Harnessing Smoothness to Accelerate Distributed Optimization

Guannan Qu, Na Li

Abstract

There has been a growing effort in studying the distributed optimization problem over a network. The objective is to optimize a global function formed by a sum of local functions, using only local computation and communication. Literature has developed consensus-based distributed (sub)gradient descent (DGD) methods and has shown that they have the same convergence rate $O(\frac{1}{\sqrt{t}})$ as the centralized (sub)gradient methods (CGD), when the function is convex but possibly nonsmooth. However, when the function is convex and smooth, under the framework of DGD, it is unclear how to harness the smoothness to obtain a faster convergence rate comparable to CGD's convergence rate. In this paper, we propose a distributed algorithm that, despite using the same amount of communication per iteration as DGD, can effectively harnesses the function smoothness and converge to the optimum with a rate of $O(\frac{1}{t})$. If the objective function is further strongly convex, our algorithm has a linear convergence rate. Both rates match the convergence rate of CGD. The key step in our algorithm is a novel gradient estimation scheme that uses history information to achieve fast and accurate estimation of the average gradient. To motivate the necessity of history information, we also show that it is impossible for a class of distributed algorithms like DGD to achieve a linear convergence rate without using history information even if the objective function is strongly convex and smooth.

I. Introduction

Given a set of agents $\mathcal{N} = \{1, 2, \ldots, n\}$, each of which has a local convex cost function $f_i(x) : \mathbb{R}^N \to \mathbb{R}$, the objective of distributed optimization is to find x that minimizes the average of all the functions,

$$\min_{x \in \mathbb{R}^N} f(x) \triangleq \frac{1}{n} \sum_{i=1}^{n} f_i(x)$$

using local communication and local computation. The local communication is defined through an undirected communication graph $\mathcal{G} = (V, E)$, where the nodes $V = \mathcal{N}$ and edges $E \subset V \times V$. Agent i and j can send information to each other if and only if i and j are connected in graph $\mathcal{G}$. The local computation means that each agent i can only make their decision based on their local function f_i and the information obtained from their neighbors.

This problem has recently received much attention and has found various applications in multi-agent control, distributed state estimation over sensor networks, large scale computation in machine/statistical learning, etc. As a concrete example, in the setting of distributed statistical learning, x is the parameter to infer, and f_i is the empirical

Guannan Qu and Na Li are affiliated with John A. Paulson School of Engineering and Applied Sciences at Harvard University. Email: gqu@g.harvard.edu, nali@seas.harvard.edu

loss function of the local dataset of agent i. Then minimizing f means empirical loss minimization that uses datasets of all the agents.

The early work of this problem can be found in [1], [2]. Recently, [3] (see also [4]) proposes a consensus-based distributed (sub)gradient descent (DGD) method where each agent performs a consensus step and then a descent step along the local (sub)gradient direction.[1] [3] shows that, when using a diminishing step size, the method converges with a rate of $O(\frac{1}{\sqrt{t}})$ for convex Lipshitz and possibly nonsmooth objective functions. This matches the convergence rate of centralized subgradient descent algorithm. [5] applies a similar idea to develop a distributed dual averaging algorithm which converges to the optimal solution with a similar rate but (almost) independent of network size. Recently, this line of work has been extended to distributed optimization under various realistic conditions, such as stochastic subgradient errors [6], directed or random communication graph [7]–[9], linear scaling in network size [10], heterogeneous local constraints [11], [12].

All the above work features a diminishing step size and is designated for possibly nonsmooth functions. Though it can also be applied to smooth functions, it does not fully exploit the function smoothness and has a slower convergence rate compared with the normal Centralized Gradient Descent (CGD) method. For example, [13] shows that, when applied to smooth functions, DGD can not be faster than $\Omega(\frac{1}{t^{2/3}})$, slower than CGD's $O(\frac{1}{t})$. One way to speed up DGD is to use a fixed step size. However, as shown in [9], [14], using a fixed step size will make DGD only converge to a neighborhood of the optimizer. In fact, we prove in this paper that for strongly convex and smooth functions, it is impossible for DGD-like distributed algorithms to achieve the same linear convergence rate as CGD (Theorem 4).

Another way to achieve a faster convergence rate [13], [15] is to perform multiple consensus steps after each gradient evaluation. This method is able to converge to the optimal solution with fixed step size and achieves a fast convergence rate in terms of the number of gradient evaluation steps. However, it places too much communication burden: the further the algorithm proceeds, the larger number of consensus steps per gradient evaluation is required. This drawback poses the need for alternative distributed algorithms that effectively harness the smoothness to achieve faster convergence, using only *one* communication step per gradient evaluation iteration.

In this paper, we propose a distributed algorithm that can effectively harness the smoothness, and achieve a convergence rate that matches CGD, using only one communication step per gradient evaluation. Specifically, our algorithm achieves a $O(\frac{1}{t})$ rate for smooth convex functions (Theorem 3), and a linear convergence rate ($O(\gamma^t)$ for some $\gamma \in (0,1)$) for smooth and strongly convex functions (Theorem 1).[2] The convergence rates match the convergence rates of CGD, but with worse constants due to the effect of local communication and computation. Our algorithm is a combination of gradient descent and a novel gradient estimation scheme that utilizes history information to achieve fast and accurate estimation of the average gradient. To show the necessity of history

[1] In this paper, by local (sub)gradients we mean the (sub)gradients of each f_i, and by average (sub)gradient we mean the average of the local (sub)gradients.

[2] A recent paper [16] also achieves similar convergence rate results using a different algorithm. However, to the best of the author's knowledge, our algorithm is the first to theoretically achieve the $O(\frac{1}{t})$ convergence rate for convex smooth functions in terms of objective error. [16] achieves a $O(\frac{1}{t})$ rate in terms of the first order residual instead. A detailed comparison between our algorithm and [16] will be given in Section III-C.

information, we also prove that it is impossible for a class of distributed algorithms like DGD to achieve a linear convergence rate without using history information even if we restrict the class of objective functions to be strongly convex and smooth (Theorem 4).

Lastly, we note that our way of harnessing smoothness is different from Nesterov accelerated gradient descent [17]. We focus on how to decentralize CGD (the normal centralized gradient descent). We expect our idea can be extended to Nesterov gradient descent method with the potential of achieving convergence rates similar to Nesterov gradient descent.[3]

The rest of the paper is organized as follows. In Section II, we will formally define the problem and present our algorithm and results. In Section III, we will give a review of previous methods, introduce an impossibility result and the motivation of our approach. In Section IV we will prove our convergence results. Due to space limit, for parts of the proof we will only give the proof idea, and the complete proof can be found in the full version of this paper [18]. Section V and VI will show simulation results and conclude the paper.

Notations. Throughout the rest of the paper, n is the number of agents, and N is the dimension of the domain of the f_i's. $i, j \in \{1, 2, \ldots, n\}$ are indices for agents, while $t, k, \ell \in \mathbb{N}$ are indices for iteration steps. We use x^* and f^* to denote the minimizer and the minimal value of f, respectively. If f has multiple minimizers, x^* can be any of them. $\|\cdot\|$ denotes 2-norm for vectors, and Frobenius norm for matrices. $\langle \cdot, \cdot \rangle$ denotes inner product for vectors. $\rho(\cdot)$ denotes spectral radius for square matrices, and $\mathbb{1}$ denotes a n-dimensional all one column vector. All vectors, when having dimension N (the dimension of the domain of the f_i's), will all be regarded as row vectors. As a special case, all gradients, $\nabla f_i(x)$ and $\nabla f(x)$ are interpreted as N-dimensional row vectors. '$\leq$', when applied to vectors of the same dimension, denotes element wise 'less than or equal to'.

II. Problem and Algorithm

A. Problem Formulation

Consider n agents, $\mathcal{N} = \{1, 2, \ldots, n\}$, each of which has a convex function $f_i : \mathbb{R}^N \to \mathbb{R}$. The objective of distributed optimization is to find x to minimize the average of all the functions, i.e.

$$\min_{x \in \mathbb{R}^N} f(x) \triangleq \frac{1}{n} \sum_{i=1}^{n} f_i(x) \tag{1}$$

using local communication and local computation. The local communication is defined through an undirected communication graph $\mathcal{G} = (V, E)$, where the nodes $V = \mathcal{N}$. Agent i and j can send information to each other if and only if i and j are connected in graph $\mathcal{G}$, which is denoted as $(i, j) \in E$. The local computation means that each agent i can only make their decision based on their local function f_i and the information obtained from their neighbors.

Throughout the paper, we assume that the set of minimizers of f is non-empty and compact. We denote x^* as one of the minimizers and f^* as the minimal value. We will study the case where each f_i is convex and β-smooth (Assumption 1) and we will further study the case where each f_i is in addition α-strongly convex (Assumption 2).

[3] [13] extends the DGD idea to decentralize Nesterov accelerated gradient descent. It is shown the convergence rate is $O(\frac{\log t}{t})$, slower than the centralized Nesterov accelerated gradient descent method ($O(\frac{1}{t^2})$).

