Proceedings of Machine Learning Research vol 247:1–27, 2024 37th Annual Conference on Learning Theory

Nearly Optimal Regret for Decentralized Online Convex Optimization

Yuanyu Wan WANYY@ZJU.EDU.CN
The State Key Laboratory of Blockchain and Data Security, Zhejiang University, Hangzhou, China
School of Software Technology, Zhejiang University, Ningbo, China
Hangzhou High-Tech Zone (Binjiang) Institute of Blockchain and Data Security, Hangzhou, China

Tong Wei WEIT@SEU.EDU.CN
School of Computer Science and Engineering, Southeast University, Nanjing, China

Mingli Song BROOKSONG@ZJU.EDU.CN
The State Key Laboratory of Blockchain and Data Security, Zhejiang University, Hangzhou, China
Hangzhou High-Tech Zone (Binjiang) Institute of Blockchain and Data Security, Hangzhou, China

Lijun Zhang ZHANGLJ@LAMDA.NJU.EDU.CN
National Key Laboratory for Novel Software Technology, Nanjing University, Nanjing, China
The State Key Laboratory of Blockchain and Data Security, Zhejiang University, Hangzhou, China

Editors: Shipra Agrawal and Aaron Roth

Abstract

We investigate decentralized online convex optimization (D-OCO), in which a set of local learners are required to minimize a sequence of global loss functions using only local computations and communications. Previous studies have established $O(n^{5/4}\rho^{-1/2}\sqrt{T})$ and $O(n^{3/2}\rho^{-1}\log T)$ regret bounds for convex and strongly convex functions respectively, where n is the number of local learners, $\rho < 1$ is the spectral gap of the communication matrix, and T is the time horizon. However, there exist large gaps from the existing lower bounds, i.e., $\Omega(n\sqrt{T})$ for convex functions and $\Omega(n)$ for strongly convex functions. To fill these gaps, in this paper, we first develop novel D-OCO algorithms that can respectively reduce the regret bounds for convex and strongly convex functions to $\tilde{O}(n\rho^{-1/4}\sqrt{T})$ and $\tilde{O}(n\rho^{-1/2}\log T)$. The primary technique is to design an online accelerated gossip strategy that enjoys a faster average consensus among local learners. Furthermore, by carefully exploiting the spectral properties of a specific network topology, we enhance the lower bounds for convex and strongly convex functions to $\Omega(n\rho^{-1/4}\sqrt{T})$ and $\Omega(n\rho^{-1/2})$, respectively. These lower bounds suggest that our algorithms are nearly optimal in terms of T, n, and ρ.

Keywords: Online Convex Optimization, Decentralized Optimization, Optimal Regret, Accelerated Gossip Strategy

1. Introduction

Decentralized online convex optimization (D-OCO) (Yan et al., 2013; Hosseini et al., 2013; Zhang et al., 2017; Wan et al., 2020, 2022) is a powerful learning framework for distributed applications with streaming data, such as distributed tracking in sensor networks (Li et al., 2002; Lesser et al., 2003) and online packet routing (Awerbuch and Kleinberg, 2004, 2008). Specifically, it can be formulated as a repeated game between an adversary and a set of local learners numbered by $1, \ldots, n$ and connected by a network, where the network is defined by an undirected graph $\mathcal{G} = ([n], E)$ with the edge set $E \subseteq [n] \times [n]$. In the t-th round, each learner $i \in [n]$ first chooses a decision $\mathbf{x}_i(t)$ from a convex set $\mathcal{K} \subseteq \mathbb{R}^d$, and then receives a convex loss function $f_{t,i}(\mathbf{x}) : \mathcal{K} \to \mathbb{R}$ selected by the adversary. The goal of each learner i is to minimize the regret in terms of the global function $f_t(\mathbf{x}) = \sum_{j=1}^{n} f_{t,j}(\mathbf{x})$ at each round t, i.e.,

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

where T denotes the time horizon.

Note that in the special case with $n = 1$, D-OCO reduces to the classical online convex optimization (OCO) (Shalev-Shwartz, 2011; Hazan, 2016). There already exist many online algorithms with optimal regret bounds for convex and strongly convex functions, e.g., online gradient descent (OGD) (Zinkevich, 2003). However, these algorithms cannot be applied to the general D-OCO problem, because they need direct access to the global function $f_t(\mathbf{x})$, which is unavailable for the local learners. To be precise, there exist communication constraints in D-OCO: the learner i only has local access to the function $f_{t,i}(\mathbf{x})$, and can only communicate with its immediate neighbors via a single step of the gossip protocol (Xiao and Boyd, 2004; Boyd et al., 2006) based on a weight matrix $P \in \mathbb{R}^{n\times n}$ at each round.[1] To address this limitation, the pioneering work of Yan et al. (2013) extends OGD into the D-OCO setting, and achieves $O(n^{5/4}\rho^{-1/2}\sqrt{T})$ and $O(n^{3/2}\rho^{-1}\log T)$ regret bounds for convex and strongly convex functions respectively, where $\rho < 1$ is the spectral gap of P. The key idea is to first apply a standard gossip step (Xiao and Boyd, 2004) over the decisions of these local learners, and then perform a gradient descent step based on the local function. Later, there has been a growing research interest in developing and analyzing D-OCO algorithms based on the standard gossip step for different scenarios (Hosseini et al., 2013; Zhang et al., 2017; Lei et al., 2020; Wan et al., 2020, 2021, 2022; Wang et al., 2023). However, the best regret bounds for D-OCO with convex and strongly convex functions remain unchanged. Moreover, there exist large gaps from the lower bounds recently established by Wan et al. (2022), i.e., $\Omega(n\sqrt{T})$ for convex functions and $\Omega(n)$ for strongly convex functions.

To fill these gaps, this paper first proposes two novel D-OCO algorithms that respectively achieve a regret bound of $\tilde{O}(n\rho^{-1/4}\sqrt{T})$ for convex functions and an improved regret bound of $\tilde{O}(n\rho^{-1/2}\log T)$ for strongly convex functions.[2] Different from previous D-OCO algorithms that rely on the standard gossip step, we make use of an accelerated gossip strategy (Liu and Morse, 2011) to weaken the impact of decentralization on the regret. In the studies of offline and stochastic optimization, it is well-known that the accelerated strategy enjoys a faster average consensus among decentralized nodes (Lu and Sa, 2021; Ye and Chang, 2023; Ye et al., 2023). However, applying the accelerated strategy to D-OCO is more challenging because it requires multiple communications in each round, which violates the communication protocol of D-OCO. To tackle this issue, we design an online accelerated gossip strategy by further incorporating a blocking update mechanism, which allows us to allocate the communications required by each update into every round of a block. Furthermore, we establish nearly matching lower bounds of $\Omega(n\rho^{-1/4}\sqrt{T})$ and $\Omega(n\rho^{-1/2})$ for convex and strongly convex functions, respectively. Compared with the existing lower bounds (Wan et al., 2022), our bounds additionally uncover the effect of the spectral gap, by carefully exploiting the spectral properties of a specific network topology. To highlight the significance of this work, we compare our results with previous studies in Table 1.

1. More specifically, the essence of a single gossip step is to compute a weighted average of some parameters of these local learners based on the matrix P. Moreover, following previous studies (Yan et al., 2013; Hosseini et al., 2013), P is given beforehand, instead of being a choice of the algorithm.
2. We use the $\tilde{O}(\cdot)$ notation to hide constant factors as well as polylogarithmic factors in n, but not in T.

Table 1: Summary of our results and the best previous results for D-OCO. Abbreviations: convex → cvx, strongly convex → scvx.

