and Control (CDC)
December 13-15, 2023. Marina Bay Sands, Singapore

Event-triggered distributed nonconvex optimization with progress-based threshold

Changxin Liu[1] and Eric Shi[2]

Abstract— **This work studies the distributed nonconvex optimization problem in bandwidth-limited communication environments. We develop a communication-efficient algorithm based on the gradient-tracking based distributed optimization method, where each computation node is equipped with a new event-triggered communication scheduler. Such scheduler approves the broadcasting only when the innovation of exchanged variables exceeds the change of decision variables in two consecutive updates. Compared to the conventional scheduler with time-dependent vanishing thresholds, the proposed one adapts better to the optimization dynamics and thus leads to more significant communication reduction. Finally, we prove the convergence of the algorithm and illustrate its performance via numerical examples.**

I. Introduction

In the past decade, distributed optimization has received increasing attention from both academia and industry [1], since it enables a group of computing nodes to collaboratively solve large-scale optimization problems over communication networks. In order to tackle increasingly more sophisticated learning and control tasks, the scale and complexity of distributed optimization systems grow, resulting in significantly heavier communication load on the network during implementation. To proactively lower down the communication load, different types of communication-efficient distributed optimization algorithms have been proposed recently.

In the literature, communication-efficient distributed optimization algorithms might be roughly categorized into two groups. In the first category, the message packets are quantized deliberately to save communication while preserving the convergence property in original algorithms [2]–[4]. A typical technique to achieve this is the quantization of the variable change in two consecutive updates rather than the variable itself; see [4] for the case with stochastic gradients and [5] for the case with full gradients.

Another strategy to save communication resources is the event-triggered scheduling [6]. In this type of methods, the local broadcasting of each computing node is governed by a testing rule that compares the innovation of variables to be sent out with a prescribed threshold. The scheduler approves the broadcasting only when the up-to-date variables are innovative enough, and therefore avoids unnecessary communication usage. Most of existing event-triggered distributed optimization algorithms considered convex problems. For example, the works in [7] presented event-triggered distributed gradient algorithms with thresholds that are summable over time. For strongly convex problems, exponentially decaying thresholds are used to maintain the linear convergence of the original distributed optimization algorithm [8]. Notably, the authors in [9] designed a threshold that evolves with the consensus error, based on which the overall algorithm also presents linear convergence. Nevertheless, event-triggered distributed nonconvex optimization has rarely been explored. One recent attempt used a summable threshold and showed the convergence of the algorithm [10]. We remark that event-triggered communication is orthogonal to quantization; they can be combined to achieve more significant communication reductions, as demonstrated in [11].

In this work, our focus is placed on event-triggered distributed nonconvex optimization. We develop a new triggering threshold that is proportional to the local progress in optimization, that is, the change of decision variables in two consecutive updates. The threshold automatically decays along with the convergence of the algorithm. Compared to the time-dependent vanishing thresholds, the new one adapts well to the dynamic behavior of distributed optimization and leads to greater communication reduction. Then, this scheme is incorporated into the distributed optimization algorithm with gradient tracking (DOGT) [12] for communication reduction. We provide sufficient conditions on the step-size and the parameter in threshold, under which the event-triggered optimization algorithm converges for nonconvex problems. The performance in communication reduction is demonstrated via numerical examples.

The remaining of this work is organized as follows. We formulate the problem with some preliminaries in Section II. The algorithm is developed in Section III, followed by the convergence analysis in Section IV. Section IV presents the numerical experiments and Section V concludes this work.

II. Problem Statement and Preliminaries

A. Basic Set-up

Consider the standard distributed optimization problem, in which a group of computing nodes/agents aim at solving the following finite-sum optimization problem

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

where each f_i is a possibly nonconvex function. We assume the optimal value $f^* > -\infty$. Each agent i only has

[1]C. Liu is with the School of Electrical Engineering and Computer Science, KTH Royal Institute of Technology, SE-10044 Stockholm, Sweden changxinl@ieee.org
[2]E. Shi is with Harvard University, USA eshi@college.harvard.edu

access to the local objective function f_i, and the information exchange among the agents are restricted in the sense that the communication topology is a sparse graph. In particular, we use a doubly stochastic matrix P to describe the network topology and the weights of connected links. We denote by p_{ij} the (i,j)-th element in P, and it is positive only if the two agents i and j are neighbors. The set of i's neighbors is denoted as $\mathcal{N}_i$.

Assumption 1: i) The graph is connected; ii) P has a strictly positive diagonal, i.e., $p_{ii} > 0$; iii) $P\mathbf{1} = \mathbf{1}$ and $\mathbf{1}^T P = \mathbf{1}^T$, where $\mathbf{1}$ denotes the all-one vector of dimension n.

Assumption 1 ensures that there exists a constant $\beta \in (0,1)$ such that the second largest singular value of P

$$\sigma_2(P) \leq \beta. \tag{2}$$

For the local objective function, we make the following assumption.

Assumption 2: Each f_i is continuously differentiable, and ∇f_i is Lipschitz continuous with constant $L > 0$, i.e.,

$$\|\nabla f_i(x) - \nabla f_i(y)\| \leq L\|x - y\|, \quad \forall x, y.$$

A direct consequence of Assumption 2 is

$$f_i(y) \leq f_i(x) + \langle \nabla f_i(x), y - x\rangle + \frac{L}{2}\|y - x\|^2, \quad \forall x, y.$$

B. DOGT

Our algorithm is based on the DOGT algorithm, in which each agent i maintains two variables x_i and s_i, which are local estimates of the solution and the gradient of f, respectively. At each time t, agent i performs the following to update both variables:

$$x_i^{(t)} = \sum_{j=1}^{n} p_{ij}\left(x_j^{(t-1)} - \alpha s_j^{(t-1)}\right) \tag{3a}$$

$$s_i^{(t)} = \sum_{j=1}^{n} p_{ij} s_j^{(t-1)} + \nabla f_i(x_i^{(t)}) - \nabla f_i(x_i^{(t-1)}) \tag{3b}$$

where α is the step-size. For agent i to implement the above update, only information from its immediate neighbors is required. The convergence of DOGT has been investigated for convex and nonconvex problems [12], [13].