Assumption 1. *$\forall i$, f_i is convex. In addition, f_i is β-smooth, that is f_i is differentiable and the gradient is β-Lipschitz continuous, i.e., $\forall x, y \in \mathbb{R}^N$,*

$$\|\nabla f_i(x) - \nabla f_i(y)\| \leq \beta \|x - y\|$$

As a direct consequence, f is also β-smooth.

Assumption 2. *$\forall i$, f_i is α-strongly convex, i.e. $\forall x, y \in \mathbb{R}^N$, we have*

$$f_i(y) \geq f_i(x) + \langle \nabla f_i(x), y - x \rangle + \frac{\alpha}{2}\|y - x\|^2$$

As a result, f is also α-strongly convex.

B. Algorithm

The algorithm we will describe is a consensus-based distributed algorithm. Each agent weighs their neighbors' information to compute their local decision. To model the weighting process, we introduce a consensus weight matrix, $W = [w_{ij}] \in \mathbb{R}^{n \times n}$, which satisfies the following properties:[4]

(a) $\forall (i,j) \in E$, $w_{ij} > 0$. $\forall i$, $w_{ii} > 0$. $w_{ij} = 0$ elsewhere.

(b) W is doubly stochastic, i.e. $\sum_{i'} w_{i'j} = \sum_{j'} w_{ij'} = 1$ for all $i, j \in \mathcal{N}$.

As a result [20], $\exists \sigma \in (0,1)$ which depends on the spectrum of W, s.t. $\forall x \in \mathbb{R}^{n \times N}$, let $\bar{x} = \frac{1}{n}\mathbb{1}\mathbb{1}^T x$ (the column wise average of x), we have $\|W^t x - \mathbb{1}\bar{x}\| \leq \sigma^t \|x - \mathbb{1}\bar{x}\|$. This 'averaging' property will be frequently used in the rest of the paper.

In our algorithm, each agent i keeps an estimate of the minimizer $x_i(t) \in \mathbb{R}^{1 \times N}$, and another vector $s_i(t) \in \mathbb{R}^{1 \times N}$ which is designated to estimate the average gradient, $\frac{1}{n}\sum_{i=1}^n \nabla f_i(x_i(t))$. The algorithm starts with an arbitrary $x_i(0)$, and with $s_i(0) = \nabla f_i(x_i(0))$. The algorithm proceeds using the following update,

$$x_i(t+1) = \sum_j w_{ij} x_j(t) - \eta s_i(t) \tag{2a}$$

$$s_i(t+1) = \sum_j w_{ij} s_j(t) + \nabla f_i(x_i(t+1)) - \nabla f_i(x_i(t)) \tag{2b}$$

where $[w_{ij}]_{n \times n}$ are the consensus weights and $\eta > 0$ is a fixed step size. Because $w_{ij} = 0$ when $(i,j) \notin E$, each node i only needs to send $x_i(t)$ and $s_i(t)$ to its neighbors. Therefore, the algorithm can be operated in a fully distributed fashion, with only local communication. Note that the two consensus weight matrices in step (2a) and (2b) can be chosen differently. We use the same matrix W to carry out our analysis for the purpose of easy exposition.

The update equation (2a) is similar to the algorithm in [3] (see also (3) in Section III), except that the step size here is constant, and the subgradient is replaced with $s_i(t)$ which follows the update rule (2b). In Section III and IV-B, we will discuss the motivation and the intuition behind this algorithm.

[4]The selection of the consensus weights is an intensely studied problem, see [19], [20].

C. Convergence of the Algorithm

To state the convergence results, we need to define the following average sequences.

$$\bar{x}(t) = \frac{1}{n}\sum_{i=1}^{n} x_i(t) \in \mathbb{R}^{1\times N}, \bar{s}(t) = \frac{1}{n}\sum_{i=1}^{n} s_i(t) \in \mathbb{R}^{1\times N}$$

$$g(t) = \frac{1}{n}\sum_{i=1}^{n} \nabla f_i(x_i(t)) \in \mathbb{R}^{1\times N}$$

We also define the gradient of f evaluated at $\bar{x}(t)$,

$$h(t) = \nabla f(\bar{x}(t)) \in \mathbb{R}^{1\times N}$$

We summarize our convergence results here.

Theorem 1. *Under the smooth and strongly convex assumptions (Assumption 1 and 2), when η is such that the matrix*

$$G(\eta) = \begin{bmatrix} (\sigma+\beta\eta) & \beta(\eta\beta+2) & \eta\beta^2 \\ \eta & \sigma & 0 \\ 0 & \eta\beta & \lambda \end{bmatrix}$$

$$\textit{where } \lambda = \max(|1-\alpha\eta|, |1-\beta\eta|)$$

has spectral radius $\rho(G(\eta)) < 1$, then $\forall i$, $\|x_i(t) - x^*\|$ (distance to the optimizer), $\|x_i(t) - \bar{x}(t)\|$ (consensus error), and $\|s_i(t) - g(t)\|$ (gradient estimation error) are all decaying with rate $O(\rho(G)^t)$. As a consequence, $f(x_i(t)) - f^*$ (objective error) is $O(\rho(G)^{2t})$.

The following lemma provides a sufficient condition for the step size η to ensure $\rho(G(\eta)) < 1$.

Lemma 2. *When $0 < \eta < \eta_0 \triangleq \frac{2}{\beta}\frac{(1-\sigma)^2}{3+\sqrt{13+4\frac{\beta}{\alpha}}}$, $\rho(G(\eta)) < 1$*

Remark 1. *In experiments we find that $\eta < \frac{1}{3\beta}$ (regardless of α, σ) is usually sufficient for linear convergence. We also find that using a larger step size does not necessarily lead to a faster convergence rate. Also, we note that the convergence rate shown in the theorem appears to be conservative compared to numerical experiments.*

If we drop the strongly convex assumption, we have the following result.

Theorem 3. *Under the smooth assumption (Assumption 1), we have when η is sufficiently small,*

1) *The sequence $\{\|h(t)\|\}_{t=0}^{\infty}$ is square summable, i.e. $\sum_{t=0}^{\infty}\|h(t)\|^2 < \infty$*
2) *$f(\bar{x}(t)) - f^* = O(\frac{1}{t})$, and $\forall i$, $\min_{t'\le t} f(x_i(t')) - f^* = O(\frac{1}{t})$*

Remark 2. *Our algorithm preserves the convergence rate of CGD, in the sense that it has a linear convergence rate when the f_i's are strongly convex and smooth and a convergence rate of $O(\frac{1}{t})$ when the f_i's are just smooth. However, we note that the linear convergence rate constant $\rho(G)$ is usually worse than CGD; and moreover, in both cases, our algorithm has a worse constant in the big O terms. Moreover, compared to CGD, our step size requirement also depends on the consensus matrix W (Lemma 2).*

III. Algorithm development: Motivation

In this section, we will briefly review distributed first-order optimization algorithms that are related to our algorithm and discuss their limitations which will motivate our algorithm development. In particular, we will formally provide an impossibility result regarding the limitations. Lastly we will discuss the literature that motivates the idea of harnessing the smoothness from history information.

A. Review of Distributed First-Order Optimization Algorithms

To solve the distributed optimization problem (1), people have developed consensus-based DGD (Distributed (sub)gradient descent) methods, e.g., [3], [5]–[10], [13]–[16], that combine a consensus algorithm and a first order optimization algorithm. For a review of consensus algorithms and first order optimization algorithms, we refer to references [20] and [17], [21], [22] respectively. For the sake of concrete discussion, we focus on the algorithm in [3], where each agent i keeps an local estimate of the solution to (1), $x_i(t)$ and it updates $x_i(t)$ according to,

$$x_i(t+1) = \sum_j w_{ij} x_j(t) - \eta_t g_i(t) \tag{3}$$

where $g_i(t) \in \partial f_i(x_i(t))$ is a subgradient of f_i at $x_i(t)$ (f_i is possibly nonsmooth), and $\eta_t = \Theta(\frac{1}{\sqrt{t}})$ is the step size, and w_{ij} are some properly chosen consensus weights. (3) is essentially performing a consensus step followed by a standard subgradient descent along the local subgradient direction $g_i(t)$. [3] shows that $f(x_i(t))$ converges to the minimum f^* with rate $O(\frac{1}{\sqrt{t}})$. This is the same rate as the centralized subgradient descent algorithm.

When the f_i's are smooth, the subgradient $g_i(t)$ will equal the gradient $\nabla f_i(x_i(t))$. However, as shown in [13], even in this case the convergence rate of (3) can not be better than $\Omega(\frac{1}{t^{2/3}})$. In contrast, the CGD (centralized gradient descent) method,

$$x(t+1) = x(t) - \eta \nabla f(x) \tag{4}$$

converges to the optimum with rate $O(\frac{1}{t})$ if the stepsize η is a small enough constant. Moreover, when f is further strongly convex, CGD (4) converges to the optimal solution with a linear rate. If a fixed step size η is used in DGD (3), though the algorithm runs faster, the method only converges to a neighborhood of the optimizer [9], [14]. This is because even if $x_i(t) = x^*$ (the optimal solution), $\nabla f_i(x_i(t))$ is not necessarily zero.