$f_{t,i}(\cdot)$	Source	Upper Bound	Lower Bound	Regret Gap
cvx	Previous studies	$O(n^{5/4}\rho^{-1/2}\sqrt{T})$ Yan et al. (2013); Hosseini et al. (2013)	$\Omega(n\sqrt{T})$ Wan et al. (2022)	$O(n^{1/4}\rho^{-1/2})$
	This work	$O(n\rho^{-1/4}\sqrt{T\log n})$ Theorem 1	$\Omega(n\rho^{-1/4}\sqrt{T})$ Theorem 3	$O(\sqrt{\log n})$
scvx	Previous studies	$O(n^{3/2}\rho^{-1}\log T)$ Yan et al. (2013); Wan et al. (2021)	$\Omega(n)$ Wan et al. (2022)	$O(n^{1/2}\rho^{-1}\log T)$
	This work	$O(n\rho^{-1/2}(\log n)\log T)$ Theorem 2	$\Omega(n\rho^{-1/2})$ Theorem 4	$O((\log n)\log T)$

2. Related Work

In this section, we briefly review the related work on D-OCO, including the special case with $n = 1$ and the general case.

2.1. Special D-OCO with $n = 1$

D-OCO with $n = 1$ reduces to the classical OCO problem, which dates back to the seminal work of Zinkevich (2003). Over the past decades, this problem has been extensively studied, and various algorithms with optimal regret have been presented for convex and strongly convex functions, respectively (Zinkevich, 2003; Shalev-Shwartz and Singer, 2007; Hazan et al., 2007; Abernethy et al., 2008). The closest one to this paper is follow-the-regularized-leader (FTRL) (Shalev-Shwartz and Singer, 2007), which updates the decision (omitting the subscript of the learner 1 for brevity) as

$$\mathbf{x}(t+1) = \operatorname*{argmin}_{\mathbf{x}\in\mathcal{K}} \sum_{i=1}^{t} \langle \nabla f_i(\mathbf{x}(i)), \mathbf{x}\rangle + \frac{1}{\eta}\|\mathbf{x}\|_2^2 \tag{2}$$

where η is a parameter. By setting an appropriate η, it can achieve an optimal $O(\sqrt{T})$ regret bound for convex functions. Note that this algorithm is also known as dual averaging, especially in the filed of offline and stochastic optimization (Nesterov, 2009; Xiao, 2009). Moreover, when functions are α-strongly convex, Hazan et al. (2007) have proposed a variant of (2),[3] which makes the following update

$$\mathbf{x}(t+1) = \operatorname*{argmin}_{\mathbf{x}\in\mathcal{K}} \sum_{i=1}^{t} \langle \nabla f_i(\mathbf{x}(i)) - \alpha\mathbf{x}(i), \mathbf{x}\rangle + \frac{t\alpha}{2}\|\mathbf{x}\|_2^2. \tag{3}$$

This variant is named as follow-the-approximate-leader (FTAL), and can achieve an optimal regret bound of $O(\log T)$ for strongly convex functions.

3. Although (3) seems slightly different from the update rule of Hazan et al. (2007), it is easy to verify the equivalence between them.

2.2. General D-OCO with $n \geq 2$

D-OCO is a generalization of OCO with $n \geq 2$ local learners in the network defined by an undirected graph $\mathcal{G} = ([n], E)$. The main challenge of D-OCO is that each learner $i \in [n]$ is required to minimize the regret in terms of the global function $f_t(\mathbf{x}) = \sum_{j=1}^{n} f_{t,j}(\mathbf{x})$, i.e., $R_{T,i}$ in (1), but except the direct access to $f_{t,i}(\mathbf{x})$, it can only estimate the global information from the gossip communication occurring via the weight matrix P. To tackle this challenge, Yan et al. (2013) propose a decentralized variant of OGD (D-OGD) by first applying the standard gossip step (Xiao and Boyd, 2004) over the decisions of these local learners, and then performing a gradient descent step according to the local function. For convex and strongly convex functions, D-OGD can achieve $O(n^{5/4}\rho^{-1/2}\sqrt{T})$ and $O(n^{3/2}\rho^{-1}\log T)$ regret bounds, respectively.

Later, Hosseini et al. (2013) propose a decentralized variant of FTRL (D-FTRL), which performs the following update for each learner i at round t

$$
\begin{aligned}
\mathbf{z}_i(t+1) &= \sum_{j \in N_i} P_{ij}\mathbf{z}_j(t) + \nabla f_{t,i}(\mathbf{x}_i(t)) \\
\mathbf{x}_i(t+1) &= \operatorname*{argmin}_{\mathbf{x} \in \mathcal{K}} \langle \mathbf{z}_i(t+1), \mathbf{x} \rangle + \frac{1}{\eta}\|\mathbf{x}\|_2^2
\end{aligned}
\tag{4}
$$

where $N_i = \{j \in [n] | (i,j) \in E\} \cup \{i\}$ denotes the set including the immediate neighbors of the learner i and itself. Notice that the cumulative gradients $\sum_{i=1}^{t} \nabla f_i(\mathbf{x}(i))$ utilized in (2) is replaced by a local variable $\mathbf{z}_i(t+1)$ that is computed by first applying the standard gossip step over $\mathbf{z}_i(t)$ of these local learners and then adding the local gradient $\nabla f_{t,i}(\mathbf{x}_i(t))$. As proved by Hosseini et al. (2013), D-FTRL can also achieve the $O(n^{5/4}\rho^{-1/2}\sqrt{T})$ regret bound for convex functions. Recently, Wan et al. (2021) develop a decentralized variant of FTAL (D-FTAL) for α-strongly convex functions, where each learner i updates as

$$
\begin{aligned}
\mathbf{z}_i(t+1) &= \sum_{j \in N_i} P_{ij}\mathbf{z}_j(t) + (\nabla f_{t,i}(\mathbf{x}_i(t)) - \alpha\mathbf{x}_i(t)) \\
\mathbf{x}_i(t+1) &= \operatorname*{argmin}_{\mathbf{x} \in \mathcal{K}} \langle \mathbf{z}_i(t+1), \mathbf{x} \rangle + \frac{t\alpha}{2}\|\mathbf{x}\|_2^2
\end{aligned}
\tag{5}
$$

at round t. Different from D-FTRL, the local variable $\mathbf{z}_i(t+1)$ now is utilized to replace the cumulative information $\sum_{i=1}^{t}(\nabla f_i(\mathbf{x}(i)) - \alpha\mathbf{x}(i))$ in (3), which is critical for exploiting the strong convexity of functions to achieve a better regret bound of $O(n^{3/2}\rho^{-1}\log T)$.[4]

Additionally, many other OCO algorithms have also been extended into the decentralized setting for handling various scenarios such as complex decision sets (Zhang et al., 2017; Wan et al., 2020, 2022; Wang et al., 2023), time-varying constraints (Yi et al., 2020, 2021), and dynamic environments (Shahrampour and Jadbabaie, 2018). However, the current best regret bounds for general D-OCO with convex and strongly convex functions are still $O(n^{5/4}\rho^{-1/2}\sqrt{T})$ and $O(n^{3/2}\rho^{-1}\log T)$ respectively, and it is not clear whether these upper bounds are optimal or not. Although Wan et al. (2022) recently establish $\Omega(n\sqrt{T})$ and $\Omega(n)$ lower bounds for D-OCO with convex and strongly convex functions respectively, their results only imply that these upper bounds are (nearly) tight in

4. At first glance, Wan et al. (2021) only argue an $O(n^{3/2}\rho^{-1}T^{2/3}\log T)$ regret bound for strongly convex functions. However, this is because they focus on a projection-free property, and it is easy to verify that their algorithm can achieve the $O(n^{3/2}\rho^{-1}\log T)$ regret bound if projection is conducted in each round (see Appendix F for details).

terms of T. There still exist gaps in terms of n and ρ between existing upper and lower bounds. Notice that the value of ρ actually reflects the connectivity of the network: a larger ρ implies better connectivity, and could even be $\Omega(n^{-2})$ for "poorly connected" networks such as the 1-connected cycle graph (Duchi et al., 2011). Therefore, these gaps on n and ρ cannot be ignored, especially for large-scale distributed systems. In this paper, we fill these gaps up to polylogarithmic factors.

