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

Distributed optimization on directed graphs based on inexact ADMM with partial participation

Dingran Yi and Nikolaos M. Freris

***Abstract*— We consider the problem of minimizing the sum of cost functions pertaining to agents over a network whose topology is captured by a directed graph (i.e., asymmetric communication). We cast the problem into the ADMM setting, via a consensus constraint, for which both primal subproblems are solved inexactly. In specific, the computationally demanding local minimization step is replaced by a single gradient step, while the averaging step is approximated in a distributed fashion. Furthermore, partial participation is allowed in the implementation of the algorithm. Under standard assumptions on strong convexity and Lipschitz continuous gradients, we establish linear convergence and characterize the rate in terms of the connectivity of the graph and the conditioning of the problem. Our line of analysis provides a sharper convergence rate compared to Push-DIGing. Numerical experiments corroborate the merits of the proposed solution in terms of superior rate as well as computation and communication savings over baselines.**

I. INTRODUCTION

Distributed multi-agent optimization has found paramount applications across various fields such as in control [1], signal processing [2], [3], machine learning and data mining [4], [5], and wireless sensor networks [6], [7]. The archetypal problem is to

$$\underset{\hat{x}\in\mathbb{R}^d}{\text{minimize}} \qquad f(\hat{x}) = \sum_{i=1}^{n} f_i(\hat{x}), \tag{1}$$

where $f_i(\cdot)$ is a local cost function corresponding to agent i. Distributed optimization amounts to solving (1) over the common decision vector $\hat{x}$ by a synergy of local computations and communication exchanges. In specific, agent i holds a local variable x_i which is updated based on its local cost $f_i(\cdot)$ alongside information obtained by its neighbors (e.g., their local variables or gradients), and *consensus* (i.e., $x_1 = \cdots = x_n$) is achieved asymptotically.

There has been extensive work on the subject, especially for the case that the communication network topology is symmetric (i.e., an undirected graph) [8], [9], [10], [11]. Nevertheless, it is quite common to have unidirectional communication links in wireless networks due to heterogeneity in transceivers or perceived levels of interference [7]. Algorithms on directed graphs can be roughly classified into: a) primal methods and b) primal-dual methods, most notably based on the *Alternating Direction Method of Multipliers* (ADMM) [12]. In the former case, a local gradient step is used along with weighted averaging across neighboring agents. A sublinear convergence is established [13], [14] even for strongly convex problems, while linear convergence can be retrieved using gradient tracking [15].

Recent work has developed ADMM-based methods for distributed optimization on directed graphs [16], [17], [18]. In specific, [16] uses dynamically updated weights for local averaging and establishes linear convergence under strong convexity. The authors in [17] adopted $\epsilon-$inexact consensus and proposed an asynchronous method that requires a finite number of communication steps per round. [18] allows for both equality and inequality constraints and establishes either sublinear rate to the exact solution or linear rate to a neighborhood of optimality.

Notwithstanding, the prior art based on ADMM requires an exact solution of the local subproblems, which may incur heavy computational burden unsuitable for resource-constrained devices. Besides, it is not amenable to partial agent participation, an imperative requirement in real scenarios where user unavailability is common (due to variable operating conditions such as battery level and network bandwidth) and synchronization is difficult [19] . In order to address these challenges, we propose *IPD (Inexact, Partial participation, Directed graph)*.

Contributions:

1) We propose a primal-dual method for distributed optimization on directed graphs that alleviates the computational load by inexact solution of the local subproblems (using a single gradient step).
2) Under standard assumptions, we establish linear convergence and characterize the rate with respect to the graph connectivity and the conditioning of the problem (Thm. 1 and Cor. 1).
3) The method allows for partial user participation at each iteration of the algorithm. Thm. 2 establishes linear convergence with high probability and reveals its dependency on the activation probability.
4) Our analysis provides a sharper characterization of the rate compared to the state-of-the-art Push-DIGing method with which it shares comparable communication and computation costs.
5) Experiments on two real-life machine learning datasets demonstrate merits in terms of a) faster rate compared to Push-DIGing, b) computation and communication savings over baselines.

School of Computer Science, University of Science and Technology of China, Hefei, Anhui, 230027, China. Emails: ydr0826@mail.ustc.edu.cn, nfr@ustc.edu.cn. This research was supported by the USTC Research Department under grant WK2150110025. Correspondence to N. Freris

