

Decentralized Optimization in Networks with Arbitrary Delays

Tomàs Ortega, Hamid Jafarkhani
Center for Pervasive Communications and Computing
University of California, Irvine
 Irvine, USA
 Emails: tomaso@uci.edu, hamidj@uci.edu

Abstract—We consider the problem of decentralized optimization in networks with communication delays. To accommodate delays, we need decentralized optimization algorithms that work on directed graphs. Existing approaches require nodes to know their out-degree to achieve convergence. We propose a novel gossip-based algorithm that circumvents this requirement, allowing decentralized optimization in networks with communication delays. We prove that our algorithm converges on non-convex objectives, with the same main complexity order term as centralized Stochastic Gradient Descent (SGD), and show that the graph topology and the delays only affect the higher order terms. We provide numerical simulations that illustrate our theoretical results.

Index Terms—Decentralized optimization, gossip algorithms, networks with delays, collaborative machine learning, non-convex optimization.

I. INTRODUCTION

In decentralized optimization, the nodes in a network cooperate to minimize a global objective function that is the average of nodes' local objective functions. The motivation for these problems comes from a variety of applications, including, but not limited to, decentralized estimation in sensor networks, collaborative machine learning, and decentralized coordination of multi-agent systems. We formalize these problems by considering a collection of N nodes, which can communicate with each other through the edges of a graph G – see Fig. 1 for an example. Each node has a local cost function $f_n : \mathbb{R}^d \rightarrow \mathbb{R}$.

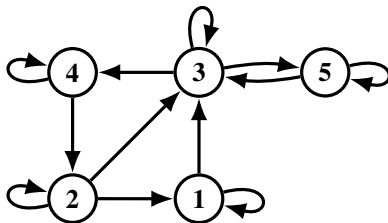


Fig. 1. Directed communication graph example, with $N = 5$.

The goal is for all nodes to find a common model $x \in \mathbb{R}^d$ that minimizes a global cost function f of the form

$$f(x) := \frac{1}{N} \sum_{n=1}^N f_n(x). \quad (1)$$

This work was supported in part by the NSF Award ECCS-2207457.

Previous research has proposed several techniques to minimize (1). The first study of this problem can be tracked back at least to [1]. For the problem of finding a common solution among nodes, the seminal works in gossip algorithms [2], [3] brought ideas from mixing in Markov chains to allow averaging over graphs. To minimize (1), it is standard to either use a combination of gradient descent with gossip steps [4], or problem-specific methods such as alternating direction method of multipliers (ADMM) [5]. In the case where G is undirected, there are multiple efforts to solve this problem in various settings, like asynchronous communications [6], time-varying graphs [7], [8], or quantized communications [9], [10]. In the case where G is directed, nodes must know their out-degree to achieve consensus, as stated in [11]. Under such assumption, a family of so-called Push-Sum methods was proposed, starting with [4], [12]. There are also multiple extensions for various settings, like quantized communications [9], and asynchronous communications [13].

There exist several nice surveys that detail the contributions in this field, see for example [11] for synchronous methods, and [14] for asynchronous methods.

The common assumption for existing algorithms for directed decentralized optimization is that nodes in the network must know their out-degree. In this paper, we propose a gossip-based optimization approach where nodes *do not* need to know their out-degree to achieve convergence. Such a property is desired in many scenarios which arise naturally. For example, consider networks where links are directed and established without a handshake, in other words, networks where the transmitter sends messages and the receivers do not need to acknowledge their reception. Such a setting is likely in scenarios where some nodes may have a transmit power that is larger than others, making some bidirectional communication impossible. Another possible application is in networks where nodes simply broadcast messages periodically, without knowledge of who receives them. Our algorithm, DT-GO (Delay Tolerant Gossiped Optimization), allows decentralized optimization in such networks.

II. SETUP AND PROPOSED ALGORITHM

A. Problem setup

As described in the introduction, we consider a set of N nodes that share their information over a network modeled as

a *directed* graph G . Since the shared information can only reach a limited number of neighboring nodes, not necessarily known to the transmitter, we assume that the nodes can *only* know their *in-degree*.

Our approach to minimize (1) is to iteratively perform two phases: (i) local optimization, and (ii) consensus. At each local optimization phase, nodes optimize their models based on their local data. This makes local models drift from the average. Therefore, we introduce a consensus phase, where nodes communicate among themselves to converge to the average of their models. The problem in the consensus phase is known in the literature as decentralized averaging.

B. Decentralized averaging

Let us assume that every Node n starts with an initial model, or initial state, $x_n(0) \in \mathbb{R}^d$. At each time t , nodes broadcast their models and collect the models of their neighbors. They then perform a weighted average with the models they receive, which is called a *gossip* step. Specifically, Node n weighs information received from Node m with weight W_{nm} . A natural choice for these weights is the inverse of the in-degree of each node. Other choices also work, as long as they satisfy our definition of a gossip matrix.

