Distributed Optimization with Global Constraints Using Noisy Measurements

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

***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 shares 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, 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 in telecommunication systems) that couple the decisions of more than one agent; hence satisfying these constraints requires 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 function 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. We refer a reader to [12] for a more detailed example of such settings (and an application of the proposed algorithm presented in Section II-A).

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], [14], [16], [20], [21] and references therein), because it has broad applications and also serves as a foundation for distributed machine learning. Here, we summarize only 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, distributed stochastic optimization has been an active research area. Srivastava and Nedić [19] proposed a distributed stochastic optimization algorithm for solving constrained optimization problems, in which the feasible set is 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 function is separable and local optimization variables are coupled via global

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

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 do not want to share their local optimization variables with others. For instance, autonomous systems or domains with their own private networks do 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 (e.g., [3], [7], [15]). More recently, Yi et al. [22] 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 (see [9] for examples of such constraint functions). In order to handle nonlinear global constraints without requiring exchange of local optimization variables among agents, our approach combines 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 [18]; 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 (a) the analytic expressions for the local objective functions of agents or their gradients are unavailable and instead need to be estimated using noisy observations and (b) global constraint functions are convex. 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 show that our algorithm converges almost surely to an optimal point of an approximated problem, which also corresponds to a Nash equilibrium of the aforementioned state-based potential game in [10] (see Remark 2 in Section III).

Preliminary results of this paper were reported in [12], where the focus was on (i) the application of the proposed approach to distributed resource allocation in 5G/6G telecommunication systems comprising multiple autonomous networks that must satisfy *end-to-end* delay constraints, and (ii) the relation between the proposed approach and the state-based potential games approach [10]. In addition, [12] only considers a simple setting in which the network graph is fixed and the local constraint sets are projection-friendly, and states a convergence result without any proof. In this paper, we extend the approach to time-varying graphs and local constraint sets that are not necessarily projection-friendly, and provide a detailed convergence analysis, including the convergence rate of the algorithm.

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 $\mathbf{x} \in \mathcal{D}$. When it is clear from the context, we 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 $\mathcal{S}$ 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 $\mathbf{R}(t)$, $t \in \mathbb{N}$, we use $\mathbf{R}^t$ to denote the collection $\{\mathbf{R}(\tau) : \tau \in \{0, \ldots, t\}\}$.

The rest of the paper is organized as follows. Section II presents the problem formulation and our proposed algorithm. Section III details the convergence analysis of our algorithm. Finally, Section IV concludes the paper.

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:

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

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

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 variables, and $\mathbf{g} = (g_1, \ldots, g_K)$ is the vector consisting of K global constraint functions. Here,

$$\mathcal{G} = \textstyle\prod_{i \in \mathcal{A}} \mathcal{G}_i$$

is the global feasible set, where $\mathcal{G}_i$ is the local constraint set of agent i. Note that this formulation includes various resource allocation problems studied in the literature, see, e.g., [3], [7], [10], [12], [15], [22]. 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

$$\tilde{\mathcal{G}}_i = \{\mathbf{y}_i \mid c_{ik}(\mathbf{y}_i) \leq 0,\ k \in \tilde{\mathcal{K}}_i\},$$

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] and [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 [11]. 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. Clearly, Φ is convex on $\mathcal{G}$ provided that ϕ_i and $\mathbf{g}_i$, $i \in \mathcal{A}$, are convex.

In general, the original optimization problem in (1) is recovered only as $\mu \to \infty$. Hence, for sufficiently large μ, an optimal solution to (2) is reasoned to be a good approximated solution to (1) in practice. However, quantifying this suboptimality of the penalty method analytically for general problems is difficult and is out of the scope of this paper. Moreover, a practical choice of μ is often problem-dependent and may require domain knowledge. When such domain knowledge is unavailable, one can solve (2) for an increasing sequence of μ [11]. With this in mind, in this paper, we consider fixed μ which is assumed known to all agents. We refer interested readers to, e.g., [10], [12] for applications of penalty methods.

Note that 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 $\mathbf{g}(\mathbf{y}(t))$. 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 γ_t is a step size at iteration t,[1] $\mathbf{V}_i(t)$ are random vectors, $\mathbf{e}_i(t) = (e_{ik}(t) : k \in \mathcal{K}_G)$, $\mathbf{g}_i(\mathbf{y}_i) = (g_{ik}(\mathbf{y}_i) : k \in \mathcal{K}_G)$, and

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

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

Here, $\mathcal{K}_{it} \subset \tilde{\mathcal{K}}_i$ 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.