To fix this problem of non-convergence, it has been proposed to use multiple consensus steps after each gradient descent [13], [15]. One example is provided as follows:

$$y_i(t,0) = x_i(t) - \eta \nabla f_i(x_i(t)) \tag{5a}$$

$$y_i(t,k) = \sum_j w_{ij} y_j(t,k-1), k = 1,2,\ldots,c_t \tag{5b}$$

$$x_i(t+1) = y_i(t,c_t) \tag{5c}$$

For each gradient descent step (5a), after c_t consensus steps ($c_t = \Theta(\log t)$ in [13], and $c_t = \Theta(t)$ in [15]), the agents' estimates $x_i(t+1)$ are sufficiently averaged, and it is as if each agent has performed a descent along the average gradient $\frac{1}{n}\sum_i \nabla f_i(x_i(t))$. As a result, algorithm (5) addresses the non-convergence problem mentioned above. However, it places too much communication burden on the agents: the further the algorithm proceeds, the

more consensus steps after each gradient descent are required. In addition, even if the algorithm already reaches the optimizer $x_i(t) = x^*$, because of (5a) and because $\nabla f_i(x^*)$ might be non-zero, $y_i(t, 0)$ will deviate from the optimizer, and then a large number of consensus steps in (5b) are needed to average out the deviation. All these drawbacks pose the need for alternative distributed algorithms that effectively harness the smoothness to achieve faster convergence, using only *one* (or a constant number of) communication step(s) per gradient evaluation.

B. An Impossibility Result

To compliment the preceding discussion, here we provide an impossibility result for a class of distributed first-order algorithms which includes algorithms like (3). We use notation $-i$ to denote the set $\mathcal{N}/\{i\}$. The class of algorithms we consider obeys the following updating rule,

$$x_i(t) = A(g(x_i(t-1), x_{-i}(t-1), \mathcal{G}), \eta_t \nabla f_i(x_i(t-1)), \quad \forall i \in \mathcal{N}. \tag{6}$$

Here both g and A denote general functions with the following properties. Function g captures how agents use their neighbors' information, and g is assumed to be a continuous function on the component $x_j(t)$, $j \in \mathcal{N}$. Note that g can be interpreted as the consensus step. A is a function of g and the scaled gradient direction $\eta_t \nabla f_i(x_i(t-1))$, and A is assumed to be L-Lipschitz continuous. Note that A can be interpreted as a first-order update, which encompasses the usual (projected) gradient descent, mirror descent, and proximal algorithms, etc. η_t can be considered as the step size, and we assume it has a limit η^* as $t \to \infty$. We will show that given strongly convex and smooth cost functions, any algorithm belonging to this class will not have a linear convergence rate, which is in contrast to the linear convergence of the centralized methods.

Theorem 4. *Consider the simple case where $\mathcal{N} = \{1, 2\}$, i.e. there are only two agents. Assume the objective functions $f_1, f_2 : \mathbb{R}^N \to \mathbb{R}$ are α-strongly convex and β-smooth. Suppose for any $f_1, f_2, x_1(0), x_2(0)$, $\lim_{t\to\infty} x_i(t) = x^*$ under algorithm (6), where x^* is the minimizer of $f_1 + f_2$. Then there exist $f_1, f_2, x_1(0), x_2(0)$ such that for any $\delta \in (0, 1)$ and $T \geq 0$, there exist $t \geq T$, s.t. $\|x_i(t+1) - x^*\| \geq \delta \|x_i(t) - x^*\|$.*

Proof. We first show $\eta^* = 0$. Assume the contrary holds, $\eta^* \neq 0$, then for any objective functions f_1, f_2, and any starting point, we have $x_1(t), x_2(t) \to x^*$, which implies $A(g(x_1(t), x_2(t)), \eta_t \nabla f_1(x_1(t))) \to x^*$. By the continuity of A and g and ∇f_1, we have $x^* = A(g(x^*, x^*), \eta^* \nabla f_1(x^*))$. We can choose f_1, f_2 to be simple quadratic functions such that $(x^*, \nabla f_1(x^*))$ can be any point in $\mathbb{R}^N \times \mathbb{R}^N$. Hence, since $\eta^* \neq 0$, we have, for any $x, y \in \mathbb{R}^N$, $x = A(g(x, x), y)$. This is impossible, because if we let the objective functions be $f_1(x) = f_2(x) = \frac{\alpha}{2}\|x\|^2$, and we start from $x_1(0) = x_2(0) \neq 0$, we will have the trajectory $x_i(t)$ stays fixed $x_1(t) = x_2(t) = x_1(0) = x_2(0)$, not converging to the minimizer 0. This is a contradiction. Hence, $\eta^* = 0$.

Now we focus on the case $f_1 = f_2 = f$, and $x_1(0) = x_2(0)$. Then $x_1(t)$ always equals $x_2(t)$ and we define $x(t) \triangleq x_1(t) = x_2(t)$. Let $\tilde{A}(x, y) = A(g(x, x), y)$, then $x(t)$ satisfies $x(t+1) = \tilde{A}(x(t), \eta_t \nabla f(x(t)))$. For $\tilde{A}$, we first show that $x = \tilde{A}(x, 0)$ for any x. This is because if we consider any x^* and a function $f(x) = \frac{\alpha}{2}\|x - x^*\|^2$ (thus

x^* is the minimizer of $f_1 + f_2 = 2f$), then the fact of $x(t) \to x^*$ and $\eta_t \nabla f(x(t)) \to 0$ implies that $x^* = \tilde{A}(x^*, 0)$ (by the continuity of A, g).

Now we are ready to prove the Theorem. Notice that for any objective functions $f_1 = f_2 = f$, if we start from $x_1(0) = x_2(0) \neq x^*$ (x^* is the minimizer of both f and $f_1 + f_2$), then the generated sequence $x(t) = x_1(t) = x_2(t)$ satisfies

$$\begin{aligned}
\|x(t+1) - x^*\| &= \|\tilde{A}(x(t), \eta_t \nabla f(x(t))) - x^*\| \\
&\geq \|\tilde{A}(x(t), 0) - x^*\| \\
&\quad - \|\tilde{A}(x(t), \eta_t \nabla f(x(t))) - \tilde{A}(x(t), 0)\| \\
&\geq \|x(t) - x^*\| - L\eta_t \|\nabla f(x(t))\| \\
&\geq (1 - \eta_t L\beta)\|x(t) - x^*\|
\end{aligned}$$

The theorem follows from the fact that $\eta_t L\beta \to 0$. □

C. Harnessing Smoothness via History Information

Motivated by the previous discussion and the impossibility results, we seek for alternative methods to exploit smoothness to develop faster distributed algorithms. Firstly we note that one major reason for the slow convergence of DGD is the decreasing step size η_t. This motivates us to use a constant step size η in our algorithm (2a). But we have discussed that constant η will lead to optimization error due to the fact that $\nabla f_i(x_i(t))$ could be very different from the average gradient $g(t) = \frac{1}{n}\sum_i \nabla f_i(x_i(t))$. However, because of smoothness, $\nabla f_i(x_i(t+1))$ and $\nabla f_i(x_i(t))$ would be close (as well as $g(t+1)$ and $g(t)$) if $x_i(t+1)$ and $x_i(t)$ are close, which is exactly the case when the algorithm is coming close to the minimizer x^*. This motivates the second step of our algorithm (2b), using history information to get an accurate estimation of the average gradient $g(t)$ which is a better descent direction than ∇f_i. Similar ideas of using history information traces back to [23], in which the previous gradient is used to narrow down the possible values of the current gradient to save communication complexity for a two-agent optimization problem.

A very recent paper [16] proposes an algorithm that achieves similar convergence results as our algorithm. The algorithm in [16] can be regarded as adding an integration type correction term to (3) while using a fixed step size. This correction term also involves history information in a certain way, which is consistent with our impossibility result. The differences of our algorithm with [16] are summarized below. Firstly, the two consensus matrices in [16] need to be symmetric and also satisfy a predefined spectral relationship, while our algorithm has a looser requirement on the consensus matrices. Secondly, without assuming strongly convex, [16] achieves a $O(\frac{1}{t})$ convergence rate in terms of the optimality residuals, which can be loosely defined as $\|\nabla f(x_i(t))\|^2$ and $\|x_i(t) - \bar{x}(t)\|^2$. Our algorithm not only achieves $O(\frac{1}{t})$ for the optimality residuals, but also achieves $O(\frac{1}{t})$ in terms of the objective error $f(x_i(t)) - f^*$, which is a more direct measure of optimality. But one downside of our current results is that [16] gives an explicit step size bound that only depends on β and not on W, whereas our step size bound for the

strongly convex case (Lemma 2) depends on W, and we do not have an explicit step size bound for the non-strongly convex case (Theorem 3).

