Distributed Optimization with Global Constraints Using Noisy Measurements

Van Sy Mai, Richard J. La, Tao Zhang, Abdella Battou

arXiv:2106.07703v1 [math.OC] 14 Jun 2021

***Abstract*— We propose a new distributed optimization algorithm for solving a class of constrained optimization problems in which (a) the objective function is separable (i.e., the sum of local objective functions of agents), (b) the optimization variables of distributed agents, which are subject to nontrivial local constraints, are coupled by global constraints, and (c) only noisy observations are available to estimate (the gradients of) local objective functions. In many practical scenarios, agents may not be willing to share their optimization variables with others. For this reason, we propose a distributed algorithm that does not require the agents to share their optimization variables with each other; instead, each agent maintains a local estimate of the global constraint functions and share the estimate only with its neighbors. These local estimates of constraint functions are updated using a consensus-type algorithm, while the local optimization variables of each agent are updated using a first-order method based on noisy estimates of gradient. We prove that, when the agents adopt the proposed algorithm, their optimization variables converge with probability 1 to an optimal point of an approximated problem based on the penalty method.**

***Index Terms*— Distributed optimization, penalty method, potential game, stochastic optimization.**

I. Introduction

Large engineered systems, such as telecommunication networks and electric power grids, are becoming more complex and often comprise subsystems that use different technologies or are controlled autonomously or by different entities. Consequently, centralized resource management or system optimization becomes increasingly impractical or impossible. This naturally calls for a distributed optimization framework that will enable distributed subsystems or agents to optimize both their local performance and, in the process, the overall system performance. What complicates the problem further is that many practical systems have global constraints (e.g., end-to-end delay requirements) that couple the decisions of more than one agent; hence satisfying these constraints will require coordination among the agents.

We study the problem of designing a distributed algorithm for a set of agents to solve a constrained optimization problem. The problem has both (a) separate *local* constraints for each agent and (b) *global* constraints that agents must satisfy together and, hence, couple the optimization variables of the agents. Moreover, the analytic expression for the global objective function or its gradient is not assumed known to every agent. For example, in many practical cases, the objective functions or constraint functions are summable (e.g., the aggregate cost of all agents). In such cases, the local objective functions of an agent may be unknown to other agents. Also, in many engineered systems, the actual costs for optimizing resource utilization need to be estimated based on measurements, which contain observation noise. Consequently, the agents must rely on *noisy* observations to update their optimization variables.

This setting applies to a wide range of real-world applications, including communication networks. For example, in the fifth-generation (5G) and future 6G systems, many heterogeneous subsystems (e.g., radio access networks vs. wired core networks) wish to minimize their own costs of delivering services and, at the same time, need to collectively assure end-to-end quality-of-service (QoS), e.g., end-to-end delays or packet loss rates, for different applications such as automated manufacturing, telesurgery, and remote controlled aerial and ground vehicles.

Since the global constraint functions depend on the optimization variables of more than one agent, their values are not always known to all agents. Thus, handling such global constraints will require the agents to exchange information. In many cases, computing and disseminating the exact values of global constraint functions to all agents will incur excessive computing and communication overheads, or cause prohibitive delays. To cope with the challenge, we adopt the penalty method and propose a new consensus-based algorithm for agents to estimate the global constraint function values in a distributed manner. We prove that, under some technical conditions, the proposed algorithm ensures almost sure convergence to an optimal point when the problem with a suitable penalty function is convex.

A. Related Literature

Distributed optimization has attracted extensive attention and accumulated a large body of literature (e.g., [3], [7], [12], [14], [18], [19] and references therein), because it has broad applications and also serves as a foundation for distributed machine learning, Here, we summarize only the most closely related studies that considered stochastic constrained optimization problems with a summable objective function, and point out the key differences between them and our study. We emphasize that this is not meant to be an exhaustive list.

Recently, consensus optimization has been an active research area. Srivastava and Nedić [17] proposed a distributed stochastic optimization algorithm for solving constrained optimization problems, in which the feasible set is assumed to

V.S. Mai, T. Zhang and A. Battou are with the National Institute of Standards and Technology (NIST). R.J. La is with NIST and the University of Maryland, College Park.

be the intersection of the feasible sets of individual agents. Each agent maintains a local copy of the global optimization variables, which are updated using a consensus-type algorithm. They showed that these local copies converge asymptotically to a common optimal point with probability 1 (w.p.1). Bianchi and Jakubowicz [2] proposed a stochastic gradient algorithm for solving a non-convex optimization problem. Every agent maintains a local copy of global optimization variables and updates its local copy using a projected stochastic gradient algorithm that requires the knowledge of the global feasible set. The agents then exchange their local copies with neighbors and update them using a consensus-type algorithm. They proved that, under the proposed algorithm, agents' local copies of optimization variables converge w.p.1 to the set of stationary or Karush-Kuhn-Tucker (KKT) points of the global objective function. In another study, Chatzipanagiotis and Zavlanos [5] proposed a distributed algorithm for solving a convex optimization problem with affine equality constraints in the presence of noise. This algorithm is based on their earlier work called accelerated distributed augmented Lagrangians [4], [6], and requires exchanges of global optimization variables among all agents.

We consider different settings in which the objective functions are separable and local optimization variables are coupled via global constraints. In order to cope with the coupling introduced by global constraints, each agent maintains local estimates of global constraint functions along with local optimization variables, and only the estimates of global constraint functions are exchanged with neighbors. Our algorithm is better suited for scenarios in which agents may not want to share their own local optimization variables with each other. For instance, autonomous systems or domains with their own private networks may not wish to reveal how they manage their networks. Furthermore, a large system likely contains subsystems that utilize different networking technologies (e.g., WiFi networks, cellular radio access networks, and wired core networks), which are managed differently. We also note that, as explained in [3], the dual problem of the constrained optimization problem we consider in this paper can be formulated as a consensus optimization problem. However, existing algorithms and results in the literature are not applicable to our settings, especially with noisy observations.

A more closely related line of research is distributed resource allocation. Many prior studies in this area focused on deterministic cases with linear resource constraints, where local objective functions are differentiable and strictly convex with Lipschitz continuous gradients that are available to the agents (e.g., [3], [7], [13]). More recently, Yi et al. [20] considered stochastic resource allocation problems with local constraints and global *affine* equality constraints in the presence of both observation and communication noise, where local constraints are determined by continuously differentiable convex functions. They proposed a new algorithm based on a primal-dual approach: each agent maintains local multipliers and auxiliary variables shared with other agents. These multipliers and auxiliary variables along with (primal) variables are updated using a combination of stochastic approximation (with decreasing step sizes) and consensus-type algorithms.

The distributed optimization problems we study can also be viewed as state-based games: Li and Marden [10] formulated the problem of designing a distributed algorithm for solving an optimization problem with affine inequality constraints as one of designing *decoupled* utility functions for distributed agents, using a penalty or barrier method. This approach leads to a state-based potential game, whose potential function is the objective function of an approximated problem, and the Nash equilibrium solves the approximated problem. Consequently, under the assumption that the analytic expressions for the objective functions are known and the exact gradients can be computed with no noise, they proved that when each agent tries to maximize its own local utility independently, their local decision variables converge to an optimal point of the approximated problem.

Our contributions: All the above studies require exact local projections by assuming simple local constraint sets. In this paper, we relax this assumption and consider a distributed optimization problem where local objective functions are convex (possibly nonsmooth) and local constraints are not assumed to be projection-friendly. We combine a stochastic subgradient method with a dynamic consensus tracking approach. Also, to deal with general local functional constraints, we adopt the idea of approximate projections from [16]; a similar approach is also used in [9] for distributed consensus optimization.

In addition, our work extends the study by Li and Marden [10] to the case where the analytic expressions for the local objective functions of agents or their gradients are unavailable and instead need to be estimated using noisy observations. We demonstrate that the agents can track the global constraint functions using a simpler dynamic consensus tracking algorithm, even for nonlinear global constraint functions. We prove that our algorithm converges to a Nash equilibrium of the aforementioned state-based potential game in [10], and no agent can increase its utility (equivalently, decrease its cost) in the state-based potential game via unilateral deviation.

Notation: Define $\mathbb{N} := \{0, 1, 2 \ldots\}$ to be set of nonnegative integers. We use x^+ to denote $\max(0, x)$. Unless stated otherwise, all vectors are column vectors and $\|\cdot\|$ denotes the ℓ_2 norm, i.e., $\|\cdot\| = \|\cdot\|_2$. Given a vector $\mathbf{x}$ or a vector function $\mathbf{f}$, we denote the kth element by x_k and f_k, respectively. The vector of zeros (resp. ones) of appropriate dimension is denoted by $\mathbf{0}$ (resp. $\mathbf{1}$). For a convex function $f : \mathcal{D} \subseteq \mathbb{R}^n \to \mathbb{R}$, we use $\partial f(\mathbf{x})$ to denote a subgradient of f at a point $\mathbf{x} \in \mathcal{D}$. When it is clear from the context, we will also use $\partial f(\mathbf{x})$ to denote the subdifferential at $\mathbf{x}$. Given a closed convex set $\mathcal{S}$ and a vector $\mathbf{x}$, $P_{\mathcal{S}}(\mathbf{x})$ denotes the projection of $\mathbf{x}$ onto the set $\mathcal{S}$ and $\mathsf{d}_{\mathcal{S}}(\mathbf{x}) := \|\mathbf{x} - P_{\mathcal{S}}(\mathbf{x})\|$ is the distance from $\mathbf{x}$ to $\mathcal{S}$. If $\mathcal{S}$ is a finite set, $|\mathcal{S}|$ denotes its cardinality. Given a sequence of random variables (RVs) or vectors $\mathbf{R}(t)$, $t \in \mathbf{N}$, we use $\mathbf{R}^t$ to denote the collection $\{\mathbf{R}(\tau) : \tau \in \{0, \ldots, t\}\}$.