Let us discuss the above algorithm in detail. First, (3a) resembles a projected stochastic gradient step, where $\mathbf{q}_i(t) + \mathbf{V}_i(t)$ is used as an approximated subgradient for

$$\partial_{\mathbf{y}_i}\Phi(\mathbf{y}) = \partial\phi_i(\mathbf{y}_i(t)) + \frac{\mu}{N}\sum_{k\in\mathcal{K}_G} \partial g_{ik}(\mathbf{y}_i(t))g_k^+(\mathbf{y}(t)).$$

Since the agents do not have access to the *global* constraint violation $g_k^+(\mathbf{y}(t)) = \big(\sum_{j\in\mathcal{A}} g_{jk}(\mathbf{y}_j(t))\big)^+$ for all $k \in \mathcal{K}_G$, each agent i substitutes its *local* estimate $e_{ik}^+(t)$ for $g_k^+(\mathbf{y}(t))/N$ instead, leading to (4). In addition, for the update in (3a), agent i only has access to a noisy estimate of the subgradient given by the sum $\partial\phi_i(\mathbf{y}_i(t)) + \mathbf{V}_i(t)$, while it can compute the second term in (4). In other words, the subgradients $\partial\phi_i(\mathbf{y}_i(t))$ are unavailable to the agents.

Second, the random projection step in (3b) is borrowed from [18] (see also [13]), 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., [18]).

In addition, in (3b) we assume, for simplicity, 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 $s_i \geq 1$, 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.

Finally, the update rule for $\mathbf{e}_i(t+1)$ in (3c) 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; this allows us to approximate $g_k^+(\mathbf{y}(t))/N$ using $e_{ik}^+(t)$ in (4) to carry out an approximated stochastic gradient step in (3a) as explained above. We emphasize that, to carry out the update of $\mathbf{e}_i(t+1)$ in (3c), agent $i \in \mathcal{A}$ only needs the estimates $\mathbf{e}_j(t)$ from its *direct* neighbors for which $w_{ij}(t) > 0$ in the weight matrix $W(t) = [w_{ij}(t) : i, j \in \mathcal{A}]$, which is allowed to be time-varying.

To analyze the convergence of our algorithm, let us define $\mathcal{F}_t$ for each $t \in \mathbf{N}$ to be the σ-field generated by $\{\mathbf{y}(0), \mathbf{V}^{t-1}, (\mathcal{K}_i^{t-1} : i \in \mathcal{A})\}$, where $\mathcal{K}_i^{t-1} = \{\mathcal{K}_{is} : 0 \leq s < t\}$ and $\mathcal{K}_{is}$, $s \in \mathbf{N}$, 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 on the optimization problem in (2).

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

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

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 $i \in \mathcal{A}$,*

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

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 [13] 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 $t \in \mathbf{N}$, define $\mathcal{E}_t = \{(i,j) \in \mathcal{A}\times\mathcal{A} : w_{ij}(t) > 0\}$.

Assumption 3. *There exists $Q \in \mathbf{N}$ 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 $t \in \mathbf{N}$, and there exists $w_{\min} > 0$ such that, for all $t \in \mathbf{N}$, all nonzero weights $w_{ij}(t)$ lie in $[w_{\min}, 1]$.*

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

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

C. Preliminaries

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

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

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

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

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

We will also use the following result on the mixing property of a sequence of doubly stochastic weight matrices [21, proof of Theorem 4.2].

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

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

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

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

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 fixed 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 $i \in \mathcal{A}$,*

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

Proof. We drop the dependence on t for notational simplicity. It follows from (2), (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, by Lemma 3

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

which follows from the relationship between 1- and 2-norms. □

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

Lemma 5. *[17, Lemma 10, p. 49] Let $\{v_t : t \in \mathbf{N}\}$ be a sequence of nonnegative RVs with $\mathbb{E}[v_0] < \infty$. Suppose $\{\alpha_t : t \in \mathbf{N}\}$ and $\{\beta_t : t \in \mathbf{N}\}$ 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*

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

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

Lemma 6. *[17, 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*

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

Then, $\sum_{t\in\mathbf{N}} u_t < \infty$ w.p.1 and v_t converges to a nonnegative RV.

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 the proposed algorithm in (3), $\mathbf{y}(t)$ converges to the optimal set $\mathcal{Y}^*$ w.p.1. (Theorem 1). We prove this result with the help of a series of auxiliary results.

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

First, the projection in (3b) satisfies the following inequality, the proof of which is inspired by [9], [13], [18].

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

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

$$c_{ik}^+(\mathbf{z}_i(t))^2 \ge \frac{3}{4}c_{ik}^+(\mathbf{y}_i(t))^2 - 8\gamma_t^2L^2\big(2L^2+\|\mathbf{V}_i(t)\|^2\big) - 16\gamma_t^2L^4\mu^2K\|\mathbf{e}_i(t)-\bar{\mathbf{e}}(t)\|^2, \quad (11b)$$

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

Proof. The proof of (11a) is similar to those given in [13], [18], but we provide it here for completeness. For any $\mathbf{y}_i \in \mathcal{G}_i$,

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

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

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

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

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

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

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

$$\begin{aligned}&c_{ik}^+(\mathbf{z}_i(t))^2\\ &\ge c_{ik}^+(\mathbf{y}_i(t))^2+2(c_{ik}^+(\mathbf{z}_i(t))-c_{ik}^+(\mathbf{y}_i(t)))c_{ik}^+(\mathbf{y}_i(t))\\ &\ge c_{ik}^+(\mathbf{y}_i(t))^2-2|c_{ik}^+(\mathbf{z}_i(t))-c_{ik}^+(\mathbf{y}_i(t))|c_{ik}^+(\mathbf{y}_i(t))\\ &\ge \tfrac{3}{4}c_{ik}^+(\mathbf{y}_i(t))^2-4|c_{ik}^+(\mathbf{z}_i(t))-c_{ik}^+(\mathbf{y}_i(t))|^2, \quad (13)\end{aligned}$$

where the last inequality follows from (8) (with $a = |c_{ik}^+(\mathbf{z}_i(t)) - c_{ik}^+(\mathbf{y}_i(t))|$, $b = c_{ik}^+(\mathbf{y}_i(t))$, and $\eta = 4$). In addition,

