2025 IEEE 64th Conference on Decision and Control (CDC)
December 10-12, 2025. Rio de Janeiro, Brazil

An Optimistic Gradient Tracking Method for Distributed Minimax Optimization

Yan Huang, Jinming Xu, Jiming Chen, and Karl Henrik Johansson

2025 IEEE 64th Conference on Decision and Control (CDC) | 979-8-3315-2627-6/25/$31.00 ©2025 IEEE | DOI: 10.1109/CDC57313.2025.11313000

***Abstract*— This paper studies the distributed minimax optimization problem over networks. To enhance convergence performance, we propose a distributed optimistic gradient tracking method, termed DOGT, which solves a surrogate function that captures the similarity between local objective functions to approximate a centralized optimistic approach locally. Leveraging a Lyapunov-based analysis, we prove that DOGT achieves linear convergence to the optimal solution for strongly convex-strongly concave objective functions while remaining robust to the heterogeneity among them. Moreover, by integrating an accelerated consensus protocol, the accelerated DOGT (ADOGT) algorithm achieves an optimal convergence rate of $\mathcal{O}\left(\kappa \log\left(\epsilon^{-1}\right)\right)$ and communication complexity of $\mathcal{O}\left(\kappa \log\left(\epsilon^{-1}\right)/\sqrt{1-\sqrt{\rho_W}}\right)$ for a suboptimality level of $\epsilon > 0$, where κ is the condition number of the objective function and ρ_W is the spectrum gap of the network. Numerical experiments illustrate the effectiveness of the proposed algorithms.**

I. Introduction

Minimax optimization has gained significant attention over the past decade due to its broad applications in robust optimization [1], [2], game theory [3], and generative adversarial networks (GANs) [4], [5], among others. With the increasing scale of data and devices [6], distributed optimization has emerged as a key approach in large-scale optimization [7], machine learning [8], and control [9].

In this paper, we consider the distributed minimax optimization problem jointly solved by a network of n nodes:

$$\min_{x\in\mathbb{R}^p}\max_{y\in\mathbb{R}^d} f(x,y) := \frac{1}{n}\sum_{i=1}^{n} f_i(x,y), \tag{1}$$

where f_i, $i = 1,\dots,n$, is the local objective function, $x \in \mathbb{R}^p$ and $y \in \mathbb{R}^d$ are primal and dual variables to be minimized and maximized, respectively. This formulation is general, as it recovers the traditional distributed minimization problem [7] as a special case, thereby capturing a broader class of applications. For instance, in distributed training of Wasserstein GANs [10], [11], nodes collaborate to minimize the generator while simultaneously maximizing the discriminator, thus enabling data-parallel accelerated training. A straightforward approach to solving this problem is to apply centralized minimax algorithms, such as gradient descent ascent (GDA) [12], optimistic gradient descent ascent (OGDA) [13] and extragradient (EG) [14], in combination with the averaging consensus protocol [15]. For instance, Deng and Mahdavi [16] proposed Local SGDA, which integrates FedAvg from federated learning [17] with stochastic GDA to reduce communication costs. Chen et al. [18] introduced a variance-reduction-based decentralized GDA method, achieving improved sample complexity under the stronger assumption of average smoothness for each sample. However, as shown in [19], centralized GDA methods do not achieve linear convergence for bilinear functions. We will show later that this limitation also applies to the distributed GDA (DGDA) algorithm.

OGDA and EG methods have gained popularity in recent literature due to their ability to achieve linear convergence rates for strongly convex-strongly concave and bilinear objective functions [19], [20], as well as their effectiveness in training GANs [13], [21]. In decentralized settings, Liu et al. [10] proposed a distributed optimistic gradient descent ascent method, DPOSG, for training large-scale GANs and analyzed its convergence for nonconvex-nonconcave (NC-NC) objective functions. Beznosikov et al. [22] introduced an EG-based FedAvg method and proved that with an accelerated consensus protocol, it achieves a near-optimal convergence rate, matching the lower bound for (strongly) convex-(strongly) concave objective functions up to logarithmic factors. Liu et al. [23] developed a decentralized proximal point method for NC-NC problems, which requires solving a subproblem at each iteration. However, the above methods establish convergence under assumptions such as uniformly bounded gradient norms [10] or bounded feasible domains [23], [22]. Otherwise, a decaying stepsize is required to remove a steady-state error [24]. These assumptions restrict the applicability of their theoretical results in many real-world tasks due to their inability to handle the heterogeneity of local objective functions [8], i.e., the gradients of each f_i differ, leading to distinct local optimas. This heterogeneity is recognized as a fundamental challenge in distributed optimization.

Gradient tracking (GT) methods [25], [26] are widely used in distributed minimization to mitigate the effects of data/function heterogeneity. Recent studies have explored their application to minimax optimization. For instance, Wai et al. [27] introduced a GDA-based GT method for multi-agent reinforcement learning and established a linear convergence rate for SC-SC settings. Mukherjee et al. [28] combined GT with EG methods, achieving linear convergence for strongly convex-strongly concave (SC-SC) objectives

This work has been supported in parts by National Natural Science Foundation of China under Grants 62373323, 62088101; and in parts by the Knut and Alice Wallenberg Foundation Wallenberg Scholar Grant, the Swedish Research Council Distinguished Professor Grant 2017-01078, and the Swedish Foundation for Strategic Research SUCCESS FUS21-0026.

Yan Huang and Karl Henrik Johansson are with the Division of Decision and Control Systems, School of EECS, KTH Royal Institute of Technology, SE-100 44 Stockholm, Sweden. Email: {yahuang, kallej}@kth.se

Jinming Xu and Jiming Chen are with the College of Control Science and Engineering, Zhejiang University, Hangzhou 310027, China. Email: {jimmyxu, cjm}@zju.edu.cn

without assuming a bounded gradient norm or feasible domain. However, there remains a gap between the convergence rates in [27], [28] and the established lower bound in [22]. Another line of research utilizes second-order similarity based on the mirror-descent method [29], [30], [31], where the difference between the second-order derivative matrices of f_i is bounded by a finite quantity δ. This approach improves algorithmic efficiency when δ is smaller than the smoothness condition number. However, these methods often incur high computational costs due to the need to solve a subproblem.