Definition 1 (Gossip matrix). Given a strongly connected directed graph G , its gossip matrix W is an $N \times N$ real matrix whose entries satisfy:

- (i) W is row-stochastic: the sum of the entries in every row is one, i.e., $\sum_{j=1}^N W_{ij} = 1$ for all i .
- (ii) Every entry W_{nm} is non-negative, and zero only if there is no directed edge from Node m to Node n in G .
- (iii) Entries in the diagonal are positive. This is equivalent to asking each node in G to have a self-loop.

Formally, each gossip iteration is defined as $x_n(t+1) = \sum_{m=1}^N W_{nm} x_m(t)$, which can be written in a matrix form as $X(t+1) = WX(t)$. Here, $X(t)$ is the $N \times d$ real matrix whose rows are the node states. Notice from item (ii) in Def. 1 that W_{nm} is only positive when Node n can actually receive model $x_n(t)$. When $t \rightarrow \infty$, we want the sequence of models to tend to the average, i.e., $x_n(t) \rightarrow \bar{x} = \frac{1}{N} \sum_{n=1}^N x_n(0)$. However, if the gossip matrix W is row-stochastic nodes *do not* necessarily converge to the average. An illustration of this phenomenon can be seen in Fig. 2, where the nodes using a regular gossip algorithm, marked by the “non-corrected node values”, clearly do not converge to $\bar{x}$. For row-stochastic matrices, it can be shown that the nodes converge to a weighted average

$$\tilde{x} = \sum_{n=1}^N \pi_n x_n(0), \quad (2)$$

where π_n are non-negative weights that add up to one [15, Lemma 5]. We prove that $\pi_1, \dots, \pi_N$ are positive in Section III. In this section, first, we present an algorithm that “corrects” the values such that the nodes converge to the average. Then, we show that this algorithm is capable of handling links with communication delays.

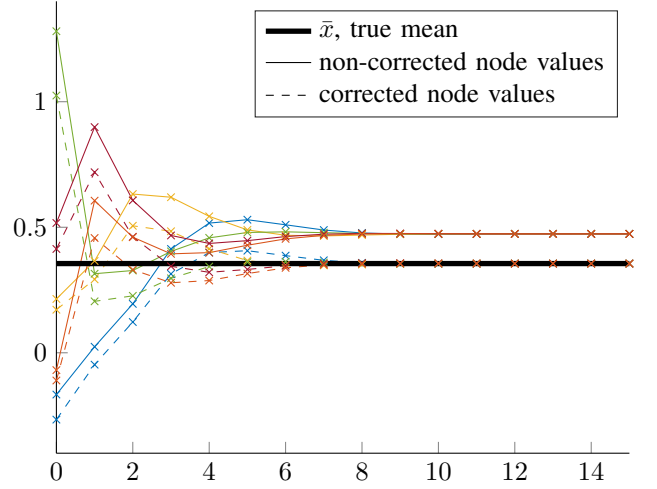


Fig. 2. Plot of corrected and non-corrected node values throughout time. Initial node values are chosen at random from a normal $\mathcal{N}(0, 5)$. The gossip weights are the inverse of the in-degrees, where G is shown in Fig. 1.

C. Algorithm design

We present DT-GO, which corrects the weighted average described in Eq. (2) and allows the consensus phase to converge to the true average. Then, we can minimize the cost function in Eq. (1) judiciously by interchanging consensus and local optimization phases.

The key idea is described in the following Lemma.

Lemma II.1. Consider a digraph G with an associated gossip matrix W as in Def. 1. If every node n multiplies its initial state $x_n(0)$ by a factor $d_n = \frac{1}{N\pi_n}$, then the gossip iterations $x_n(t+1) = \sum_{m=1}^N W_{nm} x_m(t)$ converge to the true mean, i.e.,

$$\sum_{n=1}^N \pi_n d_n x_n(0) = \sum_{n=1}^N \frac{1}{N} x_n(0) = \bar{x}. \quad (3)$$

The proof of this lemma follows immediately from Eq. (2) by inserting the re-weighted initial states. See Fig. 2 for an illustration of this lemma in practice, where the corrected node values correspond to a run where every initial state from Node n is re-weighted with d_n .

Observe that if every node has initial state $x_n(0) = e_n$, a real vector of length N that includes all zeros, except a one in coordinate n , the gossip iterations converge to

$$\tilde{x} = \sum_{n=1}^N \pi_n e_n, \quad (4)$$

that is, a vector with each π_n in the n -th coordinate. Therefore, each node can compute d_n when convergence is achieved. Leveraging this observation, we propose Alg. 1. Note that we do not use vectors e_n , because nodes do not know the network size N a priori. Instead, we use dictionaries that follow the same principle, and are implementable in practice. Once the warm-up period is over, nodes can obtain N from the dictionary size.

