

Distributed Difference of Convex Optimization

Vivek Khatana, Murti V. Salapaka

Abstract—In this article, we focus on solving a class of distributed optimization problems involving n agents with the local objective function at every agent i given by the difference of two convex functions f_i and g_i (difference-of-convex (DC) form), where f_i and g_i are potentially nonsmooth. The agents communicate via a directed graph containing n nodes. We create smooth approximations of the functions f_i and g_i and develop a distributed algorithm utilizing the gradients of the smooth surrogates and a finite-time approximate consensus protocol. We term this algorithm as DDC-Consensus. The developed DDC-Consensus algorithm allows for non-symmetric directed graph topologies and can be synthesized distributively. We establish that the DDC-Consensus algorithm converges to a stationary point of the nonconvex distributed optimization problem. The performance of the DDC-Consensus algorithm is evaluated via a simulation study to solve a nonconvex DC-regularized distributed least squares problem. The numerical results corroborate the efficacy of the proposed algorithm.

keywords: Distributed optimization, nonconvex optimization, difference-of-convex (DC) functions, DC programming, distributed gradient descent, directed graphs.

I. INTRODUCTION

Due to the increase in size and complexity of modern systems, solving problems involving many agents via distributed methods is highly desirable. An effective way towards a distributed solution is to cast the problems in the framework of distributed optimization [1], [2]. In this article, we focus on the following distributed optimization problem,

$$\underset{x \in \mathbb{R}^p}{\text{minimize}} \quad F(x) = \frac{1}{n} \sum_{i=1}^n (f_i(x) - g_i(x)), \quad (1)$$

where $x \in \mathbb{R}^p$ is a common decision, and $f_i : \mathbb{R}^p \rightarrow \mathbb{R} \cup \{-\infty, \infty\}$ and $g_i : \mathbb{R}^p \rightarrow \mathbb{R}$ are private objective functions of agent i . The agents are connected through a directed graph, $\mathcal{G}(\mathcal{V}, \mathcal{E})$, where $\mathcal{V}$ and $\mathcal{E}$ are the set of vertices and edges respectively. The terms in the aggregate objective function in (1) are of a difference-of-convex (DC) form. Due to the richness of the set of DC functions (see the properties mentioned in [3]), DC functions can be used to model most of all practical non-convex optimization problems. Many applications in statistical learning and estimation [4], power systems [5], computational biology [6], signal restoration [7], network optimization [8], combinatorial optimization [9] can be posed as DC programs.

The initial study of distributed optimization can be traced back to the seminal works [10], [11]. Since then the special case with f_i being convex and $g_i = 0$ has received significant development (see [12]–[14] and references therein). As a relaxation of the convex objective functions most works

in the literature are restricted to Lipschitz differentiable not necessarily convex functions f_i with $g_i = 0$ (some examples include [15]–[21]) and do not address the general problem (1). Article [22] focuses on a special case with weakly convex objective functions f_i and $g_i = 0$. Article [23] considers a DC form however the functions f_i and g_i are assumed to have Lipschitz differentiable gradients, and does not ensure globally agreed decisions among all the agents. To the best of the authors' knowledge, the general *distributed* optimization problem (1) with DC functions is not addressed in prior works.

However, DC programming is studied in the centralized optimization literature [24]–[30]. Articles [24]–[28] utilize an affine minorization g_i^m of the functions $g_i, i = 1$ and minimize the resulting convex function $f_i - g_i^m, i = 1$. The work in [29] provides an accelerated version of the algorithms proposed in [24]–[28] using adaptive step-sizes and extrapolation of the algorithm iterates. Article [30] proposed an alternating direction of multipliers method based algorithm for solving centralized DC optimization problems with Lipschitz differentiable functions and a linear equality constraint. The current article is closest to the centralized algorithms [31], [32], where a majorization of the objective functions via the proximal mapping is solved.

In this article, we propose an iterative algorithm termed **D**istributed **DC**-Consensus DDC-Consensus. The DDC-Consensus algorithm proceeds in two steps: first, all the agents in parallel solve a local optimization problem involving $f_i - g_i$, then each agent $i \in \mathcal{V}$ shares the obtained solution with its neighbors where the estimates of the global solution are updated by utilizing an average consensus protocol (the algorithm is presented in detail in Section II-B). The main contributions of the current article are as follows:

- i) We develop a distributed algorithm, DDC-Consensus, based on distributed gradient descent to solve (1) for non-differentiable weakly convex f_i and non-differentiable convex function g_i . These assumptions are weaker than the existing literature on distributed optimization.
- ii) We establish that the DDC-Consensus converges to a stationary solution of the nonconvex problem (1).
- iii) The proposed DDC-Consensus algorithm has desirable properties as it (a) is suitable for directed graphs and non-symmetric communication topologies, and (b) can be implemented as well as synthesized distributively.

We demonstrate the performance of the DDC-Consensus algorithm by solving a DC-regularized least squares problem. Empirical results from the numerical simulation study corroborate that the proposed DDC-Consensus algorithm performs well in practice.

The rest of the article is organized as follows: Some essential definitions and notations are presented in Sec-

Vivek Khatana (Email: {khata010}@umn.edu) and Murti V. Salapaka (Email: {murtis}@umn.edu) are with the Department of Electrical and Computer Engineering, University of Minnesota, Minneapolis, USA. This research is supported by the United States Department of Energy via grant number DE-CR0000040.

tion I-A. Section II provides the details of the proposed DDC-Consensus algorithm. Convergence analysis of DDC-Consensus is presented in Section III. Section IV presents the numerical simulation study followed by concluding remarks in Section V.

A. Definitions and Notations

Definition 1. (Directed Graph) A directed graph $\mathcal{G}$ is a pair $(\mathcal{V}, \mathcal{E})$ where $\mathcal{V}$ is a set of vertices and $\mathcal{E}$ is a set of edges, which are ordered subsets of two distinct elements of $\mathcal{V}$. If an edge from $j \in \mathcal{V}$ to $i \in \mathcal{V}$ exists it is denoted as $(i, j) \in \mathcal{E}$.

Definition 2. (Strongly Connected Graph) A directed graph is strongly connected if for any pair (i, j) , $i \neq j$, there is a directed path from node i to node j .

Definition 3. (Diameter of a Graph) The longest shortest directed path between any two nodes in the graph.

Definition 4. (In-Neighborhood) The set of in-neighbors of node $i \in \mathcal{V}$ is called the in-neighborhood of node i and is denoted by $\mathcal{N}_i^- = \{j \mid (i, j) \in \mathcal{E}\}$ not including the node i .

Definition 5. (Column-Stochastic Matrix) A matrix $M = [m_{ij}] \in \mathbb{R}^{n \times n}$ is called a column-stochastic matrix if $0 \leq m_{ij} \leq 1$ and $\sum_{i=1}^n m_{ij} = 1$ for all $1 \leq j \leq n$.

Definition 6. (Lipschitz Continuity and Differentiability) A function $f : \mathbb{R}^p \rightarrow \mathbb{R}$ is called Lipschitz continuous with constant $L > 0$ and Lipschitz differentiable with constant $L_f > 0$ respectively, if the following inequalities hold:

$$|f(x) - f(y)| \leq L\|x - y\|, \quad \forall x, y \in \mathbb{R}^p$$

$$\|\nabla f(x) - \nabla f(y)\| \leq L_f\|x - y\|, \quad \forall x, y \in \mathbb{R}^p.$$

Definition 7. (Strongly Convex Function) A differentiable function f is called strongly convex with parameter σ , if there exists $\sigma > 0$ such that for all x, y in the domain of f :

$$\langle \nabla f(x) - \nabla f(y), x - y \rangle \geq \sigma\|x - y\|^2, \quad \forall x, y \in \text{dom } f.$$

Definition 8. (Weakly Convex Function) For some $m_f \geq 0$, we say that function $f : \mathbb{R}^p \rightarrow \mathbb{R}$ is m_f -weakly convex if $f + \frac{m_f}{2}\|x\|^2$ is convex.

Definition 9. (Level-bounded Function) Function $f : \mathbb{R}^p \rightarrow \mathbb{R}$ is called level-bounded if the β -level-set $\{x : f(x) \leq \beta\}$ is bounded (possibly empty) for all $\beta \in \mathbb{R}$.

Definition 10. (Lower Semicontinuity) A function $f : \mathbb{R}^p \rightarrow \mathbb{R} \cup \{-\infty, \infty\}$ is called lower semi-continuous at a point x_0 if for every real $y < f(x_0)$ there exists a neighborhood U of x_0 such that $y < f(x)$ for all $x \in U$. The function f is lower semicontinuous at every point of its domain, $\text{dom}(f)$.

Definition 11. (Proper Function) A function $f : \mathbb{R}^p \rightarrow \mathbb{R} \cup \{-\infty, \infty\}$ is called proper if $f(x) > -\infty$ for every x and and if there also exists some point x_0 such that $f(x_0) < \infty$.

$\|x\|_1$ and $\|x\|$ denote the 1-norm and 2-norm of the vector $x \in \mathbb{R}^p$ respectively. The notation $\lceil \cdot \rceil$ denotes the least integer function or the ceiling function, defined as: given $x \in \mathbb{R}$, $\lceil x \rceil = \min\{m \in \mathbb{Z} \mid m \geq x\}$, where $\mathbb{Z}$ is the set of integers. $\text{dom}(f) = \{x : f(x) < \infty\}$.