II. Notation

The network topology is captured by a directed graph $\mathcal{G} = \{\mathcal{V}, \mathcal{E}\}$, where $\mathcal{V}$ is the set of agents (with cardinality $n := |\mathcal{V}|$) and $\mathcal{E}$ is the set of directed communication links: $(i,j) \in \mathcal{E}$ if and only if agent i can send a message to agent j. We define the set of agent i's in-neighbors as $\mathcal{N}_i^{\text{in}} := \{j : (j,i) \in \mathcal{E}\}$, and its out-degree by $d_i := |\{j : (i,j) \in \mathcal{E}\}|$. The maximum out-degree is denoted by $d_{\max} := \max_{i \in \mathcal{V}} d_i$, while $D := \text{diag}(d_1, d_2, \ldots, d_n)$ is a diagonal matrix with entries the out-degrees of the corresponding agents. The adjacency matrix $A \in \mathbb{R}^{n \times n}$ satisfies $A_{ij} = 1$ if $(j,i) \in \mathcal{E}$ and $A_{ij} = 0$ otherwise. We further define $P := \left(I + AD^{-1}\right)/2$ and for $\lambda_2(P)$ its second largest eigenvalue, while we use ϕ for the diameter of the graph. All vectors are meant as column vectors. The notation $x_i \in \mathbb{R}^d$ is for the local vector of agent i, while $x \in \mathbb{R}^{nd}$ is reserved for the concatenation, i.e., $x^T := [x_1^T, \ldots, x_n^T]^T$ (and analogously for other variables). We let $\bar{x}^k := \frac{1}{n} \sum_{i=1}^{n} x_i^k \otimes \mathbf{1}_n$, where $\mathbf{1}$ means the all-one vector, and $x_\perp^k := x^k - \bar{x}^k$ where superscript k corresponds to the $k-$th iterates; analogous definitions apply for $\bar{z}^k$ and $z_\perp^k$. Additionally, we define $F(x) := \sum_{i=1}^{n} f_i(x_i)$ and the consensus set $\mathcal{C} := \{x : x_1 = \cdots = x_n\}$ with corresponding indicator function:

$$I_{\mathcal{C}}(x) = \begin{cases} 0, & x \in \mathcal{C} \\ \infty, & \text{else} \end{cases}$$

Finally, we let $[n] := \{1, \ldots, n\}$ for $n \in \mathbb{N}$.

III. Proposed Method

To cast problem (1) into the setting of ADMM, we re-write is as:

$$\begin{aligned} \underset{x,z \in \mathbb{R}^{nd}}{\text{minimize}} \quad & F(x) + I_{\mathcal{C}}(z) \\ \text{subject to} \quad & x = z \end{aligned} \tag{2}$$

The augmented Lagrangian (AL) for (2) is given by:

$$L_\rho(x,z,y) = F(x) + I_{\mathcal{C}}(z) + y^T(x-z) + \frac{\rho}{2}\|x-z\|^2,$$

where $y \in \mathbb{R}^{nd}$ is the dual variable and $\rho > 0$. The iterations of ADMM are given by sequential alternating minimization of the AL over x, z plus a dual ascent step, as follows:

$$x^{k+1} = \underset{x}{\text{argmin}} \; L_\rho\left(x, z^k, y^k\right), \tag{3a}$$

$$z^{k+1} = \underset{z}{\text{argmin}} \; L_\rho\left(x^{k+1}, z, y^k\right), \tag{3b}$$

$$y^{k+1} = y^k + \rho\left(x^{k+1} - z^{k+1}\right). \tag{3c}$$

Step (3a) decomposes to local optimization subproblems at the agents. Solving these exactly is generally computationally burdensome, therefore *inexact* minimization is invoked in the form of a single gradient descent step (with $\eta > 0$) as:

$$x_i^{k+1} = x_i^k - \eta\left(\nabla f_i\left(x_i^k\right) + y_i^k + \rho\left(x_i^k - z_i^k\right)\right). \tag{4}$$

Algorithm 1 IPD

Initialization: $x_i^0 = z_i^0 = y_i^0 = 0, w_i^0(0) \in \left(0, d_{\max}^{-(2\phi+1)}\right]$

1: **for** $k = 0, 1, \ldots$ **do**
2: **for** each active agent i **do**
3: Compute x_i^{k+1} using (4)
4: Initialize $\xi_i^{k+1}(0) = x_i^{k+1}$
5: **for** $b = 0, 1, \ldots, B-1$ **do**
6: Broadcast $w_i^k(b)$ and $\xi_i^{k+1}(b)$
7: Compute $w_i^k(b+1)$ using (5)
8: Compute $\xi_i^{k+1}(b+1)$ using (6)
9: **end for**
10: Set $w_i^{k+1}(0) = w_i^k(B)$ and $z_i^{k+1} = \xi_i^{k+1}(B)$
11: Compute $y_i^{k+1} = y_i^k + \rho\left(x_i^{k+1} - z_i^{k+1}\right)$
12: **end for**
13: **end for**

Problem (3b) is a quadratic program with closed form solution given by:

$$z_i^{k+1} = \frac{1}{n} \sum_{j=1}^{n} \left(x_j^{k+1} + \frac{y_j^k}{\rho}\right).$$

The challenge here is that this is a global averaging step that can not be computed in a distributed manner without (prohibitively) extensive message-passing. For this reason, we also opt to solve (3b) *inexactly* by B distributed averaging steps with weights obtained by the weight balancing method proposed in [16]. The latter requires each agent to initialize its local weight $w_i^0(0) \in \left(0, d_{\max}^{-(2\phi+1)}\right]$ and update it (for $k \geq 0, b \in \{0, \ldots, B-1\}$) using:

$$w_i^k(b+1) = \frac{1}{2}\left(w_i^k(b) + \frac{1}{d_i} \sum_{j \in \mathcal{N}_i^{\text{in}}} w_j^k(b)\right), \tag{5}$$

setting $w_i^{k+1}(0) = w_i^k(B)$.

We use $\xi_i^k(\cdot)$ as the proxy for z_i^k, which is initialized as $\xi_i^{k+1}(0) = x_i^{k+1}$ and updated ($b \in \{0, \ldots, B-1\}$) using:

$$\xi_i^{k+1}(b+1) = \left(1 - d_i w_i^k(b)\right)\xi_i^{k+1}(b) + \sum_{j \in \mathcal{N}_i^{\text{in}}} w_j^k(b)\xi_j^{k+1}(b), \tag{6}$$

whence we let $z_i^{k+1} := \xi_i^{k+1}(B)$.