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

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

Although inequalities in the lemma are similar to those in [9] and [13], it contains an extra term $\|\mathbf{e}_i(t)-\bar{\mathbf{e}}(t)\|^2$ in (11b), which is critical in the analysis of our algorithm. This term measures the disagreements in the agent's local estimate of the global constraint functions. Let us define

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

In order to prove the convergence of the algorithm, we need to show a_t decays to 0 with increasing t. As will be clear, much of the challenge in establishing convergence lies in properly bounding this term, which is not bounded a priori.

The following intermediate result relates a_t to $\|\mathbf{y}(t)-\mathbf{w}\|^2$ and $\Phi(\mathbf{w})-\Phi\big(P_{\mathcal{G}}(\mathbf{y}(t))\big)$ for arbitrary $\mathbf{w} \in \mathcal{G}$ and any choice of step size γ_t to provide a bound on the progress per iteration of the algorithm. Define $\tilde{\beta} := \inf_{i,t}\tilde{\beta}_{it}$ and $\rho := \frac{\tilde{\beta}}{4C}$ with C in Assumption 1-c.

Lemma 8. *Under Assumptions 1 and 2, the following holds for any $t \in \mathbf{N}$, $\mathbf{w} \in \mathcal{G}$, and positive sequence $\{\kappa_t : t \in \mathbf{N}\}$:*

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

where $D_1 = L^2\rho^{-1}+10(2L^2N+\nu)$ and $D_2 = 20L^2K\mu^2$.

Proof. First, for any $\mathbf{w}_i \in \mathcal{G}_i$, by (11a) in 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 right-hand side (RHS) can be expanded further as

$$\begin{aligned}\|\mathbf{z}_i(t)-\mathbf{w}_i\|^2 &\le \big\|\mathbf{y}_i(t)-\gamma_t\big(\mathbf{q}_i(t)+\mathbf{V}_i(t)\big)-\mathbf{w}_i\big\|^2\\ &= \|\mathbf{y}_i(t)-\mathbf{w}_i\|^2+\gamma_t^2\|\mathbf{q}_i(t)+\mathbf{V}_i(t)\|^2\\ &\quad+2\gamma_t\big(\mathbf{q}_i(t)+\mathbf{V}_i(t)\big)^{\mathsf{T}}(\mathbf{w}_i-\mathbf{y}_i(t)). \quad (16)\end{aligned}$$

Thus, from (2), (4) and (16), we get

$$\begin{aligned}&\|\mathbf{y}_i(t+1)-\mathbf{w}_i\|^2\\ &\le\|\mathbf{y}_i(t)-\mathbf{w}_i\|^2-\tilde{\beta}_{it}\sum_{k\in\mathcal{K}_{it}}c_{ik}^+(\mathbf{z}_i(t))^2\\ &\quad+2\gamma_t\big(\partial_{\mathbf{y}_i}\Phi(\mathbf{y}(t))+\mathbf{V}_i(t)\big)^{\mathsf{T}}(\mathbf{w}_i-\mathbf{y}_i(t))\\ &\quad+2\gamma_t\mu\sum_{k\in\mathcal{K}_G}\big|e_{ik}^+(t)-\bar{e}_k^+(t)\big|\cdot\big|\partial g_{ik}(\mathbf{y}_i(t))^{\mathsf{T}}(\mathbf{w}_i-\mathbf{y}_i(t))\big|\\ &\quad+\gamma_t^2\|\mathbf{q}_i(t)+\mathbf{V}_i(t)\|^2. \quad (17)\end{aligned}$$

We bound the last two terms in (17) as follows. First,

$$\begin{aligned}&\gamma_t^2\|\mathbf{q}_i(t)+\mathbf{V}_i(t)\|^2 \le 2\gamma_t^2\big(\|\mathbf{q}_i(t)\|^2+\|\mathbf{V}_i(t)\|^2\big)\\ &\le 2\gamma_t^2\big(2L^2\mu^2K\|\mathbf{e}_i(t)-\bar{\mathbf{e}}(t)\|^2+2L^2+\|\mathbf{V}_i(t)\|^2\big), \quad (18)\end{aligned}$$

where the second inequality follows from Lemma 4. Second, using the Lipschitz continuity of g_{ik} and then the inequality in (8) of Lemma 3, we have, for any $\kappa_t > 0$,

