Journal of Machine Learning Research 23 (2022) 1-53 Submitted 11/20; Revised 7/22; Published 7/22

Projection-free Distributed Online Learning with Sublinear Communication Complexity

Yuanyu Wan WANYY9@GMAIL.COM
School of Software Technology, Zhejiang University
Ningbo 315048, China
Guanghui Wang GWANG369@GATECH.EDU
College of Computing, Georgia Tech
Atlanta, GA 30332, USA
Wei-Wei Tu TUWEIWEI@4PARADIGM.COM
4Paradigm Inc.
Beijing 100000, China
Lijun Zhang* ZHANGLJ@LAMDA.NJU.EDU.CN
National Key Laboratory for Novel Software Technology, Nanjing University
Nanjing 210023, China

Editor: Csaba Szepesvari

Abstract

To deal with complicated constraints via locally light computations in distributed online learning, a recent study has presented a projection-free algorithm called distributed online conditional gradient (D-OCG), and achieved an $O(T^{3/4})$ regret bound for convex losses, where T is the number of total rounds. However, it requires T communication rounds, and cannot utilize the strong convexity of losses. In this paper, we propose an improved variant of D-OCG, namely D-BOCG, which can attain the same $O(T^{3/4})$ regret bound with only $O(\sqrt{T})$ communication rounds for convex losses, and a better regret bound of $O(T^{2/3}(\log T)^{1/3})$ with fewer $O(T^{1/3}(\log T)^{2/3})$ communication rounds for strongly convex losses. The key idea is to adopt a delayed update mechanism that reduces the communication complexity, and redefine the surrogate loss function in D-OCG for exploiting the strong convexity. Furthermore, we provide lower bounds to demonstrate that the $O(\sqrt{T})$ communication rounds required by D-BOCG are optimal (in terms of T) for achieving the $O(T^{3/4})$ regret with convex losses, and the $O(T^{1/3}(\log T)^{2/3})$ communication rounds required by D-BOCG are near-optimal (in terms of T) for achieving the $O(T^{2/3}(\log T)^{1/3})$ regret with strongly convex losses up to polylogarithmic factors. Finally, to handle the more challenging bandit setting, in which only the loss value is available, we incorporate the classical one-point gradient estimator into D-BOCG, and obtain similar theoretical guarantees.

Keywords: projection-free, distributed online learning, communication complexity, conditional gradient

1. Introduction

Conditional gradient (CG) (Frank and Wolfe, 1956; Jaggi, 2013) (also known as Frank-Wolfe) is a simple yet efficient offline algorithm for solving high-dimensional problems with

*. Corresponding Author.

complicated constraints. To find a feasible solution, instead of performing the time-consuming projection step, CG utilizes the linear optimization step, which can be carried out much more efficiently. For example, in the matrix completion problem (Hazan and Kale, 2012), where the feasible set consists of all matrices with bounded trace norm, the projection step needs to compute the singular value decomposition (SVD) of a matrix. In contrast, the linear optimization step in CG only requires computing the top singular vector pair of a matrix, which is at least an order of magnitude faster than the SVD. Due to the emergence of large-scale problems, online conditional gradient (OCG) (Hazan and Kale, 2012; Hazan, 2016) (also known as online Frank-Wolfe) was proposed for online convex optimization (OCO)—a multi-round game between a learner and an adversary (Zinkevich, 2003), and achieved a regret bound of $O(T^{3/4})$ for convex losses, where T is the number of total rounds. In each round, OCG updates the learner by utilizing one linear optimization step to minimize a surrogate loss function. Different from CG that requires all data related to the objective function are given beforehand, OCG only requires a single data point per round.

Recently, Zhang et al. (2017) further proposed D-OCG by extending OCG into a more practical scenario—distributed OCO over a network. It is well motivated by many distributed applications such as multi-agent coordination and distributed tracking in sensor networks (Li et al., 2002; Xiao et al., 2007; Nedić et al., 2009; Duchi et al., 2011; Yang et al., 2019). Specifically, by defining the network as an undirected graph, each node of the graph represents a local learner, and can only communicate with its neighbors. The key idea of D-OCG is to maintain OCG for each local learner, and update it according to the local gradient as well as those received from its neighbors in each round. Compared with projection-based distributed algorithms (Ram et al., 2010; Hosseini et al., 2013; Yan et al., 2013), D-OCG significantly reduces the time cost for solving high-dimensional problems with complicated constraints, because it only utilizes one linear optimization step for each update of local learners. Moreover, D-OCG is more scalable than OCG, since it can utilize many locally light computation resources to handle large-scale problems.

However, there exist two interesting questions about D-OCG. First, the local learners of D-OCG communicate with their neighbors to share the local gradients in each round, so it requires T communication rounds in total. Since the communication overhead is often the performance bottleneck in distributed systems, it is natural to ask whether the communication complexity of D-OCG can be reduced without increasing its regret. Second, similar to OCG in the standard OCO, Zhang et al. (2017) have proved that D-OCG in the distributed OCO achieves an $O(T^{3/4})$ regret bound for convex losses. Note that recent studies (Garber and Kretzu, 2021; Wan and Zhang, 2021) in the standard OCO have proposed variants of OCG to attain better regret for strongly convex losses. It is thus natural to ask whether the strong convexity can also be utilized to improve the regret of D-OCG.

In this paper, we provide affirmative answers for these two questions by developing an improved variant of D-OCG, namely distributed block online conditional gradient (D-BOCG), which can attain the same $O(T^{3/4})$ regret bound with only $O(\sqrt{T})$ communication rounds for convex losses, and a better regret bound of $O(T^{2/3}(\log T)^{1/3})$ with fewer $O(T^{1/3}(\log T)^{2/3})$ communication rounds for strongly convex losses. Compared with the original D-OCG, there exist three critical changes.

- To further utilize the strong convexity of losses, a more general surrogate loss function is introduced in our D-BOCG, which is inspired by the surrogate loss function used in

strongly convex variants of OCG (Garber and Kretzu, 2021; Wan and Zhang, 2021) and is able to cover that used in D-OCG.

- To reduce the communication complexity, our D-BOCG adopts a delayed update mechanism, which divides the total T rounds into a smaller number of equally-sized blocks, and only updates the local learners for each block. In this way, the local learners only need to communicate with their neighbors once for each block, and the total number of communication rounds is reduced from T to the number of blocks.
- According to the delayed update mechanism, the number of updates in our D-BOCG could be much smaller than that in D-OCG, which brings a new challenge, i.e., only performing 1 linear optimization step as D-OCG for each update will increase the $O(T^{3/4})$ regret of D-OCG for convex losses. To address this problem, we perform iterative linear optimization steps for each update. Specifically, the number of linear optimization steps for each update is set to be the same as the block size, which ensures that the total number of linear optimization steps required by our D-BOCG is the same as that required by D-OCG.

Note that the delayed update mechanism and the iterative linear optimization steps in the last two changes are borrowed from Garber and Kretzu (2020) that employed them to improve projection-free bandit convex optimization. In contrast, we apply them to the distributed setting considered here.

Furthermore, to complement theoretical guarantees of our D-BOCG, for any distributed online algorithm with C communication rounds, we provide an $\Omega(T/\sqrt{C})$ lower regret bound with convex losses, and an $\Omega(T/C)$ lower regret bound with strongly convex losses, respectively. These lower bounds imply that the $O(\sqrt{T})$ communication rounds required by D-BOCG are optimal (in terms of T) for achieving the $O(T^{3/4})$ regret with convex losses, and the $O(T^{1/3}(\log T)^{2/3})$ communication rounds required by D-BOCG are near-optimal (in terms of T) for achieving the $O(T^{2/3}(\log T)^{1/3})$ regret with strongly convex losses up to polylogarithmic factors. To the best of our knowledge, we are the first to study lower regret bounds for distributed online algorithms with limited communication rounds. Finally, to handle the more challenging bandit setting, we propose distributed block bandit conditional gradient (D-BBCG) by combining D-BOCG with the classical one-point gradient estimator (Flaxman et al., 2005), which can approximate the gradient with a single loss value. Our theoretical analysis first reveals that in expectation, D-BBCG can also attain a regret bound of $O(T^{3/4})$ with $O(\sqrt{T})$ communication rounds for convex losses, and a regret bound of $O(T^{2/3}(\log T)^{1/3})$ with $O(T^{1/3}(\log T)^{2/3})$ communication rounds for strongly convex losses. Moreover, for convex losses, we show that D-BBCG enjoys a high-probability regret bound of $O(T^{3/4}(\log T)^{1/2})$ with $O(\sqrt{T})$ communication rounds.

A preliminary version of this paper was presented at the 37th International Conference on Machine Learning in 2020 (Wan et al., 2020). In this paper, we have significantly enriched the preliminary version by adding the following extensions.

- Different from Wan et al. (2020) that only considered convex losses, we generalize D-BOCG and D-BBCG to further exploit the strong convexity, and establish improved theoretical guarantees for strongly convex losses.
- Different from Wan et al. (2020) that only studied upper regret bounds, we provide lower bounds on the regret of distributed online algorithms with limited communication rounds for convex losses as well as strongly convex losses.

- We provide more experiments including new results for distributed online binary classification, different networks topologies, different network sizes, and three additional data sets.

2. Related Work

In this section, we briefly review existing projection-free algorithms for OCO and the distributed OCO.

2.1 Projection-free Algorithms for OCO

OCO is a general framework for online learning, which covers a variety of problems such as online portfolio selection (Blum and Kalai, 1999; Agarwal et al., 2006), online routing (Awerbuch and Kleinberg, 2004, 2008), online metric learning (Jain et al., 2008; Tsagkatakis and Savakis, 2011), and learning with expert advice (Cesa-Bianchi et al., 1997; Freund et al., 1997). It is generally viewed as a repeated game between a leaner and an adversary. In each round t, the learner first chooses a decision $\mathbf{x}(t)$ from a convex decision set $\mathcal{K} \subseteq \mathbb{R}^d$. Then, the adversary reveals a convex function $f_t(\mathbf{x}) : \mathcal{K} \to \mathbb{R}$, which incurs a loss $f_t(\mathbf{x}(t))$ to the learner. The goal of the learner is to minimize the regret with respect to any fixed optimal decision, which is defined as

$$R_T = \sum_{t=1}^{T} f_t(\mathbf{x}(t)) - \min_{\mathbf{x} \in \mathcal{K}} \sum_{t=1}^{T} f_t(\mathbf{x}).$$

OCG (Hazan and Kale, 2012; Hazan, 2016) is the first projection-free algorithm for OCO, which enjoys a regret bound of $O(T^{3/4})$ for convex losses and updates as the following linear optimization step

$$\begin{aligned} \mathbf{v} &= \operatorname*{argmin}_{\mathbf{x} \in \mathcal{K}} \nabla F_t(\mathbf{x}(t))^\top \mathbf{x} \\ \mathbf{x}(t+1) &= \mathbf{x}(t) + s_t(\mathbf{v} - \mathbf{x}(t)) \end{aligned} \tag{1}$$

where

$$F_t(\mathbf{x}) = \eta \sum_{k=1}^{t-1} \nabla f_k(\mathbf{x}(k))^\top \mathbf{x} + \|\mathbf{x} - \mathbf{x}(1)\|_2^2 \tag{2}$$

is a surrogate loss function, and s_t, η are two parameters.

Recent studies have proposed to improve the regret of OCG by exploiting the additional curvature of loss functions including smoothness and strong convexity. For convex and smooth losses, Hazan and Minasyan (2020) proposed an improved projection-free algorithm, which can attain an expected regret bound of $O(T^{2/3})$ as well as a high-probability regret bound of $O(T^{2/3} \log T)$. If the losses are α-strongly convex, Wan and Zhang (2021) proposed a strongly convex variant of OCG by redefining the surrogate loss function as

$$F_t(\mathbf{x}) = \sum_{k=1}^{t-1} \left(\nabla f_k(\mathbf{x}(k))^\top \mathbf{x} + \frac{\alpha}{2} \|\mathbf{x} - \mathbf{x}(k)\|_2^2 \right) \tag{3}$$

and using a line search rule to select the original parameter s_t in (1). This algorithm can enjoy a regret bound of $O(T^{2/3})$ for strongly convex losses, and a very similar algorithm was concurrently proposed by Garber and Kretzu (2021). Moreover, when the decision set is polyhedral (Garber and Hazan, 2016) or smooth (Levy and Krause, 2019), projection-free algorithms have been proposed to enjoy an $O(\sqrt{T})$ regret bound for convex losses and an $O(\log T)$ regret bound for strongly convex losses, respectively. If the decision set is strongly convex, Wan and Zhang (2021) have proved that OCG can achieve an $O(T^{2/3})$ regret bound for convex losses, and the strongly convex variant of OCG can achieve an $O(\sqrt{T})$ regret bound for strongly convex losses.

Furthermore, OCG has been extended to handle the more challenging bandit setting, where only the loss value is available to the learner. Due to the lack of the gradient, Chen et al. (2019) proposed a bandit variant of OCG by combining with the one-point gradient estimator (Flaxman et al., 2005), which can approximate the gradient with a single loss value. For convex losses, the bandit variant of OCG achieves an expected regret bound of $O(T^{4/5})$, which is worse than the $O(T^{3/4})$ regret bound of OCG. Later, by dividing the total rounds into several equally-sized blocks and performing iterative linear optimization steps for each block, Garber and Kretzu (2020) improved the bandit variant of OCG, and reduced the expected regret bound for convex losses from $O(T^{4/5})$ to $O(T^{3/4})$. For strongly convex losses, Garber and Kretzu (2021) further developed a projection-free bandit algorithm that attains an expected regret bound of $O(T^{2/3} \log T)$.

We also note that Chen et al. (2018) developed a projection-free algorithm for another interesting setting where the learner is allowed to access to the stochastic gradients.

2.2 Projection-free Algorithms for the Distributed OCO

According to previous studies (Hosseini et al., 2013; Zhang et al., 2017), distributed OCO is a variant of OCO over a network defined by an undirected graph $\mathcal{G} = (V, E)$, where $V = [n]$ is the node set and $E \subseteq V \times V$ is the edge set. Different from OCO where only 1 learner exists, in the distributed OCO, each node $i \in V$ is a local learner, and can communicate with its immediate neighbors

$$N_i = \{j \in V | (i, j) \in E\}.$$

In each round t, each local learner $i \in V$ chooses a decision $\mathbf{x}_i(t) \in \mathcal{K}$, and then it receives a convex loss function $f_{t,i}(\mathbf{x}) : \mathcal{K} \to \mathbb{R}$ chosen by the adversary. Moreover, the global loss function $f_t(\mathbf{x})$ is defined as the sum of local loss functions

$$f_t(\mathbf{x}) = \sum_{j=1}^{n} f_{t,j}(\mathbf{x}).$$

The goal of each local learner $i \in V$ is to minimize the regret measured by the global loss with respect to the optimal fixed decision, which is defined as

$$R_{T,i} = \sum_{t=1}^{T} f_t(\mathbf{x}_i(t)) - \min_{\mathbf{x} \in \mathcal{K}} \sum_{t=1}^{T} f_t(\mathbf{x}).$$

Since the local loss function $f_{t,i}(\mathbf{x})$ is only available to the local learner i, to achieve the global goal for all local learners, it is necessary to utilize both their local gradients and those received from their neighbors.

Therefore, to make OCG distributed, Zhang et al. (2017) first introduced a non-negative weight matrix $P \in \mathbb{R}^{n\times n}$, and redefined the surrogate loss function $F_t(\mathbf{x})$ in OCG as

$$F_{t,i}(\mathbf{x}) = \eta \mathbf{z}_i(t)^\top \mathbf{x} + \|\mathbf{x} - \mathbf{x}_1(1)\|_2^2 \tag{4}$$

for each local learner i by replacing $\sum_{k=1}^{t-1} \nabla f_k(\mathbf{x}(k))$ with $\mathbf{z}_i(t)$, where $\mathbf{z}_i(1) = \mathbf{0}$ and

$$\mathbf{z}_i(t+1) = \sum_{j\in N_i} P_{ij}\mathbf{z}_j(t) + \nabla f_{t,i}(\mathbf{x}_i(t)). \tag{5}$$

Note that the matrix P is also referred to as a gossip, consensus, or averaging matrix (Bénézit et al., 2007; Tsianos and Rabbat, 2012; Koloskova et al., 2019). Moreover, $\mathbf{z}_i(t)$ is a weighted sum of historical local gradients and those received from neighbors, which could be viewed as an approximation for the sum of global gradients and is critical for minimizing the global regret.

Then, with a time-varying parameter $s_t = 1/\sqrt{t}$, Zhang et al. (2017) proposed D-OCG updating as follows

$$\begin{aligned}
&\textbf{for } \text{each local learner } i \in V \textbf{ do}\\
&\quad \mathbf{v}_i = \operatorname*{argmin}_{\mathbf{x}\in\mathcal{K}} \nabla F_{t,i}(\mathbf{x}_i(t))^\top \mathbf{x}\\
&\quad \mathbf{x}_i(t+1) = \mathbf{x}_i(t) + s_t(\mathbf{v}_i - \mathbf{x}_i(t))\\
&\textbf{end for}
\end{aligned} \tag{6}$$

and established a regret bound of $O(T^{3/4})$ for convex losses. However, in each round t, each local learner i needs to compute $\mathbf{z}_i(t+1)$ by communicating with its neighbors, which requires T communication rounds in total.

3. Preliminaries

In this section, we introduce necessary preliminaries including standard definitions, basic algorithmic ingredients, and common assumptions.

3.1 Definitions

We first recall the standard definitions for smooth and strongly convex functions (Boyd and Vandenberghe, 2004).

Definition 1 *Let $f(\mathbf{x}) : \mathcal{K} \to \mathbb{R}$ be a function over $\mathcal{K}$. It is called β-smooth over $\mathcal{K}$ if for all $\mathbf{x} \in \mathcal{K}, \mathbf{y} \in \mathcal{K}$*

$$f(\mathbf{y}) \le f(\mathbf{x}) + \nabla f(\mathbf{x})^\top(\mathbf{y} - \mathbf{x}) + \frac{\beta}{2}\|\mathbf{y} - \mathbf{x}\|_2^2.$$

Definition 2 *Let $f(\mathbf{x}) : \mathcal{K} \to \mathbb{R}$ be a function over $\mathcal{K}$. It is called α-strongly convex over $\mathcal{K}$ if for all $\mathbf{x} \in \mathcal{K}, \mathbf{y} \in \mathcal{K}$*

$$f(\mathbf{y}) \ge f(\mathbf{x}) + \nabla f(\mathbf{x})^\top(\mathbf{y} - \mathbf{x}) + \frac{\alpha}{2}\|\mathbf{y} - \mathbf{x}\|_2^2.$$

Algorithm 1 CG

Input: feasible set $\mathcal{K}$, L, $F(\mathbf{x})$, $\mathbf{x}_{\text{in}}$
$\mathbf{c}_0 = \mathbf{x}_{\text{in}}$
for $\tau = 0, \ldots, L-1$ **do**
  $\mathbf{v}_\tau \in \operatorname{argmin}_{\mathbf{x}\in\mathcal{K}} \nabla F(\mathbf{c}_\tau)^\top \mathbf{x}$
  $s_\tau = \operatorname{argmin}_{s\in[0,1]} F(\mathbf{c}_\tau + s(\mathbf{v}_\tau - \mathbf{c}_\tau))$
  $\mathbf{c}_{\tau+1} = \mathbf{c}_\tau + s_\tau(\mathbf{v}_\tau - \mathbf{c}_\tau)$
end for
return $\mathbf{x}_{\text{out}} = \mathbf{c}_L$

3.2 Algorithmic Ingredients

Then, we present conditional gradient (CG) (Frank and Wolfe, 1956; Jaggi, 2013), which will be utilized to minimize surrogate loss functions in our algorithms. Given a function $F(\mathbf{x}) : \mathcal{K} \to \mathbb{R}$ and an initial point $\mathbf{c}_0 = \mathbf{x}_{\text{in}} \in \mathcal{K}$, it iteratively performs the linear optimization step as follows

$$\mathbf{v}_\tau \in \operatorname*{argmin}_{\mathbf{x}\in\mathcal{K}} \nabla F(\mathbf{c}_\tau)^\top \mathbf{x}$$
$$\mathbf{c}_{\tau+1} = \mathbf{c}_\tau + s_\tau(\mathbf{v}_\tau - \mathbf{c}_\tau)$$

for $\tau = 0, \ldots, L-1$, where L is the number of iterations and s_τ is selected by line search

$$s_\tau = \operatorname*{argmin}_{s\in[0,1]} F(\mathbf{c}_\tau + s(\mathbf{v}_\tau - \mathbf{c}_\tau)).$$

The detailed procedures of CG are summarized in Algorithm 1, and its convergence rate is presented in the following lemma.

Lemma 1 *(Derived from Theorem 1 of Jaggi 2013) If $F(\mathbf{x}) : \mathcal{K} \to \mathbb{R}$ is a convex and β-smooth function and $\|\mathbf{x}\|_2 \leq R$ for any $\mathbf{x} \in \mathcal{K}$, Algorithm 1 with $L \geq 1$ ensures*

$$F(\mathbf{x}_{\text{out}}) - F(\mathbf{x}^*) \leq \frac{8\beta R^2}{L+2}$$

where $\mathbf{x}^* \in \operatorname{argmin}_{\mathbf{x}\in\mathcal{K}} F(\mathbf{x})$.

According to Lemma 1, when L is large enough, CG can output a point $\mathbf{x}_{\text{out}}$ such that the approximation error $F(\mathbf{x}_{\text{out}}) - F(\mathbf{x}^*)$ is very small. As a result, with an appropriate $L > 1$, we can minimize our surrogate loss functions more accurately than only performing 1 linear optimization step, which is critical for achieving our desired regret bounds with only sublinear communication complexity. Moreover, CG has been employed to develop projection-free algorithms with improved regret bounds for bandit convex optimization (Garber and Kretzu, 2020, 2021). In this paper, we introduce it into the distributed OCO to propose projection-free algorithms with sublinear communication complexity for the full information and bandit settings, respectively.

Additionally, we introduce a standard technique called one-point gradient estimator (Flaxman et al., 2005), which can approximate the gradient with a single loss value and will

be utilized in the bandit setting. Specifically, for a function $f(\mathbf{x})$ and a constant $\delta > 0$, we can define its δ-smoothed version as

$$\widehat{f}_\delta(\mathbf{x}) = \mathbb{E}_{\mathbf{u}\sim\mathcal{B}^d}[f(\mathbf{x}+\delta\mathbf{u})]$$

where $\mathcal{B}^d$ denotes the unit Euclidean ball centered at the origin in $\mathbb{R}^d$, and notice that it satisfies the following lemma.

Lemma 2 *(Lemma 1 in Flaxman et al. 2005) Let $\mathcal{B}^d$ denote the unit Euclidean ball centered at the origin in $\mathbb{R}^d$, and $\mathcal{S}^d$ denote the unit sphere centered at the origin in $\mathbb{R}^d$. For any function $f(\mathbf{x}) : \mathcal{K} \mapsto \mathbb{R}$ and $\delta > 0$, define its δ-smoothed version as $\widehat{f}_\delta(\mathbf{x}) = \mathbb{E}_{\mathbf{u}\sim\mathcal{B}^d}[f(\mathbf{x}+\delta\mathbf{u})]$. We have*

$$\nabla\widehat{f}_\delta(\mathbf{x}) = \mathbb{E}_{\mathbf{u}\sim\mathcal{S}^d}\left[\frac{d}{\delta}f(\mathbf{x}+\delta\mathbf{u})\mathbf{u}\right]. \tag{7}$$

Lemma 2 provides an unbiased estimator of the gradient $\nabla\widehat{f}_\delta(\mathbf{x})$ by only observing the single value $f(\mathbf{x}+\delta\mathbf{u})$. Note that there also exist two-point and $(d+1)$-point gradient estimators (Agarwal et al., 2010; Duchi et al., 2015), which can approximate the gradient more accurately than the one-point gradient estimator. However, by querying two or $(d+1)$ points per round, in expectation, Agarwal et al. (2010) have recovered the best regret bounds established in the full information setting for both convex and strong convex losses, which actually implies that the bandit setting with two or $(d+1)$ points is almost as simple as the full information setting. So, in this paper, we only consider the most challenging bandit setting, where only one point is available per round. Moreover, we will show that the one-point gradient estimator is sufficient for projection-free algorithms to obtain theoretical guarantees similar to those obtained in the full information setting.

3.3 Assumptions

Next, similar to previous studies on the distributed OCO (Tsianos and Rabbat, 2012; Zhang et al., 2017), we introduce the following assumptions about loss functions, the decision set, and the non-negative weight matrix P that will be utilized to model the communication between local learners.

Assumption 1 *At each round t, each local loss function $f_{t,i}(\mathbf{x})$ is G-Lipschitz over $\mathcal{K}$, i.e., $|f_{t,i}(\mathbf{x}) - f_{t,i}(\mathbf{y})| \leq G\|\mathbf{x}-\mathbf{y}\|_2$ for any $\mathbf{x} \in \mathcal{K}, \mathbf{y} \in \mathcal{K}$.*

Assumption 2 *The convex decision set $\mathcal{K}$ is full dimensional and contains the origin. Moreover, there exist two constants $r, R > 0$ such that $r\mathcal{B}^d \subseteq \mathcal{K} \subseteq R\mathcal{B}^d$ where $\mathcal{B}^d$ denotes the unit Euclidean ball centered at the origin in $\mathbb{R}^d$.*

Assumption 3 *The non-negative weight matrix $P \in \mathbb{R}^{n\times n}$ is supported on the graph $\mathcal{G} = (V, E)$, symmetric, and doubly stochastic, which satisfies*

- *$P_{ij} > 0$ only if $(i, j) \in E$;*
- *$\sum_{j=1}^n P_{ij} = \sum_{j\in N_i} P_{ij} = 1, \forall i \in V$; $\sum_{i=1}^n P_{ij} = \sum_{i\in N_j} P_{ij} = 1, \forall j \in V$.*

Moreover, the second largest singular value of P denoted by $\sigma_2(P)$ is strictly smaller than 1.

Assumption 4 *At each round t, each local loss function $f_{t,i}(\mathbf{x})$ is α-strongly convex over $\mathcal{K}$.*

Note that if Assumption 4 holds with $\alpha = 0$, it reduces to the case with only convex losses.

Moreover, the following assumption is required in the bandit setting (Flaxman et al., 2005; Garber and Kretzu, 2020).

Assumption 5 *At each round t, each local loss function $f_{t,i}(\mathbf{x})$ is bounded over $\mathcal{K}$, i.e., $|f_{t,i}(\mathbf{x})| \leq M$ for any $\mathbf{x} \in \mathcal{K}$. Moreover, all local loss functions are chosen beforehand, i.e., the adversary is oblivious.*

More specifically, according to the one-point gradient estimator (Flaxman et al., 2005), our algorithm in the bandit setting will approximate the gradient $\nabla f_{t,i}(\mathbf{x})$ by

$$\frac{d}{\delta} f_{t,i}(\mathbf{x} + \delta \mathbf{u})\mathbf{u}$$

where $\mathbf{u}$ is uniformly sampled from the unit sphere centered at the origin in $\mathbb{R}^d$ and $\delta > 0$ is a constant such that $\mathbf{x} + \delta \mathbf{u} \in \mathcal{K}$. Then, under Assumption 5, the approximate gradient satisfies the following two facts, which are commonly used for analyzing algorithms in the bandit setting (Flaxman et al., 2005; Garber and Kretzu, 2020).