IV. Convergence Analysis

A. Analysis Setup

We first stack the $x_i(t)$, $s_i(t)$ and $\nabla f_i(x_i(t))$ in (2) into matrices. Define $x(t), s(t), \nabla(t) \in \mathbb{R}^{n\times N}$ as

$$x(t) = \begin{bmatrix} x_1(t) \\ x_2(t) \\ \vdots \\ x_n(t) \end{bmatrix}, s(t) = \begin{bmatrix} s_1(t) \\ s_2(t) \\ \vdots \\ s_n(t) \end{bmatrix}$$

$$\nabla(t) = \begin{bmatrix} \nabla f_1(x_1(t)) \\ \nabla f_2(x_2(t)) \\ \vdots \\ \nabla f_n(x_n(t)) \end{bmatrix}$$

In this way we can compactly write the update equation (2) as

$$x(t+1) = Wx(t) - \eta s(t) \tag{7a}$$

$$s(t+1) = Ws(t) + \nabla(t+1) - \nabla(t) \tag{7b}$$

and also $s(0) = \nabla(0)$. We start by introducing two straightforward lemmas. Lemma 5 derives update equations that govern the average sequence $\bar{x}(t)$ and $\bar{s}(t)$. Lemma 6 gives inequalities that are direct consequences of smoothness.

Lemma 5. *The following equalities hold*

(a) $\bar{s}(t+1) = \bar{s}(t) + g(t+1) - g(t) = g(t+1)$

(b) $\bar{x}(t+1) = \bar{x}(t) - \eta\bar{s}(t) = \bar{x}(t) - \eta g(t)$

Proof. W is doubly stochastic, which means $\mathbb{1}^T W = \mathbb{1}^T$, therefore,

$$\begin{aligned} \bar{x}(t+1) &= \frac{1}{n}\mathbb{1}^T x(t+1) \\ &= \frac{1}{n}\mathbb{1}^T(Wx(t) - \eta s(t)) \\ &= \bar{x}(t) - \eta\bar{s}(t) \end{aligned}$$

$$\begin{aligned} \bar{s}(t+1) &= \frac{1}{n}\mathbb{1}^T s(t+1) \\ &= \frac{1}{n}\mathbb{1}^T[Ws(t) + \nabla(t+1) - \nabla(t)] \\ &= \bar{s}(t) + g(t+1) - g(t) \end{aligned}$$

Do this recursively, we can write

$$\bar{s}(t+1) = \bar{s}(0) + g(t+1) - g(0)$$

Since $s(0) = \nabla(0)$, hence $\bar{s}(0) = g(0)$. This finishes the proof. □

Lemma 6. *Under Assumption 1, the following inequalities hold*

(a) $\|\nabla(t) - \nabla(t-1)\| \leq \beta\|x(t) - x(t-1)\|$

(b) $\|g(t) - g(t-1)\| \leq \beta\frac{1}{\sqrt{n}}\|x(t) - x(t-1)\|$

(c) $\|g(t) - h(t)\| \leq \beta\frac{1}{\sqrt{n}}\|x(t) - \mathbb{1}\bar{x}(t)\|$

Proof. (a)

$$\begin{aligned}
\|\nabla(t) - \nabla(t-1)\| &= \sqrt{\sum_{i=1}^{n} \|\nabla f_i(x_i(t)) - \nabla f_i(x_i(t-1))\|^2} \\
&\leq \sqrt{\sum_{i=1}^{n} \beta^2 \|x_i(t) - x_i(t-1)\|^2} \\
&= \beta\|x(t) - x(t-1)\|
\end{aligned}$$

(b)

$$\begin{aligned}
\|g(t) - g(t-1)\| &= \|\sum_{i=1}^{n} \frac{\nabla f_i(x_i(t)) - \nabla f_i(x_i(t-1))}{n}\| \\
&\leq \sum_{i=1}^{n} \frac{\|\nabla f_i(x_i(t)) - \nabla f_i(x_i(t-1))\|}{n} \\
&\leq \beta \sum_{i=1}^{n} \frac{\|x_i(t) - x_i(t-1)\|}{n} \\
&\leq \beta \sqrt{\sum_{i=1}^{n} \frac{\|x_i(t) - x_i(t-1)\|^2}{n}} \\
&= \beta\frac{1}{\sqrt{n}}\|x(t) - x(t-1)\|
\end{aligned}$$

(c)

$$\begin{aligned}
\|g(t) - h(t)\| &= \|\sum_{i=1}^{n} \frac{\nabla f_i(x_i(t)) - \nabla f_i(\bar{x}(t))}{n}\| \\
&\leq \sum_{i=1}^{n} \frac{\|\nabla f_i(x_i(t)) - \nabla f_i(\bar{x}(t))\|}{n} \\
&\leq \beta \sum_{i=1}^{n} \frac{\|x_i(t) - \bar{x}(t)\|}{n} \\
&\leq \beta \sqrt{\sum_{i=1}^{n} \frac{\|x_i(t) - \bar{x}(t)\|^2}{n}} \\
&= \beta\frac{1}{\sqrt{n}}\|x(t) - \mathbb{1}\bar{x}(t)\|
\end{aligned}$$

□

B. Why the Algorithm Works: An Intuitive Explanation

We give our intuition that partially explains, under the assumption of strongly-convex and smooth, the algorithm (7) can achieve a linear convergence rate. In fact we can prove the following proposition.

Proposition 7. *The following is true*

- *Assuming* $\|s(t) - \mathbb{1}g(t)\| = O(\mu^t)$ *decays at a linear rate, then* $\|x(t) - \mathbb{1}x^*\| = O(\gamma^t)$ *also decays at a linear rate.*
- *Assuming* $\|x(t) - \mathbb{1}x^*\| = O(\gamma^t)$ *decays at a linear rate, then* $\|s(t) - \mathbb{1}g(t)\| = O(\mu^t)$ *also decays at a linear rate.*

Proof. On one hand, assume $\|s(t) - \mathbb{1}g(t)\| \leq C\mu^t$, for some constant C and $\mu \in (0, 1)$, i.e. the estimation error decays at a linear rate. Then, the consensus error

$$\begin{aligned}
\|x(t) - \mathbb{1}\bar{x}(t)\| &\leq \|Wx(t-1) - \mathbb{1}\bar{x}(t-1)\| \\
&\quad + \eta\|s(t-1) - \mathbb{1}g(t-1)\| \\
&\leq \sigma\|x(t-1) - \mathbb{1}\bar{x}(t-1)\| + \eta C\mu^{t-1} \\
&\leq \sigma^t\|x(0) - \mathbb{1}\bar{x}(0)\| + \eta C\sum_{k=0}^{t-1}\gamma^k\mu^{t-1-k} \\
&= \sigma^t\|x(0) - \mathbb{1}\bar{x}(0)\| + \eta C\frac{\sigma^t - \mu^t}{\sigma - \mu}
\end{aligned}$$

Therefore, the consensus error also decays at a linear rate. Then, by Lemma 6 (c), $\|g(t) - h(t)\|$ also decays at a linear rate, and we have $\|g(t) - h(t)\| \leq C'\mu'^t$. By Lemma 5, $\bar{x}(t) = \bar{x}(t-1) - \eta h(t-1) - \eta(g(t-1) - h(t-1))$. Since $h(t-1) = \nabla f(\bar{x}(t-1))$, $\bar{x}(t-1) - \eta h(t-1)$ is a standard gradient step for function f. Since f is strongly convex and smooth, a standard gradient descent step shrinks the distance to the minimizer by a least a fixed ratio, hence we have

$$\|\bar{x}(t-1) - \eta h(t-1) - x^*\| \leq \lambda\|\bar{x}(t-1) - x^*\|$$

for some $\lambda \in (0, 1)$. Hence,

$$\begin{aligned}
\|\bar{x}(t) - x^*\| &\leq \lambda\|\bar{x}(t-1) - x^*\| + \eta\|h(t-1) - g(t-1)\| \\
&\leq \lambda\|\bar{x}(t-1) - x^*\| + \eta C'\mu'^{t-1} \\
&= \lambda^t\|\bar{x}(0) - x^*\| + \eta C'\frac{\lambda^t - \mu'^t}{\lambda - \mu'}
\end{aligned}$$

Therefore $\|\bar{x}(t) - x^*\|$, the distance of the average $\bar{x}(t)$ to the minimizer decays at a linear rate. Combined this with the fact that the consensus error decays at a linear rate, we have $\|x(t) - \mathbb{1}x^*\|$, the distance to the optimizer, decays at a linear rate.