$$\begin{aligned}&2\gamma_t\mu\sum_{k\in\mathcal{K}_G}\big|e_{ik}^+(t)-\bar{e}_k^+(t)\big|\cdot\big|\partial g_{ik}(\mathbf{y}_i(t))^{\mathsf{T}}(\mathbf{w}_i-\mathbf{y}_i(t))\big|\\ &\le\sum_{k\in\mathcal{K}_G}2\gamma_t\mu L\big|e_{ik}(t)-\bar{e}_k(t)\big|\cdot\|\mathbf{w}_i-\mathbf{y}_i(t)\|\\ &\le\sum_{k\in\mathcal{K}_G}\big(\gamma_t^2\mu^2L^2\kappa_t^{-1}|e_{ik}(t)-\bar{e}_k(t)|^2+\kappa_t\|\mathbf{w}_i-\mathbf{y}_i(t)\|^2\big)\\ &=\gamma_t^2\mu^2L^2\kappa_t^{-1}\|\mathbf{e}_i(t)-\bar{\mathbf{e}}(t)\|^2+\kappa_tK\|\mathbf{w}_i-\mathbf{y}_i(t)\|^2. \quad (19)\end{aligned}$$

Using the bounds of (18) and (19) in (17) and then summing over $i \in \mathcal{A}$ gives us the following bound.

$$\begin{aligned}
&\|\mathbf{y}(t+1) - \mathbf{w}\|^2 \\
&\le (1+\kappa_t K)\|\mathbf{y}(t) - \mathbf{w}\|^2 - \sum_{i\in\mathcal{A}} \tilde{\beta}_{it} \sum_{k\in\mathcal{K}_{it}} c_{ik}^+(\mathbf{z}_i(t))^2 \\
&\quad + 4NL^2\gamma_t^2 + 2\gamma_t^2\|\mathbf{V}(t)\|^2 + \gamma_t^2\mu^2(L^2\kappa_t^{-1} + 4L^2K)a_t \\
&\quad + 2\gamma_t\big(\partial\Phi(\mathbf{y}(t)) + \mathbf{V}(t)\big)^{\mathsf{T}}(\mathbf{w} - \mathbf{y}(t)). \qquad (20)
\end{aligned}$$

Now, using the convexity and Lipschitz continuity of Φ and Lemma 3, we have, for any $\rho > 0$,

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

Taking $\mathbf{u} = P_{\mathcal{G}}(\mathbf{y}(t))$ yields

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

Using this bound in (20) and then taking conditional expectation with Assumption 2 in place, we obtain

$$\begin{aligned}
&\mathbb{E}_t\left[\|\mathbf{y}(t+1) - \mathbf{w}\|^2\right] \\
&\le (1+\kappa_t K)\|\mathbf{y}(t) - \mathbf{w}\|^2 \\
&\quad - \sum_{i\in\mathcal{A}} \tilde{\beta}_{it}\mathbb{E}_t\left[\textstyle\sum_{k\in\mathcal{K}_{it}} c_{ik}^+(\mathbf{z}_i(t))^2\right] \\
&\quad + \gamma_t^2(L^2\rho^{-1} + 4NL^2 + 2\nu) + \gamma_t^2\mu^2(L^2\kappa_t^{-1} + 4L^2K)a_t \\
&\quad + 2\gamma_t\big[\Phi(\mathbf{w}) - \Phi\big(P_{\mathcal{G}}(\mathbf{y}(t))\big)\big] + \rho\mathsf{d}_{\mathcal{G}}^2(\mathbf{y}(t)). \qquad (21)
\end{aligned}$$

We proceed to bound the second term on the RHS of (21). Using (11b) in Lemma 7

$$\begin{aligned}
&\sum_{i\in\mathcal{A}} \tilde{\beta}_{it}\mathbb{E}_t\left[\sum_{k\in\mathcal{K}_{it}} c_{ik}^+(\mathbf{z}_i(t))^2\right] \\
&\ge \sum_{i\in\mathcal{A}} \tilde{\beta}_{it}\mathbb{E}_{\mathcal{K}\sim\mathcal{U}_i}\left[\sum_{k\in\mathcal{K}} \frac{3}{4} c_{ik}^+(\mathbf{y}_i(t))^2\right] \\
&\quad - \sum_{i\in\mathcal{A}} 8\tilde{\beta}_{it}s_i\gamma_t^2L^2\big(2L^2 + \mathbb{E}_t\left[\|\mathbf{V}_i(t)\|^2\right]\big) \\
&\quad - \sum_{i\in\mathcal{A}} 16\tilde{\beta}_{it}s_i\gamma_t^2L^4\mu^2K\|\mathbf{e}_i(t) - \bar{\mathbf{e}}(t)\|^2.
\end{aligned}$$

Using Assumption 1-c and inequality $\tilde{\beta}_{it}s_i = \frac{(2-\beta_{it}s_i)\beta_{it}s_i}{L^2} \le \frac{1}{L^2}$ (see Lemma 7),

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

Using this bound in (21) with $\rho = \frac{\tilde{\beta}}{4C}$ and then rearranging terms yields (15). This completes the proof. □

Recall that $\beta_{it} \in [\underline{\beta}_i, \bar{\beta}_i] \subset (0, 2s_i^{-1})$. Thus, for all $i \in \mathcal{A}$ and $t \in \mathbf{N}$, we have $\tilde{\beta}_{it} = \frac{(2-s_i\beta_{it})\beta_{it}}{L^2} \ge \frac{(2-s_i\bar{\beta}_i)\underline{\beta}_i}{L^2} > 0$. As a result, $\tilde{\beta} = \inf_{i,t}\tilde{\beta}_{it}$ is strictly positive, and so is ρ.