II. PROPOSED METHODOLOGY

The following assumption is satisfied throughout the article,

Assumption 1. 1. Functions $f_i : \mathbb{R}^p \rightarrow \mathbb{R} \cup \{-\infty, \infty\}$ are proper, lower semi-continuous, and m_{f_i} -weakly convex.
2. Functions $g_i : \mathbb{R}^p \rightarrow \mathbb{R}$ are convex and finite everywhere.
3. The set of global minimizers of (1), $\arg \text{minimize}_x F(x)$, is nonempty, and the global minimum value F^* is finite.

A vector $\tilde{y} \in \mathbb{R}^p$ is called a stationary point of F if:

$$0 \in \partial \left(\frac{1}{n} \sum_{i=1}^n f_i(\tilde{y}) \right) - \partial \left(\frac{1}{n} \sum_{i=1}^n g_i(\tilde{y}) \right), \quad (2)$$

or equivalently, $\partial \frac{1}{n} \sum_{i=1}^n f_i(\tilde{y}) \cap \partial \frac{1}{n} \sum_{i=1}^n g_i(\tilde{y}) \neq \emptyset$. Furthermore, given $\varepsilon > 0$, we say $\tilde{y} \in \mathbb{R}^p$ is an ε -stationary point of F if there exists $(\xi; y) \in \mathbb{R}^p \times \mathbb{R}^p$ such that

$$\xi \in \partial \left(\frac{1}{n} \sum_{i=1}^n f_i(\tilde{y}) \right) - \partial \left(\frac{1}{n} \sum_{i=1}^n g_i(y) \right), \quad \text{and} \quad \max \{ \|\xi\|, \|\tilde{y} - y\| \} \leq \varepsilon. \quad (3)$$

We use $\partial f(x)$ to denote the general subdifferential of the function f at x ([33], Definition 8.3).

A. Approach Roadmap and Supporting Results

Note that the objective function has a DC structure. Under Assumption 1 we do not impose differentiability of the functions f_i and g_i . Thus, to develop an algorithm with desirable convergence properties, we first utilize a smooth function mapping to obtain a differentiable DC approximation of the objective function $f_i - g_i$. Then we utilize the differentiable DC approximation to develop first-order algorithms to solve problem (1). We take the approach in [31] and utilize separate Moreau envelopes of f_i and g_i to obtain a smooth approximation of the function $f_i - g_i$. To this end, we define the Moreau envelopes: Given $0 < \mu_i < 1/m_{f_i}$, the Moreau envelopes, $M_{\mu_i f_i} : \mathbb{R}^p \rightarrow \mathbb{R}$, and $M_{\mu_i g_i} : \mathbb{R}^p \rightarrow \mathbb{R}$ of functions f_i and g_i respectively are given by

$$M_{\mu_i f_i}(y) := \underset{x \in \mathbb{R}^p}{\text{minimize}} \left\{ f_i(x) + \frac{1}{2\mu_i} \|x - y\|^2 \right\}, \quad (4)$$

$$M_{\mu_i g_i}(y) := \underset{x \in \mathbb{R}^p}{\text{minimize}} \left\{ g_i(x) + \frac{1}{2\mu_i} \|x - y\|^2 \right\}, \quad (5)$$

and form the smooth function $F_{i, \mu_i} := M_{\mu_i f_i} - M_{\mu_i g_i}$, for all $i \in \{1, \dots, n\}$. The corresponding proximal mappings $x_{\mu_i f_i} : \mathbb{R}^p \rightarrow \mathbb{R}^p$, and $x_{\mu_i g_i} : \mathbb{R}^p \rightarrow \mathbb{R}^p$ are defined as,

$$x_{\mu_i f_i}(y) := \arg \underset{x \in \mathbb{R}^p}{\text{minimize}} \left\{ f_i(x) + \frac{1}{2\mu_i} \|x - y\|^2 \right\}, \quad (6)$$

$$x_{\mu_i g_i}(y) := \arg \underset{x \in \mathbb{R}^p}{\text{minimize}} \left\{ g_i(x) + \frac{1}{2\mu_i} \|x - y\|^2 \right\}. \quad (7)$$

Define, $m_f := \max_{1 \leq i \leq n} m_{f_i}$. Given, $0 < \mu < 1/m_f$, let $F_{i, \mu}(x) := M_{\mu f_i}(x) - M_{\mu g_i}(x)$, and

$$F_\mu(x) := \frac{1}{n} \sum_{i=1}^n F_{i, \mu}(x) = \frac{1}{n} \sum_{i=1}^n (M_{\mu f_i}(x) - M_{\mu g_i}(x)). \quad (8)$$

Next, we summarize the properties of $F_{i, \mu}$, the Moreau envelopes $M_{\mu f_i}$, $M_{\mu g_i}$ the mappings $x_{\mu f_i}$, $x_{\mu g_i}$.

Lemma 1. ([31], Properties of Moreau envelopes and proximal mappings) Let Assumption 1 holds. Let $0 < \mu < 1/m_f$ and $M_{\mu f_i}$, $M_{\mu g_i}$, $x_{\mu f_i}$, $x_{\mu g_i}$ be given by definitions (4)-(7) with parameter μ . Then, the following claims hold:

1. $x_{\mu f_i}, x_{\mu g_i}$ are Lipschitz continuous with modulus, $\frac{1}{1-\mu m_f}$ and 1 respectively.
2. $M_{\mu f_i}, M_{\mu g_i}$ are differentiable with gradient $\nabla M_{\mu f_i}(y) = \mu^{-1}(y - x_{\mu f_i}(y))$, $\nabla M_{\mu g_i}(y) = \mu^{-1}(y - x_{\mu g_i}(y))$.
3. $\nabla M_{\mu f_i}, \nabla M_{\mu g_i}$ are Lipschitz continuous with modulus $L_{\mu M_f} = \frac{2-\mu m_f}{\mu-\mu^2 m_f}$ and $L_{\mu M_g} = \frac{2}{\mu}$ respectively.
4. $F_{i,\mu}$ is differentiable, and $\nabla F_{i,\mu}(y) = \nabla M_{\mu f_i}(y) - \nabla M_{\mu g_i}(y) = \mu^{-1}(x_{\mu g_i}(y) - x_{\mu f_i}(y))$ is Lipschitz continuous with the modulus $L_{\mu F} := \frac{2-\mu m_f}{\mu-\mu^2 m_f}$.

Next, we introduce the following optimization problem,

$$\underset{x \in \mathbb{R}^p}{\text{minimize}} \quad F_\mu(x) = \frac{1}{n} \sum_{i=1}^n (M_{\mu f_i}(x) - M_{\mu g_i}(x)). \quad (9)$$

Lemma 2. ([31], Relation between (1) and (9)) *Let Assumption 1 holds and $0 < \mu < 1/m_f$. Then,*

1. *The set of global minimizers of (9), $\arg \text{minimize}_x F_\mu(x)$, is non-empty, and $F^* = \text{minimize}_{x \in \mathbb{R}^p} F_\mu(x)$.*
2. *$\tilde{y}$ is a stationary point of F_μ , i.e. $\nabla F_\mu(\tilde{y}) = 0$, if and only if $\tilde{x} = \frac{1}{n} \sum_{i=1}^n x_{\mu f_i}(\tilde{y}) = \frac{1}{n} \sum_{i=1}^n x_{\mu g_i}(\tilde{y})$ is a stationary point of F in the sense of (2).*
3. *$y^* \in \arg \text{minimize}_y F_\mu(y)$ if and only if $\frac{1}{n} \sum_{i=1}^n x_{\mu f_i}(y^*) = \frac{1}{n} \sum_{i=1}^n x_{\mu g_i}(y^*) = x^* \in \arg \text{minimize}_y F(y)$.*

The minimization problem (1) is generally challenging due to the nonconvexity and nonsmoothness of the objective functions. In contrast, problem (9) with the approximation F_μ provides an attractive surrogate: as ∇F_μ is Lipschitz continuous, which is a desirable property for a wide range of first-order methods. Moreover, from Lemmas 1 and 2, F_μ also largely preserves the geometric structure of F . Obtaining a stationary solution of F_μ , we can recover its counterpart for F via the proximal mappings $x_{\mu f_i}$ and $x_{\mu g_i}$. Thus, we focus on solving problem (9).