II. Setup and Proposed Algorithm

We are interested in solving a constrained optimization problem of the form given below using a distributed algorithm, where (a) the objective function is separable, and (b) the global (inequality) constraints couple the optimization variables of the

agents:

$$\min_{\mathbf{y}\in\mathcal{G}} \quad \sum_{i\in\mathcal{A}} \phi_i(\mathbf{y}_i) \tag{1a}$$

$$\text{subject to} \quad \mathbf{g}(\mathbf{y}) := \sum_{i\in\mathcal{A}} \mathbf{g}_i(\mathbf{y}_i) \le \mathbf{0} \tag{1b}$$

$$\mathbf{y}_i \in \mathcal{G}_i \quad \text{for all } i \in \mathcal{A}, \tag{1c}$$

where $\mathcal{A}$ is the set of N agents, ϕ_i and $\mathbf{y}_i$ are the local objective function and the local optimization variables, respectively, of agent i, $\mathbf{y} = (\mathbf{y}_i : i \in \mathcal{A})$ is the vector of optimization variables, $\mathcal{G}_i$ is the local constraint set of agent i, $\mathcal{G} := \prod_{i\in\mathcal{A}} \mathcal{G}_i$ is the feasible set, and $\mathbf{g} = (g_1, \ldots, g_K)$ is the vector consisting of K global constraint functions.

We assume $\mathcal{G}_i = \mathcal{G}_{i0} \bigcap \tilde{\mathcal{G}}_i$, where $\mathcal{G}_{i0}$ is a nonempty convex set assumed to be projection friendly (e.g., box, simplex, ball constraints), and the set $\tilde{\mathcal{G}}_i$ is given by $\tilde{\mathcal{G}}_i = \{\mathbf{y}_i \mid c_{ik}(\mathbf{y}_i) \le 0,\ k \in \tilde{\mathcal{K}}_i\}$, where c_{ik}'s are local constraint functions, and $\tilde{\mathcal{K}}_i$ is the set of agent i's local (inequality) constraints. We are interested in scenarios in which the constraint set $\tilde{\mathcal{K}}_i$ is large or $\tilde{\mathcal{G}}_i$ is not projection friendly. Here, $\mathcal{G}_{i0}$ can be regarded as a hard constraint while $\tilde{\mathcal{G}}_i$ represents soft constraints.

Unlike in [5], [10], we do not assume that the constraint functions are affine. Instead, we assume that each agent i knows its contributions to global constraint function given by $\mathbf{g}_i$, but not $\mathbf{g}_j$, $j \in \mathcal{A} \setminus \{i\}$. This assumption is reasonable in many practical problems with distributed decision makers.

A popular approach to constrained optimization is the penalty method, which adds a penalty to the objective function when one or more constraints are violated. Although a general penalty function can be used in our problem, we assume a specific penalty function to facilitate our exposition. Consider the following approximated problem with a penalty function:

$$\min_{\mathbf{y}\in\mathcal{G}} \quad \sum_{i\in\mathcal{A}} \phi_i(\mathbf{y}_i) + \frac{\mu}{2N} \sum_{k\in\mathcal{K}_G} g_k^+(\mathbf{y})^2 =: \Phi(\mathbf{y}), \tag{2}$$

where $\mu > 0$ is a penalty parameter, $\mathcal{K}_G$ is the set of K global constraints, and the second term in Φ is the penalty function. The original problem in (1) is recovered by letting $\mu \to \infty$. Throughout the paper, we assume that μ is fixed and known to all the agents. The function Φ is convex on $\mathcal{G}$ provided that ϕ_i and $\mathbf{g}_i$, $i \in \mathcal{A}$, are convex.

The usual gradient projection method does not lead to a distributed algorithm because the penalty function and its gradient are coupled. Furthermore, we are interested in scenarios in which (a) the analytic expression of local objective function ϕ_i and its gradient are unknown to agent i and (b) only noisy measurements are available to estimate them. To address these issues, we propose a new algorithm that requires each agent i to keep an estimate $\mathbf{e}_i(t)$ of the constraint functions $\mathbf{g}(\mathbf{y}(t))$, which are updated via dynamic consensus tracking. These estimates are used to approximate the gradient of the penalty function and are updated using a consensus-type algorithm.

A. Proposed Algorithm

Every agent $i \in \mathcal{A}$ will first randomly choose its initial local optimization variables $\mathbf{y}_i(0))$ in $\mathcal{G}_{i0}$ and its estimate of (average) global constraint functions $\mathbf{e}_i(0) = \mathbf{g}_i(\mathbf{y}_i(0))$ based on its own contributions. Subsequently, at each iteration, it updates them according to the following algorithm:

$$\mathbf{z}_i(t) = P_{\mathcal{G}_{i0}}\Big(\mathbf{y}_i(t) - \gamma_t\big(\mathbf{q}_i(t) + \mathbf{V}_i(t)\big)\Big) \tag{3a}$$

$$\mathbf{y}_i(t+1) = P_{\mathcal{G}_{i0}}\Big(\mathbf{z}_i(t) - \beta_{it} \sum_{k\in\mathcal{K}_{it}} \frac{c_{ik}^+(\mathbf{z}_i(t))}{\|\mathbf{d}_{ik}\|^2}\mathbf{d}_{ik}\Big) \tag{3b}$$

$$\mathbf{e}_i(t+1) = \sum_{j\in\mathcal{A}} w_{ij}(t)\mathbf{e}_j(t) + \mathbf{g}_i(\mathbf{y}_i(t+1)) - \mathbf{g}_i(\mathbf{y}_i(t)) \tag{3c}$$

where

$$\mathbf{q}_i(t) = \partial\phi_i(\mathbf{y}_i(t)) + \mu \sum_{k\in\mathcal{K}_G} \partial g_{ik}(\mathbf{y}_i(t)) e_{ik}^+(t), \tag{4}$$

$\mathbf{V}_i(t)$ are random vectors, γ_t is a step size at iteration t,[1] $\mathcal{K}_{it} \subset \tilde{\mathcal{K}}_i$ in (3b) is a random set of indices selected by agent i independently at each iteration t,[2] and $\mathbf{d}_{ik}$ is a subgradient of c_{ik} at $\mathbf{z}_i(t)$ when $c_{ik}(\mathbf{z}_i(t)) > 0$ and $\mathbf{d}_{ik} = \mathbf{d} \neq \mathbf{0}$ otherwise. The value of $\mathbf{d}$ is unimportant since the summand is equal to zero in this case.

For simplicity, we assume that $\mathcal{K}_{it}$ is chosen according to a uniform distribution $\mathcal{U}_i$ over the set $\mathcal{S}_i := \{S \subset \tilde{\mathcal{K}}_i \mid |S| = s_i\}$ for some $s_i \ge 1$, and take any β_{it} in some $[\underline{\beta}_i, \bar{\beta}_i] \subset (0, 2/s_i)$ for all iterations t. In practice, we may choose $s_i \ll |\tilde{\mathcal{K}}_i|$ to limit the number of constraints considered in (3b) (even $s_i = 1$). We assume that agent i uses the same weight β_{it} for all chosen local constraints in (3b). But, it can employ different weights for different constraints which, for example, depend on the value of constraint functions. We assume that $\mathcal{G}_{i0}$ is simple enough so that $P_{\mathcal{G}_{i0}}(\cdot)$ can be evaluated efficiently.

Note that agent $i \in \mathcal{A}$ needs to exchange its estimate $\mathbf{e}_i$ only with its neighbors reflected in the weight matrix $W(t) = [w_{ij}(t) : i, j \in \mathcal{A}]$, which is allowed to be time-varying. This update rule is designed based on a dynamic consensus algorithm (e.g., [8]) in order to track the average of the global constraint functions. As a result, we can view $\mathbf{e}_i(t)$ as a local estimate of the average global constraint functions.

We note that, in our setting, the subgradient $\partial\phi_i(\mathbf{y}_i(t))$ in (4) is unavailable to agent i; instead, only the second term in (4) can be computed by agent i and a noisy estimate of the subgradient given by the sum $\partial\phi_i(\mathbf{y}_i(t)) + \mathbf{V}_i(t)$ is available to the agent for the update in (3a).

The random projection step in (3b) is borrowed from [16] (see also [11]), but we do not restrict to selecting only a single constraint at each iteration. The idea is that if the constraint c_{ik} is violated at $\mathbf{z}_i(t)$ for some randomly selected k, i.e., $c_{ik}(\mathbf{z}_i(t)) > 0$, in order to reduce this violation, the algorithm updates the optimization variables in the direction of $-\mathbf{d}_{ik}$ with a step size $\beta_{it}\frac{c_{ik}(\mathbf{z}_i(t))}{\|\mathbf{d}_{ik}\|^2}$ according to (3b). This projection step, though simple, is beneficial in dealing with nontrivial and computationally intractable constraints, including linear matrix inequalities and robust linear inequalities (e.g., [16]).

[1] We assume that the agents adopt a common step size sequence, for example, based on a global clock.

[2] Note that one could replace the set of constraints $c_{ik}(\mathbf{y}_i) \le 0$, $k \in \tilde{\mathcal{K}}_i$ with one constraint $\bar{c}_i(\mathbf{y}_i) \le 0$ where $\bar{c}_i(\mathbf{y}) = \max_{k\in\tilde{\mathcal{K}}_i} c_{ik}(\mathbf{y}_i)$. Computationally, this, however, requires evaluations of all the constraint functions, which may be impractical when $|\tilde{\mathcal{K}}_i|$ is large.

For each $t \in \mathbb{N}$, we define $\mathcal{F}_t$ to be the σ-field generated by $\{\mathbf{y}(0), \mathbf{V}^{t-1}, (\mathcal{K}_i^{t-1} : i \in \mathcal{A})\}$, where $\mathcal{K}_i^{t-1} = \{\mathcal{K}_{is} : 0 \le s < t\}$ and $\mathcal{K}_{is}$, $s \in \mathbb{N}$, is the random index set of the constraints chosen by agent i in (3b) at iteration s. Throughout the paper, we use $\mathbb{E}_t[\cdot]$ to denote the conditional expectation $\mathbb{E}[\cdot \mid \mathcal{F}_t]$.

B. Assumptions

We impose the following assumptions on the optimization problem in (2).