Note that we introduced a positive sequence $\{\kappa_t : t \in \mathbf{N}\}$ in Lemma 8, which will be chosen appropriately for various intermediate results below. The following corollary is obtained by choosing $\mathbf{w} = P_{\mathcal{G}}(\mathbf{y}(t))$ and a constant sequence $\{\kappa_t = \rho/K : t \in \mathbf{N}\}$ in Lemma 8.

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

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

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

Because $\mathsf{d}_{\mathcal{G}}^2(\mathbf{y}(t+1)) \le \|\mathbf{y}(t+1) - P_{\mathcal{G}}(\mathbf{y}(t))\|^2$, an immediate consequence of Corollary 1 is

$$\mathbb{E}_t\left[\mathsf{d}_{\mathcal{G}}^2(\mathbf{y}(t+1))\right] \le \epsilon_d\mathsf{d}_{\mathcal{G}}^2(\mathbf{y}(t)) + \gamma_t^2D_1 + \gamma_t^2D_3a_t. \qquad (22)$$

This result hints at a form of contraction property of the sequence $\{\mathsf{d}_{\mathcal{G}}^2(\mathbf{y}(t)) : t \in \mathbf{N}\}$. If the sequence $\{a_t : t \in \mathbf{N}\}$ were bounded, together with Assumption 4, (22) would immediately imply $\lim_{t\to\infty}\mathsf{d}_{\mathcal{G}}^2(\mathbf{y}(t)) = 0$. 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 (22). Below, we shed some light on this through a customized linear coupling argument.

Lemma 9. *Suppose that Assumptions 1–3 hold. There exist $\epsilon, \epsilon_a, \epsilon_b \in (0,1)$ such that the following holds:*

$$\begin{aligned}
&\mathbb{E}_t\left[\mathsf{d}_{\mathcal{G}}^2(\mathbf{y}(t+1)) + \epsilon_b b_{t+1} + \epsilon_a a_{t+1}\right] \\
&\le \epsilon\big(\mathsf{d}_{\mathcal{G}}^2(\mathbf{y}(t)) + \epsilon_b b_t\big) + \gamma_t^2 a_t(1+\epsilon_b)D_3 + \bar{D}p_t, \qquad (23)
\end{aligned}$$

when $\gamma_t^2 \le \frac{\epsilon_a}{\epsilon_b D_3}$, where D_3 is the constant given in Corollary 1, $\bar{D} = \max\{2D_1, 3\epsilon_aC_1^2, 3\epsilon_aKL^2D_1\}$, and

$$\begin{aligned}
b_t &= \textstyle\sum_{s=0}^t \sigma^{t-s}\big(\mathsf{d}_{\mathcal{G}}^2(\mathbf{y}(s)) + \frac{1}{2}\gamma_s^2a_sD_3\big), \qquad (24)\\
p_t &= \gamma_t^2 + \sigma^{2(t+1)} + \textstyle\sum_{s=0}^t \sigma^{t-s}\gamma_s^2
\end{aligned}$$

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

Proof. Let us apply 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 sum over $i \in \mathcal{A}$:

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

This bound and Cauchy-Schwartz inequality imply

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

For the last inequality, we make use of the following inequalities: i) $\big(\sum_k h_km_k\big)^2 \le \big(\sum_k h_k^2\big)\big(\sum_k m_k^2\big)$ 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 $t \in \mathbf{N}$, 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 $\|\Delta\mathbf{g}_i(s)\| \le \sqrt{K}\|\Delta\mathbf{g}_i(s)\|_\infty$. Together with the assumed L-Lipschitz continuity of g_{ik} in Assumption 1-b,

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

Summing over $i \in \mathcal{A}$ and taking conditional expectation,

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

where the last inequality follows from Corollary 1 and $\epsilon_d=1-\rho<1$. Using this bound in (25) after taking the expectation,

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

with $C_3=4C_2^2NKL^2(1-\sigma)^{-1}$ and b_t given in (24), Note that

$$b_{t+1}=\sigma b_t+\mathsf{d}^2_{\mathcal{G}}(\mathbf{y}(t+1))+\gamma_{t+1}^2a_{t+1}\tfrac{D_3}{2}. \quad (27)$$

By taking conditional expectation of both sides and using (26),

$$\mathbb{E}_t\left[b_{t+1}\right]\le(\sigma+\gamma_{t+1}^2C_3D_3)b_t+\mathbb{E}_t\left[\mathsf{d}^2_{\mathcal{G}}(\mathbf{y}(t+1))\right]+\gamma_{t+1}^2D_3\omega_t,$$

where $\omega_t=C_1^2\sigma^{2t+2}+KL^2D_1\sum_{s=0}^t\sigma^{t-s}\gamma_s^2$. Using the bound in (22) for the second term on the RHS of the above inequality yields

$$\begin{aligned}\mathbb{E}_t\left[b_{t+1}\right]&\le(\sigma+\gamma_{t+1}^2C_3D_3)b_t+\epsilon_d\mathsf{d}^2_{\mathcal{G}}(\mathbf{y}(t))\\&\quad+\gamma_t^2a_tD_3+\tilde{p}_t,\end{aligned} \quad (28)$$

where $\tilde{p}_t=\gamma_{t+1}^2D_3\omega_t+\gamma_t^2D_1$.