Discussions Different from D-OCO, previous studies have proposed optimal algorithms for many different scenarios of decentralized offline and stochastic optimization (Scaman et al., 2017, 2018, 2019; Gorbunov et al., 2020; Kovalev et al., 2020; Dvinskikh and Gasnikov, 2021; Lu and Sa, 2021; Song et al., 2023; Ye and Chang, 2023; Ye et al., 2023). The closest work to this paper is Scaman et al. (2019), which investigates decentralized offline optimization with convex and strongly convex functions. Let $\hat{\rho}$ be the normalized eigengap of P, which could be close to ρ. Scaman et al. (2019) have established optimal convergence rates of $O(\epsilon^{-2}+\epsilon^{-1}\hat{\rho}^{-1/2})$ and $O(\epsilon^{-1}+\epsilon^{-1/2}\hat{\rho}^{-1/2})$ to reach an ϵ precision for convex and strongly convex functions, respectively. However, it is worth noting that D-OCO is more challenging than the offline setting due to the change of functions. Actually, it is easy to apply a standard online-to-batch conversion (Cesa-Bianchi et al., 2004) of any D-OCO algorithm with regret $R_{T,i}$ to achieve an approximation error of $O(R_{T,i}/(nT))$ for decentralized offline optimization, but not vice versa. Moreover, one may notice that due to the online-to-batch conversion, it is possible to utilize existing lower bounds in the offline setting (Scaman et al., 2019) to derive $\Omega(n\sqrt{T}+n\hat{\rho}^{-1/2})$ and $\Omega(n+n\hat{\rho}^{-1}T^{-1})$ lower bounds for the regret of D-OCO with convex and strongly convex functions, respectively. However, for D-OCO, it is common to consider the case where T is much larger than other problem constants, and these lower bounds will also reduce to those specially established for D-OCO (Wan et al., 2022). In addition, we want to emphasize that although the accelerated gossip strategy (Liu and Morse, 2011) has been widely utilized in these previous studies on decentralized offline and stochastic optimization, this paper is the first work to apply it in D-OCO.

3. Main Results

We first introduce the necessary preliminaries including common assumptions and an algorithmic ingredient. Then, we present our two algorithms with improved regret bounds, and establish nearly matching lower bounds.

3.1. Preliminaries

Similar to previous studies on D-OCO (Yan et al., 2013; Hosseini et al., 2013), we first introduce the following assumptions.

Assumption 1 *The communication matrix $P \in \mathbb{R}^{n\times n}$ is supported on the graph $\mathcal{G} = ([n], E)$, symmetric, and doubly stochastic, which satisfies*

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

Moreover, P is positive semidefinite, and its second largest singular value denoted by $\sigma_2(P)$ is strictly smaller than 1.

Assumption 2 *At each round $t \in [T]$, the loss function $f_{t,i}(\mathbf{x})$ of each learner $i \in [n]$ is G-Lipschitz over $\mathcal{K}$, i.e., it holds that*

$$|f_{t,i}(\mathbf{x}) - f_{t,i}(\mathbf{y})| \leq G\|\mathbf{x} - \mathbf{y}\|_2, \ \forall \mathbf{x}, \mathbf{y} \in \mathcal{K}.$$

Assumption 3 *The set $\mathcal{K}$ contains the origin, i.e., $\mathbf{0} \in \mathcal{K}$, and its radius is bounded by R, i.e., it holds that*

$$\|\mathbf{x}\|_2 \leq R, \ \forall \mathbf{x} \in \mathcal{K}.$$

Assumption 4 *At each round $t \in [T]$, the loss function $f_{t,i}(\mathbf{x})$ of each learner $i \in [n]$ is α-strongly convex over $\mathcal{K}$, i.e., it holds that*

$$f_{t,i}(\mathbf{y}) \geq f_{t,i}(\mathbf{x}) + \langle \nabla f_{t,i}(\mathbf{x}), \mathbf{y} - \mathbf{x}\rangle + \frac{\alpha}{2}\|\mathbf{y} - \mathbf{x}\|_2^2, \ \forall \mathbf{x}, \mathbf{y} \in \mathcal{K}.$$

Note that Assumption 4 with $\alpha = 0$ reduces to the case with general convex functions.

Then, we briefly introduce the accelerated gossip strategy (Liu and Morse, 2011), which will be utilized to develop our algorithms. Given a set of vectors denoted as $\nabla_1, \ldots, \nabla_n \in \mathbb{R}^d$, a naive idea for approximating the average $\bar{\nabla} = \frac{1}{n}\sum_{i=1}^n \nabla_n$ in the decentralized setting is to perform multiple standard gossip steps (Xiao and Boyd, 2004), i.e., setting $\nabla_i^0 = \nabla_i$ and updating as

$$\nabla_i^{k+1} = \sum_{j \in N_i} P_{ij}\nabla_j^k \text{ for } k = 0, 1, \ldots, L-1 \tag{6}$$

where $L \geq 1$ is the number of iterations. Under Assumption 1, it is well-known that ∇_i^L generated by (6) provably converges to the average $\bar{\nabla}$ with the increase of L. However, Liu and Morse (2011) have shown that it is not the most efficient way, and proposed an accelerated gossip strategy by mixing the standard gossip step with an old averaging estimation, i.e., setting $\nabla_i^0 = \nabla_i^{-1} = \nabla_i$ and updating as

$$\nabla_i^{k+1} = (1+\theta)\sum_{j \in N_i} P_{ij}\nabla_j^k - \theta\nabla_i^{k-1} \text{ for } k = 0, 1, \ldots, L-1 \tag{7}$$

where $\theta > 0$ is the mixing coefficient. Let $X^k = \left[(\nabla_i^k)^\top; \ldots; (\nabla_n^k)^\top\right] \in \mathbb{R}^{n \times d}$ for any integer $k \geq -1$. Due to (7), it is not hard to verify that

$$X^{k+1} = (1+\theta)PX^k - \theta X^{k-1} \tag{8}$$

for any integer $k \geq 0$. We notice that this process enjoys the following convergence property, where $\bar{X} = \frac{1}{n}\mathbf{1}\mathbf{1}^\top X^0 = \left[\bar{\nabla}^\top; \ldots; \bar{\nabla}^\top\right]$ and $\mathbf{1}$ is the all-ones vector in $\mathbb{R}^n$.

Lemma 1 *(Proposition 1 in Ye et al. (2023)) Under Assumption 1, for $L \geq 1$, the iterations of (8) with $\theta = (1 + \sqrt{1 - \sigma_2^2(P)})^{-1}$ ensure that*

$$\left\|X^L - \bar{X}\right\|_F \leq \sqrt{14}\left(1 - \left(1 - \frac{1}{\sqrt{2}}\right)\sqrt{1 - \sigma_2(P)}\right)^L \left\|X^0 - \bar{X}\right\|_F.$$

3.2. Our Algorithms with Improved Regret Bounds

In the following, we first propose our algorithm for D-OCO with convex functions, and then show how to exploit the strong convexity.