To address these challenges, inspired by the mirror-descent-type method in decentralized settings [29], [30], we propose a distributed optimistic gradient tracking method, termed DOGT, for solving Problem (1). This algorithm solves a surrogate objective function that captures the similarity between the local gradient and an estimator of the global gradient, using one-step approximate solutions to avoid the need for inner loops to solve subproblems. Theoretically, we prove that DOGT achieves a linear convergence rate and outperforms the result in [28] by a factor of $\mathcal{O}\left(\kappa^{1/3}\right)$ when the graph connectivity is strong. Furthermore, by integrating an accelerated consensus protocol with prior knowledge of the graph, we propose ADOGT and prove that it achieves an optimal convergence rate of $\mathcal{O}\left(\kappa \log\left(\epsilon^{-1}\right)\right)$ and communication complexity of $\mathcal{O}\left(\kappa \log\left(\epsilon^{-1}\right)/\sqrt{1-\sqrt{\rho_W}}\right)$, matching the lower bound established in [22] for deterministic settings under the assumptions considered in this work. Numerical results validate our theoretical findings.

Paper organization. The rest of the paper is structured as follows: Section II formulates the distributed minimax optimization problem with several common assumptions and presents the algorithm design of DOGT and ADOGT. Section III provides the main convergence results of the proposed algorithms. Section IV presents numerical experiments to validate our theoretical findings. Finally, Section V concludes the paper, and several supporting lemmas for the proof of the main results are provided in the appendix.

Notations. Throughout this paper, we adopt the following notations: $\langle\cdot,\cdot\rangle$ is the inner product of vectors, $\|\cdot\|$ represents the Frobenius norm, $\lceil\cdot\rceil$ indicates the ceiling operation, $\mathbf{1}$ represents the all-ones vector, $\mathbf{I}$ denotes the identity matrix, and $\mathbf{J} = \mathbf{1}\mathbf{1}^\top/n$ denotes the averaging matrix.

II. Problem Formulation and Algorithm Design

A. Distributed Minimax Optimization Problem

To solve the distributed minimax optimization problem (1) in a decentralized manner, we consider each node $i \in [n]$ maintaining a local copy of the global primal and dual variables, i.e., $x_i \in \mathbb{R}^p, y_i \in \mathbb{R}^d$, and optimize the following constrained problem, which has the same optimal solution as the original problem (1):

$$\begin{aligned} &\min_{\mathbf{x}\in\mathbb{R}^{n\times p}} \max_{\mathbf{y}\in\mathbb{R}^{n\times d}} F(\mathbf{x},\mathbf{y}) := \frac{1}{n}\sum_{i=1}^{n} f_i(x_i,y_i), \\ &\text{s.t.} \quad x_i = x_j,\ y_i = y_j,\ \forall i,j = 1,\ldots,n \end{aligned} \tag{2}$$

where

$$\begin{aligned} \mathbf{x} &:= [x_1, x_2, \ldots, x_n]^\top \in \mathbb{R}^{n\times p}, \\ \mathbf{y} &:= [y_1, y_2, \ldots, y_n]^\top \in \mathbb{R}^{n\times d} \end{aligned}$$

are the collections of the primal and dual variables, respectively, which are required to reach consensus. Moreover, the nodes are connected over a decentralized network, and its topology is modeled as an undirected graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$, where $\mathcal{V} = \{1, 2, ..., n\}$ denotes the set of agents, and $\mathcal{E} \subseteq \mathcal{V} \times \mathcal{V}$ denotes the set of edges consisting of ordered pairs (i, j) modeling the communication link from j to i. Each node i only communicates with its neighbor $\mathcal{N}_i = \{j \,|\, j \neq i,\ (i,j) \in \mathcal{E}\}$, including i itself. Then, we make the following commonly used assumptions on the differentiable objective function f_i, its gradient ∇f_i, and the graph $\mathcal{G}$.

Assumption 1 (Convexity and Concavity): Each objective function $f_i(x,y)$ is μ-strongly convex in x and μ-strongly concave in y, i.e., $\forall x, x' \in \mathbb{R}^p$, $\forall y, y' \in \mathbb{R}^d$ and $\mu > 0$,

$$\begin{aligned} &f_i(x',y) - f_i(x,y) \\ &\geqslant \langle \nabla_x f_i(x,y), x' - x\rangle + \frac{\mu}{2}\|x - x'\|^2, \\ &f_i(x,y') - f_i(x,y) \\ &\leqslant \langle \nabla_y f_i(x,y), y' - y\rangle - \frac{\mu}{2}\|y - y'\|^2. \end{aligned}$$

Assumption 2 (Smoothness): Each objective function $f_i(x,y)$ is jointly L-smooth in x and y, i.e., $\forall x, x' \in \mathbb{R}^p$ and $\forall y, y' \in \mathbb{R}^d$, there exists a constant $L > 0$ such that for $z \in \{x, y\}$,

$$\begin{aligned} &\|\nabla_z f_i(x,y) - \nabla_z f_i(x',y')\|^2 \\ &\leqslant L^2\left(\|x - x'\|^2 + \|y - y'\|^2\right). \end{aligned}$$

Assumption 3 (Graph connectivity): The weight matrix $W = [w_{i,j}]_{i,j=1}^n$ induced by graph $\mathcal{G}$ is doubly stochastic, i.e., $W\mathbf{1} = \mathbf{1}, \mathbf{1}^\top W = \mathbf{1}^\top$ and $\rho_W := \|W - \mathbf{J}\|_2^2 < 1$.

