This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TAC.2021.3056535, IEEE Transactions on Automatic Control

On Distributed Non-convex Optimization: Projected Subgradient Method For Weakly Convex Problems in Networks

Shixiang Chen, Alfredo Garcia, and Shahin Shahrampour

***Abstract*—The stochastic subgradient method is a widely-used algorithm for solving large-scale optimization problems arising in machine learning. Often these problems are neither smooth nor convex. Recently, Davis et al. [1], [2] characterized the convergence of the stochastic subgradient method for the weakly convex case, which encompasses many important applications (e.g., robust phase retrieval, blind deconvolution, biconvex compressive sensing, and dictionary learning). In practice, distributed implementations of the projected stochastic subgradient method (stoDPSM) are used to speed-up risk minimization. In this paper, we propose a distributed implementation of the stochastic subgradient method with a theoretical guarantee. Specifically, we show the global convergence of stoDPSM using the Moreau envelope stationarity measure. Furthermore, under a so-called sharpness condition, we show that deterministic DPSM (with a proper initialization) converges linearly to the sharp minima, using geometrically diminishing step-size. We provide numerical experiments to support our theoretical analysis.**

I. Introduction

Optimization in multi-agent networks has received a great deal of attention in the past few years in control, signal processing, and machine learning. A wide range of networked problems such as distributed detection [3], estimation [4], and localization [5] can be formulated via distributed optimization with applications in wireless sensor networks [6], robotic networks [7], power networks [8], and social networks [9]. In such decentralized frameworks, a number of agents in a network need to accomplish a global task, which is formulated as an optimization. Each individual agent, however, has limited information about the objective function. Therefore, agents locally interact with each other to solve the global problem. Decentralized techniques have gained popularity over time due to robustness to individual failures, imposing low computational burden on individual agents, and promoting privacy.

In a (constrained) multi-agent optimization, we deal with a problem of the form

$$\min_x f(x) = \frac{1}{N}\sum_{i=1}^{N} f_i(x) \quad \text{s.t.} \quad x \in \mathcal{X}, \tag{I.1}$$

The authors are with the Wm Michael Barnes '64 Department of Industrial and Systems Engineering, Texas A&M University, College Station, TX 77843. Email addresses: sxchen@tamu.edu (S. Chen), alfredo.garcia@tamu.edu (A. Garcia), shahin@tamu.edu (S. Shahrampour).
This work was supported by NSF ECCS-1933878.

where $\mathcal{X} \subset \mathbb{R}^n$ is a closed convex set known to all agents, and $f_i(x)$ is only available to agent i. Given this partial knowledge, agents must communicate with each other to minimize $f(x)$. There exists a large body of works on distributed convex optimization, where each $f_i(x)$ is convex (see e.g., the seminal work of [10] and its following papers). The literature has witnessed various algorithms for solving (I.1), which come with theoretical guarantees. In this paper, we depart from the classical convex setting and focus on *weakly convex and non-smooth* problems. In particular, we assume that every $f_i(x)$ is $\rho-$weakly convex and the subgradient $\|\partial f_i(x)\|$ is uniformly bounded. The definition of $\rho-$weakly convex function is as follows.

Definition I.1. *A function $f(x)$ is $\rho-$weakly convex ($\rho > 0$) if there exists a convex function $h(x)$ such that $f(x) = h(x) - \frac{\rho}{2}\|x\|^2$.*

Weakly convex problems play a key role in important machine learning applications such as robust phase retrieval [11]–[13], blind deconvolution [14], [15], biconvex compressive sensing [16], and dictionary learning [1]. Recently, Davis et al. [1], [2] characterized the convergence of the stochastic subgradient method for the weakly convex case. However, training the aforementioned models on a single device can take a significant amount of time for large-scale data. In practice, distributed implementations of the stochastic (sub)gradient method (stoDPSM) are used to speed-up the training time (see e.g. [17], [18]) and they can achieve a linear speedup [19].

In this paper, we focus on developing a theoretical convergence result for the distributed projected subgradient method (DPSM)

$$x_{i,k+1} = \mathrm{Proj}_{\mathcal{X}}\left(\sum_{j=1}^{N} a_{i,j}(k)x_{j,k} - \alpha_k g_{i,k}\right), \tag{I.2}$$

where $\mathrm{Proj}_{\mathcal{X}}$ denotes the orthogonal projection onto $\mathcal{X}$, $a_{i,j}(k)$ is the weight agent i associates to information received from agent j at time k, α_k is the stepsize, and $g_{i,k}$ is any subgradient of f_i evaluated at $\sum_{j=1}^{N} a_{i,j}(k)x_{j,k}$. This algorithmic scheme was first proposed for unconstrained distributed convex optimization [10] and it was then extended to the constrained scenario [20], [21]. Under standard assumptions on the weights $a_{i,j}(k)$ and α_k, it can be shown that (I.2) converges to a minimizer of $f(x)$ in (I.1) for convex problems. Nevertheless,

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TAC.2021.3056535, IEEE Transactions on Automatic Control

the weakly convex problem (which is essentially non-convex) has not been addressed in the literature. (i) We provide global convergence results for DPSM using the notion of Moreau envelope and show that the infimum of its gradient approaches zero with a rate $\mathcal{O}(k^{-1/4})$ if the stepsize is $\alpha_k = \mathcal{O}(1/\sqrt{k})$. (ii) We show a linear convergence rate of DPSM under a so-called sharpness condition. Specifically, the linear rate also relies on the fact that each local variable is sufficiently close to a sharp minimizer and the stepsize is given by $\alpha_k = \mu_0\gamma^k, \mu_0 \in (0,1), \gamma \in (0,1)$, where μ_0 and γ also depend on both function and network parameters. (iii) We also extend the global convergence results to the stochastic setting (stoDPSM). This paper is the first work providing convergence analysis of *distributed* subgradient for *weakly convex, non-smooth* problems, extending the classical convex analysis [10] to a *non-convex* setting. In other words, it generalizes the centralized results of [1], [2] to the decentralized setting. The technical challenges are as follows. For the global convergence of (sto)DPSM using Moreau envelope, we show a novel and crucial weakly convex inequality in Lemma II.1. For proving the linear convergence rate under the presence of sharpness property, we need to carefully identify the relationship between the geometrically diminishing stepsize with the function and network parameters. Moreover, since the sharpness property holds in a local region, each local variable should stay in the neighborhood of that region, posing another technical challenge.

A. Related work

We now briefly review the existing work on distributed (sub)gradient method. When f_i is convex and non-smooth, the distributed subgradient method converges in terms of function value [10] in unconstrained scenario and the distributed stochastic subgradient projection algorithms [22] can deal with a common constraint. In the case that each agent only knows its own constraint information, convergence guarantee was proved in [20]. In all above results, a diminishing stepsize (that is square-summable but not summable) is required and the convergence for constrained problem is measured by the distance between the sequence and the optimal set, i.e., the limit of $\mathrm{dist}(x_{i,k}, \mathcal{X}^*)$ is zero for any i, where $\mathcal{X}^*$ is the optimal set. The square-summable condition was relaxed in [21], provided the optimum set is bounded. The best convergence rate of non-smooth convex problem is given by $\inf_{0\le k\le T} f(\bar{x}_k) - f^* = \mathcal{O}(\log T/\sqrt{T})$ with $\alpha_k = \mathcal{O}(1/\sqrt{k})$ [10], where $\bar{x}_k := 1/N\sum_{i=1}^N x_{i,k}$ is the average point and f^* is the optimal function value. Moreover, a convergence rate of $\mathrm{dist}(x_{i,k}, \mathcal{X}^*) = \mathcal{O}(1/\sqrt{k})$ is shown in [21] if the stepsize is set to $\alpha_k = \mathcal{O}(1/k)$ for strongly convex f_i. For smooth convex and unconstrained case, the convergence is established in [23]. We refer to the survey [24] for a complete review of decentralized optimization of convex problems.

If f_i is non-convex and its gradient ∇f_i is Lipschitz continuous, f_i is automatically weakly convex. Algorithms such as [19], [25]–[28] have been proposed for the non-convex setting. For example, the convergence of distributed projected stochastic gradient was discussed in [25]; an ergodic convergence rate was established in [26] for proximal gradient method. A push-sum stochastic gradient method was proposed to train deep neural networks in [27] and the sublinear rate was established. When the objective is non-smooth, previous work, e.g., [26], [28] studied the composite form, i.e., $f_i(x) = g_i(x) + h_i(x)$, where g_i is smooth but h_i is non-smooth with a closed-form proximal mapping. In contrast, the non-smooth objective in (I.1) generally does not follow an easy proximal mapping. When the computation of subgradient of f_i is inexpensive, algorithm (I.2) is a better choice.

Recently, for weakly convex f_i, the centralized proximal-type subgradient methods have been shown to converge in finite time in terms of a stationarity measure using Moreau envelope (see [1], [12]). Under the presence of sharpness property, the centralized subgradient converges linearly in a local region [2]. We summarize the convergence results for distributed projected subgradient method in table I.

II. Preliminaries

Notation: We use $\langle x, y\rangle = x^\top y$ to denote the Euclidean inner product and $\|x\|$ to denote the Euclidean norm of x. We denote by $\partial h(x)$ the subgradient set of a convex function $h(x)$. Abusing notation, we use $\langle \partial h(x), y\rangle$ to denote the inner product of any elements of $\partial h(x)$ and a vector y.

A. Network Model

We consider a time-varying network of agents that can exchange information locally. To model the network, we use a time-varying graph $(\mathcal{V}, E_k)$, where $\mathcal{V} = \{1,\dots,N\}$ denotes the set of nodes and $E_k \subseteq \mathcal{N}\times\mathcal{N}$ is the set of links connecting nodes at time $k > 0$. Let $A(k) = [a_{i,j}(k)]$ denote the matrix of weights associated with links in the graph at time $k > 0$. For node i, $\mathcal{N}_i(k)$ denotes the neighborhood of i in which $a_{i,j}(k) > 0$. Define $\Phi(k,s) = A(s)A(s+1)\cdots A(k-1)A(k)$ for $k \ge s$, $\Phi(k,k) = A(k)$ and $\Phi(k,s) = I$ for $k < s$.

B. Weak Convexity, Optimality Measure and Sharpness

We assume the local objective $f_i(x)$ in (I.1) is $\rho-$weakly convex for some $\rho \ge 0$; i.e., there exists a convex function $h_i(x)$ such that $f_i(x) = h_i(x) - \frac{\rho}{2}\|x\|^2$. Although $f_i(x)$ is not convex, we may define its subdifferential by

$$\partial f_i(x) = \partial h_i(x) - \rho x, \quad \forall x \in \mathcal{X}; \tag{II.1}$$

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TAC.2021.3056535, IEEE Transactions on Automatic Control

f_i	strongly convex	convex	weakly convex
Measure	$\text{dist}(x_{i,k}, \mathcal{X}^*)$	$\text{dist}(x_{i,k}, \mathcal{X}^*)$ or $\inf_{1\le k\le T} f(\bar{x}_k) - f^*$	$\|\nabla \varphi_t(\bar{x}_k)\|$
Convergence	$\text{dist}(x_{i,k}, \mathcal{X}^*) = \mathcal{O}(\frac{1}{\sqrt{k}})$ with $\alpha_k = \mathcal{O}(1/k)$ [21]	$\lim_{k\to\infty} \text{dist}(x_{i,k}, \mathcal{X}^*) = 0$ [20], [21]; $\inf_{1\le k\le T} f(\bar{x}_k) - f^* = \mathcal{O}(\frac{\log T}{\sqrt{T}})$ with $\alpha_k = \mathcal{O}(\frac{1}{\sqrt{k}})$ [10]	This paper: $\inf_{1\le k\le T} \|\nabla \varphi_t(\bar{x}_k)\| = \mathcal{O}(\frac{1}{T^{1/4}})$ with $\alpha_k = \mathcal{O}(1/\sqrt{k})$.

TABLE I
CONVERGENCE RESULTS: DISTRIBUTED PROJECTED SUBGRADIENT METHOD FOR CONSTRAINED OPTIMIZATION

(see [29]). Here, $\partial h_i(x)$ is the subdifferential in the convex sense. The following lemma states an equivalent definition of weakly convex functions and strongly convex functions. We are particularly interested in (II.2) for the analysis and we establish (II.3) to prove (II.2). The proof is given in the Appendix.