Assumption 1. *Problem* (2) *satisfies the following:*

a. *The local constraint sets $\mathcal{G}_i$, $i \in \mathcal{A}$, are nonempty and convex, and $\mathcal{G}_{i0}$ are closed and convex.*
b. *The functions ϕ_i and g_{ik}, $i \in \mathcal{A}$ and $k \in \mathcal{K}_G$, are convex and L-Lipschitz continuous on $\mathcal{G}_{i0}$ for some $L > 0$. Moreover, Φ is L-Lipschitz continuous on $\prod_{i\in\mathcal{A}} \mathcal{G}_{i0}$.*[3]
c. *The local constraint functions c_{ik}, $i \in \mathcal{A}$ and $k \in \tilde{\mathcal{K}}_i$, are convex and also L-Lipschitz continuous on $\mathcal{G}_{i0}$. Moreover, there exists $C > 0$ such that, for all $i \in \mathcal{A}$,*

$$\mathrm{d}^2_{\mathcal{G}_i}(\mathbf{z}) \le C\, \mathbb{E}_{\mathcal{K}\sim\mathcal{U}_i}\left[\sum_{k\in\mathcal{K}} c_{ik}^+(\mathbf{z})^2\right], \quad \mathbf{z} \in \mathcal{G}_{i0}, \tag{5}$$

where $\mathbb{E}_{\mathcal{K}\sim\mathcal{U}_i}[\cdot]$ denotes the expectation when the random set $\mathcal{K}$ is chosen according to $\mathcal{U}_i$ over the set $\mathcal{S}_i$.
d. *The optimal set of* (2)*, denoted by $\mathcal{Y}^*$, is nonempty.*

Here, without loss of generality we assume the same Lipschitz constant L. Note that we do not assume differentiability of any involved functions or compactness of constraint sets. Assumption 1-c, in particular condition (5), is known to be quite general and plays an important role in the convergence analysis of algorithms involving random projection in (3b); see [11] for a further discussion on this assumption as well as sufficient conditions for (5) to hold.

We allow the distribution of the perturbation $\mathbf{V}(t) := (\mathbf{V}_i(t) : i \in \mathcal{A})$ to depend on the optimization variables $\mathbf{y}(t)$. The distribution of $\mathbf{V}(t)$ when $\mathbf{y}(t) = \mathbf{y}$ is denoted by $\mu_{\mathbf{y}}$.

Assumption 2. *The perturbation satisfies (i) $\int \mathbf{v}\, \mu_{\mathbf{y}}(d\mathbf{v}) = \mathbf{0}$ for all $\mathbf{y} \in \mathcal{G}$ and (ii) $\sup_{\mathbf{y}\in\mathcal{G}} \int \|\mathbf{v}\|^2\, \mu_{\mathbf{y}}(d\mathbf{v}) =: \nu < \infty$.*

For each $t \in \mathbb{N}$, define $\mathcal{E}_t = \{(i,j) \in \mathcal{A}\times\mathcal{A} : w_{ij}(t) > 0\}$.

Assumption 3. *There exists $Q \in \mathbb{N}$ such that, for all $k \ge 0$, the graph $(\mathcal{A}, \bigcup_{l=1}^{Q} \mathcal{E}_{k+l})$ is strongly connected (i.e., Q-strongly connected). In addition, $W(t)$ is doubly stochastic for all $t \in \mathbb{N}$, and there exists $w_{\min} > 0$ such that, for all $t \in \mathbb{N}$, all nonzero weights $w_{ij}(t)$ lie in $[w_{\min}, 1]$.*

We will use below the following standard assumptions on the step sizes.

Assumption 4. *Steps sizes γ_t, $t \in \mathbb{N}$, are positive and satisfy (i) $\sum_{t\in\mathbb{N}} \gamma_t = \infty$ and (ii) $\sum_{t\in\mathbb{N}} \gamma_t^2 < \infty$.*

[3]This assumption holds, for example, when $\mathcal{G}_{i0}$, $i \in \mathcal{A}$, are compact.

C. Preliminaries

Let us first introduce some notation. For every $t \in \mathbb{N}$, define $\mathbf{e}(t) := (\mathbf{e}_i(t) : i \in \mathcal{A})$, $\mathbf{g}(t) := (\mathbf{g}_i(\mathbf{y}_i(t)) : i \in \mathcal{A})$,

$$\bar{\mathbf{e}}(t) := \frac{1}{N}\sum_{i\in\mathcal{A}} \mathbf{e}_i(t), \quad \text{and} \quad \bar{\mathbf{g}}(t) := \frac{1}{N}\mathbf{g}(\mathbf{y}(t)). \tag{6}$$

The following result is a direct consequence of the assumption that weight matrices $W(t)$, $t \in \mathbb{N}$, are doubly stochastic.

Lemma 1. *Under Assumption 3, the following holds:*

$$\bar{\mathbf{e}}(t) = \bar{\mathbf{g}}(t), \quad t \in \mathbb{N}. \tag{7}$$

We will also use the following result on the mixing property of a sequence of doubly stochastic weight matrices [19, Theorem 4.2].

Lemma 2. *Consider the following iterates for all $i \in \mathcal{A}$.*

$$\boldsymbol{\theta}_i(t+1) = \textstyle\sum_{j\in\mathcal{A}} w_{ij}(t)\boldsymbol{\theta}_j(t) + \boldsymbol{\varepsilon}_i(t+1), \quad t \in \mathbb{N},$$

where $\boldsymbol{\varepsilon}_i(t) \in \mathbb{R}^K$ is arbitrary, and $W(t)$ satisfies Assumption 3. Then, there exist $C_1 > 0$ and $C_2 > 0$ such that

$$\begin{aligned}\|\boldsymbol{\theta}_i(t) - \bar{\boldsymbol{\theta}}(t)\| \le{}& C_1 N^{-1}\sigma^t + N^{-1}\textstyle\sum_{j\in\mathcal{A}} \|\boldsymbol{\varepsilon}_j(t)\| + \|\boldsymbol{\varepsilon}_i(t)\| \\ &+ C_2 N^{-1} \textstyle\sum_{s=1}^{t-1} \sigma^{t-s} \sum_{j\in\mathcal{A}} \|\boldsymbol{\varepsilon}_j(s)\|,\end{aligned}$$

where $\bar{\boldsymbol{\theta}}(t) = \sum_{j\in\mathcal{A}} \boldsymbol{\theta}_j(t)/N$, and $\sigma = \left(1 - \frac{w_{\min}}{4N^2}\right)^{1/Q}$.

The lemma states that, over a larger timescale, the Q-strong connectivity has a similar mixing effect as a static, strongly connected graph.

The following results will be used frequently in our proofs.

Lemma 3. *For any $a, b \in \mathbb{R}$ and any $\eta > 0$,*

$$2ab \le \eta a^2 + \eta^{-1} b^2. \tag{8}$$

Lemma 4. *Under Assumptions 1 and 3, for any $i \in \mathcal{A}$,*

$$\|\mathbf{q}_i(t)\|^2 \le 2L^2\left(1 + \mu^2 K \|\mathbf{e}_i(t) - \bar{\mathbf{e}}(t)\|^2\right), \quad t \in \mathbb{N}. \tag{9}$$

Proof. We drop the dependence on t for notational simplicity. It follows from (4) and (7) that

$$\begin{aligned}\|\mathbf{q}_i\| &= \|\partial_{\mathbf{y}_i}\Phi(\mathbf{y}) + \mu \sum_{k\in\mathcal{K}_G} \partial g_{ik}(\mathbf{y}_i)(e_{ik}^+ - \bar{e}_k^+)\| \\ &\le \|\partial_{\mathbf{y}_i}\Phi(\mathbf{y})\| + \mu \sum_{k\in\mathcal{K}_G} \|\partial g_{ik}(\mathbf{y}_i)\| \cdot |e_{ik}^+ - \bar{e}_k^+| \\ &\le L(1 + \mu\|\mathbf{e}_i(t) - \bar{\mathbf{e}}(t)\|_1),\end{aligned}$$

where we used the Lipschitz continuity of Φ, g_{ik} and the function $\max(\cdot, 0)$. Thus,

$$\begin{aligned}\|\mathbf{q}_i(t)\|^2 &\le 2L^2\left(1 + \mu^2\|\mathbf{e}_i(t) - \bar{\mathbf{e}}(t)\|_1^2\right) \\ &\le 2L^2\left(1 + \mu^2 K\|\mathbf{e}_i(t) - \bar{\mathbf{e}}(t)\|^2\right)\end{aligned} \tag{10}$$

which follows from the equivalence of norms. □

We also make use of the following classical results on the convergence of a sequence of nonnegative RVs.

Lemma 5. *[15, Lemma 10, p. 49] Let $\{v_t : t \in \mathbb{N}\}$ be a sequence of nonnegative RVs with $\mathbb{E}[v_0] < \infty$. Suppose $\{\alpha_t : t \in \mathbb{N}\}$ and $\{\beta_t : t \in \mathbb{N}\}$ are deterministic scalar sequences*

satisfying $\alpha_t \in [0,1]$, $\beta_t > 0$, $\sum_{t\in\mathbf{N}} \alpha_t = \infty$, $\sum_{t\in\mathbf{N}} \beta_t < \infty$, $\lim_{t\to\infty} \beta_t/\alpha_t = 0$, *and*

$$\mathbb{E}\left[v_{t+1}|v^t\right] \le (1-\alpha_t)v_t + \beta_t \quad \textit{w.p.1 for all } t \in \mathbf{N},$$

where $v^t = \{v_s : 0 \le s \le t\}$. *Then,* v_t *converges to 0 w.p.1, and* $\lim_{t\to\infty} \mathbb{E}\left[v_t\right] = 0$.

Lemma 6. *[15, Lemma 11, p. 50] Let* $\{v_t : t \in \mathbf{N}\}, \{u_t : t \in \mathbf{N}\}$, $\{\alpha_t : t \in \mathbf{N}\}$ *and* $\{\beta_t : t \in \mathbf{N}\}$ *be sequences of nonnegative RVs satisfying the following conditions w.p.1:* $\sum_{t\in\mathbf{N}} \alpha_t < \infty$, $\sum_{t\in\mathbf{N}} \beta_t < \infty$, *and*

$$\mathbb{E}\left[v_{t+1}|v^t, \alpha^t, \beta^t, u^t\right] \le (1+\alpha_t)v_t - u_t + \beta_t \quad \textit{for all } t \in \mathbf{N}.$$

Then, w.p.1, (i) $\sum_{t\in\mathbf{N}} u_t < \infty$ *and (ii)* v_t *converges to some nonnegative RV* v.