3.2.1. CONVEX FUNCTIONS

Before introducing our algorithms, we first compare the regret of D-OCO against that of OCO, which will provide insights for our improvements. Specifically, compared with $O(\sqrt{T})$ regret of OGD and FTRL for OCO, the $O(n^{5/4}\rho^{-1/2}\sqrt{T})$ regret of D-OGD and D-FTRL has an additional factor of $n^{5/4}\rho^{-1/2}$. Note that this factor reflects the effect of the network size and topology, and is caused by the approximation error of the standard gossip step. For example, a critical part of the analysis for D-FTRL (Hosseini et al., 2013) is the following bound

$$\|\mathbf{z}_i(t)-\bar{\mathbf{z}}(t)\|_2=O\left(\frac{\sqrt{n}}{\rho}\right) \tag{9}$$

where $\mathbf{z}_i(t)$ is defined in (4), $\bar{\mathbf{z}}(t)=\frac{1}{n}\sum_{i=1}^{n}\mathbf{z}_i(t)$ denotes the average $\mathbf{z}_i(t)$ of all learners, and $\rho=1-\sigma_2(P)$. Since $\bar{\mathbf{z}}(t)$ is also equal to $\sum_{\tau=1}^{t-1}\bar{\mathbf{g}}(\tau)$ where $\bar{\mathbf{g}}(\tau)=\frac{1}{n}\sum_{i=1}^{n}\nabla f_{\tau,i}(\mathbf{x}_i(\tau))$, the regret of D-FTRL can be upper bounded by the regret of a virtual centralized update with $\bar{\mathbf{z}}(t)$ plus the cumulative effect of the approximation error in (9) (Hosseini et al., 2013), i.e.,

$$R_{T,i}=O\left(\frac{n}{\eta}+n\eta T\right)+O\left(n\eta T\frac{\sqrt{n}}{\rho}\right)=O\left(\frac{n}{\eta}+\frac{n^{3/2}\eta T}{\rho}\right). \tag{10}$$

By minimizing the bound in (10) with $\eta=O(\sqrt{\rho/(\sqrt{n}T)})$, we obtain the $O(n^{5/4}\rho^{-1/2}\sqrt{T})$ regret of D-FTRL.

Thus, to reduce the regret of D-OCO, we should control the approximation error caused by the standard gossip step. To this end, we propose to improve D-FTRL via the accelerated gossip strategy in (7).[5] A natural idea is to replace the standard gossip step in (4) with multiple accelerated gossip steps, i.e., first updating

$$\mathbf{z}_i^{k+1}(t)=(1+\theta)\sum_{j\in N_i}P_{ij}\mathbf{z}_j^k(t)-\theta\mathbf{z}_i^{k-1}(t)\text{ for }k=0,1,\dots,L-1 \tag{11}$$

and then setting $\mathbf{z}_i(t)=\mathbf{z}_i^L(t)$ for $t\geq 2$, where $\mathbf{z}_i^0(t)=\mathbf{z}_i(t-1)+\nabla f_{t-1,i}(\mathbf{x}_i(t-1))$, $\mathbf{z}_i^{-1}(t)=\mathbf{z}_i^{L-1}(t-1)+\nabla f_{t-1,i}(\mathbf{x}_i(t-1))$, and $\mathbf{z}_i(1)=\mathbf{z}_i^{L-1}(1)=\mathbf{0}$. One can prove that (11) ensures (see (17) in Section 4.1 for details)

$$\mathbf{z}_i(t)=\mathbf{z}_i^L(t)=\sum_{\tau=1}^{t-1}\mathbf{g}_i^{(t-\tau)L}(\tau) \tag{12}$$

where $\mathbf{g}_i^{(t-\tau)L}(\tau)$ denotes the output of virtually performing (7) with $\nabla_i=\nabla f_{\tau,i}(\mathbf{x}_i(\tau))$ and $L=(t-\tau)L$. Due to the convergence behavior of the accelerated gossip strategy, we can control the error of approximating $\bar{\mathbf{z}}(t)=\sum_{\tau=1}^{t-1}\bar{\mathbf{g}}(\tau)$ under any desired level by using a large enough L. However, this approach requires multiple communications between these learners per round, which is not allowed by D-OCO.

To address this issue, we design an online accelerated gossip strategy with only one communication per round. The key idea is to incorporate (11) into a blocking update mechanism (Garber and Kretzu, 2020; Wan et al., 2022). Specifically, we divide the total T rounds into T/L blocks, and

5. It is also possible to refine D-OGD via the accelerated gossip strategy, but the projection operation in D-OGD will make the analysis of the approximation error more complex. Thus, we focus on improving D-FTRL and its variant.

Algorithm 1 AD-FTRL

Input: η, θ, and L
Initialization: set $\mathbf{x}_i(1) = \mathbf{z}_i(1) = \mathbf{z}_i^{L-1}(1) = \mathbf{0}, \forall i \in [n]$
for $z = 1, \ldots, T/L$ **do**
  for each local learner $i \in [n]$ **do**
    If $2 \leq z$, set $k = 0$, $\mathbf{z}_i^0(z) = \mathbf{z}_i(z-1) + \mathbf{g}_i(z-1)$, $\mathbf{z}_i^{-1}(z) = \mathbf{z}_i^{L-1}(z-1) + \mathbf{g}_i(z-1)$
    for $t = (z-1)L+1, \ldots, zL$ **do**
      Play $\mathbf{x}_i(z)$ and query $\nabla f_{t,i}(\mathbf{x}_i(z))$
      If $2 \leq z$, update $\mathbf{z}_i^{k+1}(z) = (1+\theta)\sum_{j \in N_i} P_{ij}\mathbf{z}_j^k(z) - \theta\mathbf{z}_i^{k-1}(z)$ and $k = k+1$
    end for
    Set $\mathbf{g}_i(z) = \sum_{t \in \mathcal{T}_z} \nabla f_{t,i}(\mathbf{x}_i(z))$, where $\mathcal{T}_z = \{(z-1)L+1, \ldots, zL\}$
    If $2 \leq z$, set $\mathbf{z}_i(z) = \mathbf{z}_i^L(z)$
    Compute $\mathbf{x}_i(z+1) = \operatorname{argmin}_{\mathbf{x} \in \mathcal{K}} \langle \mathbf{z}_i(z), \mathbf{x} \rangle + \frac{1}{\eta}\|\mathbf{x}\|_2^2$
  end for
end for

only maintain a fixed decision $\mathbf{x}_i(z)$ for each learner $i \in [n]$ in block z, where T/L is assumed to be an integer without loss of generality. In this way, the sum of gradients of each learner i in block z can be denoted as $\mathbf{g}_i(z) = \sum_{t \in \mathcal{T}_z} \nabla f_{t,i}(\mathbf{x}_i(z))$, where $\mathcal{T}_z = \{(z-1)L+1, \ldots, zL\}$. Moreover, we redefine $\bar{\mathbf{g}}(z) = \frac{1}{n}\sum_{i=1}^n \mathbf{g}_i(z)$ for each block z, and maintain a local variable $\mathbf{z}_i(z)$ to approximate $\sum_{\tau=1}^{z-1} \bar{\mathbf{g}}(\tau)$ for each learner i. The good news is that now L communications can be utilized to update $\mathbf{z}_i(z)$ per block by uniformly allocating them to every round in the block. As a result, we set $\mathbf{z}_i(1) = \mathbf{0}$, and maintain $\mathbf{z}_i(z)$ for $z \geq 2$ in a way similar to (11), i.e., performing the following update during block z

$$\mathbf{z}_i^{k+1}(z) = (1+\theta)\sum_{j \in N_i} P_{ij}\mathbf{z}_j^k(z) - \theta\mathbf{z}_i^{k-1}(z) \text{ for } k = 0, 1, \ldots, L-1 \tag{13}$$

and setting $\mathbf{z}_i(z) = \mathbf{z}_i^L(z)$, where $\mathbf{z}_i^0(z) = \mathbf{z}_i(z-1) + \mathbf{g}_i(z-1)$, $\mathbf{z}_i^{-1}(z) = \mathbf{z}_i^{L-1}(z-1) + \mathbf{g}_i(z-1)$, and $\mathbf{z}_i^{L-1}(1) = \mathbf{0}$. Then, inspired by D-FTRL in (4), we initialize with $\mathbf{x}_i(1) = \mathbf{0}$, and set the decision $\mathbf{x}_i(z+1)$ for any $i \in [n]$ and $z \geq 1$ as

$$\mathbf{x}_i(z+1) = \underset{\mathbf{x} \in \mathcal{K}}{\operatorname{argmin}} \langle \mathbf{z}_i(z), \mathbf{x} \rangle + \frac{1}{\eta}\|\mathbf{x}\|_2^2.$$

We name the proposed algorithm as accelerated decentralized follow-the-regularized-leader (AD-FTRL), and summarize the complete procedure in Algorithm 1.