B. Proposed DDC-Consensus Algorithm

Problem (9) can be recast by creating local copies y_i for all $i \in \{1, 2, \dots, n\}$, of the solution y to problem (9) and imposing the agreement of the solutions of all the agents via consensus constraint leading to the equivalent problem,

$$\begin{aligned} & \underset{y_1, y_2, \dots, y_n}{\text{minimize}} \quad \frac{1}{n} \sum_{i=1}^n (M_{\mu f_i}(y_i) - M_{\mu g_i}(y_i)), \quad (10) \\ & \text{subject to } y_i = y_j, \text{ for all } i, j. \end{aligned}$$

We develop a distributed gradient descent-based algorithm to solve problem (10) called the DDC-Consensus algorithm. The algorithm proceeds in the following manner:

At any iteration k of the algorithm, each agent i maintains two estimates, an *optimization variable* $y_i^{(k)} \in \mathbb{R}^p$, and a local *update variable* $z_i^{(k)} \in \mathbb{R}^p$. Every iteration k involves two updates: first, each agent i updates $z_i^{(k)}$ via local gradient descent based on the mapping $F_{i,\mu}$, with the gradient evaluated at $y_i^{(k-1)}$; next, the optimization variable $y_i^{(k)}$ is updated to an estimate which is η_k -close to the average value $\hat{z}^{(k)} := \frac{1}{n} \sum_{i=1}^n z_i^{(k)}$, i.e., $\|y_i^{(k)} - \hat{z}^{(k)}\| \leq \eta^{(k)}$, using the distributed η -consensus protocol (described in detail in the next section), initialized with $z_i^{(k)}$ as the initial condition for the agent i and tolerance η_k . The above algorithm updates

at any agent i in are summarized in the next equations:

$$z_i^{(k+1)} = y_i^{(k)} - \alpha \nabla \frac{1}{n} F_{i,\mu}(y_i^{(k)}) \quad (11)$$

$$\begin{aligned} &= y_i^{(k)} - \frac{\alpha}{n} \nabla (M_{\mu f_i}(y_i^{(k)}) - M_{\mu g_i}(y_i^{(k)})) \\ &= y_i^{(k)} - \frac{\alpha \mu^{-1}}{n} (x_{\mu g_i}(y_i^{(k)}) - x_{\mu f_i}(y_i^{(k)})), \quad (12) \end{aligned}$$

$$y_i^{(k+1)} = w_i^{(t_{k+1})}, \text{ where} \quad (13)$$

$$\|w_i^{(t_{k+1})} - \hat{z}^{(k+1)}\| < \eta^{(k+1)}, \quad \hat{z}^{(k+1)} = \frac{1}{n} \sum_{i=1}^n z_i^{(k+1)} \quad (14)$$

where, $w_i^{(t_k)}$ is the output of the η -consensus protocol and is an approximate estimate of the average $\hat{z}^{(k)}$, t_k denotes the number of communication steps utilized by the η -consensus protocol at iteration k of the DDC-Consensus algorithm. We summarize DDC-Consensus in Algorithm 1.

Algorithm 1: DDC-Consensus

Input:

$0 < \mu < 1/m_f$; step-size: $0 < \alpha \leq 1/L_{\mu F}$;
consensus tolerances $\{\eta^{(k)}\}_{k \geq 0}$

Initialize:

For each agent $i \in \mathcal{V}$, $y_i^{(0)} = z_i^{(0)} \in \mathbb{R}^p$;

Repeat for $k = 0, 1, 2, \dots$

for $i = 1, 2, 3, \dots, n$, (*In parallel*) **do**

*/ * gradient descent iteration: **

$$z_i^{(k+1)} = y_i^{(k)} - \frac{\alpha \mu^{-1}}{n} (x_{\mu g_i}(y_i^{(k)}) - x_{\mu f_i}(y_i^{(k)}))$$

*/ * consensus iterations: **

$$y_i^{(k+1)} \leftarrow \eta^{(k+1)}\text{-consensus}(z_i^{(k+1)}, i \in \mathcal{V})$$

end

until a stopping criterion is met

The finite-time η -consensus protocol is described next.

C. Finite-time η -consensus Protocol

The finite-time η -consensus protocol and its variants are proposed in earlier articles [34], [35] by the authors under various practical scenarios. The finite-time η -consensus protocol aims to design a distributed protocol so that the agents can compute an approximate estimate of the average, $\hat{u} := \frac{1}{n} \sum_{i=1}^n u_i^{(0)}$, of their initial states $u_i^{(0)}$ in finite-time. This approximate estimate is parameterized by a small tolerance η chosen apriori to make the estimate precise. In the η -consensus protocol, the agents maintain state variables $u_i^{(k)} \in \mathbb{R}^p, v_i^{(k)} \in \mathbb{R}$ and update them as:

$$u_i^{(k)} = p_{ii} u_i^{(k-1)} + \sum_{j \in \mathcal{N}_i^-} p_{ij} u_j^{(k-1)} \quad (15)$$

$$v_i^{(k)} = p_{ii} v_i^{(k-1)} + \sum_{j \in \mathcal{N}_i^-} p_{ij} v_j^{(k-1)} \quad (16)$$

$$w_i^{(k)} = \frac{1}{v_i^{(k)}} u_i^{(k)}, \quad (17)$$

where, $w_i^{(0)} = u_i^{(0)}, v_i^{(0)} = 1$ for all $i \in \mathcal{V}$ and $\mathcal{N}_i^-$ denotes the set of in-neighbors of agent i . The updates (15)-(17) are based on the push-sum (or ratio consensus) updates (see [36]). The following assumption on the graph $\mathcal{G}(\mathcal{V}, \mathcal{E})$ and weight matrix $\mathbf{P} := [p_{ij}]$ is made:

Assumption 2. *The directed graph $\mathcal{G}(\mathcal{V}, \mathcal{E})$ is strongly connected and the weight matrix $\mathbf{P}$ is column-stochastic.*

Note that $\mathbf{P}$ being a column stochastic matrix allows for a distributed synthesis of the η -consensus protocol. The

variable $w_i^{(k)} \in \mathbb{R}^p$ is an estimate of the average $\hat{u}$ with each agent i at any iteration k . The estimates $w_i^{(k)}$ converge to the average $\hat{u}$ asymptotically.

Theorem 1. ([36],[37]) *Let Assumption 2 hold. Let $\{w_i^{(k)}\}_{k \geq 0}$ be generated by (17). Then $w_i^{(k)}$ asymptotically converges to $\hat{u} = \frac{1}{n} \sum_{i=1}^n u_i^{(0)}$ for all $i \in \mathcal{V}$, i.e.,*

$$\lim_{k \rightarrow \infty} w_i^{(k)} = \frac{1}{n} \sum_{i=1}^n u_i^{(0)}, \text{ for all } i \in \mathcal{V}.$$

Every agent $i \in \mathcal{V}$ maintains an additional scalar value r_i to determine when the state $w_i^{(k)}$ are η -close to each other and hence from Theorem 1, η -close to $\hat{u}$. The radius variable w_i is designed to track the radius of the minimal ball that can enclose all the states $w_i^{(k)}$ (more details can be found in [34], [35]). The radius $w_i^{(k)}$, for all $i \in \mathcal{V}$, is updated as:

$$r_i^{(k)} := \begin{cases} 0, & \text{if } k = m\mathcal{D}, m = 0, 1, 2, \dots \\ \max_{j \in \mathcal{N}_i} \{ \|w_i^{(k)} - w_j^{(k-1)}\| + r_j^{(k-1)} \}, & \text{otherwise,} \end{cases} \quad (18)$$

where $\mathcal{D}$ is an upper bound on the diameter of the directed graph $\mathcal{G}$. Denote, $\mathcal{B}(w_i^{(k)}, r_i^{(k)})$ as the p -dimensional ball of radius $r_i^{(k)}$ centered at $w_i^{(k)}$. It is established in [34] that under Assumption 2, after $\mathcal{D}$ iterations of update (18), the ball $\mathcal{B}(w_i^{(k+\mathcal{D})}, r_i^{(k+\mathcal{D})})$ encloses the states $w_i^{(k)}$ of all the agents $i \in \mathcal{V}$. Further, it is also established that radius update sequences $\{r_i^{(k)}\}_{k \geq 0}$ converge to zero as $m \rightarrow \infty$.

Theorem 2. ([34], [35]) *Let updates (15)-(17) hold. Let $\{r_i^{(k)}\}_{k \geq 0}$ be the sequences generated by (18). Under Assumption 2,*

$$\lim_{k \rightarrow \infty} r_i^{(k)} = 0, \text{ for all } i \in \mathcal{V}.$$

Theorem 2 gives a criterion for termination of the consensus iterations (15)-(17) by utilizing the radius updates at each agent $i \in \mathcal{V}$ given by (18). Using the value of the $r_i^{(k)}, k = 0, 1, 2, \dots$ each agent can ascertain the deviation from the consensus among the agents. As a method to detect η -consensus, at every iteration of the form $m\mathcal{D}$, for $m = 1, 2, \dots$, $r_i^{(m\mathcal{D})}$ is compared to the tolerance η , if $r_i^{(m\mathcal{D})} < \eta$ then all the agent states $w_i^{((m-1)\mathcal{D})}$ were η -close to $\hat{u}$ (from Theorem 1) and the iterations (15)-(17) are terminated. Proposition 1 establishes that the η -consensus protocol converges in a finite number of iterations.

Proposition 1. *Under the Assumption 2, η -consensus is achieved in a finite number of iterations k_η at each agent $i \in \mathcal{V}$. Moreover, k_η satisfies*

$$k_\eta \geq \left\lceil \frac{\log(1/\eta)}{-\log(\delta_2)} + \frac{\log(C\|\mathbf{u}^{(0)}\|_1/\delta_1)}{-\log(\delta_2)} \right\rceil, \quad (19)$$

where, $\mathbf{u}^{(0)} := [u_1^{(0)}, \dots, u_n^{(0)}] \in \mathbb{R}^{n \times p}$ and $C, \delta_1, \delta_2 \in (0, 1)$ are constants related to the matrix $\mathbf{P}$ and graph $\mathcal{G}$.