In (3), each agent i needs to update with its neighbors about x_i and s_i at each time t, which may result in unnecessary communication. To tackle this problem in this work, we employ the event-triggered communication strategy that schedules the information exchange based on a local event test, which reduces the utilization of communication resources.

III. Algorithm Development

A. Event-triggered Communication

Let $q_i = \left[x_i; \alpha s_i\right]$, and $\hat{q}_i$ denote the latest version of q_i that has been broadcast by agent i to its neighbors. Each agent i only updates q_i with its neighbors when the deviation between q_i and $\hat{q}_i$ exceeds a given threshold depending on the progress being locally made, i.e.,

$$\|\hat{q}_i^{(t-1)} - q_i^{(t)}\|^2 \geq \eta\|\Delta x_i^{(t-1)}\|^2 \tag{4}$$

where $\Delta x_i^{(t-1)} = x_i^{(t)} - x_i^{(t-1)}$ and η is a positive constant. Here, the parameter η is made uniform across the agents for brevity, and can be extended to the heterogeneous case without much effort. Because each agent i broadcasts $\hat{q}_i^{(t)}$ when the condition in (4) is violated, there holds

$$\|\hat{q}_i^{(t)} - q_i^{(t)}\|^2 \leq \eta\|\Delta x_i^{(t-1)}\|^2. \tag{5}$$

Such progress-based triggering threshold is motivated by the following. The progress made by optimization algorithms at each iteration can be typically described by the difference between two consecutive updates. The lack of up-to-date information from neighbors in distributed optimization introduces noise to local updates. The magnitude of noise is critical to convergence. If the noise caused by triggering is restricted to be proportional to the optimization progress with a tunable parameter as in (4), then it is sensible that the algorithm remains convergent as long as the effect due to noise can be compensated by sufficient progress.

Compared to time-dependent vanishing thresholds in the literature, i.e.,

$$\|\hat{q}_i^{(t-1)} - q_i^{(t)}\|^2 \geq \rho^{(t)} \tag{6}$$

where the nonnegative sequence $\{\rho^{(t)}\}_{t\geq 0}$ is square summable, the proposed one features the following. On one hand, provided that the algorithm converges, Δx_i converges to zero and thus the noise due to triggering vanishes. Thus the new triggering strategy will not prevent the algorithm from exact convergence, similar to the time-dependent vanishing thresholds. One the other hand, the triggering threshold depends on the algorithm dynamics and therefore triggers the broadcast more wisely, as we will show in numerical results.

B. Optimization Algorithm

By incorporating the event-triggered communication, we modify the DOGT algorithm as

$$\begin{aligned} x_i^{(t)} = x_i^{(t-1)} - \alpha s_i^{(t-1)} &- \sum_{j=1}^{n} p_{ij}\left(\hat{x}_i^{(t-1)} - \hat{x}_j^{(t-1)}\right) \\ &+ \alpha \sum_{j=1}^{n} p_{ij}\left(\hat{s}_i^{(t-1)} - \hat{s}_j^{(t-1)}\right) \end{aligned} \tag{7a}$$

$$\begin{aligned} s_i^{(t)} = s_i^{(t-1)} &- \sum_{j=1}^{n} p_{ij}\left(\hat{s}_i^{(t-1)} - \hat{s}_j^{(t-1)}\right) \\ &+ \nabla f_i(x_i^{(t)}) - \nabla f_i(x_i^{(t-1)}), \end{aligned} \tag{7b}$$

where $\hat{x}_i$ and $\hat{s}_i$ denote the latest iterates of agent i that is made available to its neighbors, and $\hat{s}_i^{(0)} = s_i^{(0)}$ and $\hat{x}_i^{(0)} = x_i^{(0)}$. If $\hat{x}_i \equiv x_i$ and $\hat{s}_i \equiv s_i$, then (7) reduces to the update in (3).

Algorithm 1 Event-triggered DOGT

Input: $\alpha > 0$, $x^{(0)}$

Output: $x_i^{(t)}, t = 1, 2, \dots$

Initialize: each agent $i = 1, \dots, n$ sets $x_i^{(0)} = \hat{x}_i^{(0)} = x^{(0)}$, $s_i^{(0)} = \hat{s}_i^{(0)} = \nabla f_i(x^{(0)})$, and receives $\hat{x}_j^{(0)}$ and $\hat{s}_j^{(0)}$ from $j \in \mathcal{N}_i$

for $t = 1, 2, \dots$, each agent i synchronously **do**

  update $x_i^{(t)}$ by (7a)

  update $s_i^{(t)}$ by (7b)

  if (4) is satisfied **then**

    send $x_i^{(t)}$ and $s_i^{(t)}$ to $j \in \mathcal{N}_i$

  end if

  if new information is received from $j \in \mathcal{N}_i$ **then**

    update $\hat{x}_j^{(t)} = x_j^{(t)}$ and $\hat{s}_j^{(t)} = s_j^{(t)}$

  else

    set $\hat{x}_j^{(t)} = \hat{x}_j^{(t-1)}$ and $\hat{s}_j^{(t)} = \hat{s}_j^{(t-1)}$

  end if

end for

IV. Rate Analysis

A. Analysis Set-up

Denote

$$\overline{x} = n^{-1}\sum_{i=1}^n x_i,\ \overline{s} = n^{-1}\sum_{i=1}^n s_i,\ \overline{g} = n^{-1}\nabla\sum_{i=1}^n f_i(x_i)$$

$$\mathbf{s} = \left[s_1; \dots; s_n\right],\ \hat{\mathbf{s}} = \left[\hat{s}_1; \dots; \hat{s}_n\right],\ \tilde{\mathbf{s}} = \mathbf{s} - \mathbf{1}\otimes\overline{s}$$

$$\mathbf{x} = \left[x_1; \dots; x_n\right],\ \hat{\mathbf{x}} = \left[\hat{x}_1; \dots; \hat{x}_n\right],\ \tilde{\mathbf{x}} = \mathbf{x} - \mathbf{1}\otimes\overline{x}.$$