D. Incorporating Delays

The framework proposed for DT-GO is designed to easily accommodate links with delays, i.e., links where information takes more than one round to arrive at the receiver. To incorporate the delays, similar to [1], [15], we introduce the notion of virtual nodes, or non-computing nodes, that serve as a relay of the message for a round. If Node n sends messages to Node m with a delay of k rounds, we simply add k nodes to the graph G , each with objective function 0 to avoid modifying the objective function (1). For example, in Fig. 3, Node 4

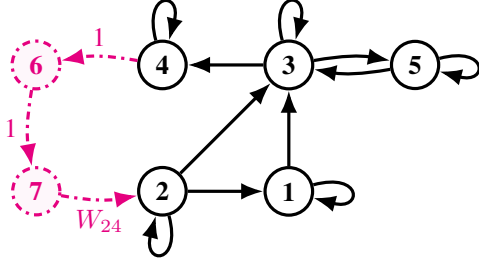


Fig. 3. Example of a graph with delayed links. We have added a delay of 2 rounds to the edge between Node 4 and Node 2 of the graph in Fig. 1.

sends messages to Node 2 with a delay of 2 rounds. Therefore, we add two nodes, 6 and 7, to the graph, and Node 4 sends messages to Node 6, which sends messages to Node 7, which finally sends messages to Node 2. The weight of the old edge from Node 4 to 2, W_{24} , is now assigned to the edge from Node 7 to 2 and the weights of the other new edges are set to 1. Note that this procedure ensures that the nodes' input

Algorithm 1 DT-GO at Node n

- 1: Generate local id number id_n .
 {WARM-UP PERIOD}
 - 2: Initialize dictionary $dict \leftarrow \{id_n : 1\}$.
 - 3: **for** $T_{\text{warm-up}}$ rounds **do**
 - 4: Broadcast $dict$ and receive neighbors' dictionaries.
 - 5: $dict \leftarrow$ weighted average of available dictionaries.
 - 6: **end for**
 - 7: From $dict$, obtain N and π_n , as in Eq. (4).
 {MINIMIZING f PERIOD}
 - 8: Initialize $x_n(0)$.
 - 9: **for** t in $0, \dots, T-1$ **do**
 - 10: Initialize auxiliary variables z_n, y .
 {Optimization phase}
 - 11: SGD step: $y \leftarrow x_n(t) - \eta \nabla F_n(x_n(t), \xi_n)$.
 {Consensus phase}
 - 12: Adjust update: $z_n \leftarrow x_n(t) + \frac{1}{N\pi_n}(y - x_n(t))$.
 - 13: **for** τ_g rounds **do**
 - 14: Broadcast z_n and receive neighbor states.
 - 15: Average states: $z_n \leftarrow \sum_{m=1}^N W_{nm} z_m$.
 - 16: **end for**
 - 17: $x_n(t+1) \leftarrow z_n$.
 - 18: **end for**
-

weights still add up to one. Also, the resulting extended graph is still strongly connected.

III. PERFORMANCE BOUNDS AND PROOF OF CONVERGENCE

In this section, first, we discuss the case without delays, as it gives a stronger and more illustrative bound. Then, we present the case with delays.

A. Case without delays

First, we introduce a proposition that is needed to show that our algorithm converges.

Proposition III.1. *Given a gossip matrix W satisfying Def. 1 we can ensure that*

- (i) *The limit $\lim_{t \rightarrow \infty} W^t := W^\infty$ exists, that is, our gossip algorithm converges to a stationary solution.*
- (ii) *The matrix W^∞ is row-stochastic, and its rows are all identical, with positive entries $\pi_1, \dots, \pi_N$ that add up to one.*
- (iii) *The squared Frobenius norm of the global state converges to the stationary solution at a geometric rate, i.e., there exist constants C and ρ , with $\rho < 1$, such that for all $t > 0$, $\|W^t - W^\infty\|_2^2 \leq C\rho^t$.*

Proof: Since W is primitive, statements (i) and (ii) are immediate consequences of applying the Perron-Frobenius theorem [16]. Furthermore, $\pi_1, \dots, \pi_N$ are the coordinates of the left Perron eigenvector of W . Statement (iii) also has a straightforward proof via eigendecomposition, as shown in [17, Fact 3]. In addition, we have $\rho = |\lambda_2|^2$, the second largest eigenvalue of W in absolute value squared. ■

B. Case with delays

Let us now assume that we have a network with arbitrary delays. The number of non-virtual nodes is N , and we add virtual nodes to obtain a network of size N_v . Similarly, the original gossip matrix W is extended as previously described to a gossip matrix W_v .

