Zeroth-Order Feedback Optimization in Multi-Agent Systems: Tackling Coupled Constraints

Yingpeng Duan and Yujie Tang*

Abstract

This paper investigates distributed zeroth-order feedback optimization in multi-agent systems with coupled constraints, where each agent operates its local action vector and observes only zeroth-order information to minimize a global cost function subject to constraints in which the local actions are coupled. Specifically, we employ two-point zeroth-order gradient estimation with delayed information to construct stochastic gradients, and leverage the constraint extrapolation technique and the averaging consensus framework to effectively handle the coupled constraints. We also provide convergence rate and oracle complexity results for our algorithm, characterizing its computational efficiency and scalability by rigorous theoretical analysis. Numerical experiments are conducted to validate the algorithm's effectiveness.

1 Introduction

Distribution optimization has gained significant attention due to its wide-ranging applications in networked systems, large-scale machine learning, and beyond [1, 2, 3, 4]. In distributed optimization problems, the systems consist of agents that aim to achieve a global objective through communicating and coordinating with other agents via a network. One of the challenges in the study of distributed optimization is handling coupled constraints among the agents. Coupled constraints occur when the feasible decisions of one agent are dependent on the actions of the other agents, making optimization problem more intricate.

A variety of algorithms have been developed to address optimization problems with constraints, among which a common and effective class of approaches is the primal-dual methods. The primal-dual methods have a long-standing history that can date back to the seminal work [5], while still being an active topic in optimization in recent years due to their flexibility and scalability in large-scale problems [6]. Some recent works on primal-dual methods include, e.g., [7] which proposed a linearized augmented Lagrangian method for solving composite convex problems with both affine

*The authors are with the College of Engineering, Peking University, China (e-mail: zxycy@stu.pku.edu.cn, yujietang@pku.edu.cn).

equality and smooth nonlinear inequality constraints. The work [8] proposed an accelerated primal-dual algorithm with momentum to solve saddle-point problems with more general coupling between the primal and the dual variables. The work [9] proposed a constraint extrapolation technique for handling both convex and nonconvex functional constraints with stochastic gradient oracles. Besides centralized optimization, researchers have also investigated how to handle functional or coupled constraints in distributed optimization. For instance, [10] presented a distributed primal-dual algorithm leveraging a consensus framework. [11] developed a distributed algorithm based on dual decomposition and proximal minimization for minimizing a separable objective function with constraint $\sum_i g_i(x_i) \leq 0$. Other works that employed primal-dual techniques to efficiently manage coupled constraints in various distributed optimization scenarios include [12, 13, 14, 15], etc.

Despite these advancements, a significant limitation persists in many practical applications where agents lack access to the explicit form of their local cost functions or their gradients. Such cases, often studied under the realm of gradient-free or zeroth-order optimization, require agents to rely solely on observed outcomes (feedback) rather than analytical gradients to make decisions. Zeroth-order techniques have become increasingly prominent in recent research. In the centralized optimization context, [16, 17] studied one-point zeroth-order gradient estimation techniques for gradient-free optimization. The works [18, 19, 20] further advanced the field with two-point zeroth-order gradient estimation. Some more recent advancements include [21] which proposed one-point gradient estimation using residual feedback, [22] which studied escaping saddle points with zeroth-order methods in nonconvex optimization, and [23] which extended the constraint extrapolation technique to zeroth-order optimization. There is also rich existing literature on distributed zeroth-order optimization, including [24, 25, 26] that considered consensus optimization problems, [27] that investigated asynchronous zeroth-order algorithms in a distributed setting, and [28, 29] that studied multi-agent zeroth-order feedback optimization.

The problem setup considered in this paper is closely related to [28], where each agent i can only control its own local action vector x^i, and each local cost function f_i is influenced by the joint action profile $x = (x^1, ..., x^n)$. However, [28] did not address the challenge of coupled constraints, which are prevalent in many practical multi-agent systems. This paper considers coupled constraints of the form $\sum_{i=1}^n g_{ij}(x^i) \leq 0$, and aims to develop a zeroth-order feedback optimization algorithm that not only addresses the zeroth-order feedback requirements in cooperative multi-agent systems but also handle coupled constraints effectively.

1.1 Our Contributions

In this paper, we propose a distributed zeroth-order feedback optimization algorithm specifically designed to handle multi-agent systems with coupled constraints. We consider a problem setup that extends the one in [28] to incorporate coupled constraints of the form $\sum_{i=1}^n g_{ij}(x^i) \leq 0$. Our algorithm employs the zeroth-order gradient estimation technique from [19], and also utilizes the constraint extrapolation technique in [9, 23]. The resulting algorithm achieves efficient coordination without requiring access to explicit gradient information, while also being able to effectively manage the coupled constraints.

We provide detailed complexity analysis of our algorithm, including its convergence in terms of the objective value gap and constraint violation assessment. Specifically, we show that the number of zeroth-order queries per agent needed for the objective value gap and the constraint violation to fall below $\epsilon > 0$ is bounded by $O(d/\epsilon^2)$, where d is the problem dimension. This complexity bound matches typical results for stochastic first-order and zeroth-order methods.

2 Problem Formulation

In this paper, we consider the situation where there are n agents connected by a communication network. The topology of the network is specified by the undirected graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$, where $\mathcal{V} \coloneqq \{1, ..., n\}$ represents the set of nodes, and $\mathcal{E} \subseteq \mathcal{V} \times \mathcal{V}$ represents the set of edges. At each time step, agent i can exchange information only with its neighbors in the communication network. We assume that $\mathcal{G}$ is connected, and use b_{ij} to denote the distance (the length of the shortest path) between nodes i and j.

Additionally, each agent i is associated with a local action vector $x^i \in \mathcal{X}_i$, where $\mathcal{X}_i \subseteq \mathbb{R}^{d_i}$ represents the set of all possible local actions for agent i. We shall assume that each $\mathcal{X}_i$ is convex and compact. The joint action profile of the group of agents is represented by $x \coloneqq (x^1, ..., x^n) \in \mathcal{X}$, where $\mathcal{X} \coloneqq \prod_{i=1}^n \mathcal{X}_i \subseteq \mathbb{R}^d$ and we denote $d \coloneqq \sum_{i=1}^n d_i$. The goal of the group of agents is to solve the following optimization problem:

$$
\begin{aligned}
\underset{x^1,\ldots,x^n}{\text{minimize}} \quad & f_0(x) = \frac{1}{n}\sum_{i=1}^n f_i(x^1, ..., x^n) \\
\text{subject to} \quad & \sum_{i=1}^n g_{ij}(x^i) \le 0, \quad j = 1, \ldots, m, \\
& x_i \in \mathcal{X}_i, \quad i = 1, \ldots, n.
\end{aligned}
\tag{1}
$$

Here $f_0 : \mathbb{R}^d \to \mathbb{R}$ denotes the global objective function, and each $f_i : \mathbb{R}^d \to \mathbb{R}$ represents the local cost function of agent i. The inequalities $\sum_{i=1}^n g_{ij}(x^i) \le 0 \, j = 1, \ldots, m$ represent m coupled constraints among n agents, where $g_{ij} : \mathbb{R}^{d_i} \to \mathbb{R}$ for each $i = 1, \ldots, n$ and $j = 1, \ldots, m$. We also introduce the vector-valued function $g_i : \mathbb{R}^{d_i} \to \mathbb{R}^m$ defined as

$$
g_i(x^i) = \begin{bmatrix} g_{i1}(x^i) \\ \vdots \\ g_{im}(x^i) \end{bmatrix};
$$

with this notation, the m coupled constraints can be concisely written as $\sum_{i=1}^n g_i(x^i) \le 0$ with $\le$ being interpreted component-wise. Note that the local cost function f_i of each agent is influenced not only by its own action vector x^i, but by the joint action profile x. On the other hand, $g_i : \mathbb{R}^{d_i} \to \mathbb{R}^m$ for agent i depends solely on its own action vector x^i.

In this paper, we study the setting where each agent i can only access zeroth-order function value of its local cost f_i and constraint function g_{ij}, and gradients of f_i and g_{ij} are not available.

Furthermore, we impose the following mechanism of how each agent can access its associated zeroth-order information: First, each agent i first determines its local action x^i. Then, after the agents take their local actions $x^1, \ldots, x^n$, each agent i will receive a corresponding local cost $f_i(x) = f_i(x^1, ..., x^n)$ together with the value of its local constraint function $g_i(x^i)$. In other words, the function values of f_i and g_i can only be obtained through observation of feedback values after actions have been taken. Similar settings without explicit coupled constraints have been considered by [28, 29, 30, 31], etc.

Remark 1. In the problem formulation (1), we assume the coupled constraints take the form $\sum_{i=1}^n g_{ij}(x^i) \leq 0$. This form of coupled constraints is common and practically relevant, and has been investigated by many existing works including [10, 11, 12, 13, 15], etc. Handling coupled constraints of more general forms is out of the scope of this paper, but will be an important future direction to explore.

We now present some assumptions related to the problem.

Assumption 1. *The functions f_0 and g_{ij} for all i, j are convex, and there exists $Z \geq 0$ such that $\|g_i(x)\| \leq Z$ for all $x \in \mathcal{X}_i$ and $i = 1, \ldots, n$.*

Furthermore, the problem (1) *has an optimal primal-dual pair (x^*, y^*).*

Assumption 2. *Each function f_i is L_0-smooth, and each function g_{ij} is L_{ij}-smooth. In other words,*

$$\|\nabla f_i(x) - \nabla f_i(y)\| \leq L_0 \|x - y\|$$

for all $x, y \in \mathbb{R}^d$, and

$$\|\nabla g_{ij}(x) - \nabla g_{ij}(y)\| \leq L_{ij} \|x - y\|$$

for all $x, y \in \mathbb{R}^{d_i}$. Additionally, define

$$L_i := \left(\sum\nolimits_{j=1}^m L_{ij}^2\right)^{\frac{1}{2}} \quad \text{and} \quad L_g := \left(\sum\nolimits_{i=1}^n L_i^2\right)^{\frac{1}{2}}.$$

Assumption 3. *Each function f_i is M_0-Lipschitz continuous over $\mathcal{X}$, and each function g_{ij} is M_{ij}-Lipschitz continuous over $\mathcal{X}_i$. In other words,*

$$|f_i(x) - f_i(y)| \leq M_0 \|x - y\|$$

for all $x, y \in \mathcal{X}$, and

$$|g_{ij}(x) - g_{ij}(y)| \leq M_{ij} \|x - y\|$$

for all $x, y \in \mathcal{X}_i$. Additionally, define

$$M_i := \left(\sum\nolimits_{j=1}^m M_{ij}^2\right)^{\frac{1}{2}} \quad \text{and} \quad M_g := \left(\sum\nolimits_{i=1}^n M_i^2\right)^{\frac{1}{2}}.$$

3 Algorithm Design

To effectively tackle optimization problems where direct access to derivatives is not feasible, we turn our attention to zeroth-order optimization. Particularly, we leverage the technique of zeroth-order gradient estimation, where one uses randomly explored function values to construct a stochastic gradient. Given a smooth function $f : \mathbb{R}^d \to \mathbb{R}$ and an arbitrary point $x \in \mathbb{R}^d$, a typical zeroth-order gradient estimator for $\nabla f(x)$ is given by

$$G_f(x, u, z) = \frac{f(x + uz) - f(x - uz)}{2u} z.$$

Here, u is a positive real number called the smoothing radius; z is a random perturbation vector. Notably, existing works [19] have shown that when z is sampled from the Gaussian distribution $\mathcal{N}(0, I_d)$, the expected value of this gradient estimator yields

$$\mathbb{E}_{z\sim\mathcal{N}(0,I_d)}[G_f(x, u, z)] = \nabla f^u(x),$$

where $f^u(x) = \mathbb{E}_{y\sim\mathcal{N}(0,I_d)}[f(x+uz)]$ is a smoothed version of the original function f. This identity is the rationale behind using $G_f(x, u, z)$ as an estimator for the true gradient, even in the absence of derivative information.

Note that the main problem (1) contains coupled constraints of the form $\sum_{i=1}^n g_{ij}(x^i) \le 0$. To handle such constraints, we consider the framework of primal-dual methods applied on the Lagrangian function. The Lagrangian function associated with (1) is given by

$$\mathcal{L}(x, y) = f_0(x) + \left\langle y, \sum_{i=1}^n g_i(x^i) \right\rangle.$$

Here, $y \in \mathbb{R}^m$ servers as the Lagrange multiplier (i.e., the dual variable), which facilitates the incorporation of constraints into the algorithm design.

Our algorithm design consists of three core ingredients:

1. Primal-dual gradient method with constraint extrapolation: In this paper, we adapt the constraint extrapolation technique proposed in [9] to our multi-agent feedback optimization setup, to effectively handle the coupled constraints. Specifically, at time step t, each agent first generates two independent perturbation variables, z_t^i and $\hat{z}_t^i$, both drawn from the Gaussian distribution $\mathcal{N}(0, I_{d_i})$. We let $z_t \in \mathbb{R}^d$ denote the concatenation of $z_t^1, \ldots, z_t^n$, which follows the Gaussian distribution $\mathcal{N}(0, I_d)$. Then one constructs

$$\begin{aligned} G_0^i(t) &= \frac{f_0(x_t + uz_t) - f_0(x_t - uz_t)}{2u} z_t^i \\ &= \frac{1}{n} \sum_{j=1}^n \frac{f_j(x_t + uz_t) - f_j(x_t - uz_t)}{2u} z_t^i. \end{aligned} \tag{2}$$

This estimator essentially estimates the partial gradient of f_0 with respect to agent i's local action. In contrast, the gradient estimator for the constraint function is more complex due to the nature of g_i being a function that maps form $\mathbb{R}^{d_i}$ to $\mathbb{R}^m$:

$$G_{ij}(t) = \frac{g_{ij}(x_t^i + u\hat{z}_t^i) - g_{ij}(x_t^i - u\hat{z}_t^i)}{2u} z_t^i.$$

Note that in constructing $G_{ij}(t)$, we employ the random perturbation $\hat{z}_t^i$ rather than z_t^i. We then assemble $G_{ij}(t)$, $j = 1, \ldots, m$ into a matrix representation

$$G_i(t) = \begin{bmatrix} G_{i1}(t)^T \\ \vdots \\ G_{im}(t)^T \end{bmatrix} \in \mathbb{R}^{m \times d_i}.$$

Next, we introduce the quantity $\ell_G^i(t)$, which can be viewed as a linear approximation of g_i at the estimate x_t^i:

$$\ell_G^i(t) = g_i(x_{t-1}^i) + G_i(t-1)(x_t^i - x_{t-1}^i).$$

We then compute an "extrapolated" estimate s_t^i that incorporates the effects of the previous estimates

$$s_t^i = (1 + \theta_t)\ell_G^i(t) - \theta_t \ell_G^i(t-1). \tag{3}$$

where θ_t is a parameter that adjusts the influence of past information. This constraint extrapolation technique lies at the core of the algorithm proposed by [9, 23], and here we tailor it to our multi-agent zeroth-order feedback optimization setup. We shall later see that the choice $\theta_t = 1$ suffices to ensure convergence of our algorithm. The quantity $\sum_{i=1}^n s_t^i$ will serve as the "gradient" of the Lagrangian with respect to the dual variable in the primal-dual framework.

Next, each agent i generate yet another perturbation vector $\bar{z}_t^i \sim \mathcal{N}(0, I_d)$, and construct

$$H_{ij}(t) = \frac{g_{ij}(x_t^i + u\bar{z}_t^i) - g_{ij}(x_t^i - u\bar{z}_t^i)}{2u} \bar{z}_t^i. \tag{4}$$

It's not hard to see that $H_{ij}(t)$ share the same form with $G_{ij}(t)$ except that in 4 we employ $\bar{z}_t^i$ as the perturbation. The reason why we introduce three sets of perturbations is rather technical; we only remark that the independence of $z_t^i, \hat{z}_t^i, \bar{z}_t^i$ will be critical for convergence analysis.