We obtain from (7) that

$$\begin{aligned}\mathbf{x}^{(t)} &= \mathbf{P}(\mathbf{x}^{(t-1)} - \alpha\mathbf{s}^{(t-1)}) \\ &\quad + (\mathbf{P} - I)\left((\hat{\mathbf{x}}^{(t-1)} - \mathbf{x}^{(t-1)}) - \alpha(\hat{\mathbf{s}}^{(t-1)} - \mathbf{s}^{(t-1)})\right) \\ \mathbf{s}^{(t)} &= \mathbf{P}\mathbf{s}^{(t-1)} + \nabla^{(t)} - \nabla^{(t-1)} \\ &\quad + (\mathbf{P} - I)(\hat{\mathbf{s}}^{(t-1)} - \mathbf{s}^{(t-1)})\end{aligned} \tag{8}$$

where $\mathbf{P} = P \otimes I$ and $\nabla^{(t)} = \left[\nabla f_1(x_1^{(t)}); \dots; \nabla f_n(x_n^{(t)})\right]$. Because of $\text{Null}(I - P^{[l]}) = \text{Span}(\mathbf{1})$ where $\text{Null}(\cdot)$ denotes the null space of a linear map, it can be verified that the following conservation property holds for (7):

$$\begin{aligned}\overline{x}^{(t+1)} &= \overline{x}^{(t)} - \alpha\overline{s}^{(t)} \\ \overline{s}^{(t+1)} &= \overline{s}^{(t)} + \overline{g}^{(t+1)} - \overline{g}^{(t)}.\end{aligned} \tag{9}$$

In addition, the update of $\{\overline{x}^{(t)}\}_{t\geq 1}$ in (9) can be taken as gradient descent with inexact gradients, whose convergence property is summarized in the following lemma.

Lemma 1: Suppose Assumption 2 holds. For $\overline{x}^{(t)}, t = 1, \dots,$ generated by (9), it holds that $\forall \varepsilon > 0$

$$\begin{aligned}n\left(f(\overline{x}^{(t)}) - f(\overline{x}^{(t-1)})\right) &\leq \left(\frac{L+\varepsilon}{2} - \frac{1}{\alpha}\right)\|\Delta\overline{\mathbf{x}}^{(t-1)}\|^2 \\ &\quad + \frac{L^2}{2\varepsilon}\|\tilde{\mathbf{x}}^{(t-1)}\|^2\end{aligned} \tag{10}$$

where $\Delta\overline{\mathbf{x}}^{(t-1)} = \overline{\mathbf{x}}^{(t)} - \overline{\mathbf{x}}^{(t-1)}$ and $\overline{\mathbf{x}} = \mathbf{1}\otimes\overline{x}$.

Lemma 1 can be taken as an unconstrained version of [14, Lemma 3]; its proof is presented in Appendix for completeness.

B. Rate Analysis

Define

$$\mathbf{q} = \left[q_1; \cdots; q_n\right],\quad \hat{\mathbf{q}} = \left[\hat{q}_1; \cdots; \hat{q}_n\right],\quad \check{\mathbf{q}} = \hat{\mathbf{q}} - \mathbf{q}.$$

Given positive constants κ and γ, we define the following Lyapunov candidate function

$$R(\overline{x}, \tilde{\mathbf{x}}, \tilde{\mathbf{s}}, \check{\mathbf{q}}) = nf(\overline{x}) + \|\tilde{\mathbf{x}}\|^2 + \kappa\|\tilde{\mathbf{s}}\|^2 + \gamma\|\check{\mathbf{q}}\|^2.$$

Clearly, this function is bounded from below by f^*. Next, we investigate the sufficient conditions under which the function value of $R(\overline{x}, \tilde{\mathbf{x}}, \tilde{\mathbf{s}}, \check{\mathbf{q}})$ monotonically decreases.

Lemma 2: Suppose Assumptions 1 and 2 hold. If the parameters α, κ, γ and η are chosen such that

$$\begin{aligned}\rho_x &= 1 - \frac{(1+\beta)^2}{4} - \kappa\zeta_2 - \frac{L^2}{2\varepsilon} - 3\eta\gamma\left(1 + \frac{(1+\beta)^2}{4}\right) \\ \rho_s &= \kappa - \kappa\zeta_1 - \frac{\beta\alpha^2(1+\beta)^2}{2(1-\beta)}(1 + 3\eta\gamma) \\ \rho_q &= \gamma - \left(\frac{6\gamma\eta(1+\beta)}{(1-\beta)} + \frac{2(1+\beta)}{1-\beta}\right)\|\mathbf{P} - I\|^2 - \kappa\zeta_3 \\ \rho_y &= \frac{1}{\alpha} - \frac{L+\varepsilon}{2} - \frac{3\kappa(1+\beta)^2L^2}{2\beta(1-\beta)} - 3\eta\gamma\end{aligned}$$

are positive, where

$$\begin{aligned}\zeta_1 &= \frac{(1+\beta)^2}{4}\left(1 + \frac{3\alpha^2L^2(1+\beta)^2}{(1-\beta)^2}\right) \\ \zeta_2 &= \frac{3(1+\beta)^2L^2}{2\beta(1-\beta)}\left(1 + \frac{(1+\beta)^2}{4}\right) \\ \zeta_3 &= \frac{1+\beta}{1-\beta}\|\mathbf{P} - I\|^2\left(\frac{1}{\alpha^2} + \frac{3(1+\beta)^2L^2}{\beta(1-\beta)}\right)\end{aligned} \tag{11}$$