- First, since each local loss function $f_{t,i}(\mathbf{x})$ is bounded over $\mathcal{K}$, the approximate gradient has a bounded norm.
- Second, since all local loss functions are chosen beforehand, the function $f_{t,i}(\mathbf{x})$ is independent on the random vector $\mathbf{u}$ sampled by our algorithm, which ensures that the approximate gradient satisfies the unbiased property in Lemma 2.

4. Distributed Block Online Conditional Gradient (D-BOCG)

In this section, we present our D-BOCG with the corresponding theoretical guarantees for convex losses and strongly convex losses.

4.1 The Algorithm

First, to reduce the communication complexity of D-OCG, we divide the total T rounds into B blocks of size K, where K is a parameter and we assume that $B = T/K$ is an integer without loss of generality. In this way, each block $m \in [B]$ consists of a set of rounds

$$\mathcal{T}_m = \{(m-1)K + 1, \ldots, mK\}.$$

For each local learner $i \in V$, its decision in each block m stays the same and is denoted by $\mathbf{x}_i(m)$. The local gradient of local learner i in each round $t \in \mathcal{T}_m$ is denoted by

$$\mathbf{g}_i(t) = \nabla f_{t,i}(\mathbf{x}_i(m)).$$

Then, the cumulative gradient of local learner i in each block m is denoted by

$$\widehat{\mathbf{g}}_i(m) = \sum_{t \in \mathcal{T}_m} \mathbf{g}_i(t).$$

Now, we describe how to compute $\mathbf{x}_i(m)$ for each local learner in each block m. Initially, we set $\mathbf{x}_i(1) = \mathbf{x}_{\text{in}}$ for each local learner i, where $\mathbf{x}_{\text{in}}$ is arbitrarily chosen from $\mathcal{K}$. For any

$m > 1$, following Zhang et al. (2017), we can update the decision $\mathbf{x}_i(m)$ by approximately minimizing a surrogate loss function $F_{m-1,i}(\mathbf{x})$. One may adopt a surrogate loss function similar to (4) used by D-OCG. However, it was only designed for convex losses, which cannot utilize the strong convexity. To address this limitation, we define a more general surrogate loss function for α-strongly convex losses, which can utilize the strong convexity for $\alpha > 0$ and cover that used in D-OCG for $\alpha = 0$. To help understanding, we start by introducing our surrogate loss function for the simple case with $n = 1$, and then extend it to the general case for any $n \geq 1$.

In the simple case with $n = 1$, the distributed OCO reduces to the standard OCO, and we only need to define a surrogate loss function $F_{m,1}(\mathbf{x})$ for each block $m \in [B]$. Note that when we assume that all local losses are α-strongly convex, the cumulative loss function $\sum_{t\in\mathcal{T}_m} f_{t,1}(\mathbf{x})$ in each block m is αK-strongly convex, because of $|\mathcal{T}_m| = K$. Therefore, to utilize the strong convexity, inspired by (3), we can define $F_{m,1}(\mathbf{x})$ as

$$F_{m,1}(\mathbf{x}) = \sum_{k=1}^{m-1} \left(\widehat{\mathbf{g}}_1(k)^\top \mathbf{x} + \frac{\alpha K}{2} \|\mathbf{x} - \mathbf{x}_1(k)\|_2^2 \right) + h\|\mathbf{x} - \mathbf{x}_{\text{in}}\|_2^2$$

where h is a parameter that allows us to recover the surrogate loss function (2) for convex losses when $\alpha = 0$, though it does not exist in (3). Since $\|\mathbf{x}_1(k)\|_2^2$ does not affect the minimizer of the function $F_{m,1}(\mathbf{x})$, we further simplify $F_{m,1}(\mathbf{x})$ to

$$F_{m,1}(\mathbf{x}) = \sum_{k=1}^{m-1} \left(\widehat{\mathbf{g}}_1(k) - \alpha K \mathbf{x}_1(k) \right)^\top \mathbf{x} + \frac{(m-1)\alpha K}{2} \|\mathbf{x}\|_2^2 + h\|\mathbf{x} - \mathbf{x}_{\text{in}}\|_2^2.$$

By initializing $\mathbf{z}_1(1) = \mathbf{0}$ and computing $\mathbf{z}_1(m+1)$ as

$$\mathbf{z}_1(m+1) = \mathbf{z}_1(m) + \widehat{\mathbf{g}}_1(m) - \alpha K \mathbf{x}_1(m) \tag{8}$$

we can rewrite the above $F_{m,1}(\mathbf{x})$ as

$$F_{m,1}(\mathbf{x}) = \mathbf{z}_1(m)^\top \mathbf{x} + \frac{(m-1)\alpha K}{2} \|\mathbf{x}\|_2^2 + h\|\mathbf{x} - \mathbf{x}_{\text{in}}\|_2^2.$$

Note that $F_{m,1}(\mathbf{x})$ and $\mathbf{z}_1(m)$ in the simple case only contain the information of the local learner 1, which cannot be used to achieve a regret bound measured by the global loss in the general case. Therefore, inspired by (5) used in D-OCG, in the general case, we update $\mathbf{z}_i(m+1)$ as

$$\mathbf{z}_i(m+1) = \sum_{j\in N_i} P_{ij} \mathbf{z}_j(m) + \widehat{\mathbf{g}}_i(m) - \alpha K \mathbf{x}_i(m) \tag{9}$$

for each local learner i, where P is a non-negative weight matrix satisfying Assumption 3. Different from (8), this update further incorporates the information from the neighbors of local learner i. Moreover, in each block m, the surrogate loss function for each local learner i is defined as

$$F_{m,i}(\mathbf{x}) = \mathbf{z}_i(m)^\top \mathbf{x} + \frac{(m-1)\alpha K}{2} \|\mathbf{x}\|_2^2 + h\|\mathbf{x} - \mathbf{x}_{\text{in}}\|_2^2. \tag{10}$$

Algorithm 2 D-BOCG

1: **Input:** feasible set $\mathcal{K}$, $\mathbf{x}_{\text{in}} \in \mathcal{K}$, α, h, L, and K
2: **Initialization:** choose $\mathbf{x}_1(1) = \cdots = \mathbf{x}_n(1) = \mathbf{x}_{\text{in}}$ and set $\mathbf{z}_1(1) = \cdots = \mathbf{z}_n(1) = \mathbf{0}$
3: **for** $m = 1, \ldots, T/K$ **do**
4: **for** each local learner $i \in V$ **do**
5: define $F_{m,i}(\mathbf{x}) = \mathbf{z}_i(m)^\top \mathbf{x} + \frac{(m-1)\alpha K}{2}\|\mathbf{x}\|_2^2 + h\|\mathbf{x} - \mathbf{x}_{\text{in}}\|_2^2$
6: $\widehat{\mathbf{g}}_i(m) = \mathbf{0}$
7: **for** $t = (m-1)K + 1, \ldots, mK$ **do**
8: play $\mathbf{x}_i(m)$ and observe $\mathbf{g}_i(t) = \nabla f_{t,i}(\mathbf{x}_i(m))$
9: $\widehat{\mathbf{g}}_i(m) = \widehat{\mathbf{g}}_i(m) + \mathbf{g}_i(t)$
10: **end for**
11: $\mathbf{x}_i(m+1) = \text{CG}(\mathcal{K}, L, F_{m,i}(\mathbf{x}), \mathbf{x}_i(m))$ //This step can be executed *in parallel* to the above *for* loop.
12: $\mathbf{z}_i(m+1) = \sum_{j \in N_i} P_{ij}\mathbf{z}_j(m) + \widehat{\mathbf{g}}_i(m) - \alpha K \mathbf{x}_i(m)$
13: **end for**
14: **end for**

Obviously, for convex losses with $\alpha = 0$, this $F_{m,i}(\mathbf{x})$ is equivalent to (4) in D-OCG by setting $K = 1$ and $h = 1/\eta$.

Finally, we need to specify how to compute $\mathbf{x}_i(m+1)$ by approximately minimizing $F_{m,i}(\mathbf{x})$ defined in (10). Similar to the update rules in (6), one may simply perform 1 linear optimization step with the above $F_{m,i}(\mathbf{x})$ to compute $\mathbf{x}_i(m+1)$ for any block m and local learner i. However, this naive update will increase the $O(T^{3/4})$ regret for convex losses established by D-OCG, since the number of updates is decreased. To address this problem, we invoke CG for each update as

$$\mathbf{x}_i(m+1) = \text{CG}(\mathcal{K}, L, F_{m,i}(\mathbf{x}), \mathbf{x}_i(m))$$

where L is an appropriate parameter. The detailed procedures of our algorithm are presented in Algorithm 2, and it is called distributed block online conditional gradient (D-BOCG).

Remark 1 We first note that Algorithm 2 requires $(T/K)L$ linear optimization steps in total. We can limit the total number of linear optimization steps to T by simply setting $L = K$, which is the same as that required by D-OCG (Zhang et al., 2017). Moreover, it is also important to note that at the step 11 in Algorithm 2, the computation about $\mathbf{x}_i(m+1)$ only depends on $\mathbf{x}_i(m)$ and $\mathbf{z}_i(m)$, which are available at the beginning of the block m. Therefore, the step 11 in Algorithm 2 can be executed in parallel to the *for* loop from steps 7 to 10 in Algorithm 2, which implies that L linear optimization steps utilized to compute $\mathbf{x}_i(m+1)$ can be uniformly allocated to all K rounds in the block m, instead of only the last round in the block m. Specifically, Algorithm 2 with $L = K$ only needs to perform 1 linear optimization step in each round, which is computationally as efficient as D-OCG. This parallel strategy is significant, because without it, Algorithm 2 needs to stop at the end of each block m and wait until L linear optimization steps are completed. It has also been utilized by Garber and Kretzu (2020, 2021) when developing improved projection-free algorithms for bandit convex optimization.

4.2 Theoretical Guarantees

In the following, we present theoretical guarantees of our D-BOCG. To help understand the effect of the CG method, we start with the following regret bound, the first term in which clearly depends on the approximation error of the CG method.

Theorem 1 *Under Assumptions 1, 2, 3, and 4, for any $i \in V$, Algorithm 2 ensures*

$$R_{T,i} \leq 3nGK \sum_{m=2}^{B} \sqrt{\frac{2\epsilon_m}{(m-2)\alpha K + 2h}} + 3nGK \sum_{m=2}^{B} \frac{K(G+\alpha R)\sqrt{n}}{((m-2)\alpha K + 2h)(1-\sigma_2(P))} + n \sum_{m=1}^{B} \frac{4K^2(G+2\alpha R)^2}{m\alpha K + 2h} + 4nhR^2$$

where $\epsilon_m = \max_{i\in[n]} \left(F_{m-1,i}(\mathbf{x}_i(m)) - \min_{\mathbf{x}\in\mathcal{K}} F_{m-1,i}(\mathbf{x})\right)$.

Remark 2 At first glance, the regret bound in Theorem 1 seems to be independent on the parameter L of D-BOCG. However, it actually depends on the parameter L through ϵ_m. Moreover, we note that D-BOCG with $K = L = 1$, $h = 1/\eta$, and $\alpha = 0$ reduces to D-OCG (Zhang et al., 2017). In that case, according to the analysis of D-OCG (see the proof of Lemmas 2 and 4 in Zhang et al. 2017 for further details), we can prove that $\epsilon_m/h = O(\sqrt{1/m})$ by setting $h = \Omega(GT^{3/4})$, where the constant G in h is introduced due to the upper bound of $F_{m,i}(\mathbf{x}) - F_{m-1,i}(\mathbf{x})$. If we further consider $L = 1$, $\alpha = 0$, and $K = \sqrt{T}$, we can similarly prove that $\epsilon_m/h = O(\sqrt{1/m})$ by setting $h = \Omega(GKB^{3/4})$, since now the upper bound of $F_{m,i}(\mathbf{x}) - F_{m-1,i}(\mathbf{x})$ in on the order of $O(GK)$ and the maximum m is changed from T to B. However, in this case, the first term in the above regret bound will satisfy

$$3nGK \sum_{m=2}^{B} \sqrt{\frac{2\epsilon_m}{2h}} = O\left(K \sum_{m=2}^{B} \frac{1}{m^{1/4}}\right) = O\left(KB^{3/4}\right) = O\left(T^{7/8}\right)$$

which is worse than $O(T^{3/4})$. Therefore, to keep the $O(T^{3/4})$ regret for convex losses with $K = \sqrt{T}$, we need to use more linear optimization steps. According to Lemma 1, if L is large enough, the approximation error ϵ_m could be very small. More specifically, by combining Theorem 1 with Lemma 1, we have the following theorem.

Theorem 2 *Under Assumptions 1, 2, 3, and 4, for any $i \in V$, Algorithm 2 ensures*

$$R_{T,i} \leq \frac{12nGRT}{\sqrt{L+2}} + \sum_{m=2}^{B} \frac{3nGK^2(G+\alpha R)\sqrt{n}}{((m-2)\alpha K + 2h)(1-\sigma_2(P))} + \sum_{m=1}^{B} \frac{4nK^2(G+2\alpha R)^2}{m\alpha K + 2h} + 4nhR^2.$$

Remark 3 Note that Theorem 2 with $K = L = 1$, $h = 1/\eta$, and $\alpha = 0$ cannot recover the $O(T^{3/4})$ regret bound of D-OCG, because $\frac{12nGRT}{\sqrt{L+2}}$ would be on the order of $O(T)$. The main reason is that for the CG method, the approximation error bound in Lemma 1 is too loose when $L = 1$. According to Zhang et al. (2017), instead of using Lemma 1, a more complicated analysis is required for bounding the approximation error when only utilizing 1 linear optimization step. To recover the $O(T^{3/4})$ regret bound with $K = L = 1$, $h = 1/\eta$,

and $\alpha = 0$, one potential way is to extend the analysis of Zhang et al. (2017) from the case with $L = 1$ to the case with any L. However, we find that this extension is highly non-trivial, and notice that Theorem 2 is sufficient to achieve our desired regret bounds and communication complexity.

For convex losses, we can simplify Theorem 2 to the following corollary.

Corollary 1 *Under Assumptions 1, 2, 3, and 4 with $\alpha = 0$, for any $i \in V$, Algorithm 2 with $\alpha = 0$, $K = L = \sqrt{T}$, and $h = \frac{n^{1/4}T^{3/4}G}{\sqrt{1-\sigma_2(P)}R}$ has*

$$R_{T,i} \leq \left(12n + 2\sqrt{1-\sigma_2(P)}n^{3/4} + \frac{11}{2}n^{5/4}(1-\sigma_2(P))^{-1/2}\right) GRT^{3/4}.$$

Remark 4 The above corollary shows that our D-BOCG can enjoy the $O(T^{3/4})$ regret bound with only $O(\sqrt{T})$ communication rounds for convex losses. By contrast, D-OCG (Zhang et al., 2017) obtains the $O(T^{3/4})$ regret bound with a larger number of T communication rounds for convex losses.

Moreover, for strongly convex losses, we can simplify Theorem 2 to the following corollary.

Corollary 2 *Under Assumptions 1, 2, 3, and 4 with $\alpha > 0$, for any $i \in V$, Algorithm 2 with $\alpha > 0$, $K = L = T^{2/3}(\ln T)^{-2/3}$, and $h = \alpha K$ ensures*

$$\begin{aligned} R_{T,i} \leq & \left(\frac{3n^{3/2}G(G+\alpha R)}{\alpha(1-\sigma_2(P))} + \frac{4n(G+2\alpha R)^2}{\alpha}\right) T^{2/3}((\ln T)^{-2/3} + (\ln T)^{1/3}) \\ & + 12nGRT^{2/3}(\ln T)^{1/3} + 4n\alpha R^2 T^{2/3}(\ln T)^{-2/3}. \end{aligned}$$

Remark 5 The above corollary shows that our D-BOCG can enjoy a regret bound of $O(T^{2/3}(\log T)^{1/3})$ with $O(T^{1/3}(\log T)^{2/3})$ communication rounds for strongly convex losses. Compared with Corollary 1, both the regret and communication complexity of our D-BOCG are improved by utilizing the strong convexity.

Remark 6 Besides the dependence on T, Corollaries 1 and 2 also explicitly show how the regret of our D-BOCG depends on the network size n and the spectral gap $1 - \sigma_2(P)$. First, the dependence on n shows that the regret of our D-BOCG will be larger on larger networks for convex losses and strongly convex losses. Second, the spectral gap actually reflects the connectivity of the network: a larger spectral gap value implies better connectivity (Duchi et al., 2011; Zhang et al., 2017). Therefore, the dependence on the spectral gap implies that the regret of our D-BOCG will be smaller on "well connected" networks than on "poorly connected" networks for convex losses and strongly convex losses. More specifically, with a particular choice of the matrix P, Duchi et al. (2011) have bounded the spectral gap for several classes of interesting networks, such as $1 - \sigma_2(P) = \Omega(1)$ for expanders and the complete graph, and $1 - \sigma_2(P) = \Omega(1/n^2)$ for the cycle graph (see Section 3.2 in Duchi et al. 2011 for details). By replacing the dependence on $1 - \sigma_2(P)$ with that on n, our Corollaries 1 and 2 further imply that:

- in the case with convex losses, the regret of D-BOCG can be bounded by $O(n^{5/4}T^{3/4})$ for "well connected" networks and $O(n^{9/4}T^{3/4})$ for "poorly connected" networks;
- in the case with strongly convex losses, the regret of D-BOCG can be bounded by $O(n^{3/2}T^{2/3}(\log T)^{1/3})$ for "well connected" networks and $O(n^{7/2}T^{2/3}(\log T)^{1/3})$ for "poorly connected" networks.

4.3 Analysis

In this section, we only provide the proof of Theorems 1 and 2. The proof of Corollaries 1 and 2 can be found in the Appendix.

4.3.1 Proof of Theorem 1

We first notice that if $f(\mathbf{x}) : \mathcal{K} \to \mathbb{R}$ is α-strongly convex over $\mathcal{K}$, according to Hazan and Kale (2012), it holds that

$$\frac{\alpha}{2}\|\mathbf{x} - \mathbf{x}^*\|_2^2 \leq f(\mathbf{x}) - f(\mathbf{x}^*) \tag{11}$$

for any $\mathbf{x} \in \mathcal{K}$ and $\mathbf{x}^* = \operatorname{argmin}_{\mathbf{x}\in\mathcal{K}} f(\mathbf{x})$. This property about strongly convex functions will be utilized in the following.

Then, we define several auxiliary variables. Let $\bar{\mathbf{z}}(m) = \frac{1}{n}\sum_{i=1}^n \mathbf{z}_i(m)$ for $m \in [B+1]$, and let $\mathbf{d}_i(m) = \widehat{\mathbf{g}}_i(m) - \alpha K \mathbf{x}_i(m)$ and $\bar{\mathbf{d}}(m) = \frac{1}{n}\sum_{i=1}^n \mathbf{d}_i(m)$ for $m \in [B]$. According to Assumption 3, we have

$$\begin{aligned}\bar{\mathbf{z}}(m+1) =& \frac{1}{n}\sum_{i=1}^n \mathbf{z}_i(m+1) = \frac{1}{n}\sum_{i=1}^n \left(\sum_{j\in N_i} P_{ij}\mathbf{z}_j(m) + \widehat{\mathbf{g}}_i(m) - \alpha K \mathbf{x}_i(m)\right) \\ =& \frac{1}{n}\sum_{i=1}^n\sum_{j=1}^n P_{ij}\mathbf{z}_j(m) + \bar{\mathbf{d}}(m) = \frac{1}{n}\sum_{j=1}^n\sum_{i=1}^n P_{ij}\mathbf{z}_j(m) + \bar{\mathbf{d}}(m) = \bar{\mathbf{z}}(m) + \bar{\mathbf{d}}(m).\end{aligned} \tag{12}$$

Moreover, we define $\bar{\mathbf{x}}(1) = \mathbf{x}_{\text{in}}$ and $\bar{\mathbf{x}}(m+1) = \operatorname{argmin}_{\mathbf{x}\in\mathcal{K}} \bar{F}_m(\mathbf{x})$ for any $m \in [B+1]$, where

$$\bar{F}_m(\mathbf{x}) = \bar{\mathbf{z}}(m)^\top \mathbf{x} + \frac{(m-1)\alpha K}{2}\|\mathbf{x}\|_2^2 + h\|\mathbf{x} - \mathbf{x}_{\text{in}}\|_2^2.$$

Similarly, we define $\widehat{\mathbf{x}}_i(m+1) = \operatorname{argmin}_{\mathbf{x}\in\mathcal{K}} F_{m,i}(\mathbf{x})$ for any $m \in [B+1]$, where

$$F_{m,i}(\mathbf{x}) = \mathbf{z}_i(m)^\top \mathbf{x} + \frac{(m-1)\alpha K}{2}\|\mathbf{x}\|_2^2 + h\|\mathbf{x} - \mathbf{x}_{\text{in}}\|_2^2.$$

is defined in Algorithm 2.

Then, we derive an upper bound of $\|\mathbf{x}_i(m) - \bar{\mathbf{x}}(m)\|_2$ for any $m \in [B]$. If $m = 1$, according to the definition and Algorithm 2, it is easy to verify that

$$\|\mathbf{x}_i(m) - \bar{\mathbf{x}}(m)\|_2 = 0. \tag{13}$$

For any $B \geq m \geq 2$, due to $\epsilon_m = \max_{i\in[n]} \left(F_{m-1,i}(\mathbf{x}_i(m)) - \min_{\mathbf{x}\in\mathcal{K}} F_{m-1,i}(\mathbf{x})\right)$, we have

$$\begin{aligned}\|\mathbf{x}_i(m) - \bar{\mathbf{x}}(m)\|_2 \leq & \|\mathbf{x}_i(m) - \widehat{\mathbf{x}}_i(m)\|_2 + \|\widehat{\mathbf{x}}_i(m) - \bar{\mathbf{x}}(m)\|_2 \\ \leq & \sqrt{\frac{2F_{m-1,i}(\mathbf{x}_i(m)) - 2F_{m-1,i}(\widehat{\mathbf{x}}_i(m))}{(m-2)\alpha K + 2h}} + \|\widehat{\mathbf{x}}_i(m) - \bar{\mathbf{x}}(m)\|_2 \\ \leq & \sqrt{\frac{2\epsilon_m}{(m-2)\alpha K + 2h}} + \|\widehat{\mathbf{x}}_i(m) - \bar{\mathbf{x}}(m)\|_2\end{aligned} \tag{14}$$

where the second inequality is due to the fact that $F_{m-1,i}(\mathbf{x})$ is $((m-2)\alpha K + 2h)$-strongly convex and (11).

To further bound $\|\widehat{\mathbf{x}}_i(m) - \bar{\mathbf{x}}(m)\|_2$ in (14), we introduce the following two lemmas.

Lemma 3 *(Derived from the Proof of Lemma 6 in Zhang et al. 2017) For any $i \in [n]$, let $\mathbf{d}_i(1), \ldots, \mathbf{d}_i(m) \in \mathbb{R}^d$ be a sequence of vectors. Let $\mathbf{z}_i(1) = \mathbf{0}$, $\mathbf{z}_i(m+1) = \sum_{j\in N_i} P_{ij}\mathbf{z}_j(m) + \mathbf{d}_i(m)$, and $\bar{\mathbf{z}}(m) = \frac{1}{n}\sum_{i=1}^n \mathbf{z}_i(m)$ for $m \in [B]$, where P satisfies Assumption 3. For any $i \in [n]$ and $m \in [B]$, assuming $\|\mathbf{d}_i(m)\|_2 \leq \widehat{G}$ where $\widehat{G} > 0$ is a constant, we have*

$$\|\mathbf{z}_i(m) - \bar{\mathbf{z}}(m)\|_2 \leq \frac{\widehat{G}\sqrt{n}}{1 - \sigma_2(P)}.$$

Lemma 4 *(Lemma 5 in Duchi et al. 2011) Let $\Pi_{\mathcal{K}}(\mathbf{u}, \eta) = \operatorname{argmin}_{\mathbf{x}\in\mathcal{K}} \eta\mathbf{u}^\top\mathbf{x} + \|\mathbf{x}\|_2^2$. We have*

$$\|\Pi_{\mathcal{K}}(\mathbf{u}, \eta) - \Pi_{\mathcal{K}}(\mathbf{v}, \eta)\|_2 \leq \frac{\eta}{2}\|\mathbf{u} - \mathbf{v}\|_2.$$

According to Assumptions 1 and 2, for any $m \in [B]$, we have

$$\begin{aligned}\|\mathbf{d}_i(m)\|_2 = & \|\widehat{\mathbf{g}}_i(m) - \alpha K\mathbf{x}_i(m)\|_2 = \left\|\sum_{t\in\mathcal{T}_m} \mathbf{g}_i(t) - \alpha K\mathbf{x}_i(m)\right\|_2 \\ \leq & \sum_{t\in\mathcal{T}_m} \|\mathbf{g}_i(t)\|_2 + \alpha K\|\mathbf{x}_i(m)\|_2 \leq K(G + \alpha R)\end{aligned} \tag{15}$$

where $\mathcal{T}_m = \{(m-1)K + 1, \ldots, mK\}$.