These steps can be carried in a distributed fashion using information obtained from in-neighbors which, in turn, suggests that *broadcasting* to out-neighbors suffices for communication. Our method is termed IPD (IPD: inexact, partial participation, directed graph) and is presented as Alg. 1. It supports partial participation (step 2); this is analyzed as random activation with probabilities q_i (Thm. 2). Besides, local computation amounts to an economical single gradient step (step 3). Communication is carried by broadcasting ξ_i and weight w_i (step 6; total cost is $d+1$) which are updated by distributed averaging (steps 7-8), while step 11 is for dual ascent. In view of partial participation, an implicit assumption is that the latest information received by broadcasting is stored in a buffer

so that the update of an active agent will not be affected by the inactivity of its neighbors and an active agent is available to carry all operations in steps 5-9. We emphasize that since only the latest information is stored, the storage requirements do not increase.

IV. Convergence analysis

The analysis is carried under the following two assumptions:

Assumption 1. *The directed graph $\mathcal{G}$ is strongly connected.*

Assumption 2. *The objective functions f_i satisfy:*

1) *Each $f_i, i \in [n]$ is strongly-convex, i.e., there exists $m_f > 0$, such that $\forall x, y \in \mathbb{R}^n$, $(\nabla f_i(x) - \nabla f_i(y))^T (x-y) \geq m_f \|x-y\|^2$.*
2) *The gradient of each $f_i, i \in [n]$ is Lipschitz continuous with constant $M_f > 0$, i.e., $\forall x, y \in \mathbb{R}^n$, $\|\nabla f_i(x) - \nabla f_i(y)\| \leq M_f \|x-y\|$.*

We denote the primal-dual optimal solution of (2) by $(x^\star, z^\star, y^\star)$; strong convexity guarantees unique primal solutions, while (7c) of the following KKT conditions guarantees uniqueness of the dual optimal solution:

$$x_1^\star = \cdots = x_n^\star, \tag{7a}$$
$$x^\star = z^\star, \tag{7b}$$
$$\nabla F(x^\star) + y^\star = 0. \tag{7c}$$

The following lemma establishes an error bound on the deviation of z^k from the average $\bar{z}^k$ that is key in establishing our convergence theorem. The analysis is similar with some steps in establishing [16, Theorem 1], but the obtained upper bound is more general.

Lemma 1. *Let $x_\perp^k := x^k - \bar{x}^k, z_\perp^k := z^k - \bar{z}^k$. Under Assumption 1, there exist $\bar{k} > 0, \delta \in (0,1)$, such that the sequence generated by Alg. 1 satisfies*

$$\left\|z_\perp^k\right\|^2 \leq \delta^{2B-1} \left\|x_\perp^k\right\|^2 + \Delta^k \left(\left\|x^k - x^\star\right\|^2 + \|x^\star\|^2\right),$$

$\forall k \geq \bar{k}$, and $\Delta^k \to 0$ geometrically with rate $\lambda_2(P)$.

Proof: See Appendix.

We proceed to analyze the convergence of Alg. 1 in two steps: a) Thm. 1 establishes linear convergence with full participation; b) Thm. 2 considers the case of partial participation and establishes the convergence with high probability.

Theorem 1 (Full Participation)**.** *Under Assumptions 1 and 2, by choosing $\eta \leq \frac{4}{15(M_f+m_f)}, \rho \in \left(0, \frac{4}{87}\frac{M_f m_f}{M_f+m_f}\right)$, $B \geq \max\left\{1, \frac{1}{2}\left((\ln \frac{5}{36}/\ln\delta)+1\right), \frac{1}{2}\left(\ln \frac{8M_f m_f}{9(M_f+m_f)^2}/\ln\delta + 1\right)\right\}$, if all agents are active in each iteration, the sequence generated by Alg. 1 satisfies*

$$\left\|x^k - x^\star\right\|^2 + \left\|y^k - y^\star\right\|^2 = \mathcal{O}\left(\lambda^k\right),$$

where

$$\lambda = \max\left\{\frac{2\mu_1}{1+\mu_1}, \lambda_2(P)\right\},$$

$$\mu_1 = \max\left\{\frac{c_1+c_2}{2c_2}, 1 - \frac{2}{3}\rho c_3\right\},$$
$$c_1 := \frac{1}{2\eta} + 2\rho + 3\rho\delta^{2B-1} - \frac{m_f M_f}{m_f+M_f},$$
$$c_2 := \frac{1}{2\eta} - \delta^{2B-1}\left[\frac{5\rho}{4} + \frac{1}{\eta} + \frac{3(m_f+M_f)}{4}\right] - \rho,$$
$$c_3 := \min\left\{\frac{1}{3(m_f+M_f)}, \eta^2\left(\frac{3(m_f+M_f)}{16} - \rho\right), \frac{c_2-c_1}{\rho^2(4+4\delta^{2B-1})}\right\}.$$

Proof: See Appendix.

The following corollary shows that for appropriate parameter choice the convergence rate depends on a) the condition number κ (with a dependency of $\frac{1}{\kappa}$ that is reminiscent of first order methods without acceleration) and b) the connectivity of the graph ($\lambda_2(P)$). The lower bound on B increases logarithmically with the condition number and decreases logarithmically with the parameter δ (pertaining to the graph topology).