and $\beta \in (0, 1)$ is defined in (2), respectively, then it holds that

$$\begin{aligned}&R(\overline{x}^{(t+1)}, \tilde{\mathbf{x}}^{(t+1)}, \tilde{\mathbf{s}}^{(t+1)}, \check{\mathbf{q}}^{(t+1)}) - R(\overline{x}^{(t)}, \tilde{\mathbf{x}}^{(t)}, \tilde{\mathbf{s}}^{(t)}, \check{\mathbf{q}}^{(t)}) \\ &\leq -r(\overline{x}^{(t)}, \tilde{\mathbf{x}}^{(t)}, \tilde{\mathbf{s}}^{(t)}, \check{\mathbf{q}}^{(t)})\end{aligned}$$

where

$$r(\overline{x}, \tilde{\mathbf{x}}, \tilde{\mathbf{s}}, \check{\mathbf{q}}) = \rho_x\|\tilde{\mathbf{x}}\|^2 + \rho_s\|\tilde{\mathbf{s}}\|^2 + \rho_q\|\check{\mathbf{q}}\|^2 + \rho_y\|\Delta\overline{\mathbf{x}}\|^2.$$

Proof: Using (9) and (8), we obtain

$$\begin{aligned}\tilde{\mathbf{x}}^{(t)} &= \mathbf{P}\mathbf{x}^{(t-1)} - \mathbf{1}\otimes\overline{x}^{(t-1)} - \alpha\left(\mathbf{P}\mathbf{s}^{(t-1)} - \mathbf{1}\otimes\overline{s}^{(t-1)}\right) \\ &\quad + (\mathbf{P} - I)\left((\hat{\mathbf{x}}^{(t-1)} - \mathbf{x}^{(t-1)}) - \alpha(\hat{\mathbf{s}}^{(t-1)} - \mathbf{s}^{(t-1)})\right) \\ &= (P\otimes I)\mathbf{x}^{(t-1)} - \left(\frac{1}{n}(\mathbf{1}\mathbf{1}^T)\otimes I\right)\mathbf{x}^{(t-1)} \\ &\quad - \alpha(P\otimes I)\mathbf{s}^{(t-1)} + \alpha\left(\frac{1}{n}(\mathbf{1}\mathbf{1}^T)\otimes I\right)\mathbf{s}^{(t-1)} \\ &\quad + (\mathbf{P} - I)\left((\hat{\mathbf{x}}^{(t-1)} - \mathbf{x}^{(t-1)}) - \alpha(\hat{\mathbf{s}}^{(t-1)} - \mathbf{s}^{(t-1)})\right) \\ &= \left(\left(P - \frac{1}{n}\mathbf{1}\mathbf{1}^T\right)\otimes I\right)\left(\tilde{\mathbf{x}}^{(t-1)} - \alpha\tilde{\mathbf{s}}^{(t-1)}\right) \\ &\quad + (\mathbf{P} - I)\left((\hat{\mathbf{x}}^{(t-1)} - \mathbf{x}^{(t-1)}) - \alpha(\hat{\mathbf{s}}^{(t-1)} - \mathbf{s}^{(t-1)})\right).\end{aligned}$$

Upon using

$$\begin{aligned}\|a+b+c\|^2 \leq & (1+k_1)(1+k_2)\|a\|^2 \\ & + (1+k_1)(1+k_2^{-1})\|b\|^2 + (1+k_1^{-1})\|c\|^2\end{aligned}$$

for $k_1 = k_2 = \frac{1}{2}\left(\frac{1}{\beta} - 1\right)$ and

$$\begin{aligned}&\left\|(\hat{\mathbf{x}}^{(t-1)} - \mathbf{x}^{(t-1)}) - \alpha(\hat{\mathbf{s}}^{(t-1)} - \mathbf{s}^{(t-1)})\right\|^2 \\ &\leq 2(\|\hat{\mathbf{x}}^{(t-1)} - \mathbf{x}^{(t-1)}\|^2 + \|\alpha(\hat{\mathbf{s}}^{(t-1)} - \mathbf{s}^{(t-1)})\|^2) \\ &\leq 2\|\check{\mathbf{q}}^{(t-1)}\|^2,\end{aligned}$$

we obtain that

$$\begin{aligned}\|\tilde{\mathbf{x}}^{(t)}\|^2 \leq & \frac{(1+\beta)^2}{4}\|\tilde{\mathbf{x}}^{(t-1)}\|^2 + \frac{\beta\alpha^2(1+\beta)^2}{2(1-\beta)}\|\tilde{\mathbf{s}}^{(t-1)}\|^2 \\ & + \frac{2(1+\beta)}{1-\beta}\|\mathbf{P}-I\|^2\|\check{\mathbf{q}}^{(t-1)}\|^2.\end{aligned} \tag{12}$$

Similarly, it can be verified from (8) that

$$\begin{aligned}&\|\tilde{\mathbf{s}}^{(t)}\|^2 \\ &\leq \frac{(1+\beta)^2}{4}\|\tilde{\mathbf{s}}^{(t-1)}\|^2 + \frac{1+\beta}{1-\beta}\|\mathbf{P}-I\|^2\|\hat{\mathbf{s}}^{(t-1)} - \mathbf{s}^{(t-1)}\|^2 \\ &\quad + \frac{(1+\beta)^2L^2}{2\beta(1-\beta)}\|\Delta\mathbf{x}^{(t-1)}\|^2 \\ &\leq \frac{(1+\beta)^2}{4}\|\tilde{\mathbf{s}}^{(t-1)}\|^2 + \frac{1+\beta}{1-\beta}\|\mathbf{P}-I\|^2\|\hat{\mathbf{s}}^{(t-1)} - \mathbf{s}^{(t-1)}\|^2 \\ &\quad + \frac{3(1+\beta)^2L^2}{2\beta(1-\beta)}\|\tilde{\mathbf{x}}^{(t)}\|^2 + \frac{3(1+\beta)^2L^2}{2\beta(1-\beta)}\|\tilde{\mathbf{x}}^{(t-1)}\|^2 \\ &\quad + \frac{3(1+\beta)^2L^2}{2\beta(1-\beta)}\|\Delta\overline{\mathbf{x}}^{(t-1)}\|^2,\end{aligned}$$