Lemma II.1. *If $f(x)$ is ρ-weakly convex and $g(x)$ is $\tau-$strongly convex in $\mathbb{R}^n$, then $\forall x_1, \dots, x_m \in \mathbb{R}^n$, it follows that*

$$f(\sum_{i=1}^m a_i x_i) \le \sum_{i=1}^m a_i f(x_i) + \frac{\rho}{2}\sum_{i=1}^{m-1}\sum_{j=i+1}^m a_i a_j \|x_i - x_j\|^2 \tag{II.2}$$

and

$$g(\sum_{i=1}^m a_i x_i) \le \sum_{i=1}^m a_i g(x_i) - \frac{\tau}{2}\sum_{i=1}^{m-1}\sum_{j=i+1}^m a_i a_j \|x_i - x_j\|^2, \tag{II.3}$$

where $\sum_{i=1}^m a_i = 1$ and $a_i \ge 0$ for all i.

To analyze DPSM, we follow the framework of [1], where a novel convergence analysis for *centralized* subgradient method is proposed. We extend this analysis to the distributed case for weakly convex problems. Since there exist different stationary points in non-convex problems, neither the suboptimal objective value nor the distance to the optimum set tend to be good measures for the analysis. On the other hand, the subgradient of the objective is not continuous, which makes it difficult to analyze the convergence of the subgradient norm. A surrogate stationary measure for problem (I.1) was thus defined using the Moreau envelope in [1]. We briefly review it in the sequel.

Recall that $f_i(x)$ is $\rho-$weakly convex, iff we have the following inequality [1, Lemma 2.1]

$$f_i(y) \ge f_i(x) + \langle \partial f_i(x), y - x\rangle - \frac{\rho}{2}\|y - x\|^2. \tag{II.4}$$

This inequality is also known as prox-regular inequality introduced earlier in [30]. Therefore, $f(x) = 1/N\sum_{i=1}^N f_i(x)$ is also $\rho-$weakly convex. Let $\varphi(x) = f(x) + \mathbb{I}_{\mathcal{X}}(x)$, where $\mathbb{I}_{\mathcal{X}}$ is the indicator function of $\mathcal{X}$ [1]. Define the Moreau envelope [31] as

$$\varphi_t(x) := \min_{y\in\mathbb{R}^n} \varphi(y) + \frac{1}{2t}\|y - x\|^2, \quad 0 < t < 1/\rho.$$

[1] $\mathbb{I}_{\mathcal{X}}(x) = 0$ when $x \in \mathcal{X}$, and $\mathbb{I}_{\mathcal{X}}(x) = \infty$ otherwise.

Since $t < 1/\rho$ and f_i is $\rho-$weakly convex, we have that $\varphi(y) + \frac{1}{2t}\|y - x\|^2$ is strictly convex, i.e., the minimization problem above is strictly convex and the minimizer is unique. The Moreau envelope is a C^1 smooth approximation to the non-smooth function $f(x)$ over $\mathcal{X}$. Let us define

$$\hat{x} = \operatorname*{argmin}_{y\in\mathbb{R}^n} \varphi(y) + \frac{1}{2t}\|y - x\|^2,$$

where $\text{prox}_{tf}(x) := \hat{x}$ is called the proximal mapping. Here, we omit $\mathcal{X}$ in the notation prox_{tf}, but recall that the proximal mapping is related to $\mathcal{X}$. The proximal mapping is only used in the analysis, and it is not computed in the algorithm. From the optimality condition of $\hat{x}$, one has

$$0 \in \partial f(\hat{x}) + \partial \mathbb{I}_{\mathcal{X}}(\hat{x}) + \frac{1}{t}(\hat{x} - x).$$

It follows that

$$\text{dist}(0, \partial f(\hat{x}) + \partial \mathbb{I}_{\mathcal{X}}(\hat{x})) \le \frac{1}{t}\|\hat{x} - x\|.$$

Therefore, if $\frac{1}{t}\|\hat{x} - x\| \le \varepsilon$, then $\hat{x}$ is $\varepsilon-$stationary and x is close to the $\varepsilon-$stationary point $\hat{x}$. We also have [31]

$$\|\nabla \varphi_t(x)\| = \frac{1}{t}\|\hat{x} - x\|, \tag{II.5}$$

which can be used as near-stationarity measure of x. Next, we introduce the sharpness property.

Definition II.2 (Sharpness)**.** *A function $f : \mathcal{X} \to \mathbb{R}$ possesses the local sharpness property, if there exist constants $\beta > 0$ and $B > 0$ such that the following inequality holds for a minimizer x^* of $f(x)$*

$$f(x) - \min f \ge \beta\|x - x^*\|, \quad \forall x \in \mathcal{B}, \tag{II.6}$$

where $\mathcal{B} = \{x \in \mathcal{X} : \|x - x^*\| \le B\}$. Furthermore, x^* is called the local sharp minimizer of f.

As we can see, the sharpness property intuitively suggests that the objective grows at least linearly in $\mathcal{B}$. It has been shown that the centralized subgradient method converges linearly in the neighborhood of the local sharp minimizer [2], [11], [32].

C. Assumptions

In this part, we introduce the assumptions used for our analysis. To begin with, we define the average vector

$$\bar{x}_k := \frac{1}{N}\sum_{i=1}^{N} x_{i,k},$$

and

$$v_{i,k} := \sum_{j\in\mathcal{N}_i(k)} a_{i,j}(k)x_{j,k}.$$

Then, the iteration in DPSM (I.2) can be rewritten as

$$x_{i,k+1} = \mathrm{Proj}_{\mathcal{X}}\left(v_{i,k} - \alpha_k g_{i,k}\right), \tag{II.7}$$

where $g_{i,k} \in \partial f_i(v_{i,k})$ is any element of the subdifferential set. Unlike the centralized algorithm, the distributed update does not rely on a fusion center and computing $v_{i,k}$ and $g_{i,k}$ can be done in a decentralized manner. The following assumptions on the network are commonly adopted in the literature [10], [20], [33].

Assumption 1 (Weights rule)**.** *There exists a scalar $\eta \in (0,1)$ such that for all $i,j \in \{1,\dots,N\}$,*

- *$a_{i,i}(k) \geq \eta$ for all $k \geq 0$.*
- *If $a_{i,j}(k) > 0$ then $a_{i,j}(k) \geq \eta$.*

Assumption 2 (Doubly stochasticity)**.** *The weight matrix $A(k)$ is doubly stochastic (i.e., $\sum_j a_{i,j}(k) = \sum_j a_{j,i}(k) = 1, \forall i,k$).*

Assumption 3 (Connectivity)**.** *The graph $(\mathcal{V}, E_\infty)$ is strongly connected, where E_∞ is the set of edges (j,i) representing agent pairs communicating directly infinitely many times, i.e., $E_\infty = \{(j,i) : (j,i) \in E_k$ for infinitely many indices $k\}$.*

Assumption 4 (Bounded intercommunication interval)**.** *There exists an integer $B \geq 1$ such that for every $(j,i) \in E^\infty$, agent j sends its information to the neighboring agent i at least once every B consecutive time slots, i.e., at time t_k or at time $t_k + 1$ or $\dots$ or (at latest) at time $t_k + B - 1$ for any $k \geq 0$.*

We also need some assumptions on the function $f(x)$.

Assumption 5. *$f(x)$ is lower bounded. Every f_i is $\rho-$weakly convex and we have $\|\partial f_i(x)\| \leq L$ for $x \in \mathcal{X}$.*

The bounded subgradient holds if $f_i(x)$ is globally $L-$Lipschitz continuous on $\mathbb{R}^n$ or $\mathcal{X}$ [34]. One common class of weakly convex functions is $f(x) = h(c(x))$, where h is convex and Lipschitz and c is smooth with a Lipschitz continuous Jacobian [35]. However, such $f(x) = h(c(x))$ is usually locally Lipschitz continuous. A common assumption to resolve this issue is that $\mathcal{X}$ is compact or the sequence $\{x_{i,k}\}$ is bounded. Then, the boundedness of subgradient is equivalent to the $L-$Lipschitz property for $f_i(x)$. Such assumptions are usually needed in centralized algorithms (see [1]).

Last, the stepsize for the global convergence of DPSM should be non-summable and diminishing.

Assumption 6. *The stepsize $\alpha_k > 0$ satisfies*

$$\sum_{k=0}^{\infty}\alpha_k = \infty, \lim_{k\to\infty}\alpha_k = 0 \text{ and } \lim_{k\to\infty}\frac{\alpha_{k+1}}{\alpha_k} = 1.$$

A commonly used stepsize sequence satisfying Assumption 6 can be as follows

$$\alpha_k = \frac{1}{k^q}, \quad \text{where } q \in (0,1].$$

D. Technical lemmas

In this part, we introduce some necessary lemmas. All proofs can be found in the Appendix.

Lemma II.3. *[10, Proposition 1] Under Assumptions 1 to 4, there exist constants $c > 0$ and $\lambda \in (0,1)$ such that*

$$\|\Phi(k,s) - \frac{1}{N}\mathbf{1}\mathbf{1}^\top\|_{op} \leq c\lambda^{k-s},$$

where $\|\cdot\|_{op}$ is the matrix operator norm.

Lemma II.4. *[20, Lemma 7] [21, Proposition 8] Let $\lambda \in (0,1)$ and $\{\gamma_k\}$ be a positive sequence. Suppose γ_k satisfies Assumption 6. Considering the convolution sequence $\sum_{k=0}^{T-1}\lambda^k\gamma_{T-k-1}$, we have*

$$\sum_{k=0}^{T-1}\lambda^k\gamma_{T-k-1} = \mathcal{O}(\frac{\gamma_{T-1}}{1-\lambda}). \tag{II.8}$$

For convergence analysis, we should show that the deviation of individual errors from the mean $\|x_{i,k} - \bar{x}_k\|$ goes to zero. We define a vector Δ_k where $\Delta_{k,i} := x_{i,k} - \bar{x}_k$. That is, $\Delta_k \in \mathbb{R}^{Nn}$ is the vector formed by stacking all individual deviations from the mean. The following inequality (II.9) for Δ_k was established in [20], [21], [24] for convex problems. We show that it still holds for the weakly convex case.

Lemma II.5. *Under Assumptions 1 to 6, for the distributed projected subgradient algorithm (II.7), we have the following consensus result*

$$\lim_{k\to\infty}\|\Delta_{k,i}\| = 0, \quad \forall i. \tag{II.9}$$

Furthermore, similar to the result of [21, Proposition 8], the convergence rate can be characterized as follows.

Lemma II.6. *Under Assumptions 1 to 6, for the distributed projected subgradient algorithm (II.7), we have the following error rate*

$$\|\Delta_k\| = \mathcal{O}(\sqrt{N}L\cdot\frac{\alpha_k}{1-\lambda}). \tag{II.10}$$

We also have the following well-known property of the projection onto convex sets.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TAC.2021.3056535, IEEE Transactions on Automatic Control

Lemma II.7. *[20] For the convex closed set $\mathcal{X}$, it follows that $\forall y \in \mathcal{X}$*

$$\|\mathrm{Proj}_{\mathcal{X}}(x) - y\|^2 \leq \|x-y\|^2 - \|x - \mathrm{Proj}_{\mathcal{X}}(x)\|^2.$$

We further have the following property of the proximal mapping. Although the proximal mapping is not non-expansive when $f(x)$ is weakly convex, it is still Lipschitz continuous. This can be easily proved using the same idea for convex functions [36]. We omit the proof.

Lemma II.8. *If $f(x)$ is $\rho-$weakly convex, then the proximal mapping with $t < 1/\rho$ satisfies*

$$\|\mathrm{prox}_{tf}(x_1) - \mathrm{prox}_{tf}(x_2)\| \leq \frac{1}{1-t\rho}\|x_1 - x_2\|, \quad \forall x_1, x_2 \in \mathcal{X}.$$

III. Main Results and Convergence Analysis

In this section, we state the main convergence results of DPSM. First, if the Assumptions 1 to 6 hold, the Moreau envelope sequence $\{\varphi_t(\bar{x}_k)\}$ converges and the infimum of its gradient converges to 0. Second, under the sharpness condition, DPSM converges linearly with a geometrically diminishing stepsize in a neighborhood of a sharp minimizer. Finally, we also provide the convergence result of distributed projected stochastic subgradient method.

A. Global Convergence

We now establish the first convergence result. The following lemma states the improvement after one iteration of the algorithm (II.7). The proof is given in the Appendix.