Corollary 1. *Let $\eta = \frac{4}{15(M_f+m_f)}$, $\rho = \frac{2}{87}\frac{M_f m_f}{M_f+m_f}$, and $B > \frac{1}{2} + \frac{\ln(1000(\kappa+1))}{2\ln 1/\delta}$, and define $\kappa := \frac{M_f}{m_f}$. Then Alg. 1 with full participation converges linearly with rate*

$$\lambda = \max\left\{1 - \mathcal{O}\left(\frac{1}{\kappa}\right), \lambda_2(P)\right\}.$$

In the following theorem, parameters η (stepsize), ρ (penalty coefficient), B (inner-loop rounds) are as in Thm. 1. Convergence with partial participation is established under a simple stochastic model that assumes agents are activated with probability $q_i > 0$ at each iteration (step 2 of Alg. 1).

Theorem 2 (Partial Participation)**.** *Let Assumptions 1 and 2 and parameters η, ρ, B as in Thm. 1. If at every round agent i is active with probability q_i (i.i.d. across rounds) with $q_{\min} := \min_{i\in[n]} q_i > 0$, then for any $\epsilon \in (0,1)$, with probability at least $1-\epsilon$, the following holds:*

$$\left\|x^k - x^\star\right\|^2 + \left\|y^k - y^\star\right\|^2 = \mathcal{O}\left(\lambda_p^k\right),$$

where λ_p can be arbitrarily chosen from $(\lambda_1, 1)$, $\lambda_1 = \max\left\{\lambda_2(P), 1 - q_{\min}\frac{1-\mu_1}{1+\mu_1}\right\}$.

Proof: See Appendix.

Remark 1. *Convergence in Thm. 2 is established with high probability, where the dependency on ϵ is hidden in $\mathcal{O}(\cdot)$, since the established rate analysis is asymptotic. In specific, for any $\epsilon \in (0,1)$, with probability at least $1-\epsilon$, there exists some $K = K(\epsilon, \lambda_p)$, such that $\forall k > K$, $\left\|x^k - x^\star\right\|^2 + \left\|y^k - y^\star\right\|^2 \leq \lambda_p^k$.*

V. Comparison with push-diging

Push-DIGing [15] is the most popular gradient-based method for directed graphs that achieves linear convergence through gradient tracking (second line in (11)). The updates

for agent i at iteration k are as follows:

$$u_i(k+1) = c_{ii}(u_i(k) - \eta y_i(k)) + \sum_{j \in \mathcal{N}_i^{\text{in}}} c_{ij}(k)(u_j(k) - \eta y_j(k)), \quad (8)$$

$$v_i(k+1) = c_{ii}(k)v_i(k) + \sum_{j \in \mathcal{N}_i^{\text{in}}} c_{ij}(k)v_j(k), \quad (9)$$

$$x_i(k+1) = u_i(k+1)/v_i(k+1), \quad (10)$$

$$y_i(k+1) = c_{ii}(k)y_i(k) + \sum_{j \in \mathcal{N}_i^{\text{in}}} c_{ij}(k)y_j(k) + \{\nabla f_i(x_i(k+1)) - \nabla f_i(x_i(k))\}. \quad (11)$$

where η is the stepsize, $C(k)$ is a column-stochastic matrix, and the initialization uses $v_i(0) = 1$, $x_i(0) = u_i(0)$, and $y_i(0) = \nabla f_i(x_i(0))$.

We first show that our proposed solution has comparable computation/communication cost (in fact lower communication cost for the case $B = 1$):

- for Push-DIGing, in each round the cost for one agent includes one broadcast of $2d + 1$, one gradient evaluation, plus a local averaging cost of $d_i^{\text{in}}(2d + 1)$ where d_i^{in} is the in-degree of agent i.
- for our method, the cost includes one broadcast of $B(d + 1)$, one gradient evaluation, and averaging cost of $d_i^{\text{in}}(B(d + 1))$.

We conclude that for $B = 1$, IPD has lower computation and communication costs.

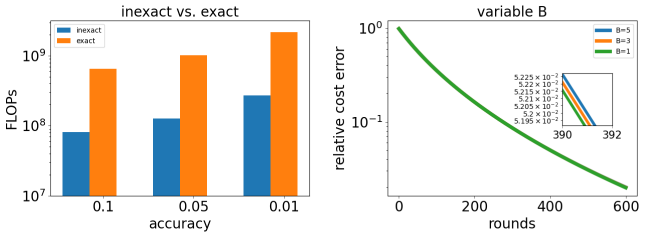
We proceed to compare our rate with the one established in [15, Theorem 18]. For Push-DIGing, by denoting $V(k) := \text{diag}(v_1(k), \ldots, v_n(k))$, it holds that $\sup_k \left\|V(k)^{-1}\right\|_{\max} = \mathcal{O}(n^n)$ [15, Equation 49], which results in an upper bound selection of stepsize as $\mathcal{O}(n^{-n})$; this, in turn, leads to a rate of $1 - \mathcal{O}(n^{-n})$. This is reminiscent of the analysis technique which was developed to address a more general problem that also considers time-varying graphs. In contrast, the stepsize in our case depends only on the scaling and not on the population (see Thm. 1). We further study the rate experimentally in Fig. 2, which depicts a substantial acceleration for the same stepsize. This is also translated to substantial computation and communication savings (since $B = 1$ is used in all comparisons) . The superior rate achieved by our proposed method can further be explained by the weight balancing process in (5) which converges to a doubly stochastic matrix. In contrast, (11) in Push-DIGing uses a column-stochastic matrix (i.e., it does not apply the push-sum protocol in the gradient estimation step).