where we use

$$\begin{aligned}&\|\Delta\mathbf{x}^{(t-1)}\|^2 \\ &= \|\mathbf{x}^{(t)} - \overline{\mathbf{x}}^{(t)} + \overline{\mathbf{x}}^{(t-1)} - \mathbf{x}^{(t-1)} + \Delta\overline{\mathbf{x}}^{(t-1)}\|^2 \\ &\leq 3\|\tilde{\mathbf{x}}^{(t)}\|^2 + 3\|\tilde{\mathbf{x}}^{(t-1)}\|^2 + 3\|\Delta\overline{\mathbf{x}}^{(t-1)}\|^2\end{aligned}$$

to get the last inequality. Using (12) and

$$\|\hat{\mathbf{s}}^{(t-1)} - \mathbf{s}^{(t-1)}\|^2 \leq \frac{\|\check{\mathbf{q}}^{(t-1)}\|^2}{\alpha^2},$$

we arrive at

$$\begin{aligned}&\|\tilde{\mathbf{s}}^{(t)}\|^2 \\ &\leq \frac{(1+\beta)^2}{4}\|\tilde{\mathbf{s}}^{(t-1)}\|^2 + \frac{1+\beta}{1-\beta}\|\mathbf{P}-I\|^2\|\check{\mathbf{s}}^{(t-1)}\|^2 \\ &\quad + \frac{3(1+\beta)^2L^2}{2\beta(1-\beta)}\Big(\frac{2(1+\beta)}{1-\beta}\|\mathbf{P}-I\|^2\|\check{\mathbf{q}}^{(t-1)}\|^2 \\ &\quad\quad + \frac{\beta\alpha^2(1+\beta)^2}{2(1-\beta)}\|\tilde{\mathbf{s}}^{(t-1)}\|^2 + \frac{(1+\beta)^2}{4}\|\tilde{\mathbf{x}}^{(t-1)}\|^2\Big) \\ &\quad + \frac{3(1+\beta)^2L^2}{2\beta(1-\beta)}\|\tilde{\mathbf{x}}^{(t-1)}\|^2 + \frac{3(1+\beta)^2L^2}{2\beta(1-\beta)}\|\Delta\overline{\mathbf{x}}^{(t-1)}\|^2 \\ &\leq \zeta_1\|\tilde{\mathbf{s}}^{(t-1)}\|^2 + \zeta_2\|\tilde{\mathbf{x}}^{(t-1)}\|^2 + \zeta_3\|\check{\mathbf{q}}^{(t-1)}\|^2 \\ &\quad + \frac{3(1+\beta)^2L^2}{2\beta(1-\beta)}\|\Delta\overline{\mathbf{x}}^{(t-1)}\|^2\end{aligned}$$

where ζ_1, ζ_2 and ζ_3 are defined in (11). In addition, we have

$$\begin{aligned}&\|\check{\mathbf{q}}^{(t)}\|^2 \leq \eta\|\Delta\mathbf{x}^{(t-1)}\|^2 \\ &\leq 3\eta\left(\|\tilde{\mathbf{x}}^{(t)}\|^2 + \|\tilde{\mathbf{x}}^{(t-1)}\|^2 + \|\Delta\overline{\mathbf{x}}^{(t-1)}\|^2\right) \\ &\leq 3\eta\left(1 + \frac{(1+\beta)^2}{4}\right)\|\tilde{\mathbf{x}}^{(t-1)}\|^2 \\ &\quad + \frac{3\eta\beta\alpha^2(1+\beta)^2}{2(1-\beta)}\|\tilde{\mathbf{s}}^{(t-1)}\|^2 \\ &\quad + \frac{6\eta(1+\beta)}{1-\beta}\|\mathbf{P}-I\|^2\|\check{\mathbf{q}}^{(t-1)}\|^2 + 3\eta\|\Delta\overline{\mathbf{x}}^{(t-1)}\|^2.\end{aligned} \tag{13}$$

where (5) is used to derive the first inequality. Upon using Lemma 1, we obtain

$$\begin{aligned}&R(\overline{x}^{(t)}, \tilde{\mathbf{x}}^{(t)}, \tilde{\mathbf{s}}^{(t)}, \check{\mathbf{q}}^{(t)}) - R(\overline{x}^{(t-1)}, \tilde{\mathbf{x}}^{(t-1)}, \tilde{\mathbf{s}}^{(t-1)}, \check{\mathbf{q}}^{(t-1)}) \\ &\leq \left(\frac{L+\varepsilon}{2} - \frac{1}{\alpha}\right)\|\Delta\overline{\mathbf{x}}^{(t-1)}\|^2 + \frac{L^2}{2\varepsilon}\|\tilde{\mathbf{x}}^{(t-1)}\|^2 \\ &\quad + \left(\frac{(1+\beta)^2}{4} + \kappa\zeta_2 - 1\right)\|\tilde{\mathbf{x}}^{(t-1)}\|^2 \\ &\quad + 3\eta\gamma\left(1 + \frac{(1+\beta)^2}{4}\right)\|\tilde{\mathbf{x}}^{(t-1)}\|^2 \\ &\quad + \left(\frac{\beta\alpha^2(1+\beta)^2}{2(1-\beta)} + \zeta_1\kappa - \kappa\right)\|\tilde{\mathbf{s}}^{(t-1)}\|^2 \\ &\quad + \frac{3\eta\gamma\beta\alpha^2(1+\beta)^2}{2(1-\beta)}\|\tilde{\mathbf{s}}^{(t-1)}\|^2 \\ &\quad + \left(\frac{6\gamma\eta(1+\beta)}{1-\beta}\|\mathbf{P}-I\|^2 - \gamma\right)\|\check{\mathbf{q}}^{(t-1)}\|^2 \\ &\quad + \left(\frac{2(1+\beta)}{1-\beta}\|\mathbf{P}-I\|^2 + \kappa\zeta_3\right)\|\check{\mathbf{q}}^{(t-1)}\|^2 \\ &\quad + \left(3\eta\gamma + \frac{3\kappa(1+\beta)^2L^2}{2\beta(1-\beta)}\right)\|\Delta\overline{\mathbf{x}}^{(t-1)}\|^2 \\ &\leq -r(y^{(t-1)}, \tilde{\mathbf{x}}^{(t-1)}, \tilde{\mathbf{s}}^{(t-1)}, \check{\mathbf{q}}^{(t-1)}).\end{aligned}$$