Lemma III.1 (One-step improvement)**.** *Let*

$$s_k := \operatorname*{argmin}_{y\in\mathcal{X}} f(y) + \frac{1}{2t}\|y - \bar{x}_k\|^2,$$
$$\hat{v}_{i,k} := \operatorname*{argmin}_{y\in\mathcal{X}} f(y) + \frac{1}{2t}\|y - v_{i,k}\|^2.$$

Under Assumption 5, we have

$$\begin{aligned}&\sum_{i=1}^N \|x_{i,k+1} - \hat{v}_{i,k}\|^2 \leq \sum_{i=1}^N \|v_{i,k} - \hat{v}_{i,k}\|^2 \\ &+ 2\alpha_k\Bigg(N(-\frac{1}{2t}+\rho)\|\bar{x}_k - s_k\|^2 \\ &+\frac{L(2-t\rho)}{1-t\rho}\sum_{i=1}^N\|x_{i,k}-\bar{x}_k\| \\ &+2\rho(1+\frac{1}{(1-t\rho)^2})\sum_{i=1}^N\|x_{i,k}-\bar{x}_k\|^2\Bigg) + NL^2\alpha_k^2.\end{aligned} \tag{III.1}$$

By invoking Lemma II.5, the distance $\|x_{i,k} - \bar{x}_k\|$ goes to 0 for all i when α_k goes to zero. Note that for $0 < t < 1/(2\rho)$, the term $(-\frac{1}{2t}+\rho)\|\bar{x}_k - s_k\|^2$ in (III.1) is strictly negative if $\|\bar{x}_k - s_k\|^2$ is not zero. Then, we may have $\sum_{i=1}^N \|x_{i,k+1} - \hat{v}_{i,k}\|^2 < \sum_{i=1}^N \|v_{i,k} - \hat{v}_{i,k}\|^2$, which means that comapred to $v_{i,k}$, the new point $x_{i,k+1}$ is closer to $\hat{v}_{i,k}$. Hence, the algorithm continues to make progress. Now, we present our first main result, which shows the decay of the gradient of the Moreau envelope (optimality measure). The proof is given in the Appendix.

Theorem III.2. *Let $0 < t < \frac{1}{2\rho}$ and $\{x_{i,k}\}$ be the sequence of projected subgradient method for solving problem (I.1). Under Assumptions 1 to 6,*

(1) If $\sum_{k=0}^\infty \alpha_k^2 < \infty$, there exists $\bar{\varphi}_t$ such that $\lim_{k\to\infty}\varphi_t(x_{i,k}) = \lim_{k\to\infty}\varphi_t(\bar{x}_k) = \bar{\varphi}_t$;

(2) There exists $b_k = \mathcal{O}(\frac{L^2\alpha_k^2}{(1-\lambda)^2})$ such that

$$\inf_k \|\nabla\varphi_t(\bar{x}_k)\|^2 \leq \frac{2}{1-2t\rho}\cdot\frac{\varphi_t(\bar{x}_0) - \inf\varphi_t(x) + \sum_{k=0}^\infty b_k + \sum_{k=0}^\infty \frac{L^2\alpha_k^2}{2t}}{\sum_{k=0}^\infty \alpha_k}.$$

Statement (1) of Theorem III.2 suggests that the Moreau envelope function value converges at the mean $\bar{x}_k$ if $\sum_{k=0}^\infty \alpha_k^2 < \infty$. Statement (2) is the decentralized counterpart of the centralized algorithm established in [1]. It also provides the convergence rate of $\inf_k \|\nabla\varphi_t(\bar{x}_k)\|^2$. For example, if $\alpha_k = \mathcal{O}(1/\sqrt{k})$, we have

$$\inf_{1\leq k\leq T}\|\nabla\varphi_t(\bar{x}_k)\|^2 = \mathcal{O}(\frac{\bar{\varphi}_{t,0} - \bar{\varphi}_t}{\sqrt{T}} + \frac{L^2}{(1-\lambda)^2}\cdot\frac{\log T}{\sqrt{T}})$$

for sufficiently large T. Compared with the centralized algorithm [1], we have an extra term $\sum_k b_k$ in the upper bound, which is the cost of decentralization as it has the constant $\frac{1}{(1-\lambda)^2}$ involving network parameters. For the convex problem, a similar result is established under the optimality measure $\inf_{1\leq k\leq T} f(\bar{x}_k) - f^*$ [24, Theorem 8]. But the dependence on the λ is $\frac{1}{1-\lambda}$. Whether $\frac{1}{(1-\lambda)^2}$ is optimal for weakly-convex problems is an interesting question that we leave for future work.

B. Local Convergence Rate with Sharpness Property

In this section, we discuss the convergence rate of the Algorithm I.2 under the presence of sharpness property defined in Definition II.2. It has been shown the centralized subgradient method converges linearly in the neighborhood of a sharp minimizer [2], [11], if the Polyak stepsize [37] or geometrically diminishing stepsize [38] are adopted. The Polyak stepsize [37] and geometrically diminishing stepsize were firstly proposed for convex problems and they also work for weakly convex problems. Since the Polyak stepsize needs the knowledge of the optimal function value and the full information of the objective f, we will only consider

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TAC.2021.3056535, IEEE Transactions on Automatic Control

the geometrically diminishing stepsize, i.e, $\alpha_k = \mu_0\gamma^k$, where $\mu_0 > 0$ and $\gamma \in (0,1)$ are constants decided by the problem parameters. Under some conditions, we can show the linear rate of DPSM in Theorem III.4. The proof idea is as follows. As the sharpness is a property of the whole function $f(x)$, we can only use the sharpness inequality at $\bar{x}_k$. We first need to estimate the deviation from mean $\|\Delta_k\|$ when using the geometrically diminishing stepsize.

Lemma III.3. *Let the stepsize α_k in Algorithm I.2 be $\alpha_k = \mu_0\gamma^k$, $k \geq 0$, where $\mu_0 > 0$, $\gamma \geq \lambda^\delta$, $\delta \in (0,1)$ and λ is the parameter given in Lemma II.3. Then, $\|\Delta_k\| = \mathcal{O}(\alpha_k)$.*

Proof. From inequality (A.7) in the proof of Lemma II.5 and the fact $\gamma \geq \lambda^\delta$, we have

$$\begin{aligned}
&\|\Delta_{k+1}\| \\
&\leq c\lambda^k\|\Delta_0\| + c\sqrt{N}L\sum_{l=0}^{k-1}\lambda^{k-l-1}\alpha_l + \sqrt{N}L\alpha_k \\
&\leq \left(\frac{c}{\lambda}\|\Delta_0\| + \frac{\sqrt{N}L}{\lambda^2}\big(c\frac{\gamma^{\frac{1}{\delta}-1}}{1-\gamma^{\frac{1}{\delta}-1}} + \lambda\big)\mu_0\right)\gamma^{k+1}.
\end{aligned} \tag{III.2}$$

□

Assumption 7. *Let $x_{1,0},\ldots,x_{N,0}$ be the initial points in Algorithm II.7. Given any constants $\Lambda \in (\lambda,1)$ and $\Gamma \geq \sqrt{2}$, define*

$$\begin{aligned}
e_0 &:= \min\left\{\max\left\{\frac{\beta}{\rho\Gamma}, \sqrt{\frac{1}{N}\sum_{i=1}^N\|x_{i,0}-x^*\|^2}\right\}, \frac{B}{\Gamma}\right\}, \\
a &:= \frac{2(L+\beta)L}{\lambda^2}, \\
q &:= \frac{2\beta}{\Gamma}e_0 - \rho e_0^2 - \frac{2(L+\beta)c}{\sqrt{N}\lambda}\|\Delta_0\|,
\end{aligned}$$

where c,λ are constants given in Lemma II.3, β and B are defined in (II.6), ρ is the weak-convexity parameter, and L is the bound on subgradients. Let the stepsize in Algorithm II.7 be given by $\alpha_k = \mu_0\gamma^k$, where $0 < \mu_0 \leq \min\{\frac{e_0}{2\beta-\rho e_0}, \frac{q}{10\sqrt{N}(a\lambda+L^2+\frac{ac\Lambda}{1-\Lambda})}\}$ and $\gamma \in (0,1)$.

We use the stepsize assumption above to prove the following theorem. The proof sketch is as follows. The sharpness property holds for the global objective $f(x)$. This motivates us to consider the full information at average point $\bar{x}_k$. With the help of Lemma III.3 and Lemma A.1 in the Appendix, we can show that $\sum_{i=1}^N\|x_{i,k}-x^*\|^2$ decays linearly, but not for $\|x_{i,k}-x^*\|$, $i \in [N]$. Using the triangle inequality

$$\|x_{i,k+1}-x^*\| \leq \|x_{i,k+1}-\bar{x}_{k+1}\| + \|\bar{x}_{k+1}-x^*\|,$$

and Lemma III.3, we can show $\|x_{i,k}-x^*\|$ also converges linearly. Meanwhile, we need to carefully consider the relation between α_k and the network and problem parameters. The proof is provided in the Appendix.

Theorem III.4. *Let $N \geq 2$ and x^* be a local sharp minimizer of problem (I.1). Suppose the initial points $x_{1,0},\ldots,x_{N,0}$ in Algorithm II.7 satisfy for all $i \in \{1,\ldots,N\}$ the three constraints*

$$\sum_{i=1}^N\|x_{i,0}-x^*\|^2 \leq \frac{N}{\Gamma^2}\min\left\{(\frac{2\beta}{\rho})^2, B^2\right\}$$

$$\|x_{i,0}-x^*\|^2 \leq \frac{\Gamma^2}{N}\sum_{i=1}^N\|x_{i,0}-x^*\|^2,$$

$$\|\Delta_0\| < \frac{\frac{2}{\Gamma}\beta e_0 - \rho e_0^2}{2(L+\beta)c}\lambda,$$

where c,λ are constants given in Lemma II.3. Under Assumptions 1 to 5 and 7, there exists sufficiently small $\delta > 0$ such that for $\gamma = \lambda^\delta$, we have

$$\sum_{i=1}^N\|x_{i,k}-x^*\|^2 \leq N\gamma^{2k}e_0^2 \tag{III.3}$$

and

$$\|x_{i,k}-x^*\|^2 \leq \Gamma^2\gamma^{2k}e_0^2 \tag{III.4}$$

for any sequence $\{x_{i,k}\}$ generated by Algorithm II.7.

The following comments about the theorem are in order:

(1) The convergence rate $\gamma = \lambda^\delta$ is the same as the decaying rate of stepsize. But it cannot be smaller than λ, which is the convergence rate of the consensus.
(2) For the centralized subgradient method [2], the local linear rate is established in the tube
$$\mathcal{T} = \{x : \mathrm{dist}(x,\mathcal{X}_*) \leq \frac{2\beta}{\rho}\},$$
where $\mathcal{X}_*$ is the set of the local sharp minimizers. In Theorem III.4, the initialization constraints ensure that the individual initial points are close enough to each other as well as a sharp minimizer (local convergence). Moreover, since we use the local sharpness property, the local region should be included in $\mathcal{B}$.
(3) An immediate corollary of Theorem III.4 is that $\|\bar{x}_k-x^*\|^2 \leq 1/N\sum_{i=1}^N\|x_{i,k}-x^*\|^2 \leq \gamma^{2k}e_0^2$ under the same conditions.
(4) If $f_i(x)$ is convex, i.e., $\rho = 0$, then the condition $e_0 \leq \frac{\beta}{\Gamma\rho}$ can be removed. The weak convexity parameter ρ restricts the initialization region, which is also clearly stated for centralized subgradient method [2].

C. Distributed Projected Stochastic Subgradient Method

In some problems, the function $f_i(x)$ at local agent is given by $f_i(x) = \frac{1}{m_i}\sum_{j=1}^{m_i}f_{i,j}$, where m_i is a large number and $f_{i,l}$ is $\rho-$weakly convex. Therefore,

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TAC.2021.3056535, IEEE Transactions on Automatic Control

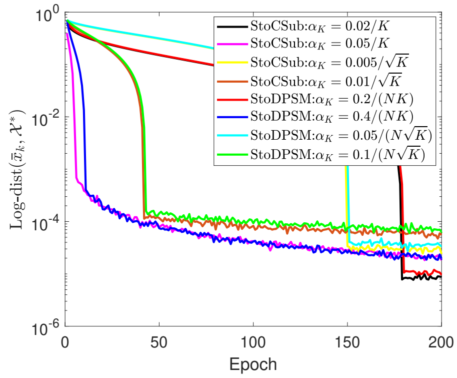


Fig. 1. Stochastic DPSM. $n = 100, N = 10, m = 1000$.

it is expensive to compute the subgradient of $f_i(x)$ in each iteration. In contrast to the algorithm (I.2), the distributed stochastic projected subgradient method iterates as follows

$$x_{i,k+1} = \mathrm{Proj}_{\mathcal{X}}\left(v_{i,k} - \alpha_k \xi_{i,k}\right), \tag{III.5}$$

where $\alpha_k > 0$ is the stepsize, and $\xi_{i,k}$ satisfies $\mathbb{E}\xi_{i,k} \in \partial f_i(v_{i,k})$. In practice, for each i, we uniformly randomly select index $i_l \in \{1, 2, \ldots, m_i\}$ and set $\xi_{i,k} \in \partial f_{i_l}(v_{i,k})$. We also assume that $\mathbb{E}\|\xi_{i,k}\|^2 \le L^2$ for all i, k, which is standard as in [1]. We have the following convergence result for distributed projected stochastic subgradient method (III.5). The proof is given in the Appendix.