B. Algorithm Design

Motivated by the SONATA algorithm [29] originally designed for minimization problems, we propose a distributed optimistic gradient tracking (DOGT) algorithm for solving the distributed minimax problem (2). In particular, we consider the following surrogate function for minimizing on $x_i \in \mathbb{R}^p$ and maximizing on $y_i \in \mathbb{R}^d$:

$$\begin{aligned} &f_i(x_i,y_i) + \frac{1}{2\gamma}\|x_i - x_{i,k}\|^2 + \frac{1}{2\gamma}\|y_i - y_{i,k}\|^2 \\ &\quad + \underbrace{\left(p_{i,k} - \nabla_x f_i(x_{i,k},y_{i,k})\right)}_{\text{similarity in } x}{}^\top (x_i - x_{i,k}) \\ &\quad + \underbrace{\left(q_{i,k} - \nabla_y f_i(x_{i,k},y_{i,k})\right)}_{\text{similarity in } y}{}^\top (y_i - y_{i,k}), \end{aligned} \tag{3}$$

where $p_{i,k}$ and $q_{i,k}$ are the gradient tracking variables used for estimating the global gradient at iteration k with respect to x and y, receptively. This surrogate function captures the difference between the local and global gradients. Solving it in a decentralized manner helps to asymptotically reduce this gap and approximate a centralized optimistic method, thereby improving the convergence [29]. Specifically, according to the

first-order optimality condition of (3), and taking the primal variable x as an example, we derive the following proximal point method:

$$\begin{aligned} x_{i,k+1} = x_{i,k} &- \gamma \nabla_x f_i\left(x_{i,k+1}, y_{i,k+1}\right) \\ &- \gamma\left(p_{i,k} - \nabla_x f_i\left(x_{i,k}, y_{i,k}\right)\right). \end{aligned} \tag{4}$$

To obtain an algorithm that is easy to deploy without requiring the exact solution of the surrogate function (3), we use the approximation

$$\begin{aligned} &\nabla_x f_i\left(x_{i,k+1}, y_{i,k+1}\right) \\ &\approx 2\nabla_x f_i\left(x_{i,k}, y_{i,k}\right) - \nabla_x f_i\left(x_{i,k-1}, y_{i,k-1}\right), \end{aligned} \tag{5}$$

which has an error bounded by $o\left(\gamma^2\right)$ [19]. Based on this, we propose the DOGT algorithm, an optimistic gradient descent ascent method with gradient tracking. The pseudo-code for DOGT is provided in Algorithm 1. Notably, DOGT adopts a single-loop structure, making it more suitable for practical deployment than mirror-descent-based methods [29], [30], particularly in scenarios with data heterogeneity.

Algorithm 1 Distributed Optimistic Gradient Tracking (DOGT)

Initialization: Initial points $x_{i,0} \in \mathbb{R}^p$, $y_{i,0} \in \mathbb{R}^d$, initial gradient tracking variables $\nabla_x f_{i,-1} = p_{i,0} = \nabla_x f_i\left(x_{i,0}, y_{i,0}\right)$, $\nabla_y f_{i,-1} = q_{i,0} = \nabla_y f_i\left(x_{i,0}, y_{i,0}\right)$, and stepsize $\gamma > 0$.

1: **for** iteration $k = 0, 1, \ldots,$ each node $i \in [n]$, **do**

2: Optimistic gradient descent ascent with gradient tracking variables:

$$\begin{aligned} x_{i,k+1} &= x_{i,k} - \gamma\left(p_{i,k} + \nabla_x f_{i,k} - \nabla_x f_{i,k-1}\right), \\ y_{i,k+1} &= y_{i,k} + \gamma\left(q_{i,k} + \nabla_y f_{i,k} - \nabla_y f_{i,k-1}\right). \end{aligned}$$

3: Compute gradient $\nabla_x f_{i,k+1}$ and $\nabla_y f_{i,k+1}$.

4: Update gradient tracking variables:

$$\begin{aligned} p_{i,k+1} &= p_{i,k} + \nabla_x f_{i,k+1} - \nabla_x f_{i,k}, \\ q_{i,k+1} &= q_{i,k} + \nabla_y f_{i,k+1} - \nabla_y f_{i,k}. \end{aligned}$$

5: Inter-node communication for packaged message $\theta_{i,k+1} := \{x_{i,k+1}, y_{i,k+1}, p_{i,k+1}, q_{i,k+1}\}$:

$$\theta_{i,k+1} \leftarrow \sum_{j \in \mathcal{N}_i} w_{i,j} \theta_{i,k+1}.$$

6: **end for**

For brevity, we introduce the following notations for the gradients of all nodes at iteration k:

$$\begin{aligned} \nabla_x F_k &:= \left[\ldots, \nabla_x f_i\left(x_{i,k}, y_{i,k}\right), \ldots\right]^\top \in \mathbb{R}^{n \times p}, \\ \nabla_y F_k &:= \left[\ldots, \nabla_y f_i\left(x_{i,k}, y_{i,k}\right), \ldots\right]^\top \in \mathbb{R}^{n \times d}. \end{aligned}$$