By applying Lemma 3 with $\|\mathbf{d}_i(m)\|_2 \leq K(G + \alpha R)$, for any $B \in [m]$, we have

$$\|\mathbf{z}_i(m) - \bar{\mathbf{z}}(m)\|_2 \leq \frac{K(G + \alpha R)\sqrt{n}}{1 - \sigma_2(P)}. \tag{16}$$

Moreover, for any $B \geq m \geq 2$, we notice that

$$\begin{aligned}\widehat{\mathbf{x}}_i(m) = & \operatorname*{argmin}_{\mathbf{x}\in\mathcal{K}} \mathbf{z}_i(m-1)^\top\mathbf{x} + \frac{(m-2)\alpha K}{2}\|\mathbf{x}\|_2^2 + h\|\mathbf{x} - \mathbf{x}_{\text{in}}\|_2^2 \\ = & \operatorname*{argmin}_{\mathbf{x}\in\mathcal{K}} (\mathbf{z}_i(m-1) - 2h\mathbf{x}_{\text{in}})^\top\mathbf{x} + \frac{(m-2)\alpha K + 2h}{2}\|\mathbf{x}\|_2^2 \\ = & \operatorname*{argmin}_{\mathbf{x}\in\mathcal{K}} \frac{2}{(m-2)\alpha K + 2h}(\mathbf{z}_i(m-1) - 2h\mathbf{x}_{\text{in}})^\top\mathbf{x} + \|\mathbf{x}\|_2^2.\end{aligned} \tag{17}$$

Similarly, for any $B \geq m \geq 2$, we have

$$\begin{aligned}\bar{\mathbf{x}}(m) &= \operatorname*{argmin}_{\mathbf{x}\in\mathcal{K}} \bar{\mathbf{z}}(m-1)^\top \mathbf{x} + \frac{(m-2)\alpha K}{2}\|\mathbf{x}\|_2^2 + h\|\mathbf{x}-\mathbf{x}_{\text{in}}\|_2^2 \\ &= \operatorname*{argmin}_{\mathbf{x}\in\mathcal{K}} (\bar{\mathbf{z}}(m-1) - 2h\mathbf{x}_{\text{in}})^\top \mathbf{x} + \frac{(m-2)\alpha K + 2h}{2}\|\mathbf{x}\|_2^2 \\ &= \operatorname*{argmin}_{\mathbf{x}\in\mathcal{K}} \frac{2}{(m-2)\alpha K + 2h}(\bar{\mathbf{z}}(m-1) - 2h\mathbf{x}_{\text{in}})^\top \mathbf{x} + \|\mathbf{x}\|_2^2.\end{aligned} \tag{18}$$

Therefore, by combining Lemma 4 with (16), for any $B \geq m \geq 2$, we have

$$\begin{aligned}\|\widehat{\mathbf{x}}_i(m) - \bar{\mathbf{x}}(m)\|_2 &\leq \frac{1}{(m-2)\alpha K + 2h}\|\mathbf{z}_i(m-1) - 2h\mathbf{x}_{\text{in}} - \bar{\mathbf{z}}(m-1) + 2h\mathbf{x}_{\text{in}}\|_2 \\ &= \frac{1}{(m-2)\alpha K + 2h}\|\mathbf{z}_i(m-1) - \bar{\mathbf{z}}(m-1)\|_2 \\ &\leq \frac{K(G+\alpha R)\sqrt{n}}{((m-2)\alpha K + 2h)(1-\sigma_2(P))}.\end{aligned}$$

By substituting the above inequality into (14), for any $B \geq m \geq 2$, we have

$$\|\mathbf{x}_i(m) - \bar{\mathbf{x}}(m)\|_2 \leq \sqrt{\frac{2\epsilon_m}{(m-2)\alpha K + 2h}} + \frac{K(G+\alpha R)\sqrt{n}}{((m-2)\alpha K + 2h)(1-\sigma_2(P))}. \tag{19}$$

Let $u_1 = 0$ and $u_m = \sqrt{\frac{2\epsilon_m}{(m-2)\alpha K+2h}} + \frac{K(G+\alpha R)\sqrt{n}}{((m-2)\alpha K+2h)(1-\sigma_2(P))}$ for any $B \geq m \geq 2$. From (13) and (19), for any $m \in [B]$, it holds that $\|\mathbf{x}_i(m) - \bar{\mathbf{x}}(m)\|_2 \leq u_m$.

Let $\mathbf{x}^* \in \operatorname{argmin}_{\mathbf{x}\in\mathcal{K}} \sum_{t=1}^T f_t(\mathbf{x})$. For any $i, j \in V$, $m \in [B]$, and $t \in \mathcal{T}_m$, according to Assumptions 1 and 4, we have

$$\begin{aligned}&f_{t,j}(\mathbf{x}_i(m)) - f_{t,j}(\mathbf{x}^*) \\ \leq& f_{t,j}(\bar{\mathbf{x}}(m)) + G\|\bar{\mathbf{x}}(m) - \mathbf{x}_i(m)\|_2 - f_{t,j}(\mathbf{x}^*) \\ \leq& f_{t,j}(\mathbf{x}_j(m)) + G\|\bar{\mathbf{x}}(m) - \mathbf{x}_j(m)\|_2 - f_{t,j}(\mathbf{x}^*) + Gu_m \\ \leq& \nabla f_{t,j}(\mathbf{x}_j(m))^\top(\mathbf{x}_j(m) - \mathbf{x}^*) - \frac{\alpha}{2}\|\mathbf{x}_j(m) - \mathbf{x}^*\|_2^2 + 2Gu_m \\ =& \nabla f_{t,j}(\mathbf{x}_j(m))^\top(\mathbf{x}_j(m) - \bar{\mathbf{x}}(m)) + \nabla f_{t,j}(\mathbf{x}_j(m))^\top(\bar{\mathbf{x}}(m) - \mathbf{x}^*) - \frac{\alpha}{2}\|\mathbf{x}_j(m) - \mathbf{x}^*\|_2^2 + 2Gu_m \\ \leq& \nabla f_{t,j}(\mathbf{x}_j(m))^\top(\bar{\mathbf{x}}(m) - \mathbf{x}^*) - \frac{\alpha}{2}\|\mathbf{x}_j(m) - \mathbf{x}^*\|_2^2 + 3Gu_m\end{aligned}$$

where the third inequality is due to the strong convexity of $f_{t,j}(\mathbf{x})$ and the last inequality is due to

$$\nabla f_{t,j}(\mathbf{x}_j(m))^\top(\mathbf{x}_j(m) - \bar{\mathbf{x}}(m)) \leq \|\nabla f_{t,j}(\mathbf{x}_j(m))\|_2\|\mathbf{x}_j(m) - \bar{\mathbf{x}}(m)\|_2 \leq Gu_m.$$

Moreover, we note that

$$\begin{aligned}\|\mathbf{x}_j(m) - \mathbf{x}^*\|_2^2 =& \|\mathbf{x}_j(m) - \bar{\mathbf{x}}(m)\|_2^2 + 2\mathbf{x}_j(m)^\top(\bar{\mathbf{x}}(m) - \mathbf{x}^*) + \|\mathbf{x}^*\|_2^2 - \|\bar{\mathbf{x}}(m)\|_2^2 \\ \geq& 2\mathbf{x}_j(m)^\top(\bar{\mathbf{x}}(m) - \mathbf{x}^*) + \|\mathbf{x}^*\|_2^2 - \|\bar{\mathbf{x}}(m)\|_2^2.\end{aligned}$$

Therefore, for any $i, j \in V$, $m \in [B]$, and $t \in \mathcal{T}_m$, we have

$$\begin{aligned}&f_{t,j}(\mathbf{x}_i(m)) - f_{t,j}(\mathbf{x}^*) - 3Gu_m\\ \leq&(\nabla f_{t,j}(\mathbf{x}_j(m)) - \alpha\mathbf{x}_j(m))^\top(\bar{\mathbf{x}}(m) - \mathbf{x}^*) - \frac{\alpha}{2}(\|\mathbf{x}^*\|_2^2 - \|\bar{\mathbf{x}}(m)\|_2^2)\end{aligned}$$

By summing over $t \in \mathcal{T}_m$ and $m \in [B]$, for any $i, j \in V$, we have

$$\begin{aligned}&\sum_{m=1}^{B}\sum_{t\in\mathcal{T}_m}(f_{t,j}(\mathbf{x}_i(m)) - f_{t,j}(\mathbf{x}^*)) - 3G\sum_{m=1}^{B}\sum_{t\in\mathcal{T}_m}u_m\\ \leq&\sum_{m=1}^{B}\sum_{t\in\mathcal{T}_m}(\nabla f_{t,j}(\mathbf{x}_j(m)) - \alpha\mathbf{x}_j(m))^\top(\bar{\mathbf{x}}(m) - \mathbf{x}^*) - \sum_{m=1}^{B}\sum_{t\in\mathcal{T}_m}\frac{\alpha}{2}(\|\mathbf{x}^*\|_2^2 - \|\bar{\mathbf{x}}(m)\|_2^2)\\ =&\sum_{m=1}^{B}(\widehat{\mathbf{g}}_j(m) - \alpha K\mathbf{x}_j(m))^\top(\bar{\mathbf{x}}(m) - \mathbf{x}^*) - \sum_{m=1}^{B}\frac{\alpha K}{2}(\|\mathbf{x}^*\|_2^2 - \|\bar{\mathbf{x}}(m)\|_2^2).\end{aligned}$$

Furthermore, by summing over $j = 1, \ldots, n$, for any $i \in V$, we have

$$\begin{aligned}&\sum_{m=1}^{B}\sum_{t\in\mathcal{T}_m}\sum_{j=1}^{n}(f_{t,j}(\mathbf{x}_i(m)) - f_{t,j}(\mathbf{x}^*)) - 3G\sum_{m=1}^{B}\sum_{t\in\mathcal{T}_m}\sum_{j=1}^{n}u_m\\ \leq&\sum_{m=1}^{B}\sum_{j=1}^{n}(\widehat{\mathbf{g}}_j(m) - \alpha K\mathbf{x}_j(m))^\top(\bar{\mathbf{x}}(m) - \mathbf{x}^*) - \frac{\alpha nK}{2}\sum_{m=1}^{B}(\|\mathbf{x}^*\|_2^2 - \|\bar{\mathbf{x}}(m)\|_2^2)\\ =&n\sum_{m=1}^{B}\left(\bar{\mathbf{d}}(m)^\top(\bar{\mathbf{x}}(m) - \mathbf{x}^*) - \frac{\alpha K}{2}(\|\mathbf{x}^*\|_2^2 - \|\bar{\mathbf{x}}(m)\|_2^2)\right).\end{aligned}$$

Then, it is easy to verify that

$$\begin{aligned}R_{T,i} =&\sum_{m=1}^{B}\sum_{t\in\mathcal{T}_m}\sum_{j=1}^{n}(f_{t,j}(\mathbf{x}_i(m)) - f_{t,j}(\mathbf{x}^*))\\ \leq&n\sum_{m=1}^{B}\left(\bar{\mathbf{d}}(m)^\top(\bar{\mathbf{x}}(m) - \mathbf{x}^*) - \frac{\alpha K}{2}(\|\mathbf{x}^*\|_2^2 - \|\bar{\mathbf{x}}(m)\|_2^2)\right) + 3G\sum_{m=1}^{B}\sum_{t\in\mathcal{T}_m}\sum_{j=1}^{n}u_m \quad (20)\\ =&n\sum_{m=1}^{B}\left(\bar{\mathbf{d}}(m)^\top(\bar{\mathbf{x}}(m) - \mathbf{x}^*) - \frac{\alpha K}{2}(\|\mathbf{x}^*\|_2^2 - \|\bar{\mathbf{x}}(m)\|_2^2)\right) + 3nGK\sum_{m=1}^{B}u_m.\end{aligned}$$

To bound $\sum_{m=1}^{B}\left(\bar{\mathbf{d}}(m)^\top(\bar{\mathbf{x}}(m) - \mathbf{x}^*) - \frac{\alpha K}{2}(\|\mathbf{x}^*\|_2^2 - \|\bar{\mathbf{x}}(m)\|_2^2)\right)$, we introduce the following lemma.

Lemma 5 *(Lemma 2.3 in Shalev-Shwartz 2011) Let* $\widehat{\mathbf{x}}_t^* = \operatorname{argmin}_{\mathbf{x}\in\mathcal{K}}\sum_{i=1}^{t-1}f_i(\mathbf{x}) + \mathcal{R}(\mathbf{x})$ *for* $t \in [T]$*, where* $\mathcal{R}(\mathbf{x})$ *is a strongly convex function. Then,* $\forall\mathbf{x} \in \mathcal{K}$*, it holds that*

$$\sum_{t=1}^{T}(f_t(\widehat{\mathbf{x}}_t^*) - f_t(\mathbf{x})) \leq \mathcal{R}(\mathbf{x}) - \mathcal{R}(\widehat{\mathbf{x}}_1^*) + \sum_{t=1}^{T}\left(f_t(\widehat{\mathbf{x}}_t^*) - f_t(\widehat{\mathbf{x}}_{t+1}^*)\right).$$

Before applying Lemma 5, we define $\widetilde{f}_m(\mathbf{x}) = \bar{\mathbf{d}}(m)^\top \mathbf{x} + \frac{\alpha K}{2}\|\mathbf{x}\|_2^2$. For any $\mathbf{x} \in \mathcal{K}$, it is easy to verify that

$$\|\nabla \widetilde{f}_m(\mathbf{x})\|_2 = \|\bar{\mathbf{d}}(m) + \alpha K \mathbf{x}\|_2 \leq \left\| \frac{1}{n}\sum_{j=1}^{n} \mathbf{d}_j(m) \right\|_2 + \|\alpha K \mathbf{x}\|_2 \leq K(G + 2\alpha R) \tag{21}$$

where the last inequality is due to Assumption 2 and (15).

Moreover, according to the definition and (12), for any $m \in [B]$, we have

$$\bar{\mathbf{x}}(m+1) = \operatorname*{argmin}_{\mathbf{x}\in\mathcal{K}} \bar{\mathbf{z}}(m)^\top \mathbf{x} + \frac{(m-1)\alpha K}{2}\|\mathbf{x}\|_2^2 + h\|\mathbf{x} - \mathbf{x}_{\text{in}}\|_2^2 = \operatorname*{argmin}_{\mathbf{x}\in\mathcal{K}} \sum_{\tau=1}^{m-1} \widetilde{f}_\tau(\mathbf{x}) + h\|\mathbf{x} - \mathbf{x}_{\text{in}}\|_2^2.$$

By applying Lemma 5 with the loss functions $\{\widetilde{f}_m(\mathbf{x})\}_{m=1}^B$, the decision set $\mathcal{K}$, and the regularizer $\mathcal{R}(\mathbf{x}) = h\|\mathbf{x} - \mathbf{x}_{\text{in}}\|_2^2$, we have

$$\begin{aligned}
&\sum_{m=1}^{B} \left(\widetilde{f}_m(\bar{\mathbf{x}}(m+1)) - \widetilde{f}_m(\mathbf{x}^*) \right) \\
&\leq h\|\mathbf{x}^* - \mathbf{x}_{\text{in}}\|_2^2 - h\|\bar{\mathbf{x}}(2) - \mathbf{x}_{\text{in}}\|_2^2 + \sum_{m=1}^{B} \left(\widetilde{f}_m(\bar{\mathbf{x}}(m+1)) - \widetilde{f}_m(\bar{\mathbf{x}}(m+2)) \right) \\
&\leq 4hR^2 + \sum_{m=1}^{B} \nabla \widetilde{f}_m(\bar{\mathbf{x}}(m+1))^\top (\bar{\mathbf{x}}(m+1) - \bar{\mathbf{x}}(m+2)) \\
&\leq 4hR^2 + \sum_{m=1}^{B} K(G + 2\alpha R)\|\bar{\mathbf{x}}(m+1) - \bar{\mathbf{x}}(m+2)\|_2
\end{aligned} \tag{22}$$

where the last inequality is due to the Cauchy-Schwarz inequality and (21).

Note that $\bar{F}_{m+1}(\mathbf{x})$ is $(m\alpha K + 2h)$-strongly convex and $\bar{\mathbf{x}}(m+2) = \operatorname{argmin}_{\mathbf{x}\in\mathcal{K}} \bar{F}_{m+1}(\mathbf{x})$. For any $m \in [B]$, we have

$$\begin{aligned}
&\frac{m\alpha K + 2h}{2}\|\bar{\mathbf{x}}(m+1) - \bar{\mathbf{x}}(m+2)\|_2^2 \\
\leq &\bar{F}_{m+1}(\bar{\mathbf{x}}(m+1)) - \bar{F}_{m+1}(\bar{\mathbf{x}}(m+2)) \\
= &\bar{F}_m(\bar{\mathbf{x}}(m+1)) + \widetilde{f}_m(\bar{\mathbf{x}}(m+1)) - \bar{F}_m(\bar{\mathbf{x}}(m+2)) - \widetilde{f}_m(\bar{\mathbf{x}}(m+2)) \\
\leq &\widetilde{f}_m(\bar{\mathbf{x}}(m+1)) - \widetilde{f}_m(\bar{\mathbf{x}}(m+2)) \\
\leq &\nabla \widetilde{f}_m(\bar{\mathbf{x}}(m+1))^\top (\bar{\mathbf{x}}(m+1) - \bar{\mathbf{x}}(m+2)) \\
\leq &K(G + 2\alpha R)\|\bar{\mathbf{x}}(m+1) - \bar{\mathbf{x}}(m+2)\|_2
\end{aligned}$$

where the first inequality is due to (11), the second inequality is due to $\bar{\mathbf{x}}(m+1) = \operatorname{argmin}_{\mathbf{x}\in\mathcal{K}} \bar{F}_m(\mathbf{x})$, and the last inequality is due to the Cauchy-Schwarz inequality and (21).

For any $m \in [B]$, the above equality can be simplified as

$$\|\bar{\mathbf{x}}(m+1) - \bar{\mathbf{x}}(m+2)\|_2 \leq \frac{2K(G + 2\alpha R)}{m\alpha K + 2h}.$$

Then, by combining the above inequality with (22), we have

$$\begin{aligned}
&\sum_{m=1}^{B}\left(\bar{\mathbf{d}}(m)^\top(\bar{\mathbf{x}}(m)-\mathbf{x}^*)-\frac{\alpha K}{2}(\|\mathbf{x}^*\|_2^2-\|\bar{\mathbf{x}}(m)\|_2^2)\right)\\
=&\sum_{m=1}^{B}\left(\widetilde{f}_m(\bar{\mathbf{x}}(m))-\widetilde{f}_m(\mathbf{x}^*)\right)\\
=&\sum_{m=1}^{B}\left(\widetilde{f}_m(\bar{\mathbf{x}}(m))-\widetilde{f}_m(\bar{\mathbf{x}}(m+1))\right)+\sum_{m=1}^{B}\left(\widetilde{f}_m(\bar{\mathbf{x}}(m+1))-\widetilde{f}_m(\mathbf{x}^*)\right)\\
\leq&K(G+2\alpha R)\sum_{m=1}^{B}(\|\bar{\mathbf{x}}(m)-\bar{\mathbf{x}}(m+1)\|_2+\|\bar{\mathbf{x}}(m+1)-\bar{\mathbf{x}}(m+2)\|_2)+4hR^2\\
\leq&2K(G+2\alpha R)\sum_{m=1}^{B}\|\bar{\mathbf{x}}(m+1)-\bar{\mathbf{x}}(m+2)\|_2+4hR^2\\
\leq&\sum_{m=1}^{B}\frac{4K^2(G+2\alpha R)^2}{m\alpha K+2h}+4hR^2
\end{aligned}\tag{23}$$

where the second inequality is due to $\bar{\mathbf{x}}(2)=\operatorname{argmin}_{\mathbf{x}\in\mathcal{K}}\bar{F}_1(\mathbf{x})=\mathbf{x}_{\text{in}}=\bar{\mathbf{x}}(1)$ and $\|\bar{\mathbf{x}}(1)-\bar{\mathbf{x}}(2)\|_2=0\leq\|\bar{\mathbf{x}}(B+1)-\bar{\mathbf{x}}(B+2)\|_2$.

Finally, we complete the proof by substituting the definition of u_m and (23) into (20).

4.3.2 Proof of Theorem 2

We note that $F_{m-1,i}(\mathbf{x})$ is $((m-2)\alpha K+2h)$-smooth, and according to our Algorithm 2, we have

$$\mathbf{x}_i(m)=\text{CG}(\mathcal{K},L,F_{m-1,i}(\mathbf{x}),\mathbf{x}_i(m-1)).$$

Therefore, for any $B\geq m\geq 2$, by applying Lemma 1, it is easy to verify that

$$\epsilon_m=\max_{i\in[n]}\left(F_{m-1,i}(\mathbf{x}_i(m))-\min_{\mathbf{x}\in\mathcal{K}}F_{m-1,i}(\mathbf{x})\right)\leq\frac{8((m-2)\alpha K+2h)R^2}{L+2}.$$

By substituting the above inequality and $K(B-1)\leq KB=T$ into Theorem 1, we have

$$R_{T,i}\leq\frac{12nGRT}{\sqrt{L+2}}+\sum_{m=2}^{B}\frac{3nGK^2(G+\alpha R)\sqrt{n}}{((m-2)\alpha K+2h)(1-\sigma_2(P))}+\sum_{m=1}^{B}\frac{4nK^2(G+2\alpha R)^2}{m\alpha K+2h}+4nhR^2.$$

5. Lower Bounds

In this section, we present lower bounds regarding the communication complexity for convex losses and strongly convex losses, respectively.

5.1 Convex Losses

Following previous studies (Hosseini et al., 2013; Zhang et al., 2017), when developing distributed online algorithms, we need to upper bound the regret of all local learners

simultaneously. Correspondingly, to establish a lower regret bound for these distributed online algorithms, we actually only need to prove that the regret of one local learner has a lower bound. For simplicity, in the following, we will consider to derive a lower regret bound for the local learner 1.

For convex losses, we present a lower bound in the following theorem.

Theorem 3 *Suppose* $\mathcal{K} = \left[-R/\sqrt{d}, R/\sqrt{d}\right]^d$, *which satisfies Assumption 2 with constants* R *and* $r = R/\sqrt{d}$. *For distributed OCO with* $n > 1$ *local learners over* $\mathcal{K}$ *and any distributed online algorithm communicating at the end of* C *rounds before the round* T, *there exists a sequence of local loss functions satisfying Assumption 1 with a constant* G *such that*

$$R_{T,1} \geq \frac{nRGT}{2\sqrt{2(C+1)}}.$$

Proof In each round t, we simply set $f_{t,1}(\mathbf{x}) = 0$ for the local learner 1, and select $f_{t,2}(\mathbf{x}), \ldots, f_{t,n}(\mathbf{x})$ for other local learners with a more careful strategy. According to this setting, the global loss function is

$$f_t(\mathbf{x}) = f_{t,1}(\mathbf{x}) + \sum_{i=2}^{n} f_{t,i}(\mathbf{x}) = \sum_{i=2}^{n} f_{t,i}(\mathbf{x}).$$

Note that the local loss function $f_{t,i}(\mathbf{x})$ is only revealed to the local learner $i \in [n]$, which implies that the local learner 1 cannot access to the global loss unless it communicates with other local learners. Therefore, we can utilize this setting to maximize the impact of communication on the regret of the local learner 1.

Without loss of generality, we denote the set of communication rounds by $\mathcal{C} = \{c_1, \ldots, c_C\}$, where $1 \leq c_1 < \cdots < c_C < T$. Let $c_0 = 0, c_{C+1} = T$. Then, we divide the total T rounds into the following $C + 1$ intervals

$$[c_0 + 1, c_1], [c_1 + 1, c_2], \ldots, [c_C + 1, c_{C+1}].$$

For any $i \in \{0, \ldots, C\}$ and $t \in [c_i + 1, c_{i+1}]$, we will set

$$f_{t,2}(\mathbf{x}) = \cdots = f_{t,n}(\mathbf{x}) = h_i(\mathbf{x}).$$

Then, the global loss can be written as

$$f_t(\mathbf{x}) = (n-1)h_i(\mathbf{x})$$

for any $i \in \{0, \ldots, C\}$ and $t \in [c_i + 1, c_{i+1}]$.

For any distributed online algorithm with communication rounds $\mathcal{C} = \{c_1, \ldots, c_C\}$, we denote the sequence of decisions made by the local learner 1 as $\mathbf{x}_1(1), \ldots, \mathbf{x}_1(T)$. For any $i \in \{0, \ldots, C\}$, we note that the decisions $\mathbf{x}_1(c_i + 1), \ldots, \mathbf{x}_1(c_{i+1})$ are made before the loss function $h_i(\mathbf{x})$ is revealed.

Inspired by the lower bound for the general OCO (Abernethy et al., 2008), we first utilize a randomized strategy to select $h_i(\mathbf{x})$ for any $i \in \{0, \ldots, C\}$, and derive an expected lower bound for $R_{T,1}$. Specifically, we independently select

$$h_i(\mathbf{x}) = \mathbf{w}_i^\top \mathbf{x}$$

for any $i \in \{0, \ldots, C\}$, where the coordinates of $\mathbf{w}_i$ are $\pm G/\sqrt{d}$ with probability $1/2$. Then, it is not hard to verify that $h_i(\mathbf{x})$ satisfies Assumption 1 and we have

$$\begin{aligned}
\mathbb{E}_{\mathbf{w}_0,\ldots,\mathbf{w}_C}[R_{T,1}] =& \mathbb{E}_{\mathbf{w}_0,\ldots,\mathbf{w}_C}\left[\sum_{t=1}^{T} f_t(\mathbf{x}_1(t)) - \min_{\mathbf{x}\in\mathcal{K}} \sum_{t=1}^{T} f_t(\mathbf{x})\right] \\
=& \mathbb{E}_{\mathbf{w}_0,\ldots,\mathbf{w}_C}\left[\sum_{i=0}^{C}\sum_{t=c_i+1}^{c_{i+1}} (n-1)h_i(\mathbf{x}_1(t)) - \min_{\mathbf{x}\in\mathcal{K}} \sum_{i=0}^{C}\sum_{t=c_i+1}^{c_{i+1}} (n-1)h_i(\mathbf{x})\right] \\
=& (n-1)\mathbb{E}_{\mathbf{w}_0,\ldots,\mathbf{w}_C}\left[\sum_{i=0}^{C}\sum_{t=c_i+1}^{c_{i+1}} \mathbf{w}_i^\top \mathbf{x}_1(t) - \min_{\mathbf{x}\in\mathcal{K}} \sum_{i=0}^{C}(c_{i+1}-c_i)\mathbf{w}_i^\top \mathbf{x}\right] \\
=& (n-1)\mathbb{E}_{\mathbf{w}_0,\ldots,\mathbf{w}_C}\left[-\min_{\mathbf{x}\in\mathcal{K}} \sum_{i=0}^{C}(c_{i+1}-c_i)\mathbf{w}_i^\top \mathbf{x}\right]
\end{aligned}$$

where the last equality is due to $\mathbb{E}_{\mathbf{w}_0,\ldots,\mathbf{w}_C}[\mathbf{w}_i^\top \mathbf{x}_1(t)] = 0$ for any $t \in [c_i + 1, c_{i+1}]$.

Due to the fact that a linear function is minimized at the vertices of the cube, we further have

$$\mathbb{E}_{\mathbf{w}_0,\ldots,\mathbf{w}_C}[R_{T,1}] = -(n-1)\mathbb{E}_{\mathbf{w}_0,\ldots,\mathbf{w}_C}\left[\min_{\mathbf{x}\in\left\{-R/\sqrt{d},R/\sqrt{d}\right\}^d} \mathbf{x}^\top \sum_{i=0}^{C}(c_{i+1}-c_i)\mathbf{w}_i\right].$$

Let $\epsilon_{01}, \ldots, \epsilon_{0d}, \ldots, \epsilon_{C1}, \ldots, \epsilon_{Cd}$ be independent and identically distributed variables with $\Pr(\epsilon_{ij} = \pm 1) = 1/2$ for $i \in \{0, \ldots, C\}$ and $j \in \{1, \ldots, d\}$. Then, we have

$$\begin{aligned}
\mathbb{E}_{\mathbf{w}_0,\ldots,\mathbf{w}_C}[R_{T,1}] =& -(n-1)\mathbb{E}_{\epsilon_{01},\ldots,\epsilon_{Cd}}\left[\sum_{j=1}^{d} -\frac{R}{\sqrt{d}}\left|\sum_{i=0}^{C}(c_{i+1}-c_i)\frac{\epsilon_{ij}G}{\sqrt{d}}\right|\right] \\
=& (n-1)RG\mathbb{E}_{\epsilon_{01},\ldots,\epsilon_{C1}}\left[\left|\sum_{i=0}^{C}(c_{i+1}-c_i)\epsilon_{i1}\right|\right] \\
\geq& \frac{(n-1)RG}{\sqrt{2}}\sqrt{\sum_{i=0}^{C}(c_{i+1}-c_i)^2} \geq \frac{(n-1)RG}{\sqrt{2}}\sqrt{\frac{(c_{C+1}-c_0)^2}{C+1}} \\
=& \frac{(n-1)RGT}{\sqrt{2(C+1)}}
\end{aligned} \tag{24}$$

where the first inequality is due to the Khintchine inequality and the second inequality is due to the Cauchy-Schwarz inequality. The expected lower bound in (24) implies that for any distributed online algorithm with communication rounds $\mathcal{C} = \{c_1, \ldots, c_C\}$, there exists a particular choice of $\mathbf{w}_0, \ldots, \mathbf{w}_C$ such that

$$R_{T,1} \geq \frac{(n-1)RGT}{\sqrt{2(C+1)}} \geq \frac{nRGT}{2\sqrt{2(C+1)}}$$

where the last inequality is due to $n - 1 \geq n/2$ for any integer $n \geq 2$. ■

Remark 7 Theorem 3 essentially establishes an $\Omega(\sqrt{T})$ lower bound on the communication rounds required by any distributed online algorithm whose all local learners attain the $O(T^{3/4})$ regret bound for convex losses, which matches (in terms of T) the $O(\sqrt{T})$ communication rounds required by our D-BOCG up to constant factors. Besides the dependence on T, the lower bound in Theorem 3 also depends on the network size n. When the number of communication rounds is limited to $O(\sqrt{T})$ and losses are convex, Theorem 3 provides a lower regret bound of $\Omega(nT^{3/4})$, but our D-BOCG only attains a regret bound of $O(n^{5/4}(1-\sigma_2(P))^{-1/2}T^{3/4})$ as shown in Corollary 1. So, in terms of the dependence on n and $1-\sigma_2(P)$, there still exists a gap between our upper and lower bounds. To eliminate this gap, one potential way is to reduce the dependence of the upper bound on n, and establish an improved lower bound depending on the spectral gap $1-\sigma_2(P)$ by carefully considering the topology of the network, which is non-trivial and will be investigated in the future.