With these components in place, the prototype of our proposed algorithm is given by

$$y_{t+1} = \operatorname*{arg\,min}_{y \geq 0, \|y\| \leq C} \left\{ -\left\langle \sum_{i=1}^n s_t^i, y \right\rangle + \frac{1}{2\mu_t} \|y - y_t\|_2^2 \right\}, \tag{5}$$

$$x_{t+1}^i = \operatorname*{arg\,min}_{x \in \mathcal{X}_i} \left\{ \langle V_t^i, x \rangle + \frac{1}{2\eta_t} \|x - x_t^i\|^2 \right\}, \tag{6}$$

where we denote

$$V_t^i = G_0^i(t) + \sum_{j=1}^m H_{ij}(t) \cdot [y_{t+1}]_j.$$

Here η_t and μ_t are the step sizes of the primal step and the dual step, respectively. $C \geq 0$ is a sufficiently large constant that bounds the norm of the optimal dual variable. $[y_{t+1}]_j$ denotes the jth component of the m-dimensional vector y_{t+1}.

However, the aforementioned algorithm prototype still faces two significant issues in our distributed scenarios:

- In constructing the partial gradient estimator (2), each agent i still needs to know the difference information $f_j(x_t+uz_t)-f_j(x_t-uz_t)$ from other agents $j \neq i$. We need to design an inter-agent communication protocol to address this issue.
- The dual iterate y_t in (5) is still a global variable whose update requires aggregating all s_t^i, which cannot be directly implemented in our distributed scenario.

To overcome these two issues, we introduce the next two core ingredients in our algorithm design.

2. Protocol for exchanging difference information between agents: To facilitate effective information exchange between agents, we draw upon the approach outlined in [28]. Specifically, for each agent i, we maintain an information array structured as follows

$$\begin{array}{|c|c|c|c|} \hline D_1^i(t) & D_2^i(t) & \cdots & D_n^i(t) \\ \hline \tau_1^i(t) & \tau_2^i(t) & \cdots & \tau_n^i(t) \\ \hline \end{array} \tag{7}$$

In this array, $D_j^i(t)$ represents the difference of f_j collected from agent j, while $\tau_j^i(t)$ denotes the time step at which $D_j^i(t)$ was collected by agent j. The pair $(D_j^i(t), \tau_j^i(t))$ represents the latest difference information available to agent i from agent j at time $\tau_j^i(t)$. The items in this array are update according to the following rules:

1. At each time step t, each agent i collects its own difference information as follows

$$D_i^i(t) = \frac{f_i(x_t+uz_t) - f_i(x_t-uz_t)}{2u} z_t^i. \tag{8}$$

This quantity can be obtained by letting each agent i apply the local actions $x_t^i \pm uz_t^i$ simultaneously and observing the associated local costs. Each agent i also records $\tau_i^i(t) = t$.

2. Each agent i retrieves the information arrays of its neighboring agents via the communication network. We let $(D_j^{k\to i}(t), \tau_j^{k\to i}(t))$ denote the jth column of agent kth array (7) received by agent i at time t.

3. Then, for each $j \neq i$, agent i identifies the neighbor k such that $\tau_j^{k\to i}(t)$ has the maximum value, which corresponds to the most up-to-date difference information of f_j. We then update $D_j^i(t)$ along with $\tau_j^i(t)$ to accord with this most up-to-date information.

The above process allows agent i to update its own array with the most recent difference information available from its peers. Particularly, we have $D_j^i(t) = D_j^j(\tau_j^i(t))$, and $\tau_j^i(t) = t - b_{ij}$ if no error occurs during communication; see [28].

Finally, we obtain a partial gradient estimator with delayed information for agent i, given by

$$G_0^i(t) = \frac{1}{n} \sum_{j=1}^{n} D_j^i(t)\, z^i_{\tau^i_j(t)}, \tag{9}$$

which replaces the original partial gradient estimator in (2).

3. Averaging consensus for the dual variable y_t: The averaging consensus method is a fundamental approach used in distributed systems to enable multiple agents to agree on a common value. We will apply the consensus technique in our context as follows. For $i = 1, \dots, n$, each agent i will keep a local copy y_t^i that serves as an estimate of the global dual variable y_t. At each time step t, agent i sends its current value y_t^i to all its neighbors j (i.e., $(i,j) \in \mathcal{E}$), and also receives the values y_t^j from its neighbors. Agent i then compute the weighted average

$$p_t^i = \sum_{j=1}^{n} W_{ij} y_t^i. \tag{10}$$

where $W = [W_{ij}] \in \mathbb{R}^{n \times n}$ is a double stochastic matrix and W_{ij} represents the weight associated with the connection between agents i and j. We assume $W_{ii} > 0$ for each i and $W_{ij} = 0$ if no communication link exists between agents i and j. We then modify the dual update step as

$$y_{t+1}^i = \operatorname*{arg\,min}_{y \geq 0, \|y\| \leq C} \left\{ - \left\langle s_t^i, y \right\rangle + \frac{1}{2\mu_t} \|y - p_t^i\|_2^2 \right\}. \tag{11}$$

It can be shown that this averaging consensus technique ensures all agents' y_t^i will converge towards a common value.

By summarizing the aforementioned ingredients, we obtain our proposed algorithm for multi-agent zeroth-order feedback optimization with coupled constraints, outlined in Algorithm 1. Note that the output of Algorithm 1 is a weighted average $\bar{x}_T = \sum_{t=0}^{T-1} \gamma_t x_{t+1} / \sum_{t=0}^{T-1} \gamma_t$; in the next section, we will present one way to set the weights γ_t that enjoys convergence guarantees.

Remark 2. Note that Step 7 of Algorithm 10 still requires projecting the dual variable back to the bounded region $\{y : \|y\| \leq C\}$, which seems common for primal-dual methods but is avoided by the original first-order constraint extrapolation method in [9]. We mention that this limitation is due to a technical point related to the bias of zeroth-order gradient estimation. We leave it as a future work to study how to eliminate the projection onto $\{y : \|y\| \leq C\}$ in the algorithm design.

4 Convergence Analysis

In this section, we present convergence analysis results of our proposed algorithm. The detailed proofs of these results will be postponed to the Supplementary Materials of this paper.

We first introduce some auxiliary quantities. We let $\bar{R}_i := \sup_{x \in \mathcal{X}_i} \|x\|$, and let $\bar{R} := \left(\sum_{i=1}^{n} \bar{R}_i^2\right)^{1/2}$; these quantities characterize the size of the feasible region for the primal variable. Then, we intro-

Algorithm 1 Multi-Agent Zeroth-order Feedback Optimization with Coupled Constraints (`MAZFO-CoupledCon`)

Parameters: Step sizes η_t, μ_t, extrapolation parameter θ_t, smoothing radius u, communication matrix W, weights γ_t.

for $t = 1, ..., T-1$ **do**

Agent i generates $z_t^i \sim \mathcal{N}(0, I_d)$, and updates $D_i^i(t)$ via (8) and sets $\tau_i^i(t) = t$.

Agent i generates $\hat{z}_t^i \sim \mathcal{N}(0, I_d)$, and constructs s_t^i via (3).

Agent i receives from its neighbor k the difference information $(D_j^{k\to i}(t), \tau_j^{k\to i}(t))$ for each $j \neq i$ and $k : (k,i) \in \mathcal{E}$, and updates

$$k_j^i(t) = \arg\max_{k:(k,i)\in\mathcal{E}} \tau_j^{k\to i}(t)$$

and $\tau_j^i(t) = \tau_j^{k_j^i(t)\to i}(t),\ D_j^i(t) = D_j^{k_j^i(t)\to i}(t)$ for $j \neq i$.

Agent i performs the averaging consensus step

$$p^i(t) = \sum_{j=1}^n W_{ij} y^j(t). \tag{10}$$

Agent i performs dual iteration

$$y_{t+1}^i = \arg\min_{y\geq 0, \|y\|\leq C} \Big\{ \langle -s_t^i, y\rangle + \frac{1}{2\mu_t}\|y - p_t^i\|_2^2 \Big\}. \tag{11}$$

Agent i generates $\bar{z} \sim \mathcal{N}(0, I_d)$ and constructs $H_{ij}(t)$ via (4).

Agent i constructs the partial gradient estimator

$$G_0^i(t) = \frac{1}{n}\sum_{j=1}^n D_j^i(t) z^i_{\tau_j^i(t)}. \tag{9}$$

Agent i updates

$$V_t^i = G_0^i(t) + \sum_{j=1}^m H_{ij}(t)\cdot [y_{t+1}^i]_j \tag{12}$$

$$x_{t+1}^i = \arg\min_{x\in\mathcal{X}_i} \Big\{ \langle V_t^i, x\rangle + \frac{1}{2\eta_t}\|x - x_t^i\|^2 \Big\}. \tag{13}$$

end for

Return: $\bar{x}_T = \sum_{t=0}^{T-1} \gamma_t x_{t+1} / \sum_{t=0}^{T-1} \gamma_t$.

duce $\bar{b}$ and $\bar{\mathfrak{b}}$ to quantify the connectivity of network, defined as

$$\bar{b} := \left(\frac{\sum_{i,j=1}^n b_{ij}^2}{n^2} \right)^{\frac{1}{2}}, \qquad \bar{\mathfrak{b}} := \left(\frac{\sum_{i,j=1}^n b_{ij}^2 d_i}{nd} \right)^{\frac{1}{2}}.$$

Intuitively, a smaller value of $\bar{b}$ or $\bar{\mathfrak{b}}$ corresponds to a higher concentration of nodes within the graph $\mathcal{G}$. We also define

$$\rho = \left\| W - \frac{1}{n}\mathbf{1}\mathbf{1}^T \right\| < 1.$$

Our main results are summarized in the following theorem.

Theorem 1. *Suppose Assumptions 1, 2 and 3 hold. Let $y_0^i = 0$ and $x_0^i = 0$, and set the parameters of Algorithm 1 as follows: $\theta_t = 1$, $\gamma_t = 1$, $\frac{1}{\eta_t} = L_0 + L_{\max} + \frac{1}{\eta}$ and $\mu_t = \mu$ with*

$$\eta = \frac{\bar{R}}{\sqrt{T\xi}}, \qquad \mu = \frac{C\sqrt{2n}}{\sqrt{T\zeta}}, \qquad u = \min\left\{ \frac{M_g}{(d+6)L_g}, \frac{1}{(d\sqrt{T}\max\{L_0, L_g\})^{\frac{1}{2}}} \right\},$$

where

$$\xi = (M_0\bar{\mathfrak{b}}\sqrt{d} + L_0\bar{b}d\bar{R} + 2\sqrt{3}\bar{\mathfrak{b}}dM_0)(24M_0^2 + 27M_g^2C^2)^{\frac{1}{2}} + 104M_0^2 d + 124M_g^2 dC^2,$$
$$\zeta = 403dM_g^2\bar{R} + \frac{1}{1-\rho}(6dZ^2 + 3M_g^2\bar{R} + 243\bar{R}dM_g^2).$$

Then, we have

$$\mathbb{E}[f_0(\bar{x}_T) - f_0(x^*)] \leq \frac{1}{\sqrt{T}}\left(2\bar{R}\sqrt{\xi} + C\sqrt{n\zeta} + \frac{2C+1}{2}\right) + \frac{(L_0 + L_{\max})\bar{R}^2}{T} + \frac{7C\sqrt{2n}}{T^{\frac{3}{2}}\sqrt{\zeta}}, \tag{14}$$

and

$$\mathbb{E}\left[\left\|\left[\sum_{i=1}^n g_i(\bar{x}_T^i)\right]_+\right\|\right] \leq \frac{1}{\sqrt{T}}\left(\bar{R}\sqrt{\xi} + \bar{R}\frac{(\xi + L_g\bar{R}C)}{\sqrt{\xi}} + 2C\sqrt{n\zeta} + C\right) + \frac{(L_0 + L_{\max})\bar{R}^2}{T} + \frac{7C\sqrt{2n}}{T^{\frac{3}{2}}\sqrt{\zeta}}, \tag{15}$$

where $\bar{x}_T = \sum_{t=0}^{T-1}\gamma_t x_{t+1} / \sum_{t=0}^{T-1}\gamma_t$, and $[\cdot]_+$ denotes taking the positive part component-wise of a vector.

Corollary 1. *Under Assumptions 1, 2 and 3, when the parameters of Algorithm 1 are properly chosen, the number of zeroth-order queries per agent to achieve $\mathbb{E}[f_0(\bar{x}_T) - f_0(x^*)] \leq \epsilon$ and $\mathbb{E}\left[\left\|\left[\sum_{i=1}^n g_i(\bar{x}_T^i)\right]_+\right\|\right] \leq \epsilon$ can be bounded by*

$$O\left(\frac{d}{\epsilon^2}\max\left\{\bar{b} + \bar{\mathfrak{b}}, \frac{n}{1-\rho}\right\}\right). \tag{16}$$

We can see that the oracle complexity of Algorithm 1 has a $O(\epsilon^{-2})$ dependence on the optimization accuracy ϵ, which is typical for stochastic first-order methods. The dimensional dependence is $O(d)$, which matches the typical results of zeroth-order optimization with two-point gradient estimation. The bound (16) also explicitly demonstrates the dependence of the oracle complexity on the network's topological characterizations $\bar{b}, \bar{\mathfrak{b}}, \rho$ as well as the number of agents n, which reveals the scalability of our proposed algorithm.

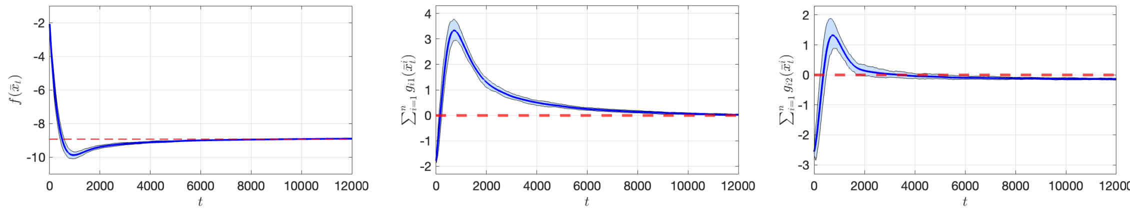

Figure 1: Convergence of Algorithm 1 on the numerical test case with constant step sizes $\eta_t = \mu_t = 1/500$.

Remark 3. In Theorem 1, we employ constant algorithmic parameters η_t, μ_t, etc. that depend on the total number of iterations T planned in advance. Such analysis paradigm is prevalent in the study of stochastic first-order methods, allowing simpler proofs while still effectively providing useful oracle complexity results. We believe that Algorithm 1 can also achieve convergence if we run the iterations indefinitely and employ diminishing step sizes, but detailed analysis seems tedious and is out of the scope of this paper.

5 Numerical Experiments

In this section, we conduct preliminary numerical experiments to test the performance of Algorithm 1. The test case consists of $n = 15$ agents, with the total dimension being $d = 40$. The local objective functions of the test case are quadratic functions of the form

$$f_i(x) = x^T A_i x + b_i^T x + c_i,$$

where $A_i \in \mathbb{R}^{d\times d}$ is positive definite, $b_i \in \mathbb{R}^d$ and $c_i \in \mathbb{R}$. The global objective function is then $f_0(x) = x^T A x + b^T x + c$ where $A = \frac{1}{n}\sum_{i=1}^n A_i,\ b = \frac{1}{n}\sum_{i=1}^n b_i,\ c = \frac{1}{n}\sum_{i=1}^n c_i$. We randomly generate one instance of A_i, b_i, c_i so that A has eigenvalues within $[0.1, 1.6]$. The test case has two inequality constraints $\sum_{i=1}^n g_{ij}(x^i) \le 0,\ i = 1, 2$ where

$$g_{ij}(x^i) = (x^i)^T P_{ij} x^i + q_{ij}^T x^i + r_{ij}.$$

Each $P_{ij} \in \mathbb{R}^{d_i \times d_i}$ is also randomly generated and positive definite with eigenvalues in $[0.1, 1.6]$. The optimal value of this test case is $f^* = -8.9368$. For the numerical experiments, we run 100 random trials of our algorithm for each choice of algorithmic parameters. We set the smoothing radius to be $u = 0.01$.