Proof. Note, $r_i^{(k)} \rightarrow 0$ as $k \rightarrow \infty$. Thus, given $\eta > 0, i \in \mathcal{V}$ there exists finite $k_{i,\eta}$ such that for $k \geq k_{i,\eta}, r_i^{(k)} < \eta$, for all $i \in \mathcal{V}$. Choosing $k_\eta := \max_{1 \leq i \leq n} k_{i,\eta}$ establishes the claim. Further, (19) is proved rigorously in [14], Lemma 3.1. $\square$

To have a global detection, each agent generates a one-bit “converged flag” indicating its detection. The flag signal can be combined using distributed one-bit consensus updates (see [34]) allowing the agents to achieve global η -consensus.

III. CONVERGENCE ANALYSIS FOR DDC-CONSENSUS

Let the average of the optimization variables at iteration k be denoted as: $\hat{y}^{(k)} := \frac{1}{n} \sum_{i=1}^n y_i^{(k)}$. Denote the gradient of the function $\frac{1}{n} \sum_{i=1}^n F_{i,\mu}$ evaluated at individual optimization variables of all the agents and the average $\hat{y}^{(k)}$ as:

$$h^{(k)} := \frac{1}{n} \sum_{i=1}^n \nabla F_{i,\mu}(y_i^{(k)}), \text{ and} \quad (20)$$

$$\hat{h}^{(k)} := \frac{1}{n} \sum_{i=1}^n \nabla F_{i,\mu}(\hat{y}^{(k)}) \quad (21)$$

respectively. A consequence of the η -consensus protocol is that the difference between $h^{(k)}$ and $\hat{h}^{(k)}$ is bounded.

Lemma 3. *Let $\eta^{(k)}$ denote the consensus tolerance in the DDC-Consensus algorithm at iteration k , then*

$$\|h^{(k)} - \hat{h}^{(k)}\| \leq 2L_{\mu F}\eta^{(k)},$$

where $L_{\mu F}$ is the constant as defined in Lemma 1.

Proof. Note that,

$$\begin{aligned} \|h^{(k)} - \hat{h}^{(k)}\| &= \left\| \frac{1}{n} \sum_{i=1}^n \nabla F_{i,\mu}(y_i^{(k)}) - \frac{1}{n} \sum_{i=1}^n \nabla F_{i,\mu}(\hat{y}^{(k)}) \right\| \\ &= \frac{1}{n} \left\| \left(\sum_{i=1}^n \nabla M_{\mu f_i}(y_i^{(k)}) - \sum_{i=1}^n \nabla M_{\mu g_i}(y_i^{(k)}) \right) \right. \\ &\quad \left. - \left(\sum_{i=1}^n \nabla M_{\mu f_i}(\hat{y}^{(k)}) - \sum_{i=1}^n \nabla M_{\mu g_i}(\hat{y}^{(k)}) \right) \right\| \\ &= \frac{1}{\mu n} \left\| \left(\sum_{i=1}^n \left(x_{\mu g_i}(y_i^{(k)}) - x_{\mu f_i}(y_i^{(k)}) \right) \right) \right. \\ &\quad \left. - \left(\sum_{i=1}^n \left(x_{\mu g_i}(\hat{y}^{(k)}) - x_{\mu f_i}(\hat{y}^{(k)}) \right) \right) \right\| \\ &= \frac{1}{\mu n} \left\| \left(\sum_{i=1}^n \left(x_{\mu f_i}(\hat{y}^{(k)}) - x_{\mu f_i}(y_i^{(k)}) \right) \right) \right. \\ &\quad \left. + \left(\sum_{i=1}^n \left(x_{\mu g_i}(y_i^{(k)}) - x_{\mu g_i}(\hat{y}^{(k)}) \right) \right) \right\| \\ &\leq \frac{1}{n\mu(1-\mu m_f)} \sum_{i=1}^n \|y_i^{(k)} - \hat{y}^{(k)}\| + \frac{1}{n\mu} \sum_{i=1}^n \|y_i^{(k)} - \hat{y}^{(k)}\| \\ &= \frac{2-\mu m_f}{n(\mu-\mu^2 m_f)} \sum_{i=1}^n \|y_i^{(k)} - \hat{y}^{(k)}\| \\ &= \frac{2-\mu m_f}{n(\mu-\mu^2 m_f)} \sum_{i=1}^n \|y_i^{(k)} - \hat{y}^{(k)} + \hat{z}^{(k)} - \hat{z}^{(k)}\| \\ &\leq \frac{L_{\mu F}}{n} \sum_{i=1}^n \|y_i^{(k)} - \hat{z}^{(k)}\| + \frac{L_{\mu F}}{n} \sum_{i=1}^n \|\hat{z}^{(k)} - \hat{y}^{(k)}\| \\ &\leq \frac{L_{\mu F}}{n} \sum_{i=1}^n (\eta^{(k)} + \eta^{(k)}) \leq 2L_{\mu F}\eta^{(k)}, \end{aligned}$$

where, $L_{\mu F} = \frac{2-\mu m_f}{(\mu-\mu^2 m_f)}$ as defined in Lemma 1 and the last step follows from (13) and (14). $\square$

Recall (11), $\frac{1}{n} \sum_{i=1}^n z_i^{(k+1)} = \frac{1}{n} \sum_{i=1}^n y_i^{(k)} - \alpha \nabla F_{i,\mu}(y_i^{(k)}) = \hat{y}^{(k)} - \alpha h^{(k)}$. Therefore, we have, $\hat{z}^{(k+1)} = \hat{y}^{(k)} - \alpha h^{(k)} + \hat{y}^{(k+1)} - \hat{y}^{(k+1)}$

$$\implies \hat{y}^{(k+1)} = \hat{y}^{(k)} - \alpha h^{(k)} + a^{(k+1)}, \quad (22)$$

with $a^{(k+1)} := \hat{y}^{(k+1)} - \hat{z}^{(k+1)}, \|a^{(k+1)}\| \leq \eta^{(k+1)}$. In the centralized setting, the information about the gradient of the function $\frac{1}{n} \sum_{i=1}^n F_{i,\mu}$, i.e. $\nabla \frac{1}{n} \sum_{i=1}^n F_{i,\mu}(y) = \frac{1}{n} \sum_{i=1}^n \nabla F_{i,\mu}(y) = h_y$ is known to the solver. Here, an iteration of the (centralized) gradient descent is of the form: $\tilde{y} = y - \alpha h_y$, where, $\tilde{y}$ denotes the updated estimate of

the solution. Due to Lemma 3, update (22) can be viewed as an “inexact” centralized gradient descent update on the gradient $\frac{1}{n} \sum_{i=1}^n \nabla F_{i,\mu}$ evaluated at the average of all the agents’ optimization variables for the function $\frac{1}{n} \sum_{i=1}^n F_{i,\mu}$,

$$\hat{y}^{(k+1)} = \hat{y}^{(k)} - \alpha \hat{h}^{(k)} + e^{(k)}, \text{ with} \quad (23)$$

$$e^{(k)} := \alpha(\hat{h}^{(k)} - h^{(k)}) + a^{(k+1)}, \quad (24)$$

$$\|e^{(k)}\| \leq 2\alpha L_{\mu F} \eta^{(k)} + \eta^{(k+1)}. \quad (25)$$

Therefore, the proposed DDC-Consensus algorithm updates are equivalent to an (approximate) centralized gradient descent update to minimize the function F_μ .

Theorem 3. *Let assumptions 1 and 2 hold, $\alpha \leq 1/L_{\mu F}$, and the tolerance sequence in DDC-Consensus is $\eta^{(k)} = 1/k^{1+\theta}$, for all k , with a positive constant θ . Suppose $\{y_i^{(k)}, x_{\mu f_i}(y_i^{(k)}), x_{\mu g_i}(y_i^{(k)})\}_{k \geq 0, i \in \mathcal{V}}$ be the sequences generated by Algorithm 1. Let $\hat{y}^{(k)} := \frac{1}{n} \sum_{i=1}^n y_i^{(k)}$, $\hat{\xi}^{(k)} := \mu^{-1}(\frac{1}{n} \sum_{i=1}^n x_{\mu g_i}(\hat{y}^{(k)}) - \frac{1}{n} \sum_{i=1}^n x_{\mu f_i}(\hat{y}^{(k)}))$, and $\xi^{(k)} := \mu^{-1}(\frac{1}{n} \sum_{i=1}^n x_{\mu g_i}(y_i^{(k)}) - \frac{1}{n} \sum_{i=1}^n x_{\mu f_i}(y_i^{(k)}))$. Then, for any positive integer K , there exists $0 \leq \bar{k} \leq K-1$ such that*

$$\begin{aligned} \hat{\xi}^{(\bar{k})} &\in \partial \frac{1}{n} \sum_{i=1}^n f_i(x_{\mu f_i}(\hat{y}^{(\bar{k})})) - \partial \frac{1}{n} \sum_{i=1}^n g_i(x_{\mu g_i}(\hat{y}^{(\bar{k})})), \\ \max\{\|\hat{\xi}^{(\bar{k})}\|, \|\frac{1}{n} \sum_{i=1}^n x_{\mu g_i}(\hat{y}^{(\bar{k})}) - \frac{1}{n} \sum_{i=1}^n x_{\mu f_i}(\hat{y}^{(\bar{k})})\|\} \\ &\leq \max\{1, \mu^{-1}\} \left(\frac{2\mu^2 (F_\mu(\hat{y}^{(0)}) - F^* + \sum_{k=0}^\infty \frac{2C}{k^{1+\theta}})}{\alpha K} \right)^{1/2}. \end{aligned}$$