We now couple (22), (26) and (28) as follows for any $\epsilon_a,\epsilon_b>0$

$$\begin{aligned}&\mathbb{E}_t\left[\mathsf{d}^2_{\mathcal{G}}(\mathbf{y}(t+1))+\epsilon_b b_{t+1}+\epsilon_a a_{t+1}\right]\\&\le\epsilon_d(\epsilon_b+1)\mathsf{d}^2_{\mathcal{G}}(\mathbf{y}(t))+\big(\epsilon_b\sigma+\gamma_{t+1}^2\epsilon_bC_3D_3+2\epsilon_aC_3\big)b_t\\&\quad+\gamma_t^2a_t(1+\epsilon_b)D_3+\bar{p}_t,\end{aligned} \quad (29)$$

where

$$\begin{aligned}\bar{p}_t&=2\epsilon_a\omega_t+\gamma_t^2D_1+\epsilon_b\tilde{p}_t\\&=2\epsilon_a\omega_t+\gamma_t^2D_1+\epsilon_b(\gamma_{t+1}^2D_3\omega_t+\gamma_t^2D_1)\\&=(2\epsilon_a+\epsilon_b\gamma_{t+1}^2D_3)\omega_t+(1+\epsilon_b)\gamma_t^2D_1.\end{aligned}$$

For any $\epsilon_b\in(0,1)$ and $\gamma_{t+1}^2\le\epsilon_a/(\epsilon_bD_3)$, we have

$$\begin{aligned}&\bar{p}_t\le 3\epsilon_a\omega_t+2\gamma_t^2D_1\\&=3\epsilon_aC_1^2\sigma^{2t+2}+3\epsilon_aKL^2D_1\sum_{s=0}^t\sigma^{t-s}\gamma_s^2+2\gamma_t^2D_1\le\bar{D}p_t,\end{aligned} \quad (30)$$

where $\bar{D}$ and p_t are given in the statement of the lemma. If we select ϵ_a and ϵ_b sufficiently small, there exists $\epsilon<1$ satisfying

$$\max\{\epsilon_d(\epsilon_b+1),\sigma+3\epsilon_a\epsilon_b^{-1}C_3\}<\epsilon. \quad (31)$$

This, together with $\gamma_{t+1}^2\le\epsilon_a/(\epsilon_bD_3)$, implies that

$$\epsilon_b\sigma+\gamma_{t+1}^2\epsilon_bC_3D_3+2\epsilon_aC_3\le\epsilon\epsilon_b. \quad (32)$$

It remains to plug this bound and (30) into (29) to obtain (23), thus proving the lemma. □

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

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. Note that under Assumption 4 we have $\gamma_t\to0$ as $t\to\infty$. This and Lemma 9 then imply

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

for all t sufficiently large. Note also that, for any $T\in\mathbf{N}$,

$$\begin{aligned}\textstyle\sum_{t=0}^T p_t&=\textstyle\sum_{t=0}^T\big(\gamma_t^2+\sigma^{2(t+1)}+\sum_{s=0}^t\sigma^{t-s}\gamma_s^2\big)\\&\le\tfrac{\sigma^2}{1-\sigma^2}+\textstyle\sum_{t=0}^T\gamma_t^2+\sum_{t=0}^T\sum_{s=0}^t\big(\sigma^{t-s}\gamma_s^2\big)\\&=\tfrac{\sigma^2}{1-\sigma^2}+\textstyle\sum_{t=0}^T\gamma_t^2+\sum_{s=0}^T\gamma_s^2\big(\sum_{k=0}^{T-s}\sigma^k\big)\\&\le\tfrac{\sigma^2}{1-\sigma^2}+\tfrac{2-\sigma}{1-\sigma}\textstyle\sum_{t=0}^T\gamma_t^2.\end{aligned} \quad (33)$$

Thus, under Assumption 4, $\sum_{t\in\mathbf{N}}p_t<\infty$. The convergence of corollary now follows from Lemma 5 because $\epsilon<1$. □

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

$$v_t=\mathsf{d}^2_{\mathcal{Y}^*}(\mathbf{y}(t))+\mathsf{d}^2_{\mathcal{G}}(\mathbf{y}(t))+\epsilon_bb_t+\epsilon_aa_t. \quad (34)$$

Define Φ^* to be the optimal value of the approximated problem in (2). We also consider the following running average terms:

$$\tilde{\mathbf{y}}(s,t)=\frac{\sum_{k=s}^t\gamma_k\mathbf{y}(k)}{\sum_{k=s}^t\gamma_k},\quad\tilde{\mathbf{x}}(s,t)=\frac{\sum_{k=s}^t\gamma_kP_{\mathcal{G}}\big(\mathbf{y}(k)\big)}{\sum_{k=s}^t\gamma_k}. \quad (35)$$