III. Convergence Analysis

In this section, we demonstrate that when all agents update their local optimization variables ($\mathbf{y}_i(t)$) and local estimates of global constraint functions ($\mathbf{e}_i(t)$) using our proposed algorithm in (3), $\mathbf{y}(t)$ converges to the optimal set $\mathcal{Y}^*$ w.p.1. (Theorem 3). We prove this result with the help of a series of auxiliary results.

First, the projection in (3b) satisfies the following inequality, the proof of which is inspired by [9], [11], [16].

Lemma 7. *Under Assumption 1, for any* $\mathbf{y}_i \in \mathcal{G}_i$, *we have*

$$\|\mathbf{y}_i(t+1) - \mathbf{y}_i\|^2 \le \|\mathbf{z}_i(t) - \mathbf{y}_i\|^2 - \tilde{\beta}_{it} \sum_{k\in\mathcal{K}_{it}} c_{ik}^+(\mathbf{z}_i(t))^2 \quad (11a)$$

$$c_{ik}^+(\mathbf{z}_i(t))^2 \ge \tfrac{\tau-1}{\tau} c_{ik}^+(\mathbf{y}_i(t))^2 - 2\tau\gamma_t^2 L^2\big(2L^2 + \|\mathbf{V}_i(t)\|^2\big) - 4\tau\gamma_t^2 L^4 \mu^2 K \|\mathbf{e}_i(t) - \bar{\mathbf{e}}(t)\|^2 \quad (11b)$$

for all $i \in \mathcal{A}$, $k \in \mathcal{K}$, *and* $\tau > 0$, *where* $\tilde{\beta}_{it} = \frac{(2-s_i\beta_{it})\beta_{it}}{L^2}$.

Proof. The proof of (11a) is similar to those given in [11], [16], but we provide it here for completeness. For any $\mathbf{y}_i \in \mathcal{G}_i$,

$$\begin{aligned}\|\mathbf{y}_i(t+1) - \mathbf{y}_i\|^2 &\le \Big\|\mathbf{z}_i(t) - \mathbf{y}_i - \beta_{it} \textstyle\sum_{k\in\mathcal{K}_{it}} \frac{c_{ik}^+(\mathbf{z}_i(t))}{\|\mathbf{d}_{ik}\|^2}\mathbf{d}_{ik}\Big\|^2 \\ &= \|\mathbf{z}_i(t) - \mathbf{y}_i\|^2 + \beta_{it}^2 \Big\| \textstyle\sum_{k\in\mathcal{K}_{it}} \frac{c_{ik}^+(\mathbf{z}_i(t))}{\|\mathbf{d}_{ik}\|^2}\mathbf{d}_{ik}\Big\|^2 \\ &\quad + 2\beta_{it} \textstyle\sum_{k\in\mathcal{K}_{it}} (\mathbf{y}_i - \mathbf{z}_i(t))^T \mathbf{d}_{ik} \frac{c_{ik}^+(\mathbf{z}_i(t))}{\|\mathbf{d}_{ik}\|^2}, \end{aligned} \quad (12)$$

where the first inequality follows from (3b) and the nonexpansiveness property of projection. Since c_{ik}^+ is a convex function and $\mathbf{d}_{ik} \in \partial c_{ik}^+(\mathbf{z}_i(t))$, it follows that

$$(\mathbf{y}_i - \mathbf{z}_i(t))^T \mathbf{d}_{ik} \le c_{ik}^+(\mathbf{y}_i) - c_{ik}^+(\mathbf{z}_i(t)) = -c_{ik}^+(\mathbf{z}_i(t)),$$

where we used $c_{ik}^+(\mathbf{y}_i) = 0$ for all $\mathbf{y}_i \in \mathcal{G}_i$. Moreover, by Cauchy-Schwarz inequality

$$\Big\| \textstyle\sum_{k\in\mathcal{K}_{it}} \frac{c_{ik}^+(\mathbf{z}_i(t))}{\|\mathbf{d}_{ik}\|^2}\mathbf{d}_{ik}\Big\|^2 \le s_i \textstyle\sum_{k\in\mathcal{K}_{it}} \frac{c_{ik}^+(\mathbf{z}_i(t))^2}{\|\mathbf{d}_{ik}\|^2}.$$

As a result, using the above bounds in (12),

$$\begin{aligned}&\|\mathbf{y}_i(t+1) - \mathbf{y}_i\|^2 \\ &\le \|\mathbf{z}_i(t) - \mathbf{y}_i\|^2 + (s_i\beta_{it}^2 - 2\beta_{it}) \textstyle\sum_{k\in\mathcal{K}_{it}} \frac{c_{ik}^+(\mathbf{z}_i(t))^2}{\|\mathbf{d}_{ik}\|^2},\end{aligned}$$

and (11a) now follows from Assumption 1-c.

For the second inequality in (11b), note that

$$\begin{aligned}c_{ik}^+(\mathbf{z}_i(t))^2 &\ge c_{ik}^+(\mathbf{y}_i(t))^2 + 2(c_{ik}^+(\mathbf{z}_i(t)) - c_{ik}^+(\mathbf{y}_i(t)))c_{ik}^+(\mathbf{y}_i(t)) \\ &\ge c_{ik}^+(\mathbf{y}_i(t))^2 - 2|c_{ik}^+(\mathbf{z}_i(t)) - c_{ik}^+(\mathbf{y}_i(t))|c_{ik}^+(\mathbf{y}_i(t)) \\ &\ge \tfrac{\tau-1}{\tau} c_{ik}^+(\mathbf{y}_i(t))^2 - \tau|c_{ik}^+(\mathbf{z}_i(t)) - c_{ik}^+(\mathbf{y}_i(t))|^2 \end{aligned} \quad (13)$$

for any $\tau > 0$, where we used (8). In addition,

$$\begin{aligned}&|c_{ik}^+(\mathbf{z}_i(t)) - c_{ik}^+(\mathbf{y}_i(t))|^2 \\ &\le L^2\|\mathbf{z}_i(t) - \mathbf{y}_i(t)\|^2 \qquad \text{(Lipschitz continuity)} \\ &\le L^2\gamma_t^2\|\mathbf{q}_i(t) + \mathbf{V}_i(t)\|^2 \qquad \text{(projection property)} \\ &\le 2L^2\gamma_t^2(\|\mathbf{q}_i(t)\|^2 + \|\mathbf{V}_i(t)\|^2) \\ &\le 4L^2\gamma_t^2\big(L^2(1+\mu^2K\|\mathbf{e}_i(t) - \bar{\mathbf{e}}(t)\|^2) + \tfrac{\|\mathbf{V}_i(t)\|^2}{2}\big). \quad \text{(by (9))}\end{aligned}$$

Substituting this bound in (13) gives us (11b). □

Although our result is similar to those in [9], [11], it contains an extra term $\|\mathbf{e}_i(t) - \bar{\mathbf{e}}(t)\|^2$ in (11b), which measures the disagreements among the agents' local estimates of the global constraint functions. Roughly speaking, for the algorithm to converge, it is necessary that this term decays to 0. Much of the challenge lies in properly bounding this term, which is not bounded a priori. To this end, let us define

$$a_t := \textstyle\sum_{i\in\mathcal{A}} \|\mathbf{e}_i(t) - \bar{\mathbf{e}}(t)\|^2, \quad t \in \mathbf{N}. \quad (14)$$

The following result relates a_t with $\|\mathbf{y}(t)$–$\mathbf{w}\|^2$ and $\Phi(\mathbf{w})$–$\Phi\big(P_{\mathcal{G}}(\mathbf{y}(t))\big)$ for any $\mathbf{w} \in \mathcal{G}$ and any choice of step size γ_t.

Theorem 1. *Under Assumptions 1 and 2, the following holds for any* $t \in \mathbf{N}$, $\mathbf{w} \in \mathcal{G}$, *and positive sequence* $\{\kappa_t : t \in \mathbf{N}\}$:

$$\begin{aligned}&\mathbb{E}_t\left[\|\mathbf{y}(t+1) - \mathbf{w}\|^2\right] \\ &\le (1+\kappa_t K)\|\mathbf{y}(t) - \mathbf{w}\|^2 + \gamma_t^2 D_1 + \gamma_t^2\big(\mu^2L^2\kappa_t^{-1} + D_2\big)a_t \\ &\quad + 2\gamma_t\big[\Phi(\mathbf{w}) - \Phi\big(P_{\mathcal{G}}(\mathbf{y}(t))\big)\big] - 2\rho\mathsf{d}_{\mathcal{G}}^2(\mathbf{y}(t)),\end{aligned} \quad (15)$$

where $\rho = \frac{\tilde{\beta}}{4C}$ *with* $\tilde{\beta} = \inf_{i,t} \tilde{\beta}_{it}$, $D_1 = \frac{L^2}{\rho} + 10(2L^2N + \nu)$ *and* $D_2 = 20L^2K\mu^2$.