Therefore, given $\varepsilon > 0$, in no more than

$$K \leq \left\lceil \frac{2 \max\{1, \mu^{-1}\}^2 \mu^2 (F_\mu(\hat{y}^{(0)}) - F^* + \sum_{k=0}^\infty \frac{2C}{k^{1+\theta}})}{\alpha \varepsilon^2} \right\rceil = O(\frac{1}{\varepsilon^2})$$

iterations, $\hat{\xi}^{(\bar{k})}$ is an ε -stationary point in the sense of (3), and Algorithm 1 estimate $\hat{\xi}^{(\bar{k})}$ is $L_{\mu F} \eta^{(\bar{k})}$ -close to an ε -stationary point in the sense of (3).

Moreover, if F is level-bounded and $\text{dom}(f_i) = \mathbb{R}^p$ for all i , then sequences $\{y_i^{(k)}\}$ stay bounded and for every limit point y^∞ of $y_i^{(k)}$ for all i , we have, $\frac{1}{n} \sum_{i=1}^n x_{\mu g_i}(y^\infty) = \frac{1}{n} \sum_{i=1}^n x_{\mu f_i}(y^\infty)$ and $\frac{1}{n} \sum_{i=1}^n x_{\mu f_i}(y^\infty)$ satisfies the exact stationary condition (2).

Proof. For any $k \in \mathbb{Z}^+$, let

$$\begin{aligned} \hat{\xi}^{(k)} &= \mu^{-1} \left(\frac{1}{n} \sum_{i=1}^n x_{\mu g_i}(\hat{y}^{(k)}) - \frac{1}{n} \sum_{i=1}^n x_{\mu f_i}(\hat{y}^{(k)}) \right) \\ &= \mu^{-1} \left(\hat{y}^{(k)} - \frac{1}{n} \sum_{i=1}^n x_{\mu g_i}(\hat{y}^{(k)}) \right) \\ &\quad - \mu^{-1} \left(\hat{y}^{(k)} - \frac{1}{n} \sum_{i=1}^n x_{\mu f_i}(\hat{y}^{(k)}) \right). \end{aligned} \quad (26)$$

From the optimality of the proximal mapping, $x_{\mu f_i}(\hat{y}^{(k)})$,

$$\begin{aligned} 0 &\in \partial \{f_i(x_{\mu f_i}(\hat{y}^{(k)})) + \frac{1}{\mu} (x_{\mu f_i}(\hat{y}^{(k)}) - \hat{y}^{(k)})\} \\ \implies \mu^{-1}(\hat{y}^{(k)} - x_{\mu f_i}(\hat{y}^{(k)})) &\in \partial f_i(x_{\mu f_i}(\hat{y}^{(k)})). \end{aligned} \quad (27)$$

Thus,

$$\begin{aligned} \frac{1}{\mu}(\hat{y}^{(k)} - \frac{1}{n} \sum_{i=1}^n x_{\mu f_i}(\hat{y}^{(k)})) &\in \partial \frac{1}{n} \sum_{i=1}^n \partial f_i(x_{\mu f_i}(\hat{y}^{(k)})) \\ &\subseteq \partial \frac{1}{n} \sum_{i=1}^n f_i(x_{\mu f_i}(\hat{y}^{(k)})). \end{aligned} \quad (28)$$

Similarly, from the optimality of $x_{\mu g_i}(\hat{y}^{(k)})$

$$\frac{1}{\mu}(\hat{y}^{(k)} - \frac{1}{n} \sum_{i=1}^n x_{\mu g_i}(\hat{y}^{(k)})) \in \partial \frac{1}{n} \sum_{i=1}^n g_i(x_{\mu g_i}(\hat{y}^{(k)})). \quad (29)$$

Therefore, from (26), (28) and (29),

$$\hat{\xi}^{(k)} \in \partial \frac{1}{n} \sum_{i=1}^n f_i(x_{\mu f_i}(\hat{y}^{(k)})) - \partial \frac{1}{n} \sum_{i=1}^n g_i(x_{\mu g_i}(\hat{y}^{(k)})). \quad (30)$$

Let $\xi^{(k)} = \mu^{-1}(\frac{1}{n} \sum_{i=1}^n x_{\mu g_i}(y_i^{(k)}) - \frac{1}{n} \sum_{i=1}^n x_{\mu f_i}(y_i^{(k)}))$. Consider, $\xi^{(k)} - \hat{\xi}^{(k)} = \mu^{-1} \left(\frac{1}{n} \sum_{i=1}^n x_{\mu g_i}(y_i^{(k)}) - \frac{1}{n} \sum_{i=1}^n x_{\mu g_i}(\hat{y}^{(k)}) \right) - \mu^{-1} \left(\frac{1}{n} \sum_{i=1}^n x_{\mu f_i}(y_i^{(k)}) - \frac{1}{n} \sum_{i=1}^n x_{\mu f_i}(\hat{y}^{(k)}) \right)$. Therefore,

$$\begin{aligned} \|\xi^{(k)} - \hat{\xi}^{(k)}\| &\leq \frac{\mu^{-1}}{n} \sum_{i=1}^n \|x_{\mu g_i}(y_i^{(k)}) - x_{\mu g_i}(\hat{y}^{(k)})\| \\ &\quad + \frac{\mu^{-1}}{n} \sum_{i=1}^n \|x_{\mu f_i}(y_i^{(k)}) - x_{\mu f_i}(\hat{y}^{(k)})\| \\ &\leq \frac{\mu^{-1}}{n} \left(\sum_{i=1}^n \|y_i^{(k)} - \hat{y}^{(k)}\| \right. \\ &\quad \left. + \frac{1}{(1-\mu m_f)} \sum_{i=1}^n \|y_i^{(k)} - \hat{y}^{(k)}\| \right) \\ &\leq \frac{L_{\mu F}}{n} \sum_{i=1}^n \eta^{(k)} = L_{\mu F} \eta^{(k)}, \end{aligned} \quad (31)$$

where we used Lipschitz continuity properties of the proximal mappings listed in Lemma 1.

Since, $F_{i,\mu}$ is $L_{\mu F}$ -Lipschitz differentiable. It can be seen that $F_\mu = \frac{1}{n} \sum_{i=1}^n F_{i,\mu}$ is Lipschitz differentiable with constant $\frac{1}{n} \sum_{i=1}^n L_{\mu F} = L_{\mu F}$. With $\alpha \leq 1/L_{\mu F}$ and (23),

$$\begin{aligned} F_\mu(\hat{y}^{(k+1)}) &\leq F_\mu(\hat{y}^{(k)}) + \langle \hat{h}^{(k)}, \hat{y}^{(k+1)} - \hat{y}^{(k)} \rangle + \frac{L_{\mu F}}{2} \|\hat{y}^{(k+1)} - \hat{y}^{(k)}\|^2 \\ &= F_\mu(\hat{y}^{(k)}) - \frac{1}{\alpha} \langle \hat{y}^{(k+1)} - \hat{y}^{(k)} - e^{(k)}, \hat{y}^{(k+1)} - \hat{y}^{(k)} \rangle \\ &\quad + \frac{L_{\mu F}}{2} \|\hat{y}^{(k+1)} - \hat{y}^{(k)}\|^2 \\ &= F_\mu(\hat{y}^{(k)}) + \left(\frac{L_{\mu F}}{2} - \frac{1}{\alpha} \right) \|\hat{y}^{(k+1)} - \hat{y}^{(k)}\|^2 \\ &\quad + \frac{1}{\alpha} \langle e^{(k)}, \hat{y}^{(k+1)} - \hat{y}^{(k)} \rangle. \end{aligned}$$