■

Remark 1: Lemma 2 presents sufficient conditions on the parameters to ensure convergence of ET-DOGT. They are stronger than that in conventional DOGT. This is because the triggering behavior introduces additional noise, i.e., (13), that should be counteracted properly. For the relation between parameters, the step-size α decreases as β and L increase, and the triggering parameter decreases as the step-size decreases. In the extreme case where $\eta = 0$, one can always find an α for any parameter $\beta \in (0, 1)$. Next, we provide a set of explicit conditions. Suppose the matrix P is designed such that

$$\beta < \frac{2\sqrt{5}}{3} - 1.$$

Take

$$\kappa = \frac{\beta(1-\beta)}{12L^2(1+\beta)^2}, \quad \varepsilon = 2L^2.$$

To ensure $\rho_x > 0$, there must hold

$$\eta\gamma < \eta^*\gamma^* := \frac{3 - (1+\beta)^2}{3(4 + (1+\beta)^2)} - \frac{1}{24} > 0.$$

Note that both ρ_s and ρ_y are monotonically decreasing over $\eta\gamma$. Then ρ_s and ρ_y can be made positive by choosing α satisfying both

$$\alpha < \sqrt{\frac{\kappa(1-\beta)^3(3+\beta)}{2\beta(1-\beta)(1+\beta)^2(1+3\eta^*\gamma^*)+3\kappa L^2(1+\beta)^4}}$$

and

$$\alpha < \frac{2\beta(1-\beta)}{\beta(1-\beta)(L+2L^2)+3\kappa(1+\beta)^2L^2+6\eta^*\gamma^*\beta(1-\beta)}.$$

Finally, to ensure $\rho_q > 0$, one sets

$$\gamma > \left(\frac{6\gamma^*\eta^*(1+\beta)}{(1-\beta)}+\frac{2(1+\beta)}{1-\beta}\right)\|\mathbf{P}-I\|^2+\kappa\zeta_3,$$

and, accordingly,

$$\eta < \frac{\eta^*\gamma^*}{\gamma}.$$

Theorem 1: Suppose the premise in Lemma 2 holds. Given $\epsilon > 0$, let $T_\epsilon = \min\{t : r(y^{(t)}, \tilde{\mathbf{z}}^{(t)}, \tilde{\mathbf{s}}^{(t)}, \check{\mathbf{q}}^{(t)}) \le \epsilon\}$. Then, it holds that

i) The sequence $\lim_{\tau\to\infty}\|\tilde{\mathbf{x}}^{(\tau)}\| = 0$ and $\lim_{\tau\to\infty}\sum_{i=1}^n \nabla f_i(x_i^{(\tau)}) = 0$;
ii) $T_\epsilon = o(1/\epsilon)$.

Proof: i) By Lemma 2 and the fact that $R(\overline{x}, \tilde{\mathbf{x}}, \tilde{\mathbf{s}}, \check{\mathbf{q}})$ is bounded from below, we have that $R(\overline{x}^{(t)}, \tilde{\mathbf{x}}^{(t)}, \tilde{\mathbf{s}}^{(t)}, \check{\mathbf{q}}^{(t)})$ converges and

$$\sum_{\tau=0}^{\infty} r(\overline{x}^{(\tau)}, \tilde{\mathbf{x}}^{(\tau)}, \tilde{\mathbf{s}}^{(\tau)}, \check{\mathbf{q}}^{(\tau)}) < \infty.$$

Therefore

$$\lim_{\tau\to\infty} r(\overline{x}^{(\tau)}, \tilde{\mathbf{x}}^{(\tau)}, \tilde{\mathbf{s}}^{(\tau)}, \check{\mathbf{q}}^{(\tau)}) = 0.$$

By the definition of r, we obtain $\lim_{\tau\to\infty}\|\tilde{x}^{(\tau)}\| = 0$ and

$$\lim_{\tau\to\infty}\|\Delta\overline{x}^{(\tau)}\| = \lim_{\tau\to\infty}\left\|\frac{\alpha}{n}\sum_{i=1}^n \nabla f_i(x_i^{(\tau)})\right\| = 0.$$

ii) Upon using Lemma 2 and the definition of T_ϵ, we obtain

$$\frac{T_\epsilon}{2}\epsilon \le \sum_{\tau=\lfloor\frac{T_\epsilon}{2}\rfloor+1}^{T_\epsilon} r^{(\tau)} \le R^{(\lfloor\frac{T_\epsilon}{2}\rfloor+1)} - R^{(T_\epsilon+1)}$$

As $\epsilon \to 0$, we consider the following two possibilities. If $T_\epsilon \to \infty$, then $T_\epsilon = o(1/\epsilon)$ because of the convergence of $R^{(\tau)}$, i.e., $\sum_{\tau=\lfloor\frac{T_\epsilon}{2}\rfloor+1}^{T_\epsilon} r^{(\tau)} \to 0$. If $T_\epsilon < \infty$, then $R^{(\tau)}$ converges in finite steps. This completes the proof. ■

V. Experiments

In this section, we present experimental results on real-world datasets to demonstrate the efficiency of ET-DOGT in saving communication resources, by comparing the performance with a few recent algorithms.