VI. Experiments

We evaluate IPD on a distributed logistic regression problem:

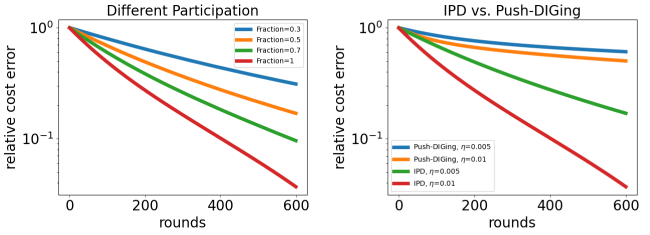
$$f_i(x) := \frac{1}{m_i} \sum_{j=1}^{m_i} \left[\ln\left(1 + e^{w_j^T x}\right) + (1 - y_j) w_j^T x\right],$$

where m_i is the number of data points held by each agent and $\{w_j, y_j\}_{j=1}^{m_i} \subset \mathbb{R}^d \times \{0, 1\}$ are labeled samples. We used two datasets from LIBSVM[1] and the UCI Machine Learning Repository[2] for Fig. 1 and Fig. 2 respectively. In each case we take 5,000 data points with dimension $d = 22$, distribute them uniformly at random across $n = 50$ agents. The communication topology is captured by a directed graph obtained by randomly adding edges to the ring graph with probability 0.2. We use the relative cost error as the metric for convergence which is defined as $\frac{\sum_{i=1}^n f(x_i^k) - f(x_i^\star)}{\sum_{i=1}^n f(x_i^0) - f(x_i^\star)}$. Fig. 1.a shows the comparison with the alternative of solving the local optimization problem (3a) exactly as in [16] : it reveals a large computational saving of 87.5% in all cases. Fig. 1.b illustrates a negligible dependency on the number of communication steps B. In fact, all other experiments for both datasets were conducted for $B = 1$. Fig. 2.a shows the effect of increasing participation in the speedup of the algorithm (as expected from Thm. 2). Last but not least, we compared against Push-DIGing in Fig. 2.b for two different stepsizes (common for both methods). The superior convergence rate of IPD translates to 90.4% computation and 94.9% communication savings for target accuracy of 0.5. In particular, to reach the target accuracy with stepsize $\eta = 0.01$, our method needs about 100 rounds while Push-DIGing needs 600 rounds and the cost per round is based on the analysis in the previous section.



(a) (b)

Fig. 1: Computation cost to reach a target accuracy compared with [16] (a) and convergence paths for variable B (b) (Full participation is considered in both cases).



(a) (b)

Fig. 2: Comparison on different levels of participation for $B = 5$ (a) and comparison (full participation) with Push-DIGing (b).

VII. Conclusion

This paper proposed proposed IPD, a primal-dual method for distributed optimization over directed graphs. The

[1]Available at https://www.csie.ntu.edu.tw/~cjlin/libsvm/.
[2]Available at https://archive.ics.uci.edu/ml/index.php.

two primal subproblems are solved inexactly: one step of gradient descent for the local optimization problem ($x-$variable) and distributed averaging based on weight balancing ($z-$variable). IPD was shown both theoretically and experimentally to have faster convergence than Push-DIGing (Sec. V) as well as substantial computation and communication savings over baseline methods. The established linear rate gives a decomposition in terms of the problem conditioning and the network connectivity (Thm. 1 and Cor. 1), similar with first order methods on undirected graphs. Furthermore, the lower bound for B (number of communications per round) was shown to have a logarithmic dependency with conditioning and connectivity (Cor. 1). A distinctive attribute of IPD is the feasibility of partial agent participation, which is crucial in large-scale real systems (the rate was established in Thm. 2).

REFERENCES

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

[2] A. Dimakis, S. Kar, J. Moura, M. Rabbat, and A. Scaglione, "Gossip algorithms for distributed signal processing," *Proceedings of the IEEE*, vol. 98, no. 11, pp. 1847–1864, 2010.

[3] P. Sopasakis, N. M. Freris, and P. Patrinos, "Accelerated reconstruction of a compressively sampled data stream," in *IEEE European Signal Processing Conference (EUSIPCO)*, 2016, pp. 1078–1082.

[4] J. Verbraeken, M. Wolting, J. Katzy, J. Kloppenburg, T. Verbelen, and J. S. Rellermeyer, "A survey on distributed machine learning," *ACM computing surveys (csur)*, vol. 53, no. 2, pp. 1–33, 2020.

[5] M. Vlachos, N. M. Freris, and A. Kyrillidis, "Compressive mining: fast and optimal data mining in the compressed domain," *The VLDB Journal*, vol. 24, no. 1, pp. 1–24, 2015.

[6] N. M. Freris, H. Kowshik, and P. R. Kumar, "Fundamentals of large sensor networks: Connectivity, capacity, clocks, and computation," *Proceedings of the IEEE*, pp. 1828–1846, 2010.