Proposition III.2. *Given a gossip matrix W satisfying Def. 1, and its extended version with delays W_v , we can ensure that*

- (i) *The limit $\lim_{t \rightarrow \infty} W_v^t := W_v^\infty$ exists, that is, our gossip algorithm with delays converges to a stationary solution.*
- (ii) *The matrix W_v^∞ is row-stochastic, and its rows are all identical, with non-negative entries $\pi_{v,1}, \dots, \pi_{v,N_v}$ that add up to one.*
- (iii) *The squared Frobenius norm of the global state converges to the stationary solution at a geometric rate, i.e., there exist constants C and ρ , with $\rho < 1$, such that for all $t > 0$, $\|W_v^t - W_v^\infty\|_2^2 \leq C\rho^t$.*
- (iv) *The weights that correspond to non-virtual nodes, i.e., $\pi_{v,1}, \dots, \pi_{v,N}$ are all positive.*

Proof: From [15, Lemma 5], Facts (i) and (ii) follow immediately. For Fact (iii), [15, Lemma 5] ensures that

$$\|W^t - W^\infty\|_F \leq 2 \frac{1 + \eta^{-B_2}}{1 - \eta^{B_2}} (1 - \eta^{B_2})^{\frac{t}{B_2}}, \quad (5)$$

where $B_2 := N - 1 + NB_1$, B_1 is the maximum number of delays between two nodes in G , and $\eta \in (0, 1)$ is a positive lower bound for all the non-zero entries of W . Squaring the expression and knowing that the Frobenius norm upper-bounds the 2-norm proves the statement. The proof of (iv) follows from [1, Lemma 5.2.1]. ■

Proposition III.2 allows us to guarantee convergence for graphs with delays, albeit it often gives us a looser bound for delay-less graphs.

C. DT-GO's convergence rate

A corollary of Propositions III.1 and III.2 is that we can choose a number of gossip rounds τ_g such that $C\rho^{\tau_g} < 1$, as needed for our convergence results.

Note that if W is doubly stochastic and we do not add delays, $C = 1$, and τ_g can be set to 1. Also, W then converges to an all-one matrix divided by N , and our algorithm becomes equivalent to Decentralized SGD (DSGD) [4]. Furthermore, if G is a complete graph and we set edge weights to $1/N$, then the gossip matrix becomes the averaging matrix, and both DT-GO and DSGD are equivalent to centralized SGD.

Assuming that the warm-up phase is complete and τ_g is large enough such that $C\rho^{\tau_g} < 1$, with a slight abuse of the notation we can write Alg. 1's iterations as

$$X(t+1) = W^{\tau_g} (X(t) - \eta D \partial F(t)), \quad (6)$$

where $\partial F(t)$ is an $N \times d$ real matrix that represents the stochastic gradients at time t , and $D = \text{diag}(\frac{1}{N\pi_1}, \dots, \frac{1}{N\pi_N})$ is the diagonal correction matrix. If there are virtual nodes, the corresponding entries in the diagonal matrix D are set to 1. This ensures that D is well defined, since $\pi_n > 0$ when Node n is non-virtual – see Proposition III.2. This matrix notation allows us to follow an analysis approach similar to [18]. Before stating our main convergence result, first, we make the following standard assumptions [4], [9].

Assumption III.3. Each function $f_n: \mathbb{R}^d \rightarrow \mathbb{R}$ is L -smooth, that is

$$\|\nabla f_n(y) - \nabla f_n(x)\| \leq L \|y - x\| \quad \forall x, y \in \mathbb{R}^d, \quad (7)$$

and we have access to unbiased stochastic gradients, $\nabla F_n(x, \xi_n)$, with bounded variance on each worker:

$$\mathbb{E} [\|\nabla F_n(x, \xi_n) - \nabla f_n(x)\|^2] \leq \sigma_n^2 \quad (8)$$

$$\mathbb{E} [\|\nabla F_n(x, \xi_n)\|^2] \leq G^2 \quad \forall x \in \mathbb{R}^d. \quad (9)$$

We denote $\bar{\sigma}^2 := \frac{1}{N} \sum_{n=1}^N \sigma_n^2$ for convenience.

We now present our main convergence result, which applies for cases with and without delays.