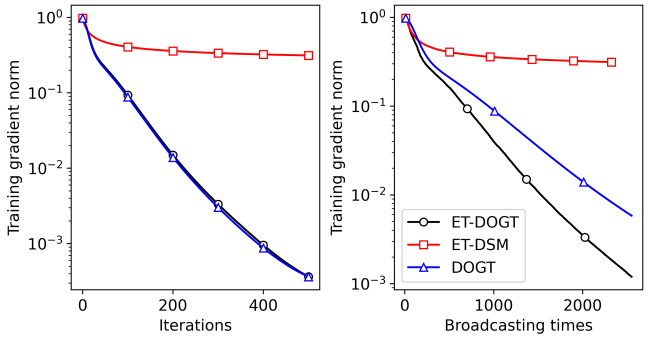

Fig. 1. Training gradient norm versus the numbers of iterations and broadcasting times.

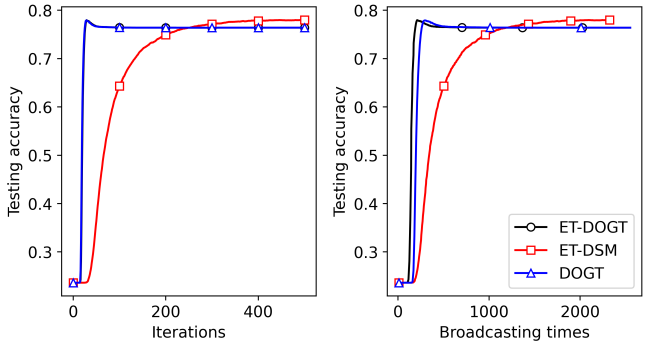

Fig. 2. Testing accuracy versus the numbers of iterations and broadcasting times.

Consider the logistic regression problem with a nonconvex regularizer on the *a9a* dataset [15]. In particular, the unshuffled dataset is evenly distributed to $n = 10$ agents, and the local objective function is given by

$$f_i(x) = \sum_{j=1}^{m_i}\log\left(1+\exp\left(-y_j^i {M_j^i}^{\mathrm{T}} x\right)\right) + \iota\sum_{l=1}^{d}\frac{(x^{[l]})^2}{1+(x^{[l]})^2}$$

where $M_j^i \in \mathbb{R}^m$ and $y_j^i \in \{-1, 1\}$ denote the j-th pair of input feature and class label for agent i, $x^{[l]}$ represents the l-th entry of x, and the regularization parameter is set as $\iota = 0.05$. For the communication network among the agents, we construct an Erdős-Rényi topology with connectivity probability 0.8. Based on it, the fastest distributed linear averaging (FDLA) matrix is chosen as the mixing matrix [16].

We compare the ET-DOGT algorithm with the original DOGT and the event-triggered distributed subgradient method (ET-DSM) in [7]. For DOGT-based algorithms, the step-size is set as $\alpha = 0.1$. For ET-DSM, the step-size is set as $1/(t+1)$ to satisfy the theoretical conditions for convergence. The parameter of the triggering condition (4) is set as $\eta = 1$ for ET-DOGT, and for ET-DSM the parameter in (6) is determined as $\rho^{(t)} = 2/(t+1)^2$. For all the methods, the initial variables are drawn from a uniform distribution over $[0, 1)$.

The results are plotted in Figs. 1 and 2. Fig. 1 depicts the change of gradient norm, i.e., $\|\sum_{i=1}^n \nabla f_i(x_i)\|$, over both

the number of iteration steps and broadcasting times under the three algorithms. The ET-DOGT algorithm presents a similar convergence speed with DOGT in terms of iteration number, however the event-triggered variant consumes much less communication resources compared to the original DOGT. ET-DSM converges much slower than the other two methods, mainly due to the decaying step-size. Fig 2 evaluates the three algorithms via testing accuracy, and the result demonstrates the same trend as in Fig. 1. In summary, the proposed ET-DOGT algorithm helps reduce the utilization of communication resources while preserving a desired convergence speed.

VI. Conclusions

This work presented an event-triggered variant of the distributed optimization algorithm with gradient tracking, and provided sufficient conditions for the step-size and the triggering parameter under which the proposed algorithm converges for nonconvex and smooth problems. The proposed triggering condition features that the threshold is determined by the change of two consecutive local updates, which adapts well to the optimization dynamics. Experimental results were presented to demonstrate the effectiveness of the proposed algorithm. Future research includes the extension to composite optimization problems, and linear convergence results under stronger assumptions such as strong convexity.

APPENDIX

A. Proof of Lemma 1

By Assumption 2, it holds that

$$\begin{aligned} &f(\overline{x}^{(t)}) - f(\overline{x}^{(t-1)}) \\ &\leq \langle \nabla f(\overline{x}^{(t-1)}), \overline{x}^{(t)} - \overline{x}^{(t-1)} \rangle + \frac{L}{2}\|\overline{x}^{(t)} - \overline{x}^{(t-1)}\|^2 \\ &= \langle \nabla f(\overline{x}^{(t-1)}) - \overline{g}^{(t-1)}, \overline{x}^{(t)} - \overline{x}^{(t-1)} \rangle \\ &\quad + \langle \overline{g}^{(t-1)}, \overline{x}^{(t)} - \overline{x}^{(t-1)} \rangle + \frac{L}{2}\|\overline{x}^{(t)} - \overline{x}^{(t-1)}\|^2. \end{aligned} \tag{14}$$

