Journal of Machine Learning Research 22 (2021) 1-42 Submitted 6/18; Revised 9/20; Published 2/21

Projection-free Decentralized Online Learning for Submodular Maximization over Time-Varying Networks

Junlong Zhu JLZHU@HAUST.EDU.CN
School of Information Engineering
Henan University of Science and Technology
Luoyang, 471023, China

Qingtao Wu WQT8921@HAUST.EDU.CN
School of Information Engineering
Henan University of Science and Technology
Luoyang, 471023, China

Mingchuan Zhang (Corresponding author) ZHANG_MCH@HAUST.EDU.CN
School of Information Engineering
Henan University of Science and Technology
Luoyang, 471023, China

Ruijuan Zheng ZHENGRUIJUAN@HAUST.EDU.CN
School of Information Engineering
Henan University of Science and Technology
Luoyang, 471023, China

Keqin Li LIK@NEWPALTZ.EDU
Department of Computer Science
State University of New York
New Paltz, NY 12561, USA

Editor: Vahab Mirrokni

Abstract

This paper considers a decentralized online submodular maximization problem over time-varying networks, where each agent only utilizes its own information and the received information from its neighbors. To address the problem, we propose a decentralized Meta-Frank-Wolfe online learning method in the adversarial online setting by using local communication and local computation. Moreover, we show that an expected regret bound of $\mathcal{O}(\sqrt{T})$ is achieved with $(1-1/e)$ approximation guarantee, where T is a time horizon. In addition, we also propose a decentralized one-shot Frank-Wolfe online learning method in the stochastic online setting. Furthermore, we also show that an expected regret bound $\mathcal{O}(T^{2/3})$ is obtained with $(1-1/e)$ approximation guarantee. Finally, we confirm the theoretical results via various experiments on different datasets.

Keywords: Expected Regret Bound, Frank-Wolfe Algorithm, Submodular Maximization.

1. Introduction

Submodular function optimization problems have received significant interest since they have found many applications in machine learning and related areas. For example, vari-

ation inference (Djolonga and Krause, 2014), diversity (Kulesza and Taskar, 2012), data summarization (Lin and Bilmes, 2011; Mirzasoleiman et al., 2013), influence maximization (Domingos and Richardson, 2001; Kempe et al., 2003), structured sparsity (Bach, 2010), dictionary learning (Krause and Cevher, 2010; Das and Kempe, 2011), and variable selection (Krause and Guestrin, 2005) etc. In order to solve such problems, we need to design effective optimization algorithms to find optimal solutions. However, the submodular optimization is a class of non-convex optimization problems. Moreover, finding a global optimum for non-convex optimization problems is NP-hard in general (Murty and Kabadi, 1987). Therefore, the design of optimization algorithms for submodular optimization is a challenging problem.

In submodular optimization, submodular functions can be minimized exactly (Iwata et al., 2001) and maximized approximately (Krause and Golovin, 2012; Nemhauser et al., 1978) in polynomial time. Classical results in submodular optimization have been mainly based on combinatorial techniques such as the greedy algorithms (Nemhauser et al., 1978; Nemhauser and Wolsey, 1978; Fisher et al., 1978). Recently, Bach (2015) demonstrated that the submodular set function can be extended to the continuous function in the context of minimization. Based on this method, various variants of submodular optimization algorithms were proposed in recent years (Hassani et al., 2017; Bian et al., 2017). Specially, Mokhtari et al. (2018a) proposed a stochastic continuous greedy algorithm with $(1 - 1/e)$ approximation guarantee for stochastic submodular maximization, which is introduced by Karimi et al. (2017). In addition, Chen et al. (2018a) proposed a Meta-Frank-Wolfe algorithm that achieves a square-root regret bound with $(1 - 1/e)$ approximation guarantee for online continuous submodular maximization, where the full gradient is available. Furthermore, the regret bound of $\mathcal{O}(\sqrt{T})$ with a weaker $1/2$ approximation is also achieved by an online stochastic gradient method, where T is a time horizon. Furthermore, Chen et al. (2018b) extended the Meta-Frank-Wolfe method that only uses the estimates of stochastic gradient, and showed that a $(1 - 1/e)$-regret bound of $\mathcal{O}(\sqrt{T})$ can be achieved under the adversarial online setting. Despite these progresses, however, the works cited above are implemented in a centralized manner.

For the "big-data" challenge, optimization algorithms have been sought for coping with high-dimensional optimization problems (Cevher et al., 2014). Moreover, since these massive data are dispersed over the nodes of networks, decentralized optimization algorithms are effective tools for tackling large-scale learning tasks, where the nodes can use the computation power in a cooperative manner (Sayed et al., 2013). Therefore, how to design efficient decentralized algorithms is desirable (Boyd et al., 2011). For these reasons, Mokhtari et al. (2018b) proposed decentralized continuous greedy optimization methods over networks for submodular maximization with $(1-1/e)$ approximation guarantee via local communication and local computation. In this work, the authors assumed that the objective functions are unchanged with time. However, the objective functions change with time in many real-world scenarios (Chen et al., 2018b). To the best of our knowledge, the decentralized online variants over time-varying networks for submoudlar maximization are barely investigated. For this reason, we focus on the design and analysis of decentralized online learning algorithms. Recently, Yan et al. (2013) proposed a distributed online projected subgradient descent algorithm for online convex optimization problems, and showed that the square-root regret and logarithmic regret are achieved for convex and strongly convex objective functions, respectively. Based on dual subgradient averaging, Hosseini et al. (2016) proposed

a distributed online algorithm over networks and established a regret bound of $\mathcal{O}(\sqrt{T})$. Shahrampour and Jadbabaie (2018) proposed a distributed mirror descent algorithm for online convex optimization in dynamic environments. Additionally, Zhang et al. (2017) proposed a projection-free distributed online learning algorithm and showed that the regret bound of $\mathcal{O}(T^{3/4})$ is achieved. Recently, Zhang et al. (2019a) proposed a distributed conditional gradient algorithm for online learning. Moreover, the regret bound of $\mathcal{O}(\sqrt{T})$ is obtained. The works cited above aim to solve online convex optimization problems, where the objective functions are convex. Moreover, these algorithms need to compute exact gradients of the objective function. For high-dimensional data, however, the computations of the exact gradients becomes expensive prohibitively. Furthermore, the close-form of the exact gradients may not exist in some cases. To avoid these issues, the decentralized online variants, which use the projection-free technique and stochastic gradient estimates, are desired. However, how to design and analyze these variants for online submodular maximization remains an open problem.

In this paper, we fill this gap and present some decentralized learning algorithms over time-varying networks for online submodular maximization. In these algorithms, we replace the exact gradients with stochastic estimates of the gradients. Moreover, each agent can exchange information with its neighbors. Furthermore, we also use the Frank-Wolfe technique to avoid projections, which are prohibitive when dealing with the high-dimensional data. The main contributions of this paper are as follows:

- We present a decentralized Meta-Frank-Wolfe online learning method over time-varying networks for submodular maximization in the adversarial online setting, where each agent only utilizes its own local information and the received information from its neighbors. Moreover, each agent has access only to stochastic gradient estimates at each iteration.

- We also show that the decentralized Meta-Frank-Wolfe online learning method can achieve $(1-1/e)$-regret with a bound $\mathcal{O}(\sqrt{T})$ via a careful estimate of gradients, where T denotes a time horizon.

- We propose a decentralized one-shot Frank-Wolfe online learning method over time-varying networks for submodular maximization in the stochastic online setting, where each agent uses local communication and local computation. Moreover, each agent has access only to a single stochastic estimate of gradient at each iteration.

- We also show that the decentralized one-shot Frank-Wolfe online learning method can achieve $(1-1/e)$-regret with a bound $\mathcal{O}(T^{2/3})$.

The remainder of this paper is organized as follows. The related works are reviewed in Section 2. In Section 3, we present some notations and mathematical background, which are used in the paper. We describe the decentralized online submodular maximization problem of our interest, design the online learning methods, and give some assumptions in Section 4. The main results of this paper are provided in Section 5. In Section 6, we analyze the performance of our proposed algorithms and provide the detailed proofs of the main results. In Section 7, we evaluate the performance of the proposed algorithms by numerical experiments on different datasets. Finally, we conclude this paper in Section 8.

2. Related Work

The framework of decentralized online convex optimization was introduced by Yan et al. (2013), in which the distributed online projected subgradient descent was proposed and showed logarithmic regret for strongly convex objective functions and square-root regret for convex objective functions. However, the projection operation can be prohibitive when dealing with high-dimensional data. For this reason, the distributed online conditional gradient algorithm was proposed by Zhang et al. (2017), in which the projection step was eschewed by exploiting the Frank-Wolfe technique. These works mainly focused on convex objective functions. To the best of our knowledge, the distributed online variant of conditional gradient over time-varying networks for non-convex objective functions is barely known.

The variance reduction method was introduced by Johnson and Zhang (2013) and independently proposed by Mahdavi et al. (2013) for accelerating stochastic gradient descent. Recently, stochastic variance reduction was used for non-convex optimization (Allen-Zhu and Hazan, 2016). Additionally, Reddi et al. (2016) also applied the stochastic variance reduction method to nonconvex optimization. Hazan and Luo (2016) proposed a projection-free stochastic optimization algorithm for convex optimization problems by using the variance reduction method. Mokhtari et al. (2018a) proposed a stochastic conditional gradient algorithm with $(1-1/e)$ approximation guarantee for stochastic submodular maximization by exploiting a different variance reduction method. Mokhtari et al. (2018b) also proposed a decentralized projection-free algorithm with $(1-1/e)$ approximation guarantee for distributed submodular maximization via this variance reduction method. However, the works cited above applied these variance reduction methods to convex optimization or submodular optimization, where the objective functions are unchanged with time.

Our paper is the first to provide the decentralized online algorithms for distributed online submodular maximization via the variance reduction methods, where the objective functions change with time. Indeed, Nemhauser et al. (1978) proposed a centralized greedy algorithm with tight approximation guarantee for maximizing submodular set functions. Moreover, its variants are studied (Feige et al., 2011; Mirzasoleiman et al., 2016; Feldman et al., 2017). However, these methods cannot scale to massive data sets since they are sequential in nature. For this reason, MapReduce style methods were proposed (Mirzasoleiman et al., 2013; Mirrokni and Zadimoghaddam, 2015). Recently, Bach (2015) extended discrete domain to continuous domain for submodular functions. Hassani et al. (2017) proposed projected gradient algorithms with 1/2 approximation guarantee for maximizing continuous submodular functions. Moreover, Bian et al. (2017) proposed a Frank-Wolfe variant for continuous submodular maximization with $(1-1/e)$ approximation guarantee. Recently, Mokhtari et al. (2018a) proposed a stochastic gradient method with $(1-1/e)$ approximation guarantee by using Frank-Wolfe technique. Decentralized conditional gradient algorithms were introduced by Mokhtari et al. (2018b). Besides, Zhang et al. (2019b) proposed quantized Frank-Wolfe algorithms to solve constrained optimization problems. Zhuo et al. (2019) presented an asynchronous stochastic Frank-Wolfe algorithm for solving an optimization problems with a nuclear norm constraint. For the online setting, Chen et al. (2018a) proposed an online variant of Frank-Wolfe for online submodular maximization and showed that a regret bound of $\mathcal{O}(\sqrt{T})$ is achieved with $(1-1/e)$ approximation guarantee. Furthermore, Chen

et al. (2018b) also proposed stochastic conditional gradient online optimization algorithms and showed that the regret bound $\mathcal{O}(\sqrt{T})$ is achieved with $(1-1/e)$ approximation guarantee. Zhang et al. (2019c) proposed three online submodular maximization algorithms, i.e., Mono-Frank-Wolfe, Bandit-Frank-Wolfe, and Responsive-Frank-Wolfe. Moreover, the $(1-1/e)$-regret bounds of $\mathcal{O}(T^{4/5})$ and $\mathcal{O}(T^{8/9})$ was also achieved, respectively. However, these methods mainly focus on centralized computational architectures for submodular maximization.

3. Preliminaries

In this section, we first provide some notations, which are used in this paper. Moreover, we also present mathematical background for submodular functions.

3.1 Notations

In this paper, all vectors are all column vectors. We use boldface to denote the vector with suitable dimension and use normal font to denote scalars. We use the notations $\mathbb{R}$ and $\mathbb{R}_+$ to denote the sets of real numbers and non-negative real numbers, respectively. Moreover, the notations $\mathbb{R}^d$ and $\mathbb{R}_+^d$ denote the real vector and non-negative real vector with dimension d, respectively. The notation $\mathbb{R}^{N\times N}$ denotes the real matrix of size $N\times N$. The notation $\|\mathbf{x}\|$ denotes the standard Euclidean norm of a vector $\mathbf{x}$. We use the notations $\mathbf{x}^\mathsf{T}$ and A^T to denote the transpose operation of a vector $\mathbf{x}$ and a matrix A, respectively. The notation $\langle \mathbf{x}, \mathbf{y}\rangle$ denotes the inner product of vectors $\mathbf{x}$ and $\mathbf{y}$. We use the notations I and $\mathbb{1}$ to denote the identity matrix and a vector that all entries are 1 with suitable size, respectively. Moreover, we use $\mathbb{E}[X]$ to denote the expectation of a random variable X. The notation $\otimes$ denotes the Kronecker product. In addition, the notations $\preceq$ and $\succeq$ denote coordinate-wise inequalities, respectively.

3.2 Mathematical Background

In this subsection, we provide some precise definitions for submodular functions. We first introduce the definition of submodular set functions. Given a ground set V, which consists of d elements. For all $A, B \subseteq V$, a set function $f: 2^V \to \mathbb{R}_+$ satisfies the following relation,

$$f(A) + f(B) \geq f(A \cap B) + f(A \cup B), \tag{1}$$

then the set function f is called submodular. Furthermore, the notion of submodularity can be extended to continuous domain. Given a subset $\mathcal{X}$ in $\mathbb{R}_+^d$, which is of the form $\mathcal{X} = \prod_{i=1}^d \mathcal{X}_i$. Moreover, each set $\mathcal{X}_i$ is a compact subset of $\mathbb{R}_+$ for $i = 1, \ldots, d$. A continuous function $F: \mathcal{X} \to \mathbb{R}_+$ is called submodular if for all $\mathbf{x}, \mathbf{y} \in \mathcal{X}$, we have

$$F(\mathbf{x}) + F(\mathbf{y}) \geq F(\mathbf{x} \vee \mathbf{y}) + F(\mathbf{x} \wedge \mathbf{y}), \tag{2}$$

where $\mathbf{x} \vee \mathbf{y} := \max\{\mathbf{x}, \mathbf{y}\}$ (coordinate-wise) and $\mathbf{x} \wedge \mathbf{y} := \min\{\mathbf{x}, \mathbf{y}\}$ (coordinate-wise). In this paper, we focus on the monotone and DR-submodular continuous function. Formally, a continuous submodular function F is called monotone on $\mathcal{X}$ if $\mathbf{x} \preceq \mathbf{y}$, we have $F(\mathbf{x}) \leq F(\mathbf{y})$ for all $\mathbf{x}, \mathbf{y} \in \mathcal{X}$. Moreover, a differentiable continuous submodular function F is DR-submodular if $\mathbf{x} \preceq \mathbf{y}$, we have $\nabla F(\mathbf{x}) \succeq \nabla F(\mathbf{y})$ for all $\mathbf{x}, \mathbf{y} \in \mathcal{X}$. Namely, $\nabla F(\cdot)$ is an

antitone mapping. Furthermore, *DR*-submodularity of function F implies that the function F is concave in positive directions, i.e., we have

$$F(\mathbf{y}) \leq F(\mathbf{x}) + \langle \nabla F(\mathbf{x}), \mathbf{y} - \mathbf{x} \rangle \tag{3}$$

for all $\mathbf{x}, \mathbf{y} \in \mathcal{X}$. In addition, when a continuous function F is twice differentiable, the function F is submodular if and only if all off-diagonal components of its Hessian matrix are non-positive. Formally, for all $\mathbf{x} \in \mathcal{X}$, we obtain

$$\forall\ i \neq j,\ \frac{\partial^2 F(\mathbf{x})}{\partial x_i \partial x_j} \leq 0. \tag{4}$$

Furthermore, if the function F is *DR*-submodular, then all elements of its Hessian matrix are non-positive. Formally, for all $\mathbf{x} \in \mathcal{X}$, we have

$$\frac{\partial^2 F(\mathbf{x})}{\partial x_i \partial x_j} \leq 0. \tag{5}$$

In addition, the twice differentiability of the function F implies that the submodular function F is smooth. Furthermore, a continuous submodular function F is L-smooth if for all $\mathbf{x}, \mathbf{y} \in \mathcal{X}$, we have

$$F(\mathbf{y}) \leq F(\mathbf{y}) + \langle \nabla F(\mathbf{x}), \mathbf{y} - \mathbf{x} \rangle + \frac{L}{2} \|\mathbf{y} - \mathbf{x}\|^2, \tag{6}$$

which implies that

$$\|\nabla F(\mathbf{x}) - \nabla F(\mathbf{y})\| \leq L \|\mathbf{x} - \mathbf{y}\|. \tag{7}$$

In this section, we present some basic notations and concepts. The formulation of our problem is described, and then we propose some efficient algorithms for this problem in the next section.

4. Problem Formulation, Algorithms Design, and Assumptions

In this section, we first formally introduce the problem of our interest, and then design decentralized online learning algorithms to solve the problem. Finally, in order to analyze the performance of the proposed algorithms, we also provide some standard assumptions.

4.1 Problem Formulation

In this paper, we consider a decentralized online optimization problem in time-varying networks, which is defined formally as follows: A graph $\mathcal{G}(t) = (\mathcal{V}, \mathcal{E}(t))$ is used to denote a time-varying network, where $\mathcal{V} = \{1, \ldots, N\}$ denotes the set of agents (nodes) and $\mathcal{E}(t) \subset \mathcal{V} \times \mathcal{V}$ is the set of edges at time t. Let $(i, j) \in \mathcal{E}(t)$ denote an edge from agent i to agent j at time t. We use notation $\mathcal{N}_i(t)$ to denote the set of neighbors of agent i at time t, where the agent i can directly communicate with the agent $j \in \mathcal{V}$. Formally, $\mathcal{N}_i(t) = \{j \in \mathcal{V} | (i, j) \in \mathcal{E}(t)\}$. In this paper, we assume that $\mathcal{N}_i(t)$ contains agent i itself. Furthermore, we also assume that each agent has only access to its local information and can receive the information from its neighbors. In decentralized online optimization, each

agent $i \in \mathcal{V}$ first chooses a decision point $\mathbf{x}_i(t)$ from the constraint set $\mathcal{K} \subset \mathbb{R}_+^d$ at each iteration $t = 1, \ldots, T$, where T denotes a time horizon. In response, the adversary replies a function $F_{t,i} : \mathcal{K} \to \mathbb{R}_+$ and the agent i receives the reward $F_{t,i}(\mathbf{x}_i(t))$. Therefore, the goal is to maximize the following decentralized online optimization problem,

$$\max_{\mathbf{x} \in \mathcal{K}} F(\mathbf{x}) := \frac{1}{N} \sum_{t=1}^{T} \sum_{i=1}^{N} F_{t,i}(\mathbf{x}), \tag{8}$$

where $F_{t,i} : \mathcal{K} \to \mathbb{R}_+$ is a submodular function, $\mathcal{K}$ is a constraint set. Note that the reward function $F_{t,i}$ becomes available to agent $i \in \mathcal{V}$ only after the agent has chosen an action at each iteration $t \in \{1, \ldots, T\}$. However, each agent can guide their choice by using the information of previously seen functions. This scenario is known as the adversarial online setting, which has an arbitrary sequence of functions $\{F_{t,1}, \ldots, F_{t,N}\}_{t=1}^{T}$. In the adversarial setting, the *adversarial regret* of agent $i \in \mathcal{V}$ with respect to any fixed choice $\mathbf{x} \in \mathcal{K}$ in hindsight is defined as

$$\alpha\text{-}\mathcal{R}_T(\mathbf{x}_i, \mathbf{x}) := \alpha \sup_{\mathbf{x} \in \mathcal{K}} \frac{1}{N} \sum_{t=1}^{T} \sum_{j=1}^{N} F_{t,j}(\mathbf{x}) - \frac{1}{N} \sum_{t=1}^{T} \sum_{j=1}^{N} F_{t,j}(\mathbf{x}_i(t)), \tag{9}$$

which is called α-regret for decentralized adversarial maximization problems. In the definition of α-regret, α is a non-negative constant. The goal is to design efficient decentralized algorithms over time-varying networks so that the upper bounded of α-regret of the algorithms are sublinear in T, i.e., $\lim_{T \to \infty} \alpha\text{-}\mathcal{R}_T / T = 0$.

If the reward functions of agent $i \in \mathcal{V}$ are the expectation of $F_{t,i}(\mathbf{x}) = F_i(\mathbf{x}, \omega_t)$, where ω_t is chosen independent and identically distributed from an unknown distribution $\mathcal{D}$, i.e., $F_i(\mathbf{x}) := \mathbb{E}_{\omega_t \sim \mathcal{D}}[F_{t,i}(\mathbf{x})]$. This scenario is known as the stochastic online setting. In the stochastic online setting, the goal of each agent $i \in \mathcal{V}$ is to maximize the *α-stochastic regret*, which is defined as

$$\alpha\text{-}\mathcal{SR}_T(\mathbf{x}_i, \mathbf{x}) := T \cdot \alpha \sup_{\mathbf{x} \in \mathcal{K}} \frac{1}{N} \sum_{j=1}^{N} F_j(\mathbf{x}) - \frac{1}{N} \sum_{t=1}^{T} \sum_{j=1}^{N} F_j(\mathbf{x}_i(t)). \tag{10}$$

Note that the best approximation guarantee is $(1 - 1/e)$ for problem (8) by using centralized online methods. In this paper, we will design decentralized online algorithms that can achieve the same approximation guarantee by using local communication and local computation.