In the following, we first present a lemma regarding the approximation error $\|\mathbf{z}_i(z) - \bar{\mathbf{z}}(z)\|_2$ of AD-FTRL, which demonstrates the advantage of utilizing the accelerated gossip strategy.

Lemma 2 *Let $\bar{\mathbf{z}}(z) = \sum_{\tau=1}^{z-1} \bar{\mathbf{g}}(\tau)$, where $\bar{\mathbf{g}}(\tau) = \frac{1}{n}\sum_{i=1}^n \mathbf{g}_i(\tau)$, and*

$$\theta = \frac{1}{1 + \sqrt{1 - \sigma_2^2(P)}}, \; L = \left\lceil \frac{\sqrt{2}\ln(\sqrt{14n})}{(\sqrt{2}-1)\sqrt{1-\sigma_2(P)}} \right\rceil. \tag{14}$$

Under Assumptions 1 and 2, for any $i \in [n]$ and $z \in [T/L]$, Algorithm 1 with θ and L defined in (14) ensures

$$\|\mathbf{z}_i(z) - \bar{\mathbf{z}}(z)\|_2 \leq 3LG.$$

Algorithm 2 AD-FTAL

Input: θ and L
Initialization: set $\mathbf{x}_i(1) = \mathbf{x}_i(2) = \mathbf{z}_i(1) = \mathbf{z}_i^{L-1}(1) = \mathbf{0}, \forall i \in [n]$
for $z = 1, \ldots, T/L$ **do**
  for each local learner $i \in [n]$ **do**
    If $2 \leq z$, set $k = 0$, $\mathbf{z}_i^0(z) = \mathbf{z}_i(z-1) + \mathbf{d}_i(z-1)$, $\mathbf{z}_i^{-1}(z) = \mathbf{z}_i^{L-1}(z-1) + \mathbf{d}_i(z-1)$
    for $t = (z-1)L+1, \ldots, zL$ **do**
      Play $\mathbf{x}_i(z)$ and query $\nabla f_{t,i}(\mathbf{x}_i(z))$
      If $2 \leq z$, update $\mathbf{z}_i^{k+1}(z) = (1+\theta)\sum_{j \in N_i} P_{ij}\mathbf{z}_j^k(z) - \theta \mathbf{z}_i^{k-1}(z)$ and $k = k+1$
    end for
    Set $\mathbf{d}_i(z) = \sum_{t \in \mathcal{T}_z}(\nabla f_{t,i}(\mathbf{x}_i(z)) - \alpha \mathbf{x}_i(z))$, where $\mathcal{T}_z = \{(z-1)L+1, \ldots, zL\}$
    If $2 \leq z$, set $\mathbf{z}_i(z) = \mathbf{z}_i^L(z)$
    Compute $\mathbf{x}_i(z+1) = \operatorname{argmin}_{\mathbf{x} \in \mathcal{K}} \langle \mathbf{z}_i(z), \mathbf{x} \rangle + \frac{(z-1)L\alpha}{2}\|\mathbf{x}\|_2^2$
  end for
end for

From Lemma 2, our AD-FTRL can enjoy an error bound of $O(\rho^{-1/2}\log n)$ for approximating $\bar{\mathbf{z}}(z)$, which is tighter than the $O(\sqrt{n}\rho^{-1})$ error bound in (9). Combining our Lemma 2 with the regret analysis for D-FTRL (Hosseini et al., 2013), we establish the regret bound of AD-FTRL.

Theorem 1 *Under Assumptions 1, 2, and 3, for any $i \in [n]$, Algorithm 1 with θ and L defined in* (14) *ensures*

$$R_{T,i} \leq \frac{nR^2}{\eta} + \frac{13n\eta LTG^2}{2}.$$

From Theorem 1, by setting $\eta = R/(\sqrt{LT}G)$, all the learners of AD-FTRL enjoy a regret bound of $O(n\rho^{-1/4}\sqrt{T\log n})$ for convex functions, which is tighter than the existing $O(n^{5/4}\rho^{-1/2}\sqrt{T})$ regret bound (Yan et al., 2013; Hosseini et al., 2013) in terms of both n and ρ.

3.2.2. STRONGLY CONVEX FUNCTIONS

Now, we proceed to exploit the strongly convex property to further reduce the regret of D-OCO. Inspired by D-FTAL in (5), we only need to make two changes to our Algorithm 1. First, each learner $i \in [n]$ should share the information $\mathbf{d}_i(z) = \sum_{t \in \mathcal{T}_z}(\nabla f_{t,i}(\mathbf{x}_i(z)) - \alpha \mathbf{x}_i(z))$ that consists of both the local gradient and decision with its neighbors, and maintain a local variable $\mathbf{z}_i(z)$ to approximate $\bar{\mathbf{z}}(z) = \sum_{\tau=1}^{z-1}\bar{\mathbf{d}}(\tau)$, where $\bar{\mathbf{d}}(\tau) = \frac{1}{n}\sum_{i=1}^n \mathbf{d}_i(\tau)$. This can be achieved by iteratively performing the update in (13) during block $z \geq 2$ with a new initialization as $\mathbf{z}_i^0(z) = \mathbf{z}_i(z-1) + \mathbf{d}_i(z-1), \mathbf{z}_i^{-1}(z) = \mathbf{z}_i^{L-1}(z-1) + \mathbf{d}_i(z-1)$, and then setting $\mathbf{z}_i(z) = \mathbf{z}_i^L(z)$. Second, each learner $i \in [n]$ at the end of block $z \geq 2$ should update the decision as

$$\mathbf{x}_i(z+1) = \operatorname*{argmin}_{\mathbf{x} \in \mathcal{K}} \langle \mathbf{z}_i(z), \mathbf{x} \rangle + \frac{(z-1)L\alpha}{2}\|\mathbf{x}\|_2^2. \tag{15}$$

Note that although (15) cannot be utilized to select $\mathbf{x}_i(1)$ and $\mathbf{x}_i(2)$, we can simply set $\mathbf{x}_1(1) = \mathbf{x}_i(2) = \mathbf{0}$.

The detailed procedures for dealing with strongly convex functions are presented in Algorithm 2, and it is called accelerated decentralized follow-the-approximate-leader (AD-FTAL). In the following, we first present a generalized version of Lemma 2 to bound the approximation error $\|\mathbf{z}_i(z) - \bar{\mathbf{z}}(z)\|_2$ of AD-FTAL.

Lemma 3 *Let $\bar{\mathbf{z}}(z) = \sum_{\tau=1}^{z-1} \bar{\mathbf{d}}(\tau)$, where $\bar{\mathbf{d}}(\tau) = \frac{1}{n}\sum_{i=1}^{n} \mathbf{d}_i(\tau)$. Under Assumptions 1, 2, and 3, for any $i \in [n]$ and $z \in [T/L]$, Algorithm 2 with θ and L defined in (14) ensures*

$$\|\mathbf{z}_i(z) - \bar{\mathbf{z}}(z)\|_2 \leq 3L(G + \alpha R).$$

Lemma 3 implies that AD-FTAL also enjoys the error bound of $O(\rho^{-1/2}\log n)$ for approximating $\bar{\mathbf{z}}(z)$. Then, combining this lemma with the regret analysis for D-FTAL (Wan et al., 2021), we prove the regret bound of AD-FTAL for strongly convex functions.