[7] C. Intanagonwiwat, R. Govindan, D. Estrin, J. Heidemann, and F. Silva, "Directed diffusion for wireless sensor networking," *IEEE/ACM Transactions on Networking*, vol. 11, no. 1, pp. 2–16, 2003.

[8] A. Nedic and A. Ozdaglar, "Distributed subgradient methods for multi-agent optimization," *IEEE Transactions on Automatic Control*, vol. 54, pp. 48–61, 2009.

[9] A. Makhdoumi and A. Ozdaglar, "Convergence rate of distributed ADMM over networks," *IEEE Transactions on Automatic Control*, vol. 62, no. 10, pp. 5082–5095, 2017.

[10] T. H. Chang, M. Hong, and X. Wang, "Multi-agent distributed optimization via inexact consensus ADMM," *IEEE Transactions on Signal Processing*, vol. 63, no. 2, pp. 482–497, 2014.

[11] N. S. Aybat, Z. Wang, T. Lin, and S. Ma, "Distributed linearized alternating direction method of multipliers for composite convex consensus optimization," *IEEE Transactions on Automatic Control*, vol. 63, no. 1, pp. 5–20, 2018.

[12] S. Boyd, N. Parikh, E. Chu, B. Peleato, and J. Eckstein, "Distributed optimization and statistical learning via the alternating direction method of multipliers," *Foundations and Trends® in Machine Learning*, vol. 3, no. 1, pp. 1–122, 2011.

[13] A. Nedić and A. Olshevsky, "Distributed optimization over time-varying directed graphs," *IEEE Transactions on Automatic Control*, vol. 60, no. 3, pp. 601–615, 2014.

[14] A. Nedić and A. Olshevsky, "Stochastic gradient-push for strongly convex functions on time-varying directed graphs," *IEEE Transactions on Automatic Control*, vol. 61, no. 12, pp. 3936–3947, 2016.

[15] 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, 2017.

[16] K. Rokade and R. K. Kalaimani, "Distributed ADMM over directed networks," *arXiv e-prints, arXiv:2010.10421*, 2020.

[17] W. Jiang, A. Grammenos, E. Kalyvianaki, and T. Charalambous, "An asynchronous approximate distributed alternating direction method of multipliers in digraphs," in *IEEE Conference on Decision and Control (CDC)*, 2021, pp. 3406–3413.

[18] V. Khatana and M. V. Salapaka, "DC-DistADMM: ADMM algorithm for contsrained distributed optimization over directed graphs," *arXiv preprint, arXiv:2003.13742*, 2020.

[19] N. M. Freris, S. R. Graham, and P. R. Kumar, "Fundamental limits on synchronizing clocks over networks," *IEEE Transactions on Automatic Control*, vol. 56, no. 6, pp. 1352–1364, 2010.

[20] D. Yi and N. M. Freris, "Distributed optimization on directed graphs based on inexact ADMM with partial participation," *arXiv e-prints, arXiv:2210.17241*, 2023.

[21] A. Makhdoumi and A. Ozdaglar, "Graph balancing for distributed subgradient methods over directed graphs," in *IEEE Conference on Decision and Control (CDC)*, 2015, pp. 1364–1371.

APPENDIX

Due to length constraints, we provide proof sketches in the following. The complete proofs are accessible at [20].

Proof of Lemma 1: For each k, we consider the $l-$th entry of each $\xi_i^k(\cdot)$ and combine them to form an $n-$dimensional vector denoted by $\widetilde{\xi}^k(\cdot)$, and we let

$$\widetilde{\xi}_\perp^k(\cdot) := \left(I - \mathbf{1}\mathbf{1}^T/n\right)\widetilde{\xi}^k(\cdot).$$

We define $W^k(b) := I - (D-A)\text{diag}\left(w^k(b)\right)$. Then, according to the first property of [16, Lemma 2], there exists $p^k(b)$, so that $W^k(b)p^k(b) = p^k(b)$ and $(p^k(b))^T\mathbf{1} = \underline{1}$. Following the steps in [16] we can obtain that for all $k \geq \bar{k}$

$$\left\|\widetilde{z}_\perp^k + (\mathbf{1}/n - p^k(B))\mathbf{1}^T\widetilde{x}^k\right\| \leq \delta^B\left\|\widetilde{x}_\perp^k + (\mathbf{1}/n - p^k(0))\mathbf{1}^T\widetilde{x}^k\right\|.$$

Letting $\hat{p}^k = \arg\max_{p\in\{p^k(0),p^k(B)\}}\|1/n - p\|$, we get

$$\begin{aligned}\left\|\widetilde{z}_\perp^k\right\| &\leq \delta^B\left\|\widetilde{x}_\perp^k\right\| + \left(1+\delta^B\right)\left\|(\mathbf{1}/n - \hat{p}^k)\mathbf{1}^T\widetilde{x}^k\right\| \\ &\leq \delta^B\left\|\widetilde{x}_\perp^k\right\| + 2\left\|(\mathbf{1}/n - \hat{p}^k)\mathbf{1}^T\widetilde{x}^k\right\|,\end{aligned}$$

which implies (using $\delta \in (0,1), (a+b)^2 \leq \frac{1}{\delta}a^2 + \frac{1}{1-\delta}b^2$ for any $a, b \in \mathbb{R}$, along with Jensen's inequality, and some algebra):

$$\begin{aligned}\left\|\widetilde{z}_\perp^k\right\|^2 &\leq \delta^{2B-1}\left\|\widetilde{x}_\perp^k\right\|^2 + \frac{4}{1-\delta}\left\|(\mathbf{1}/n - \hat{p}^k)\mathbf{1}^T\widetilde{x}^k\right\|^2 \\ &\leq \delta^{2B-1}\left\|\widetilde{x}_\perp^k\right\|^2 \\ &\quad + \frac{8}{1-\delta}\left\|\mathbf{1}/n - \hat{p}^k\right\|^2\left(n\left\|\widetilde{x}^k - \widetilde{x}^\star\right\|^2 + n\|\widetilde{x}^\star\|^2\right).\end{aligned}$$