4.2 Algorithms Design

In this subsection, we first propose a decentralized online learning method over time-varying networks in adversarial online setting. The goal is to solve the problem (8) in a decentralized and cooperative way with $(1 - 1/e)$ approximation guarantee. At each iteration t, each agent i only knows its local information and receives the information from its neighbors. Moreover, since the size of data is huge, we eschew the projection step by using Frank-Wolfe technique, which is a more efficient linear optimization step. Thus, each agent i can use local information $\mathbf{x}_i(t), \mathbf{d}_i(t) \in \mathbb{R}_+^d$ and can receive the information from its neighbors,

Algotithm 1 Decentralized Meta Frank-Wolfe Learning over Time-Varying Networks

Input: Maximum time horizon T; doubly stochastic matrix $A(t) = [a_{ij}(t)] \in \mathbb{R}^{N\times N}$; the number of agents N; parameters η_t and γ_t.

Output:$\{\mathbf{x}_i(t) : 1 \le t \le T\}$ for $i \in \{1,\ldots,N\}$

Initialize online linear optimization oracle $\mathcal{Q}_i^{(1)},\ldots,\mathcal{Q}_i^{(K)}$, $i \in \{1,\ldots,N\}$
Initialize $\mathbf{x}_i^{(0)}(t) = \mathbf{0}$ and $\mathbf{d}_i^{(0)}(t) = \mathbf{0}$
Initialize $\mathbf{x}_j^{(0)}(t) = \mathbf{0}$ and $\mathbf{d}_j^{(0)}(t) = \mathbf{0}$ for all $j \in \mathcal{N}_i(t)$
for $t = 1,\ldots,T$ **do**
  for each agent $i = 1,\ldots,N$ **do**
    for $k = 1,\ldots,K$ **do**
      Obtain $\mathbf{v}_i^{(k)}(t)$ by using the oracle $\mathcal{Q}_i^{(k)}$ in iteration $t-1$
      Update the variable $\mathbf{x}_i^{(k+1)}(t) = \sum_{j\in\mathcal{N}_i(t)} a_{ij}(t)\,\mathbf{x}_j^{(k)}(t) + \frac{1}{K}\mathbf{v}_i^{(k)}(t)$
      Exchange the variable $\mathbf{x}_i^{(k+1)}(t)$ with neighbors $j \in \mathcal{N}_i(t)$
      Compute $\mathbf{g}_i^{(k)}(t) = (1-\eta_k)\,\mathbf{g}_i^{(k-1)}(t) + \eta_k\hat{\nabla}F_{t,i}\left(\mathbf{x}_i^{(k)}(t)\right)$
      Compute $\mathbf{d}_i^{(k)}(t) = (1-\gamma_k)\sum_{j\in\mathcal{N}_i(t)} a_{ij}(t)\,\mathbf{d}_j^{(k-1)}(t) + \gamma_k\mathbf{g}_i^{(k)}(t)$
      Exchange the variable $\mathbf{d}_i^{(k)}(t)$ with the neighbors $j \in \mathcal{N}_i(t)$
      Feedback $\langle \mathbf{d}_i^{(k)}(t), \mathbf{v}\rangle$ to the oracle $\mathcal{Q}_i^{(k)}$
    end for
    Play $\mathbf{x}_i(t) = \mathbf{x}_i^{(K+1)}(t)$, then obtain the value $F_{t,i}(\mathbf{x}_i(t))$ and the unbiased estimate of $\nabla F_{t,i}$
  end for
end for

where $\mathbf{d}_i(t)$ denotes the surrogate of the gradient vector of agent i at iteration t. In this paper, we combine the consensus technique, Frank-Wolfe technique, and variance reduction technique to design the decentralized online learning methods. In the adversarial online setting, the proposed algorithm is summarized in Algorithm 1.

In Algorithm 1, we update the estimate $\mathbf{x}_i(t)$ of agent $i \in \mathcal{V}$ by running K distributed Frank-Wolfe steps. The linear optimization oracle $\mathcal{Q}_i^{(k)}$ is an efficient procedure and can return a vector $\mathbf{v}_i^{(k)}(t)$ in iteration $t-1$ for $i = 1,\ldots,N$ and $k = 1,\ldots,K$. Moreover, the weight that agent i assigns to agent j at iteration t is denoted by $a_{ij}(t)$. At each iteration $t \in \{1,\ldots,T\}$, the approximate gradient vector of agent i, $\mathbf{d}_i^{(k)}(t)$, is updated by using local gradient and the gradient information from the neighbors $j \in \mathcal{N}_i(t)$, i.e.,

$$\mathbf{d}_i^{(k)}(t) := (1-\gamma_k)\sum_{j\in\mathcal{N}_i(t)} a_{ij}(t)\,\mathbf{d}_j^{(k-1)}(t) + \gamma_k\mathbf{g}_i^{(k)}(t), \tag{11}$$

where $\gamma_k \in [0,1]$ denotes the step size and

$$\mathbf{g}_i^{(k)}(t) = (1-\eta_k)\,\mathbf{g}_i^{(k-1)}(t) + \eta_k \hat{\nabla} F_{t,i}\left(\mathbf{x}_i^{(k)}(t)\right)$$

with parameter $\eta_k \in [0,1]$. Moreover, the update rule of estimate of agent i is defined as

$$\mathbf{x}_i^{(k+1)}(t) := \sum_{j \in \mathcal{N}_i(t)} a_{ij}(t)\,\mathbf{x}_j^{(k)}(t) + \frac{1}{K}\mathbf{v}_i^{(k)}(t). \tag{12}$$

Furthermore, the estimate of agent i is obtained at each iteration t by setting $\mathbf{x}_i(t) := \mathbf{x}_i^{(K+1)}(t)$ for all $i \in \{1,\ldots,N\}$.

In the stochastic online setting, we propose a decentralized one-shot Frank-Wolfe online learning method for solving the problem (8). In our proposed algorithm, we use one Frank-Wolfe step at each iteration to avoid the projection step. Moreover, the proposed algorithm only needs to estimate the gradient and can be executed without any linear optimization oracle. Furthermore, the approximation of gradient vector is given by

$$\mathbf{d}_i(t) := (1-\gamma_t) \sum_{j \in \mathcal{N}_i(t)} a_{ij}(t)\,\mathbf{d}_j(t-1) + \gamma_t \hat{\nabla} F_{t,i}(\mathbf{x}_i(t)), \tag{13}$$

where $\gamma_t \in [0,1]$ is the step size. By using the approximate gradient vector $\mathbf{d}_i(t)$, the local ascent direction $\mathbf{v}_i(t)$ of each agent $i \in \mathcal{V}$ is obtained by solving the following linear programming,

$$\mathbf{v}_i(t) := \arg\max_{\mathbf{v} \in \mathcal{K}} \langle \mathbf{d}_i(t), \mathbf{v} \rangle. \tag{14}$$

Finally, using the local ascent directions $\mathbf{v}_i(t)$, each agent i updates its estimate as follows,

$$\mathbf{x}_i(t+1) = \sum_{j \in \mathcal{N}_i(t)} a_{ij}(t)\,\mathbf{x}_j(t) + \frac{1}{T}\mathbf{v}_i(t), \tag{15}$$

where T denotes the time horizon. The detailed description of the proposed algorithm is summarized in Algorithm 2.

Remark: When $N = 1$ in Algorithms 1 and 2, there exist only one agent i. Moreover, let $a_{ij}(t) = a_{ii}(t) = 1$ for all $t \in \{1,\ldots,T\}$ since $i = j$. Furthermore, the set of neighbors of agent i, $\mathcal{N}_i(t)$, only contains agent i itself. Then, Algorithms 1 and 2 respectively reduce to Meta-Frank-Wolfe and One-Shot Frank-Wolfe algorithms (Chen et al., 2018b), which are all implemented in the centralized setting.

4.3 Assumptions

In this subsection, we adopt some assumptions to analyze the performance of the proposed algorithms. Since each agent can exchange information with its neighbors, we model the communication between agents by a stochastic matrix $A(t) = [a_{ij}(t)] \in \mathbb{R}^{N \times N}$, which satisfies the following Assumption 1.

Assumption 1 *For all $i,j \in \mathcal{V}$ and $t \in \{1,\ldots,T\}$, $a_{ij}(t) \geq \mu$ with $\mu \in (0,1)$ if $(i,j) \in \mathcal{E}(t)$, and $a_{ij}(t) = 0$ if $(i,j) \notin \mathcal{E}(t)$. Moreover, we assume that $a_{ii}(t) \geq \mu$ for all $i \in \mathcal{V}$*

Algotithm 2 Decentralized One-Shot Frank-Wolfe Learning over Time-Varying Networks

Input: Maximum time horizon T; doubly stochastic matrix $A(t) = [a_{ij}(t)] \in \mathbb{R}^{N\times N}$; the number of agents N; step sizes γ_t.

Output:$\{\mathbf{x}_i(t) : 1 \le t \le T\}$ for $i \in \{1, \ldots, N\}$

1: Initialize $\mathbf{x}_i^{(0)}(1) = \mathbf{0}$ and $\mathbf{d}_i^{(0)}(t) = \mathbf{0}$
2: Initialize $\mathbf{x}_j^{(0)}(t) = \mathbf{0}$ and $\mathbf{d}_j^{(0)}(t) = \mathbf{0}$ for all $j \in \mathcal{N}_i(t)$
3: **for** $t = 1, \ldots, T$ **do**
4: **for** each agent $i = 1, \ldots, N$ **do**
5: Play $\mathbf{x}_i(t)$, then obtain the value $F_{t,i}(\mathbf{x}_i(t))$ and the unbiased estimate of $\nabla F_{t,i}$
6: Compute $\mathbf{d}_i(t) = (1-\gamma_t)\sum_{j\in\mathcal{N}_i(t)} a_{ij}(t)\,\mathbf{d}_j(t-1) + \gamma_t \hat{\nabla} F_{t,i}(\mathbf{x}_i(t))$
7: Exchange the variable $\mathbf{d}_i(t)$ with the neighbors $j \in \mathcal{N}_i(t)$
8: Evaluate $\mathbf{v}_i(t) = \arg\max_{\mathbf{v}\in\mathcal{K}} \langle \mathbf{d}_i(t), \mathbf{v}\rangle$
9: Update the variable $\mathbf{x}_i(t+1) = \sum_{j\in\mathcal{N}_i(t)} a_{ij}(t)\,\mathbf{x}_j(t) + \frac{1}{T}\mathbf{v}_i(t)$
10: Exchange the variable $\mathbf{x}_i(t+1)$ with neighbors $j \in \mathcal{N}_i(t)$
11: **end for**
12: **end for**

and t. Furthermore, the adjacency matrix $A(t)$ with elements $a_{ij}(t)$ satisfies the following conditions for all $t \ge 0$,

$$\sum_{i=1}^{N} a_{ij}(t) = \sum_{j=1}^{N} a_{ij}(t) = 1. \tag{16}$$

From Assumption 1, we can see that the significant weights are assigned to the estimate of each agent and the estimates of its neighbors. Moreover, the zero weights are assigned to the neighbors $j \in \mathcal{V}$ of agent i when the estimates $\mathbf{x}_j(t)$ are not available at iteration t.

Assumption 2 *The constraint set $\mathcal{K}$ is convex and compact. Furthermore, the diameter and radius of the set $\mathcal{K}$ are $D := \sup_{\mathbf{x},\mathbf{y}\in\mathcal{K}} \|\mathbf{x}-\mathbf{y}\|$ and $R := \sup_{\mathbf{x}\in\mathcal{K}} \|\mathbf{x}\|$, respectively.*

Assumption 3 *In the adversarial setting, each local function $F_{t,i}$ is monotone, DR-submodular and L-smooth. In the stochastic setting, the expected local function of each agent F_i also is monotone, DR-submodular and L-smooth. Furthermore, the gradients of $F_{t,i}$ and F_i are uniformly bounded, respectively, i.e., $\|\nabla F_{t,i}(\mathbf{x})\| \le G$ and $\|\nabla F_i(\mathbf{x})\| \le G$ for all $\mathbf{x} \in \mathbb{R}_+^d$ and $i \in \mathcal{V}$.*

Note that the objective functions $F_t = \sum_{i=1}^N F_{t,i}$ and $F = \sum_{i=1}^N F_i$ are NL-smooth. Moreover, Assumption 3 implies that the functions $F_{t,i}$ and F_i are G-Lipschitz. In addition, we also adopt the following assumption on the connectively of graph $\mathcal{G}(t)$.

Assumption 4 *There exists a constant $B \ge 1$ such that for every B consecutive rounds, agent $i \in \mathcal{V}$ can receive information from its neighboring agent $j \in \mathcal{N}_i(t)$ at least once.*

From Assumption 4, we can see that the graph $\left(\mathcal{V}, \bigcup_{l=0,\ldots,B-1} \mathcal{E}(t+l)\right)$ is strongly connected, which ensures that each agent can receive information from other agents directly and indirectly.

Assumption 5 *In the adversarial online setting, the estimate of gradient $\nabla F_{t,i}$ is unbiased for all $i \in \mathcal{V}$, i.e., $\mathbb{E}[\nabla F_{t,i}(\mathbf{x}) - \hat{\nabla} F_{t,i}(\mathbf{x})] = 0$. Moreover, the variance of the estimate gradient is bounded for all $i \in \mathcal{V}$, i.e., $\mathbb{E}[\|\nabla F_{t,i}(\mathbf{x}) - \hat{\nabla} F_{t,i}(\mathbf{x})\|^2] \leq \sigma^2$. In the stochastic online setting, the estimate of gradient ∇F_i is unbiased for all $i \in \mathcal{V}$, i.e., $\mathbb{E}[\nabla F_i(\mathbf{x}) - \hat{\nabla} F_i(\mathbf{x})] = 0$. Furthermore, the variance of the estimate gradient is bounded for all $i \in \mathcal{V}$, i.e., $\mathbb{E}[\|\nabla F_i(\mathbf{x}) - \hat{\nabla} F_i(\mathbf{x})\|^2] \leq \sigma^2$.*

In this section, we propose some decentralized online learning algorithms to solve the problem of our interest. Moreover, we also adopt some assumptions to analyze the performance of the proposed algorithms. In the next section, we will present the main results of this paper.

5. Main Results

In this section, we present the main results of this paper. In adversarial online setting, we establish a regret bound in expectation as follows.

Theorem 1 *Let Assumptions 1-5 hold. Suppose that the regret of linear optimization oracle is at most $\mathcal{R}_T^{\epsilon}$ for all $i \in \{1, \ldots, N\}$. Furthermore, assume that $\mathbf{x}^*$ is a globally optimal solution of problem (8). The sequences $\{\mathbf{x}_i(t)\}$ and $\{\mathbf{d}_i(t)\}$ are generated by Algorithm 1 for all $i \in \{1, \ldots, N\}$ and $t \in \{1, \ldots, T\}$. By choosing step sizes as $\eta_k = 2/K^{2/3}$ and $\gamma_k = 1/K^{1/2}$, we have*

$$\begin{aligned}\left(1 - \frac{1}{e}\right) \text{-}\overline{\mathcal{R}}_T(\mathbf{x}_j, \mathbf{x}^*) \leq & \left(\frac{LD^2}{2} + \frac{GND\nu}{1-\beta} + \frac{LND^2\nu}{1-\beta}\right) \cdot \frac{T}{K} + \mathcal{R}_T^{\epsilon} \\ & + \left(GD + LD^2 + \frac{ND\nu\sqrt{2(\sigma^2 + G^2)}}{1-\beta}\right) \cdot \frac{T}{K^{1/2}} \\ & + \frac{LD^2\sqrt{3 + 3\sqrt{2}N\nu/(1-\beta)}}{\sqrt{2}} \cdot \frac{T}{K^{2/3}} \\ & + \left(GD + \sqrt{2}\sigma D + \frac{LD^2\sqrt{3 + 3\sqrt{2}N\nu/(1-\beta)}}{\sqrt{2}}\right) \cdot \frac{T}{K^{1/3}},\end{aligned} \tag{17}$$

where $\overline{\mathcal{R}}_T(\mathbf{x}_j, \mathbf{x}^*) = \mathbb{E}[\mathcal{R}_T(\mathbf{x}_j, \mathbf{x}^*)]$.

The proof can be found in the next section. From Theorem 1, we choose the online linear optimization oracle as Regularized-Follow-The-Leader (RFTL) (Cohen and Hazan, 2015), then $\mathcal{R}_T^{\epsilon} = \mathcal{O}(\sqrt{T})$. Moreover, setting $K = T^{3/2}$, the square-root regret $\mathcal{O}(\sqrt{T})$ is obtained by Algorithm 1, which implies that $(1 - 1/e)\text{-}\overline{\mathcal{R}}_T(\mathbf{x}_j, \mathbf{x}^*)/T \leq \mathcal{O}(1/\sqrt{T})$. Therefore, after $\mathcal{O}(1/\epsilon^2)$ rounds of communication, the $(1 - 1/e - \epsilon)$ approximation ratio can be achieved, where ϵ is a positive constant.

In the stochastic online setting, we also establish the regret bound in expectation for Algorithm 2, which is stated as follows.

Theorem 2 *Let Assumptions 1-5 hold. Suppose that* $\mathbf{x}^*$ *is a globally optimal solution of problem* (8). *The sequences* $\{\mathbf{x}_i(t)\}$ *and* $\{\mathbf{d}_i(t)\}$ *are generated by Algorithm 2 for all* $i \in \{1,\ldots,N\}$ *and* $t \in \{1,\ldots,T\}$. *Moreover, let the step sizes be* $\eta_t = 2/T^{2/3}$ *and* $\gamma_t = 1/T^{1/2}$. *Then, we have*

$$\begin{aligned}\left(1-\frac{1}{e}\right)\text{-}\overline{\mathcal{SR}}_T(\mathbf{x}_j,\mathbf{x}^*) \leq & \left(GD+\frac{LD^2}{2}+\frac{N\nu\sqrt{\sigma^2+G^2}}{\sqrt{2}(1-\beta)}\right)\cdot T^{1/2} \\ & +\left(\frac{\sqrt{2}\sigma D}{2}+\frac{LD^2\sqrt{3+3\sqrt{2}N\nu/(1-\beta)}}{2\sqrt{2}}\right)\cdot T^{2/3} \\ & +\frac{LD^2\sqrt{3+3\sqrt{2}N\nu/(1-\beta)}}{2\sqrt{2}}\cdot T^{1/3}+\frac{LD^2}{2} \\ & +\frac{LND^2\nu}{2(1-\beta)}+\frac{GND\nu}{1-\beta},\end{aligned} \tag{18}$$

where $\overline{\mathcal{SR}}_T(\mathbf{x}_j,\mathbf{x}^*) = \mathbb{E}[\mathcal{SR}_T(\mathbf{x}_j,\mathbf{x}^*)]$.

The proof can be found in the next section. From Theorem 2, we can see that the regret bound of $\mathcal{O}(T^{2/3})$ by choosing appropriate step sizes, then $(1-1/e)\text{-}\overline{\mathcal{SR}}_T(\mathbf{x}_j,\mathbf{x}^*)/T \leq \mathcal{O}(1/T^{1/3})$. Therefore, after $\mathcal{O}(1/\epsilon^3)$ rounds of communication, the $(1-1/e-\epsilon)$ approximation ratio can be achieved.

In this section, we establish the regret bounds in the adversarial and stochastic online settings, respectively. The detailed proofs of the main results are given in the next section.

6. Performance Analysis

In this section, we analyze the performance of the proposed algorithms in the adversarial and stochastic online settings, respectively. Moreover, we also provide the detailed proof of main results of this paper. First, we study the convergence property of Algorithm 1. Afterwards, the performance of Algorithm 2 is also studied.

6.1 Adversarial Online Setting

We first analyze the performance of Algorithm 1 under the adversarial online setting. For this purpose, we introduce an auxiliary vector for all $k = 1,\ldots,K$ as follows,

$$\bar{\mathbf{x}}^{(k)}(t) := \frac{1}{N}\sum_{i=1}^{N}\mathbf{x}_i^{(k)}(t). \tag{19}$$

Moreover, we establish an upper bound of the distance between $\bar{\mathbf{x}}^{(k+1)}(t)$ and $\bar{\mathbf{x}}^{(k)}(t)$, i.e.,

Lemma 3 *Let Assumptions 1, 2, and 4 hold. The sequence* $\{\mathbf{x}_i^{(k)}(t)\}$ *is generated by Algorithm 1. For all* $t \in \{1,\ldots,T\}$ *and* $k = 1,\ldots,K$, *we have*

$$\left\|\bar{\mathbf{x}}^{(k+1)}(t)-\bar{\mathbf{x}}^{(k)}(t)\right\| \leq \frac{D}{K}. \tag{20}$$

Proof According to the definition of $\bar{\mathbf{x}}^{(k)}(t)$ and Eq. (12), we have

$$\begin{aligned}
\bar{\mathbf{x}}^{(k+1)}(t) &= \frac{1}{N}\sum_{i=1}^{N}\bar{\mathbf{x}}_i^{(k+1)}(t) \\
&= \frac{1}{N}\sum_{i=1}^{N}\sum_{j\in\mathcal{N}_i(t)} a_{ij}(t)\,\mathbf{x}_j^{(k)}(t) + \frac{1}{K}\frac{1}{N}\sum_{i=1}^{N}\mathbf{v}_i^{(k)}(t) \\
&= \frac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{N} a_{ij}(t)\,\mathbf{x}_j^{(k)}(t) + \frac{1}{K}\frac{1}{N}\sum_{i=1}^{N}\mathbf{v}_i^{(k)}(t) \\
&= \frac{1}{N}\sum_{j=1}^{N}\mathbf{x}_j^{(k)}(t)\sum_{i=1}^{N} a_{ij}(t) + \frac{1}{K}\frac{1}{N}\sum_{i=1}^{N}\mathbf{v}_i^{(k)}(t) \\
&= \frac{1}{N}\sum_{j=1}^{N}\mathbf{x}_j^{(k)}(t) + \frac{1}{K}\frac{1}{N}\sum_{i=1}^{N}\mathbf{v}_i^{(k)}(t) \\
&= \bar{\mathbf{x}}^{(k)}(t) + \frac{1}{K}\frac{1}{N}\sum_{i=1}^{N}\mathbf{v}_i^{(k)}(t),
\end{aligned} \tag{21}$$

where the fifth equality is due to $\sum_{i=1}^{N} a_{ij}(t) = 1$, in the last equality we have used the definition of $\bar{\mathbf{x}}^{(k)}(t)$. From Assumption 2, we have $\left\|\mathbf{v}_i^{(k)}(t)\right\| \leq D$ since $\mathbf{v}_i^{(k)}(t) \in \mathcal{K}$ for all i, k, t. Thus, we obtain

$$\begin{aligned}
\left\|\bar{\mathbf{x}}^{(k+1)}(t) - \bar{\mathbf{x}}^{(k)}(t)\right\| &\leq \frac{1}{K}\frac{1}{N}\sum_{i=1}^{N}\left\|\mathbf{v}_i^{(k)}(t)\right\| \\
&\leq \frac{D}{K}.
\end{aligned} \tag{22}$$

The lemma is proved completely. ■

In order to prove the main results of this paper, we also introduce a matrix, which is defined as follows:

$$\Phi(s:t) := A(s)\,A(s+1)\cdots A(t-1)\,A(t).$$

Moreover, the i-th row and the j-th column of $\Phi(s:t)$ is denoted by $[\Phi(s:t)]_{ij}$. From Assumptions 1 and 4, we have the following result, which is presented in Nedić et al. (2008) (Corollary 1).

$$\left|[\Phi(s:t)]_{ij} - \frac{1}{N}\right| \leq \nu\beta^{t-s+1}, \tag{23}$$

where $\nu = \left(1 - \frac{\mu}{4N^2}\right)^{-2}$ and $\beta = \left(1 - \frac{\mu}{4N^2}\right)^{1/B}$.