This implies,

$$\begin{aligned} \left(\frac{1}{\alpha} - \frac{L_{\mu F}}{2} \right) \|\hat{y}^{(k+1)} - \hat{y}^{(k)}\|^2 &\leq F_\mu(\hat{y}^{(k)}) - F_\mu(\hat{y}^{(k+1)}) + \frac{1}{\alpha} \|e^{(k)}\| \|\hat{y}^{(k+1)} - \hat{y}^{(k)}\| \\ &\leq F_\mu(\hat{y}^{(k)}) - F_\mu(\hat{y}^{(k+1)}) + \frac{1}{\alpha} \|e^{(k)}\| \|e^{(k)} - \alpha \hat{h}^{(k)}\| \\ &\leq F_\mu(\hat{y}^{(k)}) - F_\mu(\hat{y}^{(k+1)}) + \frac{1}{\alpha} \|e^{(k)}\|^2 + \|e^{(k)}\| \|\hat{h}^{(k)}\| \\ &\leq F_\mu(\hat{y}^{(k)}) - F_\mu(\hat{y}^{(k+1)}) + C_1 (\eta^{(k)})^2 + C_2 \eta^{(k)}, \\ &\leq F_\mu(\hat{y}^{(k)}) - F_\mu(\hat{y}^{(k+1)}) + C \eta^{(k)} \end{aligned}$$

where $C_1 := \frac{1}{\alpha} (2\alpha L_{\mu F} + 1)^2$, $C_2 := \sup_k \|\hat{h}^{(k)}\| (2\alpha L_{\mu F} + 1)$, $C := C_1 + C_2$, and we used (23)-(25). Therefore,

$$\begin{aligned} F_\mu(\hat{y}^{(k)}) - F_\mu(\hat{y}^{(k+1)}) + C \eta^{(k)} &\geq \left(\frac{1}{\alpha} - \frac{L_{\mu F}}{2} \right) \|\hat{y}^{(k+1)} - \hat{y}^{(k)}\|^2 \\ &\geq \frac{1}{2\alpha} \|\hat{y}^{(k+1)} - \hat{y}^{(k)}\|^2 \geq \frac{\alpha}{2} \|\hat{h}^{(k)}\|^2 - \frac{1}{2} C_1 (\eta^{(k)})^2. \end{aligned}$$

Thus,

$$\begin{aligned} F_\mu(\hat{y}^{(k)}) - F_\mu(\hat{y}^{(k+1)}) + \frac{3}{2} C \eta^{(k)} &\geq \frac{\alpha}{2} \|\hat{h}^{(k)}\|^2 \\ &= \frac{\alpha}{2\mu^2} \left\| \frac{1}{n} \sum_{i=1}^n x_{\mu g_i}(\hat{y}^{(k)}) - \frac{1}{n} \sum_{i=1}^n x_{\mu f_i}(\hat{y}^{(k)}) \right\|^2. \end{aligned}$$

Summing the above inequality over $k = 0, 1, \dots, K-1$, for some positive integer $K-1$, gives,

$$\begin{aligned} & \sum_{k=0}^{K-1} \left\| \frac{1}{n} \sum_{i=1}^n x_{\mu g_i}(\hat{y}^{(k)}) - \frac{1}{n} \sum_{i=1}^n x_{\mu f_i}(\hat{y}^{(k)}) \right\|^2 \\ & \leq \frac{2\mu^2}{\alpha} \left(F_\mu(\hat{y}^{(0)}) - F_\mu(\hat{y}^{(K)}) + \frac{3C}{2} \sum_{k=0}^{K-1} \eta^{(k)} \right) \\ & \leq \frac{2\mu^2}{\alpha} \left(F_\mu(\hat{y}^{(0)}) - F^* + \frac{3C}{2} \sum_{k=0}^{K-1} \eta^{(k)} \right) \\ & \leq \frac{2\mu^2}{\alpha} \left(F_\mu(\hat{y}^{(0)}) - F^* + \sum_{k=0}^{\infty} \frac{2C}{k^{1+\theta}} \right). \end{aligned} \quad (32)$$

Let $\bar{k} := \arg \min_{k=0,1,\dots,K} \left\| \frac{1}{n} \sum_{i=1}^n x_{\mu g_i}(\hat{y}^{(k)}) - \frac{1}{n} \sum_{i=1}^n x_{\mu f_i}(\hat{y}^{(k)}) \right\|^2$. Therefore, from (32),

$$\begin{aligned} & \left\| \frac{1}{n} \sum_{i=1}^n x_{\mu g_i}(\hat{y}^{(\bar{k})}) - \frac{1}{n} \sum_{i=1}^n x_{\mu f_i}(\hat{y}^{(\bar{k})}) \right\| \\ & \leq \left(\frac{2\mu^2 (F_\mu(\hat{y}^{(0)}) - F^* + \sum_{k=0}^{\infty} \frac{2C}{k^{1+\theta}})}{\alpha K} \right)^{1/2}. \end{aligned} \quad (33)$$

Thus, from (26), (30), and (33) with the claimed upper on bound K in Theorem 3 we get,

$$\begin{aligned} & \max\{\|\hat{\xi}^{(\bar{k})}\|, \left\| \frac{1}{n} \sum_{i=1}^n x_{\mu g_i}(\hat{y}^{(\bar{k})}) - \frac{1}{n} \sum_{i=1}^n x_{\mu f_i}(\hat{y}^{(\bar{k})}) \right\|\} \\ & \leq \max\{1, \mu^{-1}\} \left(\frac{2\mu^2 (F_\mu(\hat{y}^{(0)}) - F^* + \sum_{k=0}^{\infty} \frac{2C}{k^{1+\theta}})}{\alpha K} \right)^{1/2} \leq \varepsilon. \end{aligned}$$

Since F is level-bounded and $\text{dom}(f_i) = \mathbb{R}^p$ for all i , then from [31], Proposition 5, F_μ is level-bounded. As $\{F_\mu(\hat{y}^{(k)})\}_{k \in \mathbb{N}}$ is monotonically non-increasing, the sequence, $\{\hat{y}^{(k)}\}_{k \in \mathbb{N}}$ is bounded and therefore, has at least one limit point $\hat{y}^\infty$. Let $\{\hat{y}^{(k_j)}\}_{j \in \mathbb{N}}$ denote the subsequence converging to $\hat{y}^\infty$. Since, $x_{\mu f_i}$ and $x_{\mu g_i}$ are continuous, (32) implies $\frac{1}{n} \sum_{i=1}^n x_{\mu f_i}(\hat{y}^\infty) = \frac{1}{n} \sum_{i=1}^n x_{\mu g_i}(\hat{y}^\infty)$. As g_i are continuous, $\lim_{j \rightarrow \infty} \frac{1}{n} \sum_{i=1}^n g_i(x_{\mu g_i}(\hat{y}^{(k_j)})) = \frac{1}{n} \sum_{i=1}^n g_i(x_{\mu g_i}(\hat{y}^\infty))$; in addition,

$$\begin{aligned} \frac{1}{n} \sum_{i=1}^n f_i(x_{\mu f_i}(\hat{y}^\infty)) & \leq \liminf_{j \rightarrow \infty} \frac{1}{n} \sum_{i=1}^n f_i(x_{\mu f_i}(\hat{y}^{(k_j)})) \\ & \leq \limsup_{j \rightarrow \infty} \frac{1}{n} \sum_{i=1}^n f_i(x_{\mu f_i}(\hat{y}^{(k_j)})) \\ & \leq \lim_{j \rightarrow \infty} \frac{1}{n} \sum_{i=1}^n f_i(x_{\mu f_i}(\hat{y}^\infty)) \\ & \quad + \frac{1}{2\mu n} \sum_{i=1}^n \|x_{\mu f_i}(\hat{y}^\infty) - \hat{y}^{(k_j)}\|^2 \\ & \quad - \frac{1}{2\mu n} \sum_{i=1}^n \|x_{\mu f_i}(\hat{y}^{(k_j)}) - \hat{y}^{(k_j)}\|^2 \\ & = \frac{1}{n} \sum_{i=1}^n f_i(x_{\mu f_i}(\hat{y}^\infty)), \end{aligned}$$

where the first inequality is due to the lower-semicontinuity of $\frac{1}{n} \sum_{i=1}^n f_i$ and the last inequality is due to the optimality of $x_{\mu f_i}(\hat{y}^\infty)$ in each Moreau envelope evaluation, and therefore we also have $\lim_{j \rightarrow \infty} \frac{1}{n} \sum_{i=1}^n f_i(x_{\mu f_i}(\hat{y}^{(k_j)})) = \lim_{j \rightarrow \infty} \frac{1}{n} \sum_{i=1}^n f_i(x_{\mu f_i}(\hat{y}^\infty))$. Due to η -consensus there exists y^∞ such that $\lim_{j \rightarrow \infty} y_i^{(k_j)} = y^\infty$ for all $i = 1, \dots, n$. Further, for all i , $\|x_{\mu f_i}(y_i^{(k_j)}) - x_{\mu f_i}(\hat{y}^{(k_j)})\| \leq \frac{1}{1-\mu m_f} \|y_i^{(k_j)} - \hat{y}^{(k_j)}\| \leq \frac{1}{1-\mu m_f} \eta^{(k_j)}$. Thus, $x_{\mu f_i}(y_i^{(k_j)}) \rightarrow x_{\mu f_i}(\hat{y}^\infty)$ for all $i = 1, 2, \dots, n$. Since, g_i are continuous, $\lim_{j \rightarrow \infty} g_i(x_{\mu g_i}(y_i^{(k_j)})) = g_i(x_{\mu g_i}(\hat{y}^\infty))$ for all $i = 1, 2, \dots, n$. Thus, $\lim_{j \rightarrow \infty} \frac{1}{n} \sum_{i=1}^n g_i(x_{\mu g_i}(y_i^{(k_j)})) = \frac{1}{n} \sum_{i=1}^n g_i(x_{\mu g_i}(\hat{y}^\infty))$. Finally, taking the limit on (30) and (31) along the appropriate subsequences gives ξ satisfy (2) in the limit. $\square$