Proof. First, for any $\mathbf{w}_i \in \mathcal{G}_i$, by Lemma 7, we have

$$\|\mathbf{y}_i(t+1) - \mathbf{w}_i\|^2 \le \|\mathbf{z}_i(t) - \mathbf{w}_i\|^2 - \tilde{\beta}_{it} \textstyle\sum_{k\in\mathcal{K}_{it}} c_{ik}^+(\mathbf{z}_i(t))^2.$$

By the nonexpansive property of projection, the first term on the RHS can be expanded further as

$$\begin{aligned}\|\mathbf{z}_i(t) - \mathbf{w}_i\|^2 &\le \big\|\mathbf{y}_i(t) - \gamma_t\big(\mathbf{q}_i(t) + \mathbf{V}_i(t)\big) - \mathbf{w}_i\big\|^2 \\ &= \|\mathbf{y}_i(t) - \mathbf{w}_i\|^2 + \gamma_t^2\|\mathbf{q}_i(t) + \mathbf{V}_i(t)\|^2 \\ &\quad + 2\gamma_t\big(\mathbf{q}_i(t) + \mathbf{V}_i(t)\big)^{\mathsf{T}}(\mathbf{w}_i - \mathbf{y}_i(t)).\end{aligned}$$

Thus, we have

$$\begin{aligned}&\|\mathbf{y}_i(t+1) - \mathbf{w}_i\|^2 \\ &\le \|\mathbf{y}_i(t) - \mathbf{w}_i\|^2 - \tilde{\beta}_{it} \textstyle\sum_{k\in\mathcal{K}_{it}} c_{ik}^+(\mathbf{z}_i(t))^2 \\ &+ 2\gamma_t\big(\partial_{\mathbf{y}_i}\Phi(\mathbf{y}(t)) + \mathbf{V}_i(t)\big)^{\mathsf{T}}(\mathbf{w}_i - \mathbf{y}_i(t)) \\ &+ 2\gamma_t\mu \textstyle\sum_{k\in\mathcal{K}_G} \big|e_{ik}^+(t) - \bar{e}_k^+(t)\big| \cdot \big|\partial g_{ik}(\mathbf{y}_i(t))^{\mathsf{T}}(\mathbf{w}_i - \mathbf{y}_i(t))\big| \\ &+ \gamma_t^2\|\mathbf{q}_i(t) + \mathbf{V}_i(t)\|^2.\end{aligned} \quad (16)$$

We bound the last two terms in (16) as follows. First,

$$\begin{aligned}&\gamma_t^2\|\mathbf{q}_i(t) + \mathbf{V}_i(t)\|^2 \le 2\gamma_t^2\big(\|\mathbf{q}_i(t)\|^2 + \|\mathbf{V}_i(t)\|^2\big) \\ &\le 2\gamma_t^2\big(2L^2\mu^2K\|\mathbf{e}_i(t) - \bar{\mathbf{e}}(t)\|^2 + 2L^2 + \|\mathbf{V}_i(t)\|^2\big),\end{aligned} \quad (17)$$

where the second inequality follows from Lemma 4. Second,

$$\begin{aligned}
&2\gamma_t\mu\textstyle\sum_{k\in\mathcal{K}_G}\big|e_{ik}^+(t)-\bar{e}_k^+(t)\big|\cdot\big|\partial g_{ik}(\mathbf{y}_i(t))^{\mathsf{T}}(\mathbf{w}_i-\mathbf{y}_i(t))\big|\\
&\le\textstyle\sum_{k\in\mathcal{K}_G}2\gamma_t\mu L\big|e_{ik}(t)-\bar{e}_k(t)\big|\|\mathbf{w}_i-\mathbf{y}_i(t)\|\\
&\le\textstyle\sum_{k\in\mathcal{K}_G}\big(\gamma_t^2\mu^2L^2\kappa_t^{-1}|e_{ik}(t)-\bar{e}_k(t)|^2+\kappa_t\|\mathbf{w}_i-\mathbf{y}_i(t)\|^2\big)\\
&\le\gamma_t^2\mu^2L^2\kappa_t^{-1}\|\mathbf{e}_i(t)-\bar{\mathbf{e}}(t)\|^2+\kappa_tK\|\mathbf{w}_i-\mathbf{y}_i(t)\|^2 \qquad (18)
\end{aligned}$$

for any $\kappa_t>0$, where we used the Lipschitz continuity of g_{ik} in the first inequality, and (8) of Lemma 3 in the second inequality.

Using the bounds of (17) and (18) in (16) and then summing over $i\in\mathcal{A}$ gives us

$$\begin{aligned}
&\|\mathbf{y}(t+1)-\mathbf{w}\|^2\\
&\le(1+\kappa_tK)\|\mathbf{y}(t)-\mathbf{w}\|^2-\textstyle\sum_{i\in\mathcal{A}}\tilde{\beta}_{it}\sum_{k\in\mathcal{K}_{it}}c_{ik}^+(\mathbf{z}_i(t))^2\\
&\quad+4NL^2\gamma_t^2+2\gamma_t^2\|\mathbf{V}(t)\|^2+\gamma_t^2\mu^2(L^2\kappa_t^{-1}+4L^2K)a_t\\
&\quad+2\gamma_t\big(\partial\Phi(\mathbf{y}(t))+\mathbf{V}(t)\big)^{\mathsf{T}}(\mathbf{w}-\mathbf{y}(t)). \qquad (19)
\end{aligned}$$

Now, using the convexity of Φ and Lemma 3, we have

$$\begin{aligned}
&2\gamma_t\partial\Phi(\mathbf{y}(t))^{\mathsf{T}}(\mathbf{w}-\mathbf{y}(t))\le2\gamma_t\big(\Phi(\mathbf{w})-\Phi(\mathbf{y}(t))\big)\\
&=2\gamma_t\big(\Phi(\mathbf{w})-\Phi(\mathbf{u})+\Phi(\mathbf{u})-\Phi(\mathbf{y}(t))\big),\quad\forall\mathbf{u}\in\mathcal{G}\\
&\le2\gamma_t\big(\Phi(\mathbf{w})-\Phi(\mathbf{u})\big)+2\gamma_tL\|\mathbf{u}-\mathbf{y}(t)\|\\
&\le2\gamma_t\big(\Phi(\mathbf{w})-\Phi(\mathbf{u})\big)+\gamma_t^2L^2\rho^{-1}+\rho\|\mathbf{u}-\mathbf{y}(t)\|^2
\end{aligned}$$

for any $\rho>0$. Taking $\mathbf{u}=P_{\mathcal{G}}(\mathbf{y}(t))$ yields

$$\begin{aligned}
&2\gamma_t\partial\Phi(\mathbf{y}(t))^{\mathsf{T}}(\mathbf{w}-\mathbf{y}(t))\\
&\le2\gamma_t\big[\Phi(\mathbf{w})-\Phi\big(P_{\mathcal{G}}(y(t))\big)\big]+\gamma_t^2L^2\rho^{-1}+\rho\mathsf{d}_{\mathcal{G}}^2(\mathbf{y}(t)).
\end{aligned}$$

Using this bound in (19) and then taking conditional expectation with Assumption 2 in place gives us

$$\begin{aligned}
&\mathbb{E}_t\left[\|\mathbf{y}(t+1)-\mathbf{w}\|^2\right]\\
&\le(1+\kappa_tK)\|\mathbf{y}(t)-\mathbf{w}\|^2\\
&\quad-\textstyle\sum_{i\in\mathcal{A}}\tilde{\beta}_{it}\mathbb{E}_t\left[\sum_{k\in\mathcal{K}_{it}}c_{ik}^+(\mathbf{z}_i(t))^2\right]\\
&\quad+\gamma_t^2(L^2\rho^{-1}+4NL^2+2\nu)+\gamma_t^2\mu^2(L^2\kappa_t^{-1}+4L^2K)a_t\\
&\quad+2\gamma_t\big[\Phi(\mathbf{w})-\Phi\big(P_{\mathcal{G}}(\mathbf{y}(t))\big)\big]+\rho\mathsf{d}_{\mathcal{G}}^2(\mathbf{y}(t)). \qquad (20)
\end{aligned}$$

We now bound the second term on the RHS. By Lemma 7, for any $\tau>0$,

$$\begin{aligned}
&\textstyle\sum_{i\in\mathcal{A}}\tilde{\beta}_{it}\mathbb{E}_t\left[\sum_{k\in\mathcal{K}_{it}}c_{ik}^+(\mathbf{z}_i(t))^2\right]\\
&\ge\textstyle\sum_{i\in\mathcal{A}}\tilde{\beta}_{it}\mathbb{E}_{\mathcal{K}\sim\mathcal{U}_i}\left[\sum_{k\in\mathcal{K}}\frac{\tau-1}{\tau}c_{ik}^+(\mathbf{y}_i(t))^2\right]\\
&\quad-\textstyle\sum_{i\in\mathcal{A}}\tilde{\beta}_{it}s_i2\tau\gamma_t^2L^2\big(2L^2+\mathbb{E}_t\left[\|\mathbf{V}_i(t)\|^2\right]\big)\\
&\quad-\textstyle\sum_{i\in\mathcal{A}}\tilde{\beta}_{it}s_i4\tau\gamma_t^2L^4\mu^2K\|\mathbf{e}_i(t)-\bar{\mathbf{e}}(t)\|^2.
\end{aligned}$$

Taking $\tau=4$, using Assumption 1-c, $\tilde{\beta}=\inf_{i,t}\tilde{\beta}_{it}$ and the fact that $\tilde{\beta}_{it}s_i=\frac{(2-\beta_{it}s_i)\beta_{it}s_i}{L^2}\le\frac{1}{L^2}$, we then have

$$\begin{aligned}
&\textstyle\sum_{i\in\mathcal{A}}\tilde{\beta}_{it}\mathbb{E}_t\left[\sum_{k\in\mathcal{K}_{it}}c_{ik}^+(\mathbf{z}_i(t))^2\right]\\
&\ge\textstyle\sum_{i\in\mathcal{A}}\frac{3\tilde{\beta}}{4C}\mathsf{d}_{\mathcal{G}_i}^2(\mathbf{y}_i(t))-8\gamma_t^2\big(2NL^2+\nu\big)-16\gamma_t^2L^2\mu^2Ka_t\\
&\ge\tfrac{3\tilde{\beta}}{4C}\mathsf{d}_{\mathcal{G}}^2(\mathbf{y}(t))-8\gamma_t^2\big(2NL^2+\nu\big)-16\gamma_t^2L^2\mu^2Ka_t. \qquad (21)
\end{aligned}$$

Using this bound in (20) with $\rho=\frac{\tilde{\beta}}{4C}$ and then rearranging terms yields (15). This completes the proof. □

Note that $\tilde{\beta}=\inf_{i,t}\tilde{\beta}_{it}$ is strictly positive since we assume that $\beta_{it}\in[\underline{\beta}_i,\bar{\beta}_i]\subset(0,2/s_i)$, and thus, $\tilde{\beta}_{it}\ge\frac{(2-s_i\bar{\beta}_i)\underline{\beta}_i}{L^2}>0$ for all $i\in\mathcal{A}$ and $t\in\mathbb{N}$.