Then, the DOGT algorithm can be rewritten in the following compact form:

$$\begin{aligned} \mathbf{x}_{k+1} &= W\left(\mathbf{x}_k - \gamma\left(\mathbf{p}_k + \nabla_x F_k - \nabla_x F_{k-1}\right)\right), \\ \mathbf{y}_{k+1} &= W\left(\mathbf{y}_k + \gamma\left(\mathbf{q}_k + \nabla_y F_k - \nabla_y F_{k-1}\right)\right), \\ \mathbf{p}_{k+1} &= W\left(\mathbf{p}_k + \nabla_x F_{k+1} - \nabla_x F_k\right), \\ \mathbf{q}_{k+1} &= W\left(\mathbf{q}_k + \nabla_y F_{k+1} - \nabla_y F_k\right). \end{aligned} \tag{6}$$

where the collection of the gradient tracking variables for the primal and dual decision variables are denoted as follows:

$$\begin{aligned} \mathbf{p}_k &:= \left[p_{1,k}, p_{2,k}, \ldots, p_{n,k}\right]^\top \in \mathbb{R}^{n \times p}, \\ \mathbf{q}_k &:= \left[q_{1,k}, q_{2,k}, \ldots, q_{n,k}\right]^\top \in \mathbb{R}^{n \times d}. \end{aligned}$$

Furthermore, taking the average over all nodes on both sides of the algorithm yields the following equations:

$$\begin{aligned} \bar{x}_{k+1} &:= \frac{\mathbf{1}^\top}{n}\mathbf{x}_{k+1} = \bar{x}_k - \gamma\frac{\mathbf{1}^\top}{n}\left(2\nabla_x F_k - \nabla_x F_{k-1}\right), \\ \bar{y}_{k+1} &:= \frac{\mathbf{1}^\top}{n}\mathbf{y}_{k+1} = \bar{y}_k + \gamma\frac{\mathbf{1}^\top}{n}\left(2\nabla_y F_k - \nabla_y F_{k-1}\right), \\ \bar{p}_{k+1} &:= \frac{\mathbf{1}^\top}{n}\mathbf{p}_{k+1} = \frac{\mathbf{1}^\top}{n}\nabla_x F_{k+1}, \\ \bar{q}_{k+1} &:= \frac{\mathbf{1}^\top}{n}\mathbf{q}_{k+1} = \frac{\mathbf{1}^\top}{n}\nabla_y F_{k+1}. \end{aligned} \tag{7}$$

It can be observed that, on average, DOGT employs a centralized OGDA update scheme, with the gradient tracking variables asymptotically aligning with the average gradient across all nodes. The consensus of these variables is ensured by the weighting matrix W, which satisfies Assumption 3.

To further enhance the convergence of DOGT, we incorporate the accelerated gossip consensus protocol [32] and obtained ADOGT, leveraging prior knowledge of the spectrum gap ρ_W. Specifically, we replace the communication step of DOGT (cf. line 5 in Algorithm 1) with the following update, executed a finite number of T times at each iteration:

$$\theta_{i,k+1} \leftarrow (1+\eta)\sum_{j \in \mathcal{N}_i} w_{i,j}\theta_{j,k+1} - \eta\theta_{i,k+1}, \tag{8}$$

where $\eta = \left(1 - \sqrt{1-\rho_W}\right) / \left(1 + \sqrt{1-\rho_W}\right)$. Then, we obtain a generated weight matrix M_T of ADOGT as follows:

$$M_{t+1} = (1+\eta)\, W M_t - \eta M_{t-1}, \tag{9}$$

where $t = 0, 1, \ldots, T-1$, and $M_{-1} = M_0 = \mathbf{I}$. We will show that ADOGT achieves the optimal convergence rate with this accelerated weight matrix (cf., Theorem 2).

III. Convergence Analysis

In this section, we provide the convergence analysis of DOGT and ADOGT for SC-SC smooth objective functions. For simplicity, we further denote

$$\mathbf{z}_k := \left[\mathbf{x}_k, \mathbf{y}_k\right], \ \mathbf{r}_k := \left[\mathbf{p}_k, -\mathbf{q}_k\right], \ \nabla F_k := \left[\nabla_x F_k, -\nabla_y F_k\right].$$

Then, letting

$$\boldsymbol{\varepsilon}_k = \nabla F_{k+1} - \nabla F_k - \nabla F_k + \nabla F_{k-1},$$

we get the update rule of $\bar{z}_k := \mathbf{1}^\top \mathbf{z}_k / n$ as follows:

$$\bar{z}_{k+1} = \bar{z}_k - \gamma\frac{\mathbf{1}^\top}{n}\nabla F_{k+1} + \gamma\frac{\mathbf{1}^\top}{n}\boldsymbol{\varepsilon}_k, \tag{10}$$

and

$$\begin{aligned} &\bar{z}_{k+1} - \gamma\frac{\mathbf{1}^\top}{n}\left(\nabla F_{k+1} - \nabla F_k\right) \\ &= \bar{z}_k - \gamma\frac{\mathbf{1}^\top}{n}\left(\nabla F_k - \nabla F_{k-1}\right) - \gamma\frac{\mathbf{1}^\top}{n}\nabla F_{k+1}. \end{aligned} \tag{11}$$

A. Linear Convergence

To prove the convergence of DOGT, we define the following Lyapunov function:

$$\begin{aligned}\Psi_k := &\left\|\Xi_k\right\|^2 + \frac{\gamma L}{n}\left\|\mathbf{z}_k - \mathbf{z}_{k-1}\right\|^2 \\ &+ c_1\left\|\mathbf{z}_k - \mathbf{1}\bar{z}_k\right\|^2 + c_2\left\|\mathbf{r}_k - \mathbf{1}\bar{r}_k\right\|^2,\end{aligned} \tag{12}$$

where c_1 and c_2 are coefficients defined in (19), and the optimality gap to the optimal solution $z^* := [x^*, y^*]$ of problem (2) is defined as

$$\Xi_k := \bar{z}_k - \gamma\frac{\mathbf{1}^\top}{n}\left(\nabla F_k - \nabla F_{k-1}\right) - z^*. \tag{13}$$

Then, with the help of Lemmas 1-4 in the appendix, the following theorem shows a linear convergence rate of DOGT.