Remark 8 We also note that in the proof of Theorem 3, only the regret of the local learner 1 is analyzed. It is natural to ask whether the regret of other local learner $i \neq 1$ simultaneously has a lower bound similar to that of the local learner 1. Unfortunately, the answer is negative. To be precise, let us consider a distributed online algorithm, which communicates at the end of C rounds before the round T but simply computes $\mathbf{x}_i(t+1) \in \operatorname{argmin}_{\mathbf{x}\in\mathcal{K}} f_{t,i}(\mathbf{x})$. If the sequence of local losses is selected as in the proof of Theorem 3, following notations used in the proof of Theorem 3, the regret of local learner $i \neq 1$ in this algorithm can be upper bounded as

$$\begin{aligned}\sum_{t=1}^{T} f_t(\mathbf{x}_i(t)) - \sum_{t=1}^{T} f_t(\mathbf{x}^*) &= \sum_{j=0}^{C} \sum_{t=c_j+1}^{c_{j+1}} (n-1)\mathbf{w}_j^\top(\mathbf{x}_i(t) - \mathbf{x}^*) \\ &\leq \sum_{j=0}^{C} (n-1)\mathbf{w}_j^\top(\mathbf{x}_i(c_j+1) - \mathbf{x}^*) \leq 2(n-1)(C+1)RG\end{aligned} \tag{25}$$

where $\mathbf{x}^* \in \operatorname{argmin}_{\mathbf{x}\in\mathcal{K}} \sum_{t=1}^{T} f_t(\mathbf{x})$, the first inequality is due to $\mathbf{x}_i(t) \in \operatorname{argmin}_{\mathbf{x}\in\mathcal{K}} \mathbf{w}_j^\top \mathbf{x}$ for $c_{j+1} \geq t > c_j + 1$, and the last inequality is due to the fact that $\mathbf{x}_i(c_j+1) \in \mathcal{K}$ and the coordinates of $\mathbf{w}_j$ belong to $\pm G/\sqrt{d}$. When C is much smaller than $T^{2/3}$, the regret bound in (25) could be much smaller than the lower bound $nRGT/(2\sqrt{2(C+1)})$ presented in Theorem 3. However, as discussed before, deriving a lower bound for one local learner is sufficient in this paper. So, we leave the problem of simultaneously lower bounding the regret of all local learners as a future work.

5.2 Strongly Convex Losses

For strongly convex losses, we provide a lower bound in the following theorem.

Theorem 4 *Suppose $\mathcal{K} = \left[-R/\sqrt{d}, R/\sqrt{d}\right]^d$, which satisfies Assumption 2 with constants R and $r = R/\sqrt{d}$. For distributed OCO with $n > 1$ local learners over $\mathcal{K}$ and any distributed online algorithm communicating at the end of C rounds before the round T, there exists a sequence of local loss functions satisfying Assumption 4 with $\alpha > 0$ and Assumption 1 with $G = 2\alpha R$ respectively such that*

$$R_{T,1} \geq \frac{\alpha n R^2 T}{8(C+1)}.$$

Proof This proof is similar to that of Theorem 3. The main difference is to add a term $\frac{\alpha}{2}\|\mathbf{x}\|_2^2$ to previous local loss functions, which makes them α-strongly convex.

For any distributed online algorithm with C communication rounds, we still denote the set of communication rounds by $\mathcal{C} = \{c_1, \ldots, c_C\}$ where $1 \leq c_1 < \cdots < c_C < T$, and the sequence of decisions made by the local learner 1 by $\mathbf{x}_1(1), \ldots, \mathbf{x}_1(T)$. Let $c_0 = 0$ and $c_{C+1} = T$. Then, we can divide the total T rounds into $C+1$ intervals

$$[c_0 + 1, c_1], [c_1 + 1, c_2], \ldots, [c_C + 1, c_{C+1}].$$

In each round t, for the local learner 1, we simply set $f_{t,1}(\mathbf{x}) = \frac{\alpha}{2}\|\mathbf{x}\|_2^2$ that satisfies Assumption 1 with $G = 2\alpha R$ and Assumption 4. For any $i \in \{0, \ldots, C\}$ and $t \in [c_i + 1, c_{i+1}]$, we set

$$f_{t,2}(\mathbf{x}) = \cdots = f_{t,n}(\mathbf{x}) = h_i(\mathbf{x}).$$

Specifically, we independently select

$$h_i(\mathbf{x}) = \mathbf{w}_i^\top \mathbf{x} + \frac{\alpha}{2}\|\mathbf{x}\|_2^2$$

for any $i \in \{0, \ldots, C\}$, where the coordinates of $\mathbf{w}_i$ are $\pm\alpha R/\sqrt{d}$ with probability $1/2$. It is easy to verify that $h_i(\mathbf{x})$ satisfies Assumption 1 with $G = 2\alpha R$ and Assumption 4, respectively. Note that the local learner 1 does not communicate with other local learners between rounds $c_i + 1$ and c_{i+1}. Therefore, the decisions $\mathbf{x}_1(c_i + 1), \ldots, \mathbf{x}_1(c_{i+1})$ are independent of $\mathbf{w}_i$.

Let $\bar{\mathbf{w}} = \frac{1}{\alpha T}\sum_{i=0}^{C}(c_{i+1} - c_i)\mathbf{w}_i$. Then, the total loss for any $\mathbf{x} \in \mathcal{K}$ is equal to

$$\begin{aligned}
\sum_{t=1}^{T} f_t(\mathbf{x}) =& \sum_{t=1}^{T}\left(\sum_{j=2}^{n} f_{t,j}(\mathbf{x}) + \frac{\alpha}{2}\|\mathbf{x}\|_2^2\right) \\
=& \sum_{i=0}^{C}(c_{i+1} - c_i)\left((n-1)\mathbf{w}_i^\top \mathbf{x} + \frac{\alpha n}{2}\|\mathbf{x}\|_2^2\right) \\
=& \alpha(n-1)T\bar{\mathbf{w}}^\top \mathbf{x} + \frac{\alpha n T}{2}\|\mathbf{x}\|_2^2 \\
=& \frac{\alpha T}{2}\left(\left\|\sqrt{n}\mathbf{x} + \frac{(n-1)}{\sqrt{n}}\bar{\mathbf{w}}\right\|_2^2 - \left\|\frac{(n-1)}{\sqrt{n}}\bar{\mathbf{w}}\right\|_2^2\right).
\end{aligned} \tag{26}$$

Since the absolute value of each element in $\mathbf{w}_i$ is equal to $\alpha R/\sqrt{d}$, we note that the absolute value of each element in $-\frac{n-1}{n}\bar{\mathbf{w}}$ is bounded by

$$\frac{n-1}{n\alpha T}\sum_{i=0}^{C}\frac{(c_{i+1} - c_i)\alpha R}{\sqrt{d}} = \frac{(n-1)R}{n\sqrt{d}} \leq \frac{R}{\sqrt{d}}$$

which implies that $-\frac{n-1}{n}\bar{\mathbf{w}}$ belongs to $\mathcal{K} = \left[-R/\sqrt{d}, R/\sqrt{d}\right]^d$.

By combining with (26), we have

$$\operatorname*{argmin}_{\mathbf{x}\in\mathcal{K}}\sum_{t=1}^{T} f_t(\mathbf{x}) = -\frac{n-1}{n}\bar{\mathbf{w}} \text{ and } \min_{\mathbf{x}\in\mathcal{K}}\sum_{t=1}^{T} f_t(\mathbf{x}) = -\frac{\alpha T}{2}\left\|\frac{(n-1)}{\sqrt{n}}\bar{\mathbf{w}}\right\|_2^2.$$

Then, we have

$$\begin{aligned}
&\mathbb{E}_{\mathbf{w}_0,\ldots,\mathbf{w}_C}\left[\sum_{t=1}^{T} f_t(\mathbf{x}_1(t)) - \min_{\mathbf{x}\in\mathcal{K}}\sum_{t=1}^{T} f_t(\mathbf{x})\right] \\
=&\mathbb{E}_{\mathbf{w}_0,\ldots,\mathbf{w}_C}\left[\sum_{i=0}^{C}\sum_{t=c_i+1}^{c_{i+1}}\left((n-1)\mathbf{w}_i^\top\mathbf{x}_1(t) + \frac{\alpha n}{2}\|\mathbf{x}_1(t)\|_2^2\right) + \frac{\alpha T}{2}\left\|\frac{(n-1)}{\sqrt{n}}\bar{\mathbf{w}}\right\|_2^2\right] \\
\geq&\mathbb{E}_{\mathbf{w}_0,\ldots,\mathbf{w}_C}\left[\sum_{i=0}^{C}\sum_{t=c_i+1}^{c_{i+1}}(n-1)\mathbf{w}_i^\top\mathbf{x}_1(t) + \frac{\alpha(n-1)^2T}{2n}\|\bar{\mathbf{w}}\|_2^2\right] \\
=&\mathbb{E}_{\mathbf{w}_0,\ldots,\mathbf{w}_C}\left[\frac{\alpha(n-1)^2T}{2n}\|\bar{\mathbf{w}}\|_2^2\right]
\end{aligned} \tag{27}$$

where the inequality is due to $\alpha\|\mathbf{x}\|_2^2 \geq 0$ for any $\mathbf{x}$ and the last equality is due to $\mathbb{E}_{\mathbf{w}_0,\ldots,\mathbf{w}_C}[\mathbf{w}_i^\top\mathbf{x}_1(t)] = 0$ for any $t \in [c_i+1, c_{i+1}]$.

Let $\epsilon_{01},\ldots,\epsilon_{0d},\ldots,\epsilon_{C1},\ldots,\epsilon_{Cd}$ be independent and identically distributed variables with $\Pr(\epsilon_{ij} = \pm 1) = 1/2$ for $i \in \{0,\ldots,C\}$ and $j \in \{1,\ldots,d\}$. Then, we have

$$\begin{aligned}
\mathbb{E}_{\mathbf{w}_0,\ldots,\mathbf{w}_C}\left[\frac{\alpha(n-1)^2T}{2n}\|\bar{\mathbf{w}}\|_2^2\right] =& \frac{(n-1)^2}{2\alpha nT}\mathbb{E}_{\mathbf{w}_0,\ldots,\mathbf{w}_C}\left[\left\|\sum_{i=0}^{C}(c_{i+1}-c_i)\mathbf{w}_i\right\|_2^2\right] \\
=& \frac{(n-1)^2}{2\alpha nT}\mathbb{E}_{\epsilon_{01},\ldots,\epsilon_{Cd}}\left[\sum_{j=1}^{d}\left|\sum_{i=0}^{C}(c_{i+1}-c_i)\frac{\epsilon_{ij}\alpha R}{\sqrt{d}}\right|^2\right] \\
=& \frac{\alpha(n-1)^2R^2}{2nT}\mathbb{E}_{\epsilon_{01},\ldots,\epsilon_{C1}}\left[\left|\sum_{i=0}^{C}(c_{i+1}-c_i)\epsilon_{i1}\right|^2\right] \\
\geq& \frac{\alpha(n-1)^2R^2}{2nT}\sum_{i=0}^{C}(c_{i+1}-c_i)^2 \\
\geq& \frac{\alpha(n-1)^2R^2}{2nT}\cdot\frac{(c_{C+1}-c_0)^2}{C+1} = \frac{\alpha(n-1)^2R^2T}{2n(C+1)}
\end{aligned} \tag{28}$$

where the first inequality is due to the Khintchine inequality, and the second inequality is due to the Cauchy-Schwarz inequality.

By combining (27) with (28), we derive an expected lower bound as

$$\mathbb{E}_{\mathbf{w}_0,\ldots,\mathbf{w}_C}[R_{T,1}] = \mathbb{E}_{\mathbf{w}_0,\ldots,\mathbf{w}_C}\left[\sum_{t=1}^{T} f_t(\mathbf{x}_1(t)) - \min_{\mathbf{x}\in\mathcal{K}}\sum_{t=1}^{T} f_t(\mathbf{x})\right] \geq \frac{\alpha(n-1)^2R^2T}{2n(C+1)}$$

which implies that for any distributed online algorithm with communication rounds $\mathcal{C} = \{c_1,\ldots,c_C\}$, there exists a particular choice of $\mathbf{w}_0,\ldots,\mathbf{w}_C$ such that

$$R_{T,1} \geq \frac{\alpha(n-1)^2R^2T}{2n(C+1)} \geq \frac{\alpha nR^2T}{8(C+1)}$$

where the last inequality is due to $n-1 \geq n/2$ for any integer $n \geq 2$. ■

Remark 9 Theorem 4 essentially establishes an $\Omega\left(T^{1/3}(\log T)^{-1/3}\right)$ lower bound on the communication rounds required by any distributed online algorithm whose all local learners attain $O(T^{2/3}(\log T)^{1/3})$ regret bound for strongly convex losses, which almost matches (in terms of T) the $O(T^{1/3}(\log T)^{2/3})$ communication rounds required by our D-BOCG up to polylogarithmic factors. However, if we further consider the dependence on n and $1-\sigma_2(P)$, there still exists a gap between the upper bound of our D-BOCG and the lower bound in Theorem 4, which is similar to the case with convex losses. Specifically, when the number of communication rounds is limited to $O(T^{1/3}(\log T)^{2/3})$ and losses are strongly convex, Theorem 4 provides a lower regret bound of $\Omega(nT^{2/3}(\log T)^{-2/3})$, but our D-BOCG only attains a regret bound of $O(n^{3/2}(1-\sigma_2(P))^{-1}T^{2/3}(\log T)^{1/3})$ as shown in Corollary 2. As discussed in Remark 7, it is non-trivial to eliminate this gap, and we will investigate this limitation in the future.

Remark 10 Similar to the proof for Theorem 3, the proof of Theorem 4 only provides a lower bound for the regret of the local learner 1, and the regret of other local learner $i \neq 1$ is not lower bounded simultaneously. Specifically, let us consider a distributed online algorithm for α-strongly convex losses, which communicates at the end of C rounds before the round T but simply computes $\mathbf{x}_i(t+1) = \operatorname{argmin}_{\mathbf{x}\in\mathcal{K}} f_{t,i}(\mathbf{x}) + \frac{\alpha}{2(n-1)}\|\mathbf{x}\|_2^2$. If the sequence of local losses is selected as in the proof of Theorem 4, following notations used in the proof of Theorem 4, the regret of local learner $i \neq 1$ in this algorithm can be upper bounded as

$$\begin{aligned}
&\sum_{t=1}^{T} f_t(\mathbf{x}_i(t)) - \sum_{t=1}^{T} f_t(\mathbf{x}^*) \\
=&\sum_{j=0}^{C}\sum_{t=c_j+1}^{c_{j+1}} (n-1)\left(\mathbf{w}_j^\top \mathbf{x}_i(t) + \frac{n\alpha}{2(n-1)}\|\mathbf{x}_i(t)\|_2^2 - \mathbf{w}_j^\top \mathbf{x}^* - \frac{n\alpha}{2(n-1)}\|\mathbf{x}^*\|_2^2\right) \\
\leq&\sum_{j=0}^{C}(n-1)\left(\mathbf{w}_j^\top \mathbf{x}_i(c_j+1) + \frac{n\alpha}{2(n-1)}\|\mathbf{x}_i(c_j+1)\|_2^2 - \mathbf{w}_j^\top \mathbf{x}^* - \frac{n\alpha}{2(n-1)}\|\mathbf{x}^*\|_2^2\right) \\
\leq&\frac{5\alpha R^2(n-1)(C+1)}{2}
\end{aligned} \tag{29}$$

where $\mathbf{x}^* = \operatorname{argmin}_{\mathbf{x}\in\mathcal{K}} \sum_{t=1}^{T} f_t(\mathbf{x})$, the first inequality is due to $\mathbf{x}_i(t) = \operatorname{argmin}_{\mathbf{x}\in\mathcal{K}} \mathbf{w}_j^\top \mathbf{x} + \frac{n\alpha}{2(n-1)}\|\mathbf{x}\|_2^2$ for $c_{j+1} \geq t > c_j + 1$, and the last inequality is due to the fact that $\mathbf{x}_i(c_j+1) \in \mathcal{K}$ and the coordinates of $\mathbf{w}_j$ belong to $\pm\alpha R/\sqrt{d}$. When C is much smaller than $\sqrt{T}$, the regret bound in (29) could be much smaller than the lower bound $\alpha n R^2 T/(8(C+1))$ presented in Theorem 4. However, as discussed before, deriving a lower bound for one local learner is sufficient in this paper, and the problem of simultaneously lower bounding the regret of all local learners is left as a future work.

6. An Extension of D-BOCG to the Bandit Setting

In this section, we extend our D-BOCG to the bandit setting, where only the loss value is available to each local learner. The main idea is to combine D-BOCG with the one-point gradient estimator (Flaxman et al., 2005).

Algorithm 3 D-BBCG

1: **Input:** feasible set $\mathcal{K}$, δ, $\mathbf{x}_{\text{in}} \in \mathcal{K}_\delta$, α, h, L, and K
2: **Initialization:** choose $\mathbf{x}_1(1) = \cdots = \mathbf{x}_n(1) = \mathbf{x}_{\text{in}}$ and set $\mathbf{z}_1(1) = \cdots = \mathbf{z}_n(1) = \mathbf{0}$
3: **for** $m = 1, \ldots, T/K$ **do**
4: **for** each local learner $i \in V$ **do**
5: define $F_{m,i}(\mathbf{x}) = \mathbf{z}_i(m)^\top \mathbf{x} + \frac{(m-1)\alpha K}{2}\|\mathbf{x}\|_2^2 + h\|\mathbf{x} - \mathbf{x}_{\text{in}}\|_2^2$
6: $\widehat{\mathbf{g}}_i(m) = \mathbf{0}$
7: **for** $t = (m-1)K + 1, \ldots, mK$ **do**
8: play $\mathbf{y}_i(t) = \mathbf{x}_i(m) + \delta \mathbf{u}_i(t)$ where $\mathbf{u}_i(t) \sim \mathcal{S}^d$
9: observe $f_{t,i}(\mathbf{y}_i(t))$ and compute $\mathbf{g}_i(t) = \frac{d}{\delta} f_{t,i}(\mathbf{y}_i(t))\mathbf{u}_i(t)$
10: $\widehat{\mathbf{g}}_i(m) = \widehat{\mathbf{g}}_i(m) + \mathbf{g}_i(t)$
11: **end for**
12: $\mathbf{x}_i(m+1) = \text{CG}(\mathcal{K}_\delta, L, F_{m,i}(\mathbf{x}), \mathbf{x}_i(m))$ //This step can be executed *in parallel* to the above *for* loop.
13: $\mathbf{z}_i(m+1) = \sum_{j \in N_i} P_{ij}\mathbf{z}_j(m) + \widehat{\mathbf{g}}_i(m) - \alpha K \mathbf{x}_i(m)$
14: **end for**
15: **end for**

6.1 The Algorithm

By combining our D-BOCG with the one-point gradient estimator, our algorithm for the bandit setting is outlined in Algorithm 3, and named as distributed block bandit conditional gradient (D-BBCG), where $0 < \delta \leq r$ and $\mathcal{K}_\delta = (1-\delta/r)\mathcal{K} = \{(1-\delta/r)\mathbf{x} | \mathbf{x} \in \mathcal{K}\}$. Comparing D-BBCG with D-BOCG, there exist three differences as follows. First, in line 8 of D-BBCG, the actual decision $\mathbf{y}_i(t)$ is $\mathbf{x}_i(m)$ plus a random decision $\delta\mathbf{u}_i(t)$, where $\mathbf{u}_i(t)$ is uniformly sampled from the unit sphere $\mathcal{S}^d$. Second, in line 9 of D-BBCG, we can only observe the loss value $f_{t,i}(\mathbf{y}_i(t))$ instead of the gradient $\nabla f_{t,i}(\mathbf{x}_i(m))$, and adopt the one-point gradient estimator to approximate the gradient as

$$\mathbf{g}_i(t) = \frac{d}{\delta} f_{t,i}(\mathbf{y}_i(t))\mathbf{u}_i(t).$$

Third, to ensure $\mathbf{y}_i(t) \in \mathcal{K}$, in line 12 of D-BBCG, we perform

$$\mathbf{x}_i(m+1) = \text{CG}(\mathcal{K}_\delta, L, F_{m,i}(\mathbf{x}), \mathbf{x}_i(m))$$

by replacing $\mathcal{K}$ in line 11 of D-BOCG with a smaller set $\mathcal{K}_\delta \subseteq \mathcal{K}$, which limits $\mathbf{x}_i(m)$ in the set $\mathcal{K}_\delta$. Because of Assumption 2 and $0 < \delta \leq r$, it is easy to verify that $\mathbf{x} + \delta\mathbf{u} \in \mathcal{K}$ for any $\mathbf{x} \in \mathcal{K}_\delta$ and $\mathbf{u} \sim \mathcal{S}^d$ by utilizing the fact that $r\mathcal{B}^d \subseteq \mathcal{K}$.

6.2 Theoretical Guarantees

In the following, we present theoretical guarantees of our D-BBCG. We first provide expected regret bounds of D-BBCG for convex losses and strongly convex losses, respectively.

Theorem 5 *Let $\alpha = 0$, $K = L = \sqrt{T}$, $h = \frac{n^{1/4} d M T^{3/4}}{\sqrt{1-\sigma_2(P)}R}$, and $\delta = cT^{-1/4}$, where $c > 0$ is a constant such that $\delta \leq r$. Under Assumptions 1, 2, 3, and 5, for any $i \in V$, Algorithm 3*

ensures

$$\mathbb{E}\left[R_{T,i}\right] = O\left(n^{5/4}(1-\sigma_2(P))^{-1/2}T^{3/4}\right).$$

Theorem 6 *Let $\alpha > 0$, $K = L = T^{2/3}(\ln T)^{-2/3}$, $h = \alpha K$, and $\delta = cT^{-1/3}(\ln T)^{1/3}$, where $c > 0$ is a constant such that $\delta \leq r$. Under Assumptions 1, 2, 3, 4, and 5, for any $i \in V$, Algorithm 3 ensures*

$$\mathbb{E}\left[R_{T,i}\right] = O\left(n^{3/2}(1-\sigma_2(P))^{-1}T^{2/3}(\log T)^{1/3}\right).$$

Remark 11 Theorems 5 and 6 show that D-BBCG can attain an expected regret bound of $O(T^{3/4})$ with $O(\sqrt{T})$ communication rounds for convex losses, and attain an expected regret bound of $O(T^{2/3}(\log T)^{1/3})$ with $O(T^{1/3}(\log T)^{2/3})$ communication rounds, which is similar to D-BOCG in the full information setting.

Moreover, we show that D-BBCG enjoys a high-probability regret bound of $O(T^{3/4}(\log T)^{1/2})$ with $O(\sqrt{T})$ communication rounds for convex losses.

Theorem 7 *Let $\alpha = 0$, $K = L = \sqrt{T}$, and $\delta = cT^{-1/4}$, where $c > 0$ is a constant such that $\delta \leq r$. Moreover, let $h = \frac{n^{1/4}\xi_T dMT^{3/4}}{\sqrt{1-\sigma_2(P)}R}$, where $\xi_T = 1 + \sqrt{8\ln\frac{n\sqrt{T}}{\gamma}}$ and $0.5 > \gamma > 0$ is a constant. Under Assumptions 1, 2, 3, and 5, for any $i \in V$, with probability at least $1-2\gamma$, Algorithm 3 has*

$$R_{T,i} = O\left(n^{5/4}(1-\sigma_2(P))^{-1/2}T^{3/4}\xi_T\right).$$

Remark 12 While the above theorem presents a high-probability regret bound for convex losses, it is hard to extend it into the case with strongly convex losses. We note that according to the proof of Theorem 7, the high-probability regret bound of D-BBCG has a term $O(K\sqrt{B\log(1/\gamma)})$, where K is incurred by the delayed update mechanism and $\sqrt{B\log(1/\gamma)}$ is incurred by applying the classical Azuma's concentration inequality (Azuma, 1967). If we consider strongly convex losses, we would like to set $K = T^{2/3}(\ln T)^{-2/3}$ to control the communication complexity, but in this case the term $O(K\sqrt{B\log(1/\gamma)})$ is worse than the expected regret bound in Theorem 6. Therefore, to establish a high-probability regret bound for strongly convex losses, we may need some novel techniques to improve the term $O(K\sqrt{B\log(1/\gamma)})$, which will be investigated in the future.

6.3 Analysis

In this section, we only provide the proof of Theorems 5 and 6. The proof of Theorem 7 can be found in the Appendix.

6.3.1 Proof of Theorems 5 and 6

Similar to the proof of Theorem 1, we first define several auxiliary variables. Let $\bar{\mathbf{z}}(m) = \frac{1}{n}\sum_{i=1}^{n}\mathbf{z}_i(m)$ for $m \in [B+1]$, and let $\mathbf{d}_i(m) = \widehat{\mathbf{g}}_i(m) - \alpha K\mathbf{x}_i(m)$ and $\bar{\mathbf{d}}(m) = \frac{1}{n}\sum_{i=1}^{n}\mathbf{d}_i(m)$ for $m \in [B]$. Then, we define $\bar{\mathbf{x}}(1) = \mathbf{x}_{\text{in}}$ and $\bar{\mathbf{x}}(m+1) = \operatorname{argmin}_{\mathbf{x}\in\mathcal{K}_\delta}\bar{F}_m(\mathbf{x})$ for any $m \in [B+1]$, where

$$\bar{F}_m(\mathbf{x}) = \bar{\mathbf{z}}(m)^\top\mathbf{x} + \frac{(m-1)\alpha K}{2}\|\mathbf{x}\|_2^2 + h\|\mathbf{x}-\mathbf{x}_{\text{in}}\|_2^2.$$

Similarly, we define $\widehat{\mathbf{x}}_i(m+1) = \operatorname{argmin}_{\mathbf{x}\in\mathcal{K}_\delta} F_{m,i}(\mathbf{x})$ for any $m \in [B+1]$, where

$$F_{m,i}(\mathbf{x}) = \mathbf{z}_i(m)^\top \mathbf{x} + \frac{(m-1)\alpha K}{2}\|\mathbf{x}\|_2^2 + h\|\mathbf{x} - \mathbf{x}_{\text{in}}\|_2^2.$$

is defined in Algorithm 3.