On the other hand, assume $\|x(t)-\mathbb{1}x^*\|$ decays at a linear rate, then $\|x(t)-x(t-1)\|$, and subsequently by Lemma 6 (a)(b) $\|\nabla(t)-\nabla(t-1)\|$ and $\|g(t)-g(t-1)\|$, also decays at a linear rate. Let $\|x(t)-x(t-1)\| \leq C''\mu''^{t-1}$, then,

$$\begin{aligned}
&\|s(t)-\mathbb{1}g(t)\| \\
&\leq \|Ws(t-1)-\mathbb{1}g(t-1)\| + \|\nabla(t)-\nabla(t-1)\| \\
&\quad + \|\mathbb{1}g(t)-\mathbb{1}g(t-1)\| \\
&\leq \sigma\|s(t-1)-\mathbb{1}g(t-1)\| + 2\beta C''\mu''^{t-1} \\
&\leq \sigma^t\|s(0)-\mathbb{1}g(0)\| + 2\beta C''\frac{\mu''^t-\sigma^t}{\mu''-\sigma}
\end{aligned}$$

Hence the gradient estimation error $\|s(t)-\mathbb{1}g(t)\|$ decays at a linear rate. □

We now interpret the above proposition. On one hand, assuming the gradient estimation error $\|s(t)-\mathbb{1}g(t)\|$ decays at a linear rate, the distance to optimizer $\|x(t)-\mathbb{1}x^*\|$ will decay at a linear rate; on the other hand, assuming the distance to optimizer decays at a linear rate, the gradient estimation error will also decay at a linear rate. While we eventually expect to see both the gradient estimation error and the distance to optimizer decay at a linear rate, the above reasoning is a circular argument and it does not prove any thing. But it illustrates how the algorithm works: the gradient descent step (7a) and the gradient estimation step (7b) facilitate each other to converge fast in a reciprocal manner. One of them can converge at a linear rate if the other one can converge at a linear rate. This mutual dependence is distinct from many of the previous methods, where one usually bounds the consensus error at first, and then use the consensus error to bound the optimality error, and there is no mutual dependence between the two. In the next two subsections, we will rigorously prove the convergence.

C. Convergence Analysis: Strongly Convex

We start by introducing a lemma that can be found in standard optimization literature, e.g. [21]. The lemma states that if we perform a gradient descent step with a fixed step size for a strongly convex and smooth function, then the distance to optimizer shrinks by at least a fixed ratio.

Lemma 8. *$\forall x \in \mathbb{R}^N$, define $x^+ = x - \eta\nabla f(x)$ where $0 < \eta < \frac{2}{\beta}$, then*

$$\|x^+ - x^*\| \leq \lambda\|x - x^*\|$$

where $\lambda = \max(|1-\eta\alpha|, |1-\eta\beta|)$

Now we prove Theorem 1.

Proof. Our strategy is to bound $\|s(k)-\mathbb{1}g(k)\|$, $\|x(k)-\mathbb{1}\bar{x}(k)\|$, and $\|\bar{x}(k)-x^*\|$ in terms of linear combinations of their past values, and in this way obtain a recursive linear vector inequality, which will imply linear convergence.

Step 1: Bound $\|s(k)-\mathbb{1}g(k)\|$. By the update rules (7b), we have

$$s(k)-\mathbb{1}g(k) = [Ws(k-1)-\mathbb{1}g(k-1)]$$

$$+ [\nabla(k) - \nabla(k-1)] - [\mathbb{1}g(k) - \mathbb{1}g(k-1)]$$

Take the norm, and notice that the column-wise average of $s(k-1)$ is just $g(k-1)$ by Lemma 5 (a), using the property of the consensus matrix W, we have

$$\begin{aligned}
&\|s(k) - \mathbb{1}g(k)\| \\
&\leq \|Ws(k-1) - \mathbb{1}g(k-1)\| \\
&\quad + \Big\|[\nabla(k) - \nabla(k-1)] - [\mathbb{1}g(k) - \mathbb{1}g(k-1)]\Big\| \\
&\leq \sigma\|s(k-1) - \mathbb{1}g(k-1)\| \\
&\quad + \Big\|[\nabla(k) - \nabla(k-1)] - [\mathbb{1}g(k) - \mathbb{1}g(k-1)]\Big\| \qquad (8)
\end{aligned}$$

It is easy to verify

$$\begin{aligned}
&\Big\|[\nabla(k) - \nabla(k-1)] - [\mathbb{1}g(k) - \mathbb{1}g(k-1)]\Big\|^2 \\
&= \|\nabla(k) - \nabla(k-1)\|^2 - n\|g(k) - g(k-1)\|^2 \\
&\leq \|\nabla(k) - \nabla(k-1)\|^2
\end{aligned}$$

Combine this with (8), and also use Lemma 6 (a), we get

$$\begin{aligned}
&\|s(k) - \mathbb{1}g(k)\| \\
&\leq \sigma\|s(k-1) - \mathbb{1}g(k-1)\| + \beta\|x(k) - x(k-1)\| \qquad (9)
\end{aligned}$$

Step 2: Bound $\|x(k) - \mathbb{1}\bar{x}(k)\|$. Consider update rule (7a) and use Lemma 5(b) and the property of W, we have

$$\begin{aligned}
\|x(k) - \mathbb{1}\bar{x}(k)\| &\leq \sigma\|x(k-1) - \mathbb{1}\bar{x}(k-1)\| \\
&\quad + \eta\|s(k-1) - \mathbb{1}g(k-1)\| \qquad (10)
\end{aligned}$$

Step 3: Bound $\|\bar{x}(k) - x^*\|$. Notice by Lemma 5(b), the update rule for $\bar{x}(k)$ is that

$$\bar{x}(k) = \bar{x}(k-1) - \eta h(k-1) - \eta[g(k-1) - h(k-1)]$$

Since the gradient of f at $\bar{x}(k)$ is actually $h(k)$, therefore, by Lemma 8 and Lemma 6 (c), we have

$$\begin{aligned}
&\|\bar{x}(k) - x^*\| \\
&\leq \lambda\|\bar{x}(k-1) - x^*\| + \eta\|g(k-1) - h(k-1)\| \\
&\leq \lambda\|\bar{x}(k-1) - x^*\| + \eta\frac{\beta}{\sqrt{n}}\|x(k-1) - \mathbb{1}\bar{x}(k-1)\| \qquad (11)
\end{aligned}$$

where $\lambda = \max(|1 - \eta\alpha|, |1 - \eta\beta|)$.

Step 4: Bound $\|x(k) - x(k-1)\|$. Notice that by smoothness

$$\|h(k-1)\| = \|\nabla f(\bar{x}(k-1))\| \leq \beta\|\bar{x}(k-1) - x^*\|$$

Combine the above and Lemma 6(c), we have

$$\|s(k-1)\|$$

$$
\begin{aligned}
&\leq \|s(k-1) - \mathbb{1}g(k-1)\| \\
&\quad + \|\mathbb{1}g(k-1) - \mathbb{1}h(k-1)\| + \|\mathbb{1}h(k-1)\| \\
&\leq \|s(k-1) - \mathbb{1}g(k-1)\| + \beta\|x(k-1) - \mathbb{1}\bar{x}(k-1)\| \\
&\quad + \beta\sqrt{n}\|\bar{x}(k-1) - x^*\|
\end{aligned}
$$

Hence

$$
\begin{aligned}
&\|x(k) - x(k-1)\| \\
&= \|Wx(k-1) - x(k-1) - \eta s(k-1)\| \\
&= \|(W-I)(x(k-1) - \mathbb{1}\bar{x}(k-1)) - \eta s(k-1)\| \\
&\leq 2\|x(k-1) - \mathbb{1}\bar{x}(k-1)\| + \eta\|s(k-1)\| \\
&\leq \eta\|s(k-1) - \mathbb{1}g(k-1)\| \\
&+ (\eta\beta + 2)\|x(k-1) - \mathbb{1}\bar{x}(k-1)\| + \eta\beta\sqrt{n}\|\bar{x}(k-1) - x^*\|
\end{aligned} \tag{12}
$$

Step 5: Derive a recursive inequality. We combine the previous four steps into a big recursive inequality. Plug (12) into (9), we have

$$
\begin{aligned}
\|s(k) - \mathbb{1}g(k)\| \leq (\sigma + \beta\eta)&\|s(k-1) - \mathbb{1}g(k-1)\| \\
&+ \beta(\eta\beta + 2)\|x(k-1) - \mathbb{1}\bar{x}(k-1)\| \\
&+ \eta\beta^2\sqrt{n}\|\bar{x}(k-1) - x^*\|
\end{aligned} \tag{13}
$$

Combine (13), (10) and (11), we get

$$
\overbrace{\begin{bmatrix} \|s(k) - \mathbb{1}g(k)\| \\ \|x(k) - \mathbb{1}\bar{x}(k)\| \\ \sqrt{n}\|\bar{x}(k) - x^*\| \end{bmatrix}}^{\triangleq z(k) \in \mathbb{R}^3} \leq \overbrace{\begin{bmatrix} (\sigma + \beta\eta) & \beta(\eta\beta + 2) & \eta\beta^2 \\ \eta & \sigma & 0 \\ 0 & \eta\beta & \lambda \end{bmatrix}}^{\triangleq G(\eta) \in \mathbb{R}^{3\times 3}} \cdot \overbrace{\begin{bmatrix} \|s(k-1) - \mathbb{1}g(k-1)\| \\ \|x(k-1) - \mathbb{1}\bar{x}(k-1)\| \\ \sqrt{n}\|\bar{x}(k-1) - x^*\| \end{bmatrix}}^{\triangleq z(k-1) \in \mathbb{R}^3} \tag{14}
$$

where '$\leq$' means element wise less than or equal to. Since $z(k)$ and $G(\eta)$ have nonnegative entries, we can actually expand (14) recursively, and get

$$
z(k) \leq G(\eta)^k z(0)
$$

The rest is to show that all the eigenvalues of $G(\eta)$ have magnitudes less than 1 when η is sufficiently small. We first require $\eta < \frac{1}{\beta}$, in this way $\lambda = 1 - \alpha\eta$. Then, notice that when $\eta = 0$, $G(0)$'s eigenvalues are σ, σ and 1. We want to investigate how the eigenvalue 1 is perturbed if we slightly increase η. Let the characteristic equation of $G(\eta)$ be $p(y) = 0$. This equation defines a implicit map from η to y on a neighborhood of $\eta = 0$. From the

equation we can calculate $\frac{dy}{d\eta}|_{\eta=0,y=1} = -\alpha < 0$, which means that when η is positive and small enough, $G(\eta)$'s all three eigenvalues will have magnitudes less than 1. So we are done.