Next, an upper bound the sum of the distance between the local estimate $\mathbf{x}_i^{(k)}(t)$ and $\bar{\mathbf{x}}^{(k)}(t)$ is established as follows.

Lemma 4 *Let Assumptions 1, 2, and 4 hold. The sequence* $\{\mathbf{x}_i^{(k)}(t)\}$ *is generated by Algorithm 1. For all* $t \in \{1,\ldots,T\}$ *and* $k = 1,\ldots,K$*, we have*

$$\sqrt{\sum_{i=1}^{N}\left\|\mathbf{x}_i^{(k)}(t)-\bar{\mathbf{x}}^{(k)}(t)\right\|^2} \leq \frac{N\sqrt{N}D\nu}{K(1-\beta)}. \tag{24}$$

Proof We introduce two auxiliary vectors as follows:

$$\mathbf{x}^{(k)}(t) := \left[\mathbf{x}_1^{(k)}(t);\ldots;\mathbf{x}_N^{(k)}(t)\right] \in \mathbb{R}_+^{Nd}$$

and

$$\mathbf{v}^{(k)}(t) := \left[\mathbf{v}_1^{(k)}(t);\ldots;\mathbf{v}_N^{(k)}(t)\right] \in \mathbb{R}_+^{Nd}.$$

Following on from Eq. (12), we have

$$\mathbf{x}^{(k+1)}(t) = (A(t)\otimes I)\,\mathbf{x}^{(k)}(t) + \frac{1}{K}\mathbf{v}^{(k)}(t), \tag{25}$$

where the notation $\otimes$ denotes the Kronecker product, I denotes the identity matrix of size $d\times d$. By using Eq. (25) recursively, we obtain

$$\mathbf{x}^{(k)}(t) = \frac{1}{K}\sum_{s=0}^{k-1}(\Phi(s+1:k-1)\otimes I)\,\mathbf{v}^{(s)}(t), \tag{26}$$

where we have used the fact $\mathbf{x}_i^{(0)}(t) = \mathbf{0}$ for all $i = 1,\ldots,N$. In both sides of Eq. (26), we multiply by the matrix $(\frac{\mathbb{1}\mathbb{1}^\mathsf{T}}{N}\otimes I)$, and then we have

$$\left(\frac{\mathbb{1}\mathbb{1}^\mathsf{T}}{N}\otimes I\right)\mathbf{x}^{(k)}(t) = \frac{1}{K}\sum_{s=0}^{k-1}\left(\left(\frac{\mathbb{1}\mathbb{1}^\mathsf{T}}{N}\Phi(s+1:k-1)\right)\otimes I\right)\mathbf{v}^{(k)}(t). \tag{27}$$

We define a variable $\tilde{\mathbf{x}}^{(k)}(t)$ as

$$\tilde{\mathbf{x}}^{(k)}(t) := \left[\bar{\mathbf{x}}^{(k)}(t);\ldots;\bar{\mathbf{x}}^{(k)}(t)\right] \in \mathbb{R}_+^{Nd}.$$

From the definition of $\bar{\mathbf{x}}^{(k)}(t)$ in Eq. (19), we have

$$\tilde{\mathbf{x}}^{(k)}(t) = \left(\frac{\mathbb{1}\mathbb{1}^\mathsf{T}}{N}\otimes I\right)\mathbf{x}^{(k)}(t). \tag{28}$$

Since the matrix $A(t)$ is doubly stochastic, we have $\mathbb{1}\mathbb{1}^\mathsf{T}A(t) = \mathbb{1}\mathbb{1}^\mathsf{T}$. Thus, Eq. (28) can be rewritten as

$$\tilde{\mathbf{x}}^{(k)}(t) = \frac{1}{K}\sum_{s=0}^{k-1}\left(\frac{\mathbb{1}\mathbb{1}^\mathsf{T}}{N}\otimes I\right)\mathbf{v}^{(s)}(t). \tag{29}$$

Combining Eqs. (26) and (29), we obtain

$$\begin{aligned}\left\|\mathbf{x}^{(k)}(t)-\tilde{\mathbf{x}}^{(k)}(t)\right\| &= \frac{1}{K}\left\|\sum_{s=0}^{k-1}\left(\left(\Phi(s+1:k-1)\otimes I\right)-\frac{\mathbb{1}\mathbb{1}^{\mathsf{T}}}{N}\otimes I\right)\mathbf{v}^{(s)}(t)\right\| \\ &= \frac{1}{K}\left\|\sum_{s=0}^{k-1}\left(\left(\Phi(s+1:k-1)-\frac{\mathbb{1}\mathbb{1}^{\mathsf{T}}}{N}\right)\otimes I\right)\mathbf{v}^{(k)}(t)\right\| \\ &\leq \frac{1}{K}\sum_{s=0}^{k-1}\left\|\Phi(s+1:k-1)-\frac{\mathbb{1}\mathbb{1}^{\mathsf{T}}}{N}\right\|\left\|\mathbf{v}^{(k)}(t)\right\| \\ &\leq \frac{\sqrt{N}D}{K}\sum_{s=0}^{k-1}\left\|\Phi(s+1:k-1)-\frac{\mathbb{1}\mathbb{1}^{\mathsf{T}}}{N}\right\|,\end{aligned} \tag{30}$$

where in the first inequality we have used the Cauchy-Schwarz inequality, the last inequality follows from the norm equivalence $\left\|\mathbf{v}^{(k)}(t)\right\| \leq \sqrt{N}\left\|\mathbf{v}_i^{(k)}(t)\right\| \leq \sqrt{N}D$. In addition, from the property of the matrix norms (Golub and Van Loan, 2013) (see Eq. (2.3.8)), we have

$$\begin{aligned}\left\|\Phi(s+1:k-1)-\frac{\mathbb{1}\mathbb{1}^{\mathsf{T}}}{N}\right\| &\leq N\max_{i,j}\left|\left[\Phi(s+1:k-1)\right]_{ij}-\frac{1}{N}\right| \\ &\leq N\nu\beta^{k-s-1},\end{aligned} \tag{31}$$

where the last inequality holds due to Eq. (23). Plugging Eq. (31) into Eq. (30), we obtain

$$\begin{aligned}\left\|\mathbf{x}^{(k)}(t)-\tilde{\mathbf{x}}^{(k)}(t)\right\| &\leq \frac{N\sqrt{N}D\nu}{K}\sum_{s=0}^{k-1}\beta^{k-1-s} \\ &\leq \frac{N\sqrt{N}D\nu}{K(1-\beta)},\end{aligned} \tag{32}$$

where the last inequality follows from

$$\sum_{s=0}^{k-1}\beta^{k-1-s} \leq \sum_{s=0}^{\infty}\beta^{k-1-s} = 1/(1-\beta)$$

for $0<\beta<1$. Since

$$\sum_{i=1}^{N}\left\|\mathbf{x}_i^{(k)}(t)-\bar{\mathbf{x}}^{(k)}(t)\right\|^2 = \left\|\mathbf{x}^{(k)}(t)-\tilde{\mathbf{x}}^{(k)}(t)\right\|^2,$$

the statement of the lemma is obtained completely. ■

In addition, we also define the following auxiliary variable,

$$\bar{\mathbf{d}}^{(k)}(t) := \frac{1}{N}\sum_{i=1}^{N}\mathbf{d}_i^{(k)}(t). \tag{33}$$

Next, we establish an upper bound of the sum of the expected distance between the vectors $\mathbf{d}_i^{(k)}(t)$ and $\bar{\mathbf{d}}^{(k)}(t)$ for all $i=1,\ldots,N$, which is given as follows.

Lemma 5 *Let Assumption 1, 3, and 4 hold. The sequences* $\{\mathbf{x}_i^{(k)}(t)\}$ *and* $\{\mathbf{d}_i^{(k)}(t)\}$ *are generated by Algorithm 1. For all* $t \in \{1,\ldots,T\}$ *and* $k = 1,\ldots,K$, *we have*

$$\sqrt{\sum_{i=1}^{N}\mathbb{E}\left[\left\|\mathbf{d}_i^{(k)}(t)-\bar{\mathbf{d}}^{(k)}(t)\right\|^2\right]} \leq \frac{N\nu\gamma_k\sqrt{2N\left(\sigma^2+G^2\right)}}{1-\beta\left(1-\gamma_k\right)}. \tag{34}$$

Proof We define some auxiliary variables as

$$\mathbf{d}^{(k)}(t) := \left[\mathbf{d}_1^{(k)}(t);\ldots;\mathbf{d}_N^{(k)}(t)\right] \in \mathbb{R}_+^{Nd}$$

and

$$\mathbf{g}^{(k)}(t) := \left[\mathbf{g}_1^{(k)}(t);\ldots;\mathbf{g}_N^{(k)}(t)\right] \in \mathbb{R}_+^{Nd}.$$

According to Eq. (11), we have

$$\mathbf{d}^{(k)}(t) = (1-\gamma_k)\left(A(t)\otimes I\right)\mathbf{d}^{(k-1)}(t) + \gamma_k\mathbf{g}^{(k)}(t), \tag{35}$$

where $F_t\left(\mathbf{x}^{(k)}(t)\right) := \sum_{i=1}^{N} F_{t,i}\left(\mathbf{x}_i^{(k)}(t)\right)$. Since $\mathbf{d}^{(0)}(t) = \mathbf{0}$, we use Eq. (35) recursively to obtain the following equality,

$$\mathbf{d}^{(k)}(t) = \gamma_k\sum_{s=1}^{k}\left(\left(1-\gamma_k\right)^{k-s}\Phi(s+1:k)\otimes I\right)\mathbf{g}^{(s)}(t). \tag{36}$$

Multiplying the matrix $(\frac{\mathbb{1}\mathbb{1}^\mathsf{T}}{N}\otimes I)$ in both sides of Eq. (36), we have

$$\hat{\mathbf{d}}^{(k)}(t) = \gamma_k\sum_{s=0}^{k}\left(1-\gamma_k\right)^{k-s}\left(\frac{\mathbb{1}\mathbb{1}^\mathsf{T}}{N}\otimes I\right)\mathbf{g}^{(s)}(t), \tag{37}$$

where $\hat{\mathbf{d}}^{(k)}(t) := \left[\bar{\mathbf{d}}^{(k)}(t);\ldots;\bar{\mathbf{d}}^{(k)}(t)\right] \in \mathbb{R}^{Nd}$. Therefore, according to Eqs. (36) and (37), we obtain

$$\begin{aligned}
\left\|\mathbf{d}^{(k)}(t)-\hat{\mathbf{d}}^{(k)}(t)\right\| &= \gamma_k\left\|\sum_{s=1}^{k}\left(1-\gamma_k\right)^{k-s}\left(\left(\Phi(s+1:k)-\frac{\mathbb{1}\mathbb{1}^\mathsf{T}}{N}\right)\otimes I\right)\mathbf{g}^{(s)}(t)\right\| \\
&\leq \gamma_k\sum_{s=1}^{k}\left(1-\gamma_k\right)^{k-s}\left\|\Phi(s+1:k)-\frac{\mathbb{1}\mathbb{1}^\mathsf{T}}{N}\right\|\left\|\mathbf{g}^{(s)}(t)\right\| \\
&\leq N\nu\gamma_k\sum_{s=1}^{k}\left(1-\gamma_k\right)^{k-s}\beta^{k-s}\left\|\mathbf{g}^{(s)}(t)\right\|,
\end{aligned} \tag{38}$$

where the first inequality follows from the Cauchy-Schwarz inequality, the last inequality is due to Eq. (31), i.e., $\left\|\Phi(s+1:k)-\left(\mathbb{1}\mathbb{1}^\mathsf{T}\right)/N\right\| \leq N\nu\beta^{k-s}$. Taking expectation on both sides of Eq. (38), we have

$$\mathbb{E}\left[\left\|\mathbf{d}^{(k)}(t)-\hat{\mathbf{d}}^{(k)}(t)\right\|\right] \leq V\nu\gamma_k\sum_{s=1}^{k}\left(\beta\left(1-\gamma_k\right)\right)^{k-s}\mathbb{E}\left[\left\|\mathbf{g}^{(s)}(t)\right\|\right]. \tag{39}$$

In order to establish the upper bound of $\mathbb{E}\left[\left\|\mathbf{d}^{(k)}(t)-\hat{\mathbf{d}}^{(k)}(t)\right\|\right]$, we need to bound the term $\mathbb{E}\left[\left\|\mathbf{g}^{(s)}(t)\right\|\right]$. To this end, we first have

$$\begin{aligned}\left\|\hat{\nabla}F_{t,i}\left(\mathbf{x}_i^{(s)}(t)\right)\right\|^2 &= \left\|\hat{\nabla}F_{t,i}\left(\mathbf{x}_i^{(s)}(t)\right)-\nabla F_{t,i}\left(\mathbf{x}_i^{(s)}(t)\right)+\nabla F_{t,i}\left(\mathbf{x}_i^{(s)}(t)\right)\right\|^2\\ &\leq \left\|\hat{\nabla}F_{t,i}\left(\mathbf{x}_i^{(s)}(t)\right)-\nabla F_{t,i}\left(\mathbf{x}_i^{(s)}(t)\right)\right\|^2+\left\|\nabla F_{t,i}\left(\mathbf{x}_i^{(s)}(t)\right)\right\|^2\\ &+2\left\|\hat{\nabla}F_{t,i}\left(\mathbf{x}_i^{(s)}(t)\right)-\nabla F_{t,i}\left(\mathbf{x}_i^{(s)}(t)\right)\right\|\left\|\nabla F_{t,i}\left(\mathbf{x}_i^{(s)}(t)\right)\right\|\\ &\leq 2\left(\left\|\hat{\nabla}F_{t,i}\left(\mathbf{x}_i^{(s)}(t)\right)-\nabla F_{t,i}\left(\mathbf{x}_i^{(s)}(t)\right)\right\|^2+\left\|\nabla F_{t,i}\left(\mathbf{x}_i^{(s)}(t)\right)\right\|^2\right),\end{aligned} \tag{40}$$

where in the first inequality we have used the Cauchy-Schwarz inequality, the last inequality follows from the inequality $2ab\leq a^2+b^2$. Taking the expectation on both sides of (40) and using the relation $\left\|\nabla F_{t,i}(\mathbf{x}_i^{(s)}(t))\right\|\leq G$ for all $s=1,\ldots,K$, we obtain

$$\mathbb{E}\left[\left\|\hat{\nabla}F_{t,i}\left(\mathbf{x}_i^{(s)}(t)\right)\right\|^2\right]\leq 2\left(\sigma^2+G^2\right) \tag{41}$$

for all $s=1,\ldots,K$. Since $\mathbf{g}_i^{(0)}(t)=\mathbf{0}$, we have

$$\begin{aligned}\mathbb{E}\left[\left\|\mathbf{g}_i^{(1)}(t)\right\|^2|\mathbf{x}_i^{(1)}(t)\right] &= \eta_k^2\mathbb{E}\left[\hat{\nabla}F_{t,i}\left(\mathbf{x}_i^{(1)}(t)\right)|\mathbf{x}_i^{(1)}(t)\right]\\ &\leq 2\eta_k^2\left(\sigma^2+G^2\right)\leq 2\left(\sigma^2+G^2\right).\end{aligned}$$

Taking the expectation on the above inequality, we also obtain

$$\mathbb{E}\left[\left\|\mathbf{g}_i^{(1)}(t)\right\|^2\right]\leq 2\left(\sigma^2+G^2\right).$$

We assume that the inequality $\mathbb{E}\left[\left\|\mathbf{g}_i^{(k-1)}(t)\right\|^2\right]\leq 2\left(\sigma^2+G^2\right)$ holds for $k\in\{1,\ldots,K\}$. Then, we next prove that the inequality $\mathbb{E}\left[\left\|\mathbf{g}_i^{(k)}(t)\right\|^2\right]\leq 2\left(\sigma^2+G^2\right)$ holds for $k\in\{1,\ldots,K\}$. According to the definition of $\mathbf{g}_i^{(k)}(t)$, we have

$$\begin{aligned}\left\|\mathbf{g}_i^{(k)}(t)\right\|^2 &= \left\|(1-\eta_k)\,\mathbf{g}_i^{(k-1)}(t)+\eta_k\hat{\nabla}F_{t,i}\left(\mathbf{x}_i^{(k)}(t)\right)\right\|^2\\ &= (1-\eta_k)^2\left\|\mathbf{g}_i^{(k-1)}(t)\right\|^2+\eta_k^2\left\|\hat{\nabla}F_{t,i}\left(\mathbf{x}_i^{(k)}(t)\right)\right\|^2\\ &+2\eta_k(1-\eta_k)\langle\mathbf{g}_i^{(k-1)}(t),\hat{\nabla}F_{t,i}\left(\mathbf{x}_i^{(k)}(t)\right)\rangle\\ &\leq (1-\eta_k)^2\left\|\mathbf{g}_i^{(k-1)}(t)\right\|^2+\eta_k^2\left\|\hat{\nabla}F_{t,i}\left(\mathbf{x}_i^{(k)}(t)\right)\right\|^2\\ &+2\eta_k(1-\eta_k)\left\|\mathbf{g}_i^{(k-1)}(t)\right\|\left\|\hat{\nabla}F_{t,i}\left(\mathbf{x}_i^{(k)}(t)\right)\right\|,\end{aligned} \tag{42}$$

where the last inequality follows from the Cauchy-Schwartz inequality. Taking the expectation on both sides of Eq. (42) with respect to $\mathbf{x}_i^{(k)}(t)$, we obtain

$$\begin{aligned}\mathbb{E}\left[\left\|\mathbf{g}_i^{(k)}(t)\right\|^2\mid\mathbf{x}_i^{(k)}(t)\right] &\leq (1-\eta_k)^2\left\|\mathbf{g}_i^{(k-1)}(t)\right\|^2+\eta_k^2\mathbb{E}\left[\left\|\hat{\nabla}F_{t,i}\left(\mathbf{x}_i^{(k)}(t)\right)\right\|^2\right]\\ &+2\eta_k(1-\eta_k)\left\|\mathbf{g}_i^{(k-1)}(t)\right\|\mathbb{E}\left[\left\|\hat{\nabla}F_{t,i}\left(\mathbf{x}_i^{(k)}(t)\right)\right\|\mid\mathbf{x}_i^{(k)}(t)\right].\end{aligned} \tag{43}$$

In addition, using the inequality $\mathbb{E}[\|\mathbf{x}\|] \leq \sqrt{\mathbb{E}[\|\mathbf{x}\|^2]}$ and Eq. (41), we have

$$\mathbb{E}\left[\left\|\hat{\nabla} F_{t,i}\left(\mathbf{x}_i^{(k)}(t)\right)\right\|\right] \leq \sqrt{2\left(\sigma^2+G^2\right)}. \tag{44}$$

Plugging Eq. (44) into Eq. (43), we obtain

$$\begin{aligned}\mathbb{E}\left[\left\|\mathbf{g}_i^{(k)}(t)\right\|^2 \mid \mathbf{x}_i^{(k)}(t)\right] \leq & (1-\eta_k)^2\left\|\mathbf{g}_i^{(k-1)}(t)\right\|^2+2\eta_k^2\left(\sigma^2+G^2\right) \\ & +2\sqrt{2\left(\sigma^2+G^2\right)}\eta_k(1-\eta_k)\left\|\mathbf{g}_i^{(k-1)}(t)\right\|.\end{aligned} \tag{45}$$

Taking the expectation on both sides of the above inequality, we have

$$\begin{aligned}\mathbb{E}\left[\left\|\mathbf{g}_i^{(k)}(t)\right\|^2\right] & \leq (1-\eta_k)^2\,\mathbb{E}\left[\left\|\mathbf{g}_i^{(k-1)}(t)\right\|^2\right]+2\eta_k^2\left(\sigma^2+G^2\right) \\ & \quad +2\sqrt{2\left(\sigma^2+G^2\right)}\eta_k(1-\eta_k)\,\mathbb{E}\left[\left\|\mathbf{g}_i^{(k-1)}(t)\right\|\right] \\ & \leq 2(1-\eta_k)^2\left(\sigma^2+G^2\right)+2\eta_k^2\left(\sigma^2+G^2\right)+4\left(\sigma^2+G^2\right)\eta_k(1-\eta_k) \\ & =2\left(\sigma^2+G^2\right),\end{aligned} \tag{46}$$

where the last inequality follows from the induction. From Eq. (46) and using the inequality $\mathbb{E}[\|\mathbf{x}\|] \leq \sqrt{\mathbb{E}[\|\mathbf{x}\|^2]}$, we have

$$\mathbb{E}\left[\left\|\mathbf{g}^{(k)}(t)\right\|\right] \leq \sqrt{2N\left(\sigma^2+G^2\right)}, \tag{47}$$

where in the last inequality we have used the definition of $\mathbf{g}^{(k)}(t)$. Plugging Eq. (47) into Eq. (39), we obtain

$$\begin{aligned}\mathbb{E}\left[\left\|\mathbf{d}^{(k)}(t)-\hat{\mathbf{d}}^{(k)}(t)\right\|\right] & \leq N\nu\gamma_k\sqrt{2N\left(\sigma^2+G^2\right)}\sum_{s=1}^{k}\left(\beta(1-\gamma_k)\right)^{k-s} \\ & \leq \frac{N\nu\gamma_k\sqrt{2N\left(\sigma^2+G^2\right)}}{1-\beta(1-\gamma_k)}.\end{aligned} \tag{48}$$

Since $\left\|\mathbf{d}^{(k)}(t)-\hat{\mathbf{d}}^{(k)}(t)\right\|^2=\sum_{i=1}^{N}\left\|\mathbf{d}_i^{(k)}(t)-\bar{\mathbf{d}}^{(k)}(t)\right\|^2$, the conclusion of the lemma is obtained. ■

Furthermore, we also have the following lemma, which establishes an upper bound of the sum of the distance between the vectors $\nabla F_{t,i}$ and $\mathbf{g}_i^k(t)$ for all $i \in \mathcal{V}$.