Moreover, we need to introduce the following lemmas.

Lemma 6 *Let $\mathbf{d}_i(m) = \widehat{\mathbf{g}}_i(m) - \alpha K \mathbf{x}_i(m)$ for $m \in [B]$. Under Assumptions 1, 2, and 5, for any $i \in V$ and $m \in [B]$, Algorithm 3 ensures that*

$$\mathbb{E}[\|\mathbf{d}_i(m)\|_2]^2 \le \mathbb{E}[\|\mathbf{d}_i(m)\|_2^2] \le 2K\left(\frac{dM}{\delta}\right)^2 + 2K^2G^2 + 2(\alpha K R)^2.$$

Lemma 7 *(Derived from the Proof of Lemma 6 in Zhang et al. 2017) For any $i \in [n]$, let $\mathbf{d}_i(1), \ldots, \mathbf{d}_i(m) \in \mathbb{R}^d$ be a sequence of vectors. Let $\mathbf{z}_i(1) = \mathbf{0}$, $\mathbf{z}_i(m+1) = \sum_{j\in N_i} P_{ij}\mathbf{z}_j(m) + \mathbf{d}_i(m)$, and $\bar{\mathbf{z}}(m) = \frac{1}{n}\sum_{i=1}^n \mathbf{z}_i(m)$ for $m \in [B]$, where P satisfies Assumption 3. For any $i \in V$ and $m \in [B]$, assuming $\mathbb{E}[\|\mathbf{d}_i(m)\|_2] \le \widehat{G}$ where $\widehat{G} > 0$ is a constant, we have*

$$\mathbb{E}[\|\mathbf{z}_i(m) - \bar{\mathbf{z}}(m)\|_2] \le \frac{\widehat{G}\sqrt{n}}{1-\sigma_2(P)}.$$

Lemma 8 *(Lemma 2.6 in Hazan 2016 and Lemma 6 in Wan et al. 2022) Let $\mathcal{B}^d$ denote the unit Euclidean ball centered at the origin in $\mathbb{R}^d$. Let $f(\mathbf{x}) : \mathbb{R}^d \to \mathbb{R}$ be α-strongly convex and G-Lipschitz over a convex and compact set $\mathcal{K} \subset \mathbb{R}^d$. Then, its δ-smoothed version $\widehat{f}_\delta(\mathbf{x}) = \mathbb{E}_{\mathbf{u}\sim\mathcal{B}^d}[f(\mathbf{x}+\delta\mathbf{u})]$ has the following properties:*

- *$\widehat{f}_\delta(\mathbf{x})$ is α-strongly convex over $\mathcal{K}_\delta$;*
- *$|\widehat{f}_\delta(\mathbf{x}) - f(\mathbf{x})| \le \delta G$ for any $\mathbf{x} \in \mathcal{K}_\delta$;*
- *$\widehat{f}_\delta(\mathbf{x})$ is G-Lipschitz over $\mathcal{K}_\delta$.*

Now, we derive an upper bound of $\mathbb{E}[\|\mathbf{x}_i(m) - \bar{\mathbf{x}}(m)\|_2]$ for any $m \in [B]$. If $m = 1$, according to the definition and Algorithm 3, it is easy to verify that

$$\mathbb{E}[\|\mathbf{x}_i(m) - \bar{\mathbf{x}}(m)\|_2] = \mathbb{E}[0] = 0. \tag{30}$$

For any $B \ge m \ge 2$, we note that $F_{m-1,i}(\mathbf{x})$ is $((m-2)\alpha K + 2h)$-smooth, and Algorithm 3 ensures

$$\mathbf{x}_i(m) = \mathrm{CG}(\mathcal{K}_\delta, L, F_{m-1,i}(\mathbf{x}), \mathbf{x}_i(m-1)).$$

According to Lemma 1 and Assumption 2, for $B \ge m \ge 2$, it is easy to verify that

$$F_{m-1,i}(\mathbf{x}_i(m)) - F_{m-1,i}(\widehat{\mathbf{x}}_i(m)) \le \frac{8((m-2)\alpha K + 2h)R^2}{L+2}.$$

Then, for any $B \ge m \ge 2$, it is easy to verify that

$$\begin{aligned}
\|\mathbf{x}_i(m) - \bar{\mathbf{x}}(m)\|_2 \le& \|\mathbf{x}_i(m) - \widehat{\mathbf{x}}_i(m)\|_2 + \|\widehat{\mathbf{x}}_i(m) - \bar{\mathbf{x}}(m)\|_2 \\
\le& \sqrt{\frac{2F_{m-1,i}(\mathbf{x}_i(m)) - 2F_{m-1,i}(\widehat{\mathbf{x}}_i(m))}{(m-2)\alpha K + 2h}} + \|\widehat{\mathbf{x}}_i(m) - \bar{\mathbf{x}}(m)\|_2 \\
\le& \frac{4R}{\sqrt{L+2}} + \|\widehat{\mathbf{x}}_i(m) - \bar{\mathbf{x}}(m)\|_2
\end{aligned} \tag{31}$$

where the second inequality is due to the fact that $F_{m-1,i}(\mathbf{x})$ is also $((m-2)\alpha K+2h)$-strongly convex and (11).

Moreover, for any $B \geq m \geq 2$, similar to (17) and (18), we have

$$\begin{aligned}\widehat{\mathbf{x}}_i(m) =& \operatorname*{argmin}_{\mathbf{x}\in\mathcal{K}_\delta} \mathbf{z}_i(m-1)^\top \mathbf{x} + \frac{(m-2)\alpha K}{2}\|\mathbf{x}\|_2^2 + h\|\mathbf{x}-\mathbf{x}_{\text{in}}\|_2^2 \\ =& \operatorname*{argmin}_{\mathbf{x}\in\mathcal{K}_\delta} \frac{2}{(m-2)\alpha K+2h}(\mathbf{z}_i(m-1)-2h\mathbf{x}_{\text{in}})^\top \mathbf{x} + \|\mathbf{x}\|_2^2\end{aligned}$$

and

$$\begin{aligned}\bar{\mathbf{x}}(m) =& \operatorname*{argmin}_{\mathbf{x}\in\mathcal{K}_\delta} \bar{\mathbf{z}}(m-1)^\top \mathbf{x} + \frac{(m-2)\alpha K}{2}\|\mathbf{x}\|_2^2 + h\|\mathbf{x}-\mathbf{x}_{\text{in}}\|_2^2 \\ =& \operatorname*{argmin}_{\mathbf{x}\in\mathcal{K}_\delta} \frac{2}{(m-2)\alpha K+2h}(\bar{\mathbf{z}}(m-1)-2h\mathbf{x}_{\text{in}})^\top \mathbf{x} + \|\mathbf{x}\|_2^2.\end{aligned}$$

Therefore, for any $B \geq m \geq 2$, by applying Lemma 4, we have

$$\begin{aligned}\|\widehat{\mathbf{x}}_i(m)-\bar{\mathbf{x}}(m)\|_2 \leq& \frac{\|\mathbf{z}_i(m-1)-2h\mathbf{x}_{\text{in}}-\bar{\mathbf{z}}(m-1)+2h\mathbf{x}_{\text{in}}\|_2}{(m-2)\alpha K+2h} \\ =& \frac{\|\mathbf{z}_i(m-1)-\bar{\mathbf{z}}(m-1)\|_2}{(m-2)\alpha K+2h}.\end{aligned}$$

By further combining with (31), for any $B \geq m \geq 2$, we have

$$\begin{aligned}&\mathbb{E}[\|\mathbf{x}_i(m)-\bar{\mathbf{x}}(m)\|_2] \\ \leq& \frac{4R}{\sqrt{L+2}} + \mathbb{E}[\|\widehat{\mathbf{x}}_i(m)-\bar{\mathbf{x}}(m)\|_2] \\ \leq& \frac{4R}{\sqrt{L+2}} + \frac{\mathbb{E}[\|\mathbf{z}_i(m-1)-\bar{\mathbf{z}}(m-1)\|_2]}{(m-2)\alpha K+2h} \\ \leq& \frac{4R}{\sqrt{L+2}} + \sqrt{2K\left(\frac{dM}{\delta}\right)^2 + 2K^2G^2 + 2(\alpha KR)^2}\frac{\sqrt{n}}{((m-2)\alpha K+2h)(1-\sigma_2(P))}\end{aligned} \tag{32}$$

where the last inequality is due to

$$\mathbb{E}[\|\mathbf{z}_i(m-1)-\bar{\mathbf{z}}(m-1)\|_2] \leq \sqrt{2K\left(\frac{dM}{\delta}\right)^2 + 2K^2G^2 + 2(\alpha KR)^2}\frac{\sqrt{n}}{1-\sigma_2(P)}$$

which is derived by combining Lemma 6 with Lemma 7.

Let $u_1 = 0$ and

$$u_m = \frac{4R}{\sqrt{L+2}} + \sqrt{2K\left(\frac{dM}{\delta}\right)^2 + 2K^2G^2 + 2(\alpha KR)^2}\frac{\sqrt{n}}{((m-2)\alpha K+2h)(1-\sigma_2(P))}$$

for any $B \geq m \geq 2$. From (30) and (32), for any $m \in [B]$, it holds that

$$\mathbb{E}[\|\mathbf{x}_i(m)-\bar{\mathbf{x}}(m)\|_2] \leq u_m.$$

Next, let $\mathbf{x}^* \in \operatorname{argmin}_{\mathbf{x}\in\mathcal{K}} \sum_{t=1}^{T} f_t(\mathbf{x})$, $\widetilde{\mathbf{x}}^* = (1-\delta/r)\mathbf{x}^*$, and we define the δ-smoothed version of $f_{t,j}(\mathbf{x})$ as

$$\widehat{f}_{t,j,\delta}(\mathbf{x}) = \mathbb{E}_{\mathbf{u}\sim\mathcal{B}^d}[f_{t,j}(\mathbf{x}+\delta\mathbf{u})]$$

where $\mathcal{B}^d$ denotes the unit Euclidean ball centered at the origin in $\mathbb{R}^d$. For any $i, j \in V$, $m \in [B]$, and $t \in \mathcal{T}_m$, by applying Lemma 8, we have

$$\begin{aligned}
&\mathbb{E}[\widehat{f}_{t,j,\delta}(\mathbf{x}_i(m)) - \widehat{f}_{t,j,\delta}(\widetilde{\mathbf{x}}^*)] \\
\leq&\mathbb{E}[\widehat{f}_{t,j,\delta}(\bar{\mathbf{x}}(m)) - \widehat{f}_{t,j,\delta}(\widetilde{\mathbf{x}}^*) + G\|\bar{\mathbf{x}}(m) - \mathbf{x}_i(m)\|_2] \\
\leq&\mathbb{E}[\widehat{f}_{t,j,\delta}(\mathbf{x}_j(m)) - \widehat{f}_{t,j,\delta}(\widetilde{\mathbf{x}}^*) + G\|\bar{\mathbf{x}}(m) - \mathbf{x}_j(m)\|_2] + Gu_m \\
\leq&\mathbb{E}\left[\nabla\widehat{f}_{t,j,\delta}(\mathbf{x}_j(m))^\top(\mathbf{x}_j(m) - \widetilde{\mathbf{x}}^*) - \frac{\alpha}{2}\|\mathbf{x}_j(m) - \widetilde{\mathbf{x}}^*\|_2^2\right] + 2Gu_m \\
\leq&\mathbb{E}\left[\nabla\widehat{f}_{t,j,\delta}(\mathbf{x}_j(m))^\top(\bar{\mathbf{x}}(m+1) - \widetilde{\mathbf{x}}^*) - \frac{\alpha}{2}\|\mathbf{x}_j(m) - \widetilde{\mathbf{x}}^*\|_2^2\right] + 2Gu_m \\
&+ \mathbb{E}[\nabla\widehat{f}_{t,j,\delta}(\mathbf{x}_j(m))^\top(\mathbf{x}_j(m) - \bar{\mathbf{x}}(m+1))] \\
\leq&\mathbb{E}\left[\nabla\widehat{f}_{t,j,\delta}(\mathbf{x}_j(m))^\top(\bar{\mathbf{x}}(m+1) - \widetilde{\mathbf{x}}^*) - \frac{\alpha}{2}\|\mathbf{x}_j(m) - \widetilde{\mathbf{x}}^*\|_2^2\right] + 2Gu_m \\
&+ \mathbb{E}[\|\nabla\widehat{f}_{t,j,\delta}(\mathbf{x}_j(m))\|_2\|\mathbf{x}_j(m) - \bar{\mathbf{x}}(m+1)\|_2] \\
\leq&\mathbb{E}\left[\nabla\widehat{f}_{t,j,\delta}(\mathbf{x}_j(m))^\top(\bar{\mathbf{x}}(m+1) - \widetilde{\mathbf{x}}^*) - \frac{\alpha}{2}\|\mathbf{x}_j(m) - \widetilde{\mathbf{x}}^*\|_2^2\right] + 2Gu_m \\
&+ \mathbb{E}[G(\|\mathbf{x}_j(m) - \bar{\mathbf{x}}(m)\|_2 + \|\bar{\mathbf{x}}(m) - \bar{\mathbf{x}}(m+1)\|_2)] \\
\leq&\mathbb{E}\left[(\nabla\widehat{f}_{t,j,\delta}(\mathbf{x}_j(m)) - \alpha\mathbf{x}_j(m))^\top(\bar{\mathbf{x}}(m+1) - \widetilde{\mathbf{x}}^*) - \frac{\alpha}{2}(\|\widetilde{\mathbf{x}}^*\|_2^2 - \|\bar{\mathbf{x}}(m+1)\|_2^2)\right] \\
&+ \mathbb{E}\left[G\|\bar{\mathbf{x}}(m) - \bar{\mathbf{x}}(m+1)\|_2\right] + 3Gu_m
\end{aligned} \tag{33}$$

where the first two inequalities are due to the fact that $\widehat{f}_{t,j,\delta}(\mathbf{x})$ is G-Lipschitz over $\mathcal{K}_\delta$, the third inequality is due to the strong convexity of $\widehat{f}_{t,j,\delta}(\mathbf{x})$, and the last inequality is due to

$$\begin{aligned}
\|\mathbf{x}_j(m) - \widetilde{\mathbf{x}}^*\|_2^2 =&\|\mathbf{x}_j(m) - \bar{\mathbf{x}}(m+1)\|_2^2 + 2\mathbf{x}_j(m)^\top(\bar{\mathbf{x}}(m+1) - \widetilde{\mathbf{x}}^*) + \|\widetilde{\mathbf{x}}^*\|_2^2 - \|\bar{\mathbf{x}}(m+1)\|_2^2 \\
\geq&2\mathbf{x}_j(m)^\top(\bar{\mathbf{x}}(m+1) - \widetilde{\mathbf{x}}^*) + \|\widetilde{\mathbf{x}}^*\|_2^2 - \|\bar{\mathbf{x}}(m+1)\|_2^2.
\end{aligned}$$

Moreover, it is not hard to verify that

$$\begin{aligned}
R_{T,i} =& \sum_{m=1}^{B}\sum_{t\in\mathcal{T}_m}\sum_{j=1}^{n}(f_{t,j}(\mathbf{x}_i(m) + \delta\mathbf{u}_i(t)) - f_{t,j}(\mathbf{x}^*)) \\
\leq& \sum_{m=1}^{B}\sum_{t\in\mathcal{T}_m}\sum_{j=1}^{n}((f_{t,j}(\mathbf{x}_i(m)) + G\|\delta\mathbf{u}_i(t)\|_2) - (f_{t,j}(\widetilde{\mathbf{x}}^*) - G\|\widetilde{\mathbf{x}}^* - \mathbf{x}^*\|_2)) \\
\leq& \sum_{m=1}^{B}\sum_{t\in\mathcal{T}_m}\sum_{j=1}^{n}\left(f_{t,j}(\mathbf{x}_i(m)) - f_{t,j}(\widetilde{\mathbf{x}}^*) + G\|\delta\mathbf{u}_i(t)\|_2 + \frac{\delta GR}{r}\right) \\
\leq& \sum_{m=1}^{B}\sum_{t\in\mathcal{T}_m}\sum_{j=1}^{n}((\widehat{f}_{t,j,\delta}(\mathbf{x}_i(m)) + \delta G) - (\widehat{f}_{t,j,\delta}(\widetilde{\mathbf{x}}^*) - \delta G)) + \delta nGT + \frac{\delta nGRT}{r} \\
=& \sum_{m=1}^{B}\sum_{t\in\mathcal{T}_m}\sum_{j=1}^{n}(\widehat{f}_{t,j,\delta}(\mathbf{x}_i(m)) - \widehat{f}_{t,j,\delta}(\widetilde{\mathbf{x}}^*)) + 3\delta nGT + \frac{\delta nGRT}{r}
\end{aligned} \tag{34}$$

where the first inequality is due to Assumption 1, the second inequality is due to $\mathbf{x}^* \in \mathcal{K}$ and Assumption 2, and the third inequality is due to Lemma 8.

By combining (33) with (34), we have

$$\begin{aligned}
&\mathbb{E}\left[R_{T,i}\right] \\
\leq &\sum_{m=1}^{B}\sum_{t\in\mathcal{T}_m}\sum_{j=1}^{n}\mathbb{E}\left[(\nabla\widehat{f}_{t,j,\delta}(\mathbf{x}_j(m))-\alpha\mathbf{x}_j(m))^\top(\bar{\mathbf{x}}(m+1)-\widetilde{\mathbf{x}}^*)\right] \\
&-\sum_{m=1}^{B}\mathbb{E}\left[\frac{n\alpha K}{2}(\|\widetilde{\mathbf{x}}^*\|_2^2-\|\bar{\mathbf{x}}(m+1)\|_2^2)\right]+nKG\sum_{m=1}^{B}\mathbb{E}\left[\|\bar{\mathbf{x}}(m)-\bar{\mathbf{x}}(m+1)\|_2\right] \\
&+3nKG\sum_{m=1}^{B}u_m+3\delta nGT+\frac{\delta nGRT}{r}.
\end{aligned} \tag{35}$$

Let $\widetilde{f}_m(\mathbf{x})=\bar{\mathbf{d}}(m)^\top\mathbf{x}+\frac{\alpha K}{2}\|\mathbf{x}\|_2^2$. Due to Lemma 2, we have

$$\begin{aligned}
&\sum_{m=1}^{B}\sum_{t\in\mathcal{T}_m}\sum_{j=1}^{n}\mathbb{E}\left[(\nabla\widehat{f}_{t,j,\delta}(\mathbf{x}_j(m))-\alpha\mathbf{x}_j(m))^\top(\bar{\mathbf{x}}(m+1)-\widetilde{\mathbf{x}}^*)\right] \\
=&\sum_{m=1}^{B}\sum_{j=1}^{n}\mathbb{E}\left[(\widehat{\mathbf{g}}_j(m)-\alpha K\mathbf{x}_j(m))^\top(\bar{\mathbf{x}}(m+1)-\widetilde{\mathbf{x}}^*)\right] \\
=&n\sum_{m=1}^{B}\mathbb{E}\left[\bar{\mathbf{d}}(m)^\top(\bar{\mathbf{x}}(m+1)-\widetilde{\mathbf{x}}^*)\right] \\
=&n\sum_{m=1}^{B}\mathbb{E}\left[\widetilde{f}_m(\bar{\mathbf{x}}(m+1))-\widetilde{f}_m(\widetilde{\mathbf{x}}^*)\right]+\sum_{m=1}^{B}\mathbb{E}\left[\frac{n\alpha K}{2}(\|\widetilde{\mathbf{x}}^*\|_2^2-\|\bar{\mathbf{x}}(m+1)\|_2^2)\right].
\end{aligned} \tag{36}$$

According to the definition and (12), we have

$$\bar{\mathbf{x}}(m+1)=\operatorname*{argmin}_{\mathbf{x}\in\mathcal{K}_\delta}\bar{\mathbf{z}}(m)^\top\mathbf{x}+\frac{(m-1)\alpha K}{2}\|\mathbf{x}\|_2^2+h\|\mathbf{x}-\mathbf{x}_{\text{in}}\|_2^2=\operatorname*{argmin}_{\mathbf{x}\in\mathcal{K}_\delta}\sum_{\tau=1}^{m-1}\widetilde{f}_\tau(\mathbf{x})+h\|\mathbf{x}-\mathbf{x}_{\text{in}}\|_2^2.$$

By applying Lemma 5 with the loss functions $\{\widetilde{f}_m(\mathbf{x})\}_{m=1}^B$, the decision set $\mathcal{K}_\delta$, and the regularizer $\mathcal{R}(\mathbf{x})=h\|\mathbf{x}-\mathbf{x}_{\text{in}}\|_2^2$, we have

$$\begin{aligned}
&\sum_{m=1}^{B}\left(\widetilde{f}_m(\bar{\mathbf{x}}(m+1))-\widetilde{f}_m(\widetilde{\mathbf{x}}^*)\right) \\
\leq&h\|\widetilde{\mathbf{x}}^*-\mathbf{x}_{\text{in}}\|_2^2-h\|\bar{\mathbf{x}}(2)-\mathbf{x}_{\text{in}}\|_2^2+\sum_{m=1}^{B}\left(\widetilde{f}_m(\bar{\mathbf{x}}(m+1))-\widetilde{f}_m(\bar{\mathbf{x}}(m+2))\right) \\
\leq&4hR^2+\sum_{m=1}^{B}\|\nabla\widetilde{f}_m(\bar{\mathbf{x}}(m+1))\|_2\|\bar{\mathbf{x}}(m+1)-\bar{\mathbf{x}}(m+2)\|_2 \\
\leq&4hR^2+\sum_{m=1}^{B}\|\bar{\mathbf{d}}(m)+\alpha K\bar{\mathbf{x}}(m+1)\|_2\|\bar{\mathbf{x}}(m+1)-\bar{\mathbf{x}}(m+2)\|_2.
\end{aligned} \tag{37}$$

Note that $\bar{F}_{m+1}(\mathbf{x})$ is $(m\alpha K + 2h)$-strongly convex and $\bar{\mathbf{x}}(m+2) = \operatorname{argmin}_{\mathbf{x}\in\mathcal{K}_\delta} \bar{F}_{m+1}(\mathbf{x})$. For any $m \in [B]$, we have

$$\begin{aligned}
&\frac{m\alpha K + 2h}{2}\|\bar{\mathbf{x}}(m+1) - \bar{\mathbf{x}}(m+2)\|_2^2 \\
\leq&\bar{F}_{m+1}(\bar{\mathbf{x}}(m+1)) - \bar{F}_{m+1}(\bar{\mathbf{x}}(m+2)) \\
=&\bar{F}_m(\bar{\mathbf{x}}(m+1)) + \widetilde{f}_m(\bar{\mathbf{x}}(m+1)) - \bar{F}_m(\bar{\mathbf{x}}(m+2)) - \widetilde{f}_m(\bar{\mathbf{x}}(m+2)) \\
\leq&\nabla\widetilde{f}_m(\bar{\mathbf{x}}(m+1))^\top(\bar{\mathbf{x}}(m+1) - \bar{\mathbf{x}}(m+2)) \\
\leq&\|\bar{\mathbf{d}}(m) + \alpha K\bar{\mathbf{x}}(m+1)\|_2\|\bar{\mathbf{x}}(m+1) - \bar{\mathbf{x}}(m+2)\|_2
\end{aligned}$$

where the first inequality is due to (11) and the second inequality is due to $\bar{\mathbf{x}}(m+1) = \operatorname{argmin}_{\mathbf{x}\in\mathcal{K}_\delta} \bar{F}_m(\mathbf{x})$ and the convexity of $\widetilde{f}_m(\mathbf{x})$.

Moreover, for any $m \in [B]$, the above inequality can be simplified as

$$\|\bar{\mathbf{x}}(m+1) - \bar{\mathbf{x}}(m+2)\|_2 \leq \frac{2\|\bar{\mathbf{d}}(m) + \alpha K\bar{\mathbf{x}}(m+1)\|_2}{m\alpha K + 2h}. \tag{38}$$

By combining (35), (36), (37), and (38), we have

$$\begin{aligned}
&\mathbb{E}\left[R_{T,i}\right] \\
\leq&4nhR^2 + n\sum_{m=1}^{B}\mathbb{E}\left[\frac{2\|\bar{\mathbf{d}}(m) + \alpha K\bar{\mathbf{x}}(m+1)\|_2^2}{m\alpha K + 2h}\right] + nKG\sum_{m=1}^{B}\mathbb{E}\left[\|\bar{\mathbf{x}}(m) - \bar{\mathbf{x}}(m+1)\|_2\right] \\
&+ 3nKG\sum_{m=1}^{B}u_m + 3\delta nGT + \frac{\delta nGRT}{r} \\
\leq&n\sum_{m=1}^{B}\mathbb{E}\left[\frac{2\|\bar{\mathbf{d}}(m) + \alpha K\bar{\mathbf{x}}(m+1)\|_2^2}{m\alpha K + 2h}\right] + nKG\sum_{m=2}^{B}\mathbb{E}\left[\frac{2\|\bar{\mathbf{d}}(m-1) + \alpha K\bar{\mathbf{x}}(m)\|_2}{(m-1)\alpha K + 2h}\right] \\
&+ 3nKG\sum_{m=1}^{B}u_m + 3\delta nGT + \frac{\delta nGRT}{r} + 4nhR^2 \\
\leq&n\sum_{m=1}^{B}\mathbb{E}\left[\frac{2\|\bar{\mathbf{d}}(m) + \alpha K\bar{\mathbf{x}}(m+1)\|_2^2}{m\alpha K + 2h}\right] + nKG\sum_{m=1}^{B}\mathbb{E}\left[\frac{2\|\bar{\mathbf{d}}(m) + \alpha K\bar{\mathbf{x}}(m+1)\|_2}{m\alpha K + 2h}\right] \\
&+ 3nKG\sum_{m=1}^{B}u_m + 3\delta nGT + \frac{\delta nGRT}{r} + 4nhR^2
\end{aligned} \tag{39}$$

where the second inequality is derived by bounding $\|\bar{\mathbf{x}}(m) - \bar{\mathbf{x}}(m+1)\|_2$ using (38) for $m > 1$ and $\bar{\mathbf{x}}(2) = \operatorname{argmin}_{\mathbf{x}\in\mathcal{K}_\delta} \bar{F}_1(\mathbf{x}) = \mathbf{x}_{\text{in}} = \bar{\mathbf{x}}(1)$ for $m = 1$.

With the above inequality, we can establish the specific regret bound for convex losses and strongly convex losses, respectively.

In the following, we first consider the case with convex losses, in which the parameters of our Algorithm 3 are set to $\alpha = 0$, $K = L = \sqrt{T}$, $h = \frac{n^{1/4}dMT^{3/4}}{\sqrt{1-\sigma_2(P)}R}$, $\delta = cT^{-1/4}$.