Figure 1 shows the convergence of Algorithm 1 with constant step sizes $\eta_t = \mu_t = 1/500$. We plot the global function value $f(\bar{x}_t)$ as well as the constraint function values $\sum_i g_{ij}(\bar{x}_t^i)$, where $\bar{x}_t = \frac{1}{t}\sum_{\tau=1}^t x_\tau$ and $\bar{x}_t^i = \frac{1}{t}\sum_{\tau=1}^t x_\tau^i$. The dark blue curve represents the average of the 100 random trials, while the light blue shade represents the the corresponding light shade represents the 5% to 95% quantile interval among these trials. We can see that, with proper constant step sizes,

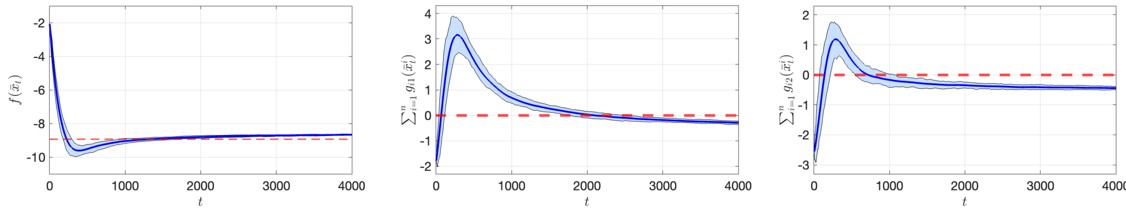

Figure 2: Convergence of Algorithm 1 on the numerical test case with constant step sizes $\eta_t = \mu_t = 1/200$.

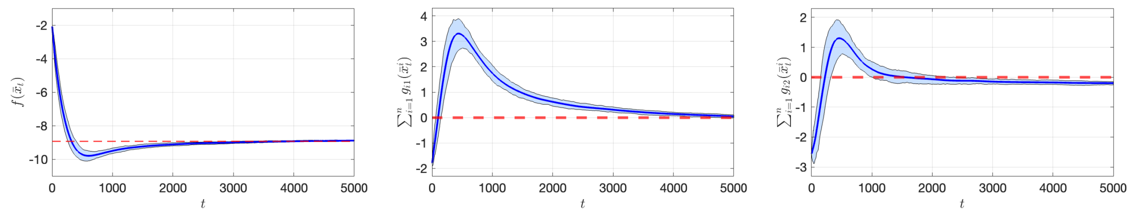

Figure 3: Convergence of Algorithm 1 on the numerical test case with diminishing step sizes $\eta_t = \mu_t = 1/(\sqrt{t} + 300)$.

Algorithm 1 is able to converge to a solution with small optimality gap and constraint violations that is consistent with our theoretical result.

Figure 2 shows the convergence of Algorithm 1 with larger constant step sizes $\eta_t = \mu_t = 1/200$. It is noteworthy that while the large step size facilitated faster convergence, it still exhibited a larger gap from the optimal value compared to Figure 1. In contrast, the smaller step size led to better accuracy but required an excessive number of iterations, resulting in slower convergence of the algorithm. We point out that these phenomena are typical for stochastic-gradient-descent-type algorithms with constant step sizes, and are also consistent with our theoretical result.

We have also tested the algorithm with diminishing step sizes $\eta_t = \mu_t = 1/(\sqrt{t} + 300)$, and the results are shown in Figure 3. We can see that the diminishing step size setting demonstrated an initial convergence rate comparable to that of the large step size, while ultimately achieving superior convergence accuracy. This observation suggests that in practice, one may use properly chosen diminishing step sizes to for better convergence behavior; it also indicates that the diminishing step size setting could be a promising topic for future investigation.

6 Conclusion

In this study, we proposed a distributed zeroth-order feedback optimization algorithm specifically designed for cooperative multi-agent systems facing coupled constraints of the form $\sum_i g_{ij}(x^i) \le 0$.

Our approach utilizes constraint extrapolation techniques and the averaging consensus framework to effectively tackle the challenges posed by coupled constraints in decentralized settings. Additionally, we provided theoretical results on its convergence rate and oracle complexity. Numerical experiments were conducted, revealing that employing a diminishing step size may yield even better performance. Some potential directions for future research include extension to coupled constraints of more general forms, removing projection of the dual iterate onto the bounded set $\{y : \|y\| \leq C\}$, analysis of the algorithm with diminishing step sizes, etc.

References

[1] T. Yang, X. Yi, J. Wu, Y. Yuan, D. Wu, Z. Meng, Y. Hong, H. Wang, Z. Lin, and K. H. Johansson, "A survey of distributed optimization," *Annual Reviews in Control*, vol. 47, pp. 278–305, 2019.

[2] A. Nedic, "Distributed gradient methods for convex machine learning problems in networks: Distributed optimization," *IEEE Signal Processing Magazine*, vol. 37, no. 3, pp. 92–101, 2020.

[3] S. Mao, Z. Dong, P. Schultz, Y. Tang, K. Meng, Z. Y. Dong, and F. Qian, "A finite-time distributed optimization algorithm for economic dispatch in smart grids," *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, vol. 51, no. 4, pp. 2068–2079, 2019.

[4] G. Wen, W. X. Zheng, and Y. Wan, "Distributed robust optimization for networked agent systems with unknown nonlinearities," *IEEE Transactions on Automatic Control*, vol. 68, no. 9, pp. 5230–5244, 2022.

[5] K. J. Arrow, L. Hurwicz, H. Uzawa, H. B. Chenery, S. Johnson, and S. Karlin, *Studies in linear and non-linear programming.* Stanford University Press Stanford, 1958, vol. 2.

[6] A. Nemirovski, "Prox-method with rate of convergence $o(1/t)$ for variational inequalities with lipschitz continuous monotone operators and smooth convex-concave saddle point problems," *SIAM Journal on Optimization*, vol. 15, no. 1, pp. 229–251, 2004.

[7] Y. Xu, "First-order methods for constrained convex programming based on linearized augmented Lagrangian function," *INFORMS Journal on Optimization*, vol. 3, no. 1, pp. 89–117, 2021.

[8] E. Y. Hamedani and N. S. Aybat, "A primal-dual algorithm with line search for general convex-concave saddle point problems," *SIAM Journal on Optimization*, vol. 31, no. 2, pp. 1299–1329, 2021.

[9] D. Boob, Q. Deng, and G. Lan, "Stochastic first-order methods for convex and nonconvex functional constrained optimization," *Mathematical Programming*, vol. 197, no. 1, pp. 215–279, 2023.

[10] T.-H. Chang, A. Nedić, and A. Scaglione, "Distributed constrained optimization by consensus-based primal-dual perturbation method," *IEEE Transactions on Automatic Control*, vol. 59, no. 6, pp. 1524–1538, 2014.

[11] A. Falsone, K. Margellos, S. Garatti, and M. Prandini, "Dual decomposition for multi-agent distributed optimization with coupling constraints," *Automatica*, vol. 84, pp. 149–158, 2017.

[12] I. Notarnicola and G. Notarstefano, "Constraint-coupled distributed optimization: A relaxation and duality approach," *IEEE Transactions on Control of Network Systems*, vol. 7, no. 1, pp. 483–492, 2019.

[13] S. Liang, G. Yin *et al.*, "Distributed smooth convex optimization with coupled constraints," *IEEE Transactions on Automatic Control*, vol. 65, no. 1, pp. 347–353, 2019.

[14] Y. Huang and J. Hu, "A primal decomposition approach to globally coupled aggregative optimization over networks," in *Proceedings of the 60th IEEE Conference on Decision and Control*, 2021, pp. 3830–3835.

[15] X. Wu, H. Wang, and J. Lu, "Distributed optimization with coupling constraints," *IEEE Transactions on Automatic Control*, vol. 68, no. 3, pp. 1847–1854, 2022.

[16] A. S. Nemirovskij and D. B. Yudin, *Problem Complexity and Method Efficiency in Optimization.* Wiley-Interscience, 1983.

[17] A. D. Flaxman, A. T. Kalai, and H. B. McMahan, "Online convex optimization in the bandit setting: gradient descent without a gradient," in *Proceedings of the Sixteenth Annual ACM-SIAM Symposium on Discrete Algorithms*, 2005, pp. 385–394.

[18] J. C. Duchi, M. I. Jordan, M. J. Wainwright, and A. Wibisono, "Optimal rates for zero-order convex optimization: The power of two function evaluations," *IEEE Transactions on Information Theory*, vol. 61, no. 5, pp. 2788–2806, 2015.

[19] Y. Nesterov and V. Spokoiny, "Random gradient-free minimization of convex functions," *Foundations of Computational Mathematics*, vol. 17, no. 2, pp. 527–566, 2017.

[20] O. Shamir, "An optimal algorithm for bandit and zero-order convex optimization with two-point feedback," *Journal of Machine Learning Research*, vol. 18, no. 52, pp. 1–11, 2017.

[21] Y. Zhang, Y. Zhou, K. Ji, and M. M. Zavlanos, "A new one-point residual-feedback oracle for black-box learning and control," *Automatica*, vol. 136, p. 110006, 2022.

[22] Z. Ren, Y. Tang, and N. Li, "Escaping saddle points in zeroth-order optimization: The power of two-point estimators," in *Proceedings of the 40th International Conference on Machine Learning*, 2023, pp. 28 914–28 975.

[23] A. Nguyen and K. Balasubramanian, “Stochastic zeroth-order functional constrained optimization: Oracle complexity and applications,” *INFORMS Journal on Optimization*, vol. 5, no. 3, pp. 256–272, 2023.

[24] D. Hajinezhad, M. Hong, and A. Garcia, “Zone: Zeroth-order nonconvex multiagent optimization over networks,” *IEEE Transactions on Automatic Control*, vol. 64, no. 10, pp. 3995–4010, 2019.

[25] Y. Tang, J. Zhang, and N. Li, “Distributed zero-order algorithms for nonconvex multiagent optimization,” *IEEE Transactions on Control of Network Systems*, vol. 8, no. 1, pp. 269–281, 2020.

[26] D. Yuan, L. Wang, A. Proutiere, and G. Shi, “Distributed zeroth-order optimization: Convergence rates that match centralized counterpart,” *Automatica*, vol. 159, p. 111328, 2024.

[27] Y. Shen, Y. Zhang, S. c. Nivison, Z. I. Bell, and M. M. Zavlanos, “Asynchronous zeroth-order distributed optimization with residual feedback,” in *2021 60th IEEE Conference on Decision and Control (CDC)*, 2021, pp. 3349–3354.

[28] Y. Tang, Z. Ren, and N. Li, “Zeroth-order feedback optimization for cooperative multi-agent systems,” *Automatica*, vol. 148, p. 110741, 2023.

[29] X. Zheng, T. Javidi, and B. Touri, “Zeroth-order non-convex optimization for cooperative multi-agent systems with diminishing step size and smoothing radius,” *IEEE Control Systems Letters*, vol. 7, pp. 2785–2790, 2023.

[30] J. R. Marden, S. D. Ruben, and L. Y. Pao, “A model-free approach to wind farm control using game theoretic methods,” *IEEE Transactions on Control Systems Technology*, vol. 21, no. 4, pp. 1207–1214, 2013.

[31] A. Menon and J. S. Baras, “Convergence guarantees for a decentralized algorithm achieving pareto optimality,” in *Proceedings of 2013 American Control Conference*, 2013, pp. 1932–1937.

7 Supplementary Materials

In this section, we provide detailed proofs for our main result Theorem 1.

7.1 Some Auxiliary Lemmas

Lemma 1. *Suppose $h : \mathbb{R}^d \to \mathbb{R}$ is an L-smooth and M-Lipschitz continuous function. Then we have*

$$|h^u(x) - h(x)| \le \min\left\{uM\sqrt{d}, \frac{1}{2}u^2Ld\right\}, \tag{17}$$

where $h^u(x) = \mathbb{E}_{z\sim\mathcal{N}(0,I_d)}[h(x+uz)]$.

Proof. Since h is an $L-$smooth function, we have

$$|h^u(x) - h(x)| = |\mathbb{E}_{z\sim\mathcal{N}(0,I_d)}[h(x+uz) - h(x)]| \le \left|\mathbb{E}_{z\sim\mathcal{N}(0,I_d)}[\langle\nabla h(x), uz\rangle + \frac{1}{2}\|uz\|^2]\right| = \frac{1}{2}u^2Ld,$$

and

$$|h^u(x) - h(x)| = |\mathbb{E}_{z\sim\mathcal{N}(0,I_d)}[h(x+uz) - h(x)]| \le uM\mathbb{E}_{z\sim\mathcal{N}(0,I_d)}[\|z\|] = uM\sqrt{d}.$$

□

Lemma 2. *Consider the following optimization problem*

$$\begin{aligned}
\text{minimize} \quad & f_0^u(x) = \sum_{i=1}^n \frac{1}{n} f_i^u(x^1, ..., x^n) \\
\text{subject to} \quad & \sum_{i=1}^n g_i^u(x^i) = \sum_{i=1}^n \begin{pmatrix} g_{i1}^u(x^i) \\ g_{i2}^u(x^i) \\ \cdots \\ g_{im}^u(x^i) \end{pmatrix} \preceq 0_m,
\end{aligned} \tag{18}$$

where $f_i^u = \mathbb{E}_{z\sim\mathcal{N}(0,I_d)}[f_i(x+uz)]$ and $g_{ij}^u = \mathbb{E}_{z\sim\mathcal{N}(0,I_d)}[g_{ij}(x+uz)]$, for all $i = 1, \dots, n$ and $j = 1, \dots, m$. Let an optimal primal-dual pair be denoted as $(x^{*,u}, y^{*,u})$, then we have

$$f_0(x^*) \le f_0^u(x^{*,u}) \le f_0^u(x^*). \tag{19}$$

Proof. Given the objective function and the constraints, we can establish the following inequality chain

$$\begin{aligned}
f_0(x_1^*, ..., x_n^*) + \langle y^*, \sum g_i(x_i^*)\rangle &\le f_0(x^{*,u}) + \langle y^*, \sum g_i(x^{*,u})\rangle \\
&\le f_0^u(x^{*,u}) + \langle y^*, \sum g_i^u(x^{*,u})\rangle \\
&\le f_0^u(x^{*,u}) + \langle y^{*,u}, \sum g_i^u(x^{*,u})\rangle.
\end{aligned}$$

By the complementary slackness condition, we have (19). □

7.2 A Critical Lemma for Establishing Convergence

Lemma 3. *Suppose $\{\gamma_t, \eta_t, \mu_t, \theta_t\}$ is a non-negative sequence that satisfying*

$$\gamma_t \theta_t = \gamma_{t-1}, \quad \frac{\gamma_t}{\mu_t} = \frac{\gamma_{t-1}}{\mu_{t-1}} \tag{20}$$

and

$$4M_i^2 \frac{\theta_t}{\theta_{t-1}} \leq \frac{\frac{1}{\eta_{t-2}} - L_0 - L_i}{12\mu_t}, \quad M_i^2 \theta_t \leq \frac{\frac{1}{\eta_{t-1}} - L_0 - L_i}{12\mu_t}, \tag{21}$$

$$\frac{4M_i^2}{\theta_{T-1}} \leq \frac{\frac{1}{\eta_{T-2}} - L_0 - L_i}{12\mu_{T-1}}, \quad M_i^2 \leq \frac{\frac{1}{\eta_{T-1}} - L_0 - L_i}{12\mu_{T-1}}. \tag{22}$$

Then for arbitrary $x \in \mathbb{R}^d$ and $y \in \mathbb{R}^m$, we have

$$\begin{aligned}
&\sum_{t=0}^{T-1} \gamma_t \Big[f_0^u(x_{t+1}) - f_0^u(x) + \sum_{i=1}^n \langle g_i^u(x_{t+1}^i), y \rangle - \sum_{i=1}^n \langle g_i^u(x^i), y_{t+1}^i \rangle \Big] \\
&\quad + \sum_{t=0}^{T-1} \gamma_t \Big[\sum_{i=1}^n \langle \delta_t^{G_i}, x_t^i - x^i \rangle - \sum_{i=1}^n \langle \delta_{t+1}^{F_i}, y_{t+1}^i - y \rangle \Big] \\
&\leq \frac{\gamma_0}{2\eta_0} \|x - x_0\|_2^2 - \frac{\gamma_{T-1}}{2\eta_{T-1}} \|x - x_T\|_2^2 + \sum_{i=1}^n \Big[\frac{\gamma_0}{2\mu_0} \|y - y_0^i\|^2 - \frac{\gamma_{T-1}}{12\mu_{T-1}} \|y - y_T^i\|^2 \Big] \\
&\quad + \sum_{i=1}^n \Bigg[\sum_{t=0}^{T-1} \frac{2\gamma_t}{\frac{1}{\eta_t} - L_0 - L_i} \Big[||\delta_t^{G_i}||^2 + \big(\frac{L_i \bar{R}_i}{2} [\|y\| - 1]_+ \big)^2 \Big] + \frac{3\gamma_{T-1}\mu_{T-1}}{2} \|q_T^i - \bar{q}_T^i\|^2 \\
&\quad + \sum_{t=1}^{T-1} \frac{3\gamma_t \theta_t^2 \mu_t}{2} \|q_t^i - \bar{q}_t^i\|^2 \Bigg] + \sum_{i=1}^n \sum_{t=0}^{T-1} \gamma_t \theta_t \langle q_t^i, p_t^i - y_t^i \rangle.
\end{aligned} \tag{23}$$

Here $q_t^i = \ell_G^i(t) - \ell_G^i(t-1)$ and $\bar{q}_t^i = \ell_{g^u}^i(t) - \ell_{g^u}^i(t-1)$ in which we denote

$$\ell_{g^u}^i(t) = g_i^u(x_{t-1}^i) + \nabla g_i^u(x_{t-1}^i)(x_t^i - x_{t-1}^i),$$

and

$$\delta_t^{G_i} = G_0^i(t) - \frac{\partial f_0^u}{\partial x^i}(x_t) + \sum_{j=1}^m [H_{ij}(t) - \nabla g_{ij}^u(x_t^i)][y_{t+1}^i]_j, \quad \delta_t^{F_i} = \ell_G^i(t) - \ell_{g^u}^i(t)$$

for all $i = 1, \ldots, n$.