Lemma 6 *Let Assumptions 1-5 hold. The sequences $\{\mathbf{x}_i^{(k)}(t)\}$ and $\{\mathbf{g}_i^{(k)}(t)\}$ are generated by Algorithm 1. For all $t \in \{1,\ldots,T\}$, $i \in \mathcal{V}$ and $k=1,\ldots,K$, we set $\eta_k = 2/K^{2/3}$ and have*

$$\begin{aligned}\mathbb{E}\left[\sum_{i=1}^{N}\left\|\nabla F_{t,i}\left(\mathbf{x}_i^{(k)}(t)\right)-\mathbf{g}_i^{(k)}(t)\right\|^2\right] & \leq \left(1-\frac{2}{K^{2/3}}\right)^k NG^2+\frac{3\kappa NL^2D^2}{2K^{4/3}} \\ & \quad +\frac{4N\sigma^2+3\kappa NL^2D^2}{2K^{2/3}},\end{aligned} \tag{49}$$

where $\kappa := 1+2N^2\nu^2/(1-\beta)^2$.

Proof According to the update rule of $\mathbf{g}_i^{(k)}(t)$, we have

$$\begin{aligned}\left\|\nabla F_{t,i}\left(\mathbf{x}_i^{(k)}(t)\right)-\mathbf{g}_i^{(k)}(t)\right\|^2 &= \left\|\nabla F_{t,i}\left(\mathbf{x}_i^{(k)}(t)\right)-(1-\eta_k)\,\mathbf{g}_i^{(k-1)}(t)-\eta_k\hat{\nabla}F_{t,i}\left(\mathbf{x}_i^{(k)}(t)\right)\right\|^2\\ &= \left\|\eta_k\left(\nabla F_{t,i}\left(\mathbf{x}_i^{(k)}(t)\right)-\hat{\nabla}F_{t,i}\left(\mathbf{x}_i^{(k)}(t)\right)\right)\right.\\ &\quad+(1-\eta_k)\left(\nabla F_{t,i}\left(\mathbf{x}_i^{(k)}(t)\right)-\nabla F_{t,i}\left(\mathbf{x}_i^{(k-1)}(t)\right)\right)\\ &\quad\left.+(1-\eta_k)\left(\nabla F_{t,i}\left(\mathbf{x}_i^{(k-1)}(t)\right)-\mathbf{g}_i^{(k-1)}(t)\right)\right\|^2,\end{aligned} \tag{50}$$

where we have added and subtracted the term $(1-\eta_k)\nabla F_{t,i}\big(\mathbf{x}_i^{(k-1)}(t)\big)$ in the last equality. Let $\mathcal{F}_{t,k}$ denote all the information of random variables generated by Algorithm 1 up to time k and iteration t. Thus, taking the expectation on both sides of (50) with respect to $\mathcal{F}_{t,k}$ and using some algebraic manipulations, we have

$$\begin{aligned}\mathbb{E}\left[\left\|\nabla F_{t,i}\left(\mathbf{x}_i^{(k)}(t)\right)-\mathbf{g}_i^{(k)}(t)\right\|^2\mid\mathcal{F}_{t,k}\right] &\le \eta_k^2\mathbb{E}\left[\left\|\nabla F_{t,i}\left(\mathbf{x}_i^{(k)}(t)\right)-\hat{\nabla}F_{t,i}\left(\mathbf{x}_i^{(k)}(t)\right)\right\|^2\mid\mathcal{F}_{t,k}\right]\\ &\quad+(1-\eta_k)^2\,\mathbb{E}\left[\left\|\nabla F_{t,i}\left(\mathbf{x}_i^{(k)}(t)\right)-\nabla F_{t,i}\left(\mathbf{x}_i^{(k-1)}(t)\right)\right\|^2\mid\mathcal{F}_{t,k}\right]\\ &\quad+(1-\eta_k)^2\,\mathbb{E}\left[\left\|\nabla F_{t,i}\left(\mathbf{x}_i^{(k-1)}(t)\right)-\mathbf{g}_i^{(k-1)}(t)\right\|^2\right]+2\,(1-\eta_k)^2\\ &\quad\times\mathbb{E}\left[\left\langle\nabla F_{t,i}\left(\mathbf{x}_i^{(k)}(t)\right)-\nabla F_{t,i}\left(\mathbf{x}_i^{(k-1)}(t)\right),\nabla F_{t,i}\left(\mathbf{x}_i^{(k-1)}(t)\right)-\mathbf{g}_i^{(k-1)}(t)\right\rangle\mid\mathcal{F}_{t,k}\right],\end{aligned} \tag{51}$$

where the inequality follows from the fact that $\hat{\nabla}F_{t,i}$ is an unbiased estimate of $\nabla F_{t,i}$. Taking the expectation on both sides of Eq. (51), we obtain

$$\begin{aligned}\mathbb{E}\left[\left\|\nabla F_{t,i}\left(\mathbf{x}_i^{(k)}(t)\right)-\mathbf{g}_i^{(k)}(t)\right\|^2\right] &\le \eta_k^2\mathbb{E}\left[\left\|\nabla F_{t,i}\left(\mathbf{x}_i^{(k)}(t)\right)-\hat{\nabla}F_{t,i}\left(\mathbf{x}_i^{(k)}(t)\right)\right\|^2\right]\\ &\quad+(1-\eta_k)^2\,\mathbb{E}\left[\left\|\nabla F_{t,i}\left(\mathbf{x}_i^{(k)}(t)\right)-\nabla F_{t,i}\left(\mathbf{x}_i^{(k-1)}(t)\right)\right\|^2\right]\\ &\quad+(1-\eta_k)^2\,\mathbb{E}\left[\left\|\nabla F_{t,i}\left(\mathbf{x}_i^{(k-1)}(t)\right)-\mathbf{g}_i^{(k-1)}(t)\right\|^2\right]+2\,(1-\eta_k)^2\\ &\quad\times\mathbb{E}\left[\left\langle\nabla F_{t,i}\left(\mathbf{x}_i^{(k)}(t)\right)-\nabla F_{t,i}\left(\mathbf{x}_i^{(k-1)}(t)\right),\nabla F_{t,i}\left(\mathbf{x}_i^{(k-1)}(t)\right)-\mathbf{g}_i^{(k-1)}(t)\right\rangle\right].\end{aligned} \tag{52}$$

Following on from the Young's inequality, we give

$$\begin{aligned}&2\left\langle\nabla F_{t,i}\left(\mathbf{x}_i^{(k)}(t)\right)-\nabla F_{t,i}\left(\mathbf{x}_i^{(k-1)}(t)\right),\nabla F_{t,i}\left(\mathbf{x}_i^{(k-1)}(t)\right)-\mathbf{g}_i^{(k-1)}(t)\right\rangle\\ &\quad\le\rho_k\left\|\nabla F_{t,i}\left(\mathbf{x}_i^{(k-1)}(t)\right)-\mathbf{g}_i^{(k-1)}(t)\right\|^2+\frac{1}{\rho_k}\left\|\nabla F_{t,i}\left(\mathbf{x}_i^{(k)}(t)\right)-\nabla F_{t,i}\left(\mathbf{x}_i^{(k-1)}(t)\right)\right\|^2\\ &\quad\le\rho_k\left\|\nabla F_{t,i}\left(\mathbf{x}_i^{(k-1)}(t)\right)-\mathbf{g}_i^{(k-1)}(t)\right\|^2+\frac{L^2}{\rho_k}\left\|\mathbf{x}_i^{(k)}(t)-\mathbf{x}_i^{(k-1)}(t)\right\|^2,\end{aligned} \tag{53}$$

where in the last inequality we have used the fact that the function $F_{t,i}$ is L-smooth. Plugging Eq. (53) into Eq. (52), we have

$$\mathbb{E}\left[\left\|\nabla F_{t,i}\left(\mathbf{x}_i^{(k)}(t)\right)-\mathbf{g}_i^{(k)}(t)\right\|^2\right]\leq \eta_k^2\sigma^2+(1-\eta_k)^2\left(1+\rho_k^{-1}\right)L^2\mathbb{E}\left[\left\|\mathbf{x}_i^{(k)}(t)-\mathbf{x}_i^{(k-1)}(t)\right\|^2\right] + (1-\eta_k)^2(1+\rho_k)\,\mathbb{E}\left[\left\|\nabla F_{t,i}\left(\mathbf{x}_i^{(k-1)}(t)\right)-\mathbf{g}_i^{(k-1)}(t)\right\|^2\right], \tag{54}$$

where we use Assumption 5 to obtain the above inequality. Summing up both sides of Eq. (54) and setting $\rho_k=\eta_k/2$, we obtain

$$\begin{aligned}\mathbb{E}\left[\sum_{i=1}^N\left\|\nabla F_{t,i}\left(\mathbf{x}_i^{(k)}(t)\right)-\mathbf{g}_i^{(k)}(t)\right\|^2\right]\leq{}& N\eta_k^2\sigma^2\\ &+L^2\left(1+2\eta_k^{-1}\right)\mathbb{E}\left[\sum_{i=1}^N\left\|\mathbf{x}_i^{(k)}(t)-\mathbf{x}_i^{(k-1)}(t)\right\|^2\right]\\ &+(1-\eta_k)\,\mathbb{E}\left[\sum_{i=1}^N\left\|\nabla F_{t,i}\left(\mathbf{x}_i^{(k-1)}(t)\right)-\mathbf{g}_i^{(k-1)}(t)\right\|^2\right].\end{aligned} \tag{55}$$

To bound Eq. (55), we need to estimate the term $\mathbb{E}[\sum_{i=1}^N\|\mathbf{x}_i^{(k)}(t)-\mathbf{x}_i^{(k-1)}(t)\|^2]$. For this purpose, using the Cauchy-Schwarz inequality, we have

$$\begin{aligned}\sum_{i=1}^N\left\|\mathbf{x}_i^{(k)}(t)-\mathbf{x}_i^{(k-1)}(t)\right\|^2\leq{}&\sum_{i=1}^N 3\left[\left\|\mathbf{x}_i^{(k)}(t)-\bar{\mathbf{x}}^{(k)}(t)\right\|^2+\left\|\bar{\mathbf{x}}^{(k-1)}(t)-\bar{\mathbf{x}}^{(k-1)}(t)\right\|^2\right]\\ &+3\sum_{i=1}^N\left\|\bar{\mathbf{x}}^{(k-1)}(t)-\bar{\mathbf{x}}_i^{(k-1)}(t)\right\|^2\\ \leq{}&\frac{3N^3\nu^2D^2}{K^2(1-\beta)^2}+\frac{3ND^2}{K^2}+\frac{3N^3\nu^2D^2}{K^2(1-\beta)^2}\\ ={}&\frac{3ND^2}{K^2}\left(1+\frac{2N^2\nu^2}{(1-\beta)^2}\right),\end{aligned} \tag{56}$$

where the last inequality is due to Lemma 3 and Lemma 4. Plugging Eq. (56) into Eq. (55), we obtain

$$\mathbb{E}\left[\sum_{i=1}^N\left\|\nabla F_{t,i}\left(\mathbf{x}_i^{(k)}(t)\right)-\mathbf{g}_i^{(k)}(t)\right\|^2\right]\leq N\eta_k^2\sigma^2+\left(1+2\eta_k^{-1}\right)\frac{3NL^2D^2}{K^2}\left(1+\frac{2N^2\nu^2}{(1-\beta)^2}\right) + (1-\eta_k)\,\mathbb{E}\left[\sum_{i=1}^N\left\|\nabla F_{t,i}\left(\mathbf{x}_i^{(k-1)}(t)\right)-\mathbf{g}_i^{(k-1)}(t)\right\|^2\right]. \tag{57}$$

Let $\Delta_i^{(k)}(t) := \left\|\nabla F_{t,i}\left(\mathbf{x}_i^{(k)}(t)\right) - \mathbf{g}_i^{(k)}(t)\right\|^2$ and set $\eta_k = 2/K^{2/3}$. Thus, we obtain

$$\mathbb{E}\left[\sum_{i=1}^{N}\Delta_i^{(k)}(t)\right] \leq \left(1-\frac{2}{K^{2/3}}\right)\mathbb{E}\left[\sum_{i=1}^{N}\Delta_i^{(k-1)}(t)\right] + \frac{4N\sigma^2}{K^{4/3}} + \frac{3\kappa NL^2D^2}{K^2} + \frac{3\kappa NL^2D^2}{K^{4/3}}, \tag{58}$$

where $\kappa := \left(1+2N^2\nu^2/(1-\beta)^2\right)$. From Eq. (58), we also have

$$\begin{aligned}
\mathbb{E}\left[\sum_{i=1}^{N}\Delta_i^{(k)}(t)\right] &\leq \left(1-\frac{2}{K^{2/3}}\right)^k \mathbb{E}\left[\sum_{i=1}^{N}\Delta_i^{(0)}(t)\right] \\
&\quad + \left(\frac{4N\sigma^2}{K^{4/3}} + \frac{3\kappa NL^2D^2}{K^2} + \frac{3\kappa NL^2D^2}{K^{4/3}}\right)\sum_{s=0}^{k-1}\left(1-\frac{2}{K^{2/3}}\right)^s \\
&\leq \left(1-\frac{2}{K^{2/3}}\right)^k \mathbb{E}\left[\sum_{i=1}^{N}\Delta_i^{(0)}(t)\right] + \frac{2N\sigma^2}{K^{2/3}} + \frac{3\kappa NL^2D^2}{2K^{4/3}} + \frac{3\kappa NL^2D^2}{2K^{2/3}} \\
&\leq \left(1-\frac{2}{K^{2/3}}\right)^k NG^2 + \frac{2N\sigma^2}{K^{2/3}} + \frac{3\kappa NL^2D^2}{2K^{4/3}} + \frac{3\kappa NL^2D^2}{2K^{2/3}},
\end{aligned} \tag{59}$$

where the second inequality follows from the relation $\sum_{s=0}^{k-1}\left(1-2/K^{2/3}\right)^s \leq \frac{K^{2/3}}{2}$. Therefore, the lemma is obtained. ■

With Lemma 6 in place, we also establish an upper bound of the expected distance between the vector $\bar{\mathbf{d}}^{(k)}(t)$ and the gradient of the objective function ∇F_t, which is stated in the following lemma.

Lemma 7 *Let Assumptions 1-5 hold. The sequences $\{\mathbf{x}_i(t)\}$ and $\{\mathbf{d}_i(t)\}$ are generated by Algorithm 1. For all $k \in \{1,\ldots,K\}$ and $i \in \{1,\ldots,N\}$, we have*

$$\begin{aligned}
\mathbb{E}\left[\left\|\bar{\mathbf{d}}^{(k)}(t) - \frac{1}{N}\sum_{i=1}^{N}\nabla F_{t,i}\left(\bar{\mathbf{x}}^{(k)}(t)\right)\right\|\right] &\leq (1-\gamma_k)^k G + \frac{(1-\gamma_k)LD}{K\gamma_k} + \frac{LND\nu}{K(1-\beta)} \\
&\quad + G\left(1-\frac{2}{K^{2/3}}\right)^{k/2} + \frac{LD\sqrt{3\kappa}}{\sqrt{2}K^{2/3}} + \frac{\sigma\sqrt{2}}{K^{1/3}} + \frac{LD\sqrt{3\kappa}}{\sqrt{2}K^{1/3}},
\end{aligned} \tag{60}$$

where $\kappa := 1+2N^2\nu^2/(1-\beta)^2$.

Proof From the update rule in Eq. (11), we have

$$\begin{aligned}
\sum_{i=1}^{N}\mathbf{d}_i^{(k)}(t) &= (1-\gamma_k)\sum_{i=1}^{N}\sum_{j=1}^{N}a_{ij}(t)\,\mathbf{d}_j^{(k-1)}(t) + \gamma_k\sum_{i=1}^{N}\mathbf{g}_i^{(k)}(t) \\
&= (1-\gamma_k)\sum_{j=1}^{N}\mathbf{d}_j^{(k-1)}(t)\sum_{i=1}^{N}a_{ij}(t) + \gamma_k\sum_{i=1}^{N}\mathbf{g}_i^{(k)}(t) \\
&= (1-\gamma_k)\sum_{j=1}^{N}\mathbf{d}_j^{(k-1)}(t) + \gamma_k\sum_{i=1}^{N}\mathbf{g}_i^{(k)}(t),
\end{aligned} \tag{61}$$

where the last equality follows from the relation $\sum_{i=1}^{N} a_{ij}(t) = 1$. According to the above equality (61), we obtain

$$
\begin{aligned}
&\left\| \sum_{i=1}^{N} \mathbf{d}_i^{(k)}(t) - \sum_{i=1}^{N} \nabla F_{t,i}\left(\bar{\mathbf{x}}^{(k)}(t)\right) \right\| \\
&\quad = \left\| (1-\gamma_k) \sum_{j=1}^{N} \mathbf{d}_j^{(k-1)}(t) + \gamma_k \sum_{i=1}^{N} \mathbf{g}_i^{(k)}(t) - \sum_{i=1}^{N} \nabla F_{t,i}\left(\bar{\mathbf{x}}^{(k)}(t)\right) \right\| \\
&\quad = \left\| (1-\gamma_k) \sum_{j=1}^{N} \mathbf{d}_j^{(k-1)}(t) - (1-\gamma_k) \sum_{i=1}^{N} \nabla F_{t,i}\left(\bar{\mathbf{x}}^{(k-1)}(t)\right) \right. \\
&\quad \left. + (1-\gamma_k) \sum_{i=1}^{N} \nabla F_{t,i}\left(\bar{\mathbf{x}}^{(k-1)}(t)\right) + \gamma_k \sum_{i=1}^{N} \mathbf{g}_i^{(k)}(t) - \sum_{i=1}^{N} \nabla F_{t,i}\left(\bar{\mathbf{x}}^{(k)}(t)\right) \right\| \\
&\quad = \left\| (1-\gamma_k) \left(\sum_{j=1}^{N} \mathbf{d}_j^{(k-1)}(t) - \sum_{i=1}^{N} \nabla F_{t,i}\left(\bar{\mathbf{x}}^{(k-1)}(t)\right) \right) \right. \\
&\quad + (1-\gamma_k) \left(\sum_{i=1}^{N} \nabla F_{t,i}\left(\bar{\mathbf{x}}^{(k-1)}(t)\right) - \sum_{i=1}^{N} \nabla F_{t,i}\left(\bar{\mathbf{x}}^{(k)}(t)\right) \right) \\
&\quad \left. + \gamma_k \left(\sum_{i=1}^{N} \mathbf{g}_i^{(k)}(t) - \sum_{i=1}^{N} \nabla F_{t,i}\left(\bar{\mathbf{x}}^{(k)}(t)\right) \right) \right\| \\
&\quad \leq (1-\gamma_k) \left\| \sum_{j=1}^{N} \mathbf{d}_j^{(k-1)}(t) - \sum_{i=1}^{N} \nabla F_{t,i}\left(\bar{\mathbf{x}}^{(k-1)}(t)\right) \right\| \\
&\quad + (1-\gamma_k) \left\| \sum_{i=1}^{N} \nabla F_{t,i}\left(\bar{\mathbf{x}}^{(k-1)}(t)\right) - \sum_{i=1}^{N} \nabla F_{t,i}\left(\bar{\mathbf{x}}^{(k)}(t)\right) \right\| \\
&\quad + \gamma_k \left\| \sum_{i=1}^{N} \mathbf{g}_i^{(k)}(t) - \sum_{i=1}^{N} \nabla F_{t,i}\left(\bar{\mathbf{x}}^{(k)}(t)\right) \right\| \\
&\quad \leq (1-\gamma_k) \left\| \sum_{j=1}^{N} \mathbf{d}_j^{(k-1)}(t) - \sum_{i=1}^{N} \nabla F_{t,i}\left(\bar{\mathbf{x}}^{(k-1)}(t)\right) \right\| \\
&\quad + (1-\gamma_k) \left\| \sum_{i=1}^{N} \nabla F_{t,i}\left(\bar{\mathbf{x}}^{(k-1)}(t)\right) - \sum_{i=1}^{N} \nabla F_{t,i}\left(\bar{\mathbf{x}}^{(k)}(t)\right) \right\| \\
&\quad + \gamma_k \left\| \sum_{i=1}^{N} \mathbf{g}_i^{(k)}(t) - \sum_{i=1}^{N} \nabla F_{t,i}\left(\mathbf{x}_i^{(k)}(t)\right) \right\| \\
&\quad + \gamma_k \left\| \sum_{i=1}^{N} \nabla F_{t,i}\left(\mathbf{x}_i^{(k)}(t)\right) - \sum_{i=1}^{N} \nabla F_{t,i}\left(\bar{\mathbf{x}}^{(k)}(t)\right) \right\|,
\end{aligned}
\tag{62}
$$