Since $\overline{g}^{(t-1)} = \alpha^{-1}(\overline{x}^{(t-1)} - \overline{x}^{(t-1)})$, we have

$$\begin{aligned} f(\overline{x}^{(t)}) - f(\overline{x}^{(t-1)}) \leq \langle \nabla f(\overline{x}^{(t-1)}) - \overline{g}^{(t-1)}, \overline{x}^{(t)} - \overline{x}^{(t-1)} \rangle \\ - \left(\frac{1}{\alpha} - \frac{L}{2}\right)\|\overline{x}^{(t)} - \overline{x}^{(t-1)}\|^2. \end{aligned} \tag{15}$$

In addition, we have

$$\begin{aligned} &\|\nabla f(\overline{x}^{(t-1)}) - \overline{g}^{(t-1)}\|^2 \\ &= \left\| n^{-1}\sum_{i=1}^{n} \nabla f(\overline{x}^{(t-1)}) - \nabla f_i(x_i^{(t-1)}) \right\|^2 \\ &\leq \frac{1}{n}\sum_{i=1}^{n} \|\nabla f(\overline{x}^{(t-1)}) - \nabla f_i(x_i^{(t-1)})\|^2 \\ &\leq \frac{L^2}{n}\|\tilde{\mathbf{x}}^{(t-1)}\|^2, \end{aligned}$$

where the first inequality is due to the convexity of norm square and Jensen's inequality, and the second inequality follows from Assumption 2. Therefore, it holds that

$$\begin{aligned} &\langle \nabla f(\overline{x}^{(t-1)}) - \overline{g}^{(t-1)}, \overline{x}^{(t)} - \overline{x}^{(t-1)} \rangle \\ &\leq \frac{\varepsilon}{2}\|\overline{x}^{(t)} - \overline{x}^{(t-1)}\|^2 + \frac{1}{2\varepsilon}\|\nabla f(\overline{x}^{(t-1)}) - \overline{g}^{(t-1)}\|^2 \\ &\leq \frac{\varepsilon}{2}\|\overline{x}^{(t)} - \overline{x}^{(t-1)}\|^2 + \frac{L^2}{2n\varepsilon}\|\tilde{\mathbf{x}}^{(t-1)}\|^2, \forall \varepsilon > 0. \end{aligned} \tag{16}$$

Combining (15) and (16) completes the proof.

References

[1] T. Yang, X. Yi, J. Wu, Y. Yuan, D. Wu, Z. Meng, Y. Hong, H. Wang, Z. Lin, and K. H. Johansson, "A survey of distributed optimization," *Annual Reviews in Control*, vol. 47, pp. 278–305, 2019.

[2] A. Kashyap, T. Basar, and R. Srikant, "Quantized consensus," *Automatica*, vol. 43, pp. 1192–1203, 2007.

[3] H. Tang, S. Gan, C. Zhang, T. Zhang, and J. Liu, "Communication compression for decentralized training," *Advances in Neural Information Processing Systems*, vol. 31, 2018.

[4] A. Koloskova, S. Stich, and M. Jaggi, "Decentralized stochastic optimization and gossip algorithms with compressed communication," in *International Conference on Machine Learning*, pp. 3478–3487, PMLR, 2019.

[5] Y. Xiong, L. Wu, K. You, and L. Xie, "Quantized distributed gradient tracking algorithm with linear convergence in directed networks," *IEEE Transactions on Automatic Control*, 2022.

[6] D. V. Dimarogonas, E. Frazzoli, and K. H. Johansson, "Distributed event-triggered control for multi-agent systems," *IEEE Transactions on automatic control*, vol. 57, no. 5, pp. 1291–1297, 2011.

[7] Y. Kajiyama, N. Hayashi, and S. Takai, "Distributed subgradient method with edge-based event-triggered communication," *IEEE Transactions on Automatic Control*, vol. 63, no. 7, pp. 2248–2255, 2018.

[8] C. Liu, H. Li, and Y. Shi, "Resource-aware exact decentralized optimization using event-triggered broadcasting," *IEEE Transactions on Automatic Control*, vol. 66, no. 7, pp. 2961–2974, 2020.

[9] M. Li, L. Su, and T. Liu, "Distributed optimization with event-triggered communication via input feedforward passivity," *IEEE Control Systems Letters*, vol. 5, no. 1, pp. 283–288, 2020.

[10] T. Adachi, N. Hayashi, and S. Takai, "Distributed gradient descent method with edge-based event-driven communication for non-convex optimization," *IET Control Theory & Applications*, vol. 15, no. 12, pp. 1588–1598, 2021.

[11] N. Singh, D. Data, J. George, and S. Diggavi, "Sparq-sgd: Event-triggered and compressed communication in decentralized optimization," *IEEE Transactions on Automatic Control*, 2022.

[12] A. Daneshmand, G. Scutari, and V. Kungurtsev, "Second-order guarantees of distributed gradient algorithms," *SIAM Journal on Optimization*, vol. 30, no. 4, pp. 3029–3068, 2020.

[13] J. Xu, S. Zhu, Y. C. Soh, and L. Xie, "Augmented distributed gradient methods for multi-agent optimization under uncoordinated constant stepsizes," in *2015 54th IEEE Conference on Decision and Control (CDC)*, pp. 2055–2060, IEEE, 2015.

[14] C. Liu, X. Wu, X. Yi, Y. Shi, and K. H. Johansson, "Rate analysis of dual averaging for nonconvex distributed optimization," *arXiv preprint arXiv:2211.06914*, 2022.

[15] C.-C. Chang and C.-J. Lin, "Libsvm: a library for support vector machines," *ACM transactions on intelligent systems and technology (TIST)*, vol. 2, no. 3, pp. 1–27, 2011.

[16] L. Xiao and S. Boyd, "Fast linear iterations for distributed averaging," *Systems & Control Letters*, vol. 53, no. 1, pp. 65–78, 2004.