Theorem 1: Consider the DOGT algorithm as depicted in (6). Suppose Assumptions 1-3 hold. Let the stepsize

$$\gamma \leqslant \min\left\{\frac{1}{64L}, \frac{\left(1-\rho_W\right)^2}{144L\sqrt{\rho_W}}\right\}. \tag{14}$$

Then, we have for all $k \geqslant 0$,

$$\Psi_k \leqslant \left(1 - \min\left\{\frac{3\gamma\mu}{4}, \frac{1-\rho_W}{8}\right\}\right)^k \Psi_0. \tag{15}$$

Proof: By Lemma 1 and Lemma 4, we can obtain that

$$\begin{aligned}&\left\|\Xi_{k+1}\right\|^2 + \frac{\gamma L}{n}\left\|\mathbf{z}_{k+1} - \mathbf{z}_k\right\|^2 \\ &\leqslant \left(1 - \frac{3\gamma\mu}{4}\right)\left\|\Xi_k\right\|^2 + \frac{4\gamma^3 L^3}{n}\left\|\mathbf{z}_k - \mathbf{z}_{k-1}\right\|^2 \\ &\quad + \frac{9\gamma L}{n}\left\|\mathbf{z}_k - \mathbf{1}\bar{z}_k\right\|^2 + \frac{9\gamma^3 L}{n\left(1-\rho_W\right)}\left\|\mathbf{r}_k - \mathbf{1}\bar{r}_k\right\|^2 \\ &\quad - \left(\frac{\gamma^2}{4n} - \frac{8\gamma^3 L}{n}\right)\left\|\nabla F\left(\mathbf{1}\bar{z}_k\right)\right\|^2.\end{aligned} \tag{16}$$

Letting $\gamma \leqslant 1/(8L)$ such that

$$1 - \frac{3\gamma\mu}{4} \geqslant 1 - \frac{3\mu}{32L} \geqslant 4\gamma^2 L^2 + 4\gamma L,$$

we can obtain

$$\begin{aligned}&\left\|\Xi_{k+1}\right\|^2 + \frac{\gamma L}{n}\left\|\mathbf{z}_{k+1} - \mathbf{z}_k\right\|^2 \\ &\leqslant \left(1 - \frac{3\gamma\mu}{4}\right)\left(\left\|\Xi_k\right\|^2 + \left\|\mathbf{z}_k - \mathbf{z}_{k-1}\right\|^2\right) \\ &\quad + \frac{9\gamma L}{n}\left\|\mathbf{z}_k - \mathbf{1}\bar{z}_k\right\|^2 + \frac{9\gamma^3 L}{n\left(1-\rho_W\right)}\left\|\mathbf{r}_k - \mathbf{1}\bar{r}_k\right\|^2 \\ &\quad - \frac{4\gamma L}{n}\left\|\mathbf{z}_k - \mathbf{z}_{k-1}\right\|^2 - \left(\frac{\gamma^2}{4n} - \frac{8\gamma^3 L}{n}\right)\left\|\nabla F\left(\mathbf{1}\bar{z}_k\right)\right\|^2.\end{aligned} \tag{17}$$

Then, combining Lemma 2 and 3 , we can obtain the contraction of the Lyapunov function:

$$\begin{aligned}&\Psi_{k+1} \\ &\leqslant \left(1 - \min\left\{\frac{3\gamma\mu}{4}, \frac{1-\rho_W}{8}\right\}\right)\Psi_k \\ &\quad + \left(\frac{8\gamma^2 L^4\rho_W}{1-\rho_W}c_2 + \frac{4\gamma^2\rho_W L^2}{1-\rho_W}c_1 - \frac{4\gamma^2 L^2}{n}\right)\left\|\mathbf{z}_k - \mathbf{z}_{k-1}\right\|^2 \\ &\quad + \left(\frac{9L^2\rho_W}{1-\rho_W}c_2 + \frac{9\gamma L}{n} - \frac{1-\rho_W}{4}c_1\right)\left\|\mathbf{z}_k - \mathbf{1}\bar{z}_k\right\|^2 \\ &\quad + \left(\frac{4\gamma^2\rho_W}{1-\rho_W}c_1 + \frac{9\gamma^3 L}{n\left(1-\rho_W\right)} - \frac{1-\rho_W}{8}c_2\right)\left\|\mathbf{r}_k - \mathbf{1}\bar{r}_k\right\|^2 \\ &\quad - \left(\frac{\gamma^2}{4n} - \frac{8\gamma^3 L}{n} - c_2\frac{16\gamma^2 L^2\rho_W}{1-\rho_W}\right)\left\|\nabla F\left(\mathbf{1}\bar{z}_k\right)\right\|^2.\end{aligned} \tag{18}$$

Set the parameters of the Lyapunov function as follows:

$$c_1 = \frac{72\gamma L}{n\left(1-\rho_W\right)}, \quad c_2 = \frac{4608\gamma^3 L}{n\left(1-\rho_W\right)^3}. \tag{19}$$

Then, letting the stepsize further satisfy (14) such that the coefficients of the last four terms on the right-hand side of the inequality are all less than or equal to 0, we complete the proof. ■

Remark 1: Theorem 1 shows that DOGT converges to the optimal solution of problem (2) at a linear rate. And, by the upper bound of the stepsize (14), we can drive that it reaches a suboptimality level of $\epsilon > 0$ in at most K iterations:

$$K = \mathcal{O}\left(\left(\kappa\left(1 + \frac{\sqrt{\rho_W}}{\left(1-\rho_W\right)^2}\right) + \frac{1}{1-\rho_W}\right)\log\left(\epsilon^{-1}\right)\right), \tag{20}$$

where $\kappa := L/\mu$ denotes the condition number of the overall objective function, $\mathcal{O}(\cdot)$ hides the constants. Compared to the GT-EG method [28], DOGT achieves at least the same convergence rate and improves it by a factor of $\mathcal{O}\left(\kappa^{1/3}\right)$ when ρ_W is small, indicating strong network connectivity.