Proof. Considering the first-order optimality condition of (11), we have the following

$$\Big\langle -s_t^i + \frac{1}{\mu_t}(y_{t+1}^i - p_t^i), y - y_{t+1}^i \Big\rangle \geq 0.$$

This implies that

$$\langle -s_t^i, y_{t+1}^i - y \rangle \leq \frac{1}{\mu_t} \langle y_{t+1}^i - p_t^i, y - y_{t+1}^i \rangle = \frac{1}{2\mu_t} \Big[\|y - p_t^i\|_2^2 - \|y_{t+1}^i - p_t^i\|_2^2 - \|y - y_{t+1}\|_2^2 \Big]. \tag{24}$$

Similarly, for (13), utilizing its first-order optimality condition, we obtain

$$\langle V_t^i, x_{t+1}^i - x^i \rangle \leq \frac{1}{\eta_t} \langle x_{t+1} - x_t^i, x - x_{t+1}^i \rangle = \frac{1}{2\eta_t} \Big[\|x - x_t^i\|_2^2 - \|x_{t+1}^i - x_t^i\|_2^2 - \|x - x_{t+1}^i\|_2^2 \Big]. \quad (25)$$

Denote $v_t^i = \frac{\partial f_0^u}{\partial x^i}(x_t) + \sum_{j=1}^m \nabla g_{ij}^u(x_t^i)[y_{t+1}^i]_j$. Due to the convexity of g_i, we have

$$\begin{aligned} & \langle v_t^i, x_{t+1}^i - x^i \rangle \\ = & \left\langle \frac{\partial f_0^u}{\partial x^i}(x_t), x_{t+1}^i - x^i \right\rangle + \sum_{j=1}^m [y_{t+1}^i]_j \langle \nabla g_{ij}^u(x_t^i), x_{t+1}^i - x_t^i \rangle + \sum_{j=1}^m [y_{t+1}^i]_j \langle \nabla g_{ij}^u(x_t^i), x_t^i - x^i \rangle \\ \geq & \left\langle \frac{\partial f_0^u}{\partial x^i}(x_t), x_{t+1}^i - x^i \right\rangle + \langle y_{t+1}^i, \ell_{g^u}^i(t+1) - g_i^u(x^i) \rangle. \end{aligned} \quad (26)$$

Noting that $\delta_t^{G_i} = V_t^i - v_t^i$ and considering (24), (25) and (26), we get

$$\begin{aligned} & \left\langle \frac{\partial f_0^u}{\partial x^i}(x_t), x_{t+1}^i - x^i \right\rangle + \langle \delta_t^{G_i}, x_{t+1}^i - x^i \rangle \\ & - \langle g_i^u(x^i), y_{t+1}^i \rangle + \langle \ell_{g^u}^i(t+1), y_{t+1}^i \rangle - \langle s_t^i, y_{t+1}^i - y \rangle \\ \leq & \frac{1}{2\eta_t} \Big[\|x^i - x_t^i\|_2^2 - \|x_{t+1}^i - x_t^i\|_2^2 - \|x^i - x_{t+1}^i\|_2^2 \Big] \\ & + \frac{1}{2\mu_t} \Big[\|y - p_t^i\|_2^2 - \|y_{t+1}^i - p_t^i\|_2^2 - \|y - y_{t+1}^i\|_2^2 \Big]. \end{aligned} \quad (27)$$

Since g_{ij}^u is an L_{ij}-smooth function and $L_i = \sqrt{\sum_{j=1}^m L_{ij}^2}$, we can see that

$$g_i^u(x_{t+1}^i) - \ell_{g^u}^i(t+1) = g_i^u(x_{t+1}^i) - \Big[g_i^u(x_t) + \nabla g_i^u(x_t^i)(x_{t+1}^i - x_t^i) \Big] \leq \frac{L_i}{2} \|x_{t+1}^i - x_t^i\|^2,$$

and according to the Cauchy–Schwarz inequality, we further get

$$\langle g_i^u(x_{t+1}^i) - \ell_{g^u}^i(t+1), y \rangle \leq \|y\| C_{t+1}^i,$$

where $C_{t+1}^i = \frac{L_i}{2} \|x_{t+1}^i - x_t^i\|^2$. Then

$$\begin{aligned} \|y\| C_{t+1}^i &= \frac{L_i}{2} (\|y\| - 1) \|x_{t+1}^i - x_t^i\|^2 + \frac{L_i}{2} \|x_{t+1}^i - x_t^i\|^2 \\ &\leq \frac{L_i}{2} [\|y\| - 1]_+ \|x_{t+1}^i - x_t^i\|^2 + \frac{L_f}{2} \|x_{t+1}^i - x_t^i\|^2 \\ &\leq \frac{L_i}{2} \|x_{t+1}^i - x_t^i\|^2 + \frac{L_i \bar{R}_i}{2} [\|y\| - 1]_+ \|x_{t+1}^i - x_t^i\|. \end{aligned} \quad (28)$$

Next, from $q_t^i = \ell_G^i(t) - \ell_G^i(t-1)$ and $\delta_t^{F_i} = \ell_G^i(t) - \ell_{g^u}^i(t)$, we can see that

$$
\begin{aligned}
&\langle \ell_{g^u}^i(t+1), y_{t+1}^i\rangle - \langle g_i^u(x_{t+1}^i), y\rangle - \langle s_t^i, y_{t+1}^i - y\rangle \\
&= \langle \ell_{g^u}^i(t+1), y_{t+1}^i\rangle - \left[\langle g_i^u(x_{t+1}^i) - \ell_{g^u}^i(t+1), y\rangle\right] - \langle \ell_{g^u}^i(t+1), y\rangle - \langle s_t^i, y_{t+1}^i - y\rangle \\
&\geq \langle \ell_{g^u}^i(t+1), y_{t+1}^i\rangle - \langle \ell_{g^u}^i(t+1), y\rangle - \langle s_t^i, y_{t+1}^i - y\rangle - \|y\|C_{t+1}^i \\
&= \langle \ell_{g^u}^i(t+1) - s_t^i, y_{t+1}^i - y\rangle - \|y\|C_{t+1}^i \\
&= \langle \ell_{g^u}^i(t+1) - \ell_G^i(t) - \theta_t q_t^i, y_{t+1}^i - y\rangle - \|y\|C_{t+1}^i \\
&= \langle \ell_{g^u}^i(t+1) + \ell_G^i(t+1) - \ell_G^i(t) - \ell_G^i(t+1) - \theta_t q_t^i, y_{t+1}^i - y\rangle - \|y\|C_{t+1}^i \\
&= \langle q_{t+1}^i, y_{t+1}^i - y\rangle - \theta_t\langle q_t^i, y_{t+1}^i - y\rangle - \langle \delta_{t+1}^{F_i}, y_{t+1}^i - y\rangle - \|y\|C_{t+1}^i \\
&= \langle q_{t+1}^i, y_{t+1}^i - y\rangle - \theta_t\langle q_t^i, y_{t+1}^i - p_t^i\rangle - \theta_t\langle q_t^i, p_t^i - y\rangle - \langle \delta_{t+1}^{F_i}, y_{t+1}^i - y\rangle - \|y\|C_{t+1}^i.
\end{aligned}
\tag{29}
$$

Combining (27), (28) and (29), we obtain

$$
\begin{aligned}
&\left\langle \frac{\partial f_0^u}{\partial x^i}(x_t), x_{t+1}^i - x^i \right\rangle + \langle g_i^u(x_{t+1}^i), y\rangle - \langle g_i^u(x^i), y_{t+1}^i\rangle + \langle \delta_t^{G_i}, x_t^i - x^i\rangle - \langle \delta_{t+1}^{F_i}, y_{t+1}^i - y\rangle \\
&\quad + \langle q_{t+1}^i, y_{t+1}^i - y\rangle - \theta_t\langle q_t^i, p_t^i - y\rangle + \frac{1}{2}L_0\|x_{t+1}^i - x_t^i\|^2 \\
&\leq \frac{1}{2\eta_t}\left[\|x^i - x_t^i\|_2^2 - \|x^i - x_{t+1}^i\|_2^2\right] + \left[\frac{1}{2\mu_t}\|y - y_t^i\|_2^2 - \frac{1}{2\mu_t}\|y - y_{t+1}^i\|_2^2\right] + \theta_t\langle q_t^i, y_{t+1}^i - p_t^i\rangle \\
&\quad - \frac{1}{2\mu_t}\|y_{t+1}^i - p_t^i\|_2^2 + \langle \delta_t^{G_i}, x_t^i - x_{t+1}^i\rangle + \frac{1}{2\mu_t}\left[\|y - p_t^i\|_2^2 - \|y - y_t^i\|_2^2\right] \\
&\quad + \mathcal{H}^i(y)\|x_{t+1}^i - x_t^i\| - \frac{1}{2}\left(\frac{1}{\eta_t} - L_i - L_0\right)\|x_{t+1}^i - x_t^i\|_2^2.
\end{aligned}
\tag{30}
$$

where we denote $\mathcal{H}^i(y) = \frac{L_i \bar{R}_i}{2}[\|y\| - 1]_+$. Now since $\bar{q}_t^i = \ell_{g^u}^i(t) - \ell_{g^u}^i(t-1)$, by multiplying both sides of the above inequality by γ_t and then summing over t, we obtain

$$
\begin{aligned}
&\sum_{t=0}^{T-1}\gamma_t\left[\left\langle \frac{\partial f_0^u}{\partial x^i}(x_t), x_{t+1}^i - x^i \right\rangle + \langle g_i^u(x_{t+1}^i), y\rangle - \langle g_i^u(x^i), y_{t+1}^i\rangle + \frac{L_0}{2}\|x_{t+1}^i - x_t^i\|_2^2\right] \\
&\quad + \sum_{t=0}^{T-1}\gamma_t\left[\langle \delta_t^{G_i}, x_t^i - x^i\rangle - \langle \delta_{t+1}^{F_i}, y_{t+1}^i - y\rangle + \langle q_{t+1}^i, y_{t+1}^i - y\rangle - \theta_t\langle q_t^i, p_t^i - y\rangle\right] \\
&\leq \frac{\gamma_0}{2\eta_0}\|x^i - x_0^i\|_2^2 - \frac{\gamma_{T-1}}{2\eta_{T-1}}\|x^i - x_T^i\|_2^2 + \frac{\gamma_0}{2\mu_0}\|y - y_0^i\|_2^2 - \frac{\gamma_{T-1}}{2\mu_{T-1}}\|y - y_T^i\|_2^2 \\
&\quad + \sum_{t=0}^{T-1}\left[\gamma_t\theta_t\langle q_t - \bar{q}_t^i, y_{t+1}^i - p_t^i\rangle + \gamma_t\theta_t\langle \bar{q}_t^i, y_{t+1}^i - p_t^i\rangle\right] - \sum_{t=0}^{T-1}\frac{\gamma_t}{2\mu_t}\|y_{t+1}^i - p_t^i\|^2 \\
&\quad + \sum_{t=0}^{T-1}\gamma_t\langle \delta_t^{G_i}, x_t^i - x_{t+1}^i\rangle + \sum_{t=0}^{T-1}\frac{\gamma_t}{2\mu_t}\left[\|y - p_t^i\|_2^2 - \|y - y_t^i\|_2^2\right] \\
&\quad - \sum_{t=0}^{T-1}\left[\frac{\gamma_t}{2}\left(\frac{1}{\eta_t} - L_0 - L_i\right)\|x_{t+1}^i - x_t^i\|_2^2 - \gamma_t\mathcal{H}^i(y)\|x_{t+1}^i - x_t^i\|\right].
\end{aligned}
\tag{31}
$$

Based on (20), we have

$$
\begin{aligned}
&\sum_{t=0}^{T-1}[\gamma_t\langle q_{t+1}^i, y_{t+1}^i - y\rangle - \gamma_t\theta_t\langle q_t^i, p_t^i - y\rangle] \\
&= \sum_{t=0}^{T-1}\Big[\gamma_t\langle q_{t+1}^i, y_{t+1}^i - y\rangle - \gamma_t\theta_t\langle q_t^i, y_t^i - y\rangle - \gamma_t\theta_t\langle q_t^i, p_t^i - y_t^i\rangle\Big] \\
&= \gamma_{T-1}\langle q_T^i, y_T^i - y\rangle - \sum_{t=0}^{T-1}\gamma_t\theta_t\langle q_t^i, p_t^i - y_t^i\rangle.
\end{aligned}
\tag{32}
$$

Combining (32) and (31) leads to

$$
\begin{aligned}
&\sum_{t=0}^{T-1}\gamma_t\left[\left\langle\frac{\partial f_0^u}{\partial x^i}(x_t), x_{t+1}^i - x^i\right\rangle + \langle g_i^u(x_{t+1}^i), y\rangle - \langle g_i^u(x^i), y_{t+1}^i\rangle + \frac{L_0}{2}\|x_{t+1}^i - x_t^i\|^2\right] \\
&\quad + \sum_{t=0}^{T-1}\gamma_t\Big[\langle\delta_t^{G_i}, x_t^i - x^i\rangle - \langle\delta_{t+1}^{F_i}, y_{t+1}^i - y\rangle\Big] + \gamma_{T-1}\langle q_T^i, y_T^i - y\rangle \\
&\le \frac{\gamma_0}{2\eta_0}\|x^i - x_0^i\|_2^2 - \frac{\gamma_{T-1}}{2\eta_{T-1}}\|x^i - x_T^i\|_2^2 + \frac{\gamma_0}{2\mu_0}\|y - y_0^i\|_2^2 - \frac{\gamma_{T-1}}{2\mu_{T-1}}\|y - y_T^i\|_2^2 \\
&\quad + \sum_{t=0}^{T-1}\Big[\gamma_t\theta_t\langle q_t - \bar{q}_t^i, y_{t+1}^i - p_t^i\rangle + \gamma_t\theta_t\langle\bar{q}_t^i, y_{t+1}^i - p_t^i\rangle\Big] - \sum_{t=0}^{T-1}\frac{\gamma_t}{2\mu_t}\|y_{t+1}^i - p_t^i\|^2 \\
&\quad + \sum_{t=0}^{T-1}\gamma_t\langle\delta_t^{G_i}, x_t^i - x_{t+1}^i\rangle + \sum_{t=0}^{T-1}\frac{\gamma_t}{2\mu_t}\Big[\|y - p_t^i\|_2^2 - \|y - y_t^i\|_2^2\Big] + \sum_{t=0}^{T-1}\gamma_t\theta_t\langle q_t^i, p_t^i - y_t^i\rangle. \\
&\quad + \sum_{t=0}^{T-1}\left[\frac{\gamma_t}{2}\left(\frac{1}{\eta_t} - L_0 - L_i\right)\|x_{t+1}^i - x_t^i\|_2^2 - \gamma_t\mathcal{H}^i(y)\|x_{t+1}^i - x_t^i\|\right].
\end{aligned}
\tag{33}
$$

For $\bar{q}_t^i$, note that

$$
\begin{aligned}
\|\bar{q}_t^i\| &= \|\ell_{g^u}^i(t) - \ell_{g^u}^i(t-1)\| \\
&= \|g_i^u(x_{t-1}^i) + \nabla g_i^u(x_{t-1}^i)(x_t^i - x_{t-1}^i) - g_i^u(x_{t-2}^i) - \nabla g_i^u(x_{t-2}^i)(x_{t-1}^i - x_{t-2}^i)\| \\
&\le \|g_i^u(x_{t-1}^i) - g_i^u(x_{t-2}^i)\| + \|\nabla g_i^u(x_{t-1}^i)^T(x_t^i - x_{t-1}^i)\| + \|\nabla g_i^u(x_{t-2}^i)^T(x_{t-1}^i - x_{t-2}^i)\| \\
&\le 2M_i\|x_{t-1}^i - x_{t-2}^i\| + M_i\|x_t^i - x_{t-1}^i\|,
\end{aligned}
$$

where we used the fact that g_{ij} is M_{ij}-Lipschitz and let $M_i = \sqrt{\sum_{j=1}^m M_{ij}^2}$.