Note also that we introduced a positive sequence $\{\kappa_t:t\in\mathbb{N}\}$ in (15), which will be chosen appropriately for various intermediate results below. The following is obtained by choosing a constant sequence.

Corollary 1. *For all $t\in\mathbb{N}$, we have*

$$\begin{aligned}
&\mathbb{E}_t\left[\|\mathbf{y}(t+1)-P_{\mathcal{G}}(\mathbf{y}(t))\|^2\right]\\
&\qquad\le\epsilon_d\mathsf{d}_{\mathcal{G}}^2(\mathbf{y}(t))+\gamma_t^2D_1+\gamma_t^2D_3a_t, \qquad (22)
\end{aligned}$$

where $\epsilon_d=1-\rho$, and $D_3=\mu^2L^2K\rho^{-1}+D_2$.

Proof. From (15) in Theorem 1 with $\mathbf{w}=P_{\mathcal{G}}(\mathbf{y}(t))$, we have, for any $\kappa_t>0$ and $t\in\mathbb{N}$,

$$\begin{aligned}
&\mathbb{E}_t\left[\|\mathbf{y}(t+1)-P_{\mathcal{G}}(\mathbf{y}(t))\|^2\right]\\
&\le(1+\kappa_tK-2\rho)\mathsf{d}_{\mathcal{G}}^2(\mathbf{y}(t))+\gamma_t^2D_1+\gamma_t^2\big(\mu^2L^2\kappa_t^{-1}+D_2\big)a_t.
\end{aligned}$$

Thus, (22) follows by taking $\kappa_t\equiv\frac{\rho}{K}$ for all $t\in\mathbb{N}$. □

Since $\mathsf{d}_{\mathcal{G}}^2(\mathbf{y}(t+1))\le\|\mathbf{y}(t+1)-P_{\mathcal{G}}(\mathbf{y}(t))\|^2$, a direct consequence of this corollary is that

$$\mathbb{E}_t\left[\mathsf{d}_{\mathcal{G}}^2(\mathbf{y}(t+1))\right]\le\epsilon_d\mathsf{d}_{\mathcal{G}}^2(\mathbf{y}(t))+\gamma_t^2D_1+\gamma_t^2D_3a_t. \quad (23)$$

This result hints at a form of contraction property of the sequence $\{\mathsf{d}_{\mathcal{G}}^2(\mathbf{y}(t)):t\in\mathbb{N}\}$. If $\{a_t:t\in\mathbb{N}\}$ were bounded, together with Assumption 4, (23) would immediately imply that $\lim_{t\to\infty}\mathsf{d}_{\mathcal{G}}^2(\mathbf{y}(t))=0$. However, as a_t depends on $\|\mathbf{g}(\mathbf{y}(t+1))-\mathbf{g}(\mathbf{y}(t))\|$, which in turn depends on $\|\mathbf{y}(t+1)-\mathbf{y}(t)\|$, such convergence does not follow from (23). Below, we shed some light on this through a customized linear coupling argument.

Theorem 2. *Suppose that Assumptions 1–3 hold and step sizes satisfy $\lim_{t\to\infty}\gamma_t=0$. Then, there exist $\epsilon,\epsilon_a,\epsilon_b\in(0,1)$ such that the following holds for all t sufficiently large:*

$$\begin{aligned}
&\mathbb{E}_t\left[\mathsf{d}_{\mathcal{G}}^2(\mathbf{y}(t+1))+\epsilon_bb_{t+1}+\epsilon_aa_{t+1}\right]\\
&\le\epsilon\big(\mathsf{d}_{\mathcal{G}}^2(\mathbf{y}(t))+\epsilon_bb_t\big)+\gamma_t^2a_t(1+\epsilon_b)D_3+\mathcal{O}(p_t), \qquad (24)
\end{aligned}$$

where D_3 is given in Corollary 1,

$$\begin{aligned}
&b_t=\textstyle\sum_{s=0}^t\sigma^{t-s}\big(\mathsf{d}_{\mathcal{G}}^2(\mathbf{y}(s))+\frac{1}{2}\gamma_s^2a_sD_3\big),\quad\textit{and}\\
&p_t=\gamma_t^2+\sigma^{2(t+1)}+\textstyle\sum_{s=0}^t\sigma^{t-s}\gamma_s^2
\end{aligned}$$

with $\sigma\in(0,1)$ given in Lemma 2.

Proof. Applying Lemma 2 with $\boldsymbol{\theta}_i(t)=\mathbf{e}_i(t)$ and $\boldsymbol{\varepsilon}_i(t+1)=\Delta\mathbf{g}_i(t):=\mathbf{g}_i(\mathbf{y}_i(t+1))-\mathbf{g}_i(\mathbf{y}_i(t))$ and then summing over $i\in\mathcal{A}$ yields

$$\sum_{i\in\mathcal{A}}\|\mathbf{e}_i(t)-\bar{\mathbf{e}}(t)\|\le C_1\sigma^t+C_2\sum_{s=0}^{t-1}\sigma^{t-1-s}\sum_{i\in\mathcal{A}}\|\Delta\mathbf{g}_i(s)\|.$$

This relation and Cauchy-Schwartz inequality imply

$$\begin{aligned}
&\tfrac{1}{2}a_{t+1}\le\tfrac{1}{2}\big(\textstyle\sum_{i\in\mathcal{A}}\|\mathbf{e}_i(t+1)-\bar{\mathbf{e}}(t+1)\|\big)^2\\
&\le C_1^2\sigma^{2(t+1)}+C_2^2\big(\textstyle\sum_{s=0}^t\sigma^{t-s}\sum_{i\in\mathcal{A}}\|\Delta\mathbf{g}_i(s)\|\big)^2\\
&\le C_1^2\sigma^{2(t+1)}+\tfrac{C_2^2N}{1-\sigma}\textstyle\sum_{s=0}^t\sigma^{t-s}\sum_{i\in\mathcal{A}}\|\Delta\mathbf{g}_i(s)\|^2. \qquad (25)
\end{aligned}$$

For the last inequality, we make use of the following inequalities: i) $(\sum_k h_k m_k)^2 \le (\sum_k h_k^2)(\sum_k m_k^2)$ with $h_k = \sqrt{\sigma^{t-k}}$ and $m_k = \sqrt{\sigma^{t-k}}\sum_{i\in\mathcal{A}}\|\Delta\mathbf{g}_i(k)\|$, ii) $\sum_{s=0}^t \sigma^{t-s} \le \frac{1}{1-\sigma}$ for all $t \in \mathbb{N}$, and iii) $\big(\sum_{i\in\mathcal{A}}\|\Delta\mathbf{g}_i(s)\|\big)^2 \le N\sum_{i\in\mathcal{A}}\|\Delta\mathbf{g}_i(s)\|^2$.

Next, we bound the last term in (25). Note that $\|\Delta\mathbf{g}_i(s)\| \le \sqrt{K}\|\Delta\mathbf{g}_i(s)\|_\infty$. Thus, we have

$$\begin{aligned}&\|\Delta\mathbf{g}_i(s)\|^2 \le KL^2\|\mathbf{y}_i(s+1)-\mathbf{y}_i(s)\|^2\\ &\le 2KL^2\big(\|\mathbf{y}_i(s+1)-P_{\mathcal{G}_i}(\mathbf{y}_i(s))\|^2+\mathsf{d}^2_{\mathcal{G}_i}(\mathbf{y}_i(s))\big).\end{aligned}$$

Therefore,

$$\begin{aligned}&\mathbb{E}_s\left[\textstyle\sum_{i\in\mathcal{A}}\|\Delta\mathbf{g}_i(s)\|^2\right]\\ &\le 2KL^2\big(\mathbb{E}_s\left[\|\mathbf{y}(s+1)-P_{\mathcal{G}}(\mathbf{y}(s))\|^2\right]+\mathsf{d}^2_{\mathcal{G}}(\mathbf{y}(s))\big)\\ &\le 2KL^2\big(2\mathsf{d}^2_{\mathcal{G}}(\mathbf{y}(s))+\gamma_s^2D_1+\gamma_s^2a_sD_3\big),\end{aligned}$$

where the last inequality follows from (22) and the fact that $\epsilon_d < 1$. Using this bound in (25) after taking the expectation,

$$\mathbb{E}_t\left[a_{t+1}\right] \le 2\big(C_1^2\sigma^{2t+2}+C_3b_t+KL^2D_1\textstyle\sum_{s=0}^t\sigma^{t-s}\gamma_s^2\big) \quad (26)$$

where $b_t = \sum_{s=0}^t \sigma^{t-s}\big(\mathsf{d}^2_{\mathcal{G}}(\mathbf{y}(s)) + \gamma_s^2 a_s\frac{D_3}{2}\big)$ and $C_3 = 4C_2^2NKL^2(1-\sigma)^{-1}$. Note that

$$b_{t+1} = \sigma b_t + \mathsf{d}^2_{\mathcal{G}}(\mathbf{y}(t+1)) + \gamma_{t+1}^2 a_{t+1}\tfrac{D_3}{2}$$

which, when taking conditional expectation of both sides and using (26), implies

$$\begin{aligned}\mathbb{E}_t\left[b_{t+1}\right] &\le (\sigma+\gamma_{t+1}^2C_3D_3)b_t+\mathbb{E}_t\left[\mathsf{d}^2_{\mathcal{G}}(\mathbf{y}(t+1))\right]\\ &\quad+\gamma_{t+1}^2D_3\big(C_1^2\sigma^{2t+2}+KL^2D_1\textstyle\sum_{s=0}^t\sigma^{t-s}\gamma_s^2\big).\end{aligned}$$

Then, in view of the bound in (23), we further have

$$\begin{aligned}\mathbb{E}_t\left[b_{t+1}\right] &\le (\sigma+\gamma_{t+1}^2C_3D_3)b_t+\epsilon_d\mathsf{d}^2_{\mathcal{G}}(\mathbf{y}(t))\\ &\quad+\gamma_t^2a_tD_3+\tilde{p}_t,\end{aligned} \quad (27)$$

where $\tilde{p}_t = \gamma_{t+1}^2D_3\big(C_1^2\sigma^{2t+2}+KL^2D_1\sum_{s=0}^t\sigma^{t-s}\gamma_s^2\big)+\gamma_t^2D_1$.