Theorem III.4. *Given a gossip matrix as in Def. 1 (possibly extended with delays as in Section II-D), a number of iterations T , a large enough τ_g such that $\|W^{\tau_g} - W^\infty\|_2^2 \leq C\rho^{\tau_g} := 1 - c < 1$, and under Assumption III.3, there exists a constant*

(depending on T) stepsize η such that the DT-GO iterations $\tilde{x}^{(t)} := \sum_{n=1}^N \pi_n x_n(t)$ satisfy:

$$\frac{1}{T} \sum_{t=0}^{T-1} \mathbb{E} [\|\nabla f(\tilde{x}(t))\|_2^2] = \mathcal{O} \left(\left(\frac{LF_0 \bar{\sigma}^2}{NT} \right)^{1/2} + \left(\frac{\|D\|_2 GLF_0}{cT} \right)^{2/3} + \frac{LF_0}{T} \right),$$

where $F_0 := f(\tilde{x}^{(0)}) - f^*$, f^* is a lower bound for f , and D is the diagonal correction matrix, with virtual nodes having 1 in their corresponding entries.

We present a sketch of Theorem III.4's proof, which is not included with full details due to the lack of space. The first key insight is that

$$\mathbb{E} [f(\tilde{x}(t+1))] = \mathbb{E} \left[f(\tilde{x}(t)) - \frac{\eta}{N} \sum_{n=1}^N \nabla F_n(x_n(t), \xi_n) \right], \quad (10)$$

since the weighted averages are preserved at each iteration. We can check this as follows: let us define $\tilde{X}(t+1)$ as an $N \times d$ matrix whose rows are $\tilde{x}(t)$. From Eq. (6), we obtain

$$\begin{aligned} \tilde{X}(t+1) &= W^\infty X(t+1) = W^\infty (X(t) - \eta D \partial F(t)) \\ &= \tilde{X}(t) - \eta J \partial F(t), \end{aligned}$$

where J is the averaging matrix of all ones divided by N . Notice that this is exactly Eq. (10) in matrix form. Using this insight, we expand Eq. (10) leveraging L -smoothness:

$$\begin{aligned} \mathbb{E} [f(\tilde{x}(t+1))] &\leq \mathbb{E} [f(\tilde{x}(t))] \\ &\quad - \mathbb{E} \left[\langle \nabla f(\tilde{x}(t)), \frac{\eta}{N} \sum_{n=1}^N \nabla F_n(x_n(t), \xi_n) \rangle \right] \\ &\quad + \eta^2 \frac{L}{2} \mathbb{E} \left[\left\| \frac{1}{N} \sum_{n=1}^N \nabla F_n(x_n(t), \xi_n) \right\|_2^2 \right]. \end{aligned}$$

After algebraic manipulation, and using the bounded variance assumption as well as L -smoothness, we obtain

$$\begin{aligned} \mathbb{E} [f(\tilde{x}(t+1))] &\leq \mathbb{E} [f(\tilde{x}(t))] - \frac{\eta}{4} \mathbb{E} [\|\nabla f(\tilde{x}(t))\|_2^2] \\ &\quad + \frac{3\eta L^2}{4N} \sum_{n=1}^N \mathbb{E} [\|\tilde{x}(t) - x_n(t)\|_2^2] + \eta^2 \frac{L\bar{\sigma}^2}{2N}. \end{aligned}$$

Rearranging terms and averaging over T iterations yield

$$\begin{aligned} \frac{1}{T} \sum_{t=0}^{T-1} \mathbb{E} [\|\nabla f(\tilde{x}(t))\|_2^2] &\leq 4 \frac{f(\tilde{x}(0)) - f^*}{\eta T} + \eta \frac{2L\bar{\sigma}^2}{N} \\ &\quad + \frac{3L^2}{NT} \sum_{t=0}^{T-1} \sum_{n=1}^N \mathbb{E} [\|\tilde{x}(t) - x_n(t)\|_2^2]. \end{aligned}$$

We bound the third term on the RHS using Lemma III.5. Then, choosing the appropriate stepsize, following arguments similar to [18, Lemma A.5] concludes the proof.

Notice that the first term in Theorem III.4 is the same as that of centralized SGD, and only the higher order terms are

affected by the graph topology and the delays between nodes, which affect D and c as described in Proposition III.2.

We now present Lemma III.5, used in the main proof, which guarantees that all nodes converge to the same solution when $\eta \rightarrow 0$ as $T \rightarrow \infty$.

Lemma III.5. *Given a gossip matrix as in Def. 1 (possibly extended with delays as in Section II-D), a large enough τ_g such that $\|W^{\tau_g} - W^\infty\|_2^2 \leq C\rho^{\tau_g} := 1 - c < 1$, and under Assumption III.3, with constant stepsize η , DT-GO iterations satisfy*

$$\sum_{n=1}^N \|\tilde{x}(t) - x_n(t)\|_2^2 \leq \eta^2 \frac{4N \|D\|_2^2 G^2}{c^2},$$

where D is the diagonal correction matrix, with virtual nodes having 1 in their corresponding entries.