where we have used the triangle inequality in the first inequality and the last inequality. Furthermore, since the functions $F_{t,i}$ are L-smooth, we obtain

$$\begin{aligned}\left\|\sum_{i=1}^{N}\mathbf{d}_i^{(k)}(t)-\sum_{i=1}^{N}\nabla F_{t,i}\left(\bar{\mathbf{x}}^{(k)}(t)\right)\right\| &\leq (1-\gamma_k)\left\|\sum_{j=1}^{N}\mathbf{d}_j^{(k-1)}(t)-\sum_{i=1}^{N}\nabla F_{t,i}\left(\bar{\mathbf{x}}^{(k-1)}(t)\right)\right\| \\ &+ (1-\gamma_k)L\sum_{i=1}^{N}\left\|\bar{\mathbf{x}}^{(k-1)}(t)-\bar{\mathbf{x}}^{(k)}(t)\right\| \\ &+ \gamma_k L\sum_{i=1}^{N}\left\|\mathbf{x}_i^{(k)}(t)-\bar{\mathbf{x}}^{(k)}(t)\right\| \\ &+ \gamma_k\left\|\sum_{i=1}^{N}\mathbf{g}_i^{(k)}(t)-\sum_{i=1}^{N}\nabla F_{t,i}\left(\mathbf{x}_i^{(k)}(t)\right)\right\| \\ &\leq (1-\gamma_k)\left\|\sum_{j=1}^{N}\mathbf{d}_j^{(k-1)}(t)-\sum_{i=1}^{N}\nabla F_{t,i}\left(\bar{\mathbf{x}}^{(k-1)}(t)\right)\right\| \\ &+ \frac{(1-\gamma_k)LND}{K}+\frac{\gamma_k LN^2D\nu}{K(1-\beta)} \\ &+ \gamma_k\left\|\sum_{i=1}^{N}\mathbf{g}_i^{(k)}(t)-\sum_{i=1}^{N}\nabla F_{t,i}\left(\mathbf{x}_i^{(k)}(t)\right)\right\|,\end{aligned} \tag{63}$$

where the first inequality follows from the triangle inequality, the last inequality follows from the Cauchy-Schwarz inequality and lemmata 3 and 4. To bound the expectation of the left term in Eq. (47), i.e., $\mathbb{E}[\|\sum_{i=1}^{N}\mathbf{d}_i^{(k)}(t)-\sum_{i=1}^{N}\nabla F_{t,i}(\bar{\mathbf{x}}^{(k)}(t))\|]$, we need to bound the term $\mathbb{E}[\|\sum_{i=1}^{N}\mathbf{g}_i^{(k)}(t)-\sum_{i=1}^{N}\nabla F_{t,i}(\mathbf{x}_i^{(k)}(t))\|]$. To this end, applying the Cauchy-Schwarz inequality, we first have

$$\begin{aligned}&\mathbb{E}\left[\left\|\sum_{i=1}^{N}\mathbf{g}_i^{(k)}(t)-\sum_{i=1}^{N}\nabla F_{t,i}\left(\mathbf{x}_i^{(k)}(t)\right)\right\|\right] \\ &\qquad\leq\sqrt{N}\mathbb{E}\left[\left(\sum_{i=1}^{N}\left\|\mathbf{g}_i^{(k)}(t)-\nabla F_{t,i}\left(\mathbf{x}_i^{(k)}(t)\right)\right\|^2\right)^{1/2}\right] \\ &\qquad\leq\sqrt{N}\left(\mathbb{E}\left[\sum_{i=1}^{N}\left\|\mathbf{g}_i^{(k)}(t)-\nabla F_{t,i}\left(\mathbf{x}_i^{(k)}(t)\right)\right\|^2\right]\right)^{1/2},\end{aligned} \tag{64}$$

where the last inequality follows from the Jensen's inequality. Furthermore, according to Lemma 6 and Eq. (64), we have

$$\begin{aligned}\mathbb{E}\left[\sum_{i=1}^{N}\left\|\mathbf{g}_i^{(k)}(t)-\nabla F_{t,i}\left(\mathbf{x}_i^{(k)}(t)\right)\right\|^2\right]\\ \leq\sqrt{N}\left(\left(1-\frac{2}{K^{2/3}}\right)^k NG^2+\frac{3\kappa NL^2D^2}{2K^{4/3}}+\frac{4N\sigma^2+3\kappa NL^2D^2}{2K^{2/3}}\right)^{1/2}.\end{aligned}\tag{65}$$

Using the inequality $\sum_{i=1}^{N} r_i^2 \leq (\sum_{i=1}^{N} r_i)^2$, we have

$$\begin{aligned}\mathbb{E}\left[\sum_{i=1}^{N}\left\|\mathbf{g}_i^{(k)}(t)-\nabla F_{t,i}\left(\mathbf{x}_i^{(k)}(t)\right)\right\|^2\right]\leq NG\left(1-\frac{2}{K^{2/3}}\right)^{k/2}+\frac{NLD\sqrt{3\kappa}}{\sqrt{2}K^{2/3}}\\ +\frac{N\sigma\sqrt{2}}{K^{1/3}}+\frac{NLD\sqrt{3\kappa}}{\sqrt{2}K^{1/3}}.\end{aligned}\tag{66}$$

Taking the expectation of both sides in Eq. (63) and using the expression (66), we obtain

$$\begin{aligned}\mathbb{E}\left[\left\|\sum_{i=1}^{N}\mathbf{d}_i^{(k)}(t)-\sum_{i=1}^{N}\nabla F_{t,i}\left(\bar{\mathbf{x}}^{(k)}(t)\right)\right\|\right]\leq\frac{(1-\gamma_k)LND}{K}+\frac{\gamma_k LN^2D\nu}{K(1-\beta)}\\ +(1-\gamma_k)\,\mathbb{E}\left[\left\|\sum_{j=1}^{N}\mathbf{d}_j^{(k-1)}(t)-\sum_{i=1}^{N}\nabla F_{t,i}\left(\bar{\mathbf{x}}^{(k-1)}(t)\right)\right\|\right]\\ +\gamma_k NG\left(1-\frac{2}{K^{2/3}}\right)^{k/2}+\frac{\gamma_k NLD\sqrt{3\kappa}}{\sqrt{2}K^{2/3}}+\frac{\gamma_k N\sigma\sqrt{2}}{K^{1/3}}+\frac{\gamma_k NLD\sqrt{3\kappa}}{\sqrt{2}K^{1/3}}.\end{aligned}\tag{67}$$

Furthermore, multiplying both sides of Eq. (67) by $1/N$ and applying the resulted expression recursively, we have

$$\begin{aligned}\mathbb{E}\left[\left\|\frac{1}{N}\sum_{i=1}^{N}\mathbf{d}_i^{(k)}(t)-\frac{1}{N}\sum_{i=1}^{N}\nabla F_{t,i}\left(\bar{\mathbf{x}}^{(k)}(t)\right)\right\|\right]\leq\left[\frac{(1-\gamma_k)LD}{K}+\frac{\gamma_k LND\nu}{K(1-\beta)}\right]\sum_{s=0}^{k-1}(1-\gamma_k)^s\\ +(1-\gamma_k)^k\left\|\frac{1}{N}\sum_{i=1}^{N}\mathbf{d}_i^{(0)}(t)-\frac{1}{N}\sum_{i=1}^{N}\nabla F_{t,i}\left(\bar{\mathbf{x}}^{(0)}(t)\right)\right\|\\ +\gamma_k\left[G\left(1-\frac{2}{K^{2/3}}\right)^{k/2}+\frac{LD\sqrt{3\kappa}}{\sqrt{2}K^{2/3}}+\frac{\sigma\sqrt{2}}{K^{1/3}}+\frac{LD\sqrt{3\kappa}}{\sqrt{2}K^{1/3}}\right]\sum_{s=0}^{k-1}(1-\gamma_k)^s\\ \leq(1-\gamma_k)^k G+\frac{(1-\gamma_k)LD}{K\gamma_k}+\frac{LND\nu}{K(1-\beta)}\\ +G\left(1-\frac{2}{K^{2/3}}\right)^{k/2}+\frac{LD\sqrt{3\kappa}}{\sqrt{2}K^{2/3}}+\frac{\sigma\sqrt{2}}{K^{1/3}}+\frac{LD\sqrt{3\kappa}}{\sqrt{2}K^{1/3}},\end{aligned}\tag{68}$$

where the last inequality follows from the inequality $\sum_{s=0}^{k-1}(1-\gamma_k)^s \leq 1/\gamma_k$. Therefore, the statement of the lemma is proved completely. ■

With Lemma 7 in place, we now start to prove Theorem 1.

Proof of Theorem 1. Since the functions $F_{t,i}$ are L-smooth, we have

$$\begin{aligned}
\frac{1}{N}\sum_{i=1}^{N} F_{t,i}\left(\bar{\mathbf{x}}^{(k+1)}(t)\right) &\geq \frac{1}{N}\sum_{i=1}^{N} F_{t,i}\left(\bar{\mathbf{x}}^{(k)}(t)\right) - \frac{L}{2}\left\|\bar{\mathbf{x}}^{(k+1)}(t) - \bar{\mathbf{x}}^{(k)}(t)\right\|^2 \\
&\quad + \frac{1}{N}\left\langle \sum_{i=1}^{N} \nabla F_{t,i}\left(\bar{\mathbf{x}}^{(k)}(t)\right), \bar{\mathbf{x}}^{(k+1)}(t) - \bar{\mathbf{x}}^{(k)}(t)\right\rangle \\
&= \frac{1}{N}\sum_{i=1}^{N} F_{t,i}\left(\bar{\mathbf{x}}^{(k)}(t)\right) - \frac{L}{2K^2}\left\|\frac{1}{N}\sum_{i=1}^{N}\mathbf{v}_i^{(k)}(t)\right\|^2 \\
&\quad + \frac{1}{K}\left\langle \frac{1}{N}\sum_{i=1}^{N} \nabla F_{t,i}\left(\bar{\mathbf{x}}^{(k)}(t)\right), \frac{1}{N}\sum_{i=1}^{N}\mathbf{v}_i^{(k)}(t)\right\rangle \\
&\geq \frac{1}{N}\sum_{i=1}^{N} F_{t,i}\left(\bar{\mathbf{x}}^{(k)}(t)\right) - \frac{LD^2}{2K^2} \\
&\quad + \frac{1}{K}\left\langle \frac{1}{N}\sum_{i=1}^{N} \nabla F_{t,i}\left(\bar{\mathbf{x}}^{(k)}(t)\right), \frac{1}{N}\sum_{i=1}^{N}\mathbf{v}_i^{(k)}(t)\right\rangle,
\end{aligned} \tag{69}$$

where the last inequality follows from the relation $\|(1/N)\sum_{i=1}^{N}\mathbf{v}_i^{(k)}(t)\|^2 \leq D^2$ and Eq. (21). Setting $\bar{\mathbf{v}}^{(k)}(t) := (1/N)\sum_{i=1}^{N}\mathbf{v}_i^{(k)}(t)$ and $F_t := (1/N)\sum_{i=1}^{N} F_{t,i}$. By adding and subtracting $\langle \bar{\mathbf{d}}^{(k)}(t), \bar{\mathbf{v}}^{(k)}(t)\rangle$ in the last term of the right side of Eq. (69), we obtain

$$\begin{aligned}
\left\langle \frac{1}{N}\sum_{i=1}^{N} \nabla F_{t,i}\left(\bar{\mathbf{x}}^{(k)}(t)\right), \frac{1}{N}\sum_{i=1}^{N}\mathbf{v}_i^{(k)}(t)\right\rangle &= \left\langle \nabla F_t\left(\bar{\mathbf{x}}^{(k)}(t)\right), \bar{\mathbf{v}}^{(k)}(t)\right\rangle \\
&= \left\langle \nabla F_t\left(\bar{\mathbf{x}}^{(k)}(t)\right) - \bar{\mathbf{d}}^{(k)}(t), \bar{\mathbf{v}}^{(k)}(t)\right\rangle \\
&\quad + \left\langle \bar{\mathbf{d}}^{(k)}(t), \bar{\mathbf{v}}^{(k)}(t)\right\rangle \\
&= \left\langle \nabla F_t\left(\bar{\mathbf{x}}^{(k)}(t)\right) - \bar{\mathbf{d}}^{(k)}(t), \bar{\mathbf{v}}^{(k)}(t)\right\rangle \\
&\quad + \left\langle \bar{\mathbf{d}}^{(k)}(t), \mathbf{x}^*\right\rangle + \left\langle \bar{\mathbf{d}}^{(k)}(t), \bar{\mathbf{v}}^{(k)}(t) - \mathbf{x}^*\right\rangle \\
&= \left\langle \nabla F_t\left(\bar{\mathbf{x}}^{(k)}(t)\right) - \bar{\mathbf{d}}^{(k)}(t), \bar{\mathbf{v}}^{(k)}(t) - \mathbf{x}^*\right\rangle \\
&\quad + \left\langle \nabla F_t\left(\bar{\mathbf{x}}^{(k)}(t)\right), \mathbf{x}^*\right\rangle + \left\langle \bar{\mathbf{d}}^{(k)}(t), \bar{\mathbf{v}}^{(k)}(t) - \mathbf{x}^*\right\rangle,
\end{aligned} \tag{70}$$

where we have added and subtracted the term $\langle \bar{\mathbf{d}}^{(k)}(t), \bar{\mathbf{v}}^{(k)}(t)\rangle$ to obtain the second equality, the third equality is obtained by adding and subtracting the term $\langle \bar{\mathbf{d}}^{(k)}(t), \mathbf{x}^*\rangle$, in the last equality we have added and subtracted the term $\langle \nabla F_t\left(\bar{\mathbf{x}}^{(k)}(t)\right), \mathbf{x}^*\rangle$. Since the submodular functions $F_{t,i}$ are monotonic and concave along non-negative directions, we obtain

$$\begin{aligned} F_{t,i}\left(\mathbf{x}^*\right) - F_{t,i}\left(\bar{\mathbf{x}}^{(k)}\left(t\right)\right) &\leq F_{t,i}\left(\mathbf{x}^* \vee \bar{\mathbf{x}}^{(k)}\left(t\right)\right) - F_{t,i}\left(\bar{\mathbf{x}}^{(k)}\left(t\right)\right) \\ &\leq \left\langle \nabla F_{t,i}\left(\bar{\mathbf{x}}^{(k)}\left(t\right)\right), \mathbf{x}^* \vee \bar{\mathbf{x}}^{(k)}\left(t\right) - \bar{\mathbf{x}}^{(k)}\left(t\right)\right\rangle \\ &= \left\langle \nabla F_{t,i}\left(\bar{\mathbf{x}}^{(k)}\left(t\right)\right), \left(\mathbf{x}^* - \bar{\mathbf{x}}^{(k)}\left(t\right)\right) \vee \mathbf{0}\right\rangle \\ &\leq \left\langle \nabla F_{t,i}\left(\bar{\mathbf{x}}^{(k)}\left(t\right)\right), \mathbf{x}^*\right\rangle. \end{aligned} \tag{71}$$

Following on from the definition of F_t and using the above inequality (71), we also have

$$F_t\left(\mathbf{x}^*\right) - F_t\left(\bar{\mathbf{x}}^{(k)}\left(t\right)\right) \leq \left\langle \nabla F_t\left(\bar{\mathbf{x}}^{(k)}\left(t\right)\right), \mathbf{x}^*\right\rangle. \tag{72}$$

Plugging Eq. (72) into Eq. (70), we obtain

$$\begin{aligned} \left\langle \nabla F_t\left(\bar{\mathbf{x}}^{(k)}\left(t\right)\right), \bar{\mathbf{v}}^{(k)}\left(t\right)\right\rangle \geq{}& \left\langle \nabla F_t\left(\bar{\mathbf{x}}^{(k)}\left(t\right)\right) - \bar{\mathbf{d}}^{(k)}\left(t\right), \bar{\mathbf{v}}^{(k)}\left(t\right) - \mathbf{x}^*\right\rangle \\ &+ \left\langle \bar{\mathbf{d}}^{(k)}\left(t\right), \bar{\mathbf{v}}^{(k)}\left(t\right) - \mathbf{x}^*\right\rangle + \left(F_t\left(\mathbf{x}^*\right) - F_t\left(\bar{\mathbf{x}}^{(k)}\left(t\right)\right)\right). \end{aligned} \tag{73}$$

In addition, by the Cauchy-Schwarz inequality, we have

$$\begin{aligned} \left\langle \nabla F_t\left(\bar{\mathbf{x}}^{(k)}\left(t\right)\right) - \bar{\mathbf{d}}^{(k)}\left(t\right), \bar{\mathbf{v}}^{(k)}\left(t\right) - \mathbf{x}^*\right\rangle &\geq -\left\|\nabla F_t\left(\bar{\mathbf{x}}^{(k)}\left(t\right)\right) - \bar{\mathbf{d}}^{(k)}\left(t\right)\right\| \left\|\bar{\mathbf{v}}^{(k)}\left(t\right) - \mathbf{x}^*\right\| \\ &\geq -D\left\|\nabla F_t\left(\bar{\mathbf{x}}^{(k)}\left(t\right)\right) - \bar{\mathbf{d}}^{(k)}\left(t\right)\right\|, \end{aligned} \tag{74}$$

where the last inequality follows from the fact that $\left(\bar{\mathbf{v}}^{(k)}\left(t\right) - \mathbf{x}^*\right) \in \mathcal{K}$. Plugging Eq. (74) into Eq. (73), we obtain

$$\begin{aligned} \left\langle \nabla F_t\left(\bar{\mathbf{x}}^{(k)}\left(t\right)\right), \bar{\mathbf{v}}^{(k)}\left(t\right)\right\rangle \geq{}& -D\left\|\nabla F_t\left(\bar{\mathbf{x}}^{(k)}\left(t\right)\right) - \bar{\mathbf{d}}^{(k)}\left(t\right)\right\| \\ &+ \left\langle \bar{\mathbf{d}}^{(k)}\left(t\right), \bar{\mathbf{v}}^{(k)}\left(t\right) - \mathbf{x}^*\right\rangle + \left(F_t\left(\mathbf{x}^*\right) - F_t\left(\bar{\mathbf{x}}^{(k)}\left(t\right)\right)\right). \end{aligned} \tag{75}$$

Substituting Eq. (75) into Eq. (69), we have

$$\begin{aligned} \frac{1}{N}\sum_{i=1}^{N} F_{t,i}\left(\bar{\mathbf{x}}^{(k+1)}\left(t\right)\right) \geq{}& \frac{1}{N}\sum_{i=1}^{N} F_{t,i}\left(\bar{\mathbf{x}}^{(k)}\left(t\right)\right) - \frac{LD^2}{2K^2} \\ &- \frac{D}{K}\left\|\nabla F_t\left(\bar{\mathbf{x}}^{(k)}\left(t\right)\right) - \bar{\mathbf{d}}^{(k)}\left(t\right)\right\| \\ &+ \frac{1}{K}\left\langle \bar{\mathbf{d}}^{(k)}\left(t\right), \bar{\mathbf{v}}^{(k)}\left(t\right) - \mathbf{x}^*\right\rangle + \frac{1}{K}\left(F_t\left(\mathbf{x}^*\right) - F_t\left(\bar{\mathbf{x}}^{(k)}\left(t\right)\right)\right). \end{aligned} \tag{76}$$

By adding and subtracting the term $(1/N)\sum_{i=1}^{N} F_{t,i}(\mathbf{x}^*)$ in both sides of Eq. (76) and using some algebraic manipulations, we obtain

$$\begin{aligned}\frac{1}{N}\sum_{i=1}^{N} F_{t,i}(\mathbf{x}^*) - \frac{1}{N}\sum_{i=1}^{N} F_{t,i}\left(\bar{\mathbf{x}}^{(k+1)}(t)\right) &\\ \leq \left(1-\frac{1}{K}\right)\left[\frac{1}{N}\sum_{i=1}^{N} F_{t,i}(\mathbf{x}^*) - \frac{1}{N}\sum_{i=1}^{N} F_{t,i}\left(\bar{\mathbf{x}}^{(k)}(t)\right)\right] &\\ + \frac{D}{K}\left\|\nabla F_t\left(\bar{\mathbf{x}}^{(k)}(t)\right) - \bar{\mathbf{d}}^{(k)}(t)\right\| &\\ - \frac{1}{K}\left\langle \bar{\mathbf{d}}^{(k)}(t), \bar{\mathbf{v}}^{(k)}(t) - \mathbf{x}^*\right\rangle + \frac{LD^2}{2K^2}.&\end{aligned} \tag{77}$$

Therefore, using the relation (77) recursively, we have

$$\begin{aligned}\frac{1}{N}\sum_{i=1}^{N} F_{t,i}(\mathbf{x}^*) - \frac{1}{N}\sum_{i=1}^{N} F_{t,i}\left(\bar{\mathbf{x}}^{(K+1)}(t)\right) &\\ \leq \left(1-\frac{1}{K}\right)^{K}\left[\frac{1}{N}\sum_{i=1}^{N} F_{t,i}(\mathbf{x}^*) - \frac{1}{N}\sum_{i=1}^{N} F_{t,i}\left(\bar{\mathbf{x}}^{(1)}(t)\right)\right] &\\ + \frac{D}{K}\sum_{k=1}^{K}\left\|\nabla F_t\left(\bar{\mathbf{x}}^{(k)}(t)\right) - \bar{\mathbf{d}}^{(k)}(t)\right\| &\\ - \frac{1}{K}\sum_{k=1}^{K}\left\langle \bar{\mathbf{d}}^{(k)}(t), \bar{\mathbf{v}}^{(k)}(t) - \mathbf{x}^*\right\rangle + \frac{LD^2}{2K}.&\end{aligned} \tag{78}$$