Since l is arbitrarily chosen from $\{1,\ldots,d\}$, the inequality above holds for each entry position. Adding the d inequalities together and defining $\Delta^k := \frac{8n}{1-\delta}\left\|\mathbf{1}/n - \hat{p}^k\right\|^2$, which in view of the fact that both $\left\|\mathbf{1}/n - p^k(0)\right\|^2$ and $\left\|\mathbf{1}/n - p^k(B)\right\|^2$ tend to zero geometrically with the rate to be $\lambda_2(P)$ [21], completes the proof. ■

Proof of Theorem 1: For notational simplification, we denote $\nabla F\left(x^k\right)$ by ∇F^k and $\nabla F\left(x^\star\right)$ by $\nabla F^\star$. From (4) and (7c) we have

$$\nabla F^k - \nabla F^\star = -\frac{1}{\eta}(x^{k+1} - x^k) - (y^k - y^\star) - \rho(x^k - z^k)$$

and

$$(x^k - x^\star)^T(\nabla F^k - \nabla F^\star) = \underbrace{(x^{k+1} - x^\star)^T(\nabla F^k - \nabla F^\star)}_{(i)} + (x^k - x^{k+1})^T(\nabla F^k - \nabla F^\star).$$

$$(i) = \underbrace{-\frac{1}{\eta}(x^{k+1} - x^\star)^T(x^{k+1} - x^k)}_{(ii)} \underbrace{-(x^{k+1} - x^\star)^T(y^k - y^\star)}_{(iii)} \underbrace{-\rho(x^{k+1} - x^\star)^T(x^k - z^k)}_{(iv)}.$$

The following is to bound the three terms. The detailed calculations can be found in [20].

$$\begin{aligned}
(ii) =& \frac{1}{2\eta}\left(\left\|x^k - x^\star\right\|^2 - \left\|x^{k+1} - x^\star\right\|^2 - \left\|x^{k+1} - x^k\right\|^2\right),\\
(iii) =& -(z_\perp^{k+1})^T(y^k - y^\star) - \frac{1}{2\rho}\left(\left\|y^{k+1} - y^\star\right\|^2 \right.\\
& \left. -\left\|y^k - y^{k+1}\right\|^2 - \left\|y^k - y^\star\right\|^2\right),\\
(iv) =& \frac{\rho}{2}\left(\left\|x^\star - z^k\right\|^2 + \left\|x^k - x^{k+1}\right\|^2 - \left\|x^\star - x^k\right\|^2 \right.\\
& \left. -\left\|x^{k+1} - z^k\right\|^2\right)\\
\le& \frac{\rho}{2}\left(\left\|x^k - x^{k+1}\right\|^2 - \left\|x^{k+1} - z^k\right\|^2 \right.\\
& \left. +\Delta^k\left(\left\|x^k - x^\star\right\|^2 + \|x^\star\|^2\right)\right).
\end{aligned}$$

Adding the three terms gives:

$$\begin{aligned}
(i) \le& \frac{1}{2\eta}\left(\left\|x^k - x^\star\right\|^2 - \left\|x^{k+1} - x^\star\right\|^2\right)\\
&+ \frac{1}{2\rho}\left(\left\|y^k - y^\star\right\|^2 - \left\|y^{k+1} - y^\star\right\|^2\right)\\
&+ (\frac{\rho}{2} - \frac{1}{2\eta})\left\|x^{k+1} - x^k\right\|^2 + \frac{\rho}{2}\Delta^k\left(\left\|x^k - x^\star\right\|^2 + \|x^\star\|^2\right)\\
&\underbrace{+\frac{1}{2\rho}\left\|y^k - y^{k+1}\right\|^2 - \frac{\rho}{2}\left\|x^{k+1} - z^k\right\|^2}_{(v)} \underbrace{- (z_\perp^{k+1})^T(y^k - y^\star)}_{(vi)}.
\end{aligned}$$

By bounding (v) and (vi) and invoking the cocoercivity of the gradient, the choices for B, η, ρ in Thm. 1 and moreover $\gamma = 4, \tau = \frac{3(M_f + m_f)}{4}, \zeta = 3(M_f + m_f)$ guarantee that the following inequalities hold:

$$\frac{1}{4\tau} + \frac{1}{\zeta} - \frac{1}{m_f + M_f} < 0, \tag{12a}$$

$$\tau + \rho + \frac{1}{\gamma\eta} - \frac{1}{2\eta} < 0, \tag{12b}$$

$$\frac{1}{2\eta} - \left[\delta^{2B-1}\left(\frac{5\rho}{4} + \frac{\gamma}{4\eta} + \frac{\zeta}{4}\right) + \rho\right] > 0, \tag{12c}$$

$$\delta^{2B-1}\left(\frac{17\rho}{4} + \frac{\gamma}{4\eta} + \frac{\zeta}{4}\right) + 3\rho - \frac{m_f M_f}{m_f + M_f} < 0. \tag{12d}$$