Theorem III.5. *Let $t < \frac{1}{2\rho}$ and $\{x_{i,k}\}$ be the sequence of algorithm* (III.5)*. Under Assumptions 1 to 6,*

$$\lim_{T\to\infty} \inf_{k \le T} \mathbb{E}\|\nabla \varphi_t(\bar{x}_k)\| = 0.$$

IV. Numerical Experiment

We conduct simulations on robust phase retrieval problem

$$\min_{x\in\mathbb{R}^n} f(x) = \frac{1}{N}\sum_{i=1}^{N}\Big(\frac{1}{m}\sum_{j=1}^{m} |\left\langle w_{i,j}, x\right\rangle^2 - y_{i,j}|\Big). \tag{IV.1}$$

The problem is to recover the random signal $\tilde{x} \in \mathbb{R}^n$ using Gaussian measurements $w_{i,j}$. We generate the measurements $w_{i,j}$ and the observations $y_{i,j}$ following the work [12]. For simplicity, we only consider the noiseless case. More specifically, the ground truth $\tilde{x}$ is drawn from $N(0, I_n)$ and $y_{i,j} = \langle w_{i,j}, \tilde{x}\rangle^2$, where $w_{i,j}$ are i.i.d standard Gaussian random variables. As suggested by [12], the recovery rate is 100% when $N \times m \ge 2.7n$ for the proximal linear algorithm. Therefore, we use $N \times m \ge 3n$ for subgradient method in all tests. The initialization follows from the procedure proposed in [12, Section 4.2] and we set $x_{1,0} = x_{2,0} = \ldots = x_{N,0}$.

The robust phase retrieval formulation (IV.1) was shown to be weakly convex [12] and have sharpness property w.h.p under mild probabilistic assumptions in [12], [13]. Different from the Definition II.2, the sharpness condition is given by

$$f(x) - \min f \ge \kappa\|x - \tilde{x}\|\|x + \tilde{x}\|,$$

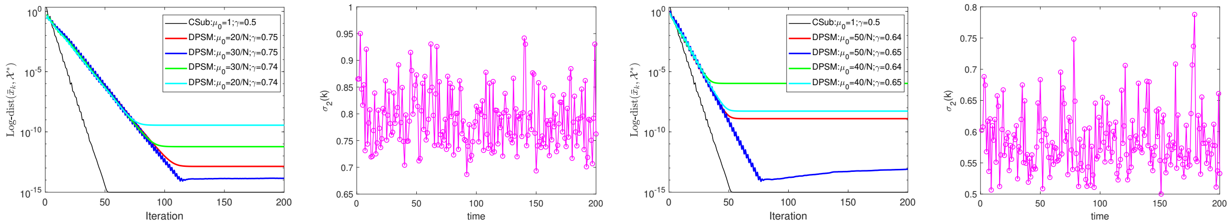
where $\kappa > 0$ is some number. Hence, $\pm\tilde{x}$ are also the global minimizers.

The global minimizers set is $\{\tilde{x}, -\tilde{x}\}$. According to [11, Lemma 3.1], there is no other critical points in the tube $\{x : \mathrm{dist}(x, \mathcal{X}^*) \le \frac{2\beta}{\rho}\}$. Since 0 is also a critical point to the population function $f_P(x) = \mathbb{E}_a[|\left\langle a, x\right\rangle^2 - \left\langle a, \tilde{x}\right\rangle^2|]$ [11, Theorem 5.1]. We have $\{x : \|x - \tilde{x}\| \le \frac{2\beta}{\rho}\} \cap \{x : \|x+\tilde{x}\| \le \frac{2\beta}{\rho}\} = \emptyset$. To satisfy Definition II.2, we let $\beta = \kappa\|\tilde{x}\|$ and choose $\mathcal{B} = \{x : \|x - x^*\| \le \frac{2\beta}{\rho}\}$, where x^* is $\tilde{x}$ or $-\tilde{x}$ and the sign is decided by the initialization.

Synthetic data First, we solve the robust phase retrieval problem (IV.1) by stochastic DPSM using diminishing stepsize. We generate an Erdös-Rényi model $\mathsf{ER}(N, 0.3)$ and $A(k) = A$ is time-invariant Metropolis Hasting matrix associated with the graph. Therefore, we have λ is the second largest singular value of A in Lemma II.3. In each epoch K, the stepsize is set to $\alpha_K = \mathcal{O}(1/K)$ or $\alpha_K = \mathcal{O}(1/\sqrt{K})$. We plot the log distance v.s. epoch K in fig. 1. We also compare stoDPSM with the stochastic centralized subgradient method(StoCSub) [1]. To make a fair comparison, we set the mini-batch size in StoCSub to $N = 10$. We tune the stepsize such that the best performance is achieved. We see that the convergence of StoDPSM is comparable with StoCSub in the epoch.

Next, we demonstrate the linear rate of DPSM. The data size is fixed with $N = 100, n = 100, m = 50$. Like the centralized subgradient method [2], μ_0 and γ should be tuned. In fig. 2, 'CSub' represents the centralized subgradient method [2]. The graph is time varying. Specifically, we generate an Erdös-Rényi model $\mathsf{ER}(100, p)$ and Metropolis Hasting matrix associated with the graph at each iteration. In fig. 2 (a) and (b), the probability p is 0.1. We demonstrate the linear convergence of CSub and DPSM with different stepsize in fig. 2 (a). We see that $\gamma = 0.5$ works for CSub but not for DPSM, since the smallest γ is 0.75 for $\mu_0 = 30/N$. For $\mu_0 = 20/N, \gamma = 0.75$, DPSM does not converge to the same precision as $\mu_0 = 30/N$, so the largest μ_0 may be $30/N$. This indicates that convergence rate of DPSM is slower than CSub. In fig. 2 (b), we plot the $\sigma_2(k)$ w.r.t the iteration, where $\sigma_2(k)$ is the second largest singular value of the matrix $A(k)$. We see that most singular values $\sigma_2(k)$ are larger than 0.7. And $\gamma = 0.75 > 0.7$ also demonstrates that the convergence rate cannot be faster than consensus. In fig. 2 (c) and (d), we show the similar results for $p = 0.2$. Since the connectivity is stronger, we find that smaller $\gamma = 0.65$ can guarantee the linear rate. Although DPSM is not faster than CSub

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TAC.2021.3056535, IEEE Transactions on Automatic Control



(a) convergence, $p = 0.1$ (b) $\sigma_2(k)$, $p = 0.1$ (c) convergence, $p = 0.2$ (d) $\sigma_2(k)$, $p = 0.2$

Fig. 2. Linear rate with different stepsize. Data size: $n = 100, N = 100, m = 50$.

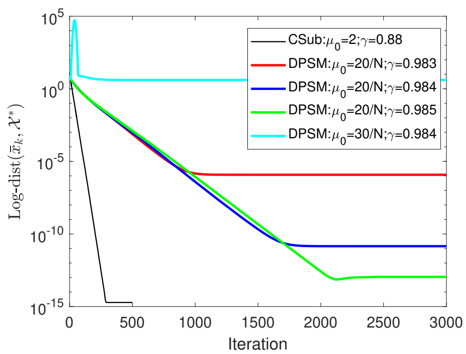


Fig. 3. Linear rate with different stepsize. Data size: $n = 100, N = 400, m = 1$. $\lambda = 0.28067$.

in the iteration number, DPSM has the advantage of parallel computation. And if the data number $m \times N$ is large, the computation of the whole subgradient is not affordable. We also test the case $m = 1$, i.e., there is only single data at each node. In this case, the graph is time-invariant. We fix the graph following $\mathsf{ER}(400, 0.3)$. However, the smallest $\gamma = 0.985$ is observed, which is away from $\lambda = 0.28067$ compared with fig. 2. This could be because single data in local node contributes little information about the sharpness.

Real-world image We use digit images from the MNIST data set [39]. The gray image dimension is $n = 28 \times 28 = 784$ and we set $m = 84, N = 28$ so that the number of Gaussian measurements is $m \times N = 3 \times n$. Other settings are the same as previous synthetic data. We fix a graph following $\mathsf{ER}(28, 0.3)$. In fig. 4, we show the original, initial guess and the recovered image. We see that the recovery is identical to the true image. The convergence plot of DPSM and stoDPSM is shown in fig. 5.

V. Conclusion

We analyzed the distributed subgradient method for solving constrained weakly convex optimization. We presented the global convergence of the average point using the notion of Moreau envelope. Moreover, we proved a linear convergence rate under the sharpness property. Numerical results on robust phase retrieval illustrate our theory.

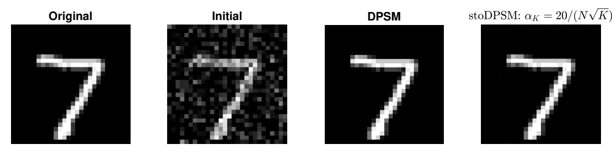
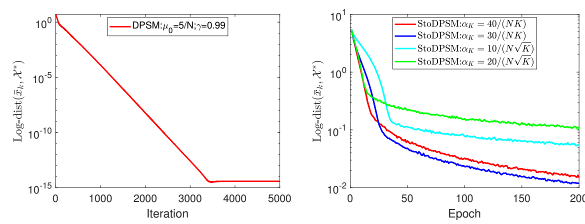


Fig. 4. Digit recovery; the first one is the true digit, the second is the initial guess, the third is the digit produced by DPSM and the last is produced by stoDPSM with $\alpha_K = 20/(N\sqrt{K})$. Data size: $n = 784, N = 28, m = 84$.



(a) DPSM (b) stoDPSM

Fig. 5. Linear rate of DPSM. MINIST Data: $n = 784, N = 28, m = 84$.

A natural extension of this work is to consider the directed network. For example, the convergence of directed distributed subgradient method for convex problems was analyzed in [40]. It will also be interesting to see whether it is possible to deal with different constraints at each local node (see e.g., the convex constraints in [20]). Finally, it will be worth considering non-convex constraints (e.g., sphere constraint [32]).

Appendix

Proof of Lemma II.1. We prove it by induction. For $m = 2$, let $y = a_1 x_1 + a_2 x_2$, where $a_1 + a_2 = 1, a_1 \geq 0$ and $a_2 \geq 0$. From the subgradient inequality (II.4), we have

$$f(x_1) \geq f(y) + \langle \partial f(y), x_1 - y \rangle - \frac{\rho}{2}\|x_1 - y\|^2$$

and

$$f(x_2) \geq f(y) + \langle \partial f(y), x_2 - y \rangle - \frac{\rho}{2}\|x_2 - y\|^2.$$

Multiplying the two above inequalities by a_1 and a_2, respectively, and summing them, yield

$$a_1 f(x_1) + a_2 f(x_2) \geq f(y) - \frac{\rho}{2} a_1 a_2 \|x_1 - x_2\|^2.$$

Similarly, we also have

$$a_1 g(x_1) + a_2 g(x_2) \geq g(y) + \frac{\tau}{2} a_1 a_2 \|x_1 - x_2\|^2.$$

Therefore, inequality (II.2) holds for $m = 2$. Suppose they hold for $m = k$. For $m = k+1$, let $z = \sum_{i=1}^{k+1} a_i x_i$ and $b = \sum_{i=1}^{k} a_i$. We have

$$\begin{aligned} f(z) &= f(b \sum_{i=1}^{k} \frac{a_i}{b} x_i + a_{k+1} x_{k+1}) \\ &\leq b f(\sum_{i=1}^{k} \frac{a_i}{b} x_i) + a_{k+1} f(x_{k+1}) + \\ &\quad \frac{\rho}{2} a_{k+1} b \| \sum_{i=1}^{k} \frac{a_i}{b}(x_i - x_{k+1}) \|^2 \\ &\leq b \left(\sum_{i=1}^{k} \frac{a_i}{b} f(x_i) + \frac{\rho}{2} \sum_{i=1}^{k-1} \sum_{j=i+1}^{k} \frac{a_i a_j}{b^2} \|x_i - x_j\|^2 \right) \\ &\quad + a_{k+1} f(x_{k+1}) + \frac{\rho}{2} a_{k+1} b \| \sum_{i=1}^{k} \frac{a_i}{b}(x_i - x_{k+1}) \|^2, \end{aligned} \tag{A.1}$$

where the first inequality follows from $b + a_{k+1} = 1$ and the second from assumption step. Notice that since $\|\cdot\|^2$ is $2-$strongly convex, it follows from the assumption for strongly convex function that

$$\begin{aligned} &\| \sum_{i=1}^{k} \frac{a_i}{b}(x_i - x_{k+1}) \|^2 \\ &\leq \sum_{i=1}^{k} \frac{a_i}{b} \|x_i - x_{k+1}\|^2 - \sum_{i=1}^{k-1} \sum_{j=i+1}^{k} \frac{a_i a_j}{b^2} \|x_i - x_j\|^2. \end{aligned}$$