Proof: We follow the proof from [18, Lemma A.2].

Define $r_t = \mathbb{E} \left[\|X(t) - \tilde{X}(t)\|_F^2 \right]$. Then,

$$r_{t+1} = \mathbb{E} \left[\|(W^{\tau_g} - W^\infty)(X(t) - \eta D \partial F(t) - \tilde{X}(t))\|_F^2 \right],$$

since $(W^{\tau_g} - W^\infty)\tilde{X}(t) = 0$. Using the contractive property of $W^{\tau_g} - W^\infty$, it follows immediately that

$$\begin{aligned} r_{t+1} &\leq (1-c) \mathbb{E} \left[\|X(t) - \tilde{X}(t) - \eta D \partial F(t)\|_F^2 \right] \\ &\leq (1-c)(1+\alpha) \mathbb{E} \left[\|X(t) - \tilde{X}(t)\|_F^2 \right] \\ &\quad + (1-c)(1+\alpha^{-1}) \eta^2 \mathbb{E} \left[\|D \partial F(t)\|_F^2 \right] \\ &\leq \left(1 - \frac{c}{2}\right) r_t + \frac{2}{c} \eta^2 \|D\|_2^2 N G^2. \end{aligned}$$

Finally, we can check by induction that $r_t \leq \frac{\eta^2 4 \|D\|_2^2 N G^2}{c^2}$ satisfies the recursion and the proof is complete. ■

IV. EXPERIMENTAL RESULTS

We perform a set of experiments on the following problem: Each node n has an objective function $f_n(x) = (x - n)^2$. This defines a function f that has a unique global minimum at $x^* = \frac{1+2+\dots+N}{N}$. We generate random directed graphs with $N = 100$ nodes. Every possible edge in the graph exists with a probability p and non-strongly connected generated graphs are discarded. In this section, we report DT-GO's mean performance over $I = 100$ randomly generated graphs. Note that if $p = 1$, the graph is complete and DT-GO is equivalent to centralized SGD, which serves as our benchmark. Every node averages all received messages with equal weights, i.e., the inverse of the in-degree. We use $\eta_t = \frac{0.1}{\sqrt{t}}$ and $\tau_g = 1$, and report the cost and consensus suboptimality. The cost and consensus are defined as $\frac{1}{N} \sum_{n=1}^N (n - x_n(t))^2$ and $\frac{1}{N} \sum_{n=1}^N (\bar{x}(t) - x_n(t))^2$, respectively. For both measures, the Suboptimality is defined as the difference between the value and that of the centralized SGD method, which is equivalent to the case of $p = 1$. Figs. 4 and 5 show how the graph

topology, measured by the edge probability p affects DT-GO's performance. Higher graph connectivity implies faster

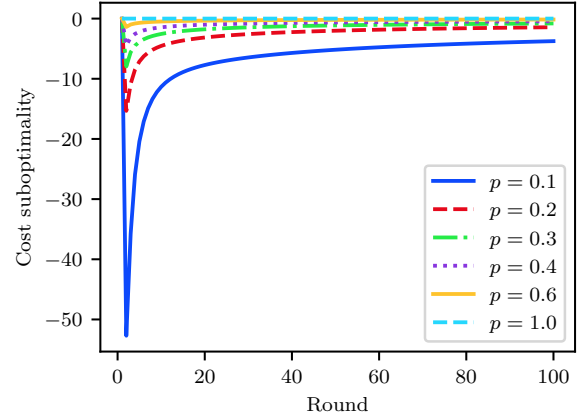


Fig. 4. DT-GO's cost comparison for strongly connected directed random graphs with edge probability p . The cost is $\frac{1}{N} \sum_{n=1}^N (n - x_n(t))^2$, and the suboptimality is the difference with respect to the baseline, $p = 1$, which is centralized SGD.

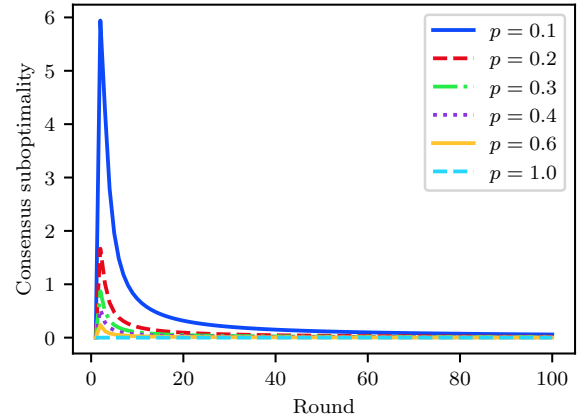


Fig. 5. DT-GO's consensus comparison for strongly connected directed random graphs with edge probability p . Consensus is defined as $\frac{1}{N} \sum_{n=1}^N (\bar{x}(t) - x_n(t))^2$, and the suboptimality is the difference with respect to the baseline, $p = 1$, which is centralized SGD.

consensus. Also, a smaller consensus value allows solutions that differ more at each node, which make local costs lower. In turn, the cost suboptimality has negative peaks before reaching consensus.