B. Optimal Convergence Rate

For the ADOGT algorithm with accelerated consensus protocol (8), the following theorem gives the optimal convergence rate and communication complexity matching the existing lower bound in deterministic settings [22].

Theorem 2: Suppose Assumptions 1-3 hold. Let the stepsize satisfy (14), and the number of communication steps at each round $T = \lceil \ln(2)/\sqrt{1-\sqrt{\rho_W}} \rceil$. Then, for a given ρ_W, the ADOGT algorithm achieves a linear rate of $\mathcal{O}\left(\kappa\log\left(\epsilon^{-1}\right)\right)$, and the number of communication rounds R required to reach a suboptimality level of $\epsilon > 0$ is

$$R = \mathcal{O}\left(\frac{\kappa}{\sqrt{1-\sqrt{\rho_W}}}\log\left(\epsilon^{-1}\right)\right). \tag{21}$$

Proof: The proof of ADOGT follows a similar approach to that of DOGT, with the weight matrix W replaced by M_T. In specific, after executing T steps of the accelerated consensus (8), we obtain an equivalent weight matrix M_T generated by (9). Then, by Proposition 3 in [32], we get

$$\rho_M := \left\|M_T - \mathbf{J}\right\|^2 \leqslant 2\left(1 - \sqrt{1-\sqrt{\rho_W}}\right)^{2T}. \tag{22}$$

Letting the number of communication steps at each round $T = \lceil \ln(2)/\sqrt{1-\sqrt{\rho_W}} \rceil$, we have $1-\rho_M \geqslant 1/2$. Then, replacing ρ_W in the iteration complexity of DOGT in (20) with ρ_M and noticing that the total number of communication steps equals the iterations multiplied by T, we obtain the optimal linear convergence rate and communication complexity. ■

IV. Numerical Experiments

In this section, we conduct numerical experiments to compare the convergence performance between DGDA, D-OGDA, DOGT, and ADOGT. To this end, we consider the following synthetic example:

$$f_i(x_i, y_i) = x_i^T y_i + \frac{\mu}{2}\|x_i - a_i\|^2 - \frac{\mu}{2}\|y_i - b_i\|^2, \quad (23)$$

where $i \in [n]$, $a_i \in \mathbb{R}^p$ and $b_i \in \mathbb{R}^d$ are different for each nodes. We consider $n = 16$ nodes connected as an undirected ring graph. We set the dimensions $p = d = 2$, the regularization constant $\mu = 0.1$, and the stepsize $\gamma = 0.1$, respectively. For ADOGT, we set the number of communication steps at each round $T = 4$ as suggested by Theorem 2. Moreover, we set the optimal solution $z^* = [0, 0; 0, 0]$ by choosing $\sum_{i=1}^n a_i = \sum_{i=1}^n b_i = [0; 0]$.

In Fig. 1, we plot the trajectory with respect to the primal and dual decision variables, the residual $1/n\|\mathbf{z}_k - \mathbf{1}z^*\|^2$, and the consensus error $1/n\|\mathbf{z}_k - \mathbf{1}\bar{z}\|$ for each algorithm, respectively. The trajectories show that DOGT and ADOGT converge to the optimal solution, with ADOGT achieving faster convergence due to the integration of an accelerated consensus protocol. This demonstrates the robustness of GT-based methods against function heterogeneity. Instead, D-OGDA converges to a non-optimal point, and DGDA exhibits periodic oscillations without achieving convergence. Their non-zero residuals further support this behavior. Additionally, we observe that the consensus errors of DGDA and D-OGDA do not converge to zero. This is due to the heterogeneity of the objective functions, as also observed in [24], which leads to inconsistent optimal points across nodes. In contrast, DOGT and ADOGT remain robust to this issue. These results demonstrate the effectiveness of the proposed algorithms.

V. Conclusion

In this paper, we have proposed a distributed optimistic gradient tracking method, DOGT, along with its accelerated variant, ADOGT, for solving distributed minimax optimization problems over networks. We have also provided rigorous theoretical analysis to show that DOGT achieves a linear convergence rate to the optimal solution for SC-SC objective functions. Furthermore, the proposed ADOGT integrated with accelerated consensus protocol achieves an optimal convergence rate and communication complexity that matches the existing lower bound. Numerical experiments demonstrate the effectiveness of the proposed algorithms. Future work will focus on the design and analysis of algorithms in stochastic and non-convex settings, as well as exploring their potential applications in the distributed training of GANs.

Appendix

In this section, we introduce several supporting lemmas for the proof of the main results, highlighting important contraction properties of the consensus error, gradient tracking error, and optimality gap.

Lemma 1: Suppose Assumptions 2 and 3 hold. Then, we have for all $k \geqslant 0$,

$$\begin{aligned}&\|\mathbf{z}_{k+1} - \mathbf{z}_k\|^2 \\ &\leqslant 4\gamma^2 L^2 \|\mathbf{z}_k - \mathbf{z}_{k-1}\|^2 \quad + \left(4 + 8\gamma^2 L^2\right)\|\mathbf{z}_k - \mathbf{1}\bar{z}_k\|^2 \\ &\quad + 8\gamma^2\|\mathbf{r}_k - \mathbf{1}\bar{r}_k\|^2 + 8n\gamma^2\|\nabla f(\bar{z}_k)\|^2.\end{aligned} \quad (24)$$

Proof: See Appendix in [33]. ■

Lemma 2 (Consensus error): Suppose Assumptions 2 and 3 hold. Let the stepsize satisfy $\gamma \leqslant 1/(4L)$, we get $\forall k \geqslant 0$,

$$\begin{aligned}&\|\mathbf{z}_{k+1} - \mathbf{1}\bar{z}_{k+1}\|^2 \\ &\leqslant \frac{1+\rho_W}{2}\|\mathbf{z}_k - \mathbf{1}\bar{z}_k\|^2 + \frac{2\gamma^2(1+\rho_W)\rho_W}{1-\rho_W}\|\mathbf{r}_k - \mathbf{1}\bar{r}_k\|^2 \\ &\quad + \frac{2\gamma^2(1+\rho_W)\rho_W L^2}{1-\rho_W}\|\mathbf{z}_k - \mathbf{z}_{k-1}\|^2.\end{aligned} \quad (25)$$

Proof: See Appendix in [33]. ■

Next, we have the following lemma showing the contraction of the gradient tracking error.