Lemma 9. *We have $\rho(G(\eta)) < 1$ when η satisfies the following criterion.*

$$0 < \eta < \eta_0 \triangleq \frac{2}{\beta}\frac{(1-\sigma)^2}{3+\sqrt{13+4\frac{\beta}{\alpha}}}$$

Proof. Since $\eta_0 < \frac{1}{\beta}$, in the rest of the proof we will let $\lambda = 1 - \alpha\eta$. Define $\gamma(\eta) = \rho(G(\eta))$ over $\eta \in [0, \eta_0)$. When $\eta = 0$, $G(0)$'s eigenvalues are σ, σ and 1. We now investigate how the eigenvalue 1 is perturbed if we slightly increase η from 0. Let the characteristic equation of $G(\eta)$ be $p(y) = 0$. This equation defines a implicit map from η to y on a neighborhood of $\eta = 0$. From the equation we can calculate $\gamma'(0) = \frac{dy}{d\eta}|_{\eta=0,y=1} = -\alpha < 0$. Therefore, there exists a $\eta_1 \in (0, \eta_0)$ such that $\gamma(\eta) < 1$ on $(0, \eta_1)$. Assume there exists a $\eta' \in (0, \eta_0)$ s.t. $\gamma(\eta') \geq 1$, by continuity of γ there exists a $\eta'' \in [\eta_1, \eta']$ such that $\gamma(\eta'') = 1$. Since $G(\eta'')$ is a nonnegative matrix, by Perron-Frobenius Theorem, 1 is a eigenvalue of $G(\eta'')$, i.e. $\det(I - G(\eta'')) = 0$, i.e.

$$\eta''\left[\beta^2(\alpha+\beta)\eta''^2 + \alpha\beta(3-\sigma)\eta' - \alpha(1-\sigma)^2\right] = 0$$

Solve this equation for η'', we get three solutions, one is negative, one is 0, and the last one is

$$\eta'' = \frac{2\alpha(1-\sigma)^2}{\sqrt{\Delta} + \alpha\beta(3-\sigma)}$$

where $\Delta = \alpha^2\beta^2(3-\sigma)^2 + 4\alpha(\alpha+\beta)\beta^2(1-\sigma)^2$. It is easy to verify such a $\eta'' > \eta_0$. But $\eta_0 > \eta' \geq \eta''$. This is a contradiction.

□

□

D. Convergence Analysis: Non-Strongly Convex

Proof. We first change the starting configuration to be that, all nodes agree upon a vector $x_0 \in \mathbb{R}^{1\times N}$ first, and then set $x_i(0) = x_0$. Then, each agent calculates $\nabla f_i(x_i(0))$ and sets $s_i(0) = \frac{1}{n}\sum_{i=1}^n \nabla f_i(x_i(0))$ (which might need some initial coordination). We do this for simplifying the proof, and the theorem is still true if we use the original starting configuration. We also note here that the lemmas we are to cite in this proof, Lemma 5 and Lemma 6, are still true after the change of starting configuration.

We divide the proof into 3 steps. Step 1 will be devoted to bound what we call 'relative consensus error' (15). After this is done, as will be shown in step 2 and 3, our proof will follow almost the same as the proof of the $O(\frac{1}{t})$ rate for CGD.

Step 1: Bound relative consensus error. In this step, we prove the inequality (15), which is essentially upper bounding the consensus error by the gradient $\|h(k)\|$ (thus the name 'relative consensus error')

$$\|x(k) - \mathbb{1}\bar{x}(k)\| \leq \frac{1}{2}\frac{\sqrt{n}}{\beta}\|h(k)\| \tag{15}$$

We first prove the following lemma.

Lemma 10. *The following inequality is true.*

$$\overbrace{\begin{bmatrix} \|s(k)-\mathbb{1}g(k)\| \\ \|x(k)-\mathbb{1}\bar{x}(k)\| \end{bmatrix}}^{\triangleq z'(k)\in\mathbb{R}^2} \leq \overbrace{\begin{bmatrix} (\sigma+\beta\eta) & \beta(\eta\beta+2) \\ \eta & \sigma \end{bmatrix}}^{\triangleq G'(\eta)\in\mathbb{R}^{2\times 2}} \cdot \overbrace{\begin{bmatrix} \|s(k-1)-\mathbb{1}g(k-1)\| \\ \|x(k-1)-\mathbb{1}\bar{x}(k-1)\| \end{bmatrix}}^{\triangleq z'(k-1)\in\mathbb{R}^2} + \begin{bmatrix} \eta\beta\sqrt{n}\|h(k-1)\| \\ 0 \end{bmatrix} \tag{16}$$

Moreover, It can be shown that, there exists $\eta_0 > 0$ such that $\rho(G'(\eta_0)) < 1$, and the eigenvector vectors associated with $G'(\eta_0)$'s leading eigenvalue can be made to have all positive entries.

Proof. It is easy to check that (9) and (10) (copied below) still holds if we remove the strongly convex assumption.

$$\begin{aligned} \|s(k)-\mathbb{1}g(k)\| &\leq \sigma\|s(k-1)-\mathbb{1}g(k-1)\| \\ &\quad + \beta\|x(k)-x(k-1)\| \end{aligned} \tag{17}$$

$$\begin{aligned} \|x(k)-\mathbb{1}\bar{x}(k)\| &\leq \eta\|s(k-1)-\mathbb{1}g(k-1)\| \\ &\quad + \sigma\|x(k-1)-\mathbb{1}\bar{x}(k-1)\| \end{aligned} \tag{18}$$

Notice we have

$$\begin{aligned} &\|s(k-1)\| \\ &\leq \|s(k-1)-\mathbb{1}g(k-1)\| \\ &\quad + \|\mathbb{1}g(k-1)-\mathbb{1}h(k-1)\| + \|\mathbb{1}h(k-1)\| \\ &= \|s(k-1)-\mathbb{1}g(k-1)\| + \beta\|x(k-1)-\mathbb{1}\bar{x}(k-1)\| \\ &\quad + \sqrt{n}\|h(k-1)\| \end{aligned} \tag{19}$$

And hence

$$\begin{aligned} &\|x(k)-x(k-1)\| \\ &= \|Wx(k-1)-x(k-1)-\eta s(k-1)\| \\ &= \|(W-I)(x(k-1)-\mathbb{1}\bar{x}(k-1))-\eta s(k-1)\| \\ &\leq 2\|x(k-1)-\mathbb{1}\bar{x}(k-1)\| + \eta\|s(k-1)\| \end{aligned} \tag{20}$$

Combining (17), (19) and (20) yields

$$\begin{aligned} &\|s(k)-\mathbb{1}g(k)\| \\ &\leq (\sigma+\eta\beta)\|s(k-1)-\mathbb{1}g(k-1)\| \end{aligned}$$

$$+ \beta(\eta\beta + 2)\|x(k-1) - \mathbb{1}\bar{x}(k-1)\| + \eta\beta\sqrt{n}\|h(k-1)\|$$

Combining the above and (18) yields (16)

In the statement of the lemma we claimed that $\exists\ \eta_0 > 0$, such that $\rho(G'(\eta_0)) < 1$, and its leading eigenvector can have all positive entries. We now prove the claim. We notice that $G'(0)$ has eigenvalue σ with multiplicity 2. It is easy to check, by explicitly solving the characteristic equation, that there exists $\eta_0 > 0$ such that $G'(\eta_0)$ will have two distinct *real* eigenvalues $\lambda_1 < \lambda_2$, both in $(0, 1)$. Then, since $\lambda_1 + \lambda_2 = trace(G'(\eta_0)) = 2\sigma + \beta\eta_0$, we have $\lambda_2 > \sigma$. Therefore, the second row of $G'(\eta_0) - \lambda_2 I$ have entries that are both non-zero and have different signs. Let $\mu \in \mathbb{R}^2$ be a non-zero solution to $(G'(\eta_0) - \lambda_2 I)\mu = 0$, i.e. μ is a eigenvector of $G'(\eta_0)$ associated with $G'(\eta_0)$'s leading eigenvalue. μ cannot have any zero entry because otherwise, since the second row of $G'(\eta_0) - \lambda_2 I$ have entries that are both non-zero, μ would be zero. Since the second row of $G'(\eta_0) - \lambda_2 I$ have opposite signs, the two entries of μ must have the same sign. Therefore, we can choose μ to have all positive signs. This proves the claim.