Since $\mathbf{x}_i^{(K+1)}(t) = \mathbf{x}_i(t)$ in iteration t, we obtain

$$\begin{aligned}\frac{1}{N}\sum_{i=1}^{N} F_{t,i}(\mathbf{x}^*) - \frac{1}{N}\sum_{i=1}^{N} F_{t,i}(\bar{\mathbf{x}}(t)) &\leq \frac{1}{e}\left[\frac{1}{N}\sum_{i=1}^{N} F_{t,i}(\mathbf{x}^*) - \frac{1}{N}\sum_{i=1}^{N} F_{t,i}\left(\bar{\mathbf{x}}^{(1)}(t)\right)\right] &\\ &+ \frac{D}{K}\sum_{k=1}^{K}\left\|\nabla F_t\left(\bar{\mathbf{x}}^{(k)}(t)\right) - \bar{\mathbf{d}}^{(k)}(t)\right\| &\\ &- \frac{1}{K}\sum_{k=1}^{K}\left\langle \bar{\mathbf{d}}^{(k)}(t), \bar{\mathbf{v}}^{(k)}(t) - \mathbf{x}^*\right\rangle + \frac{LD^2}{2K} &\\ &\leq \frac{1}{e}\left(\frac{1}{N}\sum_{i=1}^{N} F_{t,i}(\mathbf{x}^*) - \frac{1}{N}\sum_{i=1}^{N} F_{t,i}(\mathbf{0})\right) &\\ &+ \frac{D}{K}\sum_{k=1}^{K}\left\|\nabla F_t\left(\bar{\mathbf{x}}^{(k)}(t)\right) - \bar{\mathbf{d}}^{(k)}(t)\right\| &\\ &- \frac{1}{K}\sum_{k=1}^{K}\left\langle \bar{\mathbf{d}}^{(k)}(t), \bar{\mathbf{v}}^{(k)}(t) - \mathbf{x}^*\right\rangle + \frac{LD^2}{2K}.&\end{aligned} \tag{79}$$

Furthermore, according to the fact that $F_{t,i}(\mathbf{0}) \geq 0$ for all $i \in \mathcal{V}$ and $t \in \{1, \ldots, T\}$, we have

$$\begin{aligned}\left(1-\frac{1}{e}\right)\frac{1}{N}\sum_{i=1}^{N}F_{t,i}\left(\mathbf{x}^{*}\right)-\frac{1}{N}\sum_{i=1}^{N}F_{t,i}\left(\bar{\mathbf{x}}\left(t\right)\right) \leq \frac{1}{K}\sum_{k=1}^{K}\left\langle \bar{\mathbf{d}}^{(k)}\left(t\right), \mathbf{x}^{*}-\bar{\mathbf{v}}^{(k)}\left(t\right)\right\rangle + \frac{LD^{2}}{2K} \\ + \frac{D}{K}\sum_{k=1}^{K}\left\|\nabla F_{t}\left(\bar{\mathbf{x}}^{(k)}\left(t\right)\right)-\bar{\mathbf{d}}^{(k)}\left(t\right)\right\|.\end{aligned} \tag{80}$$

Using the following relation

$$\begin{aligned}\left\langle \sum_{i=1}^{N}\mathbf{d}_{i}^{(k)}(t), \sum_{i=1}^{N}\mathbf{v}_{i}^{(k)}(t)\right\rangle &= \sum_{i=1}^{N}\sum_{j=1}^{N}\left\langle \mathbf{d}_{i}^{(k)}(t), \mathbf{v}_{j}^{(k)}(t)\right\rangle \\ &= \sum_{j=1}^{N}\left\langle \sum_{i=1}^{N}\mathbf{d}_{i}^{(k)}(t), \mathbf{v}_{j}^{(k)}(t)\right\rangle,\end{aligned}$$

the term $\langle \bar{\mathbf{d}}^{(k)}\left(t\right), \mathbf{x}^{*}-\bar{\mathbf{v}}^{(k)}\left(t\right)\rangle$ can be rewritten as

$$\begin{aligned}\left\langle \bar{\mathbf{d}}^{(k)}\left(t\right), \mathbf{x}^{*}-\bar{\mathbf{v}}^{(k)}\left(t\right)\right\rangle &= \left\langle \frac{1}{N}\sum_{i=1}^{N}\mathbf{d}_{i}^{(k)}\left(t\right), \mathbf{x}^{*}-\frac{1}{N}\sum_{i=1}^{N}\mathbf{v}_{i}^{(k)}\left(t\right)\right\rangle \\ &= \frac{1}{N^{2}}\sum_{j=1}^{N}\left\langle \sum_{i=1}^{N}\mathbf{d}_{i}^{(k)}\left(t\right), \mathbf{x}^{*}-\mathbf{v}_{j}^{(k)}\left(t\right)\right\rangle \\ &= \frac{1}{N}\sum_{j=1}^{N}\left\langle \left(\frac{1}{N}\sum_{i=1}^{N}\mathbf{d}_{i}^{(k)}\left(t\right)-\mathbf{d}_{j}^{(k)}\left(t\right)\right), \mathbf{x}^{*}-\mathbf{v}_{j}^{(k)}\left(t\right)\right\rangle \\ &+ \frac{1}{N}\sum_{j=1}^{N}\left\langle \mathbf{d}_{j}^{(k)}\left(t\right), \mathbf{x}^{*}-\mathbf{v}_{j}^{(k)}\left(t\right)\right\rangle \\ &\leq \frac{1}{N}\sum_{j=1}^{N}\left\|\frac{1}{N}\sum_{i=1}^{N}\mathbf{d}_{i}^{(k)}\left(t\right)-\mathbf{d}_{j}^{(k)}\left(t\right)\right\|\left\|\mathbf{x}^{*}-\mathbf{v}_{j}^{(k)}\left(t\right)\right\| \\ &+ \frac{1}{N}\sum_{j=1}^{N}\left\langle \mathbf{d}_{j}^{(k)}\left(t\right), \mathbf{x}^{*}-\mathbf{v}_{j}^{(k)}\left(t\right)\right\rangle \\ &\leq \frac{D}{N}\sum_{j=1}^{N}\left\|\frac{1}{N}\sum_{i=1}^{N}\mathbf{d}_{i}^{(k)}\left(t\right)-\mathbf{d}_{j}^{(k)}\left(t\right)\right\| \\ &+ \frac{1}{N}\sum_{j=1}^{N}\left\langle \mathbf{d}_{j}^{(k)}\left(t\right), \mathbf{x}^{*}-\mathbf{v}_{j}^{(k)}\left(t\right)\right\rangle\end{aligned} \tag{81}$$

where the first inequality follows from the Cauchy-Schwarz inequality, the last inequality follows from the fact that $(\mathbf{x}^{*}-\mathbf{v}_{j}^{(k)}\left(t\right)) \in \mathcal{K}$ for all $j \in \mathcal{V}$. Thus, plugging Eq. (81) into

Eq. (80), we have

$$\left(1-\frac{1}{e}\right)\frac{1}{N}\sum_{i=1}^{N}F_{t,i}\left(\mathbf{x}^*\right)-\frac{1}{N}\sum_{i=1}^{N}F_{t,i}\left(\bar{\mathbf{x}}\left(t\right)\right)\leq\frac{1}{KN}\sum_{k=1}^{K}\sum_{j=1}^{N}\left\langle\mathbf{d}_j^{(k)}\left(t\right),\mathbf{x}^*-\mathbf{v}_j^{(k)}\left(t\right)\right\rangle+\frac{LD^2}{2K}$$
$$+\frac{D}{K}\sum_{k=1}^{K}\left\|\nabla F_t\left(\bar{\mathbf{x}}^{(k)}\left(t\right)\right)-\bar{\mathbf{d}}^{(k)}\left(t\right)\right\|$$
$$+\frac{1}{K}\sum_{k=1}^{K}\frac{D}{N}\sum_{j=1}^{N}\left\|\frac{1}{N}\sum_{i=1}^{N}\mathbf{d}_i^{(k)}\left(t\right)-\mathbf{d}_j^{(k)}\left(t\right)\right\|. \tag{82}$$

Summing up Eq. (82) over t from 1 to T, we obtain

$$\left(1-\frac{1}{e}\right)\frac{1}{N}\sum_{t=1}^{T}\sum_{i=1}^{N}F_{t,i}\left(\mathbf{x}^*\right)-\frac{1}{N}\sum_{t=1}^{T}\sum_{i=1}^{N}F_{t,i}\left(\bar{\mathbf{x}}\left(t\right)\right)\leq\frac{TLD^2}{2K}$$
$$+\frac{1}{KN}\sum_{k=1}^{K}\sum_{j=1}^{N}\sum_{t=1}^{T}\left\langle\mathbf{d}_j^{(k)}\left(t\right),\mathbf{x}^*-\mathbf{v}_j^{(k)}\left(t\right)\right\rangle$$
$$+\frac{D}{K}\sum_{k=1}^{K}\sum_{t=1}^{T}\left\|\nabla F_t\left(\bar{\mathbf{x}}^{(k)}\left(t\right)\right)-\bar{\mathbf{d}}^{(k)}\left(t\right)\right\| \tag{83}$$
$$+\frac{D}{KN}\sum_{j=1}^{N}\sum_{k=1}^{K}\sum_{t=1}^{T}\left\|\frac{1}{N}\sum_{i=1}^{N}\mathbf{d}_i^{(k)}\left(t\right)-\mathbf{d}_j^{(k)}\left(t\right)\right\|.$$

According to Algorithm 1 and the definition of the regret, we can see that

$$\sum_{t=1}^{T}\left\langle\mathbf{d}_j^{(k)}\left(t\right),\mathbf{x}^*-\mathbf{v}_j^{(k)}\left(t\right)\right\rangle\leq\mathcal{R}_T^{\epsilon} \tag{84}$$

for all $j\in\{1,\ldots,N\}$. Plugging Eq. (84) into Eq. (83), we have

$$\left(1-\frac{1}{e}\right)\frac{1}{N}\sum_{t=1}^{T}\sum_{i=1}^{N}F_{t,i}\left(\mathbf{x}^*\right)-\frac{1}{N}\sum_{t=1}^{T}\sum_{i=1}^{N}F_{t,i}\left(\bar{\mathbf{x}}\left(t\right)\right)\leq\frac{TLD^2}{2K}+\mathcal{R}_T^{\epsilon}$$
$$+\frac{D}{K}\sum_{k=1}^{K}\sum_{t=1}^{T}\left\|\nabla F_t\left(\bar{\mathbf{x}}^{(k)}\left(t\right)\right)-\bar{\mathbf{d}}^{(k)}\left(t\right)\right\| \tag{85}$$
$$+\frac{D}{KN}\sum_{j=1}^{N}\sum_{k=1}^{K}\sum_{t=1}^{T}\left\|\frac{1}{N}\sum_{i=1}^{N}\mathbf{d}_i^{(k)}\left(t\right)-\mathbf{d}_j^{(k)}\left(t\right)\right\|.$$

Since the functions $F_{t,i}$ are G-Lipschitz for all $t\in\{1,\ldots,T\}$ and $i\in\{1,\ldots,N\}$, we know that

$$\left|\frac{1}{N}\sum_{i=1}^{N}F_{t,i}\left(\bar{\mathbf{x}}\left(t\right)\right)-\frac{1}{N}\sum_{i=1}^{N}F_{t,i}\left(\mathbf{x}_j\left(t\right)\right)\right|\leq\frac{G}{N}\sum_{i=1}^{N}\left\|\bar{\mathbf{x}}\left(t\right)-\mathbf{x}_j\left(t\right)\right\|\leq\frac{GND\nu}{K\left(1-\beta\right)}, \tag{86}$$

where in the last inequality we have used Eq. (24) in Lemma 4 and the Cauchy-Schwarz inequality. Plugging Eq. (86) into Eq. (85), we obtain

$$\begin{aligned}\left(1-\frac{1}{e}\right)\frac{1}{N}\sum_{t=1}^{T}\sum_{i=1}^{N}F_{t,i}\left(\mathbf{x}^{*}\right)-\frac{1}{N}\sum_{t=1}^{T}\sum_{i=1}^{N}F_{t,i}\left(\mathbf{x}_{j}\left(t\right)\right)\leq\frac{TLD^{2}}{2K}+\mathcal{R}_{T}^{\epsilon}+\frac{TGND\nu}{K\left(1-\beta\right)}\\+\frac{D}{K}\sum_{k=1}^{K}\sum_{t=1}^{T}\left\|\nabla F_{t}\left(\bar{\mathbf{x}}^{(k)}\left(t\right)\right)-\bar{\mathbf{d}}^{(k)}\left(t\right)\right\|\\+\frac{D}{KN}\sum_{j=1}^{N}\sum_{k=1}^{K}\sum_{t=1}^{T}\left\|\frac{1}{N}\sum_{i=1}^{N}\mathbf{d}_{i}^{(k)}\left(t\right)-\mathbf{d}_{j}^{(k)}\left(t\right)\right\|.\end{aligned}\tag{87}$$

Taking expectation in Eq. (87), we have

$$\begin{aligned}\left(1-\frac{1}{e}\right)\frac{1}{N}\sum_{t=1}^{T}\sum_{i=1}^{N}\mathbb{E}\left[F_{t,i}\left(\mathbf{x}^{*}\right)\right]-\frac{1}{N}\sum_{t=1}^{T}\sum_{i=1}^{N}\mathbb{E}\left[F_{t,i}\left(\mathbf{x}_{j}\left(t\right)\right)\right]\leq\frac{TLD^{2}}{2K}+\mathcal{R}_{T}^{\epsilon}+\frac{TGND\nu}{K\left(1-\beta\right)}\\+\frac{D}{K}\sum_{k=1}^{K}\sum_{t=1}^{T}\mathbb{E}\left[\left\|\nabla F_{t}\left(\bar{\mathbf{x}}^{(k)}\left(t\right)\right)-\bar{\mathbf{d}}^{(k)}\left(t\right)\right\|\right]\\+\frac{D}{KN}\sum_{j=1}^{N}\sum_{k=1}^{K}\sum_{t=1}^{T}\mathbb{E}\left[\left\|\frac{1}{N}\sum_{i=1}^{N}\mathbf{d}_{i}^{(k)}\left(t\right)-\mathbf{d}_{j}^{(k)}\left(t\right)\right\|\right].\end{aligned}\tag{88}$$

Furthermore, we also have

$$\begin{aligned}\frac{1}{N}\sum_{j=1}^{N}\mathbb{E}\left[\left\|\frac{1}{N}\sum_{i=1}^{N}\mathbf{d}_{i}^{(k)}\left(t\right)-\mathbf{d}_{j}^{(k)}\left(t\right)\right\|\right]&\leq\frac{1}{\sqrt{N}}\sqrt{\mathbb{E}\left[\left\|\frac{1}{N}\sum_{i=1}^{N}\mathbf{d}_{i}^{(k)}\left(t\right)-\mathbf{d}_{j}^{(k)}\left(t\right)\right\|^{2}\right]}\\&\leq\frac{\gamma_{k}N\nu\sqrt{2\left(\sigma^{2}+G^{2}\right)}}{1-\beta\left(1-\gamma_{k}\right)},\end{aligned}\tag{89}$$

where in the first inequality we have used the Cauchy-Schwarz inequality and the last inequality follows from Eq. (34) in Lemma 5. Moreover, plugging Eq. (60) in Lemma 7 and Eq. (89) into Eq. (88), we obtain

$$\begin{aligned}\left(1-\frac{1}{e}\right)\frac{1}{N}\sum_{t=1}^{T}\sum_{i=1}^{N}\mathbb{E}\left[F_{t,i}\left(\mathbf{x}^{*}\right)\right]-\frac{1}{N}\sum_{t=1}^{T}\sum_{i=1}^{N}\mathbb{E}\left[F_{t,i}\left(\mathbf{x}_{j}\left(t\right)\right)\right]\leq\frac{TLD^{2}}{2K}+\mathcal{R}_{T}^{\epsilon}+\frac{TGND\nu}{K\left(1-\beta\right)}\\+\frac{D}{K}\sum_{k=1}^{K}\sum_{t=1}^{T}\left[\left(1-\gamma_{k}\right)^{k}G+\frac{\left(1-\gamma_{k}\right)LD}{K\gamma_{k}}+\frac{LND\nu}{K\left(1-\beta\right)}+G\left(1-\frac{2}{K^{2/3}}\right)^{k/2}\right.\\\left.+\frac{LD\sqrt{3\kappa}}{\sqrt{2}K^{2/3}}+\frac{\sigma\sqrt{2}}{K^{1/3}}+\frac{LD\sqrt{3\kappa}}{\sqrt{2}K^{1/3}}\right]+\frac{D}{K}\sum_{k=1}^{K}\sum_{t=1}^{T}\frac{\gamma_{k}N\nu\sqrt{2\left(\sigma^{2}+G^{2}\right)}}{1-\beta\left(1-\gamma_{k}\right)}.\end{aligned}\tag{90}$$

Setting $\gamma_k = 1/\sqrt{K}$ for all $k = 1, \ldots, K$ and using Eq. (90), we have

$$\begin{aligned}\left(1-\frac{1}{e}\right)\frac{1}{N}\sum_{t=1}^{T}\sum_{i=1}^{N}\mathbb{E}\left[F_{t,i}\left(\mathbf{x}^*\right)\right] - \frac{1}{N}\sum_{t=1}^{T}\sum_{i=1}^{N}\mathbb{E}\left[F_{t,i}\left(\mathbf{x}_j\left(t\right)\right)\right] &\leq \frac{TLD^2}{2K} + \mathcal{R}_T^{\epsilon} + \frac{TGND\nu}{K\left(1-\beta\right)} \\ &+ \frac{TD}{K}\sum_{k=1}^{K}\left[\left(1-\frac{1}{K^{1/2}}\right)^k G + \frac{LD}{K^{1/2}} + \frac{LND\nu}{K\left(1-\beta\right)} + G\left(1-\frac{2}{K^{2/3}}\right)^{k/2}\right. \\ &\left. + \frac{LD\sqrt{3\kappa}}{\sqrt{2}K^{2/3}} + \frac{\sigma\sqrt{2}}{K^{1/3}} + \frac{LD\sqrt{3\kappa}}{\sqrt{2}K^{1/3}}\right] + \frac{TND\nu}{K}\sum_{k=1}^{K}\frac{\sqrt{2\left(\sigma^2+G^2\right)}}{K^{1/2}\left(1-\beta\right)} \\ &\leq \frac{TLD^2}{2K} + \frac{GNDT\nu}{K\left(1-\beta\right)} + \frac{GDT}{K^{1/2}} + \frac{LD^2T}{K^{1/2}} + \frac{LND^2T\nu}{K\left(1-\beta\right)} + \frac{GDT}{K^{1/3}} \\ &+ \frac{LD^2T\sqrt{3\kappa}}{\sqrt{2}K^{2/3}} + \frac{\sigma DT\sqrt{2}}{K^{1/3}} + \frac{LD^2T\sqrt{3\kappa}}{\sqrt{2}K^{1/3}} + \frac{NDT\nu\sqrt{2\left(\sigma^2+G^2\right)}}{K^{1/2}\left(1-\beta\right)} + \mathcal{R}_T^{\epsilon},\end{aligned} \tag{91}$$

where the last inequality follows from the inequalities

$$\sum_{k=1}^{K}(1-1/K^{1/2}) \leq \sum_{k=0}^{\infty}(1-1/K^{1/2}) = K^{1/2}$$

and

$$\sum_{k=1}^{K}(1-2/K^{2/3})^{k/2} \leq \sum_{k=0}^{\infty}(1-2/K^{2/3})^{k/2} = \frac{1}{1-(1-2/K^{2/3})^{1/2}} \leq K^{2/3}.$$

In addition, since $\kappa = \left(1+2N^2\nu^2/\left(1-\beta\right)^2\right)$, we have $\sqrt{\kappa} \leq 1 + \sqrt{2}N\nu/(1-\beta)$. Therefore, combining the result and the expression (91), the statement of Theorem 1 is proved completely. ■

6.2 Stochastic Online Setting

In this subsection, we analyze the performance of Algorithm 2. Furthermore, we also provide the detailed proof of Theorem 2.