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

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

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

$$\begin{aligned}&\mathbb{E}\left[\Phi\big(\tilde{\mathbf{x}}(s,t)\big)\right]-\Phi^*+\frac{\rho}{\bar{\gamma}}\mathbb{E}\left[\|\tilde{\mathbf{y}}(s,t)-\tilde{\mathbf{x}}(s,t)\|^2\right]\le E_{s,t}\\&\text{with } E_{s,t}=\frac{\mathbb{E}\left[v_s\right]+\mathcal{O}(\sum_{k=s}^t p_k)}{2\sum_{k=s}^t\gamma_k},\quad 0\le s\le t.\end{aligned} \quad (36)$$

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

Proof. First, we use Lemma 8 with $\mathbf{w}=P_{\mathcal{Y}^*}(\mathbf{y}(t))$ for each $t\in\mathbf{N}$ to bound $\mathbb{E}_t\left[\mathsf{d}^2_{\mathcal{Y}^*}(\mathbf{y}(t+1))\right]$: under Assumptions 1–3,

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

where $u_t:=2\gamma_t\big[\Phi\big(P_{\mathcal{G}}(\mathbf{y}(t))\big)-\Phi^*\big]+2\rho\mathsf{d}^2_{\mathcal{G}}(\mathbf{y}(t))\ge0$. Choosing a positive sequence $\{\kappa_t=\gamma_t^2\mu^2L^2/(\epsilon\epsilon_a):t\in\mathbf{N}\}$,

$$\begin{aligned}&\mathbb{E}_t\left[\mathsf{d}^2_{\mathcal{Y}^*}(\mathbf{y}(t+1))\right]\\&\le\big(1+\kappa_tK\big)\mathsf{d}^2_{\mathcal{Y}^*}(\mathbf{y}(t))-u_t+\epsilon\epsilon_aa_t+\gamma_t^2(D_1+a_tD_2).\end{aligned} \quad (37)$$

By adding (23) and (37), we get the following bound on $\mathbb{E}_t[v_{t+1}]$:

$$\begin{aligned}&\mathbb{E}_t\left[\mathsf{d}^2_{\mathcal{Y}^*}(\mathbf{y}(t+1))+\mathsf{d}^2_{\mathcal{G}}(\mathbf{y}(t+1))+\epsilon_bb_{t+1}+\epsilon_aa_{t+1}\right]\\&\le\big(1+\kappa_tK\big)\mathsf{d}^2_{\mathcal{Y}^*}(\mathbf{y}(t))+\epsilon\big(\mathsf{d}^2_{\mathcal{G}}(\mathbf{y}(t))+\epsilon_bb_t+\epsilon_aa_t\big)\\&\quad-u_t+\gamma_t^2a_t(D_2+(1+\epsilon_b)D_3)+\bar{D}p_t\\&=\big(1+\kappa_tK\big)\mathsf{d}^2_{\mathcal{Y}^*}(\mathbf{y}(t))+\epsilon\big(\mathsf{d}^2_{\mathcal{G}}(\mathbf{y}(t))+\epsilon_bb_t\big)\\&\quad+\big(\epsilon+\gamma_t^2D_4\big)\epsilon_aa_t-u_t+\mathcal{O}(p_t),\end{aligned} \quad (38)$$

where $D_4=(D_2+(1+\epsilon_b)D_3)/\epsilon_a$ and $\mathcal{O}(p_t)=\bar{D}p_t+D_1\gamma_t^2$.

Part (i): From the definition of v_t in (34) and (38), we have

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

where $\alpha_t=\max\{\kappa_tK,\gamma_t^2D_4\}=\gamma_t^2\big(\max\{\mu^2L^2K/\epsilon,(D_2+(1+\epsilon_b)D_3)\}/\epsilon_a\big)$. Since $\sum_{t\in\mathbf{N}}p_t<\infty$ (from (33) in the proof of Corollary 2) and $\sum_{t\in\mathbf{N}}\gamma_t^2<\infty$ (Assumption 4), Lemma 6 tells us that the following holds w.p.1: (a) $\sum_{t\in\mathbf{N}}u_t<\infty$, hence $\sum_{t\in\mathbf{N}}\gamma_t(\Phi\big(P_{\mathcal{G}}(\mathbf{y}(t))\big)-\Phi^*)<\infty$ and $\sum_{t\in\mathbf{N}}\mathsf{d}^2_{\mathcal{G}}(\mathbf{y}(t))<\infty$,

and (b) $v_t \to v$ for some nonnegative RV v. Because $\sum_{t\in\mathbf{N}} \gamma_t = \infty$ (Assumption 4), (a) implies $\liminf_t \Phi\big(P_{\mathcal{G}}(\mathbf{y}(t))\big) = \Phi^*$. Also, as $\mathsf{d}^2_{\mathcal{G}}(\mathbf{y}(t)) + \epsilon_b b_t + \epsilon_a a_t \to 0$ (Corollary 2), (b) implies $\mathsf{d}^2_{\mathcal{Y}^*}(\mathbf{y}(t)) \to v$ as $t \to \infty$. Together with the continuity of Φ, they imply $v = 0$ w.p.1.