Figs. 6 and 7 show the effects of delay on DT-GO's performance. We consider a complete graph of $N = 100$ nodes and add delays at each node according to a Poisson distribution with parameter λ . We generate the delays for $I = 100$ cases for each λ value and report the mean performance. Note that $\lambda = 0$ implies no delays, and for a complete graph without delays DT-GO is the same as centralized SGD. This will serve as the baseline for our experiments. As expected, higher values of λ make DT-GO converge slower. Figs. 6 and 7 only include the first 30 iterations of the algorithm to highlight the

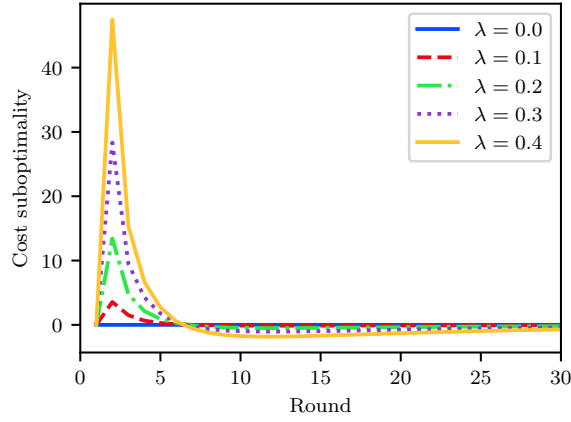


Fig. 6. DT-GO's cost comparison for complete graphs with Poisson-distributed delays among the edges. The cost is $\frac{1}{N} \sum_{n=1}^N (n - x_n(t))^2$, averaging only over non-virtual nodes. The suboptimality is the difference with respect to the baseline, $\lambda = 0$, which is centralized SGD without delays.

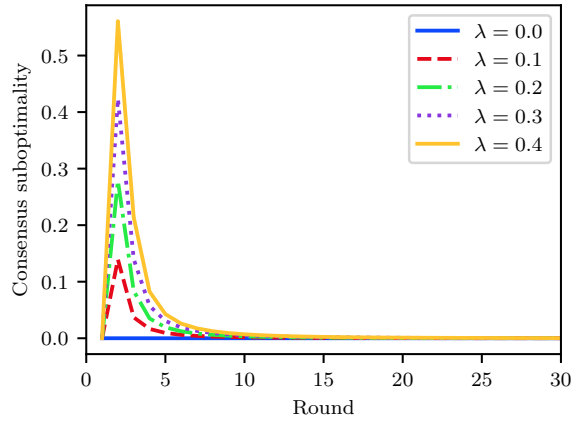


Fig. 7. DT-GO's cost comparison for complete graphs with Poisson-distributed delays among the edges. Consensus is $\frac{1}{N} \sum_{n=1}^N (\bar{x}(t) - x_n(t))^2$, averaging only over non-virtual nodes. The suboptimality is the difference with respect to the baseline, $\lambda = 0$, which is centralized SGD without delays.

delays' effects. The code for both experiments is available at [19].

V. CONCLUSIONS

We have presented DT-GO, a decentralized algorithm for solving optimization problems in networks with arbitrary delays. Nodes in the network need not have knowledge of their out-degree for DT-GO to converge. We have shown that, under standard assumptions, DT-GO is guaranteed to converge to the problem's optimal solution at the same complexity order as the plain decentralized SGD with a fully connected graph. We have also provided simulation results that back up our theoretical findings.

REFERENCES

[1] J. N. Tsitsiklis, "Problems in decentralized decision making and computation," Ph.D. dissertation, Massachusetts Institute of Technology, 1984.

[2] D. Kempe, A. Dobra, and J. Gehrke, "Gossip-based computation of aggregate information," in *44th Annual IEEE Symposium on Foundations of Computer Science, 2003. Proceedings.* Cambridge, MA, USA: IEEE Computer. Soc, 2003, pp. 482–491. [Online]. Available: <http://ieeexplore.ieee.org/document/1238221/>

[3] S. Boyd, A. Ghosh, B. Prabhakar, and D. Shah, "Randomized gossip algorithms," *IEEE Transactions on Information Theory*, vol. 52, no. 6, pp. 2508–2530, Jun. 2006. [Online]. Available: <http://ieeexplore.ieee.org/document/1638541/>

[4] A. Nedic and A. Ozdaglar, "Distributed subgradient methods for multi-agent optimization," *IEEE Transactions on Automatic Control*, vol. 54, no. 1, pp. 48–61, Jan 2009.