Theorem 2 *Under Assumptions 1, 2, 3, and 4, for any $i \in [n]$, Algorithm 2 with θ and L defined in (14) ensures*

$$R_{T,i} \leq \frac{3nLG(7G + 13\alpha R)(\ln(T/L) + 1)}{\alpha}.$$

From Theorem 2, all the learners of AD-FTAL can exploit the strong convexity of functions to achieve a regret bound of $O(n\rho^{-1/2}(\log n)\log T)$. Compared with the $O(n\rho^{-1/4}\sqrt{T\log n})$ regret of AD-FTRL for convex functions, this bound has a much tighter dependence on T. Moreover, it is better than the existing $O(n^{3/2}\rho^{-1}\log T)$ regret bound for strongly convex functions (Yan et al., 2013; Wan et al., 2021) in terms of both n and ρ.

3.2.3. Discussions on Extreme Choices of Parameters

One may have noticed that the values of θ and L are carefully selected to establish the theoretical guarantees of our AD-FTRL and AD-FTAL. To emphasize their significance, we further consider two extreme cases: one with $\theta = 0$, and the other with $L = 1$.

First, in the extreme case with $\theta = 0$, our AD-FTRL and AD-FTAL is equivalent to improving D-FTRL and D-FTAL by only using multiple standard gossip steps, respectively. It is easy to verify that due to the slower convergence of standard gossip steps (Xiao and Boyd, 2004), this extreme case requires a larger $L = O(\rho^{-1}\log n)$ to achieve an even worse error bound of $O(\rho^{-1}\log n)$ for approximating $\bar{\mathbf{z}}(z)$. Based on this result, the regret bound of AD-FTRL and AD-FTAL will degenerate to $O(n\rho^{-1/2}\sqrt{T\log n})$ and $O(n\rho^{-1}(\log n)\log T)$, respectively. Although these bounds are also tighter than the existing $O(n^{5/4}\rho^{-1/2}\sqrt{T})$ and $O(n^{3/2}\rho^{-1}\log T)$ regret bounds for convex and strongly convex functions respectively, their dependence on ρ is worse than the regret bounds achieved in Theorems 1 and 2.

Second, if $L = 1$, our AD-FTRL and AD-FTAL become a non-blocked combination of D-FTRL and D-FTAL with the accelerated gossip strategy, respectively. Following previous studies on D-OCO (Yan et al., 2013; Hosseini et al., 2013), such a non-blocked combination may be more natural than the blocked version. However, in this way, the distance between $\mathbf{z}_i(z)$, which satisfies (12) with $t = z$ and $L = 1$, and $\bar{\mathbf{z}}(z)$ cannot be controlled as desired. Specifically, due to the newly added component in $\mathbf{z}_i(z)$, i.e., $\mathbf{g}_i^{z-\tau}(\tau)$ in (12) for τ close to $z-1$, we can only modify the analysis of Lemmas 2 and 3 to derive a worse error bound of $O(\rho^{-1/2}\sqrt{n})$. Correspondingly, the regret bound of AD-FTRL and AD-FTAL will degenerate to $O(n^{5/4}\rho^{-1/4}\sqrt{T})$ and $O(n^{3/2}\rho^{-1/2}\log T)$ respectively, whose dependence on n is much worse than the regret bounds achieved in Theorems 1 and 2.

From the above discussions, both the accelerated gossip strategy and the blocking update mechanism are critical for our desired regret bounds.

3.3. Lower Bounds

Although there still exist gaps between our improved regret bounds and the lower bounds established by Wan et al. (2022), this is mainly because they do not take the decentralized structure into account. To fill these gaps, we maximize the hardness of D-OCO by considering the 1-connected cycle graph (Duchi et al., 2011), i.e., constructing the graph $\mathcal{G}$ by placing the n nodes on a circle and only connecting each node to one neighbor on its right and left. In this topology, the adversary can make at least one learner, e.g., learner 1, suffer $O(n)$ communication delays for receiving the information of the global function $f_t(\mathbf{x})$. Because of the $O(n)$ communication delays, we can establish $\Omega(n\sqrt{nT})$ and $\Omega(n^2)$ lower bounds for convex and strongly convex functions, respectively. Finally, by exploiting the dependence of spectral properties on the network size n, we obtain lower bounds involving the spectral gap, which are presented in the following theorems.

Theorem 3 *Suppose $\mathcal{K} = [-R/\sqrt{d}, R/\sqrt{d}]^d$ which satisfies Assumption 3, and $n = 2(m+1)$ for some positive integer m. For any D-OCO algorithm, there exists a sequence of loss functions satisfying Assumption 2, a graph $\mathcal{G} = ([n], E)$, and a matrix P satisfying Assumption 1 such that*

$$\text{if } n \leq 8T+8, R_{T,1} \geq \frac{nRG\sqrt{\pi T}}{16(1-\sigma_2(P))^{1/4}}, \text{ and otherwise, } R_{T,1} \geq \frac{nRGT}{4}.$$

Theorem 4 *Suppose $\mathcal{K} = [-R/\sqrt{d}, R/\sqrt{d}]^d$, which satisfies Assumption 3 and $n = 2(m+1)$ for some positive integer m. For any D-OCO algorithm, there exists a sequence loss functions satisfying Assumption 4 and Assumption 2 with $G = 2\alpha R$, a graph $\mathcal{G} = ([n], E)$, and a matrix P satisfying Assumption 1 such that*

$$\text{if } n \leq 8T+8, R_{T,1} \geq \frac{\alpha\pi nR^2}{256\sqrt{1-\sigma_2(P)}}, \text{ and otherwise, } R_{T,1} \geq \frac{\alpha nR^2T}{16}.$$

Note that in both previous studies (Yan et al., 2013; Hosseini et al., 2013) and this paper, the upper regret bounds of D-OCO algorithms generally hold for all possible graphs and communication matrices P satisfying Assumption 1. Therefore, although lower bounds in our Theorems 3 and 4 only hold for a specific choice of the graph and P, they are sufficient to prove the tightness of the upper bound in general. Specifically, Theorem 3 establishes a lower bound of $\Omega(n\rho^{-1/4}\sqrt{T})$ for D-OCO with convex functions, which matches the $O(n\rho^{-1/4}\sqrt{T\log n})$ regret of our AD-FTRL up to polylogarithmic factors. For D-OCO with strongly convex functions, Theorem 4 establishes a lower bound of $\Omega(n\rho^{-1/2})$, which matches the $O(n\rho^{-1/2}(\log n)\log T)$ regret of our AD-FTAL up to polylogarithmic factors. To the best of our knowledge, this paper provides the first lower bounds that reveal the effect of the spectral gap on D-OCO.

4. Analysis

We here present the proof of Lemmas 2 and 3, and the omitted proofs can be found in the appendix.

4.1. Proof of Lemmas 2 and 3

Let $\mathbf{g}_i^0(z) = \mathbf{g}_i^{-1}(z) = \mathbf{g}_i(z)$. For any $i \in [n]$, $z \in [T/L-1]$, and any non-negative integer k, we first define a virtual update as

$$\mathbf{g}_i^{k+1}(z) = (1+\theta)\sum_{j\in N_i} P_{ij}\mathbf{g}_j^k(z) - \theta\mathbf{g}_i^{k-1}(z). \tag{16}$$

In the following, we will prove that for any $z = 2, \ldots, T/L$, Algorithm 1 ensures

$$\mathbf{z}_i^k(z) = \sum_{\tau=1}^{z-1} \mathbf{g}_i^{(z-\tau-1)L+k}(\tau),\ \forall k = 1, \ldots, L \tag{17}$$

by the induction method.