We now couple (23), (26) and (27) as follows: for any $\epsilon_a > 0$ and $\epsilon_b > 0$, add (23), (26) multiplied by ϵ_a, and (27) multiplied by ϵ_b, and then simplify the expression.

$$\begin{aligned}&\mathbb{E}_t\left[\mathsf{d}^2_{\mathcal{G}}(\mathbf{y}(t+1))+\epsilon_bb_{t+1}+\epsilon_aa_{t+1}\right]\\ &\le\epsilon_d(\epsilon_b+1)\mathsf{d}^2_{\mathcal{G}}(\mathbf{y}(t))+\big(\epsilon_b\sigma+\gamma_{t+1}^2\epsilon_bC_3D_3+2\epsilon_aC_3\big)b_t\\ &\quad+\gamma_t^2a_t(1+\epsilon_b)D_3+\bar{p}_t,\end{aligned}$$

where

$$\begin{aligned}\bar{p}_t &= \epsilon_a2\big(C_1^2\sigma^{2t+2}+KL^2D_1\textstyle\sum_{s=0}^t\sigma^{t-s}\gamma_s^2\big)+\gamma_t^2D_1+\epsilon_b\tilde{p}_t\\ &= \mathcal{O}\big(\gamma_t^2+\sigma^{2(t+1)}+\textstyle\sum_{s=0}^t\sigma^{t-s}\gamma_s^2\big).\end{aligned} \quad (28)$$

Let us now choose ϵ_a and ϵ_b sufficiently small so that there exists $\epsilon \in (0,1)$ satisfying

$$\epsilon_d(\epsilon_b+1) < \epsilon \quad\text{and}\quad \epsilon_b(\sigma+\gamma_{t+1}^2C_3D_3)+2\epsilon_aC_3 < \epsilon\epsilon_b$$

for sufficiently large t (with $\gamma_{t+1}^2 < \frac{1}{C_3D_3}(\epsilon-\sigma-\epsilon_a\frac{2C_3}{\epsilon_b})$). Thus, we conclude that (24) holds for sufficiently large t. □

Note that the exact form of the term $\mathcal{O}(p_t)$, given by $\bar{p}_t$ in (28), is not important for us, as we will see later that this term behaves like γ_t^2 and, thus, is summable under Assumption 4.

It is now clear that the sequence $\{\mathsf{d}^2_{\mathcal{G}}(\mathbf{y}(t))+\epsilon_bb_t+\epsilon_aa_t : t\in\mathbb{N}\}$ possesses a contraction property. As a result, we have the following corollary.

Corollary 2. *Under Assumptions 1–4, w.p.1.,*

$$\lim_{t\to\infty}\big(\mathsf{d}^2_{\mathcal{G}}(\mathbf{y}(t))+\epsilon_bb_t+\epsilon_aa_t\big)=0.$$

Proof. For sufficiently large t, Theorem 2 tells us

$$\begin{aligned}&\mathbb{E}_t\left[\mathsf{d}^2_{\mathcal{G}}(\mathbf{y}(t+1))+\epsilon_bb_{t+1}+\epsilon_aa_{t+1}\right]\\ &\le\epsilon\big(\mathsf{d}^2_{\mathcal{G}}(\mathbf{y}(t))+\epsilon_bb_t+\epsilon_aa_t\big)+\mathcal{O}(p_t).\end{aligned}$$

Note that, for any $T\in\mathbb{N}$,

$$\begin{aligned}\textstyle\sum_{t=0}^T p_t &= \textstyle\sum_{t=0}^T\big(\gamma_t^2+\sigma^{2(t+1)}+\sum_{s=0}^t\sigma^{t-s}\gamma_s^2\big)\\ &\le\tfrac{\sigma^2}{1-\sigma^2}+\textstyle\sum_{t=0}^T\gamma_t^2+\sum_{t=0}^T\sum_{s=0}^t\sigma^{t-s}\gamma_s^2\\ &=\tfrac{\sigma^2}{1-\sigma^2}+\textstyle\sum_{t=0}^T\gamma_t^2+\sum_{s=0}^T\gamma_s^2\sum_{k=0}^{T-s}\sigma^k\\ &\le\tfrac{\sigma^2}{1-\sigma^2}+\tfrac{2-\sigma}{1-\sigma}\textstyle\sum_{t=0}^T\gamma_t^2.\end{aligned} \quad (29)$$

Thus, under Assumption 4, $\sum_{t\in\mathbb{N}}p_t<\infty$. The convergence of corollary now follows from Lemma 5. □

We are now ready to present the main convergence result of the paper, which will be proved using Lemma 6 with

$$v_t = \mathsf{d}^2_{\mathcal{Y}^*}(\mathbf{y}(t))+\mathsf{d}^2_{\mathcal{G}}(\mathbf{y}(t))+\epsilon_bb_t+\epsilon_aa_t.$$

We also consider the following running average terms:

$$\tilde{\mathbf{y}}(s,t)=\frac{\sum_{k=s}^t\gamma_k\mathbf{y}(k)}{\sum_{k=s}^t\gamma_k},\ \ \tilde{\mathbf{x}}(s,t)=\frac{\sum_{k=s}^t\gamma_kP_{\mathcal{G}}\big(\mathbf{y}(k)\big)}{\sum_{k=s}^t\gamma_k} \quad (30)$$

Theorem 3. *Suppose that Assumptions 1–3 hold and step sizes satisfy* $\lim_{t\to\infty}\gamma_t=0$. *Let* $\bar{\gamma}=\sup_t\gamma_t$.

(i) *If* $\mathbb{E}\left[\mathsf{d}^2_{\mathcal{Y}^*}(\mathbf{y}(t))\right]$ *is bounded for all* $t\in\mathbb{N}$, *then*

$$\mathbb{E}\left[\Phi\big(\tilde{\mathbf{x}}(s,t)\big)\right]-\Phi^*+\frac{\rho}{\bar{\gamma}}\mathbb{E}\left[\|\tilde{\mathbf{y}}(s,t)-\tilde{\mathbf{x}}(s,t)\|^2\right]\le E_{s,t}$$

$$\text{with } E_{s,t}=\frac{\mathbb{E}\left[v_s\right]+\mathcal{O}(\sum_{k=s}^t p_k)}{2\sum_{k=s}^t\gamma_k},\quad 0\le s\le t. \quad (31)$$

Moreover, $\lim_{t\to\infty}E_{s,t}=0$ *for all* $s\in\mathbb{N}$ *if* $\sum_{t\in\mathbb{N}}\gamma_t=\infty$.

(ii) *If Assumption 4 holds, then* $\mathsf{d}_{\mathcal{Y}^*}(\mathbf{y}(t))\to0$ *w.p.1.*

Proof. First, we make use of Theorem 1 with $\mathbf{w}=P_{\mathcal{Y}^*}(\mathbf{y}(t))$ for each t. Under Assumptions 1–3,

$$\begin{aligned}&\mathbb{E}_t\left[\mathsf{d}^2_{\mathcal{Y}^*}(\mathbf{y}(t+1))\right]\\ &\le(1+\kappa_tK)\mathsf{d}^2_{\mathcal{Y}^*}(\mathbf{y}(t))-u_t+\gamma_t^2\mu^2L^2\kappa_t^{-1}a_t+\gamma_t^2(D_1+a_tD_2),\end{aligned}$$

where

$$u_t:=2\gamma_t\big[\Phi\big(P_{\mathcal{G}}(\mathbf{y}(t))\big)-\Phi^*\big]+2\rho\mathsf{d}^2_{\mathcal{G}}(\mathbf{y}(t))\ge0.$$

Taking $\kappa_t=\gamma_t^2\mu^2L^2/\epsilon\epsilon_a$, we can write the above relation as

$$\begin{aligned}&\mathbb{E}_t\left[\mathsf{d}^2_{\mathcal{Y}^*}(\mathbf{y}(t+1))\right]\\ &\le\big(1+\kappa_tK\big)\mathsf{d}^2_{\mathcal{Y}^*}(\mathbf{y}(t))-u_t+\epsilon\epsilon_aa_t+\gamma_t^2(D_1+a_tD_2).\end{aligned}$$

Adding this relation with (24), we obtain

$$\begin{aligned}&\mathbb{E}_t\left[\mathsf{d}^2_{\mathcal{Y}^*}(\mathbf{y}(t+1))+\mathsf{d}^2_{\mathcal{G}}(\mathbf{y}(t+1))+\epsilon_bb_{t+1}+\epsilon_aa_{t+1}\right]\\ &\le\big(1+\kappa_tK\big)\mathsf{d}^2_{\mathcal{Y}^*}(\mathbf{y}(t))+\epsilon\big(\mathsf{d}^2_{\mathcal{G}}(\mathbf{y}(t))+\epsilon_bb_t+\epsilon_aa_t\big)\\ &\quad-u_t+\gamma_t^2a_t(D_2+D_3+\epsilon_bD_3)+\mathcal{O}(p_t).\end{aligned} \quad (32)$$

Under the assumption that $\mathbb{E}\left[\mathsf{d}^2_{\mathcal{Y}^*}(\mathbf{y}(t))\right]$ is bounded for all $t \in \mathbb{N}$, i.e., $\mathbb{E}\left[\mathsf{d}^2_{\mathcal{Y}^*}(\mathbf{y}(t))\right] = \mathcal{O}(1)$, taking the expectation of (32), we obtain the following for all sufficiently large t.