Substituting it into (A.1) yields

$$\begin{aligned} f(z) &\leq \sum_{i=1}^{k} a_i f(x_i) \\ &\quad + \frac{\rho}{2} \sum_{i=1}^{k-1} \sum_{j=i+1}^{k} (\frac{a_i a_j}{b} - \frac{a_{k+1} a_i a_j}{b}) \|x_i - x_j\|^2 \\ &\quad + a_{k+1} f(x_{k+1}) + \frac{\rho}{2} \sum_{i=1}^{k} a_i a_{k+1} \|x_i - x_{k+1}\|^2 \\ &= \sum_{i=1}^{k+1} a_i f(x_i) + \frac{\rho}{2} \sum_{i=1}^{k} \sum_{j=i+1}^{k+1} a_i a_j \|x_i - x_j\|^2, \end{aligned}$$

where we use $a_{k+1} = 1 - b$ in the equality. Therefore, inequality (II.2) holds for $m = k+1$. Using the same argument and noticing that $-\|\cdot\|^2$ is $2-$weakly convex, we have that (II.3) also holds for $m = k+1$. Hence, we obtain the desired results. $\square$

Proof of Lemma II.3. It is shown in [10, Proposition 1] that there exist η such that

$$\left| [\Phi(k,s)]_j^i - \frac{1}{N} \right| \leq 2 \frac{1 + \eta^{-B_0}}{1 - \eta^{B_0}} (1 - \eta^{B_0})^{(k-s)/B_0}$$

for all s and k with $k \geq s$, where $[\Phi(k,s)]_j^i-$ denotes the $i-$th row and $j-$th column element of $\Phi(k,s)$, $B_0 = (N-1)B$ and B is the intercommunication interval bound of Assumption 3. By using the matrix norm inequality

$$\|A\|_{op} \leq \|A\|_F \leq N \|A\|_\infty$$

for any symmetric real matrix $A \in \mathbb{R}^{N \times N}$, where $\|A\|_F$ is the Frobenius norm and $\|A\|_\infty = \max_{i,j} |[A]_j^i|$, we have the desired result, where $c = 2N \frac{1+\eta^{-B_0}}{1-\eta^{B_0}}$ and $\lambda = (1 - \eta^{B_0})^{B_0^{-1}}$. $\square$

Proof of Lemma II.4. This can be proved following the similar argument of [20, Lemma 7]. Since $\lim_{T\to\infty} \gamma_T = 0$, there exists $M > 0$ such that γ_k is uniformly bounded, i.e., $\gamma_k \leq M, \forall k \geq 0$. For each T, we have $\lambda^k \leq \gamma_{T-1}$ for any $k \geq K_0(T) := \lceil \frac{\log \gamma_{T-1}}{\log \lambda} \rceil$. It follows that

$$\begin{aligned} &\sum_{k=0}^{T-1} \lambda^k \gamma_{T-k-1} \\ &= \sum_{k=0}^{K_0(T)-1} \lambda^k \gamma_{T-k-1} + \sum_{k=K_0(T)}^{T-1} \lambda^k \gamma_{T-k-1} \\ &\leq \frac{1}{1-\lambda} \cdot \max_{0 \leq k \leq K_0(T)-1} \gamma_{T-k-1} + \frac{\lambda^{K_0(T)}}{1-\lambda} \cdot M \\ &\leq \frac{1}{1-\lambda} \left(\max_{0 \leq k \leq K_0(T)-1} \gamma_{T-k-1} + M \gamma_{T-1} \right). \end{aligned}$$

Recall $\lim_{T\to\infty} \gamma_T = 0$ and $\sum_T \gamma_T = \infty$. It is clear that $K_0(T) = \lceil \frac{\log \gamma_{T-1}}{\log \lambda} \rceil = o(T)$, otherwise there exists a subsequence decreasing geometrically, which contradicts with $\lim_{k\to\infty} \gamma_{k+1}/\gamma_k = 1$ and $\sum_T \gamma_T = \infty$. Then, we have

$$\lim_{T\to\infty} \frac{\max_{0 \leq k \leq K_0(T)-1} \gamma_{T-k-1}}{\gamma_{T-1}} = 1.$$

Therefore, $\sum_{k=0}^{T-1} \lambda^k \gamma_{T-k-1} = \mathcal{O}(\frac{\gamma_{T-1}}{1-\lambda})$ holds for sufficiently large T and thus we have the desired result. $\square$

Proof of Lemma II.5 and Lemma II.6. The inequality (II.9) is the same as [20, Lemma 8]. We provide the proof for completeness. Without loss of generality, we assume $n = 1$. Define

$$\begin{aligned} x_k &= [x_{1,k}, x_{2,k}, \ldots, x_{N,k}], \\ v_k &= [v_{1,k}, v_{2,k}, \ldots, v_{N,k}], \\ e_k &= [e_{1,k}, e_{2,k}, \ldots, e_{N,k}], \end{aligned} \tag{A.2}$$

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TAC.2021.3056535, IEEE Transactions on Automatic Control

where $e_{i,k} = \mathrm{Proj}_{\mathcal{X}}(v_{i,k} - \alpha_k g_{i,k}) - v_{i,k}$. The iteration (II.7) can be rewritten as

$$x_{k+1} = v_k + e_k = A(k)x_k + e_k. \tag{A.3}$$

That is, the iteration is split into a linear term $A(k)x_k$ and a nonlinear term e_k. Using Lemma II.7 and Assumption 5, it follows that

$$\|e_{i,k}\|^2 \le \|v_{i,k} - \alpha_k g_{i,k} - v_{i,k}\|^2 \le \alpha_k^2 L^2. \tag{A.4}$$

Therefore, we have

$$\|e_k\| \le \sqrt{N} L \alpha_k. \tag{A.5}$$

Let $J = \frac{1}{N}\mathbf{1}\mathbf{1}^T$, where $\mathbf{1} \in \mathbb{R}^N$ is a column vector with all elements 1. Then, $\Delta_k = x_k - Jx_k$. We have

$$\begin{aligned} \Delta_{k+1} &= (I-J)x_{k+1} \\ &= (I-J)A(k)x_k + (I-J)e_k \\ &= A(k)x_k - A(k)Jx_k + (I-J)e_k \\ &= A(k)\Delta_k + (I-J)e_k, \end{aligned} \tag{A.6}$$

where the third equality is due to $JA(k) = J = A(k)J$. Therefore, the following recursion holds for $k \ge s \ge 0$

$$\Delta_{k+1} = \Phi(k,s)\Delta_s + \sum_{l=s}^{k-1} \Phi(k,l+1)(I-J)e_l + (I-J)e_k.$$

Since $\mathbf{1}^\top \Delta_l = \mathbf{1}^\top (I-J)e_l = 0,\ \forall l$, we have

$$\begin{aligned} \Delta_{k+1} &= (\Phi(k,s) - J)\Delta_s \\ &\quad + \sum_{l=s}^{k-1} (\Phi(k,l+1) - J)(I-J)e_l + (I-J)e_k. \end{aligned}$$

It follows from Lemma II.3 that there exist $c > 0$ and $\lambda \in (0,1)$, where λ is independent of k, such that

$$\|\Delta_{k+1}\| \le c\lambda^k \|\Delta_0\| + c\sqrt{N} L \sum_{l=0}^{k-1} \lambda^{k-l-1}\alpha_l + \sqrt{N} L \alpha_k. \tag{A.7}$$

With the Lemma II.4 and $\lim_{k\to\infty} \alpha_{k+1}/\alpha_k = 1$, we have (II.10) as desired. □

Proof of Lemma III.1. The following inequality holds because of the non-expansiveness of the projector

$$\begin{aligned} \|x_{i,k+1} - \hat{v}_{i,k}\|^2 &= \|\mathrm{Proj}_{\mathcal{X}}(v_{i,k} - \alpha_k g_{i,k}) - \hat{v}_{i,k}\|^2 \\ &\le \|v_{i,k} - \alpha_k g_{i,k} - \hat{v}_{i,k}\|^2 \\ &= \|v_{i,k} - \hat{v}_{i,k}\|^2 - 2\alpha_k \langle v_{i,k} - \hat{v}_{i,k}, g_{i,k}\rangle \\ &\quad + \alpha_k^2 \|g_{i,k}\|^2. \end{aligned}$$

Recall the weak convexity of f_i and the boundedness of $g_{i,k}$. It follows that

$$\begin{aligned} &\|x_{i,k+1} - \hat{v}_{i,k}\|^2 \\ &\le \|v_{i,k} - \hat{v}_{i,k}\|^2 + 2\alpha_k (f_i(\hat{v}_{i,k}) - f_i(v_{i,k}) \\ &\quad + \frac{\rho}{2}\|v_{i,k} - \hat{v}_{i,k}\|^2) + L^2\alpha_k^2. \end{aligned} \tag{A.8}$$

Using the Lipschitz continuity of f_i and Lemma II.8, we have

$$\begin{aligned} &f_i(\hat{v}_{i,k}) - f_i(v_{i,k}) \\ &= f_i(\hat{v}_{i,k}) - f_i(s_k) + f_i(s_k) - f_i(\bar{x}_k) + f_i(\bar{x}_k) - f_i(v_{i,k}) \\ &\le L\|\hat{v}_{i,k} - s_k\| + f_i(s_k) - f_i(\bar{x}_k) + f_i(\bar{x}_k) - f_i(v_{i,k}) \\ &\le L(\frac{1}{1-t\rho} + 1)\|v_{i,k} - \bar{x}_k\| + f_i(s_k) - f_i(\bar{x}_k) \\ &\le \frac{L(2-t\rho)}{1-t\rho} \sum_{j=1}^N a_{i,j}(k)\|x_{j,k} - \bar{x}_k\| + f_i(s_k) - f_i(\bar{x}_k) \end{aligned} \tag{A.9}$$

and

$$\begin{aligned} &\frac{\rho}{2}\|v_{i,k} - \hat{v}_{i,k}\|^2 \\ &= \frac{\rho}{2}\|v_{i,k} - \bar{x}_k + \bar{x}_k - s_k + s_k - \hat{v}_{i,k}\|^2 \\ &\le \rho\|\bar{x}_k - s_k\|^2 + \rho\|v_{i,k} - \bar{x}_k + s_k - \hat{v}_{i,k}\|^2 \\ &\le \rho\|\bar{x}_k - s_k\|^2 + 2\rho(1 + \frac{1}{(1-t\rho)^2})\|v_{i,k} - \bar{x}_k\|^2 \\ &\le \rho\|\bar{x}_k - s_k\|^2 \\ &\quad + 2\rho(1 + \frac{1}{(1-t\rho)^2})\sum_{j=1}^N a_{i,j}(k)\|x_{j,k} - \bar{x}_k\|^2. \end{aligned} \tag{A.10}$$

Summing inequalities (A.9) and (A.10) for $i = 1, \dots, N$, yields

$$\begin{aligned} &\sum_{i=1}^N \Big(f_i(\hat{v}_{i,k}) - f_i(v_{i,k}) + \frac{\rho}{2}\|v_{i,k} - \hat{v}_{i,k}\|^2\Big) \\ \le &\frac{L(2-t\rho)}{1-t\rho}\sum_{i=1}^N \|x_{i,k} - \bar{x}_k\| + N(f(s_k) - f(\bar{x}_k)) \\ &+ N\rho\|\bar{x}_k - s_k\|^2 + 2\rho(1 + \frac{1}{(1-t\rho)^2})\sum_{i=1}^N \|x_{i,k} - \bar{x}_k\|^2. \end{aligned} \tag{A.11}$$

From the definition of s_k, if $t < \frac{1}{2\rho}$, one has

$$\begin{aligned} &f(s_k) - f(\bar{x}_k) + \rho\|\bar{x}_k - s_k\|^2 \\ &= f(s_k) - f(\bar{x}_k) + (\frac{1}{2t} - \frac{1}{2t} + \rho)\|\bar{x}_k - s_k\|^2 \\ &\le (-\frac{1}{2t} + \rho)\|\bar{x}_k - s_k\|^2. \end{aligned} \tag{A.12}$$

Therefore, we have (III.1) by combining (A.8), (A.11) and (A.12). □

Proof of Theorem III.2. (1). From the definition of $\varphi_t(x_{i,k+1})$, we have

$$\varphi_t(x_{i,k+1}) \le f(z) + \frac{1}{2t}\|x_{i,k+1} - z\|^2, \quad \forall z \in \mathcal{X}. \tag{A.13}$$