Lemma 3 (Gradient tracking error): Suppose Assumptions 2 and 3 hold. Let the stepsize satisfy

$$\gamma \leqslant \left\{\frac{1}{4L}, \frac{1-\rho_W}{8L\sqrt{\rho_W}}\right\}. \quad (26)$$

Then, we get for all $k \geqslant 0$,

$$\begin{aligned}&\|\mathbf{r}_{k+1} - \mathbf{1}\bar{r}_{k+1}\|^2 \\ &\leqslant \frac{3+\rho_W}{4}\|\mathbf{r}_k - \mathbf{1}\bar{r}_k\|^2 + \frac{8\gamma^2 L^4\rho_W}{1-\rho_W}\|\mathbf{z}_k - \mathbf{z}_{k-1}\|^2 \\ &\quad + \frac{9L^2\rho_W}{1-\rho_W}\|\mathbf{z}_k - \mathbf{1}\bar{z}_k\|^2 + \frac{16n\gamma^2 L^2\rho_W}{1-\rho_W}\|\nabla f(\bar{z}_k)\|^2.\end{aligned} \quad (27)$$

Proof: See Appendix in [33]. ■

We are now in a position to show the contraction of the optimality gap defined in (13).

Lemma 4 (Optimality gap): Suppose Assumptions 1-3 hold. Let the stepsize satisfy

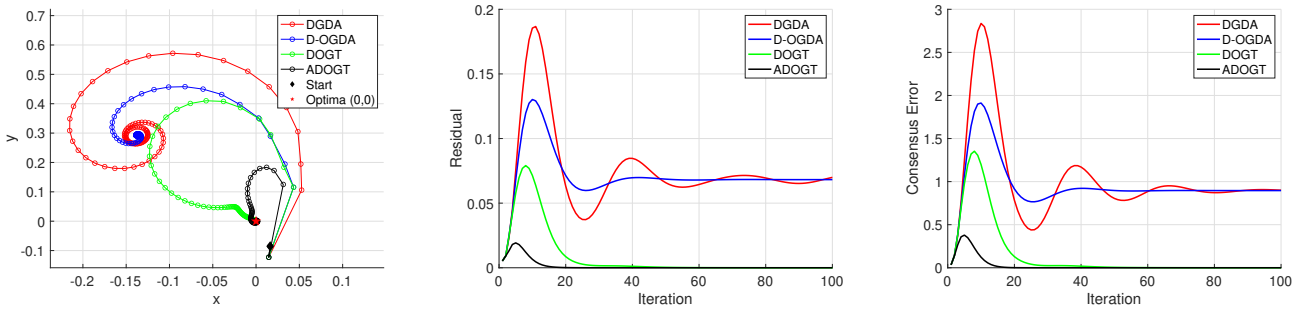
$$\gamma \leqslant \left\{\frac{1}{8L}, \frac{1-\rho_W}{8L\rho_W}\right\}. \quad (28)$$

Then, we have for all $k \geqslant 0$,

$$\begin{aligned}&\|\Xi_{k+1}\|^2 \\ &\leqslant \left(1 - \frac{3\gamma\mu}{4}\right)\|\Xi_k\|^2 + \frac{5\gamma^2 L^2}{4n}\|\mathbf{z}_k - \mathbf{z}_{k-1}\|^2 \\ &\quad + \frac{4\gamma L}{n}\|\mathbf{z}_k - \mathbf{1}\bar{z}_k\|^2 + \frac{9\gamma^3 L\rho_W}{n(1-\rho_W)}\|\mathbf{r}_k - \mathbf{1}\bar{r}_k\|^2 \\ &\quad - \frac{\gamma^2}{4n}\|\nabla F(\mathbf{1}\bar{z}_k)\|^2.\end{aligned} \quad (29)$$

Proof: See Appendix in [33].

■



(a) Trajectory (b) Residual (c) Consensus error

Fig. 1: Comparison of the convergence performance between DGDA, D-OGDA, DOGT and ADOGT.

REFERENCES

[1] M. Mohri, G. Sivek, and A. T. Suresh, "Agnostic federated learning," in *International Conference on Machine Learning*, pp. 4615–4625, PMLR, 2019.

[2] A. Sinha, H. Namkoong, R. Volpi, and J. Duchi, "Certifying some distributional robustness with principled adversarial training," *arXiv preprint arXiv:1710.10571*, 2017.

[3] K. G. Vamvoudakis, F. Fotiadis, J. P. Hespanha, R. Chinchilla, G. Yang, M. Liu, J. S. Shamma, and L. Pavel, "Game theory for autonomy: From min-max optimization to equilibrium and bounded rationality learning," in *2023 American Control Conference (ACC)*, pp. 4363–4380, IEEE, 2023.

[4] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial nets," *Advances in neural information processing systems*, vol. 27, 2014.

[5] I. Gulrajani, F. Ahmed, M. Arjovsky, V. Dumoulin, and A. C. Courville, "Improved training of wasserstein gans," *Advances in neural information processing systems*, vol. 30, 2017.

[6] J. Konečnỳ, H. B. McMahan, D. Ramage, and P. Richtárik, "Federated optimization: Distributed machine learning for on-device intelligence," *arXiv preprint arXiv:1610.02527*, 2016.