Proof of Theorem 2. According to the smoothness of the functions $F_{t,i}$ for all $i \in \{1, \ldots, N\}$ and $t \in \{1, \ldots, T\}$, and using the expression (15), we have

$$\begin{aligned}\frac{1}{N}\sum_{i=1}^{N}F_i\left(\bar{\mathbf{x}}\left(t+1\right)\right) &\geq \frac{1}{N}\sum_{i=1}^{N}F_i\left(\bar{\mathbf{x}}\left(t\right)\right) - \frac{L}{2T^2}\left\|\frac{1}{N}\sum_{i=1}^{N}\mathbf{v}_i\left(t\right)\right\|^2 \\ &+ \frac{1}{T}\left\langle\frac{1}{N}\sum_{i=1}^{N}\nabla F_i\left(\bar{\mathbf{x}}\left(t\right)\right), \frac{1}{N}\sum_{i=1}^{N}\mathbf{v}_i\left(t\right)\right\rangle \\ &\geq \frac{1}{N}\sum_{i=1}^{N}F_i\left(\bar{\mathbf{x}}\left(t\right)\right) - \frac{LD^2}{2T^2} \\ &+ \frac{1}{T}\left\langle\frac{1}{N}\sum_{i=1}^{N}\nabla F_i\left(\bar{\mathbf{x}}\left(t\right)\right), \frac{1}{N}\sum_{i=1}^{N}\mathbf{v}_i\left(t\right)\right\rangle,\end{aligned} \tag{92}$$

where we use the relation $\|(1/N)\sum_{i=1}^{N}\mathbf{v}_i(t)\| \leq D^2$ to obtain the last inequality. To bound the relation (92), we first establish the bound of the term $\langle \frac{1}{N}\sum_{i=1}^{N}\nabla F_i\left(\bar{\mathbf{x}}\left(t\right)\right), \frac{1}{N}\sum_{i=1}^{N}\mathbf{v}_i\left(t\right)\rangle$. Thus, adding and subtracting the term $\langle (1/N)\sum_{i=1}^{N}\mathbf{d}_i\left(t\right), (1/N)\sum_{i=1}^{N}\mathbf{v}_i\left(t\right)\rangle$, we obtain

$$\begin{aligned}
\left\langle \frac{1}{N}\sum_{i=1}^{N}\nabla F_i\left(\bar{\mathbf{x}}\left(t\right)\right), \frac{1}{N}\sum_{i=1}^{N}\mathbf{v}_i\left(t\right)\right\rangle &= \left\langle \frac{1}{N}\sum_{i=1}^{N}\mathbf{d}_i\left(t\right), \frac{1}{N}\sum_{i=1}^{N}\mathbf{v}_i\left(t\right)\right\rangle \\
&+ \left\langle \frac{1}{N}\sum_{i=1}^{N}\nabla F_i\left(\bar{\mathbf{x}}\left(t\right)\right) - \frac{1}{N}\sum_{i=1}^{N}\mathbf{d}_i\left(t\right), \frac{1}{N}\sum_{i=1}^{N}\mathbf{v}_i\left(t\right)\right\rangle \\
&= \frac{1}{N^2}\sum_{j=1}^{N}\left\langle \sum_{i=1}^{N}\mathbf{d}_i\left(t\right), \mathbf{v}_j\left(t\right)\right\rangle + \left\langle \frac{1}{N}\sum_{i=1}^{N}\nabla F_i\left(\bar{\mathbf{x}}\left(t\right)\right) - \frac{1}{N}\sum_{i=1}^{N}\mathbf{d}_i\left(t\right), \frac{1}{N}\sum_{i=1}^{N}\mathbf{v}_i\left(t\right)\right\rangle \\
&= \frac{1}{N}\sum_{j=1}^{N}\left\langle \left(\frac{1}{N}\mathbf{d}_i\left(t\right) - \mathbf{d}_j\left(t\right)\right), \mathbf{v}_j\left(t\right)\right\rangle + \frac{1}{N}\sum_{j=1}^{N}\left\langle \mathbf{d}_j\left(t\right), \mathbf{v}_j\left(t\right)\right\rangle \\
&+ \left\langle \frac{1}{N}\sum_{i=1}^{N}\nabla F_i\left(\bar{\mathbf{x}}\left(t\right)\right) - \frac{1}{N}\sum_{i=1}^{N}\mathbf{d}_i\left(t\right), \frac{1}{N}\sum_{i=1}^{N}\mathbf{v}_i\left(t\right)\right\rangle,
\end{aligned} \tag{93}$$

where the second equality is obtained by using the following equality,

$$\left\langle \sum_{i=1}^{N}\mathbf{d}_i\left(t\right), \sum_{i=1}^{N}\mathbf{v}_i\left(t\right)\right\rangle = \sum_{i=1}^{N}\sum_{j=1}^{N}\left\langle \mathbf{d}_i\left(t\right), \mathbf{v}_j\left(t\right)\right\rangle = \sum_{j=1}^{N}\left\langle \sum_{i=1}^{N}\mathbf{d}_i\left(t\right), \mathbf{v}_j\left(t\right)\right\rangle,$$

in the last equality we have added and subtracted the term $(1/N)\sum_{j=1}^{N}\langle \mathbf{d}_j\left(t\right), \mathbf{v}_j\left(t\right)\rangle$. Since $\mathbf{v}_i(t) = \arg\max_{\mathbf{v}\in\mathcal{K}}\langle \mathbf{d}_i(t), \mathbf{v}\rangle$, we obtain $\langle \mathbf{d}_i(t), \mathbf{v}_i(t)\rangle \geq \langle \mathbf{d}_i(t), \mathbf{x}^*\rangle$ for all $i \in \{1, \ldots, N\}$. Therefore, plugging the result into Eq. (93), we have

$$\begin{aligned}
\left\langle \frac{1}{N}\sum_{i=1}^{N}\nabla F_i\left(\bar{\mathbf{x}}\left(t\right)\right), \frac{1}{N}\sum_{i=1}^{N}\mathbf{v}_i\left(t\right)\right\rangle &\geq \frac{1}{N}\sum_{j=1}^{N}\left\langle \left(\frac{1}{N}\mathbf{d}_i\left(t\right) - \mathbf{d}_j\left(t\right)\right), \mathbf{v}_j\left(t\right)\right\rangle \\
&+ \frac{1}{N}\sum_{j=1}^{N}\left\langle \mathbf{d}_j\left(t\right), \mathbf{x}^*\right\rangle + \left\langle \frac{1}{N}\sum_{i=1}^{N}\nabla F_i\left(\bar{\mathbf{x}}\left(t\right)\right) - \frac{1}{N}\sum_{i=1}^{N}\mathbf{d}_i\left(t\right), \frac{1}{N}\sum_{i=1}^{N}\mathbf{v}_i\left(t\right)\right\rangle \\
&= \frac{1}{N}\sum_{j=1}^{N}\left\langle \left(\frac{1}{N}\mathbf{d}_i\left(t\right) - \mathbf{d}_j\left(t\right)\right), \mathbf{v}_j\left(t\right)\right\rangle + \frac{1}{N}\sum_{j=1}^{N}\left\langle \mathbf{d}_j\left(t\right) - \frac{1}{N}\sum_{i=1}^{N}\mathbf{d}_i\left(t\right), \mathbf{x}^*\right\rangle \\
&+ \frac{1}{N^2}\sum_{j=1}^{N}\left\langle \sum_{i=1}^{N}\mathbf{d}_i\left(t\right), \mathbf{x}^*\right\rangle + \left\langle \frac{1}{N}\sum_{i=1}^{N}\nabla F_i\left(\bar{\mathbf{x}}\left(t\right)\right) - \frac{1}{N}\sum_{i=1}^{N}\mathbf{d}_i\left(t\right), \frac{1}{N}\sum_{i=1}^{N}\mathbf{v}_i\left(t\right)\right\rangle,
\end{aligned} \tag{94}$$

where we add and subtract the term $(1/N^2)\sum_{j=1}^{N}\langle\sum_{i=1}^{N}\mathbf{d}_i(t),\mathbf{x}^*\rangle$ in the last equality. Furthermore, adding and subtracting the term $1/N^2\sum_{j=1}^{N}\langle\sum_{i=1}^{N}\nabla F_i(\bar{\mathbf{x}}(t)),\mathbf{x}^*\rangle$, we have

$$\begin{aligned}
\left\langle \frac{1}{N}\sum_{i=1}^{N}\nabla F_i(\bar{\mathbf{x}}(t)), \frac{1}{N}\sum_{i=1}^{N}\mathbf{v}_i(t)\right\rangle &\geq \frac{1}{N}\sum_{j=1}^{N}\left\langle\left(\frac{1}{N}\mathbf{d}_i(t)-\mathbf{d}_j(t)\right),\mathbf{v}_j(t)\right\rangle \\
&+\frac{1}{N}\sum_{j=1}^{N}\left\langle\mathbf{d}_j(t)-\frac{1}{N}\sum_{i=1}^{N}\mathbf{d}_i(t),\mathbf{x}^*\right\rangle+\frac{1}{N^2}\sum_{j=1}^{N}\left\langle\sum_{i=1}^{N}\nabla F_i(\bar{\mathbf{x}}(t)),\mathbf{x}^*\right\rangle \\
&+\frac{1}{N^2}\sum_{j=1}^{N}\left\langle\sum_{i=1}^{N}\mathbf{d}_i(t)-\sum_{i=1}^{N}\nabla F_i(\bar{\mathbf{x}}(t)),\mathbf{x}^*\right\rangle \\
&+\left\langle\frac{1}{N}\sum_{i=1}^{N}\nabla F_i(\bar{\mathbf{x}}(t))-\frac{1}{N}\sum_{i=1}^{N}\mathbf{d}_i(t),\frac{1}{N}\sum_{i=1}^{N}\mathbf{v}_i(t)\right\rangle \\
&=\frac{1}{N}\sum_{j=1}^{N}\left\langle\left(\frac{1}{N}\mathbf{d}_i(t)-\mathbf{d}_j(t)\right),\mathbf{v}_j(t)-\mathbf{x}^*\right\rangle+\frac{1}{N}\left\langle\sum_{i=1}^{N}\nabla F_i(\bar{\mathbf{x}}(t)),\mathbf{x}^*\right\rangle \\
&+\frac{1}{N}\left\langle\sum_{i=1}^{N}\mathbf{d}_i(t)-\sum_{i=1}^{N}\nabla F_i(\bar{\mathbf{x}}(t)),\mathbf{x}^*-\frac{1}{N}\sum_{i=1}^{N}\mathbf{v}_i(t)\right\rangle.
\end{aligned} \tag{95}$$

Plugging Eq. (95) into Eq. (92), we have

$$\begin{aligned}
\frac{1}{N}\sum_{i=1}^{N}F_i(\bar{\mathbf{x}}(t+1)) &\geq \frac{1}{N}\sum_{i=1}^{N}F_i(\bar{\mathbf{x}}(t))+\frac{1}{NT}\left\langle\sum_{i=1}^{N}\nabla F_i(\bar{\mathbf{x}}(t)),\mathbf{x}^*\right\rangle-\frac{LD^2}{2T^2} \\
&+\frac{1}{NT}\sum_{j=1}^{N}\left\langle\left(\frac{1}{N}\mathbf{d}_i(t)-\mathbf{d}_j(t)\right),\mathbf{v}_j(t)-\mathbf{x}^*\right\rangle \\
&+\frac{1}{NT}\left\langle\sum_{i=1}^{N}\mathbf{d}_i(t)-\sum_{i=1}^{N}\nabla F_i(\bar{\mathbf{x}}(t)),\mathbf{x}^*-\frac{1}{N}\sum_{i=1}^{N}\mathbf{v}_i(t)\right\rangle.
\end{aligned} \tag{96}$$

Similar to the relation (71), we obtain

$$\left\langle\frac{1}{N}\sum_{i=1}^{N}\nabla F_i(\bar{\mathbf{x}}(t)),\mathbf{x}^*\right\rangle\geq\frac{1}{N}\sum_{i=1}^{N}F_i(\mathbf{x}^*)-\frac{1}{N}\sum_{i=1}^{N}F_i(\bar{\mathbf{x}}(t)). \tag{97}$$

Plugging Eq. (97) into Eq. (96), we have

$$\begin{aligned}
\frac{1}{N}\sum_{i=1}^{N}F_i(\bar{\mathbf{x}}(t+1)) &\geq \frac{1}{N}\sum_{i=1}^{N}F_i(\bar{\mathbf{x}}(t))+\frac{1}{T}\left(\frac{1}{N}\sum_{i=1}^{N}F_i(\mathbf{x}^*)-\frac{1}{N}\sum_{i=1}^{N}F_i(\bar{\mathbf{x}}(t))\right) \\
&+\frac{1}{NT}\sum_{j=1}^{N}\left\langle\left(\frac{1}{N}\mathbf{d}_i(t)-\mathbf{d}_j(t)\right),\mathbf{v}_j(t)-\mathbf{x}^*\right\rangle-\frac{LD^2}{2T^2} \\
&+\frac{1}{NT}\left\langle\sum_{i=1}^{N}\mathbf{d}_i(t)-\sum_{i=1}^{N}\nabla F_i(\bar{\mathbf{x}}(t)),\mathbf{x}^*-\frac{1}{N}\sum_{i=1}^{N}\mathbf{v}_i(t)\right\rangle.
\end{aligned} \tag{98}$$

Furthermore, by using the Cauchy-Schwarz inequality, we give

$$\begin{aligned}\frac{1}{N}\sum_{i=1}^{N}F_i\left(\bar{\mathbf{x}}\left(t+1\right)\right) \geq \frac{1}{N}\sum_{i=1}^{N}F_i\left(\bar{\mathbf{x}}\left(t\right)\right) + \frac{1}{T}\left(\frac{1}{N}\sum_{i=1}^{N}F_i\left(\mathbf{x}^*\right) - \frac{1}{N}\sum_{i=1}^{N}F_i\left(\bar{\mathbf{x}}\left(t\right)\right)\right) \\ - \frac{1}{NT}\sum_{j=1}^{N}\left\|\frac{1}{N}\mathbf{d}_i\left(t\right) - \mathbf{d}_j\left(t\right)\right\|\left\|\mathbf{v}_j\left(t\right) - \mathbf{x}^*\right\| - \frac{LD^2}{2T^2} \\ - \frac{1}{NT}\left\|\sum_{i=1}^{N}\mathbf{d}_i\left(t\right) - \sum_{i=1}^{N}\nabla F_i\left(\bar{\mathbf{x}}\left(t\right)\right)\right\|\left\|\mathbf{x}^* - \frac{1}{N}\sum_{i=1}^{N}\mathbf{v}_i\left(t\right)\right\|.\end{aligned} \tag{99}$$

Since $\mathbf{v}_i\left(t\right) \in \mathcal{K}$ for all $i \in \mathcal{V}$, we know that $(1/N)\sum_{i=1}^{N}\mathbf{v}_i\left(t\right) \in \mathcal{K}$. Therefore, $\|\mathbf{v}_j\left(t\right) - \mathbf{x}^*\| \leq D$ and $\|(1/N)\sum_{i=1}^{N}\mathbf{v}_i\left(t\right) - \mathbf{x}^*\| \leq D$. Furthermore, subtracting the term $(1/N)\sum_{i=1}^{N}F_i(\mathbf{x}^*)$ on both sides of Eq. (99) and using some algebraic manipulations, we have

$$\begin{aligned}\frac{1}{N}\sum_{i=1}^{N}F_i\left(\mathbf{x}^*\right) - \frac{1}{N}\sum_{i=1}^{N}F_i\left(\bar{\mathbf{x}}\left(t+1\right)\right) &\leq \left(1-\frac{1}{T}\right)\left(\frac{1}{N}\sum_{i=1}^{N}F_i\left(\mathbf{x}^*\right) - \frac{1}{N}\sum_{i=1}^{N}F_i\left(\bar{\mathbf{x}}\left(t\right)\right)\right) \\ &+ \frac{1}{NT}\sum_{j=1}^{N}\left\|\frac{1}{N}\mathbf{d}_i\left(t\right) - \mathbf{d}_j\left(t\right)\right\|\left\|\mathbf{v}_j\left(t\right) - \mathbf{x}^*\right\| + \frac{LD^2}{2T^2} \\ &+ \frac{1}{NT}\left\|\sum_{i=1}^{N}\mathbf{d}_i\left(t\right) - \sum_{i=1}^{N}\nabla F_i\left(\bar{\mathbf{x}}\left(t\right)\right)\right\|\left\|\mathbf{x}^* - \frac{1}{N}\sum_{i=1}^{N}\mathbf{v}_i\left(t\right)\right\| \\ &\leq \left(1-\frac{1}{T}\right)\left(\frac{1}{N}\sum_{i=1}^{N}F_i\left(\mathbf{x}^*\right) - \frac{1}{N}\sum_{i=1}^{N}F_i\left(\bar{\mathbf{x}}\left(t\right)\right)\right) \\ &+ \frac{D}{NT}\sum_{j=1}^{N}\left\|\frac{1}{N}\mathbf{d}_i\left(t\right) - \mathbf{d}_j\left(t\right)\right\| + \frac{LD^2}{2T^2} \\ &+ \frac{D}{NT}\left\|\sum_{i=1}^{N}\mathbf{d}_i\left(t\right) - \sum_{i=1}^{N}\nabla F_i\left(\bar{\mathbf{x}}\left(t\right)\right)\right\|.\end{aligned} \tag{100}$$

Applying the above relation (100) recursively, we obtain

$$\begin{aligned}\frac{1}{N}\sum_{i=1}^{N}F_i\left(\mathbf{x}^*\right) - \frac{1}{N}\sum_{i=1}^{N}F_i\left(\bar{\mathbf{x}}\left(t+1\right)\right) &\leq \left(1-\frac{1}{T}\right)^t\left(\frac{1}{N}\sum_{i=1}^{N}F_i\left(\mathbf{x}^*\right) - \frac{1}{N}\sum_{i=1}^{N}F_i\left(\bar{\mathbf{x}}\left(1\right)\right)\right) \\ &+ \frac{D}{NT}\sum_{\tau=1}^{t}\sum_{j=1}^{N}\left\|\frac{1}{N}\mathbf{d}_i\left(\tau\right) - \mathbf{d}_j\left(\tau\right)\right\| + \sum_{\tau=1}^{t}\left(1-\frac{1}{T}\right)^{\tau}\frac{LD^2}{2T^2} \\ &+ \frac{D}{NT}\sum_{\tau=1}^{t}\left\|\sum_{i=1}^{N}\mathbf{d}_i\left(\tau\right) - \sum_{i=1}^{N}\nabla F_i\left(\bar{\mathbf{x}}\left(\tau\right)\right)\right\|.\end{aligned} \tag{101}$$

Since $(1-1/T)^T \leq 1/e$ and $\sum_{\tau=1}^{t}(1-1/T)^\tau \leq T$, we have

$$\begin{aligned}\frac{1}{N}\sum_{i=1}^{N}F_i\left(\mathbf{x}^*\right)-\frac{1}{N}\sum_{i=1}^{N}F_i\left(\bar{\mathbf{x}}\left(t+1\right)\right) \leq \frac{1}{e}\frac{1}{N}\sum_{i=1}^{N}F_i\left(\mathbf{x}^*\right)+\frac{LD^2}{2T}\\ +\frac{D}{NT}\sum_{\tau=1}^{t}\sum_{j=1}^{N}\left\|\frac{1}{N}\mathbf{d}_i\left(\tau\right)-\mathbf{d}_j\left(\tau\right)\right\| \\ +\frac{D}{NT}\sum_{\tau=1}^{t}\left\|\sum_{i=1}^{N}\mathbf{d}_i\left(\tau\right)-\sum_{i=1}^{N}\nabla F_i\left(\bar{\mathbf{x}}\left(\tau\right)\right)\right\|.\end{aligned} \tag{102}$$

Moreover, similar to the expressions (34) and (60) in lemmata 5 and 7, and taking the expectation on both sides of Eq. (102), we obtain

$$\begin{aligned}\left(1-\frac{1}{e}\right)\frac{1}{N}\sum_{i=1}^{N}\mathbb{E}\left[F_i\left(\mathbf{x}^*\right)\right]-\frac{1}{N}\sum_{i=1}^{N}\mathbb{E}\left[F_i\left(\bar{\mathbf{x}}\left(t+1\right)\right)\right] \leq \frac{ND\nu}{T}\sum_{\tau=1}^{t}\frac{\gamma_t\sqrt{2\left(\sigma^2+G^2\right)}}{1-\beta\left(1-\gamma_t\right)}+\frac{LD^2}{2T}\\ +\frac{D}{T}\sum_{\tau=1}^{t}\left(\left(1-\gamma_t\right)^\tau G+\frac{\left(1-\gamma_t\right)LD}{T\gamma_t}+\frac{LND\nu}{T\left(1-\beta\right)}+\frac{LD\sqrt{3\kappa}}{\sqrt{2}T^{2/3}}+\frac{\sqrt{2}\sigma}{T^{1/3}}+\frac{LD\sqrt{3\kappa}}{\sqrt{2}T^{1/3}}\right).\end{aligned} \tag{103}$$

Setting $\gamma_t = 1/\sqrt{T}$, and summing up the inequality (103) from $t=1$ to $t=T$, we have

$$\begin{aligned}\left(1-\frac{1}{e}\right)\frac{1}{N}\sum_{t=1}^{T}\sum_{i=1}^{N}\mathbb{E}\left[F_i\left(\mathbf{x}^*\right)\right]-\frac{1}{N}\sum_{t=1}^{T}\sum_{i=1}^{N}\mathbb{E}\left[F_i\left(\bar{\mathbf{x}}\left(t+1\right)\right)\right] \leq \frac{N\nu\sqrt{\sigma^2+G^2}}{\sqrt{2}\left(1-\beta\right)}\sqrt{T}+\frac{LD^2}{2}\\ +GD\sqrt{T}+\frac{LD^2}{2}\sqrt{T}+\frac{LND^2\nu}{2\left(1-\beta\right)}+\frac{LD^2\sqrt{3\kappa}}{2\sqrt{2}}T^{1/3}+\frac{\sqrt{2}\sigma D}{2}T^{2/3}+\frac{LD^2\sqrt{3\kappa}}{2\sqrt{2}}T^{2/3}.\end{aligned} \tag{104}$$

Since the functions F_i are G-Lipschitz for all $t \in \{1,\ldots,T\}$ and $i \in \{1,\ldots,N\}$, we obtain

$$\left|\frac{1}{N}\sum_{i=1}^{N}F_i\left(\bar{\mathbf{x}}\left(t\right)\right)-\frac{1}{N}\sum_{i=1}^{N}F_i\left(\mathbf{x}_j\left(t\right)\right)\right| \leq \frac{G}{N}\sum_{i=1}^{N}\left\|\bar{\mathbf{x}}\left(t\right)-\mathbf{x}_j\left(t\right)\right\| \leq \frac{GND\nu}{T\left(1-\beta\right)}. \tag{105}$$

Combining Eqs. (104) and (105), we have

$$\begin{aligned}\left(1-\frac{1}{e}\right)\frac{1}{N}\sum_{t=1}^{T}\sum_{i=1}^{N}\mathbb{E}\left[F_i\left(\mathbf{x}^*\right)\right]-\frac{1}{N}\sum_{t=1}^{T}\sum_{i=1}^{N}\mathbb{E}\left[F_i\left(\mathbf{x}_j\left(t+1\right)\right)\right] \leq \frac{N\nu\sqrt{\sigma^2+G^2}}{\sqrt{2}\left(1-\beta\right)}\sqrt{T}+\frac{LD^2}{2}\\ +GD\sqrt{T}+\frac{LD^2}{2}\sqrt{T}+\frac{LND^2\nu}{2\left(1-\beta\right)}+\frac{LD^2\sqrt{3\kappa}}{2\sqrt{2}}T^{1/3}+\frac{\sqrt{2}\sigma D}{2}T^{2/3}+\frac{LD^2\sqrt{3\kappa}}{2\sqrt{2}}T^{2/3}+\frac{GND\nu}{1-\beta}.\end{aligned} \tag{106}$$

Furthermore, since $\kappa = \left(1+2N^2\nu^2/\left(1-\beta\right)^2\right)$, we know that $\sqrt{\kappa} \leq 1+\sqrt{2}N\nu/(1-\beta)$. Plugging the result into Eq. (106), and after some algebraic manipulations, we prove the Theorem 2 completely. ■

In this section, we present the proofs of the main results in detail. The performance evaluation of the proposed algorithms are provided in the next section.