Let $\hat{v}_{i,k} = \mathrm{argmin}_{y\in\mathcal{X}}\, f(y) + \frac{1}{2t}\|y - v_{i,k}\|^2$ and $\hat{x}_{i,k} = \mathrm{argmin}_{y\in\mathcal{X}}\, f(y) + \frac{1}{2t}\|y - x_{i,k}\|^2$.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TAC.2021.3056535, IEEE Transactions on Automatic Control

Substituting $z = \hat{v}_{i,k}$ into (A.13), we obtain

$$\varphi_t\left(x_{i,k+1}\right) \le f(\hat{v}_{i,k}) + \frac{1}{2t}\left\|x_{i,k+1} - \hat{v}_{i,k}\right\|^2. \tag{A.14}$$

Summing the above inequality for i and using inequality (III.1) yields

$$\begin{aligned}
&\sum_{i=1}^{N}\varphi_t\left(x_{i,k+1}\right)\\
&\le \sum_{i=1}^{N}\varphi_t\left(v_{i,k}\right) + \frac{\alpha_k}{t}\Bigg(N(-\frac{1}{2t}+\rho)\|\bar{x}_k - s_k\|^2\\
&\quad + \frac{L(2-t\rho)}{1-t\rho}\sum_{i=1}^{N}\|x_{i,k}-\bar{x}_k\|\\
&\quad + 2\rho(1+\frac{1}{(1-t\rho)^2})\sum_{i=1}^{N}\|x_{i,k}-\bar{x}_k\|^2\Bigg) + \frac{NL^2\alpha_k^2}{2t}.
\end{aligned} \tag{A.15}$$

Noticing $v_{i,k} = \sum_{j=1}^{N} a_{i,j}(k)x_{j,k}$, we have

$$\begin{aligned}
&\varphi_t(v_{i,k})\\
&= f(\sum_{j=1}^{N} a_{i,j}(k)\hat{v}_{i,k}) + \frac{1}{2t}\|\sum_{j=1}^{N} a_{i,j}(k)(\hat{v}_{i,k} - x_{j,k})\|^2\\
&\le f(\sum_{j=1}^{N} a_{i,j}(k)\hat{x}_{j,k}) + \frac{1}{2t}\|\sum_{j=1}^{N} a_{i,j}(k)(\hat{x}_{j,k} - x_{j,k})\|^2\\
&\le \sum_{j=1}^{N} a_{i,j}(k) f(\hat{x}_{j,k})\\
&\quad + \frac{\rho}{2}\sum_{j=1}^{N-1}\sum_{l=j+1}^{N} a_{i,j}(k)a_{i,l}(k)\|\hat{x}_{j,k} - \hat{x}_{l,k}\|^2\\
&\quad + \sum_{j=1}^{N} a_{i,j}(k)\frac{1}{2t}\|\hat{x}_{j,k} - x_{j,k}\|^2\\
&\le \sum_{j=1}^{N} a_{i,j}(k)\varphi_t(x_{j,k})\\
&\quad + \frac{\rho}{2(1-t\rho)^2}\sum_{j=1}^{N-1}\sum_{l=j+1}^{N} a_{i,j}(k)a_{i,l}(k)\|x_{j,k} - x_{l,k}\|^2,
\end{aligned}$$

where the first inequality is because of the definition of $\hat{v}_{i,k}$ and $\sum_{j=1}^{N} a_{i,j}(k)\hat{x}_{j,k} \in \mathcal{X}$, the second inequality follows from inequality (II.2) in Lemma II.1 and the convexity of $\|\cdot\|^2$ and the last inequality holds due to Lemma II.8. Letting $\bar{\varphi}_{t,k+1} := \frac{1}{N}\sum_{i=1}^{N}\varphi_t\left(x_{i,k+1}\right)$ together with (A.15) gives

$$\begin{aligned}
&\bar{\varphi}_{t,k+1} \le \bar{\varphi}_{t,k}\\
&\quad + \frac{\rho}{2N(1-t\rho)^2}\cdot\\
&\quad \sum_{i=1}^{N}\sum_{j=1}^{N-1}\sum_{l=j+1}^{N} a_{i,j}(k)a_{i,l}(k)\|x_{j,k} - x_{l,k}\|^2\\
&\quad + \frac{\alpha_k}{t}\Bigg((\rho - \frac{1}{2t})\|\bar{x}_k - s_k\|^2\\
&\quad + \frac{L(2-t\rho)}{N(1-t\rho)}\sum_{i=1}^{N}\|x_{i,k}-\bar{x}_k\|\\
&\quad + \frac{2\rho}{N}(1+\frac{1}{(1-t\rho)^2})\sum_{i=1}^{N}\|x_{i,k}-\bar{x}_k\|^2\Bigg) + \frac{L^2\alpha_k^2}{2t}\\
&\le \bar{\varphi}_{t,k} + b_k + \frac{L^2\alpha_k^2}{2t},
\end{aligned} \tag{A.16}$$

where

$$\begin{aligned}
&b_k\\
&:= \frac{\rho}{2N(1-t\rho)^2}\sum_{i=1}^{N}\sum_{j=1}^{N-1}\sum_{l=j+1}^{N} a_{i,j}(k)a_{i,l}(k)\|x_{j,k} - x_{l,k}\|^2\\
&\quad + \frac{\alpha_k}{t}\Bigg(\frac{L(2-t\rho)}{N(1-t\rho)}\sum_{i=1}^{N}\|x_{i,k}-\bar{x}_k\|\\
&\quad + \frac{2\rho}{N}(1+\frac{1}{(1-t\rho)^2})\sum_{i=1}^{N}\|x_{i,k}-\bar{x}_k\|^2\Bigg).
\end{aligned}$$

and the last inequality in (A.16) follows from $-\frac{1}{2t}+\rho<0$. By invoking Lemma II.6, we have

$$\sum_{i=1}^{N}\sum_{j=1}^{N-1}\sum_{l=j+1}^{N} a_{i,j}(k)a_{i,l}(k)\|x_{j,k}-x_{l,k}\|^2 = \mathcal{O}(\frac{NL^2\alpha_k^2}{(1-\lambda)^2}), \tag{A.17}$$

$$\alpha_k\sum_{i=1}^{N}\|x_{i,k}-\bar{x}_k\| = \mathcal{O}(\frac{NL\alpha_k^2}{1-\lambda}), \tag{A.18}$$

$$\alpha_k\sum_{i=1}^{N}\|x_{i,k}-\bar{x}_k\|^2 = \mathcal{O}(\frac{NL^2\alpha_k^3}{(1-\lambda)^2}) \tag{A.19}$$

and thus $b_k = \mathcal{O}(\frac{L^2\alpha_k^2}{(1-\lambda)^2})$. Because $f(x)$ is lower bounded on $\mathcal{X}$, we have $\varphi_t(x)$ is also lower bounded on $\mathcal{X}$. From (A.16) it follows that

$$\bar{\varphi}_{t,k+1} - \inf\varphi_t(x) \le \bar{\varphi}_{t,k} - \inf\varphi_t(x) + \mathcal{O}(\alpha_k^2).$$

Since $\sum_{k=0}^{\infty}\alpha_k^2 < \infty$, using Lemma 2[2] in [41, Chapter 2.2] we have $\{\bar{\varphi}_{t,k}\}$ converges to some value $\bar{\varphi}_t$.

Recall that $\varphi_t(x)$ is continuous differentiable. Since $\|x_{i,k}-\bar{x}_k\| \to 0$, it follows that

$$|\varphi_t(x_{i,k}) - \varphi_t(\bar{x}_k)|^2 \to 0$$

and

$$\begin{aligned}
|\bar{\varphi}_{t,k} - \varphi_t(\bar{x}_k)|^2 &= |\frac{1}{N}\sum_{i=1}^{N}\varphi_t(x_{i,k}) - \varphi_t(\bar{x}_k)|^2\\
&\le \frac{1}{N}\sum_{i=1}^{N}|\varphi_t(x_{i,k}) - \varphi_t(\bar{x}_k)|^2 \to 0.
\end{aligned}$$

[2]The lemma is stated as follows. Let $u_{k+1} \ge 0$ and let $u_{k+1} \le (1+\alpha_k)u_k + \beta_k$, $\sum_{k=0}^{\infty}\alpha_k < \infty$, $\sum_{k=0}^{\infty}\beta_k < \infty$. Then $u_k \to u \ge 0$.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TAC.2021.3056535, IEEE Transactions on Automatic Control

Thus, $\varphi_t(\bar{x}_k) \to \bar{\varphi}_t$.

(2). The inequality (A.16) can be re-written as

$$\frac{\alpha_k}{t}(\frac{1}{2t}-\rho)\|\bar{x}_k - s_k\|^2 \le \bar{\varphi}_{t,k} - \bar{\varphi}_{t,k+1} + b_k + \frac{L^2\alpha_k^2}{2t}. \tag{A.20}$$

Using (A.20), we have

$$\sum_{k=0}^{\infty} \frac{\alpha_k}{t}(\frac{1}{2t}-\rho)\|\bar{x}_k - s_k\|^2 \le \bar{\varphi}_{t,0} - \bar{\varphi}_t + \sum_{k=0}^{\infty} b_k + \sum_{k=0}^{\infty}\frac{L^2\alpha_k^2}{2t}.$$

Dividing both sides by $\sum_{k=0}^{\infty}\alpha_k$ yields

$$\inf_{k=1,\ldots,\infty}\|\nabla\varphi_t(\bar{x}_k)\|^2 \le \frac{2}{1-2t\rho}\frac{\bar{\varphi}_{t,0} - \bar{\varphi}_t + \sum_{k=0}^{\infty} b_k + \sum_{k=0}^{\infty}\frac{L^2\alpha_k^2}{2t}}{\sum_{k=0}^{\infty}\alpha_k}.$$

Since $b_k = \mathcal{O}(\frac{L^2\alpha_k^2}{(1-\lambda)^2})$, if $\alpha_k = \mathcal{O}(1/\sqrt{k})$, for sufficiently large T we have

$$\inf_{1\le t\le T}\|\nabla\varphi_t(\bar{x}_k)\|^2 = \mathcal{O}(\frac{\bar{\varphi}_{t,0} - \bar{\varphi}_t}{\sqrt{T}} + \frac{L^2}{(1-\lambda)^2}\cdot\frac{\log T}{\sqrt{T}}).$$

□

Before proving Theorem III.4, we need the following technical lemma.

Lemma A.1. *Given $a > 0$, $0 < 2b \le a$ and $c \ge 1$, the lower bound of the minimum value in (P_N) is given by $-\frac{1}{2}Na^2 + \frac{Nba}{c}$.*

$$\begin{aligned} \min_{x_1,\ldots,x_N} \quad & -\frac{1}{2}\sum_{i=1}^{N}(x_i^2 - 2bx_i) \\ \text{s.t.} \quad & \sum_{i=1}^{N} x_i^2 \le Na^2, \\ & 0 \le x_i \le ca, \quad \forall i. \end{aligned} \tag{P_N}$$

Proof of Lemma A.1. The dual function is given by

$$g(\lambda) := \min_{0\le x_i\le ca} -\frac{1}{2}\sum_{i=1}^{N}(x_i^2 - 2bx_i) + \lambda(\sum_{i=1}^{N}x_i^2 - Na^2),$$

where $\lambda \ge 0$. We have

$$\begin{aligned} & g(\lambda) \\ &= N\cdot\min_{0\le x\le ca}\{(\lambda - \frac{1}{2})x^2 + bx\} - \lambda Na^2 \\ &= \begin{cases} -\lambda Na^2 & \text{if} \quad \lambda \ge \frac{1}{2} - \frac{b}{ca}, \\ N\left[(\lambda - \frac{1}{2})c^2a^2 + cba\right] - \lambda Na^2 & \text{otherwise.} \end{cases} \end{aligned}$$

Note that $\frac{1}{2} - \frac{b}{ca} \ge 0$. Therefore, we have

$$\max_{\lambda\ge 0} g(\lambda) = g(\frac{1}{2} - \frac{b}{ca}) = -\frac{1}{2}Na^2 + \frac{Nba}{c}.$$