[7] A. Nedić and J. Liu, "Distributed optimization for control," *Annual Review of Control, Robotics, and Autonomous Systems*, vol. 1, pp. 77–103, 2018.

[8] 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," *Advances in neural information processing systems*, vol. 30, 2017.

[9] D. K. Molzahn, F. Dörfler, H. Sandberg, S. H. Low, S. Chakrabarti, R. Baldick, and J. Lavaei, "A survey of distributed optimization and control algorithms for electric power systems," *IEEE Transactions on Smart Grid*, vol. 8, no. 6, pp. 2941–2962, 2017.

[10] M. Liu, W. Zhang, Y. Mroueh, X. Cui, J. Ross, T. Yang, and P. Das, "A decentralized parallel algorithm for training generative adversarial nets," *Advances in Neural Information Processing Systems*, vol. 33, pp. 11056–11070, 2020.

[11] Y. Huang, X. Li, Y. Shen, N. He, and J. Xu, "Achieving near-optimal convergence for distributed minimax optimization with adaptive stepsizes," *Advances in Neural Information Processing Systems*, vol. 37, pp. 19740–19782, 2024.

[12] V. F. Dem'yanov and A. B. Pevnyi, "Numerical methods for finding saddle points," *USSR Computational Mathematics and Mathematical Physics*, vol. 12, no. 5, pp. 11–52, 1972.

[13] C. Daskalakis, A. Ilyas, V. Syrgkanis, and H. Zeng, "Training gans with optimism," *arXiv preprint arXiv:1711.00141*, 2017.

[14] G. M. Korpelevich, "The extragradient method for finding saddle points and other problems," *Matecon*, vol. 12, pp. 747–756, 1976.

[15] L. Xiao, S. Boyd, and S. Lall, "Distributed average consensus with time-varying metropolis weights," *Automatica*, vol. 1, pp. 1–4, 2006.

[16] Y. Deng and M. Mahdavi, "Local stochastic gradient descent ascent: Convergence analysis and communication efficiency," in *International Conference on Artificial Intelligence and Statistics*, pp. 1387–1395, PMLR, 2021.

[17] B. McMahan, E. Moore, D. Ramage, S. Hampson, and B. A. y Arcas, "Communication-efficient learning of deep networks from decentralized data," in *Artificial intelligence and statistics*, pp. 1273–1282, PMLR, 2017.

[18] L. Chen, H. Ye, and L. Luo, "An efficient stochastic algorithm for decentralized nonconvex-strongly-concave minimax optimization," in *International Conference on Artificial Intelligence and Statistics*, pp. 1990–1998, PMLR, 2024.

[19] A. Mokhtari, A. Ozdaglar, and S. Pattathil, "A unified analysis of extra-gradient and optimistic gradient methods for saddle point problems: Proximal point approach," in *International Conference on Artificial Intelligence and Statistics*, pp. 1497–1507, PMLR, 2020.

[20] W. Azizian, I. Mitliagkas, S. Lacoste-Julien, and G. Gidel, "A tight and unified analysis of gradient-based methods for a whole spectrum of differentiable games," in *International conference on artificial intelligence and statistics*, pp. 2863–2873, PMLR, 2020.

[21] T. Liang and J. Stokes, "Interaction matters: A note on non-asymptotic local convergence of generative adversarial networks," in *The 22nd International Conference on Artificial Intelligence and Statistics*, pp. 907–915, PMLR, 2019.

[22] A. Beznosikov, V. Samokhin, and A. Gasnikov, "Distributed saddle-point problems: Lower bounds, near-optimal and robust algorithms," *arXiv preprint arXiv:2010.13112*, 2020.

[23] W. Liu, A. Mokhtari, A. Ozdaglar, S. Pattathil, Z. Shen, and N. Zheng, "A decentralized proximal point-type method for saddle point problems," *arXiv preprint arXiv:1910.14380*, 2019.

[24] H. Cai, S. A. Alghunaim, and A. H. Sayed, "Diffusion stochastic optimization for min-max problems," *IEEE Transactions on Signal Processing*, 2024.

[25] 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.

[26] S. Pu and A. Nedić, "Distributed stochastic gradient tracking methods," *Mathematical Programming*, vol. 187, no. 1, pp. 409–457, 2021.

[27] H.-T. Wai, Z. Yang, Z. Wang, and M. Hong, "Multi-agent reinforcement learning via double averaging primal-dual optimization," *Advances in neural information processing systems*, vol. 31, 2018.

[28] S. Mukherjee and M. Chakraborty, "A decentralized algorithm for large scale min-max problems," in *2020 59th IEEE Conference on Decision and Control (CDC)*, pp. 2967–2972, IEEE, 2020.

[29] Y. Sun, G. Scutari, and A. Daneshmand, "Distributed optimization based on gradient tracking revisited: Enhancing convergence rate via surrogation," *SIAM Journal on Optimization*, vol. 32, no. 2, pp. 354–385, 2022.

[30] A. Beznosikov, G. Scutari, A. Rogozin, and A. Gasnikov, "Distributed saddle-point problems under data similarity," *Advances in Neural Information Processing Systems*, vol. 34, pp. 8172–8184, 2021.

[31] Q. Zhou, H. Ye, and L. Luo, "Near-optimal distributed minimax optimization under the second-order similarity," *Advances in Neural Information Processing Systems*, vol. 37, pp. 28009–28050, 2024.

[32] J. Liu and A. S. Morse, "Accelerated linear iterations for distributed averaging," *Annual Reviews in Control*, vol. 35, pp. 160–165, 2011.

[33] Y. Huang, J. Xu, J. Chen, and K. H. Johansson, "An optimistic gradient tracking method for distributed minimax optimization," *arXiv preprint arXiv:2508.21431*, 2025.