Because of $\alpha = 0$, $K = \sqrt{T}$, and $\delta = cT^{-1/4}$, we have

$$\begin{aligned}\mathbb{E}[\|\bar{\mathbf{d}}(m) + \alpha K \bar{\mathbf{x}}(m+1)\|_2^2] =& \mathbb{E}[\|\bar{\mathbf{d}}(m)\|_2^2] \leq 2K \left(\frac{dM}{\delta}\right)^2 + 2K^2G^2 + 2(\alpha KR)^2 \\ =& \left(\frac{2d^2M^2}{c^2} + 2G^2\right) T\end{aligned}$$

where the first inequality is due to Lemma 6.

Therefore, with $\alpha = 0$, $K = \sqrt{T}$, $h = \frac{n^{1/4}dMT^{3/4}}{\sqrt{1-\sigma_2(P)}R}$, and $\delta = cT^{-1/4}$, we have

$$\begin{aligned}n \sum_{m=1}^{B} \mathbb{E}\left[\frac{2\|\bar{\mathbf{d}}(m) + \alpha K \bar{\mathbf{x}}(m+1)\|_2^2}{m\alpha K + 2h}\right] \leq& \left(\frac{d^2M^2}{c^2} + G^2\right) \frac{2n^{3/4}\sqrt{1-\sigma_2(P)}RT^{3/4}}{dM} \\ =& O(n^{3/4}T^{3/4}).\end{aligned} \tag{40}$$

Similarly, with $\alpha = 0$, $K = \sqrt{T}$, $h = \frac{n^{1/4}dMT^{3/4}}{\sqrt{1-\sigma_2(P)}R}$, and $\delta = cT^{-1/4}$, we have

$$\begin{aligned}nKG \sum_{m=1}^{B} \mathbb{E}\left[\frac{2\|\bar{\mathbf{d}}(m) + \alpha K \bar{\mathbf{x}}(m+1)\|_2}{m\alpha K + 2h}\right] \leq& \sqrt{\frac{2d^2M^2}{c^2} + 2G^2} \frac{n^{3/4}\sqrt{1-\sigma_2(P)}GRT^{3/4}}{dM} \\ =& O(n^{3/4}T^{3/4}).\end{aligned} \tag{41}$$

Note that $u_1 = 0$ and $u_m = \frac{4R}{\sqrt{L+2}} + \sqrt{2K\left(\frac{dM}{\delta}\right)^2 + 2K^2G^2 + 2(\alpha KR)^2} \frac{\sqrt{n}}{((m-2)\alpha K + 2h)(1-\sigma_2(P))}$ for any $B \geq m \geq 2$. With $\alpha = 0$, $K = L = \sqrt{T}$, $h = \frac{n^{1/4}dMT^{3/4}}{\sqrt{1-\sigma_2(P)}R}$, and $\delta = cT^{-1/4}$, we have

$$\begin{aligned}3nKG \sum_{m=1}^{B} u_m =& 3nKG \sum_{m=2}^{B} \left(\frac{4R}{\sqrt{L+2}} + \sqrt{2K\left(\frac{dM}{\delta}\right)^2 + 2K^2G^2} \frac{\sqrt{n}}{2h(1-\sigma_2(P))}\right) \\ \leq& \frac{12nK(B-1)GR}{\sqrt{L+2}} + \sqrt{2K\left(\frac{dM}{\delta}\right)^2 + 2K^2G^2} \frac{3n^{3/2}K(B-1)G}{2h(1-\sigma_2(P))} \\ \leq& 12nGRT^{3/4} + \sqrt{\frac{2d^2M^2}{c^2} + 2G^2} \frac{3n^{5/4}GRT^{3/4}}{2dM\sqrt{1-\sigma_2(P)}} \\ =& O\left(n^{5/4}(1-\sigma_2(P))^{-1/2}T^{3/4}\right)\end{aligned} \tag{42}$$

Moreover, with $K = \sqrt{T}$, $h = \frac{n^{1/4}dMT^{3/4}}{\sqrt{1-\sigma_2(P)}R}$, and $\delta = cT^{-1/4}$, we have

$$\begin{aligned}3\delta nGT + \frac{\delta nGRT}{r} + 4nhR^2 =& 3cnGT^{3/4} + \frac{cnGRT^{3/4}}{r} + \frac{4n^{5/4}dMRT^{3/4}}{\sqrt{1-\sigma_2(P)}} \\ =& O\left(n^{5/4}(1-\sigma_2(P))^{-1/2}T^{3/4}\right).\end{aligned} \tag{43}$$

By combining (39), (40), (41), (42), and (43), our Algorithm 3 with $\alpha = 0$, $K = L = \sqrt{T}$, $h = \frac{n^{1/4}dMT^{3/4}}{\sqrt{1-\sigma_2(P)}R}$, and $\delta = cT^{-1/4}$ ensures

$$\mathbb{E}\left[R_{T,i}\right] = O\left(n^{5/4}(1-\sigma_2(P))^{-1/2}T^{3/4}\right)$$

for convex losses, which completes the proof of Theorem 5.

We continue to consider the case with the strongly convex losses, in which the parameters of our Algorithm 3 are set to $\alpha > 0$, $K = L = T^{2/3}(\ln T)^{-2/3}$, $\delta = cT^{-1/3}(\ln T)^{1/3}$, and $h = \alpha K$.

With $K = T^{2/3}(\ln T)^{-2/3}$ and $\delta = cT^{-1/3}(\ln T)^{1/3}$, we have

$$\begin{aligned}\mathbb{E}[\|\bar{\mathbf{d}}(m) + \alpha K\bar{\mathbf{x}}(m)\|_2]^2 \leq& \mathbb{E}[\|\bar{\mathbf{d}}(m) + \alpha K\bar{\mathbf{x}}(m)\|_2^2] \leq \mathbb{E}[2\|\bar{\mathbf{d}}(m)\|_2^2] + \mathbb{E}[2\|\alpha K\bar{\mathbf{x}}(m)\|_2^2] \\ \leq& 4K\left(\frac{dM}{\delta}\right)^2 + 4K^2G^2 + 6(\alpha KR)^2 \\ =& \left(\frac{4d^2M^2}{c^2} + 4G^2 + 6\alpha^2R^2\right)\left(\frac{T}{\ln T}\right)^{4/3}\end{aligned}$$

where the third inequality is due to Lemma 6 and Assumption 2.

For brevity, let $C = \frac{4d^2M^2}{c^2} + 4G^2 + 6\alpha^2R^2$. With $\alpha > 0$, $K = T^{2/3}(\ln T)^{-2/3}$, $\delta = cT^{-1/3}(\ln T)^{1/3}$, and $h = \alpha K$, we have

$$\begin{aligned}& n\sum_{m=1}^{B}\mathbb{E}\left[\frac{2\|\bar{\mathbf{d}}(m) + \alpha K\bar{\mathbf{x}}(m+1)\|_2^2}{m\alpha K + 2h}\right] \\ \leq& \frac{2nC}{\alpha K}\left(\frac{T}{\ln T}\right)^{4/3}\sum_{m=1}^{B}\frac{1}{m+2} \leq \frac{2nC}{\alpha K}\left(\frac{T}{\ln T}\right)^{4/3}\sum_{m=1}^{B}\frac{1}{m} \leq \frac{2nCT^{2/3}}{\alpha(\ln T)^{2/3}}(1 + \ln B) \\ \leq& \frac{2nCT^{2/3}}{\alpha(\ln T)^{2/3}} + \frac{2nCT^{2/3}(\ln T)^{1/3}}{\alpha} = O(nT^{2/3}(\log T)^{1/3})\end{aligned} \tag{44}$$

where the last inequality is due to $B \leq T$.

Similarly, with $\alpha > 0$, $K = T^{2/3}(\ln T)^{-2/3}$, $\delta = cT^{-1/3}(\ln T)^{1/3}$, and $h = \alpha K$, we have

$$\begin{aligned}& nKG\sum_{m=1}^{B}\mathbb{E}\left[\frac{2\|\bar{\mathbf{d}}(m) + \alpha K\bar{\mathbf{x}}(m+1)\|_2}{m\alpha K + 2h}\right] \\ \leq& \frac{2nG\sqrt{C}}{\alpha}\left(\frac{T}{\ln T}\right)^{2/3}\sum_{m=1}^{B}\frac{1}{m+2} \leq \frac{2nG\sqrt{C}}{\alpha}\left(\frac{T}{\ln T}\right)^{2/3}\sum_{m=1}^{B}\frac{1}{m} \\ \leq& \frac{2nG\sqrt{C}}{\alpha}\left(\frac{T}{\ln T}\right)^{2/3}(1 + \ln B) = O(nT^{2/3}(\log T)^{1/3}).\end{aligned} \tag{45}$$

Moreover, with $\alpha > 0$, $K = L = T^{2/3}(\ln T)^{-2/3}$, $\delta = cT^{-1/3}(\ln T)^{1/3}$, and $h = \alpha K$, we have

$$\begin{aligned}u_m =& \frac{4R}{\sqrt{L+2}} + \sqrt{2K\left(\frac{dM}{\delta}\right)^2 + 2K^2G^2 + 2(\alpha KR)^2}\frac{\sqrt{n}}{m\alpha K(1-\sigma_2(P))} \\ \leq& \frac{4R(\ln T)^{1/3}}{T^{1/3}} + \frac{\sqrt{Cn}}{\sqrt{2}m\alpha(1-\sigma_2(P))}\end{aligned}$$

for any $B \geq m \geq 2$.

Then, with $u_1 = 0$, $\alpha > 0$, $K = T^{2/3}(\ln T)^{-2/3}$, $\delta = cT^{-1/3}(\ln T)^{1/3}$, and $h = \alpha K$, we have

$$\begin{aligned}3nKG\sum_{m=1}^{B} u_m \leq & \frac{12nK(B-1)GR(\ln T)^{1/3}}{T^{1/3}} + \frac{3n\sqrt{Cn}G}{\sqrt{2}\alpha(1-\sigma_2(P))}\left(\frac{T}{\ln T}\right)^{2/3}\sum_{m=1}^{B}\frac{1}{m} \\ \leq & 12nGRT^{2/3}(\ln T)^{1/3} + \frac{3n\sqrt{Cn}G}{\sqrt{2}\alpha(1-\sigma_2(P))}\left(\frac{T}{\ln T}\right)^{2/3}(1+\ln T) \\ = & O\left(n^{3/2}(1-\sigma_2(P))^{-1}T^{2/3}(\log T)^{1/3}\right).\end{aligned} \tag{46}$$

Moreover, with $\alpha > 0$, $K = T^{2/3}(\ln T)^{-2/3}$, $\delta = cT^{-1/3}(\ln T)^{1/3}$, and $h = \alpha K$, we have

$$\begin{aligned}&3\delta nGT + \frac{\delta nGRT}{r} + 4nhR^2 \\ &= \left(3cnG + \frac{cnGR}{r}\right)T^{2/3}(\ln T)^{1/3} + 4n\alpha R^2 T^{2/3}(\ln T)^{-2/3} = O(nT^{2/3}(\ln T)^{1/3}).\end{aligned} \tag{47}$$

Finally, by combining (39), (44), (45), (46), and (47), our Algorithm 3 with $\alpha > 0$, $K = L = T^{2/3}(\ln T)^{-2/3}$, $\delta = cT^{-1/3}(\ln T)^{1/3}$, and $h = \alpha K$ ensures

$$\mathbb{E}\left[R_{T,i}\right] = O\left(n^{3/2}(1-\sigma_2(P))^{-1}T^{2/3}(\log T)^{1/3}\right)$$

for α-strongly convex losses, which completes the proof of Theorem 6.

6.3.2 Proof of Lemma 6

We first notice that

$$\|\mathbf{d}_i(m)\|_2^2 = \|\widehat{\mathbf{g}}_i(m) - \alpha K\mathbf{x}_i(m)\|_2^2 \leq 2\|\widehat{\mathbf{g}}_i(m)\|_2^2 + 2\|\alpha K\mathbf{x}_i(m)\|_2^2 \leq 2\|\widehat{\mathbf{g}}_i(m)\|_2^2 + 2(\alpha KR)^2$$

where the last inequality is due to Assumption 2.

Moreover, it is easy to provide an upper bound of $\mathbb{E}[\|\widehat{\mathbf{g}}_i(m)\|_2^2]$ by following the proof of Lemma 5 in Garber and Kretzu (2020). We include the detailed proof for completeness.

Let $t_j = (m-1)K + j$ for $j = 1, \ldots, K$. We have

$$\begin{aligned}&\mathbb{E}\left[\|\widehat{\mathbf{g}}_i(m)\|_2^2|\mathbf{x}_i(m)\right] \\ =&\mathbb{E}\left[\sum_{j=1}^{K}\mathbf{g}_i(t_j)^\top\mathbf{g}_i(t_j)\middle|\mathbf{x}_i(m)\right] + \mathbb{E}\left[\sum_{j=1}^{K}\sum_{k\in[K]\cap k\neq j}\mathbf{g}_i(t_j)^\top\mathbf{g}_i(t_k)\middle|\mathbf{x}_i(m)\right] \\ =&\mathbb{E}\left[\sum_{j=1}^{K}\|\mathbf{g}_i(t_j)\|_2^2\middle|\mathbf{x}_i(m)\right] + \sum_{j=1}^{K}\sum_{k\in[K]\cap k\neq j}\mathbb{E}\left[\mathbf{g}_i(t_j)|\,\mathbf{x}_i(m)\right]^\top\mathbb{E}\left[\mathbf{g}_i(t_k)|\,\mathbf{x}_i(m)\right] \\ \leq&K\left(\frac{dM}{\delta}\right)^2 + \sum_{j=1}^{K}\sum_{k\in[K]\cap k\neq j}\|\mathbb{E}\left[\mathbf{g}_i(t_j)|\,\mathbf{x}_i(m)\right]\|_2\|\mathbb{E}\left[\mathbf{g}_i(t_k)|\,\mathbf{x}_i(m)\right]\|_2 \\ \leq&K\left(\frac{dM}{\delta}\right)^2 + (K^2-K)G^2 \\ \leq&K\left(\frac{dM}{\delta}\right)^2 + K^2G^2\end{aligned} \tag{48}$$

Data Set	# Features	# Classes	# Examples
a9a	123	2	32561
ijcnn1	22	2	49990
aloi	128	1000	108000
news20	62061	20	15935

Table 1: Summary of data sets.

where the second inequality is due to Lemmas 2 and 8.

Therefore, we have

$$\begin{aligned}\mathbb{E}[\|\mathbf{d}_i(m)\|_2^2] \leq & 2\mathbb{E}[\|\widehat{\mathbf{g}}_i(m)\|_2^2] + 2(\alpha KR)^2 = 2\mathbb{E}\left[\mathbb{E}\left[\|\widehat{\mathbf{g}}_i(m)\|_2^2 | \mathbf{x}_i(m)\right]\right] + 2(\alpha KR)^2 \\ \leq & 2K\left(\frac{dM}{\delta}\right)^2 + 2K^2G^2 + 2(\alpha KR)^2.\end{aligned}$$

Moreover, according to the Jensen's inequality, we have

$$\mathbb{E}[\|\mathbf{d}_i(m)\|_2]^2 \leq \mathbb{E}[\|\mathbf{d}_i(m)\|_2^2].$$

7. Experiments

In this section, we perform simulation experiments on the multiclass classification problem and the binary classification problem to verify the performance of our proposed algorithms.

7.1 Data Sets and Topologies of the Networks

We conduct experiments on four publicly available data sets—aloi, news20, a9a, and ijcnn1 from the LIBSVM repository (Chang and Lin, 2011), and the details of these data sets are summarized in Table 1. Specifically, aloi and news20 are used in the multiclass classification problem, and the other two data sets are used in the binary classification problem. For any data set, let T_e denote the number of examples. We first divide it into n equally-sized parts where each part contains $\lfloor T_e/n \rfloor$ examples, and then distribute them onto n computing nodes in the network,[1] where $n = 9$ for the multiclass classification problem and $n = 100$ for the binary classification problem. Moreover, each part of the data set will be reused n times, which implies that the number of rounds T is equal to $n\lfloor T_e/n \rfloor$.

To model the distributed network, we will use three types of graphs including a complete graph, a two-dimensional grid graph, and a cycle graph. The complete graph is a "well connected" network, where each node is connected to all other nodes. In contrast, the cycle graph is a "poorly connected" network, each node of which is only connected to two other nodes. Moreover, in the two-dimensional grid graph, each node not in the boundary is connected to its four nearest neighbors in axis-aligned directions. Its connectivity is between that of the complete graph and the cycle graph.

For the weight matrix P, we first compute P_{ij} for $i \neq j$ as

$$P_{ij} = \begin{cases} 0, & \text{if } j \notin N_i, \\ 1/\max(|N_i|, |N_j|), & \text{if } j \in N_i. \end{cases}$$

1. The remaining $T_e - n\lfloor T_e/n \rfloor$ examples are not used.

Then, we compute $P_{ij} = 1 - \sum_{q \in N_i, q \neq i} P_{iq}$ for $i = j$. In this way, we can ensure that P satisfies Assumption 3 for all three types of graphs.

7.2 Multiclass Classification

Following Zhang et al. (2017), we first compare our D-BOCG against their D-OCG by conducting experiments on distributed online multiclass classification. Let k be the number of features, and let v be the number of classes. In the t-th round, after receiving a single example $\mathbf{e}_i(t) \in \mathbb{R}^k$, each local learner i chooses a decision matrix $X_i(t) = [\mathbf{x}_1^\top; \mathbf{x}_2^\top; \ldots; \mathbf{x}_v^\top] \in \mathbb{R}^{v \times k}$ from the convex set

$$\mathcal{K} = \{X \in \mathbb{R}^{v \times k} | \|X\|_* \leq \tau\}$$

where $\|X\|_*$ denotes the trace norm of X and τ is set to be 50. Note that $X_i(t)$ can be utilized to predict the class label of $\mathbf{e}_i(t)$ by computing $\operatorname{argmax}_{\ell \in [v]} \mathbf{x}_\ell^\top \mathbf{e}_i(t)$. Then, the true class label $y_i(t) \in \{1, \ldots, v\}$ is revealed, which incurs the multivariate logistic loss

$$f_{t,i}(X_i(t)) = \ln \left(1 + \sum_{\ell \neq y_i(t)} e^{\mathbf{x}_\ell^\top \mathbf{e}_i(t) - \mathbf{x}_{y_i(t)}^\top \mathbf{e}_i(t)} \right).$$

The average loss of node i at the t-th round is defined as

$$AL(t, i) = \frac{1}{tn} \sum_{q=1}^{t} \sum_{j=1}^{n} f_{q,j}(X_i(q)). \tag{49}$$

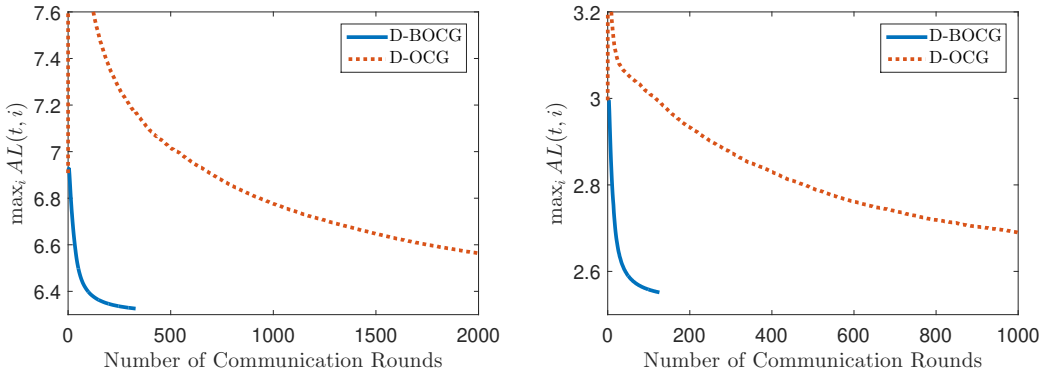
For both methods, we simply initialize $X_i(1) = \mathbf{0}_{v \times k}, \forall i \in [n]$. According to Zhang et al. (2017), we set $s_t = 1/\sqrt{t}$ and $\eta = cT^{-3/4}$ for D-OCG by tuning the constant c. Because the multivariate logistic loss is not strongly convex, the parameters of our D-BOCG are selected according to Corollary 1. Specifically, we set $\alpha = 0$, $K = L = \lfloor \sqrt{T} \rfloor$, and $h = T^{3/4}/c$ by tuning the constant c. For both methods, the constant c is selected from $\{0.01, \ldots, 1e5\}$.

Fig. 1 shows the comparisons of our D-BOCG and D-OCG on distributed online multiclass classification over the complete graph. We find that the average loss of the worst local node in D-BOCG decreases faster than that of the worst local node in D-OCG with the increasing of communication rounds, which verifies our theoretical results about the regret bound and communication complexity of D-BOCG. Furthermore, Fig. 2 shows comparisons of D-BOCG on distributed online multiclass classification over different graphs. We find that with the improvement of the graph connectivity, the convergence of our D-BOCG is slightly improved, which is also consistent with our theoretical results about the regret bound of D-BOCG.

7.3 Binary Classification

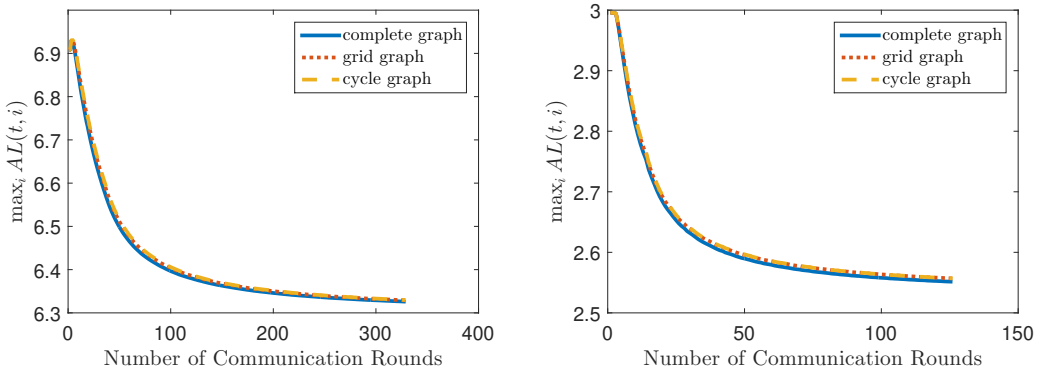
We also consider the problem of binary classification in the distributed online learning setting. In the t-th round, each local learner i receives a single example $\mathbf{e}_i(t) \in \mathbb{R}^d$ and chooses a decision $\mathbf{x}_i(t) \in \mathbb{R}^d$ from the convex set

$$\mathcal{K} = \{\mathbf{x} \in \mathbb{R}^d | \|\mathbf{x}\|_1 \leq \tau\}$$



(a) aloi (b) news20

Figure 1: Comparisons of D-BOCG and D-OCG on distributed online multiclass classification over the complete graph.



(a) aloi (b) news20

Figure 2: Comparisons of D-BOCG on distributed online multiclass classification over different graphs.

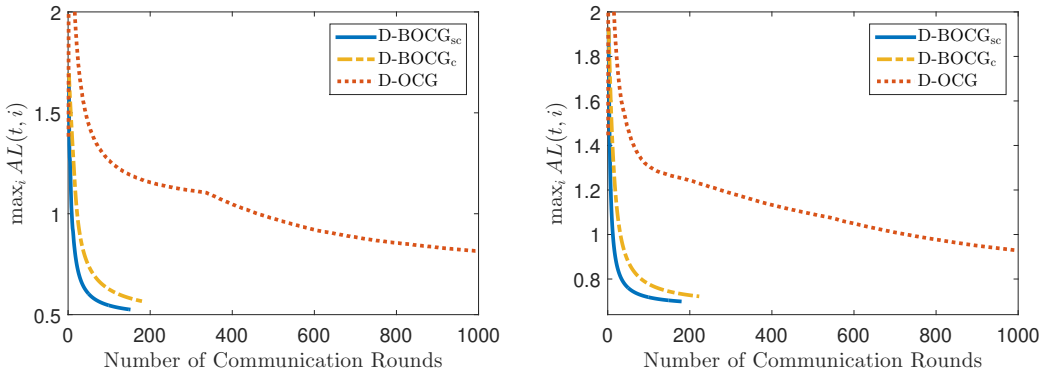
where τ is set to be 10. Then, the true class label $y_i(t) \in \{-1, 1\}$ is revealed, and it suffers the regularized hinge loss

$$f_{t,i}(\mathbf{x}_i(t)) = \max\Big\{1 - y_i(t)\mathbf{e}_i(t)^\top \mathbf{x}_i(t), 0\Big\} + \lambda\|\mathbf{x}_i(t)\|_2^2$$

where λ is set to be 0.1. Similar to (49), the average loss of node i at the t-th round is defined as

$$AL(t,i) = \frac{1}{tn}\sum_{q=1}^{t}\sum_{j=1}^{n} f_{q,j}(\mathbf{x}_i(q)).$$

Note that the regularized hinge loss is 2λ-strongly convex. To utilize the strong convexity, we can set parameters of D-BOCG according to Corollary 2. Moreover, to show the advantage



(a) a9a (b) ijcnn1

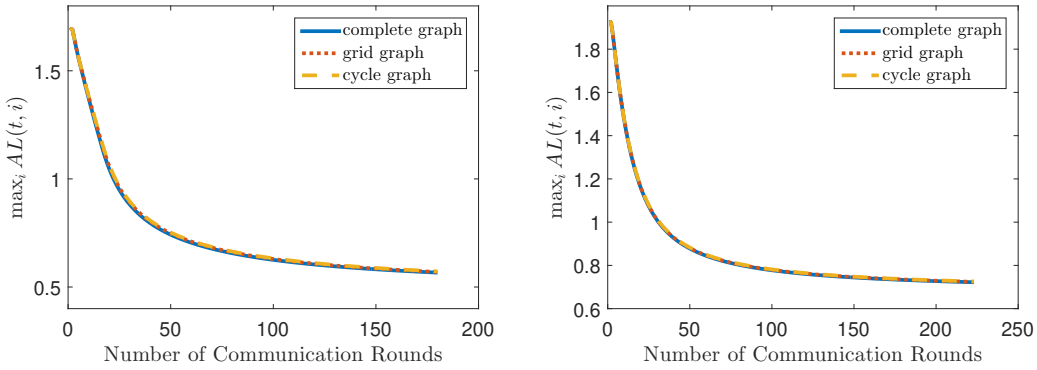
Figure 3: Comparisons of D-OCG, D-BOCG$_c$, and D-BOCG$_{sc}$ on distributed online binary classification over the complete graph.

of utilizing the strong convexity, we also run D-BOCG with parameters in Corollary 1, which only utilizes the convexity condition. To distinguish these two different instances of D-BOCG, we denote D-BOCG with parameters in Corollary 2 as D-BOCG$_{sc}$, and D-BOCG with parameters in Corollary 1 as D-BOCG$_c$.

For D-OCG, D-BOCG$_c$, and D-BOCG$_{sc}$, we simply initialize $\mathbf{x}_i(1) = \tau\mathbf{1}/d, \forall i \in [n]$, where $\mathbf{1}$ denotes the vector with each entry equal 1. Other parameters of D-OCG are set in the same way as D-OCG in previous experiments, and other parameters of D-BOCG$_c$ are set in the same way as D-BOCG in previous experiments. For D-BOCG$_{sc}$, according to Corollary 2, we set $\alpha = 2\lambda$ and $K = L = \lfloor T^{2/3}(\ln T)^{-2/3} \rfloor$. Moreover, although we use $h = \alpha K$ in Corollary 2, in the experiments, we set $h = c'\alpha K$ by tuning the constant c' from $\{1, 2, 3, 4, 5\}$. It is easy to verify that the modified h only affects the constant factor of the original regret bound in Corollary 2.