[5] E. Wei and A. Ozdaglar, "Distributed alternating direction method of multipliers," in *2012 IEEE 51st IEEE Conference on Decision and Control (CDC)*. Maui, HI, USA: IEEE, Dec. 2012, pp. 5445–5450. [Online]. Available: <http://ieeexplore.ieee.org/document/6425904/>

[6] K. Srivastava and A. Nedic, "Distributed asynchronous constrained stochastic optimization," *IEEE Journal of Selected Topics in Signal Processing*, vol. 5, no. 4, pp. 772–790, Aug. 2011. [Online]. Available: <https://ieeexplore.ieee.org/abstract/document/5719290>

[7] A. Nedić, A. Olshevsky, and W. Shi, "Achieving geometric convergence for distributed optimization over time-varying graphs," *SIAM Journal on Optimization*, vol. 27, no. 4, pp. 2597–2633, Jan. 2017. [Online]. Available: <https://epubs.siam.org/doi/10.1137/16M1084316>

[8] A. Koloskova, N. Loizou, S. Boreiri, M. Jaggi, and S. Stich, "A unified theory of decentralized sgd with changing topology and local updates," in *Proceedings of the 37th International Conference on Machine Learning*. PMLR, Nov 2020, pp. 5381–5393. [Online]. Available: <https://proceedings.mlr.press/v119/koloskova20a.html>

[9] A. Koloskova, S. Stich, and M. Jaggi, "Decentralized stochastic optimization and gossip algorithms with compressed communication," in *Proceedings of the 36th International Conference on Machine Learning*. PMLR, May 2019, pp. 3478–3487. [Online]. Available: <https://proceedings.mlr.press/v97/koloskova19a.html>

[10] T. Ortega and H. Jafarkhani, "Gossiped and quantized online multi-kernel learning," *IEEE Signal Processing Letters*, vol. 30, pp. 468–472, 2023.

[11] A. Nedić, A. Olshevsky, and M. G. Rabbat, "Network topology and communication-computation tradeoffs in decentralized optimization," *Proceedings of the IEEE*, vol. 106, no. 5, pp. 953–976, May 2018. [Online]. Available: <https://ieeexplore.ieee.org/document/8340193>

[12] F. Benezit, V. Blondel, P. Thiran, J. Tsitsiklis, and M. Vetterli, "Weighted gossip: Distributed averaging using non-doubly stochastic matrices," in *2010 IEEE International Symposium on Information Theory*. Austin, TX, USA: IEEE, Jun. 2010, pp. 1753–1757. [Online]. Available: <http://ieeexplore.ieee.org/document/5513273/>

[13] M. Assran, N. Loizou, N. Ballas, and M. Rabbat, "Stochastic gradient push for distributed deep learning," in *Proceedings of the 36th International Conference on Machine Learning*, ser. Proceedings of Machine Learning Research, K. Chaudhuri and R. Salakhutdinov, Eds., vol. 97. PMLR, 09–15 Jun 2019, pp. 344–353. [Online]. Available: <https://proceedings.mlr.press/v97/assran19a.html>

[14] B. M. Assran, A. Aytekin, H. R. Feyzmahdavian, M. Johansson, and M. G. Rabbat, "Advances in asynchronous parallel and distributed optimization," *Proceedings of the IEEE*, vol. 108, no. 11, pp. 2013–2031, Nov. 2020. [Online]. Available: <https://ieeexplore.ieee.org/document/9217472/>

[15] A. Nedić and A. Ozdaglar, "Convergence rate for consensus with delays," *Journal of Global Optimization*, vol. 47, no. 3, pp. 437–456, Jul 2010. [Online]. Available: <https://doi.org/10.1007/s10898-008-9370-2>

[16] C. D. Meyer and I. Stewart, *Matrix Analysis and Applied Linear Algebra, Second Edition*, ser. Other Titles in Applied Mathematics. Society for Industrial and Applied Mathematics, Jan. 2023. [Online]. Available: <https://epubs.siam.org/doi/book/10.1137/1.9781611977448>

[17] J. S. Rosenthal, "Convergence rates for markov chains," *SIAM Review*, vol. 37, no. 3, pp. 387–405, 1995. [Online]. Available: <http://www.jstor.org/stable/2132659>

[18] A. Koloskova, T. Lin, S. U. Stich, and M. Jaggi, "Decentralized deep learning with arbitrary communication compression," in *International Conference on Learning Representations*, 2019.

[19] T. Ortega. (2023, Oct.) DT-GO. GitHub. [Online]. Available: <https://github.com/TomasOrtega/DT-GO>