IV. NUMERICAL SIMULATIONS

This section presents a simulation study for the proposed DDC-Consensus algorithm. We consider a network of 10 agents with the underlying communication network generated using the Erdos-Renyi model [38] with a connectivity probability of 0.2. We consider a ℓ_{1-2} regularized distributed least squares problem [28]:

$$\underset{x \in \mathbb{R}^p}{\text{minimize}} \quad \frac{1}{10} \sum_{i=1}^{10} \frac{1}{2} \|A_i x - b_i\|^2 + \rho \|x\|_1 - \rho \|x\|_2.$$

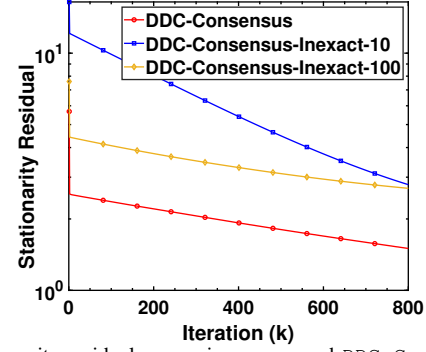


Fig. 1: Stationarity residual comparison: proposed DDC-Consensus and DDC-Consensus-Inexact- q

The problem data is generated as follows: first, we create the data matrices $A_i \in \mathbb{R}^{m \times p}$, for all i , with entries generated from the standard Gaussian distribution. Then the columns of the matrices are normalized to the unit norm. We generate $x^* \in \mathbb{R}^p$ such that x^* is a sparse vector with s non-zero entries. The non-zero entries of x^* are drawn from the standard Gaussian distribution. Finally, the vectors $b_i \in \mathbb{R}^m$ are determined as $b_i = A_i x^* + 0.01 \zeta_i$, where, $\zeta_i \in \mathbb{R}^m$ has standard Gaussian entries. $F^* := \frac{1}{10} \sum_{i=1}^{10} \frac{1}{2} \|A_i x^* - b_i\|^2 + \rho \|x^*\|_1 - \rho \|x^*\|_2$. The parameters used during the simulations are: $m = 720, p = 2560, \rho = 0.1, \alpha = 0.01, \mu = 1/\lambda_{\max}(A_i^\top A_i)$, where $\lambda_{\max}(M)$ is the maximum eigenvalue of matrix M . We choose $\eta^{(k)} = 1/k^{1.1}$ as the tolerance sequence in Algorithm 1.

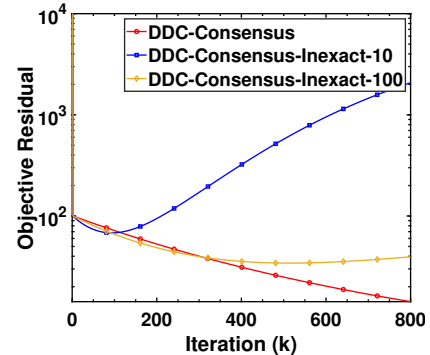


Fig. 2: Objective residual comparison: proposed DDC-Consensus and DDC-Consensus-Inexact- q

We compare the performance of the proposed algorithm DDC-Consensus with its following two relaxations: (i) DDC-Consensus-Inexact- q , where we solve the minimization problems (4) and (5) inexactly via a gradient descent method for q iterations, (ii) DDC-Mixing, where the η -consensus is replaced with one step mixing (applying one iteration of the updates (15)-(17) to y_i estimates of the optimization algorithms). To demonstrate the performance of the proposed algorithm we present utilize the following resid-

ual metrics: *Solution Residual*(k): $\frac{1}{10} \sum_{i=1}^{10} \|y_i^{(k)} - x^*\|_2$, *Stationarity Residual*(k): $\frac{1}{10} \sum_{i=1}^{10} \|x_{\mu g_i}(y_i^{(k)}) - x_{\mu f_i}(y_i^{(k)})\|_2$, *Objective Residual*(k): $F(y_i^{(k)}) - F^*$, *Consensus Residual*(k): $\frac{1}{10} \sum_{i=1}^n \sum_{j=1}^n \|y_i^{(k)} - y_j^{(k)}\|_2$.

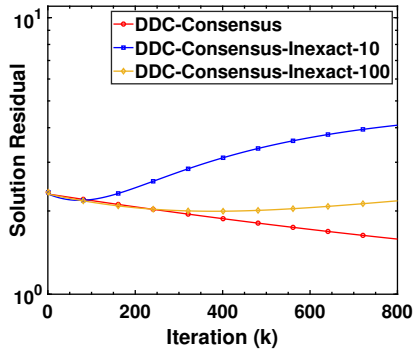


Fig. 3: Solution residual comparison: proposed DDC-Consensus and DDC-Consensus-Inexact- q

A. Comparison with DDC-Consensus-Inexact- q

Figs. 1-3 demonstrate the performance comparisons between the proposed DDC-Consensus algorithm and two instances of DDC-Consensus-Inexact- q that are denoted as DDC-Consensus-Inexact-10 and DDC-Consensus-Inexact-100 respectively. This comparison aims to determine a trade-off between the accuracy and computational burden of the proximal minimization in the proposed DDC-Consensus algorithm. As expected the DDC-Consensus algorithm performs the best with respect to all the performance metrics. Fig. 1 shows the stationarity metric that gives an estimate of the rate of convergence in the sense of Theorem 3.

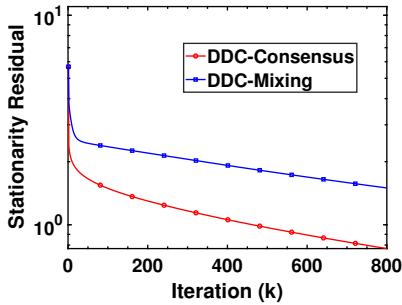


Fig. 4: Stationarity residual: DDC-Consensus and DDC-Mixing

Figs. 2 and 3 show that gradient descent with a low computational footprint can perform relatively well compared to the exact minimization in the DDC-Consensus algorithm. However, it is also established that if the proximal minimization is not performed with good accuracy the algorithm iterates to diverge as in the case of DDC-Consensus-Inexact-10.

B. Comparison with DDC-Mixing

Figs. 4-7 present comparison plots showing the different metrics for both the DDC-Consensus and DDC-Mixing algorithms. The simulation study in this section aims to study the effect of the agreement quality between the agent

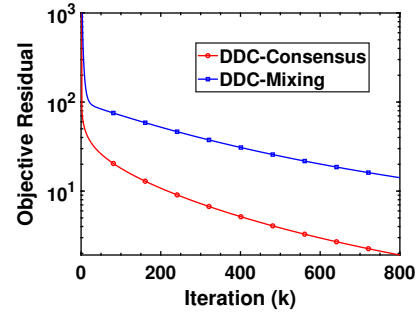


Fig. 5: Objective residual: DDC-Consensus and DDC-Mixing

estimates in the DDC-Consensus algorithm on its performance. The DDC-Mixing algorithm uses *only* one information exchange step among the agents and is theoretically the lower bound on information aggregation. Note, that both DDC-Consensus and DDC-Mixing algorithms perform exact proximal minimization steps.

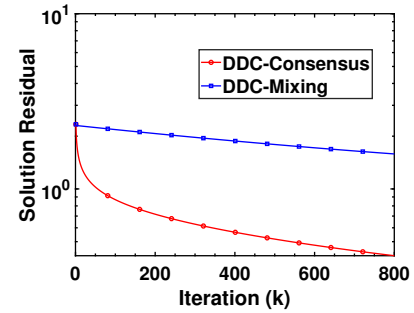


Fig. 6: Solution residual comparison: proposed DDC-Consensus and DDC-Mixing

Figs. 4-6 show that the η -consensus protocol allows for a higher degree of agreement between the agent estimates during the algorithm iterations lead to better convergence properties of the proposed DDC-Consensus algorithm compared to the DDC-Mixing algorithm. Furthermore, the η -consensus protocol adaptively controls the agent mismatch and leads to the consensus constraint satisfaction at every iterate as seen in Fig. 7.

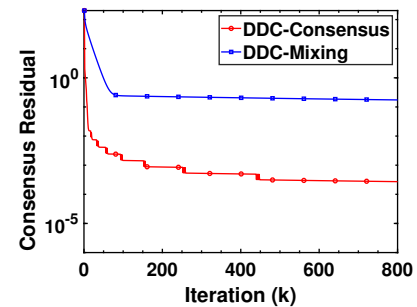


Fig. 7: Consensus residual: DDC-Consensus and DDC-Mixing

V. CONCLUSION

We focused on a distributed DC optimization problem with local objective functions given by the difference between a nonsmooth weakly convex function and a nonsmooth convex function. Based on the gradient of the smooth approximations of the weakly convex and convex functions we developed the distributed DDC-Consensus algorithm. In

the DDC-Consensus algorithm each agent performs a local gradient descent step and updates its estimates using a finite-time η -consensus protocol. The agent estimates are shared over a directed graph and updated via a column stochastic matrix; giving DDC-Consensus the desirable properties of allowing for non-symmetric communication and distributed synthesis. We established that the DDC-Consensus algorithm converges to a stationary point of the nonconvex distributed optimization problem. The numerical performance of the DDC-Consensus algorithm is presented in a simulation study solving the DC-regularized least squares problem. Results showing the behavior of DDC-Consensus and its relaxations DDC-Consensus-Inexact- q and DDC-Mixing concerning the residual metrics demonstrate the efficacy of the proposed DDC-Consensus algorithm.