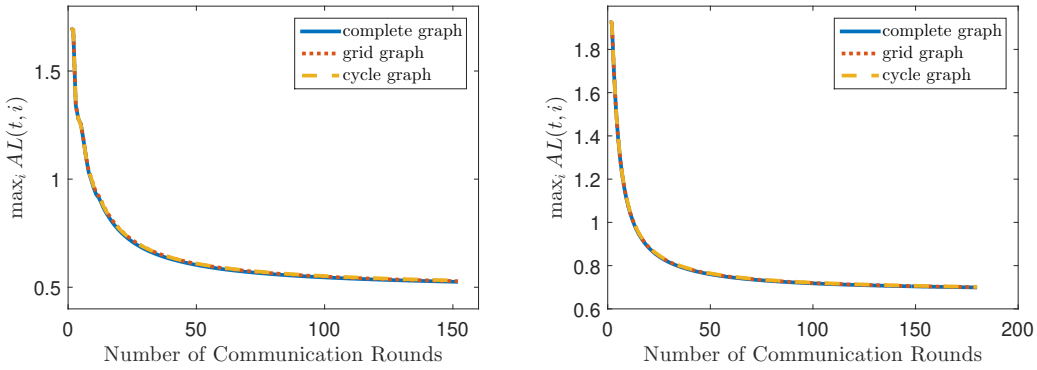
Fig. 3 shows comparisons of D-OCG, D-BOCG$_c$, and D-BOCG$_{sc}$ on distributed online binary classification over the complete graph. First, the average loss of the worst local node in D-BOCG$_c$ and D-BOCG$_{sc}$ decreases faster than that of the worst local node in D-OCG with the increasing of communication rounds, which validates our advantage in the communication complexity again. Moreover, our D-BOCG$_{sc}$ outperforms D-BOCG$_c$, which further validates the advantage of utilizing the strong convexity. Fig. 4 and 5 show comparisons of D-BOCG$_c$ and D-BOCG$_{sc}$ on distributed online binary classification over different graphs. We find that the effect of the graph connectivity is similar to that presented in Fig. 2, though the number of nodes increases from 9 to 100.

Then, to verify the performance of our D-BBCG, we compare it with our D-BOCG. Note that D-BBCG only uses approximate gradients generated by the one-point gradient estimator, the performance of which is highly affected by the dimensionality. Therefore, to make a fair comparison, we only use ijcnn1, the dimensionality of which is relatively small. Specifically, we denote D-BBCG with parameters in Theorem 5 as D-BBCG$_c$, and D-BBCG with parameters in Theorem 6 as D-BBCG$_{sc}$. According to Theorems 5 and 6, we set $\alpha = 0$,



(a) a9a (b) ijcnn1

Figure 4: Comparisons of D-BOCG$_c$ on distributed online binary classification over different graphs.

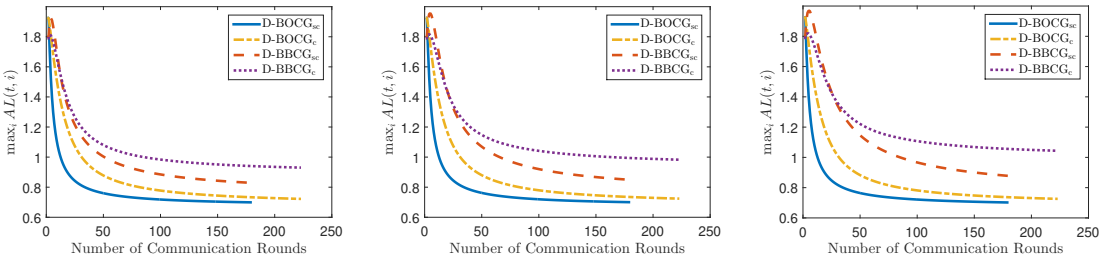


(a) a9a (b) ijcnn1

Figure 5: Comparisons of D-BOCG$_{sc}$ on distributed online binary classification over different graphs.

$K = L = \lfloor \sqrt{T} \rfloor$, $\delta = 10T^{-1/4}$, and $h = T^{3/4}/c$ for D-BBCG$_c$ where the constant c is tuned from $\{0.01, \ldots, 1e5\}$, and set $\alpha = 2\lambda$, $K = L = \lfloor T^{2/3}(\ln T)^{-2/3} \rfloor$, $\delta = 10T^{-1/3}(\ln T)^{1/3}$, and $h = c'\alpha K$ where the constant c' is tuned from $\{1, 2, 3, 4, 5\}$. Moreover, we initialize $\mathbf{x}_i(1) = (1 - \delta\sqrt{d}/\tau)\mathbf{1}/d, \forall i \in [n]$ for both D-BBCG$_c$ and D-BBCG$_{sc}$. Since D-BBCG$_c$ and D-BBCG$_{sc}$ are randomized algorithms, we repeat them 10 times and report the average results.

Fig. 6 shows comparisons of D-BOCG$_c$, D-BOCG$_{sc}$, D-BBCG$_c$, and D-BBCG$_{sc}$ on distributed online binary classification for ijcnn1. For all three types of graphs, we find that D-BBCG$_c$ is worse than D-BOCG$_c$ and D-BBCG$_{sc}$ is worse than D-BOCG$_{sc}$, which is reasonable because D-BBCG$_c$ and D-BBCG$_{sc}$ are working with the more challenging bandit



(a) complete graph (b) grid graph (c) cycle graph

Figure 6: Comparisons of D-BOCG$_{\mathrm{c}}$, D-BOCG$_{\mathrm{sc}}$, D-BBCG$_{\mathrm{c}}$, and D-BBCG$_{\mathrm{sc}}$ on distributed online binary classification for ijcnn1.

setting. Moreover, D-BBCG$_{\mathrm{sc}}$ is better than D-BBCG$_{\mathrm{c}}$, which validates the advantage of utilizing the strong convexity in the bandit setting.

8. Conclusion and Future Work

In this paper, we first propose a projection-free algorithm called D-BOCG for distributed online convex optimization. Our analysis shows that D-BOCG enjoys an $O(T^{3/4})$ regret bound with $O(\sqrt{T})$ communication rounds for convex losses, and a better regret bound of $O(T^{2/3}(\log T)^{1/3})$ with fewer $O(T^{1/3}(\log T)^{2/3})$ communication rounds for strongly convex losses. In the case with convex losses, the $O(T^{3/4})$ regret bound of D-BOCG matches the best result established by the existing projection-free algorithm with T communication rounds, and the $O(\sqrt{T})$ communication rounds required by D-BOCG match (in terms of T) the lower bound for any distributed online algorithm attaining the $O(T^{3/4})$ regret. In the case with strongly convex losses, we also provide a lower bound to show that the $O(T^{1/3}(\log T)^{2/3})$ communication rounds required by D-BOCG are nearly optimal (in terms of T) for obtaining the $O(T^{2/3}(\log T)^{1/3})$ regret bound up to polylogarithmic factors. Furthermore, to handle the bandit setting, we propose a bandit variant of D-BOCG, namely D-BBCG, and obtain similar theoretical guarantees.

Besides the future work discussed before, there are still several open problems to be investigated. First, in the standard OCO, Hazan and Minasyan (2020) have proposed a projection-free algorithm that obtains an expected regret bound of $O(T^{2/3})$ for convex and smooth losses. It is interesting to extend their algorithm to the distributed setting studied in this paper. However, their algorithm is not based on conditional gradient, which makes the extension non-trivial. Second, in this paper, the weight matrix P is assumed to be symmetric and doubly stochastic. It is appealing to consider a more practical scenario, in which P could be asymmetric or only column (or row) stochastic (Yang et al., 2019; Yi et al., 2020). Finally, we will investigate whether the regret bound for the full information setting can be improved if a few projections are allowed. We note that $O(\log T)$ projections are sufficient to achieve the optimal convergence rate for stochastic optimization of smooth and strongly convex functions (Zhang et al., 2013).

Acknowledgments

This work was partially supported by NSFC (62122037, 61921006), and JiangsuSF (BK20200 064). We are grateful for the anonymous reviewers and the editor for their helpful comments. We also thank an anonymous reviewer of ICML 2020 for suggesting us investigating the lower bound of communication complexity.

Appendix A. Proof of Corollaries 1 and 2

Corollary 1 can be proved by substituting $\alpha = 0$, $K = L = \sqrt{T}$, and $h = \frac{n^{1/4}T^{3/4}G}{\sqrt{1-\sigma_2(P)}R}$ into Theorem 2, as follows

$$\begin{aligned}
R_{T,i} \leq & \frac{12nGRT}{\sqrt{\sqrt{T}+2}} + \sum_{m=2}^{B} \frac{3n^{5/4}T^{1/4}GR}{2\sqrt{1-\sigma_2(P)}} + \sum_{m=1}^{B} 2\sqrt{1-\sigma_2(P)}n^{3/4}T^{1/4}GR + \frac{4n^{5/4}T^{3/4}GR}{\sqrt{1-\sigma_2(P)}} \\
\leq & 12nGRT^{3/4} + \frac{11n^{5/4}T^{3/4}GR}{2\sqrt{1-\sigma_2(P)}} + 2\sqrt{1-\sigma_2(P)}n^{3/4}T^{3/4}GR
\end{aligned}$$

where the last inequality is due to $B - 1 < B = T/K = \sqrt{T}$.

Corollary 2 can be proved by substituting $\alpha > 0$, $K = L = T^{2/3}(\ln T)^{-2/3}$, and $h = \alpha K$ into Theorem 2, as follows

$$\begin{aligned}
R_{T,i} \leq & \frac{12nGRT}{\sqrt{L}} + \sum_{m=2}^{B} \frac{3nGK(G+\alpha R)\sqrt{n}}{m\alpha(1-\sigma_2(P))} + \sum_{m=1}^{B} \frac{4nK(G+2\alpha R)^2}{(m+2)\alpha} + 4nhR^2 \\
\leq & \frac{12nGRT}{\sqrt{L}} + \left(\frac{3nG(G+\alpha R)\sqrt{n}}{\alpha(1-\sigma_2(P))} + \frac{4n(G+2\alpha R)^2}{\alpha}\right)\sum_{m=1}^{B}\frac{K}{m} + 4nhR^2 \\
\leq & 12nGRT^{2/3}(\ln T)^{1/3} + \left(\frac{3nG(G+\alpha R)\sqrt{n}}{\alpha(1-\sigma_2(P))} + \frac{4n(G+2\alpha R)^2}{\alpha}\right)K(1+\ln B) + 4nhR^2 \\
\leq & \left(\frac{3n^{3/2}G(G+\alpha R)}{\alpha(1-\sigma_2(P))} + \frac{4n(G+2\alpha R)^2}{\alpha}\right)T^{2/3}((\ln T)^{-2/3} + (\ln T)^{1/3}) \\
& + 12nGRT^{2/3}(\ln T)^{1/3} + 4n\alpha R^2 T^{2/3}(\ln T)^{-2/3}
\end{aligned}$$

where the last inequality is due to $K = T^{2/3}(\ln T)^{-2/3}$ and $\ln B \leq \ln T$.

Appendix B. Proof of Theorem 7

In the beginning, we define several auxiliary variables. Let $\bar{\mathbf{z}}(m) = \frac{1}{n}\sum_{i=1}^{n}\mathbf{z}_i(m)$ for any $m \in [B+1]$ and $\bar{\mathbf{g}}(m) = \frac{1}{n}\sum_{i=1}^{n}\widehat{\mathbf{g}}_i(m)$ for any $m \in [B]$. Then, we define $\bar{\mathbf{x}}(1) = \mathbf{x}_{\text{in}}$ and $\bar{\mathbf{x}}(m+1) = \operatorname{argmin}_{\mathbf{x}\in\mathcal{K}_\delta}\bar{F}_m(\mathbf{x})$ for any $m \in [B+1]$, where

$$\bar{F}_m(\mathbf{x}) = \bar{\mathbf{z}}(m)^\top\mathbf{x} + h\|\mathbf{x} - \mathbf{x}_{\text{in}}\|_2^2.$$

Similarly, we define $\widehat{\mathbf{x}}_i(m+1) = \operatorname{argmin}_{\mathbf{x}\in\mathcal{K}_\delta}F_{m,i}(\mathbf{x})$ for any $m \in [B+1]$, where

$$F_{m,i}(\mathbf{x}) = \mathbf{z}_i(m)^\top\mathbf{x} + h\|\mathbf{x} - \mathbf{x}_{\text{in}}\|_2^2$$

is defined in Algorithm 3 when $\alpha = 0$.

Moreover, let $\mathbf{x}^* \in \operatorname{argmin}_{\mathbf{x}\in\mathcal{K}} \sum_{t=1}^{T} f_t(\mathbf{x})$ and $\widetilde{\mathbf{x}}^* = (1-\delta/r)\mathbf{x}^*$. For any $j \in V$ and $t \in [T]$, we define the δ-smoothed version of $f_{t,j}(\mathbf{x})$ as

$$\widehat{f}_{t,j,\delta}(\mathbf{x}) = \mathbb{E}_{\mathbf{u}\sim\mathcal{B}^d}[f_{t,j}(\mathbf{x}+\delta\mathbf{u})]$$

where $\mathcal{B}^d$ denotes the unit Euclidean ball centered at the origin in $\mathbb{R}^d$. Note that as in (34), we have proved that Algorithm 3 ensures

$$R_{T,i} \leq \sum_{m=1}^{B}\sum_{t\in\mathcal{T}_m}\sum_{j=1}^{n}(\widehat{f}_{t,j,\delta}(\mathbf{x}_i(m)) - \widehat{f}_{t,j,\delta}(\widetilde{\mathbf{x}}^*)) + 3\delta nGT + \frac{\delta nGRT}{r}. \tag{50}$$

To bound the term $\sum_{m=1}^{B}\sum_{t\in\mathcal{T}_m}\sum_{j=1}^{n}(\widehat{f}_{t,j,\delta}(\mathbf{x}_i(m)) - \widehat{f}_{t,j,\delta}(\widetilde{\mathbf{x}}^*))$ in (50), we assume that for all $i \in V$ and $m = 1, \ldots, B$, Algorithm 3 ensures that

$$\|\widehat{\mathbf{g}}_i(m)\|_2 \leq \widehat{G} = \xi_T\frac{dM\sqrt{K}}{\delta} + KG.$$

Then, we can derive an upper bound of $\|\mathbf{x}_i(m) - \bar{\mathbf{x}}(m)\|_2$. For any $B \geq m \geq 2$, we note that $F_{m-1,i}(\mathbf{x})$ is $2h$-smooth, and Algorithm 3 ensures $\mathbf{x}_i(m) = \mathrm{CG}(\mathcal{K}_\delta, L, F_{m-1,i}(\mathbf{x}), \mathbf{x}_i(m-1))$. According to Lemma 1, Assumption 2, and $\mathcal{K}_\delta \subseteq \mathcal{K}$, for $B \geq m \geq 2$, it is easy to verify that

$$F_{m-1,i}(\mathbf{x}_i(m)) - F_{m-1,i}(\widehat{\mathbf{x}}_i(m)) \leq \frac{16hR^2}{L+2}.$$

Then, for any $B \geq m \geq 2$, we have

$$\begin{aligned}
\|\mathbf{x}_i(m) - \bar{\mathbf{x}}(m)\|_2 \leq& \|\mathbf{x}_i(m) - \widehat{\mathbf{x}}_i(m)\|_2 + \|\widehat{\mathbf{x}}_i(m) - \bar{\mathbf{x}}(m)\|_2\\
\leq& \sqrt{\frac{F_{m-1,i}(\mathbf{x}_i(m)) - F_{m-1,i}(\widehat{\mathbf{x}}_i(m))}{h}} + \|\widehat{\mathbf{x}}_i(m) - \bar{\mathbf{x}}(m)\|_2\\
\leq& \frac{4R}{\sqrt{L+2}} + \|\widehat{\mathbf{x}}_i(m) - \bar{\mathbf{x}}(m)\|_2\\
\leq& \frac{4R}{\sqrt{L+2}} + \frac{1}{2h}\|\mathbf{z}_i(m) - 2h\mathbf{x}_{\mathrm{in}} - \bar{\mathbf{z}}(m) + 2h\mathbf{x}_{\mathrm{in}}\|_2\\
\leq& \frac{4R}{\sqrt{L+2}} + \frac{\widehat{G}\sqrt{n}}{2h(1-\sigma_2(P))}
\end{aligned} \tag{51}$$

where the second inequality is due to the fact that $F_{m-1,i}(\mathbf{x})$ is $2h$-strongly convex and (11), the fourth inequality is due to Lemma 4, and the last inequality is due to Lemma 3.

For brevity, let

$$\epsilon = \frac{4R}{\sqrt{L+2}} + \frac{\widehat{G}\sqrt{n}}{2h(1-\sigma_2(P))}.$$

By combining (51) with $\mathbf{x}_i(1) = \bar{\mathbf{x}}(m) = \mathbf{x}_{\mathrm{in}}$, for any $m \in [B]$, we have

$$\|\mathbf{x}_i(m) - \bar{\mathbf{x}}(m)\|_2 \leq \epsilon. \tag{52}$$

For any $i, j \in V$, $m \in [B]$, and $t \in \mathcal{T}_m$, according to Lemma 8 and Assumption 1, $\widehat{f}_{t,j,\delta}(\mathbf{x})$ is also convex and G-Lipschitz. Then, by combining with (52), we have

$$\begin{aligned}\widehat{f}_{t,j,\delta}(\mathbf{x}_i(m)) - \widehat{f}_{t,j,\delta}(\widetilde{\mathbf{x}}^*) \leq & \widehat{f}_{t,j,\delta}(\bar{\mathbf{x}}(m)) - \widehat{f}_{t,j,\delta}(\widetilde{\mathbf{x}}^*) + G\|\bar{\mathbf{x}}(m) - \mathbf{x}_i(m)\|_2 \\ \leq & \widehat{f}_{t,j,\delta}(\mathbf{x}_j(m)) - \widehat{f}_{t,j,\delta}(\widetilde{\mathbf{x}}^*) + G\|\bar{\mathbf{x}}(m) - \mathbf{x}_j(m)\|_2 + G\epsilon \\ \leq & \nabla \widehat{f}_{t,j,\delta}(\mathbf{x}_j(m))^\top (\mathbf{x}_j(m) - \bar{\mathbf{x}}(m) + \bar{\mathbf{x}}(m) - \widetilde{\mathbf{x}}^*) + 2G\epsilon \\ \leq & \nabla \widehat{f}_{t,j,\delta}(\mathbf{x}_j(m))^\top (\bar{\mathbf{x}}(m) - \widetilde{\mathbf{x}}^*) + 3G\epsilon.\end{aligned} \tag{53}$$

By combining (50) with (53), for any $i \in V$, we have

$$R_{T,i} \leq \sum_{m=1}^{B} \sum_{t \in \mathcal{T}_m} \sum_{j=1}^{n} \nabla \widehat{f}_{t,j,\delta}(\mathbf{x}_j(m))^\top (\bar{\mathbf{x}}(m) - \widetilde{\mathbf{x}}^*) + 3nGT\epsilon + 3\delta nGT + \frac{\delta nGRT}{r}.$$

Then, to bound $\sum_{m=1}^{B} \sum_{t \in \mathcal{T}_m} \sum_{j=1}^{n} \nabla \widehat{f}_{t,j,\delta}(\mathbf{x}_j(m))^\top (\bar{\mathbf{x}}(m) - \widetilde{\mathbf{x}}^*)$, we introduce the following lemma.

Lemma 9 *Let $\bar{\mathbf{z}}(m) = \frac{1}{n}\sum_{i=1}^{n} \mathbf{z}_i(m)$ for any $m \in [B+1]$ and $\bar{\mathbf{g}}(m) = \frac{1}{n}\sum_{i=1}^{n} \widehat{\mathbf{g}}_i(m)$ for any $m \in [B]$. Define $\bar{\mathbf{x}}(1) = \mathbf{x}_{\text{in}}$, where $\mathbf{x}_{\text{in}}$ is an input of Algorithm 3. Moreover, define $\bar{F}_m(\mathbf{x}) = \bar{\mathbf{z}}(m)^\top \mathbf{x} + h\|\mathbf{x} - \mathbf{x}_{\text{in}}\|_2^2$ and $\bar{\mathbf{x}}(m+1) = \operatorname{argmin}_{\mathbf{x} \in \mathcal{K}_\delta} \bar{F}_m(\mathbf{x})$ for any $m \in [B+1]$. Under Assumptions 1, 2, 3, and an additional assumption that $\|\widehat{\mathbf{g}}_i(m)\|_2 \leq \widehat{G}$ for any $i \in V$ and $m \in [B]$, with probability at least $1-\gamma$, Algorithm 3 with $\alpha = 0$ has*

$$\sum_{m=1}^{B} \sum_{t \in \mathcal{T}_m} \sum_{j=1}^{n} \nabla \widehat{f}_{t,j,\delta}(\mathbf{x}_j(m))^\top (\bar{\mathbf{x}}(m) - \widetilde{\mathbf{x}}^*) \leq 2nR(KG + \widehat{G})\sqrt{2B \ln \frac{1}{\gamma}} + 4nhR^2 + \frac{2nB\widehat{G}^2}{h}$$

where $\widetilde{\mathbf{x}}^* = (1 - \delta/r)\mathbf{x}^*$, $\mathbf{x}^* \in \operatorname{argmin}_{\mathbf{x} \in \mathcal{K}} \sum_{t=1}^{T} f_t(\mathbf{x})$, and $\widehat{f}_{t,j,\delta}(\mathbf{x})$ denotes the δ-smoothed version of $f_{t,j}(\mathbf{x})$.

According to Lemma 9, by assuming that $\|\widehat{\mathbf{g}}_i(m)\|_2 \leq \widehat{G}$ for any $i \in V$ and $m \in [B]$, with probability at least $1-\gamma$, we have

$$R_{T,i} \leq 2nR(KG + \widehat{G})\sqrt{2B \ln \frac{1}{\gamma}} + 4nhR^2 + \frac{2nB\widehat{G}^2}{h} + 3nGT\epsilon + 3\delta nGT + \frac{\delta nGRT}{r}.$$

By substituting $\epsilon = \frac{4R}{\sqrt{L+2}} + \frac{\widehat{G}\sqrt{n}}{2h(1-\sigma_2(P))}$, $h = \frac{n^{1/4}\xi_T dMT^{3/4}}{\sqrt{1-\sigma_2(P)}R}$, $\delta = cT^{-1/4}$, $K = L = \sqrt{T}$, and $\widehat{G} = \xi_T \frac{dM\sqrt{K}}{\delta} + KG$ into the above inequality, we have

$$\begin{aligned}R_{T,i} \leq & 2nR\left(2G + \frac{\xi_T dM}{c}\right)\sqrt{2\ln\frac{1}{\gamma}}T^{3/4} + \frac{4\xi_T n^{5/4} dMR}{\sqrt{1-\sigma_2(P)}}T^{3/4} \\ & + 2n^{3/4}\sqrt{1-\sigma_2(P)}\left(\frac{R}{c} + \frac{RG}{\xi_T dM}\right)\left(\frac{\xi_T dM}{c} + G\right)T^{3/4} \\ & + 12nGRT^{3/4} + \frac{3n^{5/4}G}{2\sqrt{1-\sigma_2(P)}}\left(\frac{R}{c} + \frac{RG}{\xi_T dM}\right)T^{3/4} \\ & + 3cnGT^{3/4} + \frac{cnGR}{r}T^{3/4} \\ = & O\left(n^{5/4}(1-\sigma_2(P))^{-1/2}T^{3/4}\xi_T\right).\end{aligned}$$

Let $\mathcal{A}$ denote the event of $\|\widehat{\mathbf{g}}_i(m)\|_2 \leq \widehat{G}, \forall i \in V, m \in [B]$. Because we have used the event $\mathcal{A}$ as a fact, the above result should be formulated as

$$\Pr\left(R_{T,i} = O\left(n^{5/4}(1-\sigma_2(P))^{-1/2}T^{3/4}\xi_T\right)\Big|\,\mathcal{A}\right) \geq 1-\gamma. \tag{54}$$

Furthermore, we introduce the following lemma with respect to the probability of the event $\mathcal{A}$.

Lemma 10 *Under Assumptions 1 and 5, for all $i \in V$ and $m \in [B]$, Algorithm 3 has*

$$\|\widehat{\mathbf{g}}_i(m)\|_2 \leq \left(1+\sqrt{8\ln\frac{nB}{\gamma}}\right)\frac{dM\sqrt{K}}{\delta} + KG$$

with probability at least $1-\gamma$.

Then, by applying Lemma 10 with $B = T/K = \sqrt{T}$, we have

$$\Pr\left(\mathcal{A}\right) \geq 1-\gamma. \tag{55}$$

Finally, we complete the proof by combining (54) with (55).

Appendix C. Proof of Lemmas 3 and 7

These two lemmas can be derived by following the proof of Lemma 6 in Zhang et al. (2017). For completeness, we include the detailed proof in this paper.

Let P^s denote the s-th power of P and P_{ij}^s denote the j-th entry of the i-row in P^s for any $s \geq 0$. Note that P^0 denotes the identity matrix I_n. For $m = 1$, it is easy to verify that

$$\|\mathbf{z}_i(m) - \bar{\mathbf{z}}(m)\|_2 = 0 \leq \frac{\sqrt{n}\widehat{G}}{1-\sigma_2(P)}. \tag{56}$$

To analyze the case with $B \geq m \geq 2$, we introduce two intermediate results from Zhang et al. (2017) and Duchi et al. (2011). First, as shown in the proof of Lemma 6 at Zhang et al. (2017), for any $B \geq m \geq 2$, we have

$$\|\mathbf{z}_i(m) - \bar{\mathbf{z}}(m)\|_2 \leq \sum_{\tau=1}^{m-1}\sum_{j=1}^{n}\left|P_{ij}^{m-1-\tau} - \frac{1}{n}\right|\|\mathbf{d}_j(\tau)\|_2 \tag{57}$$

under Assumption 3. Second, as shown in Appendix B of Duchi et al. (2011), when P is a doubly stochastic matrix, for any positive integer s and any $\mathbf{x}$ in the n-dimensional probability simplex, it holds that

$$\|P^0\mathbf{x} - \mathbf{1}/n\|_1 \leq \sigma_2^s(P)\sqrt{n} \tag{58}$$

where $\mathbf{1}$ is the all-ones vector in $\mathbb{R}^n$.

Let $\mathbf{e}_i$ denote the i-th canonical basis vector in $\mathbb{R}^n$. By substituting $\mathbf{x} = \mathbf{e}_i$ into (58), we have

$$\|P^s\mathbf{e}_i - \mathbf{1}/n\|_1 \leq \sigma_2^s(P)\sqrt{n} \tag{59}$$

for any positive integer s. If $s = 0$, we also have

$$\|P^0\mathbf{e}_i - \mathbf{1}/n\|_1 = \frac{2(n-1)}{n} \leq \sqrt{n} = \sigma_2^0(P)\sqrt{n} \tag{60}$$

where the inequality is due to $n \geq 1$.