The weak duality implies the desired result. □

Proof of Theorem III.4. We prove it by induction. By the definition of e_0 and the assumptions on $k = 0$, we have $\sum_{i=1}^{N}\|x_{i,0} - x^*\|^2 \le N\gamma^{2k}e_0^2$ and $\|x_{i,0} - x^*\| \le \Gamma e_0, \forall i \in [N] := \{1,\ldots,N\}$. Assume that (III.3) and (III.4) hold for $k \ge 0$. For $k+1$, we have

$$\begin{aligned} & \sum_{i=1}^{N}\|x_{i,k+1} - x^*\|^2 \\ \le & \sum_{i=1}^{N}\|v_{i,k} - \alpha_k g_{i,k} - x^*\|^2 \\ \le & \sum_{i=1}^{N}\left(\|v_{i,k} - x^*\|^2 - 2\alpha_k\langle v_{i,k} - x^*, g_{i,k}\rangle\right) + NL^2\alpha_k^2 \\ \le & \sum_{i=1}^{N}\left(\|v_{i,k} - x^*\|^2 - 2\alpha_k(f_i(v_{i,k}) - f_i(x^*))\right. \\ & \left. + \alpha_k\rho\|v_{i,k} - x^*\|^2\right) + NL^2\alpha_k^2 \\ = & \sum_{i=1}^{N}\left(\|v_{i,k} - x^*\|^2 - 2\alpha_k(f_i(v_{i,k}) - f_i(\bar{x}_k)\right. \\ & \left. + f_i(\bar{x}_k) - f_i(x^*)) + \alpha_k\rho\|v_{i,k} - x^*\|^2\right) + NL^2\alpha_k^2 \\ \le & \sum_{i=1}^{N}\left((1+\rho\alpha_k)\|v_{i,k} - x^*\|^2 + 2L\alpha_k\|v_{i,k} - \bar{x}_k\|\right) \\ & - 2N\beta\alpha_k\|\bar{x}_k - x^*\| + NL^2\alpha_k^2, \end{aligned}$$

where the third inequality follows from the weak convexity and the last one is due to the sharpness property and Lipschitz continuity of f_i. Using the convexity of $\|\cdot\|^2$ and $\|\cdot\|$ and the stochasticity of columns of $A(k)$, we have

$$\begin{aligned} & \sum_{i=1}^{N}\|x_{i,k+1} - x^*\|^2 \\ \le & \sum_{i=1}^{N}\left((1+\rho\alpha_k)\|x_{i,k} - x^*\|^2 + 2L\alpha_k\|x_{i,k} - \bar{x}_k\|\right) \\ & - 2N\beta\alpha_k\|\bar{x}_k - x^*\| + NL^2\alpha_k^2 \\ \le & \sum_{i=1}^{N}\left((1+\rho\alpha_k)\|x_{i,k} - x^*\|^2 - 2\beta\alpha_k\|x_{i,k} - x^*\|\right) \\ & + 2(L+\beta)\alpha_k\sum_{i=1}^{N}\|\bar{x}_k - x_{i,k}\| + NL^2\alpha_k^2 \\ \le & \sum_{i=1}^{N}\left((1+\rho\mu_0)\|x_{i,k} - x^*\|^2 - 2\beta\alpha_k\|x_{i,k} - x^*\|\right) \\ & + \frac{2\sqrt{N}(L+\beta)c\|\Delta_0\|}{\lambda}\gamma^k\alpha_k \\ & + \frac{2N(L+\beta)L}{\lambda^2}(\frac{c\gamma^{1/\delta-1}}{1-\gamma^{1/\delta-1}} + \lambda)\alpha_k^2 + NL^2\alpha_k^2, \end{aligned} \tag{A.21}$$

where we use $\|\bar{x}_k - x^*\| \geq \|x_{i,k} - x^*\| - \|\bar{x}_k - x_{i,k}\|$ and $\|\cdot\|_1 \leq \sqrt{N}\|\cdot\|$ in the second inequality. The last inequality is due to (III.2). Recall the induction assumption that $\sum_{i=1}^N \|x_{i,k} - x^*\|^2 \leq Ne_0^2\gamma^{2k}$ and $\|x_{i,k} - x^*\| \leq \Gamma e_0\gamma^k$. Since

$$\mu_0 \leq \frac{e_0}{2\beta - \rho e_0}, \tag{A.22}$$

we have $\frac{2\beta\alpha_k}{1+\rho\mu_0} = \frac{2\beta\mu_0\gamma^k}{1+\rho\mu_0} \leq e_0\gamma^k$. By invoking Lemma A.1(letting $a = e_0\gamma^k$, $b = \frac{\beta\alpha_k}{1+\rho\mu_0}$ and $c = \Gamma$ in the lemma), we deduce that

$$\begin{aligned}&(1+\rho\mu_0)\sum_{i=1}^N\left(\|x_{i,k} - x^*\|^2 - \frac{2\beta\alpha_k}{(1+\rho\mu_0)}\|x_{i,k} - x^*\|\right)\\ &\leq (1+\rho\mu_0)Ne_0^2\gamma^{2k} - 2\frac{N}{\Gamma}\beta\alpha_k e_0\gamma^k.\end{aligned}$$

This, together with (A.21) yields

$$\begin{aligned}&\sum_{i=1}^N \|x_{i,k+1} - x^*\|^2\\ &\leq (1+\rho\mu_0)N(e_0\gamma^k)^2 - 2\frac{N}{\Gamma}\beta\mu_0 e_0\gamma^{2k}\\ &\quad + \frac{2\sqrt{N}(L+\beta)c\|\Delta_0\|}{\lambda}\mu_0\gamma^{2k}\\ &\quad + \frac{2N(L+\beta)L}{\lambda^2}\left(\frac{c\gamma^{1/\delta}}{1-\gamma^{1/\delta-1}} + \lambda\right)\mu_0^2\gamma^{2k} + NL^2\mu_0^2\gamma^{2k}\\ &= N\gamma^{2k}e_0^2\left(1 + \left(\rho - \frac{2\beta}{\Gamma e_0} + \frac{2(L+\beta)c\|\Delta_0\|}{\sqrt{N}\lambda e_0^2}\right)\mu_0 + \right.\\ &\quad\left. + \frac{\frac{2(L+\beta)L}{\lambda^2}\left(\frac{c\gamma^{1/\delta-1}}{1-\gamma^{1/\delta-1}} + \lambda\right) + L^2}{e_0^2}\mu_0^2\right)\\ &= N\gamma^{2k}e_0^2\left(1 - \frac{q}{e_0^2}\mu_0 + \frac{\frac{ac\gamma^{1/\delta-1}}{1-\gamma^{1/\delta-1}} + a\lambda + L^2}{e_0^2}\mu_0^2\right),\end{aligned}$$

where $a = \frac{2(L+\beta)L}{\lambda^2}, q = \frac{2\beta}{\Gamma}e_0 - \rho e_0^2 - \frac{2(L+\beta)c}{\sqrt{N}\lambda}\|\Delta_0\|$. Since $\gamma \in (0,1)$, if we have the following two conditions

1) $q > 0$

2)

$$1 > \gamma^2 \geq 1 - \frac{q\mu_0}{e_0^2} + \frac{\frac{ac\gamma^{1/\delta-1}}{1-\gamma^{1/\delta-1}} + a\lambda + L^2}{e_0^2}\mu_0^2, \tag{A.23}$$

the result follows

$$\sum_{i=1}^N \|x_{i,k+1} - x^*\|^2 \leq N\gamma^{2(k+1)}e_0^2.$$

Proof of Condition 1) Since $e_0 \leq \frac{2\beta}{\rho\Gamma}$ and

$$\|\Delta_0\| < \frac{\frac{2\beta}{\Gamma}e_0 - \rho e_0^2}{2(L+\beta)c}\lambda, \tag{A.24}$$

we have $q > 0$.

Proof of Condition 2) To ensure (A.23), it is sufficient to show

$$1 > \gamma^2 \geq 1 - \frac{q\mu_0}{10e_0^2\sqrt{N}} + \frac{\frac{ac\gamma^{1/\delta-1}}{1-\gamma^{1/\delta-1}} + a\lambda + L^2}{e_0^2}\mu_0^2, \tag{A.25}$$

for some $\gamma \in (0,1)$, which is equivalent to

$$\begin{aligned}&-\gamma^{1/\delta+1} + \gamma^2 + \left(1 - \frac{q\mu_0}{10e_0^2\sqrt{N}} + \frac{a\lambda + L^2 - ac}{e_0^2}\mu_0^2\right)\gamma^{1/\delta-1}\\ &\qquad - \left(1 - \frac{q\mu_0}{10e_0^2\sqrt{N}} + \frac{a\lambda + L^2}{e_0^2}\mu_0^2\right) \geq 0,\end{aligned}$$

if we multiply by $(1-\gamma^{1/\delta-1})$ and re-arrange the terms. Consider the function

$$\begin{aligned}&\phi(\gamma)\\ &= -\gamma^{1/\delta+1} + \left(1 - \frac{q\mu_0}{10e_0^2\sqrt{N}} + \frac{a\lambda + L^2 - ac}{e_0^2}\mu_0^2\right)\gamma^{1/\delta-1}\\ &\quad + \gamma^2 - \left(1 - \frac{q\mu_0}{10e_0^2\sqrt{N}} + \frac{a\lambda + L^2}{e_0^2}\mu_0^2\right).\end{aligned}$$

Our goal is to find $1 > \delta > 0$ such that $\phi(\lambda^\delta) \geq 0$ when $\mu_0 > 0$.

Letting $\epsilon := -\frac{q\mu_0}{10e_0^2\sqrt{N}} + \frac{a\lambda+L^2}{e_0^2}\mu_0^2$, we have $\epsilon < 0$ since $0 < \mu_0 < \frac{q}{10(a\lambda+L^2)\sqrt{N}}$ due to Assumption 7. By the same token we have

$$-\frac{ac}{e_0^2}\mu_0^2 \geq \left(\frac{1}{\Lambda} - 1\right)\epsilon, \quad \text{as} \quad \mu_0 \leq \frac{q}{10\sqrt{N}(a\lambda + L^2 + \frac{ac\Lambda}{1-\Lambda})}. \tag{A.26}$$

Therefore, if $0 < \mu_0 \leq \frac{q}{10\sqrt{N}(a\lambda+L^2+\frac{ac\Lambda}{1-\Lambda})}$, we have

$$\begin{aligned}&\phi(\lambda^\delta)\\ &= \left(1 - \lambda^{2\delta} + \epsilon - \frac{ac\mu_0^2}{e_0^2}\right)\lambda^{1-\delta} + \lambda^{2\delta} - (1+\epsilon)\\ &\geq \left(1 - \lambda^{2\delta} + \frac{1}{\Lambda}\epsilon\right)\lambda^{1-\delta} + \lambda^{2\delta} - (1+\epsilon)\\ &= (1-\lambda^{2\delta})(\lambda^{1-\delta} - 1) + \frac{1}{\Lambda}\lambda^{1-\delta}\epsilon - \epsilon.\end{aligned}$$

It is clear for every $\lambda \in (0,1)$ that

$$(1-\lambda^{2\delta})(\lambda^{1-\delta} - 1) + \frac{1}{\Lambda}\lambda^{1-\delta}\epsilon \to \frac{1}{\Lambda}\lambda\epsilon \quad \text{as} \quad \delta \to 0.$$

Therefore, there exists sufficiently small $\delta > 0$ such that $\phi(\lambda^\delta) \geq \frac{1}{\Lambda}\lambda\epsilon - \epsilon > 0$, since $\Lambda > \lambda$ and $\epsilon < 0$.

Combining (A.22), (A.26) and (A.24), we have

$$\sum_{i=1}^N \|x_{i,k+1} - x^*\|^2 \leq Ne_0^2\gamma^{2k+2},$$

if $0 < \mu_0 \leq \min\{\frac{e_0}{2\beta-\rho e_0}, \frac{q}{10\sqrt{N}(a\lambda+L^2+\frac{ac\Lambda}{1-\Lambda})}\}$, $\gamma = \lambda^\delta$ and $\|\Delta_0\| < \frac{\frac{2}{\Gamma}\beta e_0 - \rho e_0^2}{2(L+\beta)c}\lambda$.