□

From the above lemma, we define $G'_0 = G'(\eta_0)$ and let $\theta = \rho(G'_0)$. In the rest of the proof, we will require $\eta < \eta_0$, therefore

$$z'(k) \leq G'_0 z'(k-1) + \begin{bmatrix} \eta\beta\sqrt{n}\|h(k-1)\| \\ 0 \end{bmatrix} \tag{21}$$

We now focus on (15). We in fact prove a more general statement than (15). Let $\mu = [\mu_1, \mu_2]^T$ be the eigenvector corresponds to eigenvalue θ of G'_0. We can choose μ s.t. $\mu_2 = 1$ and $\mu_1 > 0$, since by step 1, G'_0's eivenvector associated with θ can have all positive entries. We now further require η to be small enough such that

$$(\theta + \frac{2}{\mu_1}\eta\beta^2)\frac{2}{1+\theta} < 1 \tag{22}$$

$$1 - \frac{3}{2}\eta\beta > \frac{1+\theta}{2} \tag{23}$$

With this requirement we prove the following statement (note (26) is exactly our original statement (15))

$$\|h(k)\| \geq \frac{1+\theta}{2}\|h(k-1)\| \tag{24}$$

$$\|s(k-1) - \mathbb{1}g(k-1)\| \leq \frac{1}{2}\frac{\sqrt{n}}{\beta}\|h(k-1)\|\mu_1 \tag{25}$$

$$\|x(k-1) - \mathbb{1}\bar{x}(k-1)\| \leq \frac{1}{2}\frac{\sqrt{n}}{\beta}\|h(k-1)\|\mu_2 \tag{26}$$

For $k = 1$, (25) and (26) are certainly true, since the left hand sides are zero. We can also verify that (24) is true for $k = 1$. In fact, we have

$$\begin{aligned} \|h(1)\| &\geq \|h(0)\| - \|h(1) - h(0)\| \\ &\geq \|h(0)\| - \beta\|\bar{x}(1) - \bar{x}(0)\| \\ &= \|h(0)\| - \eta\beta\|g(0)\| \end{aligned}$$

$$
\begin{aligned}
&\geq (1-\eta\beta)\|h(0)\| \\
&\geq (1-\frac{3}{2}\eta\beta)\|h(0)\| \\
&\geq \frac{1+\theta}{2}\|h(0)\|
\end{aligned}
$$

Suppose the statement is true for k. Notice that (25) and (26) are equivalent to $z'(k-1) \leq \frac{1}{2}\frac{\sqrt{n}}{\beta}\|h(k-1)\|\mu$. For $k+1$, we have by (21) and the induction assumption,

$$
\begin{aligned}
z'(k) &\leq G_0'\frac{1}{2}\frac{\sqrt{n}}{\beta}\|h(k-1)\|\mu + \begin{bmatrix} \eta\beta\sqrt{n}\|h(k-1)\| \\ 0 \end{bmatrix} \\
&\leq \theta\frac{1}{2}\frac{\sqrt{n}}{\beta}\|h(k-1)\|\mu + \frac{1}{\mu_1}\eta\beta\sqrt{n}\|h(k-1)\|\mu \\
&\leq (\theta + \frac{2}{\mu_1}\eta\beta^2)\frac{2}{1+\theta}\frac{1}{2}\frac{\sqrt{n}}{\beta}\|h(k)\|\mu \\
&\leq \frac{1}{2}\frac{\sqrt{n}}{\beta}\|h(k)\|\mu
\end{aligned} \tag{27}
$$

This shows that (25) and (26) is true for $k+1$. For (24), we have

$$
\begin{aligned}
\|h(k+1)\| &\geq \|h(k)\| - \|h(k+1)-h(k)\| \\
&\geq \|h(k)\| - \beta\|\bar{x}(k+1)-\bar{x}(k)\| \\
&= \|h(k)\| - \eta\beta\|g(k)\| \\
&\geq (1-\eta\beta)\|h(k)\| - \eta\beta\|g(k)-h(k)\| \\
&\geq (1-\eta\beta)\|h(k)\| - \eta\frac{\beta^2}{\sqrt{n}}\|x(k)-\mathbb{1}\bar{x}(k)\| \\
&\geq (1-\eta\beta)\|h(k)\| - \eta\frac{\beta^2}{\sqrt{n}}\frac{1}{2}\frac{\sqrt{n}}{\beta}\|h(k)\| \\
&= (1-\frac{3}{2}\eta\beta)\|h(k)\| \\
&\geq \frac{1+\theta}{2}\|h(k)\|
\end{aligned}
$$

Thus (24) is true for $k+1$. This concludes our induction.

Step 2: Follow the proof of CGD. We can show, by the β smoothness of f and after some manipulations (see [18]), when $\eta < \frac{1}{\beta}$,

$$
\begin{aligned}
&f(\bar{x}(k+1)) \\
&\leq f(\bar{x}(k)) + \langle h(k), \bar{x}(k+1)-\bar{x}(k)\rangle + \frac{\beta}{2}\|\bar{x}(k+1)-\bar{x}(k))\|^2 \\
&= f(\bar{x}(k)) - \eta\langle h(k), g(k)\rangle + \frac{\beta\eta^2}{2}\|g(k)\|^2 \\
&= f(\bar{x}(k)) - \eta\|h(k)\|^2 - \eta\langle h(k), g(k)-h(k)\rangle \\
&+ \frac{\beta\eta^2}{2}\|h(k)\|^2 + \frac{\beta\eta^2}{2}\|g(k)-h(k)\|^2 + \beta\eta^2\langle h(k), g(k)-h(k)\rangle \\
&= f(\bar{x}(k)) + (\frac{\beta\eta^2}{2}-\eta)\|h(k)\|^2 + \frac{\beta\eta^2}{2}\|g(k)-h(k)\|^2
\end{aligned}
$$

$$
\begin{aligned}
&+ (\beta\eta^2 - \eta)\langle h(k), g(k) - h(k)\rangle \\
&\leq f(\bar{x}(k)) + (\frac{\beta\eta^2}{2} - \eta)\|h(k)\|^2 + \frac{\beta\eta^2}{2}\|g(k) - h(k)\|^2 \\
&\quad + (\eta - \beta\eta^2)\frac{\|h(k)\|^2 + \|g(k) - h(k)\|^2}{2} \\
&= f(\bar{x}(k)) - \frac{\eta}{2}\|h(k)\|^2 + \frac{\eta}{2}\|g(k) - h(k)\|^2 \\
&\leq f(\bar{x}(k)) - \frac{\eta}{2}\|h(k)\|^2 + \frac{\eta}{2}\frac{\beta^2}{n}\|x(k) - \mathbb{1}\bar{x}(k)\|^2
\end{aligned} \tag{28}
$$

Now plug (15) into (28), we get

$$
\begin{aligned}
f(\bar{x}(k+1)) &\leq f(\bar{x}(k)) - \frac{3}{8}\eta\|h(k)\|^2 \\
&\leq f(\bar{x}(0)) - \frac{3}{8}\eta\sum_{\ell=0}^{k}\|h(\ell)\|^2
\end{aligned} \tag{29}
$$

Since f is lower bounded, the above inequality directly shows that $\sum_{\ell=0}^{\infty}\|h(\ell)\|^2 < \infty$, thus proving the first part of the Theorem. Next, by (29), we have $f(\bar{x}(k)) \leq f(\bar{x}(0))$, therefore $\bar{x}(k)$ belongs to f's $f(\bar{x}(0))$-level set. Since the set of minimizers of f is compact, by [22] Proposition B.9, f's level sets are compact. Therefore, $\|\bar{x}(k)\|$ is upper bounded, and so is $\|\bar{x}(k) - x^*\|$. Let $\|\bar{x}(k) - x^*\| \leq C$, and let $\delta_k = f(\bar{x}(k)) - f^*$. By convexity,

$$
\delta_k \leq \langle \nabla f(\bar{x}(k)), \bar{x}(k) - x^*\rangle \leq C\|h(k)\|
$$

Plug the above into (29), we have $\delta_{k+1} \leq \delta_k - \frac{3}{8}\eta\frac{1}{C^2}\delta_k^2$, which is equivalent to

$$
\frac{1}{\delta_{k+1}} - \frac{1}{\delta_k} \geq \frac{3}{8}\eta\frac{1}{C^2}\frac{\delta_k}{\delta_{k+1}} \geq \frac{3}{8}\eta\frac{1}{C^2}
$$

Then it immediately follows that $\delta_k = O(\frac{1}{k})$, i.e. $f(\bar{x}(k)) - f^* = O(\frac{1}{k})$.