It is easy to verify that (17) holds for $z = 2$ due to $\mathbf{z}_i^0(2) = \mathbf{z}_i^{-1}(2) = \mathbf{g}_i(1)$ and (16). Then, we assume that (17) holds for some $z > 2$, and prove it also holds for $z + 1$. From step 5 of our Algorithm 1, we have

$$\begin{aligned}
\mathbf{z}_i^0(z+1) &= \mathbf{z}_i(z) + \mathbf{g}_i(z) = \mathbf{z}_i^L(z) + \mathbf{g}_i^0(z) \overset{(17)}{=} \sum_{\tau=1}^{z} \mathbf{g}_i^{(z-\tau)L}(\tau), \\
\mathbf{z}_i^{-1}(z+1) &= \mathbf{z}_i^{L-1}(z) + \mathbf{g}_i(z) = \mathbf{z}_i^{L-1}(z) + \mathbf{g}_i^{-1}(z) \overset{(17)}{=} \sum_{\tau=1}^{z} \mathbf{g}_i^{(z-\tau)L-1}(\tau).
\end{aligned} \tag{18}$$

Combining (18) with step 8 of Algorithm 1, for $k = 1$, we have

$$\begin{aligned}
\mathbf{z}_i^k(z+1) =&(1+\theta) \sum_{j\in N_i} P_{ij}\mathbf{z}_j^{k-1}(z+1) - \theta \mathbf{z}_i^{k-2}(z+1) \\
=&(1+\theta) \sum_{j\in N_i} P_{ij} \sum_{\tau=1}^{z} \mathbf{g}_j^{(z-\tau)L+k-1}(\tau) - \theta \sum_{\tau=1}^{z} \mathbf{g}_i^{(z-\tau)L-1+k-1}(\tau) \\
=&\sum_{\tau=1}^{z} \left((1+\theta) \sum_{j\in N_i} P_{ij}\mathbf{g}_j^{(z-\tau)L+k-1}(\tau) - \theta \mathbf{g}_i^{(z-\tau)L-1+k-1}(\tau) \right) \\
\overset{(16)}{=}&\sum_{\tau=1}^{z} \mathbf{g}_i^{(z-\tau)L+k}(\tau).
\end{aligned}$$

By repeating the above equality for $k = 2, \ldots, L$, the proof of (17) for $z + 1$ is completed.

From (17), for any $i \in [n]$ and $z = 2, \ldots, T/L$, we have

$$\|\mathbf{z}_i(z) - \bar{\mathbf{z}}(z)\|_2 = \left\| \sum_{\tau=1}^{z-1} \mathbf{g}_i^{(z-\tau)L}(\tau) - \sum_{\tau=1}^{z-1} \bar{\mathbf{g}}(\tau) \right\|_2 \leq \sum_{\tau=1}^{z-1} \left\| \mathbf{g}_i^{(z-\tau)L}(\tau) - \bar{\mathbf{g}}(\tau) \right\|_2. \tag{19}$$

To further analyze the right side of (19), we define

$$X^k = \left[\mathbf{g}_1^k(\tau)^\top; \ldots; \mathbf{g}_n^k(\tau)^\top \right] \in \mathbb{R}^{n\times d}$$

for any integer $k \geq -1$, $\bar{X} = \frac{1}{n}\mathbf{1}\mathbf{1}^\top X^0 = \left[\bar{\mathbf{g}}(\tau)^\top; \ldots; \bar{\mathbf{g}}(\tau)^\top \right]$, and $c = 1 - 1/\sqrt{2}$. According to (16), it is not hard to verify that the sequence of $X^1, \ldots, X^L$ follows the update rule in (8).

Then, combining with Lemma 1, for any $\tau \leq z$, we have

$$\begin{aligned}
&\left\| X^{(z-\tau)L} - \bar{X} \right\|_F \leq \sqrt{14} \left(1 - c\sqrt{1-\sigma_2(P)} \right)^{(z-\tau)L} \left\| X^0 - \bar{X} \right\|_F \\
\leq&\sqrt{14} \left(1 - c\sqrt{1-\sigma_2(P)} \right)^{(z-\tau)L} \left(\left\| X^0 \right\|_F + \left\| \bar{X} \right\|_F \right) \\
=&\sqrt{14} \left(1 - c\sqrt{1-\sigma_2(P)} \right)^{(z-\tau)L} \left(\sqrt{\sum_{i=1}^{n} \|\mathbf{g}_i(\tau)\|_2^2} + \sqrt{n\|\bar{\mathbf{g}}(\tau)\|_2^2} \right).
\end{aligned} \tag{20}$$

Because of Assumption 2, for any $z \in [T/L]$ and $i \in [n]$, it easy to verify that

$$\|\mathbf{g}_i(z)\|_2 = \left\|\sum_{t\in\mathcal{T}_z} \nabla f_{t,i}(\mathbf{x}_i(z))\right\|_2 \leq LG, \ \|\bar{\mathbf{g}}(z)\|_2 = \left\|\frac{1}{n}\sum_{i=1}^{n} \mathbf{g}_i(z)\right\|_2 \leq LG. \tag{21}$$

Additionally, because of the value of L in (14), we have

$$\begin{aligned}\epsilon =& \left(1 - c\sqrt{1-\sigma_2(P)}\right)^L \leq \left(1 - c\sqrt{1-\sigma_2(P)}\right)^{\frac{\ln(\sqrt{14n})}{c\sqrt{1-\sigma_2(P)}}} \\ \leq& \left(1 - c\sqrt{1-\sigma_2(P)}\right)^{\frac{\ln(\sqrt{14n})}{\ln\left(1-c\sqrt{1-\sigma_2(P)}\right)^{-1}}} = \frac{1}{\sqrt{14n}}\end{aligned} \tag{22}$$

where the second inequality is due to $\ln x^{-1} \geq 1 - x$ for any $x > 0$.

Combining (20) with (21) and (22), for any $i \in [n]$ and $\tau \leq z$, we have

$$\begin{aligned}\left\|\mathbf{g}_i^{(z-\tau)L}(\tau) - \bar{\mathbf{g}}(\tau)\right\|_2 \leq \left\|X^{(z-\tau)L} - \bar{X}\right\|_F \leq& 2\sqrt{14n}\left(1 - c\sqrt{1-\sigma_2(P)}\right)^{(z-\tau)L} LG \\ \leq& 2\left(1 - c\sqrt{1-\sigma_2(P)}\right)^{(z-\tau-1)L} LG.\end{aligned} \tag{23}$$

Moreover, combining (19) with (23), for any $i \in [n]$ and $z = 2, \ldots, T/L$, we have

$$\begin{aligned}\|\mathbf{z}_i(z) - \bar{\mathbf{z}}(z)\|_2 \leq& 2LG\sum_{\tau=1}^{z-1}\left(1 - c\sqrt{1-\sigma_2(P)}\right)^{(z-\tau-1)L} = 2LG\sum_{\tau=0}^{z-2}\epsilon^\tau \\ \leq& \frac{2LG}{1-\epsilon} \overset{(22)}{\leq} 2LG + \frac{2LG}{\sqrt{14n}-1} \leq 3LG\end{aligned} \tag{24}$$

where the last inequality is due to $\sqrt{14n} > 3$ for any $n \geq 1$. Now, we can complete the proof of Lemma 2 by combining (24) with $\|\mathbf{z}_i(1) - \bar{\mathbf{z}}(1)\|_2 = 0$.