Lastly, we need to verify (III.4) for $k+1$. Since $\|\bar{x}_{k+1}-x^*\|^2 \leq \frac{1}{N}\sum_{i=1}^N \|x_{i,k+1}-x^*\|^2 \leq e_0^2\gamma^{2k+2}$, it follows from (III.2) that

$$
\begin{aligned}
&\|x_{i,k+1}-x^*\| \\
&\leq \|x_{i,k+1}-\bar{x}_{k+1}\| + \|\bar{x}_{k+1}-x^*\| \\
&\leq \|\Delta_{k+1}\| + e_0\gamma^{k+1} \\
&\leq \left(\frac{c}{\lambda}\|\Delta_0\| + \frac{\sqrt{N}L}{\lambda^2}(c\frac{\gamma^{\frac{1}{\delta}-1}}{1-\gamma^{\frac{1}{\delta}-1}}+\lambda)\mu_0\right)\gamma^{k+1} + e_0\gamma^{k+1}
\end{aligned}
$$

Using (A.25), one has

$$
(c\frac{\gamma^{\frac{1}{\delta}-1}}{1-\gamma^{\frac{1}{\delta}-1}}+\lambda)\mu_0 < \frac{q}{10\sqrt{N}a} \leq \frac{\beta e_0}{5\sqrt{N}a}.
$$

Therefore, we have

$$
\|x_{i,k+1}-x^*\| \leq \left(\frac{c}{\lambda}\|\Delta_0\| + \frac{\beta}{10(L+\beta)}e_0\right)\gamma^{k+1} + e_0\gamma^{k+1}.
$$

Since $\|\Delta_0\| < \frac{\frac{2}{\Gamma}\beta e_0 - \rho e_0^2}{2(L+\beta)c}\lambda$, it follows that

$$
\begin{aligned}
&\|x_{i,k+1}-x^*\| \\
&\leq \left(\frac{\frac{2}{\Gamma}\beta e_0 - \rho e_0^2}{2(L+\beta)} + \frac{\beta}{10(L+\beta)}e_0\right)\gamma^{k+1} + e_0\gamma^{k+1} \\
&\leq (\frac{1}{2\Gamma}+\frac{21}{20})e_0\gamma^{k+1} \\
&\leq \Gamma e_0\gamma^{k+1},
\end{aligned}
$$

where the second inequality follows from $\beta \leq L$ and the last inequality holds since $\Gamma \geq \sqrt{2}$. □

Proof of Theorem III.5. The proof is quite similar to that of algorithm (I.2). We explain the main steps below. First, we have the same consensus lemma as Lemma II.5. Substituting $z = \hat{v}_{i,k}$ into (A.13) and taking expectation conditioned on k, we obtain

$$
\begin{aligned}
&\mathbb{E}\varphi_t(x_{i,k+1}) \\
&\leq f(\hat{v}_{i,k}) + \mathbb{E}\frac{1}{2t}\|x_{i,k+1}-\hat{v}_{i,k}\|^2 \\
&\leq f(\hat{v}_{i,k}) + \frac{1}{2t}\mathbb{E}\left(\|v_{i,k}-\hat{v}_{i,k}\|^2 - 2\alpha_k\langle v_{i,k}-\hat{v}_{i,k}, \xi_{i,k}\rangle\right. \\
&\quad \left. +\alpha_k^2\|\xi_{i,k}\|^2\right) \\
&\leq \varphi_t(v_{i,k}) - \frac{\alpha_k}{t}\langle v_{i,k}-\hat{v}_{i,k}, g_{i,k}\rangle + \frac{\alpha_k^2}{t}L^2.
\end{aligned}
\tag{A.27}
$$

Then, the remaining parts of the proof are the same as that of Theorem III.2. □

References

[1] D. Davis and D. Drusvyatskiy, "Stochastic model-based minimization of weakly convex functions," *SIAM Journal on Optimization*, vol. 29, no. 1, pp. 207–239, 2019.

[2] D. Davis, D. Drusvyatskiy, K. J. MacPhee, and C. Paquette, "Subgradient methods for sharp weakly convex functions," *Journal of Optimization Theory and Applications*, vol. 179, no. 3, pp. 962–982, 2018.

[3] S. Shahrampour, A. Rakhlin, and A. Jadbabaie, "Distributed detection: Finite-time analysis and impact of network topology," *IEEE Transactions on Automatic Control*, vol. 61, no. 11, pp. 3256–3268, 2016.

[4] S. Kar, J. M. Moura, and K. Ramanan, "Distributed parameter estimation in sensor networks: Nonlinear observation models and imperfect communication," *IEEE Transactions on Information Theory*, vol. 58, no. 6, pp. 3575–3605, 2012.

[5] N. Atanasov, R. Tron, V. M. Preciado, and G. J. Pappas, "Joint estimation and localization in sensor networks," in *53rd IEEE Conference on Decision and Control*, pp. 6875–6882, IEEE, 2014.

[6] G. Mateos and G. B. Giannakis, "Distributed recursive least-squares: Stability and performance analysis," *IEEE Transactions on Signal Processing*, vol. 60, no. 7, pp. 3740–3754, 2012.

[7] F. Bullo, J. Cortes, and S. Martinez, *Distributed control of robotic networks: a mathematical approach to motion coordination algorithms*. Princeton University Press, 2009.

[8] S. Bolognani, R. Carli, G. Cavraro, and S. Zampieri, "Distributed reactive power feedback control for voltage regulation and loss minimization," *IEEE Transactions on Automatic Control*, vol. 60, no. 4, pp. 966–981, 2014.

[9] D. Acemoglu, A. Nedic, and A. Ozdaglar, "Convergence of rule-of-thumb learning rules in social networks," in *2008 47th IEEE Conference on Decision and Control*, pp. 1714–1720, IEEE, 2008.

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

[11] D. Davis, D. Drusvyatskiy, and C. Paquette, "The nonsmooth landscape of phase retrieval," *IMA Journal of Numerical Analysis*, 2020.

[12] J. C. Duchi and F. Ruan, "Solving (most) of a set of quadratic equalities: Composite optimization for robust phase retrieval," *Information and Inference: A Journal of the IMA*, vol. 8, no. 3, pp. 471–529, 2019.

[13] Y. C. Eldar and S. Mendelson, "Phase retrieval: Stability and recovery guarantees," *Applied and Computational Harmonic Analysis*, vol. 36, no. 3, pp. 473–494, 2014.

[14] T. F. Chan and C.-K. Wong, "Total variation blind deconvolution," *IEEE transactions on Image Processing*, vol. 7, no. 3, pp. 370–375, 1998.

[15] A. Levin, Y. Weiss, F. Durand, and W. T. Freeman, "Understanding blind deconvolution algorithms," *IEEE transactions on pattern analysis and machine intelligence*, vol. 33, no. 12, pp. 2354–2367, 2011.

[16] S. Ling and T. Strohmer, "Self-calibration and biconvex compressive sensing," *Inverse Problems*, vol. 31, no. 11, p. 115002, 2015.

[17] M. Abadi, P. Barham, J. Chen, Z. Chen, and A. e. a. Davis, "Tensorflow: A system for large-scale machine learning," in *12th USENIX Symposium on Operating Systems Design and Implementation*, 2016.

[18] H. Zhang, Z. Zheng, S. Xu, W. Dai, Q. Ho, X. Liang, Z. Hu, J. Wei, P. Xie, , and E. Xing, "Poseidon: An efficient communication architecture for distributed deep learning on gpu clusters," in *USENIX Annual Technical Conference*, 2017.

[19] X. Lian, C. Zhang, H. Zhang, C.-J. Hsieh, W. Zhang, and J. Liu, "Can decentralized algorithms outperform centralized algorithms? a case study for decentralized parallel stochastic gradient descent," in *Advances in Neural Information Processing Systems*, pp. 5330–5340, 2017.

[20] A. Nedic, A. Ozdaglar, and P. A. Parrilo, "Constrained consensus and optimization in multi-agent networks," *IEEE Transactions on Automatic Control*, vol. 55, no. 4, pp. 922–938, 2010.

[21] S. Liu, Z. Qiu, and L. Xie, "Convergence rate analysis of distributed optimization with projected subgradient algorithm," *Automatica*, vol. 83, pp. 162–169, 2017.

[22] 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.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TAC.2021.3056535, IEEE Transactions on Automatic Control

[23] K. Yuan, Q. Ling, and W. Yin, "On the convergence of decentralized gradient descent," *SIAM Journal on Optimization*, vol. 26, no. 3, pp. 1835–1854, 2016.

[24] A. Nedić, A. Olshevsky, and M. G. Rabbat, "Network topology and communication-computation tradeoffs in decentralized optimization," *Proceedings of the IEEE*, vol. 106, no. 5, pp. 953–976, 2018.

[25] P. Bianchi and J. Jakubowicz, "Convergence of a multi-agent projected stochastic gradient algorithm for non-convex optimization," *IEEE transactions on automatic control*, vol. 58, no. 2, pp. 391–405, 2012.

[26] J. Zeng and W. Yin, "On nonconvex decentralized gradient descent," *IEEE Transactions on signal processing*, vol. 66, no. 11, pp. 2834–2848, 2018.

[27] M. Assran, N. Loizou, N. Ballas, and M. Rabbat, "Stochastic gradient push for distributed deep learning," in *International Conference on Machine Learning*, pp. 344–353, PMLR, 2019.

[28] G. Scutari and Y. Sun, "Distributed nonconvex constrained optimization over time-varying digraphs," *Mathematical Programming*, vol. 176, no. 1-2, pp. 497–544, 2019.

[29] J.-P. Vial, "Strong and weak convexity of sets and functions," *Mathematics of Operations Research*, vol. 8, no. 2, pp. 231–259, 1983.

[30] R. Poliquin and R. Rockafellar, "Prox-regular functions in variational analysis," *Transactions of the American Mathematical Society*, vol. 348, no. 5, pp. 1805–1838, 1996.

[31] R. Rockafellar, *Convex Analysis*. Princeton: Princeton University Press, 1970.

[32] X. Li, S. Chen, Z. Deng, Q. Qu, Z. Zhu, and A. M. C. So, "Weakly convex optimization over stiefel manifold using riemannian subgradient-type methods," *arXiv preprint arXiv:1911.05047*, 2019.

[33] J. N. Tsitsiklis, "Problems in decentralized decision making and computation.," tech. rep., Massachusetts Inst of Tech Cambridge Lab for Information and Decision Systems, 1984.

[34] R. Rockafellar and R. J.-B. Wets, *Variational analysis*, vol. 317. Springer Science & Business Media, 2009.

[35] D. Drusvyatskiy and C. Paquette, "Efficiency of minimizing compositions of convex functions and smooth maps," *Mathematical Programming*, vol. 178, no. 1-2, pp. 503–558, 2019.

[36] J.-J. Moreau, "Proximité et dualité dans un espace hilbertien," *Bulletin de la Société mathématique de France*, vol. 93, pp. 273–299, 1965.

[37] B. T. Polyak, "Minimization of unsmooth functionals," *USSR Computational Mathematics and Mathematical Physics*, vol. 9, no. 3, pp. 14–29, 1969.

[38] J.-L. Goffin, "On convergence rates of subgradient optimization methods," *Mathematical programming*, vol. 13, no. 1, pp. 329–347, 1977.

[39] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-based learning applied to document recognition," *Proceedings of the IEEE*, vol. 86, no. 11, pp. 2278–2324, 1998.

[40] C. Xi and U. A. Khan, "Distributed subgradient projection algorithm over directed graphs," *IEEE Transactions on Automatic Control*, vol. 62, no. 8, pp. 3986–3992, 2016.

[41] B. T. Polyak, "Introduction to optimization. optimization software," *Inc., Publications Division, New York*, vol. 1, 1987.

Shixiang Chen is currently a postdoctoral associate in the Department of Industrial & Systems Engineering at Texas A&M University (TAMU), with Prof. Alfredo Garcia and Prof. Shahin Shahrampour. He received Ph.D. degree in Systems Engineering and Engineering Management from The Chinese University of Hong Kong in July, 2019 and B.S. degree in Mathematics from Nanjing University in 2015. His current research interests include design and analysis of optimization algorithms, and their applications in machine learning and signal processing; non-convex optimization; the theory and methods for distributed networking.

Alfredo Garcia received the Degree in electrical engineering from the Universidad de los Andes, Bogotá,Colombia,in 1991,the Diplôme d'Etudes Approfondies in automatic control from the Université Paul Sabatier, Toulouse, France, in 1992, and the Ph.D. degree in industrial and operations engineering from the University of Michigan, Ann Arbor, MI, USA, in 1997. From 1997 to 2001, he was a consultant to government agencies and private utilities in the electric power industry. From 2001 to 2015, he was a Faculty with the Department of Systems and Information Engineering, University of Virginia, Charlottesville, VA, USA. From 2015 to 2017, he was a Professor with the Department of Industrial and Systems Engineering, University of Florida, Gainesville, FL, USA. In 2018, he joined the Department of Industrial and System Engineering, Texas A&M University, College Station, TX, USA. His research interests include game theory and dynamic optimization with applications in communications and energy networks.

Shahin Shahrampour is an Assistant Professor in the Departments of Industrial & Systems Engineering and Electrical & Computer Engineering (by courtesy) at Texas A&M University (TAMU). Before joining TAMU, he was a Postdoctoral Fellow in the School of Engineering and Applied Sciences (SEAS) at Harvard University. He received the Ph.D. degree in Electrical and Systems Engineering, the M.A. degree in Statistics (The Wharton School), and the M.S.E. degree in Electrical Engineering, all from the University of Pennsylvania, in 2015, 2014, and 2012, respectively. Prior to that, he received the B.Sc. degree from Sharif University of Technology, Iran, in 2009. His research interests include machine learning, optimization, sequential decision-making, and distributed learning, with a focus on developing computationally efficient methods for data analytics.