Now, based on (21) and Young's inequality, we have

$$
\begin{aligned}
&\gamma_t\theta_t\langle \bar{q}_t^i, y_{t+1}^i - p_t^i\rangle - \frac{\gamma_t}{3\mu_t}\|y_{t+1}^i - p_t^i\|^2 - \frac{\gamma_{t-2}(\frac{1}{\eta_{t-2}} - L_0 - L_i)}{8}\|x_{t-1}^i - x_{t-2}^i\|_2^2 \\
&\quad - \frac{\gamma_{t-1}(\frac{1}{\eta_{t-1}} - L_0 - L_i)}{8}\|x_t^i - x_{t-1}^i\|_2^2 \\
&\le \gamma_t\theta_t\|\bar{q}_t^i\|_2\|y_{t+1}^i - p_t^i\|_2 - \frac{\gamma_t}{3\mu_t}\|y_{t+1}^i - p_t^i\|^2 - \frac{\gamma_{t-2}(\frac{1}{\eta_{t-2}} - L_0 - L_i)}{8}\|x_{t-1}^i - x_{t-2}^i\|_2^2 \\
&\quad - \frac{\gamma_{t-1}(\frac{1}{\eta_{t-1}} - L_0 - L_i)}{8}\|x_t^i - x_{t-1}^i\|_2^2 \\
&\le 2M_i\gamma_t\theta_t\|x_{t-1}^i - x_{t-2}^i\|\|y_{t+1}^i - p_t^i\| - \frac{\gamma_t}{6\mu_t}\|y_{t+1}^i - p_t^i\|^2 \\
&\quad + M_i\gamma_t\theta_t\|x_t^i - x_{t-1}^i\|\|y_{t+1}^i - p_t^i\| - \frac{\gamma_t}{6\mu_t}\|y_{t+1}^i - p_t^i\|^2 \\
&\quad - \frac{\gamma_{t-2}(\frac{1}{\eta_{t-2}} - L_0 - L_i)}{8}\|x_{t-1}^i - x_{t-2}^i\|^2 - \frac{\gamma_{t-1}(\frac{1}{\eta_{t-1}} L_0 - L_i)}{8}\|x_t^i - x_{t-1}^i\|^2 \\
&\le 0.
\end{aligned} \tag{34}
$$

Applying Cauchy–Schwarz inequality and Young's inequality, we can justify that

$$
\begin{aligned}
\gamma_t\theta_t\langle q_t^i - \bar{q}_t^i, y_{t+1}^i - p_t^i\rangle - \frac{\gamma_t}{6\mu_t}\|y_{t+1}^i - p_t^i\|^2 &\le \frac{3\gamma_t\theta_t^2\mu_t}{2}\|q_t^i - \bar{q}_t^i\|^2 \\
\langle \gamma_t\delta_t^{G_i}, x_t^i - x_{t+1}^i\rangle - \frac{\gamma_t(\frac{1}{\eta_t} - L_0 - L_i)}{8}\|x_{t+1}^i - x_t^i\|_2^2 &\le \frac{2\gamma_t}{\frac{1}{\eta_t} - L_0 - L_i}\|\delta_t^{G_i}\|^2 \\
\gamma_t\mathcal{H}^i(y)\|x_{t+1}^i - x_t^i\| - \frac{\gamma_t(\frac{1}{\eta_t} - L_0 - L_i)}{8}\|x_{t+1}^i - x_t^i\|_2^2 &\le \frac{2\gamma_t}{\frac{1}{\eta_t} - L_0 - L_i}\mathcal{H}^i(y)^2.
\end{aligned} \tag{35}
$$

As a result, by combining (33), (34) and (35), we can obtain

$$
\begin{aligned}
&\sum_{t=0}^{T-1}\gamma_t\left[\left\langle \frac{\partial f_0^u}{\partial x^i}(x_t), x_{t+1}^i - x^i\right\rangle + \langle g_i^u(x_{t+1}^i), y\rangle - \langle g_i^u(x^i), y_{t+1}^i\rangle + \frac{L_0}{2}\|x_{t+1}^i - x_t^i\|^2\right] \\
&\quad + \sum_{t=0}^{T-1}\gamma_t\Big[\langle \delta_t^{G_i}, x_t^i - x^i\rangle - \langle \delta_{t+1}^{F_i}, y_{t+1}^i - y\rangle\Big] + \gamma_{T-1}\langle q_T, y_T^i - y\rangle \\
&\le \frac{\gamma_0}{2\eta_0}\|x^i - x_0^i\|_2^2 - \frac{\gamma_{T-1}}{2\eta_{T-1}}\|x^i - x_T^i\|_2^2 + \frac{\gamma_0}{2\mu_0}\|y - y_0^i\|_2^2 - \frac{\gamma_{T-1}}{2\mu_{T-1}}\|y - y_T^i\|_2^2 \\
&\quad + \sum_{t=0}^{T-1}\Big[\frac{3\gamma_t\theta_t^2\mu_t}{2}\|q_t^i - \bar{q}_t^i\|^2 + \frac{2\gamma_t}{\frac{1}{\eta_t} - L_0 - L_i}\|\delta_t^{G_i}\|^2 + \frac{2\gamma_t}{\frac{1}{\eta_t} - L_0 - L_i}\mathcal{H}^i(y)^2\Big] \\
&\quad - \frac{\gamma_{T-2}(\frac{1}{\eta_{T-2}} - L_0 - L_i)}{8}\|x_{T-1}^i - x_{T-2}^i\|_2^2 - \frac{\gamma_{T-1}(\frac{1}{\eta_{T-1}} - L_0 - L_i)}{4}\|x_T^i - x_{T-1}^i\|_2^2 \\
&\quad + \sum_{t=0}^{T-1}\gamma_t\theta_t\langle q_t^i, p_t^i - y_t^i\rangle + \sum_{t=0}^{T-1}\gamma_t\Big[\frac{1}{2\mu_t}\|y - p_t^i\|_2^2 - \frac{1}{2\mu_t}\|y - y_t^i\|_2^2\Big].
\end{aligned} \tag{36}
$$

Similarly, base on (22), we get

$$\begin{aligned}
&- \gamma_{T-1}\langle \bar{q}_T^i, y_T^i - y\rangle - \frac{\gamma_{T-1}}{3\mu_{T-1}}\|y - y_T^i\|^2 \\
&- \frac{\gamma_{T-2}(\frac{1}{\eta_{T-2}} - L_0 - L_i)}{8}\|x_{T-1}^i - x_{T-2}^i\|_2^2 - \frac{\gamma_{T-1}(\frac{1}{\eta_{T-1}} - L_0 - L_i)}{4}\|x_T^i, x_{T-1}^i\|_2^2 \\
\le\ & M_i\gamma_{T-1}\|x_T^i - x_{T-1}^i\|\|y_T^i - y\| - \frac{\gamma_{T-1}}{12\mu_{T-2}}\|y - y_T^i\|^2 \\
&+ 2M_i\gamma_{T-1}\|x_{T-1}^i - x_{T-2}^i\|\|y_T^i - y\| - \frac{\gamma_{T-1}}{6\mu_{T-1}}\|y - y_T^i\|^2 - \frac{\gamma_{T-1}}{12\mu_{T-1}}\|y_T^i - y\|^2 \\
&- \frac{\gamma_{T-1}(\frac{1}{\eta_{T-1}} - L_0 - L_i)}{4}\|x_T^i - x_{T-1}^i\|_2^2 - \frac{\gamma_{T-2}(\frac{1}{\eta_{T-2}} - L_0 - L_i)}{4}\|x_{T-1}^i - x_{T-2}^i\|^2 \\
\le\ & - \frac{\gamma_{T-1}}{12\mu_{T-1}}\|y_T^i - y\|^2,
\end{aligned} \tag{37}$$

and again using Young's inequality, we have

$$-\gamma_{T-1}\langle q_T^i - \bar{q}_T^i, y_T^i - y\rangle - \frac{\gamma_{T-1}}{6\mu_{T-1}}\|y - y_T\|^2 \le \frac{3\gamma_{T-1}\mu_{T-1}}{2}\|q_T^i - \bar{q}_T^i\|^2. \tag{38}$$

Combining (36), (37) and (38), we obtain

$$\begin{aligned}
&\sum_{t=0}^{T-1}\gamma_t\left[\left\langle \frac{\partial f_0^u}{\partial x^i}(x_t), x_{t+1}^i - x^i\right\rangle + \langle g_i^u(x_{t+1}^i), y\rangle - \langle g_i^u(x^i), y_{t+1}^i\rangle + \frac{L_0}{2}\|x_{t+1}^i - x_t^i\|^2\right] \\
&+ \sum_{t=0}^{T-1}\gamma_t\Big[\langle \delta_t^{G_i}, x_t^i - x^i\rangle - \langle \delta_{t+1}^{F_i}, y_{t+1}^i - y\rangle\Big] \\
\le\ & \frac{\gamma_0}{2\eta_0}\|x^i - x_0^i\|_2^2 - \frac{\gamma_{T-1}}{2\eta_{T-1}}\|x^i - x_T^i\|_2^2 + \frac{\gamma_0}{2\mu_0}\|y - y_0^i\|_2^2 - \frac{\gamma_{T-1}}{12\mu_{T-1}}\|y - y_T^i\|_2^2 \\
&+ \sum_{t=0}^{T-1}\frac{2\gamma_t}{\frac{1}{\eta_t} - L_0 - L_i}\Big[\|\delta_t^{G_i}\|^2 + \big(\frac{L_i\bar{R}_i}{2}\big[\|y\| - 1\big]_+\big)^2\Big] + \sum_{t=0}^{T-1}\frac{3\gamma_t\theta_t^2\mu_t}{2}\|q_t^i - \bar{q}_t^i\|^2 \\
&+ \frac{3\gamma_{T-1}\mu_{T-1}}{2}\|q_T^i - \bar{q}_T^i\|^2 + \sum_{t=0}^{T-1}\gamma_t\theta_t\langle q_t^i, p_t^i - y_t^i\rangle + \sum_{t=0}^{T-1}\gamma_t\Big[\frac{1}{2\mu_t}\|y - p_t^i\|_2^2 - \frac{1}{2\mu_t}\|y - y_t^i\|_2^2\Big].
\end{aligned} \tag{39}$$

Next, we note that

$$\begin{aligned}
\sum_{i=1}^{n}\langle \frac{\partial f_0^u}{\partial x^i}(x_t), x_{t+1}^i - x^i\rangle &= \langle \nabla f_0^u(x_t), x_{t+1} - x\rangle \\
&= \langle \nabla f_0^u(x_t), x_{t+1} - x_t\rangle + \langle \nabla f_0^u(x_t), x_t - x\rangle \\
&\ge \Big[f_0^u(x_{t+1}) - f_0^u(x_t) - \frac{L_0}{2}\|x_{t+1} - x_t\|^2\Big] + \Big[f_0^u(x_t) - f_0^u(x)\Big] \\
&= f_0^u(x_{t+1}) - f_0^u(x) - \frac{L_0}{2}\|x_{t+1} - x_t\|^2,
\end{aligned} \tag{40}$$

where we used the smoothness and convexity of f_0. Then, due to convexity of $f(x) = \|x\|^2$ and using Jensen's inequality, we have

$$\sum_{i=1}^{n} \|y - p_t^i\|_2^2 \leq \sum_{i=1}^{n}\sum_{j=1}^{n} W_{ij}\|y - y_t^j\|_2^2 = \sum_{j=1}^{n} \|y - y_t^j\|_2^2.$$

The last equality holds because we exchanged the order of summation. Thus,

$$\sum_{i=1}^{n}\sum_{t=0}^{T-1} \frac{\gamma_t}{2\mu_t}\|y - p_t^i\|_2^2 - \sum_{i=1}^{n}\sum_{t=0}^{T-1} \frac{\gamma_t}{2\mu_t}\|y - y_t^i\|_2^2 \leq 0. \tag{41}$$

Finally, by combining (39) with (40), summing over i and plugging in the inequality (41), we complete the proof. □

7.3 Building Blocks for Convergence Rate and Constraint Violation Analysis

In this subsection, we will derive the initial forms of the convergence rate and constraint violation bounds using Lemma 3.

Lemma 4. *For all $T \geq 1$, we have*

$$\begin{aligned}
&\Gamma_T \mathbb{E}\Big[f_0(\bar{x}_T) - f_0(x^*)\Big] \\
\leq\ & \frac{\gamma_0}{2\eta_0}\|x^{*,u} - x_0\|_2^2 + \sum_{i=1}^{n} \frac{\gamma_0}{2\mu_0}\|y_0^i\|^2 + \sum_{i=1}^{n}\Bigg[\sum_{t=0}^{T-1} \frac{2\gamma_t}{\frac{1}{\eta_t} - L_0 - L_i}\mathbb{E}[\|\delta_t^{G_i}\|^2] \\
&+ \frac{3\gamma_{T-1}\mu_{T-1}}{2}\mathbb{E}\big[\|q_T^i - \bar{q}_T^i\|^2\big] + \sum_{t=1}^{T-1}\frac{3\gamma_t\theta_t^2\mu_t}{2}\mathbb{E}\big[\|q_t^i - \bar{q}_t^i\|^2\big]\Bigg] \\
&- \sum_{t=0}^{T-1}\gamma_t \mathbb{E}\Big[\sum_{i=1}^{n}\langle \delta_t^{G_i}, x_t^i - x^i\rangle - \sum_{i=1}^{n}\langle \delta_{t+1}^{F_i}, y_{t+1}^i - y\rangle\Big]_{x = x^{*,u}, y=0} \\
&+ \mathbb{E}\Big[\sum_{i=1}^{n}\sum_{t=0}^{T-1}\gamma_t\theta_t\langle q_t^i, p_t^i - y_t^i\rangle\Big] + \Gamma_T\frac{1}{2}u^2 L_0 d,
\end{aligned} \tag{42}$$

where $\Gamma_T = \sum_{t=0}^{T-1}\gamma_t$.