7. Numerical Experiments

In this section, the performance of the proposed algorithms are evaluated by numerical experiments on two datasets.

In the experiments, we use two datasets, i.e., MovieLens and Jester. MovieLens dataset contains $1,000,000$ ratings through $6,000$ users for $4,000$ movies. Moreover, the rating range is $[1,5]$. Jester dataset consists of $73,421$ rating though $73,421$ users for 100 jokes. The rating range is $[-10,10]$. In order to ensure that the ratings are non-negative, the rating range is re-scaled into the range $[0,20]$. In addition, we assume that the data is dispersed equally over the agents of networks.

For our experiments, we use $r_{u,m}$ and $r_{u,j}$ to represent the rating of user u for the the movie m in MovieLens dataset and the joke j in Jester dataset, respectively. Moreover, we split all users into disjoint sets $\mathcal{S}_1, \mathcal{S}_2, \ldots, \mathcal{S}_T$. Each set contains U_m users in MovieLens dataset and U_j in Jester dataset. Furthermore, each agent $i \in \mathcal{V}$ has access to the data of $\mathcal{S}_t$ for all $t \in \{1, \ldots, T\}$. Therefore, each subset is denoted by $\mathcal{S}_{i,t}$. For the sake of description, we use v to denote m in MovieLens dataset or j in Jester dataset. The facility location objective function, which is associated to each user u, is given by $\phi_u(S_v) = \max_{v \in S_v} r_{u,v}$, where S_v represents any subset of the movies in MovieLens dataset or the jokes in Jester dataset. In other words, $S_v \in \mathcal{B}_m := \{1, \ldots, 4000\}$ for $v = m$ and $S_v \in \mathcal{B}_j := \{1, \ldots, 600\}$ for $v = j$. Moreover, we use the notation $\mathcal{B}_v$ to denote $\mathcal{B}_m$ in MovieLens dataset or $\mathcal{B}_j$ in Jester dataset. Therefore, we associate to each agent i an objective function, which is defined as follows:

$$f_{i,t}(S_v) := \sum_{u \in \mathcal{S}_{i,t}} \phi(S_v).$$

Following on from Iyer et al. (2014), we obtain the multilinear extension of $f_{i,t}(S_v)$ as follows:

$$F_{i,t}(\mathbf{x}) = \sum_{u \in \mathcal{S}_{i,t}} \sum_{\ell=1}^{|\mathcal{B}_v|} r_{u,v_u^\ell} \mathbf{x}_{v_u^\ell} \prod_{\hbar=1}^{\ell-1} \left(1 - \mathbf{x}_{v_u^\hbar}\right)$$

for all $\mathbf{x} \in [0,1]^{|\mathcal{B}_v|}$, where $|\mathcal{B}_v|$ is the cardinal number of the set $\mathcal{B}_v$. Moreover, $v_u^1, \ldots, v_u^{|\mathcal{B}_v|}$ denotes a permutation of $1, \ldots, |\mathcal{B}_v|$ and satisfies the condition $r_{u,v_u^1} \geq \cdots \geq r_{u,v_u^{|\mathcal{B}_v|}}$. In addition, we set $U_m = 5$ in MovieLens dataset and $U_j = 5$ in Jester dataset, respectively. Moreover, the constraint set is $\left\{\mathbf{x} \in [0,1]^{|\mathcal{B}_v|} : \mathbb{1}^\mathsf{T}\mathbf{x} \leq 1\right\}$.

Firstly, we compare Algorithm 1 (DMFW) and Algorithm 2 (DOSFW) with the distributed online learning algorithm, i.e., DOGD, which is proposed in Yan et al. (2013). In this experiment, we set $N = 100$. Moreover, Algorithm 1 (DMFW), Algorithm 2 (DOSFW) and DOGD run on the complete graph, where each node is connected with other nodes. As shown in Figure 1, the smallest average regret is obtained by DMFW. In other words, the performance of DMFW is better than DOSFW and DOGD.

How the number of nodes affects the performance of Algorithm 1 is investigated on MovieLens and Jester datasets in the second experiment. The results is summarized in Figure 2. From Figure 2, we can observe that the average regret decrease more slowly as the number of nodes increases. Therefore, the theoretical results are confirmed by the experimental results. Compared with the centralized MFW (Chen et al., 2018b), the comparable results can be obtained by Algorithm 1, which is implemented in a decentralized way.

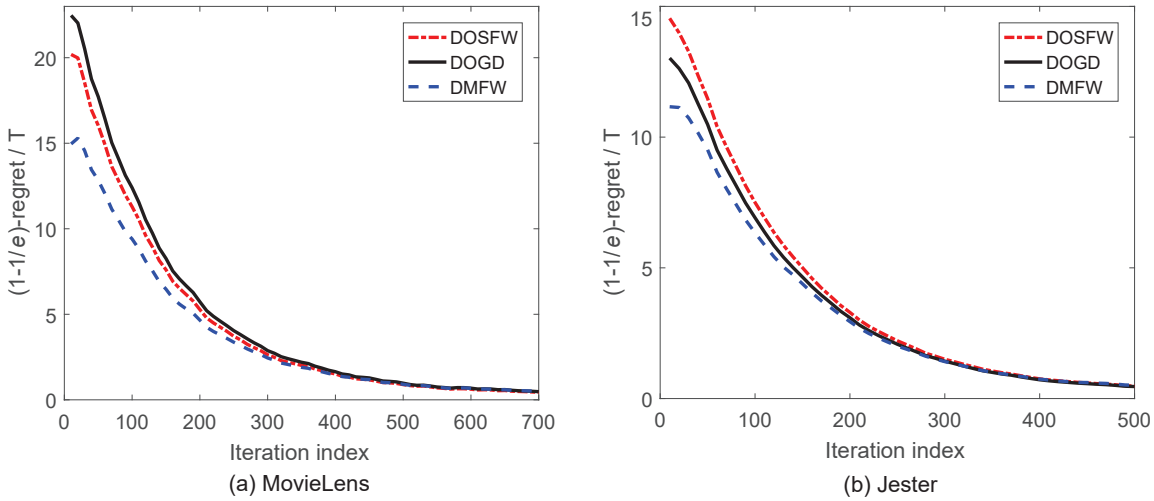

Figure 1: Comparison of DMFW, DOSFW, and DOGD on the MovieLens and Jester datasets.

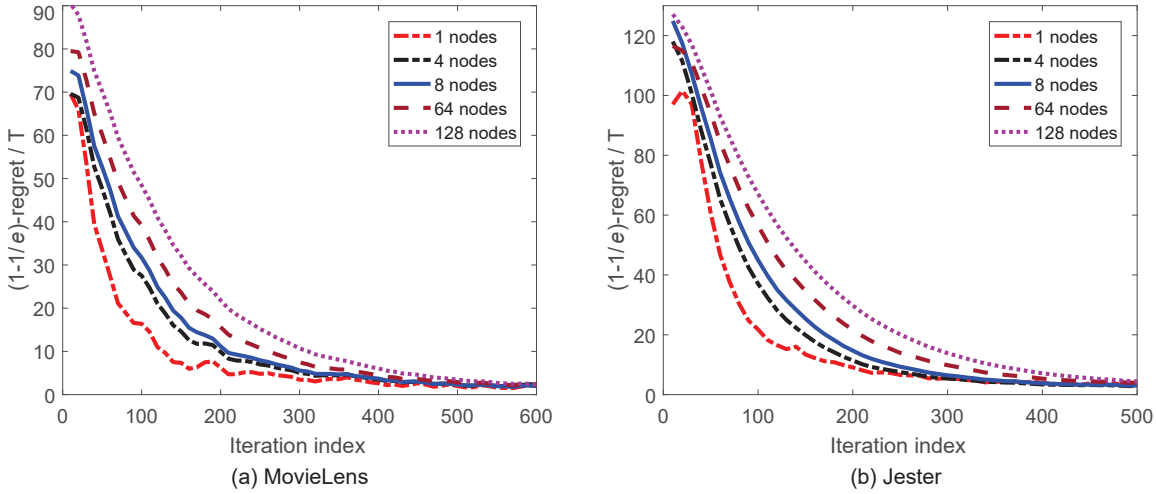

Figure 2: Comparison of DMFW with different number of nodes on the MovieLens and Jester datasets.

In the third experiment, we investigate how the network topology affects the performance of Algorithm 1 on the MovieLens and Jester datasets with 100 nodes. To this end, we construct three types of network topologies, i.e., complete graph, cycle graph, and Watts-Strogatz. The experimental results are summarized in Figure 3. Compared with the cycle graph and Watts-Strogatz, the complete graph leads to slightly faster convergence. In other words, the better connectivity can improve convergence rate of Algorithm 1.

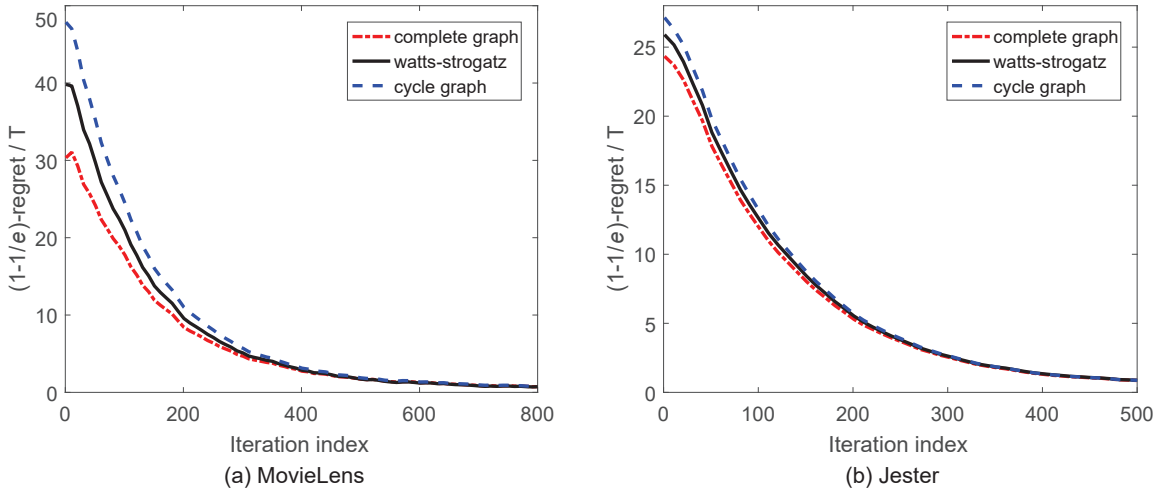

Figure 3: Comparison of Algorithm 1 with fixed 100 nodes and different network topology on the MovieLens and Jester datasets.

8. Conclusion

In this paper, we have considered the distributed online submodular optimization problems over networks, where each agent has only access to its own submodular function. For the adversarial online setting, we have proposed a distributed Meta-Frank-Wolfe online learning algorithm to solve the optimization problems using local communication and local computation. We have also showed that the proposed algorithm can achieve a expected square-root regret bound with $(1 - 1/e)$ approximation guarantee. Additionally, we have proposed a distributed one-shot Frank-Wolfe online learning algorithm for the stochastic online setting. Furthermore, we have also showed that the proposed algorithm can achieve an expected regret bound of $\mathcal{O}(T^{2/3})$ with $(1 - 1/e)$ approximation guarantee, where T is a time horizon. Finally, we have confirmed the theoretical results by various numerical experiments.

Acknowledgments

We would like to acknowledge support for this project in part by the National Natural Science Foundation of China (NSFC) under Grants no. 61976243, and no. 61971458, and in part by the Leading talents of science and technology in the Central Plain of China under Grants no. 214200510012, and in part by the Scientific and Technological Innovation Team of Colleges and Universities in Henan Province under Grants No. 20IRTSTHN018, and in part by the basic research projects in the University of Henan Province under Grants No. 19zx010.

References

Zeyuan Allen-Zhu and Elad Hazan. Variance reduction for faster non-convex optimization. In *Proceedings of the 33rd International Conference on Machine Learning*, pages 699–707, 2016.

Francis Bach. Structured sparsity-inducing norms through submodular functions. In *Advances in Neural Information Processing Systems*, pages 118–126, 2010.

Francis Bach. Submodular functions: from discrete to continuous domains. *arXiv: 1511.00394*, 2015.

Andrew An Bian, Baharan Mirzasoleiman, Joachim M Buhmann, and Andreas Krause. Guaranteed non-convex optimization: Submodular maximization over continuous domains. In *Proceedings of The 20th International Conference on Artificial Intelligence and Statistics*, pages 111–120, 2017.

Stephen Boyd, Neal Parikh, Eric Chu, Borja Peleato, and Jonathan Echstein. Distributed optimization and statistical learning via the alternating direction method of multiplies. *Foundations and Trends® in Machine Learning*, 3(1):1–122, 2011.

Volkan Cevher, Stephen Becker, and Mark Schmidt. Convex optimization for big data: Scalable, randomized, and parallel algorithms for big data analytics. *IEEE Signal Processing Magazine*, 31(5):32–43, 2014.

Lin Chen, Hamed Hassani, and Amin Karbasi. Online continuous submodular maximization. *arXiv: 1802.06052*, 2018a.

Lin Chen, Christopher Harshaw, Hamed Hassani, and Amin Karbasi. Projection-free online optimization with stochastic gradient: From convexity to submodularity. *arXiv: 1802.08183*, 2018b.

Abhimanyu Das and David Kempe. Submodular meets spectral: Greedy algorithms for subset selection, sparse approximation and dictionary selection. In *Proceedings of the 28th International Conference on Machine Learning*, pages 1057–1064, 2011.

Josip Djolonga and Andreas Krause. From map to marginals: Variational inference in bayesian submodular models. In *Advances in Neural Information Processing Systems*, pages 244–252, 2014.

Pedro Domingos and Matt Richardson. Mining the network value of customers. In *Proceedings of the Seventh ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 57–66, 2001.

Uriel Feige, Vahab S Mirrokni, and Jan Vondrak. Maximizing non-monotone submodular functions. *SIAM Journal on Computing*, 40(4):1133–1153, 2011.

Moran Feldman, Christopher Harshaw, and Amin Karbasi. Greed is good: Near-optimal submodular maximization via greedy optimization. *arXiv: 1704.01652*, 2017.

Marshall L Fisher, George L Nemhauser, and Laurence A Wolsey. An analysis of approximations for maximizing submodular set functions – ii. *Mathematical Programming Study*, 8:73–87, 1978.

Gene H Golub and Charles F Van Loan. Matrix Computations (Fourth Edition). *The Johns Hopkins University Press*, 2013.

Hamed Hassani, Mahdi Soltanolkotabi, and Amin Karbasi. Gradient methods for submodular maximization. In *Advances in Neural Information Processing Systems*, pages 5841–5851, 2017.

Alon Cohen and Elad Hazan. Following the perturbed leader for online structure learning. In *Proceedings of the 32nd International Conference on Machine Learning*, pages 1034–1042, 2015.

Elad Hazan and Haipeng Luo. Variance-reduced and projection-free stochastic optimization. In *Proceedings of the 33rd International Conference on Machine Learning*, pages 1263–1271, 2016.

Saghar Hosseini, Airlie Chapman, and Mehran Mesbahi. Online distributed convex optimization on dynamic networks. *IEEE Transactions on Automatic Control*, 61(11):3545–3550, 2016.

Satoru Iwata, Lisa Fleischer, and Satoru Fujishige. A combinatorial strongly polynomial algorithm for minimizing submodular functions. *Journal of the ACM*, 48(4):761–777, 2001.

Rishabh K Iyer, Stefanie Jegelka, and Jeff A Bilmes. Monotone closure of relaxed constraints in submodular optimization: Connections between minimization and maximization. In *Proceedings of the 30th Conference on Uncertainty in Artificial Intelligence*, pages 360–369, 2014.

Rie Johnson and Tong Zhang. Accelerating stochastic gradient descent using predictive variance reduction. In *Advances in Neural Information Processing Systems*, pages 315–323, 2013.

Mohammad Reza Karimi, Mario Lucic, Hamed Hassani, and Andreas Krause. Stochastic submodular maximization: The case of coverage functions. In *Advances in Neural Information Processing Systems*, pages 6853–6863, 2017.

David Kempe, Jon Kleinberg, and Éva Tardos. Maximizing the Spread of Influence Through a Social Network. In *Proceedings of the Ninth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 137–146, 2003.

Andreas Krause and Volkan Cevher. Submodular dictionary selection for sparse representation. In *Proceedings of the 27th International Conference on Machine Learning*, pages 567–574, 2010.

Andreas Krause and Daniel Golovin. Submodular function maximization. *Tractability: Practical Approaches to Hard Problems*, 3:19, 2012.

Andreas Krause and Carlos Guestrin. Near-optimal nonmyopic value of information in graphical models. In *Proceedings of the Twenty-First Conference on Uncertainty in Artificial Intelligence*, pages 324–331, 2005.

Alex Kulesza and Ben Taskar. *Determinantal Point Processes for Machine Learning*. Now Publishers Inc., MA, USA, 2012.

Hui Lin and Jeff Bilmes. A class of submodular functions for document summarization. In *Proceeding of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies-Volume I*, pages 510–520, 2011.

Mehrdad Mahdavi, Lijun Zhang, and Rong Jin. Mixed optimization for smooth functions. In *Advances in Neural Information Processing Systems*, pages 674–682, 2013.

Vahab S Mirrokni and Morteza Zadimoghaddam. Randomized composable core-sets for distributed submodular maximization. In *Proceedings of the Forty-Seventh Annual ACM Symposium on Theory of Computing*, pages 153–162, 2015.

Baharan Mirzasoleiman, Amin Karbasi, Rik Sarkar, and Andreas Krause. Distributed submodular maximization: Identifying representative elements in masstive data. In *Advances in Neural Information Processing Systems*, pages 2049–2057, 2013.

Baharan Mirzasoleiman, Ashwinkunar Badanidiyuru, and Amin Karbasi. Fast constrained submodular maximization: Personalized data summarization. In *Proceedings of the 33rd International Conference on Machine Learning*, pages 1358–1367, 2016.

Aryan Mokhtari, Hamed Hassani, Amin Karhasi. Stochastic conditional gradient methods: From convex minimization to submodular maximization. *arXiv: 1804.09554*, 2018a.

Aryan Mokhtari, Hamed Hassani, and Amin Karbasi. Decentralized submodular maximization: Bridging discrete and continuous settings. *arXiv: 1802.03825*, 2018b.

Katta G Murty and Santosh N Kabadi. Some NP-complete problems in quadratic and nonlinear programming. *Mathematical Programming*, 39(2):117–129, 1987.

Angelia Nedić, Alex Olshevsky, Asuman Ozdaglar, and John N. Tsitsiklis. Distributed subgradient methods and quantization effects. In *Proceedings of the 47th IEEE Conference on Decision and Control*, pages 4177–4184, 2008.

George L Nemhauser, Laurence A Wolsey, and Marshall L Fisher. An analysis of approximations for maximizing submodular set functions – i. *Mathematical Programming*, 14(1):265–294, 1978.

George L Nemhauser and Laurence A Wolsey. Best algorithms for approximating the maximum of a submodular set function. *Mathematics of Operations Research*, 3(3):177–188, 1978.

Sashank J Reddi, Ahmed Hefny, Suvrit Sra, Barnabás Póczós, and Alex Smola. Stochastic variance reduction for nonconvex optimization. In *Proceedings of the 33rd International Conference on Machine Learning*, pages 314–323, 2016.

Ali H Sayed, Sheng-Yuan Tu, Jianshu Chen, Xiaochuan Zhao, and Zaid J Towfic. Diffusion strategies for adaptation and learning over networks: An examination of distributed strategies and network behavior. *IEEE Signal Processing Magazine*, 30(3):155–171, 2013.

Shahin Shahrampour and Ali Jadbabaie. Distributed online optimization in dynamic environments using mirror descent. *IEEE Transactions on Automatic Control*, 63(3):714–715, 2018.

Feng Yan, Shreyas Sundaram, S V N Vishwanathan, and Yuan Qi. Distributed autonomous online learning: Regrets and intrinsic privacy-preserving properties. *IEEE Transactions on Knowledge and Data Engineering*, 25(11):2483–2493, 2013.

Wenpeng Zhang, Peilin Zhao, Wenwu Zhu, Steven C H Hoi, and Tong Zhang. Projection-free distributed online learning in networks. In *Proceedings of the 34th International Conference on Machine Learning*, pages 4054–4062, 2017.

Mingchuan Zhang, Wei Quan, Nan Cheng, Qingtao Wu, Junlong Zhu, Ruijuan Zheng, and Keqin Li. Distributed conditional gradient online learning for IoT optimization. *IEEE Internet of Things Journal*, 2019a.

Mingrui Zhang, Lin Chen, Aryan Mokhtari, Hamed Hassani, and Amin Karibasi. Quantized Frank-Wolfe: Faster optimization, lower communication, and projection free. *arXiv: 1902.06332*, 2019b.

Mingrui Zhang, Lin Chen, Hamed Hassani, and Amin Karibasi. Online continuous submodular maximization: From full-information to bandit feedback. In *Advances in Neural Information Processing Systems*, pages 9210–9221, 2019c.

Jiacheng Zhuo, Qi Lei, Alexandros G. Dimakis, and Constantine Caramanis. Communication-efficient asynchronous stochastic Frank-Wolfe over nuclear-norm balls. *arXiv: 1910.07703*, 2019.