$$\begin{aligned}
&\mathbb{E}\left[\mathsf{d}^2_{\mathcal{Y}^*}(\mathbf{y}(t+1)) + \mathsf{d}^2_{\mathcal{G}}(\mathbf{y}(t+1)) + \epsilon_b b_{t+1} + \epsilon_a a_{t+1}\right] \\
&\leq \mathbb{E}\left[\mathsf{d}^2_{\mathcal{Y}^*}(\mathbf{y}(t)) + \epsilon\big(\mathsf{d}^2_{\mathcal{G}}(\mathbf{y}(t)) + \epsilon_b b_t + \epsilon_a a_t\big)\right] - \mathbb{E}\left[u_t\right] \\
&\quad + \mathcal{O}(\gamma_t^2)\mathbb{E}\left[a_t\right] + \mathcal{O}(p_t). \\
&\leq \mathbb{E}\left[\mathsf{d}^2_{\mathcal{Y}^*}(\mathbf{y}(t)) + \mathsf{d}^2_{\mathcal{G}}(\mathbf{y}(t)) + \epsilon_b b_t + \epsilon_a a_t\right] - \mathbb{E}\left[u_t\right] + \mathcal{O}(p_t),
\end{aligned}$$

where we have used $\kappa_t K \mathbb{E}\left[\mathsf{d}^2_{\mathcal{Y}^*}(\mathbf{y}(t))\right] + \mathcal{O}(p_t) = \mathcal{O}(p_t)$ in the first inequality and $\big(\epsilon\epsilon_a + \mathcal{O}(\gamma_t^2)\big)\mathbb{E}\left[a_t\right] \leq \epsilon_a \mathbb{E}\left[a_t\right]$ in the second. After rearranging terms of the above relation, we have

$$\mathbb{E}\left[u_t\right] \leq \mathbb{E}\left[v_t\right] - \mathbb{E}\left[v_{t+1}\right] + \mathcal{O}(p_t)$$

which, together with the fact that $\mathbb{E}\left[v_{t+1}\right] \geq 0$, implies

$$\textstyle\sum_{k=s}^t \mathbb{E}\left[u_k\right] \leq \mathbb{E}\left[v_s\right] + \mathcal{O}(\sum_{k=s}^t p_t). \tag{33}$$

Next we bound the left-hand side. Note that

$$\begin{aligned}
&\textstyle\sum_{k=s}^t u_k = \sum_{k=s}^t \left(2\gamma_k\big[\Phi\big(P_{\mathcal{G}}(\mathbf{y}(k))\big) - \Phi^*\big] + 2\rho \mathsf{d}^2_{\mathcal{G}}(\mathbf{y}(k))\right) \\
&\geq 2\textstyle\sum_{k=s}^t \gamma_k\big[\Phi\big(P_{\mathcal{G}}(\mathbf{y}(k))\big) - \Phi^* + \frac{\rho}{\bar{\gamma}}\mathsf{d}^2_{\mathcal{G}}(\mathbf{y}(k))\big] \\
&\geq 2\big(\textstyle\sum_{k=s}^t \gamma_k\big)\big[\Phi\big(\tilde{\mathbf{x}}(s,t)\big) - \Phi^* + \frac{\rho}{\bar{\gamma}}\|\tilde{\mathbf{y}}(s,t) - \tilde{\mathbf{x}}(s,t)\|^2\big],
\end{aligned}$$

where we used convexity of $\Phi(\cdot)$ and $\|\cdot\|^2$ and the definitions of $\tilde{\mathbf{y}}(s,t)$ and $\tilde{\mathbf{x}}(s,t)$ in (30). Using this bound for (33), we have

$$\begin{aligned}
&\mathbb{E}\left[\Phi\big(\tilde{\mathbf{x}}(s,t)\big)\right] - \Phi^* + \frac{\rho}{\bar{\gamma}}\mathbb{E}\left[\|\tilde{\mathbf{y}}(s,t) - \tilde{\mathbf{x}}(s,t)\|^2\right] \\
&\leq \frac{\mathbb{E}\left[v_s\right] + \mathcal{O}(\sum_{k=s}^t p_k)}{2\sum_{k=s}^t \gamma_k}.
\end{aligned}$$

This proves statement (i) of the theorem.

To show statement (ii), note from (32) that

$$\mathbb{E}_t\left[v_{t+1}\right] \leq (1+\alpha_t)v_t - u_t + \mathcal{O}(p_t),$$

where $\alpha_t = \kappa_t K = \gamma_t^2 \mu^2 L^2 K/\epsilon\epsilon_a$. Applying Lemma 6 to this relation and noting that $\sum_t p_t < \infty$ (see (29)), the following holds w.p.1:

(a) $\sum_{t\in\mathbb{N}} \gamma_t(\Phi\big(P_{\mathcal{G}}(\mathbf{y}(t))\big) - \Phi^*) < \infty$;
(b) $\sum_{t\in\mathbb{N}} \mathsf{d}^2_{\mathcal{G}}(\mathbf{y}(t)) < \infty$; and
(c) there is some nonegative RV v such that $v_t \to v$.

Since $\sum_{t\in\mathbb{N}} \gamma_t = \infty$ and $\mathsf{d}^2_{\mathcal{G}}(\mathbf{y}(t)) + \epsilon_b b_t + \epsilon_a a_t \to 0$ (Corollary 2), the findings (a) and (c) imply $\liminf_t \Phi\big(P_{\mathcal{G}}(\mathbf{y}(t))\big) = \Phi^*$ and $\mathsf{d}^2_{\mathcal{Y}^*}(\mathbf{y}(t)) \to v$ as $t \to \infty$. Because Φ is continuous, it follows that $v = 0$ w.p.1. This proves the second part of the theorem. □

Remark 1. Note that the boundedness condition of $\mathbb{E}\left[\mathsf{d}^2_{\mathcal{Y}^*}(\mathbf{y}(t))\right]$ in Theorem 3(i) is satisfied if $\mathcal{G}_{i0}$, $i \in \mathcal{A}$, are compact, which will likely hold in many, if not most, cases of practical interest. Also, the convergence rate of the algorithm obviously depends on the choice of step sizes. For instance, if $\gamma_t = \mathcal{O}(\frac{1}{\sqrt{t}})$, then $E_{\lfloor t/2\rfloor,t} = \mathcal{O}(\frac{1}{\sqrt{t}})$ [1].

IV. Conclusion

We studied solving a constrained optimization problem using noisy observations, and proposed a new distributed algorithm that does not require sharing optimization variables among agents. Instead, the agents update their local estimates of global constraints using a consensus-type algorithm, while updating their own local optimization variable based on noisy estimates of gradients of local objective functions. We proved that (a) the optimization variables converge to an optimal point of an approximated problem and (b) the tracking errors of local estimates of constraint functions vanish asymptotically.

References

[1] A. Beck. *First-order methods in optimization*. SIAM, 2017.

[2] P. Bianchi and J. Jakubowicz. Convergence of a multi-agent projected stochastic gradient algorithm for non-convex optimization. *IEEE Trans. on Automatic Control*, 58(2):391–405, 2013.

[3] T.-H. Chang, M. Hong, and X. Wang. Multi-agent distributed optimization via inexact consensus admm. *IEEE Trans. on Signal Processing*, 63(2):482–497, 2015.

[4] N. Chatzipanagiotis, D. Dentcheva, and M. M. Zavlanos. An augmented Lagrangian method for distributed optimization. *Mathematical Programming*, 152(1):405–434, 2015.

[5] N. Chatzipanagiotis and M. M. Zavlanos. A distributed algorithm for convex constrained optimization under noise. *IEEE Trans. on Automatic Control*, 61(9):2496–2511, 2016.

[6] N. Chatzipanagiotis and M. M. Zavlanos. On the convergence of a distributed augmented lagrangian method for nonconvex optimization. *IEEE Trans. on Automatic Control*, 62(9):4405–4420, 2017.

[7] A. Cherukuri and J. Cortes. Distributed algorithms for convex network optimization under non-sparse equality constraints. In *the proc. of the 54th Annual Allerton Conference*, 2016.

[8] So. S. Kia, B. Van Scoy, J. Cortes, R. A. Freeman, K. M. Lynch, and S. Martinez. Tutorial on dynamic average consensus: The problem, its applications, and the algorithms. *IEEE Control Systems Magazine*, 39(3):40–72, 2019.

[9] S. Lee and M. M. Zavlanos. Approximate projection methods for decentralized optimization with functional constraints. *IEEE Trans. on Automatic Control*, 63(10):3248–3260, 2017.

[10] N. Li and J. R. Marden. Decoupling coupled constraints through utility design. *IEEE Trans. on Automatic Control*, 59(8):2289–2294, 2014.

[11] A. Nedić. Random algorithms for convex minimization problems. *Mathematical programming*, 129(2):225–253, 2011.

[12] A. Nedić and A. Olshevsky. Distributed optimization over time-varying directed graphs. *IEEE Trans. on Automatic Control*, 60(3):601–615, 2015.

[13] A. Nedić, A. Olshevsky, and W. Shi. Improved convergence rates for distributed resource allocation. In *the proc. of 2018 IEEE Conference on Decision and Control (CDC)*, pages 172–177. IEEE, 2018.

[14] A. Nedić, A. Ozdaglar, and P. Parrilo. Constrained consensus and optimization in multi-agent networks. *IEEE Trans. on Automatic Control*, 55(4):922–938, 2010.

[15] B. T. Polyak. *Introduction to Optimization (Translations Series in Mathematics and Engineering)*. Optimization Software Inc. Publications Division, New York, 1987.

[16] B. T. Polyak. Random algorithms for solving convex inequalities. In *Studies in Computational Mathematics*, volume 8, pages 409–422. Elsevier, 2001.

[17] K. Srivastava and A. Nedić. Distributed asynchronous constrained stochastic optimization. *IEEE Journal of Selected Topics in Signal Processing*, 5(4):772–790, 2011.

[18] S. Sundhar Ram, A. Nedić, and V. V. Veeravalli. Distributed stochastic subgradient projection algorithms for convex optimization. *Journal of Optimization Theory and Applications*, 14:514–545, 2010.

[19] S. Sundhar Ram, A. Nedić, and V. V. Veeravalli. A new class of distributed optimization algorithms: Application to regression of distributed data. *Optimization Methods and Software*, 27(1):71–88, 2012.

[20] P. Yi, J. Lei, and Y. Hong. Distributed resource allocation over random networks based on stochastic approximation. *Systems & Control Letters*, 114:44–51, 2018.