Proof. By Jensen's inequality, we have

$$f_0(\bar{x}_T) = f_0\left(\frac{\sum_{t=0}^{T-1}\gamma_t x_{t+1}}{\sum_{t=0}^{T-1}\gamma_t}\right) \leq \frac{1}{\sum_{t=0}^{T-1}\gamma_t}\sum_{t=0}^{T-1}\Big[\gamma_t f_0(x_{t+1})\Big] = \frac{1}{\Gamma_T}\sum_{t=0}^{T-1}\Big[\gamma_t f_0(x_{t+1})\Big].$$

Taking $x = x^{*,u}$ and $y = 0$, we see that the first line on the left-hand side of (23) becomes

$$\begin{aligned}
&\sum_{t=0}^{T-1} \gamma_t \Big[f_0^u(x_{t+1}) - f_0^u(x) + \sum_{i=1}^n \langle g_i^u(x_{t+1}^i), y \rangle - \sum_{i=1}^n \langle g_i^u(x^i), y_{t+1}^i \rangle \Big]_{|x=x^{*,u}, y=0} \\
&= \sum_{t=0}^{T-1} \gamma_t \Big[f_0^u(x_{t+1}) - f_0^u(x^{*,u}) - \sum_{i=1}^n \langle g_i^u((x^i)^{*,u}), y_{t+1}^i \rangle \Big].
\end{aligned}$$

Then, from (17) and (19), we have

$$|f_0^u(x^{*,u}) - f_0(x^*)| \leq \frac{1}{2} u^2 L_0 d.$$

Summarizing the above inequalities, we obtain

$$\begin{aligned}
\Gamma_T \Big[f_0(\bar{x}_T) - \big(f_0(x^*) + \frac{1}{2} u^2 L_0 d \big) \Big] &\leq \sum_{t=0}^{T-1} \gamma_t \Big[f_0^u(x_{t+1}) - f_0^u(x^{*,u}) \Big] \\
&\leq \sum_{t=0}^{T-1} \gamma_t \Big[f_0^u(x_{t+1}) - f_0^u(x^{*,u}) - \sum_{i=1}^n \langle g_i^u((x^i)^{*,u}), y_{t+1}^i \rangle \Big],
\end{aligned} \tag{43}$$

where the second inequality follows by noting that $g_i^u((x^i)^{*,u}) \leq 0$ and $y_{t+1}^i \geq 0$ leads to the inequality $\sum_{i=1}^n \langle g_i^u((x^i)^{*,u}), y_{t+1}^i \rangle \leq 0$.

By combining (23) and (43), we complete the proof. □

Lemma 5 ([9])**.** *Let $\rho_0, \ldots, \rho_j$ be a sequence in $\mathbb{R}^n$, and let S be a convex set in $\mathbb{R}^n$. Define the sequence v_t for $t = 0, 1, \ldots$ such that $v_0 \in S$ and*

$$v_{t+1} = \arg\min_{x \in S} \langle \rho_t, x \rangle + \frac{1}{2} \|x - v_t\|_2^2. \tag{44}$$

Then, for any $x \in S$ and $t \geq 0$, we have

$$\langle \rho_t, v_t - x \rangle \leq \frac{1}{2} \|x - v_t\|_2^2 - \frac{1}{2} \|x - v_{t+1}\|_2^2 + \frac{1}{2} \|\rho_t\|_2^2 \tag{45}$$

$$\sum_{t=0}^{j} \langle \rho_t, v_t - x \rangle \leq \frac{1}{2} \|x - v_0\|_2^2 + \frac{1}{2} \sum_{t=0}^{j} \|\rho_t\|_2^2. \tag{46}$$

Before proceeding to the next lemma that provides a preliminary form for the constraint violation bound, we introduce some extra notations. We let $R \coloneqq \|y^{*,u}\| + 1$ and

$$\mathcal{B}_+^2(R) = \{x \in \mathbb{R}^m : \|x\|_2 \leq R,\, x \geq 0\}.$$

We also denote

$$\hat{y} \coloneqq (\|y^{*,u}\| + 1) \frac{[\sum_{i=1}^n g_i^u(\bar{x}_T^i)]_+}{\|[\sum_{i=1}^n g_i^u(\bar{x}_T^i)]_+\|} \in \mathcal{B}_+^2(R).$$

We further define an auxiliary sequence $(y_v^i(t))_{t \geq 0}$ for each i such that $y_v^i(0) \coloneqq y_0^i$, and for all $t \geq 0$,

$$y_v^i(t+1) \coloneqq \arg\min_{y \in \mathcal{B}_+^2(R)} \Big\{ \mu_{t-1} \langle \delta_t^{F_i}, y \rangle + \frac{1}{2} \|y - y_v^i(t)\| \Big\}.$$

Lemma 6. *For all $T \geq 1$, we have*

$$\Gamma_T \mathbb{E}\left[\left\|\left[\sum_{i=1}^n g_i(\bar{x}_T^i)\right]_+\right\|\right] \leq \frac{\gamma_0}{2\eta_0}\|x^{*,u} - x_0\|_2^2 + \sum_{i=1}^n \frac{\gamma_0}{2\mu_0}\Big[\mathbb{E}\big[\|\hat{y} - y_0^i\|^2 + \|\hat{y} - y_v^i(1)\|^2\big]\Big]$$
$$+ \sum_{i=1}^n \left[\sum_{t=0}^{T-1} \frac{2\gamma_t}{\frac{1}{\eta_t} - L_0 - L_i}\Big[\mathbb{E}[\|\delta_t^{G_i}\|^2] + \big(\frac{L_i \bar{R}_i}{2}\|y^{*,u}\|_2\big)\Big]\right.$$
$$\left. + \frac{3\gamma_{T-1}\mu_{T-1}}{2}\mathbb{E}\big[\|q_T^i - \bar{q}_T^i\|^2\big] + \sum_{t=1}^{T-1}\frac{3\gamma_t\theta_t^2\mu_t}{2}\mathbb{E}\big[\|q_t^i - \bar{q}_t^i\|^2\big]\right]$$
$$- \sum_{t=0}^{T-1}\gamma_t \mathbb{E}\Big[\sum_{i=1}^n \langle \delta_t^{G_i}, x_t^i - x^i\rangle - \sum_{i=1}^n \langle \delta_{t+1}^{F_i}, y_{t+1}^i - y_v^i(t+1)\rangle\Big]_{x=x^{*,u}, y=\hat{y}}$$
$$+ \mathbb{E}\Big[\sum_{i=1}^n \sum_{t=0}^{T-1}\gamma_t\theta_t\langle q_t^i, p_t^i - y_t^i\rangle\Big] + \sum_{i=1}^n \mathbb{E}\Big[\sum_{t=0}^{T-1}\frac{\gamma_t\mu_t}{2}\|\delta_{t+1}^{F_i}\|_2^2\Big]. \tag{47}$$

Proof. According to Lemma 5, we have

$$\mu_t\langle \delta_{t+1}^{F_i}, y_v^i(t+1) - y\rangle \leq \frac{1}{2}\|y - y_v^i(t+1)\|_2^2 - \frac{1}{2}\|y - y_v^i(t+2)\|_2^2 + \frac{\mu_t^2}{2}\|\delta_{t+1}^{F_i}\|_2^2.$$

Multiplying both sides of the inequality by $\frac{\gamma_t}{\mu_t}$ and summing from $t = 0$ to $T-1$, we obtain

$$\sum_{t=0}^{T-1}\frac{\gamma_t\mu_t}{\mu_t}\langle \delta_{t+1}^{F_i}, y_v^i(t+1) - y\rangle \leq \sum_{t=0}^{T-1}\frac{\gamma_t}{2\mu_t}\|y - y_v^i(t+1)\|_2^2 - \sum_{t=0}^{T-1}\frac{\gamma_t}{2\mu_t}\|y - y_v^i(t+2)\|_2^2 + \sum_{t=0}^{T-1}\frac{\gamma_t\mu_t^2}{2\mu_t}\|\delta_{t+1}^{F_i}\|_2^2.$$

Summing over i and noting the second relation in (20), we obtain

$$\sum_{i=1}^n \sum_{t=0}^{T-1}\gamma_t\langle \delta_{t+1}^{F_i}, y_v^i(t+1) - y\rangle \leq \sum_{i=1}^n \frac{\gamma_0}{2\mu_0}\|y - y_v^i(1)\|_2^2 + \sum_{i=1}^n\sum_{t=0}^{T-1}\frac{\gamma_t\mu_t}{2}\|\delta_{t+1}^{F_i}\|_2^2. \tag{48}$$

From (48) and (23), we further get

$$\sum_{t=0}^{T-1}\gamma_t\Big[f_0^u(x_{t+1}) - f_0^u(x) + \sum_{i=1}^n \langle g_i^u(x_{t+1}^i), y\rangle - \sum_{i=1}^n \langle g_i^u(x^i), y_{t+1}^i\rangle\Big]$$
$$+ \sum_{t=0}^{T-1}\gamma_t\Big[\sum_{i=1}^n\langle \delta_t^{G_i}, x_t^i - x^i\rangle - \sum_{i=1}^n \langle \delta_{t+1}^{F_i}, y_{t+1}^i - y_v^i(t+1)\rangle\Big]$$
$$\leq \frac{\gamma_0}{2\eta_0}\|x, x_0\|_2^2 + \sum_{i=1}^n \frac{\gamma_0}{2\mu_0}\Big[\|y - y_0^i\|^2 + \|y - y_v^i(1)\|^2\Big] \tag{49}$$
$$+ \sum_{i=1}^n\left[\sum_{t=0}^{T-1}\frac{2\gamma_t}{\frac{1}{\eta_t} - L_0 - L_i}\Big[\|\delta_t^{G_i}\|^2 + \big(\frac{L_i\bar{R}_i}{2}[\|y\| - 1]_+\big)^2\Big] + \frac{3\gamma_{T-1}\mu_{T-1}}{2}\|q_T^i - \bar{q}_T^i\|^2\right.$$
$$\left. + \sum_{t=1}^{T-1}\frac{3\gamma_t\theta_t^2\mu_t}{2}\|q_t^i - \bar{q}_t^i\|^2 + \sum_{t=0}^{T-1}\frac{\gamma_t\mu_t}{2}\|\delta_{t+1}^{F_i}\|_2^2\right] + \sum_{i=1}^n\sum_{t=0}^{T-1}\gamma_t\theta_t\langle q_t^i, p_t^i - y_t^i\rangle.$$

Then, by Jensen's inequality, we have

$$
\begin{aligned}
&\mathbb{E}\Big[f_0^u(\bar{x}_T) - f_0^u(x^{*,u}) + \sum_{i=1}^n \langle g_i^u(\bar{x}_T^i), \hat{y}\rangle - \sum_{i=1}^n \langle g_i^u((x^i)^{*,u}), \bar{y}_T\rangle\Big] \\
&\leq \frac{1}{\Gamma_T}\mathbb{E}\left[\sum_{t=0}^{T-1} \gamma_t \Big[f_0^u(x_{t+1}) - f_0^u(x^{*,u}) + \sum_{i=1}^n \langle g_i^u(x_{t+1}^i), \hat{y}\rangle - \sum_{i=1}^n \langle g_i^u((x^i)^{*,u}), y_{t+1}^i\rangle\Big]\right]
\end{aligned}
\tag{50}
$$

Note that, for the problem (18), its Lagrangian function is given by

$$
\mathcal{L}^u(x, y) = f_0^u(x) + \left\langle \sum_{i=1}^n g_i^u(x^i), y \right\rangle.
$$

Since $\mathcal{L}^u(\bar{x}_T, y^{*,u}) - \mathcal{L}^u(x^{*,u}, y^{*,u}) \geq 0$, we can see that

$$
f_0^u(\bar{x}_T) + \left\langle \sum_{i=1}^n g_i^u(\bar{x}_T^i), y^{*,u} \right\rangle - f_0^u(x^{*,u}) \geq 0. \tag{51}
$$

Additionally, we have

$$
\left\langle y^{*,u}, \sum_{i=1}^n g_i^u(\bar{x}_T^i) \right\rangle \leq \left\langle y^{*,u}, \left[\sum_{i=1}^n g_i^u(\bar{x}_T^i)\right]_+ \right\rangle \leq \|y^{*,u}\|_2 \left\| \left[\sum_{i=1}^n g_i^u(\bar{x}_T^i)\right]_+ \right\|_2. \tag{52}
$$

Combining (51)) and (52), we obtain

$$
f_0^u(\bar{x}_T) + \|y^{*,u}\|_2 \left\| \left[\sum_{i=1}^n g_i^u(\bar{x}_T^i)\right]_+ \right\|_2 - f_0^u(x^{*,u}) \geq 0.
$$

Moreover,

$$
\begin{aligned}
&\mathcal{L}^u(\bar{x}_T, \hat{y}) - \mathcal{L}^u(x^{*,u}, \bar{y}_T) \\
&\geq \mathcal{L}^u(\bar{x}_T, \hat{y}) - \mathcal{L}^u(x^{*,u}, y^{*,u}) \\
&= f_0^u(\bar{x}_T) + (\|y^{*,u}\|_2 + 1) \left\| \left[\sum_{i=1}^n g_i(\bar{x}_T^i)\right]_+ \right\| - f_0^u(x^{*,u}) \\
&= f_0^u(\bar{x}_T) + \|y^{*,u}\|_2 \left\| \left[\sum_{i=1}^n g_i^u(\bar{x}_T^i)\right]_+ \right\| - f_0^u(x^{*,u}) + \left\| \left[\sum_{i=1}^n g_i^u(\bar{x}_T^i)\right]_+ \right\|.
\end{aligned}
$$

Therefore,

$$
\begin{aligned}
\left\| \left[\sum_{i=1}^n g_i^u(\bar{x}_T^i)\right]_+ \right\| &\leq \mathcal{L}^u(\bar{x}_T, \hat{y}) - \mathcal{L}^u(x^{*,u}, \bar{y}_T) \\
&= f_0^u(\bar{x}_T^i) + \left\langle \sum_{i=1}^n g_i^u(\bar{x}_T^i), \hat{y} \right\rangle - f_0^u(x^{*,u}) - \left\langle \bar{y}_T, \sum_{i=1}^n g_i^u(x^i)^{*,u} \right\rangle.
\end{aligned}
\tag{53}
$$

Combining (50) and (53), we obtain

$$
\begin{aligned}
&\mathbb{E}\left[\left\|\left[\sum_{i=1}^{n} g_i^u(\bar{x}_T^i)\right]_+\right\|\right] \\
&\leq \frac{1}{\Gamma_T}\mathbb{E}\left[\sum_{t=0}^{T-1}\gamma_t\Big[f_0^u(x_{t+1}) - f_0^u(x^{*,u}) + \sum_{i=1}^{n}\langle g_i^u(x_{t+1}^i), \hat{y}\rangle - \sum_{i=1}^{n}\langle g_i^u((x^i)^{*,u}), y_{t+1}^i\rangle\Big]\right].
\end{aligned}
\tag{54}
$$

Additionally, since

$$
\left\|\left[\sum_{i=1}^{n} g_i(\bar{x}_T^i)\right]_+\right\| \leq \left\|\left[\sum_{i=1}^{n} g_i^u(\bar{x}_T^i)\right]_+\right\|, \tag{55}
$$

by (54), (55) and (49), and setting $x = x^{*,u}$ and $y = \hat{y}$, we complete the proof. $\square$

7.4 Bounds for Certain Quantities in Convergence Rate and Constraint Violation Analysis

Lemma 7 ([28]). *For any $t \geq 0$, we have*

$$
\mathbb{E}[\|D_j(t)z_t^i\|] \leq 12M_0^2 d_i, \quad \mathbb{E}[\|G_0^i(t)\|] \leq 12M_0^2 d_i. \tag{56}
$$

Lemma 8. *Suppose $u \leq \frac{M_g}{(d+6)L_g}$. For any $t \geq 0$, we have*

$$
\mathbb{E}[\|V_t^i\|^2] \leq 24M_0^2 d_i + 27M_g^2 d_i C^2. \tag{57}
$$

Proof. Given that $V_t^i = G_0^i(t) + \sum_{j=1}^{m} H_{ij}[y_{t+1}^i]_j$, according to [19], we have

$$
\begin{aligned}
\mathbb{E}[\|V^i(t)\|^2] &\leq 2\mathbb{E}[\|G^i(t)\|^2] + 2\sum_{j=1}^{m}\mathbb{E}[\|H_{ij}(t)[y_{t+1}^i]_j\|^2] \\
&\leq 24M_0^2 d_i + \sum_{j=1}^{m} C^2[u^2 L_{ij}^2(d_i+6)^3 + 4(d_i+4)M_{ij}^2] \\
&= 24M_0^2 d_i + C^2[u^2 L_i^2(d_i+6)^3 + 4(d_i+4)M_i^2].
\end{aligned}
$$

Then, since $u \leq \frac{M_g}{(d+6)L_g}$, we conclude that

$$
\mathbb{E}[\|V_t^i\|^2] \leq 24M_0^2 d_i + C^2[(d_i+6)M_g^2 + 4(d_i+4)M_i^2] \leq 24M_0^2 d_i + 27M_g^2 d_i C^2.
$$

$\square$

Lemma 9. *Suppose $u \leq \frac{M_g}{(d+6)L_g}$ and let $\frac{1}{\eta_t} = L_0 + L_{\max} + \frac{1}{\eta}$. For any $t \geq 0$, we have*

$$
\mathbb{E}[\|x_t^i - x_{\tau_j^i(t)}^i\|^2] \leq \eta^2 b_{ij}^2\Big[24M_0^2 d_i + 27M_g^2 d_i C^2\Big] \tag{58}
$$

$$
\mathbb{E}[\|x_t - x_{\tau_j^i(t)}\|^2] \leq \eta^2 b_{ij}^2\Big[24M_0^2 d + 27M_g^2 d C^2\Big]. \tag{59}
$$

Proof. First of all, since $\frac{1}{\eta_t} = L_0 + L_{\max} + \frac{1}{\eta}$, we see that $\eta_t = \frac{1}{L_0 + L_{\max} + \frac{1}{\eta}} = \frac{\eta}{L_0\eta + L_{\max}\eta + 1} \leq \eta$.