REFERENCES

- [1] V. Khatana and M. V. Salapaka, "D-distadmm: A $\mathcal{O}(1/k)$ distributed admm for distributed optimization in directed graph topologies," in *2020 59th IEEE Conference on Decision and Control (CDC)*, 2020, pp. 2992–2997.
- [2] V. Khatana, G. Saraswat, S. Patel, and M. V. Salapaka, "Gradient-consensus method for distributed optimization in directed multi-agent networks," in *2020 American Control Conference (ACC)*, 2020, pp. 4689–4694.
- [3] P. Hartman, "On functions representable as a difference of convex functions," *Pacific Journal of Mathematics*, vol. 9, no. 3, pp. 707–713, 1959.
- [4] M. Nouiehed, J.-S. Pang, and M. Razaviyayn, "On the pervasiveness of difference-convexity in optimization and statistics," *Mathematical Programming*, vol. 174, no. 1, pp. 195–222, 2019.
- [5] S. Merkli, A. Domahidi, J. L. Jerez, M. Morari, and R. S. Smith, "Fast ac power flow optimization using difference of convex functions programming," *IEEE Transactions on Power Systems*, vol. 33, no. 1, pp. 363–372, 2017.
- [6] H. An and L. Thi, "Solving large scale molecular distance geometry problems by a smoothing technique via the gaussian transform and dc programming," *Journal of Global Optimization*, vol. 27, no. 4, pp. 375–397, 2003.
- [7] T. H. L. An and P. D. Tao, "Dc optimization approaches via markov models for restoration of signal (1-d) and (2-d)," in *Advances in Convex Analysis and Global Optimization*. Springer, 2001, pp. 303–317.
- [8] L. T. Hoai An and P. D. Tao, "Dc programming approach for multicommodity network optimization problems with step increasing cost functions," *Journal of Global Optimization*, vol. 22, no. 1, pp. 205–232, 2002.
- [9] H. A. Le Thi and T. P. Dinh, "A continuous approach for globally solving linearly constrained quadratic," *Optimization*, vol. 50, no. 1-2, pp. 93–120, 2001.
- [10] J. N. Tsitsiklis, "Problems in decentralized decision making and computation." Massachusetts Inst of Tech Cambridge Lab for Information and Decision Systems, Tech. Rep., 1984.
- [11] D. P. Bertsekas and J. N. Tsitsiklis, *Parallel and distributed computation: numerical methods*. Prentice hall Englewood Cliffs, NJ, 1989, vol. 23.
- [12] S. Pu, W. Shi, J. Xu, and A. Nedic, "Push-pull gradient methods for distributed optimization in networks," *IEEE Transactions on Automatic Control*, 2020.
- [13] V. Khatana and M. V. Salapaka, "Dc-distadmm: Admm algorithm for constrained optimization over directed graphs," *IEEE Transactions on Automatic Control*, vol. 68, no. 9, pp. 5365–5380, 2023.
- [14] V. Khatana, G. Saraswat, S. Patel, and M. V. Salapaka, "Gradconsensus: Linearly convergent algorithm for reducing disagreement in multi-agent optimization," *IEEE Transactions on Network Science and Engineering*, vol. 11, no. 1, pp. 1251–1264, 2024.
- [15] J. Zeng and W. Yin, "On nonconvex decentralized gradient descent," *IEEE Transactions on Signal Processing*, vol. 66, no. 11, pp. 2834–2848, 2018.
- [16] R. Xin, U. A. Khan, and S. Kar, "A fast randomized incremental gradient method for decentralized nonconvex optimization," *IEEE Transactions on Automatic Control*, vol. 67, pp. 5150–5165, 2020. [Online]. Available: <https://api.semanticscholar.org/CorpusID:226282498>
- [17] Y. Yan, C. Niu, Y. Ding, Z. Zheng, F. Wu, G. Chen, S. Tang, and Z. Wu, "Distributed non-convex optimization with sublinear speedup under intermittent client availability," *INFORMS J. Comput.*, vol. 36, pp. 185–202, 2020. [Online]. Available: <https://api.semanticscholar.org/CorpusID:211146497>
- [18] G. Scutari and Y. Sun, "Distributed nonconvex constrained optimization over time-varying digraphs," *Mathematical Programming*, vol. 176, pp. 497–544, 2019.
- [19] E. Gorbunov, K. P. Burlachenko, Z. Li, and P. Richtárik, "Marina: Faster non-convex distributed learning with compression," in *International Conference on Machine Learning*. PMLR, 2021, pp. 3788–3798.
- [20] X. Jiang, X. Zeng, J. Sun, and J. Chen, "Distributed proximal gradient algorithm for nonconvex optimization over time-varying networks," *IEEE Transactions on Control of Network Systems*, vol. 10, no. 2, pp. 1005–1017, 2023.
- [21] P. D. Lorenzo and G. Scutari, "Next: In-network nonconvex optimization," *IEEE Transactions on Signal and Information Processing over Networks*, vol. 2, pp. 120–136, 2016. [Online]. Available: <https://api.semanticscholar.org/CorpusID:3261082>
- [22] S. Chen, A. García, and S. Shahrampour, "On distributed nonconvex optimization: Projected subgradient method for weakly convex problems in networks," *IEEE Transactions on Automatic Control*, vol. 67, pp. 662–675, 2020. [Online]. Available: <https://api.semanticscholar.org/CorpusID:234356225>
- [23] A. Alvarado, G. Scutari, and J.-S. Pang, "A new decomposition method for multiuser dc-programming and its applications," *IEEE Transactions on Signal Processing*, vol. 62, no. 11, pp. 2984–2998, 2014.
- [24] H. Le Thi, *DC Programming and DCA*. (Homepage), 2005. [Online]. Available: <http://www.lita.univ-lorraine.fr/~lethi/index.php/en/research/dc-programming-and-dca.html>
- [25] H. A. Le Thi, T. P. Dinh, H. M. Le, and X. T. Vo, "Dc approximation approaches for sparse optimization," *European Journal of Operational Research*, vol. 244, no. 1, pp. 26–46, 2015.
- [26] P. D. Tao and L. H. An, "Convex analysis approach to dc programming: theory, algorithms and applications," *Acta mathematica vietnamica*, vol. 22, no. 1, pp. 289–355, 1997.
- [27] L. T. H. An and P. D. Tao, "The dc (difference of convex functions) programming and dca revisited with dc models of real world nonconvex optimization problems," *Annals of operations research*, vol. 133, pp. 23–46, 2005.
- [28] M. Ahn, J.-S. Pang, and J. Xin, "Difference-of-convex learning: directional stationarity, optimality, and sparsity," *SIAM Journal on Optimization*, vol. 27, no. 3, pp. 1637–1665, 2017.
- [29] P. D. Nhat, H. M. Le, and H. A. Le Thi, "Accelerated difference of convex functions algorithm and its application to sparse binary logistic regression," in *IJCAI*, 2018, pp. 1369–1375.
- [30] T. Sun, P. Yin, L. Cheng, and H. Jiang, "Alternating direction method of multipliers with difference of convex functions," *Advances in Computational Mathematics*, vol. 44, pp. 723–744, 2018.
- [31] K. Sun and X. A. Sun, "Algorithms for difference-of-convex programs based on difference-of-moreau-envelopes smoothing," *INFORMS Journal on Optimization*, vol. 5, no. 4, pp. 321–339, 2023.
- [32] B. Wen, X. Chen, and T. K. Pong, "A proximal difference-of-convex algorithm with extrapolation," *Computational optimization and applications*, vol. 69, pp. 297–324, 2018.
- [33] R. T. Rockafellar and R. J.-B. Wets, *Variational analysis*. Springer Science & Business Media, 2009, vol. 317.
- [34] J. Melbourne, G. Saraswat, V. Khatana, S. Patel, and M. V. Salapaka, "On the geometry of consensus algorithms with application to distributed termination in higher dimension," in the *proceedings of International Federation of Automatic Control (IFAC)*, 2020. [Online]. Available: http://box5779.temp.domains/~jamesmel/wp-content/uploads/2019/11/Vector_Consensus_IFAC_-1.pdf
- [35] —, "Convex decreasing algorithms: Distributed synthesis and finite-time termination in higher dimension," *IEEE Transactions on Automatic Control*, vol. 69, no. 6, pp. 3960–3967, 2024.
- [36] 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*. IEEE, 2003, pp. 482–491.
- [37] G. Saraswat, V. Khatana, S. Patel, and M. V. Salapaka, "Distributed finite-time termination for consensus algorithm in switching topologies," *IEEE Transactions on Network Science and Engineering*, vol. 10, no. 1, pp. 489–499, 2023.
- [38] P. Erdős and A. Rényi, "On the evolution of random graphs," *Publ. Math. Inst. Hung. Acad. Sci.*, vol. 5, no. 1, pp. 17–60, 1960.