Step 3: Prove the rest of the statements. $\forall i$, by smoothness and (15)

$$
\begin{aligned}
&f(x_i(k)) - f^* \\
&\leq [f(\bar{x}(k)) - f^*] + \langle h(k), x_i(k) - \bar{x}(k)\rangle + \frac{\beta}{2}\|x_i(k) - \bar{x}(k)\|^2 \\
&\leq [f(\bar{x}(k)) - f^*] + (\frac{1}{2}\frac{\sqrt{n}}{\beta} + \frac{1}{8}\frac{n}{\beta})\|h(k)\|^2
\end{aligned}
$$

The first term is $O(\frac{1}{k})$. Since $\|h(k)\|$ is square summable, it is easy to verify $\min_{k' \leq k}\|h(k')\|^2 = O(\frac{1}{k})$. Therefore,

$$
\min_{k' \leq k} f(x_i(k')) - f^* = O(\frac{1}{k})
$$

which concludes our proof. □

V. Numerical Experiments

We simulate our algorithm and compare it with other algorithms. We choose $n = 100$ agents. The algorithms we are comparing with include DGD (3) with vanishing step size and fixed step size, the algorithm proposed in [16] (with $\tilde{W} = \frac{W+I}{2}$ and step size $\frac{1}{\beta}$), and a direct optimization of f with full information using CGD (with step size $\frac{1}{\beta}$). The W matrix is chosen according to [5]. We focus on three cases: i) The functions f_i are smooth but not strongly convex; ii) The functions f_i are strongly convex and smooth, and the graph is well connected such

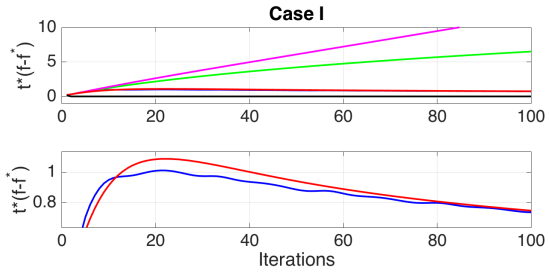

Figure 1: Simulation results for case I. The lower figure is the zoomed in version of the upper figure. Green is DGD (3) with vanishing step size; magenta is DGD (3) with fixed step size; blue is the algorithm in [16]; red is our algorithm; black is CGD.

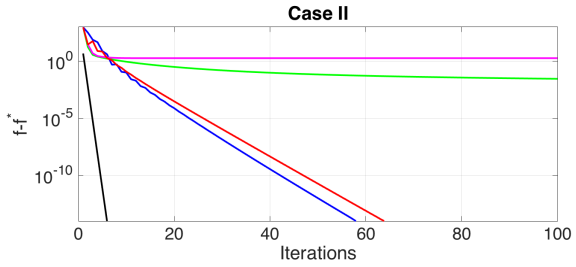

Figure 2: Simulation results for case II. Colors are the same as Figure 1.

that $\sigma < \frac{\beta-\alpha}{\beta+\alpha}$;[5] iii) The functions f_i are still strongly convex and smooth, but the graph is poorly connected such that $\sigma > \frac{\beta-\alpha}{\beta+\alpha}$. In case I, we plot $t \times (\frac{1}{n}\sum_i f(x_i(t)) - f^*)$ to check if the objective error decays as $O(\frac{1}{t})$. In case II and III, for each algorithm, we plot the average objective error, i.e. $\frac{1}{n}\sum_i f(x_i(t)) - f^*$, in the log scale. The results of the three cases are shown in Figure 1, 2 and 3 respectively. As shown in Figure 1, our algorithm, [16] and CGD all achieve a $O(\frac{1}{t})$ convergence rate for smooth but non-strongly convex functions, though CGD has a better constant in the big O. In Figure 2 and 3, our algorithm, [16] and CGD all achieve a linear convergence rate for smooth and strongly convex functions. When the graph is well connected (in Figure 2), our algorithm and [16] have roughly the same linear convergence rate. When the graph is poorly connected (in Figure 3), [16] exhibits a faster linear convergence rate, which leads to an interesting future direction to investigate reasons behind this. In both cases, CGD has a better linear convergence rate than [16] and our algorithm.

VI. Conclusion

We propose a method that can effectively harness smoothness to speed up distributed optimization. Future work includes giving better step size and convergence rate bounds, and apply the gradient estimation scheme to other first order optimization algorithms.

[5] $\frac{\beta-\alpha}{\beta+\alpha}$ is the theoretically optimal convergence rate of CGD for α-strongly convex and β-smooth functions.

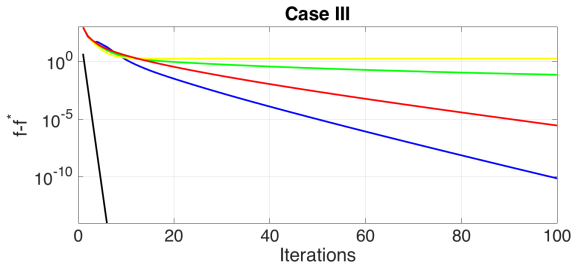

Figure 3: Simulation results for case III. Colors are the same as Figure 1.

REFERENCES

[1] J. N. Tsitsiklis, D. P. Bertsekas, and M. Athans, "Distributed asynchronous deterministic and stochastic gradient optimization algorithms," in *1984 American Control Conference*, 1984, pp. 484–489.

[2] D. P. Bertsekas and J. N. Tsitsiklis, *Parallel and distributed computation: numerical methods*. Prentice hall Englewood Cliffs, NJ, 1989, vol. 23.

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

[4] I. Lobel and A. Ozdaglar, "Convergence analysis of distributed subgradient methods over random networks," in *Communication, Control, and Computing, 2008 46th Annual Allerton Conference on*. IEEE, 2008, pp. 353–360.

[5] J. C. Duchi, A. Agarwal, and M. J. Wainwright, "Dual averaging for distributed optimization: convergence analysis and network scaling," *Automatic control, IEEE Transactions on*, vol. 57, no. 3, pp. 592–606, 2012.

[6] S. S. Ram, A. Nedić, and V. V. Veeravalli, "Distributed stochastic subgradient projection algorithms for convex optimization," *Journal of optimization theory and applications*, vol. 147, no. 3, pp. 516–545, 2010.

[7] A. Nedic and A. Olshevsky, "Stochastic gradient-push for strongly convex functions on time-varying directed graphs," *arXiv preprint arXiv:1406.2075*, 2014.

[8] ——, "Distributed optimization over time-varying directed graphs," *Automatic Control, IEEE Transactions on*, vol. 60, no. 3, pp. 601–615, 2015.

[9] I. Matei and J. S. Baras, "Performance evaluation of the consensus-based distributed subgradient method under random communication topologies," *Selected Topics in Signal Processing, IEEE Journal of*, vol. 5, no. 4, pp. 754–771, 2011.

[10] A. Olshevsky, "Linear time average consensus on fixed graphs and implications for decentralized optimization and multi-agent control," *arXiv preprint arXiv:1411.4186*, 2014.

[11] M. Zhu and S. Martínez, "On distributed convex optimization under inequality and equality constraints," *Automatic Control, IEEE Transactions on*, vol. 57, no. 1, pp. 151–164, 2012.

[12] I. Lobel, A. Ozdaglar, and D. Feijer, "Distributed multi-agent optimization with state-dependent communication," *Mathematical Programming*, vol. 129, no. 2, pp. 255–284, 2011.

[13] D. Jakovetic, J. Xavier, and J. M. Moura, "Fast distributed gradient methods," *Automatic Control, IEEE Transactions on*, vol. 59, no. 5, pp. 1131–1146, 2014.

[14] K. Yuan, Q. Ling, and W. Yin, "On the convergence of decentralized gradient descent," *arXiv preprint arXiv:1310.7063*, 2013.

[15] A. I. Chen and A. Ozdaglar, "A fast distributed proximal-gradient method," in *Communication, Control, and Computing (Allerton), 2012 50th Annual Allerton Conference on*. IEEE, 2012, pp. 601–608.

[16] W. Shi, Q. Ling, G. Wu, and W. Yin, "Extra: An exact first-order algorithm for decentralized consensus optimization," *SIAM Journal on Optimization*, vol. 25, no. 2, pp. 944–966, 2015.

[17] Y. Nesterov, *Introductory lectures on convex optimization: A basic course*. Springer Science & Business Media, 2013, vol. 87.

[18] G. Qu and N. Li. (2016) Harnessing smoothness to accelerate distributed optimization. [Online]. Available: http://scholar.harvard.edu/files/nali/files/cdc2016smooth.pdf

[19] R. Olfati-Saber, J. A. Fax, and R. M. Murray, "Consensus and cooperation in networked multi-agent systems," *Proceedings of the IEEE*, vol. 95, no. 1, pp. 215–233, Jan 2007.

[20] A. Olshevsky and J. N. Tsitsiklis, "Convergence speed in distributed consensus and averaging," *SIAM Journal on Control and Optimization*, vol. 48, no. 1, pp. 33–55, 2009.

[21] S. Bubeck *et al.*, "Convex optimization: Algorithms and complexity," *Foundations and Trends® in Machine Learning*, vol. 8, no. 3-4, pp. 231–357, 2015.

[22] D. P. Bertsekas, "Nonlinear programming," 1999.

[23] J. N. Tsitsiklis and Z.-Q. Luo, "Communication complexity of convex optimization," *Journal of Complexity*, vol. 3, no. 3, pp. 231–243, 1987.