The first-order optimality condition for (13) is

$$\left\langle V_t^i + \frac{1}{\eta_t}(x_{t+1}^i - x_t^i), x^i - x_{t+1}^i \right\rangle \geq 0$$

for any $x^i \in \mathcal{X}_i$. Setting $x^i = x_t^i$, we obtain

$$\langle \eta_t V_t^i, x_t^i - x_{t+1}^i \rangle \geq \|x_{t+1}^i - x_t^i\|^2,$$

which implies $\|x_{t+1}^i - x_t^i\| \leq \eta_t \|V_t^i\|$.

Then we have

$$\begin{aligned}
\mathbb{E}[\|x_t^i - x_{\tau_j^i(t)}^i\|^2] &\leq \mathbb{E}\left[\Big(\sum_{\tau=-b_{ij}}^{-1} \|\eta_{\tau+t} V^i(\tau+t)\|\Big)^2\right] \\
&\leq \sum_{\tau=-b_{ij}}^{-1} \eta_{\tau+t}^2 \mathbb{E}[\|V^i(\tau+t)\|^2] \\
&\leq b_{ij}\Big[24M_0^2 d_i + 27M_g^2 d_i C^2\Big] \sum_{\tau=-b_{ij}}^{-1} \eta_{\tau+t}^2 \\
&\leq \eta^2 b_{ij}^2 \Big[24M_0^2 d_i + 27M_g^2 d_i C^2\Big],
\end{aligned}$$

which further leads to

$$\mathbb{E}[\|x_t - x_{\tau_j^i(t)}\|^2] \leq \eta^2 b_{ij}^2 \Big[24M_0^2 d + 27M_g^2 dC^2\Big].$$

This completes the proof. □

Lemma 10. *Suppose $u \leq \frac{M_g}{(d+6)L_g}$ and let $\frac{1}{\eta_t} = L_0 + L_{\max} + \frac{1}{\eta}$. For any $t \geq 0$, we have*

$$\mathbb{E}\left[\sum_{i=1}^n \langle \delta_t^{G_i}, x^i - x_t^i \rangle\right] \leq \eta(M_0 \bar{\mathfrak{b}}\sqrt{d} + L_0 \bar{b} d\bar{R} + 2\sqrt{3}\bar{\mathfrak{b}} dM_0)(24M_0^2 + 27M_g^2 C^2)^{\frac{1}{2}}. \tag{60}$$

and

$$\sum_{i=1}^n \mathbb{E}[\|\delta_t^{G_i}\|^2] \leq 48M_0^2 d + 4M_0^2 n + 62M_g^2 dC^2 \tag{61}$$

Proof. Recalling the definition of $\delta_t^{G_i}$, we have

$$\begin{aligned}
\sum_{i=1}^n \langle \delta_t^{G_i}, x^i - x_t^i \rangle &= \sum_{i=1}^n \left\langle (G_0^i(t) - \frac{\partial f_0^u}{\partial x^i}(x_t)) + (H_i(t) - \nabla g_i^u(x_t^i))^T y_{t+1}^i, x^i - x_t^i \right\rangle \\
&= \sum_{i=1}^n \left\langle (G_0^i(t) - \frac{\partial f_0^u}{\partial x^i}(x_t)), x^i - x_t^i \right\rangle + \sum_{i=1}^n \langle (H_i(t) - \nabla g_i^u(x_t^i))^T y_{t+1}^i, x^i - x_t^i \rangle.
\end{aligned} \tag{62}$$

Let's consider these two terms separately. For the first term, we have

$$\mathbb{E}\left[\sum_{i=1}^{n}\left\langle G_0^i(t)-\frac{\partial f_0^u}{\partial x^i}(x_t),x^i-x_t^i\right\rangle\right]=\mathbb{E}\left[\sum_{i=1}^{n}\langle G_0^i(t),x^i-x_t^i\rangle\right]-\mathbb{E}\left[\sum_{i=1}^{n}\left\langle\frac{\partial f_0^u}{\partial x^i}(x_t),x^i-x_t^i\right\rangle\right]. \tag{63}$$

Recall that $G_0^i(t)=\frac{1}{n}\sum_{j=1}^{n}D_j^i(t)z^i_{\tau_j^i(t)}$, and we get

$$\begin{aligned}\mathbb{E}\left[\sum_{i=1}^{n}\langle G_0^i(t),x^i-x_t^i\rangle\right]&=\mathbb{E}\left[\frac{1}{n}\sum_{i,j=1}^{n}\langle D_j(\tau_j^i(t))z^i_{\tau_j^i(t)},x^i-x^i_{\tau_j^i(t)}\rangle\right]\\&\quad+\mathbb{E}\left[\frac{1}{n}\sum_{i,j=1}^{n}\langle D_j(\tau_j^i(t))z^i_{\tau_j^i(t)},x^i_{\tau_j^i(t)}-x_t^i\rangle\right].\end{aligned} \tag{64}$$

Then we note that

$$\begin{aligned}&\mathbb{E}\left[\frac{1}{n}\sum_{i,j=1}^{n}\langle D_j(\tau_j^i(t))z^i_{\tau_j^i(t)},x^i-x^i_{\tau_j^i(t)}\rangle\right]\\&=\frac{1}{n}\sum_{i,j=1}^{n}\mathbb{E}[\langle\nabla^i f_j^u(x_{\tau_j^i(t)}),x^i-x^i_{\tau_j^i(t)}\rangle]\\&=\mathbb{E}[\langle\nabla f^u(x_t),x-x_t\rangle]+\frac{1}{n}\sum_{i,j=1}^{n}\mathbb{E}[\langle\nabla^i f_j^u(x_t),x_t^i-x^i_{\tau_j^i(t)}\rangle]\\&\quad+\frac{1}{n}\sum_{i,j=1}^{n}\mathbb{E}[\langle\nabla^i f_j^u(x_{\tau_j^i(t)})-\nabla^i f_j^u(x_t),x^i-x^i_{\tau_j^i(t)}\rangle].\end{aligned} \tag{65}$$

By the Peter–Paul inequality and Lemma 9, we see that

$$\begin{aligned}&\frac{1}{n}\sum_{i,j=1}^{n}\mathbb{E}[\langle\nabla^i f_j^u(x_t),x_t^i-x^i(\tau_j^i(t))\rangle]\\&\le\frac{1}{n}\sum_{i,j=1}^{n}\left[\frac{\varepsilon}{2}\mathbb{E}[\|\nabla^i f_j^u(x_t)\|^2]+\frac{1}{2\varepsilon}\mathbb{E}[\|x^i(t)-x^i_{\tau_j^i(t)}\|^2]\right]\\&\le\frac{1}{n}\left[\frac{\varepsilon}{2}nM_0^2+\frac{1}{2\varepsilon}\sum_{i,j=1}^{n}\eta^2 b_{ij}^2 d_i(24M_0^2+27M_g^2C^2)\right]\\&=\frac{\varepsilon}{2}M_0^2+\frac{1}{2\varepsilon}\eta^2\bar{\mathfrak{b}}^2 d(24M_0^2+27M_g^2C^2).\end{aligned}$$

Taking $\varepsilon=\frac{\eta\bar{\mathfrak{b}}\sqrt{d}(24M_0^2+27M_g^2C^2)^{\frac{1}{2}}}{M_0}$, we get

$$\frac{1}{n}\sum_{i,j=1}^{n}\mathbb{E}[\langle\nabla^i f_j^u(x_t),x_t^i-x^i_{\tau_j^i(t)}\rangle]\le\eta M_0\bar{\mathfrak{b}}\sqrt{d}(24M_0^2+27M_g^2C^2)^{\frac{1}{2}}. \tag{66}$$

Next, we note that

$$\begin{aligned}
&\frac{1}{n}\sum_{i,j=1}^{n}\mathbb{E}[\langle\nabla^i f_j^u(x_{\tau_j^i(t)})-\nabla^i f_j^u(x_t),x^i-x^i_{\tau_j^i(t)}\rangle]\\
\le&\frac{1}{n}\sum_{i,j=1}^{n}\mathbb{E}[\|\nabla^i f_j^u(x_{\tau_j^i(t)})-\nabla^i f_j^u(x_t)\|\bar{R}_i]\\
\le&\frac{L_0}{n}\sum_{i,j=1}^{n}\sqrt{\mathbb{E}[\|x_{\tau_j^i(t)}-x_t\|^2]}\bar{R}_i\\
\le&\frac{L_0}{n}\sum_{i,j=1}^{n}\sqrt{\eta^2 b_{ij}^2 d\Big[24M_0^2+27M_g^2C^2\Big]}\bar{R}_i\\
=&\eta L_0\bar{b}\sqrt{nd}(24M_0^2+27M_g^2C^2)\bar{R}.
\end{aligned}\tag{67}$$

Furthermore, by Peter–Paul inequality, we have

$$\begin{aligned}
&\mathbb{E}\left[\frac{1}{n}\sum_{i,j=1}^{n}\langle D_j(\tau_j^i(t))z^i_{\tau_j^i(t)},x^i_{\tau_j^i(t)}-x^i_t\rangle\right]\\
\le&\frac{1}{n}\sum_{i,j=1}^{n}\mathbb{E}\Big[\frac{\varepsilon}{2}\|D_j(\tau_j^i(t))z^i_{\tau_j^i(t)}\|^2+\frac{1}{2\varepsilon}\|x^i_{\tau_j^i(t)}-x^i_t\|^2\Big]\\
\le&\frac{1}{n}\sum_{i,j=1}^{n}\Big[\frac{\varepsilon}{2}12M_0^2d_i+\frac{1}{2\varepsilon}\eta^2b_{ij}^2d_i\big[24M_0^2+27M_g^2C^2\big]\Big]\\
=&\frac{\varepsilon}{2}12M_0^2d+\frac{1}{2\varepsilon}\eta^2\bar{\mathfrak{b}}^2d(24M_0^2+27M_g^2C^2).
\end{aligned}$$

Taking $\varepsilon=\frac{\eta\bar{\mathfrak{b}}\sqrt{d}(24M_0^2+27M_g^2C^2)^{\frac{1}{2}}}{2\sqrt{3}M_0\sqrt{d}}$, we have

$$\mathbb{E}[\frac{1}{n}\sum_{i,j=1}^{n}\langle D_j(\tau_j^i(t))z^i_{\tau_j^i(t)},x^i_{\tau_j^i(t)}-x^i_t\rangle]\le 2\sqrt{3d}\eta\bar{\mathfrak{b}}dM_0\big[24M_0^2+27M_g^2C^2\big]^{\frac{1}{2}}.\tag{68}$$

Now, since H_i is an unbiased estimator of ∇g_i^u, we have

$$\mathbb{E}[\langle(H_i(t)-\nabla g_i^u(x_t^i))^Ty^i_{t+1},x^i-x^i_t\rangle]=0.$$

By summarizing the previous results, we obtain (60).

For $\|\delta_t^{G_i}\|^2$, we have

$$\begin{aligned}
\mathbb{E}[\|\delta_t^{G_i}\|^2]&=\mathbb{E}[\|(G_0^i(t)-\frac{\partial f_0^u}{\partial x^i}(x_t))+(H_i(t)-\nabla g_i^u(x_t^i))^Ty^i_{t+1}\|^2]\\
&\le 2\mathbb{E}[\|G_0^i(t)-\frac{\partial f_0^u}{\partial x^i}(x_t)\|^2]+2\mathbb{E}[\|(H_i(t)-\nabla g_i^u(x_t^i))^Ty^i_{t+1}\|^2]\\
&\le 4\mathbb{E}\Big[\|G_0^i(t)\|^2+\|\frac{\partial f_0^u}{\partial x^i}(x_t)\|^2\Big]+\sum_{j=1}^{m}2C^2[u^2L_{ij}(d_i+6)^3+4(d_i+5)M_{ij}^2]\\
&\le 48M_0^2d_i+4M_0^2+2C^2[u^2L_i^2(d_i+6)^3+4(d_i+5)M_i^2].
\end{aligned}$$

We complete our proof by summing over i. $\square$

Lemma 11. *Let $u \le \frac{M_g}{(d+6)L_g}$ and set $\theta_t = 1, \mu_t = \mu$, for all $t \ge 0$, we have*

$$\mathbb{E}\Big[\sum_{i=1}^{n}\langle q_t^i, p_t^i - y_t^i\rangle\Big] \le \frac{\mu}{1-\rho}(6nZ^2 + 3M_g^2\bar{R} + 243\bar{R}dM_g^2) \tag{69}$$

Proof. From the first-order optimality condition of (11), we have

$$\left\langle -s_t^i + \frac{1}{\mu_t}(y_{t+1}^i - p_t^i), y - y_{t+1}^i \right\rangle \ge 0, \qquad \forall y : y \ge 0, \|y\| \le C.$$

By plugging in $y = p_t^i$, we have

$$\langle -s_t^i, p_t^i - y_{t+1}^i\rangle \ge \frac{1}{\mu_t}\|p_t^i - y_{t+1}^i\|^2, \tag{70}$$

and applying the Cauchy-Schwarz inequality leads to

$$\|p_t^i - y_{t+1}^i\| \le \mu_t\|s_t^i\|. \tag{71}$$

Next, by the definition of s_i^t, we have

$$\begin{aligned}
\mathbb{E}[\|s_t^i\|^2] \le\ & \mathbb{E}[\|g_i(x_{t-1}^i) + G_i(t-1)(x_t^i - x_{t-1}^i) + \theta_t[g_i(x_{t-1}^i) - g_i(x_{t-2}^i)] \\
& + \theta_t[G_i(t-1)(x_t^i - x_{t-1}^i) - G_i(t-2)(x_{t-1}^i - x_{t-2}^i)]\|^2] \\
\le\ & 3\mathbb{E}[\|g_i(x_{t-1}^i) + G_i(t-1)(x_t^i - x_{t-1}^i)\|^2] \\
& + 3\theta_t^2\mathbb{E}[\|g_i(x_{t-1}^i) - g_i(x_{t-2}^i)\|^2] \\
& + 3\theta_t^2\mathbb{E}[\|G_i(t-1)(x_t^i - x_{t-1}^i) - G_i(t-2)(x_{t-1}^i - x_{t-2}^2)\|^2].
\end{aligned}$$