Part (ii): Suppose that $\mathbb{E}\left[\mathsf{d}^2_{\mathcal{Y}^*}(\mathbf{y}(t))\right]$ is bounded for all $t \in \mathbf{N}$, i.e., $\mathbb{E}\left[\mathsf{d}^2_{\mathcal{Y}^*}(\mathbf{y}(t))\right] = \mathcal{O}(1)$. Taking the expectation of (38), we obtain the following for all sufficiently large t.

$$\begin{aligned}
&\mathbb{E}\left[\mathsf{d}^2_{\mathcal{Y}^*}(\mathbf{y}(t+1)) + \mathsf{d}^2_{\mathcal{G}}(\mathbf{y}(t+1)) + \epsilon_b b_{t+1} + \epsilon_a a_{t+1}\right] \\
&\le \mathbb{E}\left[\mathsf{d}^2_{\mathcal{Y}^*}(\mathbf{y}(t)) + \epsilon\big(\mathsf{d}^2_{\mathcal{G}}(\mathbf{y}(t)) + \epsilon_b b_t + \epsilon_a a_t\big)\right] - \mathbb{E}\left[u_t\right] \\
&\quad + \gamma_t^2 \epsilon_a D_4 \mathbb{E}\left[a_t\right] + \mathcal{O}(p_t) \\
&\le \mathbb{E}\left[\mathsf{d}^2_{\mathcal{Y}^*}(\mathbf{y}(t)) + \mathsf{d}^2_{\mathcal{G}}(\mathbf{y}(t)) + \epsilon_b b_t + \epsilon_a a_t\right] - \mathbb{E}\left[u_t\right] + \mathcal{O}(p_t),
\end{aligned}$$

where we 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 $\big(\epsilon\epsilon_a + \epsilon_a D_4 \gamma_t^2\big)\mathbb{E}\left[a_t\right] \le \epsilon_a \mathbb{E}\left[a_t\right]$ (for all sufficiently large t) in the second. Using the definition of v_t in (34), after rearranging terms,

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

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

$$\textstyle\sum_{k=s}^t \mathbb{E}\left[u_k\right] \le \mathbb{E}\left[v_s\right] + \mathcal{O}(\sum_{k=s}^t p_k). \tag{39}$$

Next we bound the left-hand side of (39). Note that

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

where the last inequality follows from the convexity of $\Phi(\cdot)$ and $\|\cdot\|^2$ and the definitions of $\tilde{\mathbf{y}}(s,t)$ and $\tilde{\mathbf{x}}(s,t)$ in (35). Using this bound in (39) and dividing both sides by $2\sum_{k=2}^t \gamma_k$,

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

This proves statement (ii) of the theorem. □

Remark 1. The boundedness condition of $\mathbb{E}\left[\mathsf{d}^2_{\mathcal{Y}^*}(\mathbf{y}(t))\right]$ in Theorem 1(ii) 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].

Remark 2. One can show that, when the weight matrix W is fixed, any optimal point in $\mathcal{Y}^*$ gives rise to a Nash equilibrium of a state-based potential game similar to that of [10], in which the cost function of an agent accounts for both its own local cost and penalties based on its own and neighbors' estimates of constraint functions (eq. (8) in [10]). Thus, after the algorithm converges to an optimal point of (2) and the agents have consistent estimates of global constraint functions, no agent can decrease the aforementioned cost by unilaterally changing its local optimization variables. We refer a reader to [12, Section V] for a more detailed discussion on this connection.

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. Automat. Contr.*, 58(2):391–405, 2013.

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

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

[5] N. Chatzipanagiotis and M. M. Zavlanos. A distributed algorithm for convex constrained optimization under noise. *IEEE Trans. Automat. Contr.*, 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. Automat. Contr.*, 62(9):4405–4420, 2017.

[7] A. Cherukuri and J. Cortes. Distributed algorithms for convex network optimization under non-sparse equality constraints. In *Proc. 54th Annu. Allerton Conf.*, 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 Syst. Mag.*, 39(3):40–72, 2019.

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

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

[11] D. G. Luenberger. *Optimization by Vector Space Methods*. John Wiley & Sons, 1969.

[12] V.-S. Mai, R. J. La, T. Zhang, and A. Battou. End-to-end quality-of-service assurance with autonomous systems: 5G/6G case study. In *Proc. 19th IEEE Consumer Comm. Network. Conf.*, pages 644–651, 2022.

[13] A. Nedić. Random algorithms for convex minimization problems. *Math. Program.*, 129(2):225–253, 2011.

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

[15] A. Nedić, A. Olshevsky, and W. Shi. Improved convergence rates for distributed resource allocation. In *Proc. 2018 IEEE Conf. Decis. Contr.*, pages 172–177, 2018.

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

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

[18] B. T. Polyak. Random algorithms for solving convex inequalities. *Stud. Comput. Math.*, 8:409–422, 2001.

[19] K. Srivastava and A. Nedić. Distributed asynchronous constrained stochastic optimization. *IEEE J. Sel. Top. Signal Process.*, 5(4):772–790, 2011.

[20] S. Sundhar Ram, A. Nedić, and V. V. Veeravalli. Distributed stochastic subgradient projection algorithms for convex optimization. *J. Optim. Theory Appl.*, 14:514–545, 2010.

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

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