Consequently, there exist μ_1, μ_2, μ_3, k_1, such that $0 < \mu_1 < \mu_2$ and for all $k > k_1$, the following holds:

$$\begin{aligned}
&\mu_1\left(\left\|x^k - x^\star\right\|^2 + \left\|y^k - y^\star\right\|^2\right) + \mu_3\left(\Delta^k + \Delta^{k+1}\right)\|x^\star\|^2\\
&\ge \left\|x^{k+1} - x^\star\right\|^2 + \left\|y^{k+1} - y^\star\right\|^2,
\end{aligned}$$

where $\mu_1 = \max\left\{\frac{c_1 + c_2}{2c_2}, \left(\frac{1}{2\rho} - c_3\right) / \left(\frac{1}{2\rho}\right)\right\}$ and c_1, c_2, c_3 are specified in the statement of the theorem. ■

Proof of Theorem 2: Let $\Omega^k \in \mathbb{R}^{n\times n}$ be a $0-1$ diagonal matrix where 1 corresponds to the case agent i is active and 0 when it is inactive. Let $\mathbb{E}\left[\Omega^k\right] = \Omega := \text{diag}(q_1, \ldots, q_n)$. The update of w can be written compactly as

$$w^k(b+1) = w^k(b) + \Omega^k\left(Pw^k(b) - w^k(b)\right).$$

Let $w^\infty = \lim_{t\to\infty} P^t w^0(0)$, By taking conditional expectation on the update of w (denoted by $\mathbb{E}^k$, where conditioning is on past activations), we have

$$\begin{aligned}
&\mathbb{E}^k\left[\left\|w^k(b+1) - w^\infty\right\|_{\Omega^{-1}}^2\right]\\
=&\left\|w^k(b) - w^\infty\right\|_{\Omega^{-1}}^2 + \left\|Pw^k(b) - w^k(b)\right\|^2\\
&+ 2\left(w^k(b) - w^\infty\right)^T\left(Pw^k(b) - w^k(b)\right).
\end{aligned}$$

From [21, Lemma 1], there exists some positive $\theta < 1$ only depending on P, such that

$$\mathbb{E}^k\left[\left\|w^k(b+1) - w^\infty\right\|_{\Omega^{-1}}^2\right] \le (1 - q_{\min}\theta)\left\|w^k(b) - w^\infty\right\|_{\Omega^{-1}}^2.$$

By induction, we obtain

$$\mathbb{E}\left[\left\|w^k(b) - w^\infty\right\|^2\right] \le c_1\lambda_2^{kB+b},$$

for some constant c_1 and $\lambda_2 := 1 - q_{\min}\theta$, whence invoking Markov's inequality, we have $\forall\epsilon > 0$, $\exists K_1(\epsilon)$, which implies

$$\Pr\left[\cap_{k=K_1}^\infty \cap_{b=0}^{B-1}\left\{\left\|w^k(b) - w^\infty\right\|^2 < \epsilon_1\right\}\right] \ge 1 - \frac{\epsilon}{2}. \tag{13}$$

We define $v^k := \left[(x^k)^T, (y^k)^T\right]^T$, and let T_a to be the operator corresponding to Alg. 1 for full participation, i.e., $T_a : (x^k, y^k, z^k) \mapsto (x^{k+1}, y^{k+1}, z^{k+1})$ and $T := \begin{bmatrix} I_d & 0 & 0\\ 0 & I_d & 0\end{bmatrix} T_a$.

We define $\Phi^k := \text{blkdiag}\left(\Omega^k \otimes I_d, \Omega^k \otimes I_d\right)$ and $\Phi_a^k := \text{blkdiag}\left(\Omega^k \otimes I_d, \Omega^k \otimes I_d, \Omega^k \otimes I_d\right)$, then $\mathbb{E}\left[\Phi^k\right] = \Phi := \text{blkdiag}\left(\Omega \otimes I_d, \Omega \otimes I_d\right)$ so that

$$\Phi^k T = \Phi^k \begin{bmatrix} I_d & 0 & 0\\ 0 & I_d & 0\end{bmatrix} T_a = \begin{bmatrix} I_d & 0 & 0\\ 0 & I_d & 0\end{bmatrix} \Phi_a^k T_a,$$

which allows the analysis to carry for T only.

For partial participation, we have

$$v^{k+1} = v^k + \Phi^k\left(Tv^k - v^k\right),$$

i.e.,

$$\left\|v^{k+1} - v^\star\right\|_{\Phi^{-1}}^2 = \left\|v^k - v^\star + \Phi^k\left(Tv^k - v^k\right)\right\|_{\Phi^{-1}}^2.$$

An analogous use of Markov's inequality in the establishment of (13) concludes there exists some K_2, such that

$$\Pr\left[\cap_{k=K_2}^\infty \left\|v^{k+1} - v^\star\right\|_{\Phi^{-1}}^2 \le \lambda_p^k\right] \ge 1 - \frac{\epsilon}{2},$$

where λ_p can be arbitrarily chosen in $(\lambda_1, 1)$ and λ_1 is as in the statement of Thm. 2. ■