Recall that $\|g_i(x^i)\| \le Z$, and it follows that

$$\begin{aligned}
3\mathbb{E}[\|g_i(x_{t-1}^i) + G_i(t-1)(x_t^i - x_{t-1}^i)\|^2] &\le 6Z^2 + 6\bar{R}_i^2(\frac{u^2}{2}L_i^2(d_i+6)^3 + 2(d_i+4)M_i^2) \\
3\theta_t^2\mathbb{E}[\|g_i(x_{t-1}^i) - g_i(x_{t-2}^i)\|^2] &\le 3\theta_t^2 M_i^2\bar{R}_i^2 \\
3\theta_t^2\mathbb{E}[\|G_i(t-1)(x_t^i - x_{t-1}^i) - G_i(t-2)(x_{t-1}^i - x_{t-2}^2)\|^2] &\le 12\theta_t^2\bar{R}_i^2[\frac{u^2}{2}L_i^2(d_i+6)^3 + 2(d_i+4)M_i^2].
\end{aligned}$$

Define $e_t^i = y_{t+1}^i - p_t^i$. Since $\theta_t = 1$ and $u \le \frac{M_g}{(d+6)L_g}$, for all $t \ge 0$ we have

$$\mathbb{E}[\|e_t^i\|^2] \le \mu^2(6Z^2 + 3M_i^2\bar{R}_i^2 + 243\bar{R}_i^2 d_i M_g^2) = (\alpha^i)^2.$$

Summing over i, we obtain

$$\sum_{i=1}^{n}\mathbb{E}[\|e_t^i\|^2] \le \mu^2(6nZ^2 + 3M_g^2\bar{R} + 243\bar{R}dM_g^2) = \alpha^2.$$

Define the matrices

$$Y_t = \begin{pmatrix} -{y_t^1}^T- \\ \vdots \\ -{y_t^n}^T- \end{pmatrix}, P_t = \begin{pmatrix} -{p_t^1}^T- \\ \vdots \\ -{p_t^n}^T- \end{pmatrix}.$$

It is easy to see that

$$\begin{aligned} Y_{t+1} &= WY_t + (Y_{t+1} - WY_t) \\ \bar{Y}_{t+1} &= \frac{1}{n}\mathbf{1}\mathbf{1}^T Y_{t+1} \\ W(W-I) &= (W - \frac{1}{n}\mathbf{1}\mathbf{1}^T)(W-I). \end{aligned}$$

Thus

$$\begin{aligned} P_{t+1} - Y_{t+1} &= (W-I)Y_{t+1} \\ &= W(W-I)Y_t + (W-I)(Y_{t+1} - WY_t) \\ &= (W - \frac{1}{n}\mathbf{1}\mathbf{1}^T)(W-I)Y_t + (W-I)(Y_{t+1} - WY_t). \end{aligned}$$

Consequently,

$$\|P_{t+1} - Y_{t+1}\|_F \le \left\| W - \frac{1}{n}\mathbf{1}\mathbf{1}^T \right\|_2 \|(W-I)Y_t\|_F + \|W-I\|_2 \|Y_{t+1} - WY_t\|_F.$$

Since W is also a positive semi-definite matrix, we can check that $\|W-I\|_2 \le 1$. Define

$$\begin{aligned} \rho &= \left\| W - \frac{1}{n}\mathbf{1}\mathbf{1}^T \right\|_2 < 1 \\ \kappa_t &= \|P_t - Y_t\|_F. \end{aligned}$$

Then, we can derive that

$$\begin{aligned} \kappa_{t+1} &\le \rho\kappa_t + \varphi\alpha \\ &\le \rho[\rho\kappa_{t-1} + \varphi\alpha] + \varphi\alpha \\ &\le \cdots \\ &\le \rho^{t+1}\kappa_0 + \varphi\sum_{k=0}^{t} \rho^k \alpha. \end{aligned}$$

Since $y_0^i = 0$ for all $i = 1, ..., n$, we have

$$\begin{aligned} \kappa_{t+1} &\le \frac{\varphi\alpha}{1-\rho} \le \frac{\alpha}{1-\rho} \\ \kappa_{t+1}^2 &\le \left(\frac{\alpha}{1-\rho}\right)^2. \end{aligned}$$

Thus,

$$\|P_{t+1} - Y_{t+1}\|_F^2 = \sum_{i=1}^{n} \|p_{t+1}^i - y_{t+1}^i\|^2 \le \left(\frac{\alpha}{1-\rho}\right)^2.$$

Then,

$$\begin{aligned}
\mathbb{E}[\|q_t^i\|^2] &= \mathbb{E}[\|\ell_G^i(t) - \ell_G^i(t-1)\|^2] \\
&= \mathbb{E}[\|g_i(x_{t-1}^i) - g_i(x_{t-2}^i) + G_i(t-1)(x_t^i - x_{t-1}^i) - G_i(t-2)(x_{t-1}^i - x_{t-2}^i)\|^2] \\
&\leq 3\mathbb{E}[\|g_i(x_{t-1}^i) - g_i(x_{t-2}^i))\|^2] + 3\bar{R}_i^2\mathbb{E}[\|G_i(t-1)\|^2] + 3\bar{R}_i^2\mathbb{E}[\|G_i(t-2)\|^2] \\
&\leq 3\bar{R}_i^2 M_i^2 + 81\bar{R}_i^2 d_i M_g^2.
\end{aligned}$$

Using the Peter–Paul inequality, we get

$$\begin{aligned}
\mathbb{E}\Big[\sum_{i=1}^n \langle q_t^i, p_t^i - y_t^i\rangle\Big] &\leq \mathbb{E}\Big[\sum_{i=1}^n \Big(\frac{\|q_t^i\|^2}{2\varepsilon} + \frac{\varepsilon}{2}\|p_t^i - y_t^i\|^2\Big)\Big] \\
&= \frac{1}{2\varepsilon}\sum_{i=1}^n \Big(3\bar{R}_i^2 M_i^2 + 81\bar{R}_i^2 d_i M_g^2\Big) + \frac{\varepsilon}{2}(\frac{\alpha}{1-\rho})^2 \\
&= \frac{1}{2\varepsilon}\Big(3\bar{R}^2 M_g^2 + 81\bar{R}^2 d M_g^2\Big) \\
&\quad + \frac{\varepsilon}{2}(\frac{1}{1-\rho})^2[\mu^2(6nZ^2 + 3M_g^2\bar{R} + 243\bar{R}dM_g^2)].
\end{aligned} \tag{72}$$

Setting $\varepsilon = \frac{(1-\rho)(3\bar{R}^2M_g^2 + 81\bar{R}dM_g^2)^{\frac{1}{2}}}{\mu(6nZ^2 + 3M_g^2\bar{R} + 243\bar{R}dM_g^2)^{\frac{1}{2}}}$, we now conclude that

$$\begin{aligned}
\mathbb{E}\Big[\sum_{i=1}^n \langle q_t^i, p_t^i - y_t^i\rangle\Big] &\leq \frac{\mu}{1-\rho}(3\bar{R}^2M_g^2 + 81\bar{R}dM_g^2)^{\frac{1}{2}}(6nZ^2 + 3M_g^2\bar{R} + 243\bar{R}dM_g^2)^{\frac{1}{2}} \\
&\leq \frac{\mu}{1-\rho}(6nZ^2 + 3M_g^2\bar{R} + 243\bar{R}dM_g^2).
\end{aligned} \tag{73}$$

This completes the proof. □

With all the necessary preparations in place, we now proceed to prove our results.

Proof of Theorem 1. Given that $G_{ij}(t) = \frac{g_{ij}(x_t^i + u\hat{z}_t^i) - g_{ij}(x_t^i - u\hat{z})_t^i}{2u}\hat{z}_t^i$, based on [19], we have

$$\begin{aligned}
\mathbb{E}[\|G_{ij}(t-1) - \nabla g_{ij}^u(x_{t-1}^i)\|^2] &\leq 2\mathbb{E}[\|G_{ij}(t)\|^2] + 2\|\nabla g_{ij}^u(x_{t-1}^i)\|^2 \\
&\leq 2\Big(\frac{u^2}{2}L_{ij}^2(d_i+6)^3 + 2(d_i+4)\|\nabla g_{ij}(x)\|^2\Big) + 2M_{ij}^2 \\
&\leq u^2L_{ij}^2(d_i+6)^3 + 4(d_i+4)M_{ij}^2 + 2M_{ij}^2 \\
&\leq u^2L_{ij}^2(d_i+6)^3 + 4(d_i+5)M_{ij}^2.
\end{aligned}$$

This leads to the following inequality

$$\begin{aligned}
\mathbb{E}[\|G_i(t-1) - \nabla g_i^u(x_{t-1}^i)\|_F^2] &\leq \sum_{j=1}^m \Big[u^2L_{ij}^2(d_i+6)^3 + 4(d_i+5)M_{ij}^2\Big] \\
&= u^2L_i^2(d_i+6)^3 + 4(d_i+5)M_i^2 \leq 31d_iM_g^2.
\end{aligned} \tag{74}$$

According to Lemma 1,we obtain

$$\begin{aligned}\mathbb{E}[\|g_i(x_{t-1}^i) - g_i^u(x_{t-1}^i)\|^2] &= \mathbb{E}[\sum_{j=1}^{m}(g_{ij}(x_{t-1}^i) - g_{ij}^u(x_{t-1}^i))^2] \\ &\le \frac{1}{4}\sum_{j=1}^{m} u^4 L_{ij}^2 d_i^2 = \frac{1}{4}u^4 L_i^2 d_i^2\end{aligned}$$

where the last equality holds because $\sum_{j=1}^{m} L_{ij}^2 = L_i^2$. Then, we have

$$\begin{aligned}\mathbb{E}[\|\delta_t^{F_i}\|_2^2] &= \mathbb{E}[\|\ell_G^i(t) - \ell_{g^u}^i(t)\|_2^2] \\ &= \mathbb{E}\Big[\|(g_i(x_{t-1}^i) + G_i(t-1)(x_t^i - x_{t-1}^i)) - (g_i^u(x_{t-1}^i) + \nabla g_i^u(x_{t-1}^i)(x_t^i - x_{t-1}^i))\|_2^2\Big] \\ &= \mathbb{E}\Big[\|(g_i(x_{t-1}^i) - g_i^u(x_{t-1}^i)) + (G_i(t-1) - \nabla g_i^u(x_{t-1}^i)(x_t^i - x_{t-1}^i))\|_2^2\Big] \\ &\le 2\mathbb{E}[\|g_i(x_{t-1}^i) - g_i^u(x_{t-1}^i)\|^2] + 2\mathbb{E}[\|G_i(t-1) - \nabla g_i^u(x_{t-1}^i)(x_t^i - x_{t-1}^i)\|^2] \\ &\le u^4 L_i^2 d_i^2 + 62 d_i M_g^2 \bar{R}_i^2.\end{aligned} \tag{75}$$

Summing over i, we have

$$\sum_{i=1}^{n}\mathbb{E}[\|\delta_t^{F_i}\|_2^2] \le u^4 L_g^2 d^2 + 62 d M_g^2 \bar{R}^2. \tag{76}$$

Then, we note that

$$\mathbb{E}[\langle\delta_{t+1}^{F_i}, y_{t+1}^i - y\rangle] = \mathbb{E}\big[\langle\mathbb{E}[\delta_{t+1}^{F}|\mathcal{F}_t], y_{t+1}^i - y\rangle\big],$$

where

$$\mathbb{E}[\delta_{t+1}^{F_i}|\mathcal{F}_t] = \mathbb{E}[g_i(x_t^i) - g_i^u(x_t^i)|\mathcal{F}_t] \le \frac{u^2 d_i}{2} L_i.$$

Thus

$$\mathbb{E}[\langle\delta_{t+1}^{F_i}, y_{t+1}^i - y\rangle] = \mathbb{E}\big[\langle\mathbb{E}[\delta_{t+1}^{F}|\mathcal{F}_t], y_{t+1}^i - y\rangle\big] \le u^2 d_i L_i C.$$

Summing over i leads to

$$\sum_{i=1}^{n}\mathbb{E}[\langle\delta_{t+1}^{F_i}, y_{t+1}^i - y\rangle] \le u^2 d L_{\max} C, \tag{77}$$

where $L_{\max} = \max\{L_1, L_2, ..., L_n\}$.

Next, we note that

$$\begin{aligned}\sum_{i=1}^{n}\mathbb{E}[\|q_t^i - \bar{q}_t^i\|_2^2] &= \sum_{i=1}^{n}\mathbb{E}[\|\ell_G^i(t) - \ell_G^i(t-1) - \ell_{g^u}^i(t) + \ell_{g^u}^i(t-1)\|_2^2] \\ &\le \sum_{i=1}^{n}(2\mathbb{E}[\|\delta_t^{F_i}\|_2^2] + 2\mathbb{E}[\|\delta_{t-1}^{F_i}\|_2^2]) \\ &\le (\sum_{i=1}^{n} 4u^4 L_i^2 d_i + 248 d_i M_g^2 \bar{R}_i^2) \\ &\le 4u^4 L_g^2 d^2 + 248 d M_g^2 \bar{R}^2.\end{aligned} \tag{78}$$

Setting $\theta_t = 1$, $\mu_t = \mu$, $\gamma_t = 1$ and $\frac{1}{\eta_t} = L_0 + L_{\max} + \frac{1}{\eta}$, and combining the results from (42), (60), (61), (77), (78) and (69), we arrive at the following inequality

$$\begin{aligned}
&\mathbb{E}[f_0(\bar{x}_T) - f_0(x^*)] \\
\leq &\frac{1}{T}(L_0 + L_{\max})\bar{R}^2 + \frac{1}{T\eta}\bar{R}^2 + \eta(104M_0^2 d + 124M_g^2 dC^2) + \mu(6u^4 L_g^2 d^2 + 372dM_g^2\bar{R}^2) \\
&+ \eta(M_0\bar{\mathfrak{b}}\sqrt{d} + 2\sqrt{3}\bar{\mathfrak{b}}dM_0)(24M_0^2 + 27M_g^2C^2)^{\frac{1}{2}} + u^2 dL_{\max}C \\
&+ \frac{1}{2}u^2 L_0 d + \frac{\mu}{1-\rho}(6dZ^2 + 3M_g^2\bar{R} + 243\bar{R}dM_g^2).
\end{aligned}$$

Similarly, by combining (47), (60), (61), (69), (76), (77), (78), we derive the following result

$$\begin{aligned}
&\mathbb{E}\left[\left\|\left[\sum_{i=1}^{n} g_i(\bar{x}_T^i)\right]_+\right\|\right] \\
\leq &\frac{1}{T}(L_0 + L_{\max})\bar{R}^2 + \frac{1}{T\eta}\bar{R}^2 + \frac{1}{\mu}nC^2 + \eta(104M_0^2 d + 124M_g^2 dC^2) \\
&+ \mu(7u^4 L_g^2 d^2 + 403dM_g^2\bar{R}^2) + \eta(M_0\bar{\mathfrak{b}}\sqrt{d} + 2\sqrt{3}\bar{\mathfrak{b}}dM_0)(24M_0^2 + 27M_g^2C^2)^{\frac{1}{2}} \\
&+ u^2 dL_{\max}C + \frac{\mu}{1-\rho}(6dZ^2 + 3M_g^2\bar{R} + 243\bar{R}dM_g^2) + \eta L_g\bar{R}C.
\end{aligned}$$

To refine these results further, we set the parameters as $\eta = \frac{\bar{R}}{\sqrt{T\xi}}$ and $\mu = \frac{C\sqrt{2n}}{\sqrt{T\zeta}}$, where the constants ξ and ζ are given by

$$\begin{aligned}
\xi &= (M_0\bar{\mathfrak{b}}\sqrt{d} + L_0\bar{b}d\bar{R} + 2\sqrt{3}\bar{\mathfrak{b}}dM_0)(24M_0^2 + 27M_g^2C^2)^{\frac{1}{2}} + 104M_0^2 d + 124M_g^2 dC^2, \\
\zeta &= 403dM_g^2\bar{R} + \frac{1}{1-\rho}(6dZ^2 + 3M_g^2\bar{R} + 243\bar{R}dM_g^2).
\end{aligned}$$

With these parameter choices, we subsequently derive the convergence rate and constraint violation bounds as specified in (14) and (15). □