Then, for any $B \geq m \geq 2$, by combining (57) and $\|\mathbf{d}_i(m)\|_2 \leq \widehat{G}$, we have

$$\begin{aligned}\|\mathbf{z}_i(m) - \bar{\mathbf{z}}(m)\|_2 \leq & \widehat{G}\sum_{\tau=1}^{m-1}\sum_{j=1}^{n}\left|P_{ij}^{m-1-\tau} - \frac{1}{n}\right| = \widehat{G}\sum_{\tau=1}^{m-1}\sum_{j=1}^{n}\left|P_{ji}^{m-1-\tau} - \frac{1}{n}\right| \\ = & \widehat{G}\sum_{\tau=1}^{m-1}\left\|P^{m-1-\tau}\mathbf{e}_i - \frac{\mathbf{1}}{n}\right\|_1\end{aligned}$$

where the first equality is due to the symmetry of P.

Because of (59), (60), and $\sigma_2(P) < 1$, for any $B \geq m \geq 2$, we have

$$\|\mathbf{z}_i(m) - \bar{\mathbf{z}}(m)\|_2 \leq \widehat{G}\sum_{\tau=1}^{m-1}\sigma_2(P)^{m-1-\tau}\sqrt{n} = \frac{(1-\sigma_2(P)^{m-1})\widehat{G}\sqrt{n}}{1-\sigma_2(P)} \leq \frac{\sqrt{n}\widehat{G}}{1-\sigma_2(P)}. \tag{61}$$

By combining (56) and (61), we can complete the proof of Lemma 3.

Furthermore, by taking the expectation on the both sides of (57) and combining with $\mathbb{E}[\|\mathbf{d}_i(m)\|_2] \leq \widehat{G}$, we can prove Lemma 7 in a similar way.

Appendix D. Proof of Lemma 9

We first introduce the classical Azuma's inequality (Azuma, 1967) for martingales in the following lemma.

Lemma 11 *Suppose $D_1, \ldots, D_s$ is a martingale difference sequence and*

$$|D_j| \leq c_j$$

almost surely. Then, we have

$$\Pr\left(\sum_{j=1}^{s} D_j \geq \Delta\right) \leq \exp\left(\frac{-\Delta^2}{2\sum_{j=1}^{s} c_j^2}\right).$$

To apply Lemma 11, with $\mathcal{T}_m = \{(m-1)K+1, \ldots, mK\}$, we define

$$\begin{aligned}D_m &= \sum_{t\in\mathcal{T}_m}\sum_{j=1}^{n}\left(\nabla\widehat{f}_{t,j,\delta}(\mathbf{x}_j(m)) - \mathbf{g}_j(t)\right)^\top(\bar{\mathbf{x}}(m) - \widetilde{\mathbf{x}}^*) \\ &= \sum_{j=1}^{n}\left(\sum_{t\in\mathcal{T}_m}\nabla\widehat{f}_{t,j,\delta}(\mathbf{x}_j(m)) - \widehat{\mathbf{g}}_j(m)\right)^\top(\bar{\mathbf{x}}(m) - \widetilde{\mathbf{x}}^*).\end{aligned} \tag{62}$$

According to Algorithm 3 and Lemma 2, we have

$$\mathbb{E}\left[D_m | \mathbf{x}_1(m), \ldots, \mathbf{x}_n(m), \bar{\mathbf{x}}(m)\right] = 0$$

which further implies that $D_1, \ldots, D_B$ is a martingale difference sequence with

$$\begin{aligned}
|D_m| &= \left| \sum_{j=1}^{n} \left(\sum_{t \in \mathcal{T}_m} \nabla \widehat{f}_{t,j,\delta}(\mathbf{x}_j(m)) - \widehat{\mathbf{g}}_j(m) \right)^\top (\bar{\mathbf{x}}(m) - \widetilde{\mathbf{x}}^*) \right| \\
&\leq \sum_{j=1}^{n} \left\| \sum_{t \in \mathcal{T}_m} \nabla \widehat{f}_{t,j,\delta}(\mathbf{x}_j(m)) - \widehat{\mathbf{g}}_j(m) \right\|_2 \|(\bar{\mathbf{x}}(m) - \widetilde{\mathbf{x}}^*)\|_2 \\
&\leq 2R \sum_{j=1}^{n} \left(\left\| \sum_{t \in \mathcal{T}_m} \nabla \widehat{f}_{t,j,\delta}(\mathbf{x}_j(m)) \right\|_2 + \|\widehat{\mathbf{g}}_j(m)\|_2 \right) \\
&\leq 2R \sum_{j=1}^{n} \sum_{t \in \mathcal{T}_m} \left\| \nabla \widehat{f}_{t,j,\delta}(\mathbf{x}_j(m)) \right\|_2 + 2nR\widehat{G} \\
&\leq 2nRKG + 2nR\widehat{G}
\end{aligned}$$

where the second inequality is due to Assumption 2, and the last inequality is due to Lemma 8 and $|\mathcal{T}_m| = K$.

Then, by applying Lemma 11 with $\Delta = 2nR(KG + \widehat{G})\sqrt{2B \ln \frac{1}{\gamma}}$, with probability at least $1 - \gamma$, we have

$$\sum_{m=1}^{B} D_m \leq \Delta = 2nR(KG + \widehat{G})\sqrt{2B \ln \frac{1}{\gamma}}. \tag{63}$$

Additionally, by combining (62) with $\bar{\mathbf{g}}(m) = \frac{1}{n} \sum_{i=1}^{n} \widehat{\mathbf{g}}_i(m)$, we further have

$$\sum_{m=1}^{B} \sum_{t \in \mathcal{T}_m} \sum_{j=1}^{n} \nabla \widehat{f}_{t,j,\delta}(\mathbf{x}_j(m))^\top (\bar{\mathbf{x}}(m) - \widetilde{\mathbf{x}}^*) = \sum_{m=1}^{B} D_m + n \sum_{m=1}^{B} \bar{\mathbf{g}}(m)^\top (\bar{\mathbf{x}}(m) - \widetilde{\mathbf{x}}^*). \tag{64}$$

Therefore, we still need to bound $\sum_{m=1}^{B} \bar{\mathbf{g}}(m)^\top (\bar{\mathbf{x}}(m) - \widetilde{\mathbf{x}}^*)$. According to Assumption 3, it is easy to verify that Algorithm 3 with $\alpha = 0$ ensures

$$\begin{aligned}
\bar{\mathbf{z}}(m+1) =& \frac{1}{n} \sum_{i=1}^{n} \mathbf{z}_i(m+1) = \frac{1}{n} \sum_{i=1}^{n} \left(\sum_{j \in N_i} P_{ij} \mathbf{z}_j(m) + \widehat{\mathbf{g}}_i(m) \right) \\
=& \frac{1}{n} \sum_{i=1}^{n} \sum_{j=1}^{n} P_{ij} \mathbf{z}_j(m) + \bar{\mathbf{g}}(m) = \frac{1}{n} \sum_{j=1}^{n} \sum_{i=1}^{n} P_{ij} \mathbf{z}_j(m) + \bar{\mathbf{g}}(m) \\
=& \bar{\mathbf{z}}(m) + \bar{\mathbf{g}}(m) = \sum_{s=1}^{m} \bar{\mathbf{g}}(s).
\end{aligned}$$

Moreover, according to the definition, for any $m \in [B+1]$, we have

$$\bar{\mathbf{x}}(m+1) = \operatorname*{argmin}_{\mathbf{x} \in \mathcal{K}_\delta} \bar{F}_m(\mathbf{x}) = \operatorname*{argmin}_{\mathbf{x} \in \mathcal{K}_\delta} \bar{\mathbf{z}}(m)^\top \mathbf{x} + h\|\mathbf{x} - \mathbf{x}_{\text{in}}\|_2^2.$$

By applying Lemma 5 with the linear loss functions $\{\bar{\mathbf{g}}(m)^\top \mathbf{x}\}_{m=1}^B$, the decision set $\mathcal{K} = \mathcal{K}_\delta$, and the regularizer $\mathcal{R}(\mathbf{x}) = h\|\mathbf{x} - \mathbf{x}_{\text{in}}\|_2^2$, we have

$$\begin{aligned}\sum_{m=1}^B \bar{\mathbf{g}}(m)^\top(\bar{\mathbf{x}}(m+1) - \widetilde{\mathbf{x}}^*) &\leq h\|\widetilde{\mathbf{x}}^* - \mathbf{x}_{\text{in}}\|_2^2 + \sum_{m=1}^B \bar{\mathbf{g}}(m)^\top(\bar{\mathbf{x}}(m+1) - \bar{\mathbf{x}}(m+2)) \\ &\leq 4hR^2 + \sum_{m=1}^B \|\bar{\mathbf{g}}(m)\|_2 \|\bar{\mathbf{x}}(m+1) - \bar{\mathbf{x}}(m+2)\|_2\end{aligned} \tag{65}$$

where the last inequality is due to Assumption 2.

Note that $\bar{F}_{m+1}(\mathbf{x})$ is $2h$-strongly convex and $\bar{\mathbf{x}}(m+2) = \operatorname{argmin}_{\mathbf{x}\in\mathcal{K}_\delta} \bar{F}_{m+1}(\mathbf{x})$. For any $m \in [B]$, we have

$$\begin{aligned}&h\|\bar{\mathbf{x}}(m+1) - \bar{\mathbf{x}}(m+2)\|_2^2 \\ \leq&\bar{F}_{m+1}(\bar{\mathbf{x}}(m+1)) - \bar{F}_{m+1}(\bar{\mathbf{x}}(m+2)) \\ =&\bar{F}_m(\bar{\mathbf{x}}(m+1)) + \bar{\mathbf{g}}(m)^\top\bar{\mathbf{x}}(m+1) - \bar{F}_m(\bar{\mathbf{x}}(m+2)) - \bar{\mathbf{g}}(m)^\top\bar{\mathbf{x}}(m+2) \\ \leq&\|\bar{\mathbf{g}}(m)\|_2\|\bar{\mathbf{x}}(m+1) - \bar{\mathbf{x}}(m+2)\|_2\end{aligned}$$

where the first inequality is due to (11) and the second inequality is due to $\bar{\mathbf{x}}(m+1) = \operatorname{argmin}_{\mathbf{x}\in\mathcal{K}_\delta} \bar{F}_m(\mathbf{x})$.

The above inequality implies that for any $m \in [B]$, it holds that

$$\|\bar{\mathbf{x}}(m+1) - \bar{\mathbf{x}}(m+2)\|_2^2 \leq \frac{\|\bar{\mathbf{g}}(m)\|_2}{h}.$$

By combining with (65), we have

$$\begin{aligned}&\sum_{m=1}^B \bar{\mathbf{g}}(m)^\top(\bar{\mathbf{x}}(m) - \widetilde{\mathbf{x}}^*) \\ =&\sum_{m=1}^B \bar{\mathbf{g}}(m)^\top(\bar{\mathbf{x}}(m) - \bar{\mathbf{x}}(m+1)) + \sum_{m=1}^B \bar{\mathbf{g}}(m)^\top(\bar{\mathbf{x}}(m+1) - \widetilde{\mathbf{x}}^*) \\ \leq&\sum_{m=1}^B \|\bar{\mathbf{g}}(m)\|_2\|\bar{\mathbf{x}}(m) - \bar{\mathbf{x}}(m+1)\|_2 + 4hR^2 + \sum_{m=1}^B \|\bar{\mathbf{g}}(m)\|_2\|\bar{\mathbf{x}}(m+1) - \bar{\mathbf{x}}(m+2)\|_2 \\ \leq&4hR^2 + \frac{1}{h}\sum_{m=2}^B \|\bar{\mathbf{g}}(m)\|_2\|\bar{\mathbf{g}}(m-1)\|_2 + \|\bar{\mathbf{g}}(1)\|_2\|\bar{\mathbf{x}}(1) - \bar{\mathbf{x}}(2)\|_2 + \frac{1}{h}\sum_{m=1}^B \|\bar{\mathbf{g}}(m)\|_2^2 \\ \leq&4hR^2 + \frac{1}{h}\sum_{m=2}^B \|\bar{\mathbf{g}}(m)\|_2\|\bar{\mathbf{g}}(m-1)\|_2 + \frac{1}{h}\sum_{m=1}^B \|\bar{\mathbf{g}}(m)\|_2^2\end{aligned} \tag{66}$$

where the last inequality is due to $\bar{\mathbf{x}}(1) = \mathbf{x}_{\text{in}}$ and $\bar{\mathbf{x}}(2) = \operatorname{argmin}_{\mathbf{x}\in\mathcal{K}_\delta} \bar{F}_1(\mathbf{x}) = \mathbf{x}_{\text{in}}$.

Since $\|\widehat{\mathbf{g}}_i(m)\|_2 \leq \widehat{G}$, for any $m \in [B]$, we also have

$$\|\bar{\mathbf{g}}(m)\|_2 = \left\|\frac{1}{n}\sum_{i=1}^n \widehat{\mathbf{g}}_i(m)\right\|_2 \leq \frac{1}{n}\sum_{i=1}^n \|\widehat{\mathbf{g}}_i(m)\|_2 \leq \widehat{G}. \tag{67}$$

By substituting (67) into (66), we have

$$\sum_{m=1}^{B} \bar{\mathbf{g}}(m)^\top (\bar{\mathbf{x}}(m) - \widetilde{\mathbf{x}}^*) \leq 4hR^2 + \frac{(2B-1)\widehat{G}^2}{h} \leq 4hR^2 + \frac{2B\widehat{G}^2}{h}. \tag{68}$$

Finally, by substituting (63) and (68) into (64), we complete the proof.

Appendix E. Proof of Lemma 10

This proof is inspired by the proof of Theorem 12 in Gross (2011), which gave the classical Bernstein inequality for independent vector-valued random variables. However, the vector-valued random variables in this proof are only conditionally independent, and we do not need to use the Bernstein inequality to incorporate the variance information.

According to Algorithm 3, for any $i \in V$ and $m = 1, \ldots, B$, conditioned on $\mathbf{x}_i(m)$,

$$\mathbf{g}_i((m-1)K+1), \ldots, \mathbf{g}_i(mK)$$

are K independent random vectors. For brevity, for $j = 1, \ldots, K$, let

$$X_j = \mathbf{g}_i(t_j)$$

where $t_j = (m-1)K + j$, and let $N = \left\| \sum_{j=1}^{K} X_j \right\|_2$, $\widehat{S}_j = \sum_{k \neq j} X_k$.

To bound N by using Lemma 11, we define $\mathbf{X}_0 = \{\mathbf{x}_i(m)\}$, $\mathbf{X}_j = \{\mathbf{x}_i(m), X_1, \ldots, X_j\}$ for $j \geq 1$ and a sequence $D_1, \ldots, D_K$ as

$$D_j = \mathbb{E}[N|\mathbf{X}_j] - \mathbb{E}[N|\mathbf{X}_{j-1}].$$

It is not hard to verify that

$$\mathbb{E}[D_j|\mathbf{X}_{j-1}] = \mathbb{E}[\mathbb{E}[N|\mathbf{X}_j] - \mathbb{E}[N|\mathbf{X}_{j-1}]|\mathbf{X}_{j-1}] = 0$$

which implies that $D_1, \ldots, D_K$ is a martingale difference sequence.

Then, using the triangle inequality, we have

$$N \leq \|\widehat{S}_j\|_2 + \|X_j\|_2 \text{ and } N \geq \|\widehat{S}_j\|_2 - \|X_j\|_2. \tag{69}$$

Moreover, according to the Algorithm 3 and Assumption 5, we have

$$\|X_j\|_2 = \left\| \frac{d}{\delta} f_{t_j,i}(\mathbf{y}_i(t_j)) \mathbf{u}_i(t_j) \right\|_2 \leq \frac{dM}{\delta}.$$

Therefore, by combining with (69), we have

$$N \leq \|\widehat{S}_j\|_2 + \frac{dM}{\delta} \text{ and } N \geq \|\widehat{S}_j\|_2 - \frac{dM}{\delta}.$$

Then, we have

$$D_j \leq \mathbb{E}[\|\widehat{S}_j\|_2|\mathbf{X}_j] + \frac{dM}{\delta} - \mathbb{E}[\|\widehat{S}_j\|_2|\mathbf{X}_{j-1}] + \frac{dM}{\delta} = \frac{2dM}{\delta}$$

and

$$D_j \geq \mathbb{E}[\|\widehat{S}_j\|_2|\mathbf{X}_j] - \frac{dM}{\delta} - \mathbb{E}[\|\widehat{S}_j\|_2|\mathbf{X}_{j-1}] - \frac{dM}{\delta} = -\frac{2dM}{\delta}$$

where the above two equalities are due to $\mathbb{E}[\|\widehat{S}_j\|_2|\mathbf{X}_j] = \mathbb{E}[\|\widehat{S}_j\|_2|\mathbf{X}_{j-1}]$, because $\widehat{S}_j$ dose not depend on X_j given $\mathbf{x}_i(m)$. Therefore, we have $|D_j| \leq \frac{2dM}{\delta}$.

Let $\Delta = \frac{\sqrt{K}dM}{\delta}\sqrt{8\ln\frac{nB}{\gamma}}$. Then, by applying Lemma 11, with probability at least $1 - \frac{\gamma}{nB}$, we have

$$N - \mathbb{E}[N|\mathbf{x}_i(m)] = \mathbb{E}[N|\mathbf{X}_K] - \mathbb{E}[N|\mathbf{X}_0] = \sum_{j=1}^{K} D_j \leq \frac{\sqrt{K}dM}{\delta}\sqrt{8\ln\frac{nB}{\gamma}}$$

which implies that

$$\|\widehat{\mathbf{g}}_i(m)\|_2 = N \leq \frac{\sqrt{K}dM}{\delta}\sqrt{8\ln\frac{nB}{\gamma}} + \mathbb{E}[N|\mathbf{x}_i(m)] \leq \frac{\sqrt{K}dM}{\delta}\sqrt{8\ln\frac{nB}{\gamma}} + \sqrt{\mathbb{E}[N^2|\mathbf{x}_i(m)]}.$$

where the last inequality is due to the Jensen's inequality.

By combining the above inequality with $N^2 = \|\widehat{\mathbf{g}}_i(m)\|_2^2$ and (48), with probability at least $1 - \frac{\gamma}{nB}$, we have

$$\|\widehat{\mathbf{g}}_i(m)\|_2 \leq \left(1 + \sqrt{8\ln\frac{nB}{\gamma}}\right)\frac{dM\sqrt{K}}{\delta} + KG.$$

Finally, by using the union bound, we complete the proof for all $i \in V$ and $m = 1, \ldots, B$.

References

Jacob Abernethy, Peter L. Bartlett, Alexander Rakhlin, and Ambuj Tewari. Optimal strategies and minimax lower bounds for online convex games. In *Proceedings of the 21st Annual Conference on Learning Theory*, pages 415–423, 2008.

Alekh Agarwal, Ofer Dekel, and Lin Xiao. Optimal algorithms for online convex optimization with multi-point bandit feedback. In *Proceedings of the 23rd Annual Conference on Learning Theory*, pages 28–40, 2010.

Amit Agarwal, Elad Hazan, Satyen Kale, and Robert E. Schapire. Algorithms for portfolio management based on the newton method. In *Proceedings of the 23rd International Conference on Machine Learning*, pages 9–16, 2006.

Baruch Awerbuch and Robert D. Kleinberg. Adaptive routing with end-to-end feedback: Distributed learning and geometric approaches. In *Proceedings of the 36th Annual ACM Symposium on Theory of Computing*, pages 45–53, 2004.

Baruch Awerbuch and Robert D. Kleinberg. Online linear optimization and adaptive routing. *Journal of Computer and System Sciences*, 74(1):97–114, 2008.

Kazuoki Azuma. Weighted sums of certain dependent random variables. *Tohoku Mathematical Journal*, 19(3):357–367, 1967.

Florence Bénézit, Alexandros G. Dimakis, Patrick Thiran, and Martin Vetterli. Gossip along the way: Order-optimal consensus through randomized path averaging. In *Proceedings of the 45th Annual Allerton Conference on Communication, Control, and Computing*, 2007.

Avrim Blum and Adam Kalai. Universal portfolios with and without transaction costs. *Machine Learning*, 35(3):193–205, 1999.

Stephen Boyd and Lieven Vandenberghe. *Convex Optimization*. Cambridge University Press, 2004.

Nicolò Cesa-Bianchi, Yoav Freund, David Haussler, David P. Helmbold, Robert E. Schapire, and Manfred K. Warmuth. How to use expert advice. *Journal of the ACM*, 44(3):427–485, 1997.

Chih-Chung Chang and Chih-Jen Lin. LIBSVM: A library for support vector machines. *ACM Transactions on Intelligent Systems and Technology*, 2(27):1–27, 2011.

Lin Chen, Christopher Harshaw, Hamed Hassani, and Amin Karbasi. Projection-free online optimization with stochastic gradient: From convexity to submodularity. In *Proceedings of the 35th International Conference on Machine Learning*, pages 814–823, 2018.

Lin Chen, Mingrui Zhang, and Amin Karbasi. Projection-free bandit convex optimization. In *Proceedings of the 22nd International Conference on Artificial Intelligence and Statistics*, pages 2047–2056, 2019.

John C. Duchi, Alekh Agarwal, and Martin J. Wainwright. Dual averaging for distributed optimization: Convergence analysis and network scaling. *IEEE Transactions on Automatic Control*, 57(3):592–606, 2011.

John C. Duchi, Michael I. Jordan, Martin J. Wainwright, and Andre Wibisono. Optimal rates for zero-order convex optimization: The power of two function evaluations. *IEEE Transactions on Information Theory*, 61(5):2788–2806, 2015.

Abraham D. Flaxman, Adam Tauman Kalai, and H. Brendan McMahan. Online convex optimization in the bandit setting: Gradient descent without a gradient. In *Proceedings of the 16th Annual ACM-SIAM Symposium on Discrete Algorithms*, pages 385–394, 2005.

Marguerite Frank and Philip Wolfe. An algorithm for quadratic programming. *Naval Research Logistics Quarterly*, 3(1–2):95–110, 1956.

Yoav Freund, Robert E. Schapire, Yoram Singer, and Manfred K. Warmuth. Using and combining predictors that specialize. In *Proceedings of the 29th Annual ACM Symposium on Theory of Computing*, pages 334–343, 1997.

Dan Garber and Elad Hazan. A linearly convergent conditional gradient algorithm with applications to online and stochastic optimization. *SIAM Journal on Optimization*, 26(3): 1493–1528, 2016.

Dan Garber and Ben Kretzu. Improved regret bounds for projection-free bandit convex optimization. In *Proceedings of the 23rd International Conference on Artificial Intelligence and Statistics*, pages 2196–2206, 2020.

Dan Garber and Ben Kretzu. Revisiting projection-free online learning: The strongly convex case. In *Proceedings of the 24th International Conference on Artificial Intelligence and Statistics*, pages 3592–3600, 2021.

David Gross. Recovering low-rank matrices from few coefficients in any basis. *IEEE Transactions on Information Theory*, 57(3):1548–1566, 2011.

Elad Hazan. Introduction to online convex optimization. *Foundations and Trends in Optimization*, 2(3–4):157–325, 2016.

Elad Hazan and Satyen Kale. Projection-free online learning. In *Proceedings of the 29th International Conference on Machine Learning*, pages 1843–1850, 2012.

Elad Hazan and Edgar Minasyan. Faster projection-free online learning. In *Proceedings of the 33rd Annual Conference on Learning Theory*, pages 1877–1893, 2020.

Saghar Hosseini, Airlie Chapman, and Mehran Mesbahi. Online distributed optimization via dual averaging. In *52nd IEEE Conference on Decision and Control*, pages 1484–1489, 2013.

Martin Jaggi. Revisiting frank-wolfe: Projection-free sparse convex optimization. In *Proceedings of the 30th International Conference on Machine Learning*, pages 427–435, 2013.

Prateek Jain, Brian Kulis, Inderjit S. Dhillon, and Kristen Grauman. Online metric learning and fast similarity search. In *Advances in Neural Information Processing Systems 21*, pages 761–768, 2008.

Anastasia Koloskova, Sebastian U. Stich, and Martin Jaggi. Decentralized stochastic optimization and gossip algorithms with compressed communication. In *Proceedings of the 36th International Conference on Machine Learning*, pages 3478–3487, 2019.

Kfir Y. Levy and Andreas Krause. Projection free online learning over smooth sets. In *Proceedings of the 22nd International Conference on Artificial Intelligence and Statistics*, pages 1458–1466, 2019.

Dan Li, Kerry D. Wong, Yu H. Hu, and Akbar M. Sayeed. Detection, classification and tracking of targets in distributed sensor networks. *IEEE Signal Processing Magazine*, 19 (2):17–29, 2002.

Angelia Nedić, Alex Olshevsky, Asuman Ozdaglar, and John N. Tsitsiklis. On distributed averaging algorithms and quantization effects. *IEEE Transactions on Automatic Control*, 54(11):2506–2517, 2009.

S. Sundhar Ram, A. Nedić, and V. V. Veeravalli. Distributed stochastic subgradient projection algorithms for convex optimization. *Journal of Optimization Theory and Applications*, 147(3):516–545, 2010.

Shai Shalev-Shwartz. Online learning and online convex optimization. *Foundations and Trends in Machine Learning*, 4(2):107–194, 2011.

Grigorios Tsagkatakis and Andreas Savakis. Online distance metric learning for object tracking. *IEEE Transactions on Circuits and Systems for Video Technology*, 21(12): 1810–1821, 2011.

Konstantinos I. Tsianos and Michael G. Rabbat. Distributed strongly convex optimization. In *Proceedings of the 50th Annual Allerton Conference on Communication, Control, and Computing*, pages 593–600, 2012.

Yuanyu Wan and Lijun Zhang. Projection-free online learning over strongly convex sets. In *Proceedings of the 35th AAAI Conference on Artificial Intelligence*, pages 10076–10084, 2021.

Yuanyu Wan, Wei-Wei Tu, and Lijun Zhang. Projection-free distributed online convex optimization with $O(\sqrt{T})$ communication complexity. In *Proceedings of the 37th International Conference on Machine Learning*, pages 9818–9828, 2020.

Yuanyu Wan, Wei-Wei Tu, and Lijun Zhang. Online strongly convex optimization with unknown delays. *Machine Learning*, 111(3):871–893, 2022.

Lin Xiao, Stephen Boyd, and Seung-Jean Kim. Distributed average consensus with least-mean-square deviation. *Journal of Parallel and Distributed Computing*, 67(1):33–46, 2007.

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.

Tao Yang, Xinlei Yi, Junfeng Wu, Ye Yuan, Di Wu, Ziyang Meng, Yiguang Hong, Hong Wang, Zongli Lin, and Karl H. Johansson. A survey of distributed optimization. *Annual Reviews in Control*, 47:278–305, 2019.

Xinlei Yi, Xiuxian Li, Lihua Xie, and Karl H. Johansson. Distributed online convex optimization with time-varying coupled inequality constraints. *IEEE Transactions on Signal Processing*, 68:731–746, 2020.

Lijun Zhang, Tianbao Yang, Rong Jin, and Xiaofei He. $O(\log T)$ projections for stochastic optimization of smooth and strongly convex functions. In *Proceedings of the 30th International Conference on Machine Learning*, pages 1121–1129, 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.

Martin Zinkevich. Online convex programming and generalized infinitesimal gradient ascent. In *Proceedings of the 20th International Conference on Machine Learning*, pages 928–936, 2003.