Note that Lemma 3 can be proved by simply repeating the above procedures with the update rule of $\mathbf{z}_i(z)$ in Algorithm 2 and replacing the norm bound of gradients in (21) with

$$\begin{aligned}\|\mathbf{d}_i(z)\|_2 =& \left\|\sum_{t\in\mathcal{T}_z}(\nabla f_{t,i}(\mathbf{x}_i(z)) - \alpha\mathbf{x}_i(z))\right\|_2 \leq L(G + \alpha R), \\ \|\bar{\mathbf{d}}(z)\|_2 =& \left\|\frac{1}{n}\sum_{i=1}^{n}\mathbf{d}_i(z)\right\|_2 \leq L(G + \alpha R)\end{aligned} \tag{25}$$

where the first inequality is due to Assumptions 2 and 3.

5. Conclusion and Future Work

This paper investigates D-OCO with convex and strongly convex functions respectively, and aims to fill the gaps between the existing upper and lower bounds. First, we propose a novel algorithm for D-OCO with convex functions, namely AD-FTRL, which reduces the existing $O(n^{5/4}\rho^{-1/2}\sqrt{T})$ regret bound to $\tilde{O}(n\rho^{-1/4}\sqrt{T})$. Second, we propose a variant of AD-FTRL for strongly convex functions, namely AD-FTAL, which achieves a better regret bound of $\tilde{O}(n\rho^{-1/2}\log T)$. Finally, we demonstrate their optimality by deriving $\Omega(n\rho^{-1/4}\sqrt{T})$ and $\Omega(n\rho^{-1/2})$ lower bounds for convex and strongly convex functions, respectively.

Both AD-FTRL and AD-FTAL proposed in this paper require the projection operation to update the decision, which could be time-consuming for complex decision sets. Thus, it is interesting to design projection-free variants of our algorithms by replacing the projection with linear optimization steps. Although there already exist projection-free algorithms for D-OCO (Wan et al., 2020, 2021, 2022), they follow the standard gossip step, and thus only achieve $O(n^{5/4}\rho^{-1/2}T^{3/4})$ and $O(n^{3/2}\rho^{-1}T^{2/3}\log^{1/3}T)$ regret bounds for convex and strongly convex functions, respectively. In contrast, projection-free variants of our algorithms are expected to achieve refined regret bounds of $\tilde{O}(n\rho^{-1/4}T^{3/4})$ and $\tilde{O}(n\rho^{-1/2}T^{2/3}\log^{1/3}T)$, respectively.

Acknowledgments

This work was partially supported by the National Natural Science Foundation of China (62306275, 62122037), the Zhejiang Province High-Level Talents Special Support Program "Leading Talent of Technological Innovation of Ten-Thousands Talents Program" (No. 2022R52046), the Key Research and Development Program of Zhejiang Province (No. 2023C03192), and the Open Research Fund of the State Key Laboratory of Blockchain and Data Security, Zhejiang University. The authors would also like to thank the anonymous reviewers for their helpful comments.

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.

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.

Stephen Boyd, Arpita Ghosh, Balaji Prabhakar, and Devavrat Shah. Randomized gossip algorithms. *IEEE Transactions on Information Theory*, 52(6):2508–2530, 2006.

Nicolò Cesa-Bianchi, Alex Conconi, and Claudio Gentile. On the generalization ability of on-line learning algorithms. *IEEE Transactions on Information Theory*, 50(9):2050–2057, 2004.

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.

Darina Dvinskikh and Alexander Gasnikov. Decentralized and parallel primal and dual accelerated methods for stochastic convex programming problems. *Journal of Inverse and Ill-posed Problems*, 29(3):385–405, 2021.

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

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.

Eduard Gorbunov, Darina Dvinskikh, and Alexander Gasnikov. Optimal decentralized distributed algorithms for stochastic convex optimization. *arXiv:1911.07363*, 2020.

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, Amit Agarwal, and Satyen Kale. Logarithmic regret algorithms for online convex optimization. *Machine Learning*, 69(2):169–192, 2007.

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.

Dmitry Kovalev, Adil Salim, and Peter Richtarik. Optimal and practical algorithms for smooth and strongly convex decentralized optimization. In *Advances in Neural Information Processing Systems 33*, pages 18342–18352, 2020.

Jinlong Lei, Peng Yi, Yiguang Hong, Jie Chen, and Guodong Shi. Online convex optimization over Erdős–Rényi random networks. In *Advances in Neural Information Processing Systems 33*, pages 15591–15601, 2020.

Victor Lesser, Charles L. Ortiz, and Milind Tambe, editors. *Distributed Sensor Networks: A Multiagent Perspective*. Springer New York, 2003.

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.

Ji Liu and A. Stephen Morse. Accelerated linear iterations for distributed averaging. *Annual Reviews in Control*, 35(2):160–165, 2011.

Yucheng Lu and Christopher De Sa. Optimal complexity in decentralized training. In *Proceedings of the 38th International Conference on Machine Learning*, pages 7111–7123, 2021.

Yurii Nesterov. Primal-dual subgradient methods for convex problems. *Mathematical Programming*, 120(1):221–259, 2009.

Kevin Scaman, Francis Bach, Sébastien Bubeck, Yin Tat Lee, and Laurent Massoulié. Optimal algorithms for smooth and strongly convex distributed optimization in networks. In *Proceedings of the 34th International Conference on Machine Learning*, pages 3027–3036, 2017.

Kevin Scaman, Francis Bach, Sébastien Bubeck, Yin Tat Lee, and Laurent Massoulié. Optimal algorithms for non-smooth distributed optimization in networks. In *Advances in Neural Information Processing Systems 31*, pages 2740–2749, 2018.

Kevin Scaman, Francis Bach, Sébastien Bubeck, Yin Tat Lee, and Laurent Massoulié. Optimal convergence rates for convex distributed optimization in networks. *Journal of Machine Learning Research*, 20(159):1–31, 2019.

Shahin Shahrampour and Ali Jadbabaie. Distributed online optimization in dynamic environments using mirror descent. *IEEE Transactions on Automatic Control*, 63(3):714–725, 2018.

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

Shai Shalev-Shwartz and Yoram Singer. A primal-dual perspective of online learning algorithm. *Machine Learning*, 69(2–3):115–142, 2007.

Zhuoqing Song, Lei Shi, Shi Pu, and Ming Yan. Optimal gradient tracking for decentralized optimization. *Mathematical Programming*, pages 1–53, 2023.

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, Guanghui Wang, and Lijun Zhang. Projection-free distributed online learning with strongly convex losses. *arXiv:2103.11102v1*, 2021.

Yuanyu Wan, Guanghui Wang, Wei-Wei Tu, and Lijun Zhang. Projection-free distributed online learning with sublinear communication complexity. *Journal of Machine Learning Research*, 23 (172):1–53, 2022.

Yibo Wang, Yuanyu Wan, Shimao Zhang, and Lijun Zhang. Distributed projection-free online learning for smooth and convex losses. In *Proceedings of the 37th AAAI Conference on Artificial Intelligence*, pages 10226–10234, 2023.

Lin Xiao. Dual averaging method for regularized stochastic learning and online optimization. In *Advances in Neural Information Processing Systems 22*, pages 2116–2124, 2009.

Lin Xiao and Stephen Boyd. Fast linear iterations for distributed averaging. *Systems and Control Letters*, 53(1):65–78, 2004.

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.

Haishan Ye and Xiangyu Chang. Optimal decentralized composite optimization for strongly convex functions. *arXiv:2312.15845*, 2023.

Haishan Ye, Luo Luo, Ziang Zhou, and Tong Zhang. Multi-consensus decentralized accelerated gradient descent. *Journal of Machine Learning Research*, 24(306):1–50, 2023.

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.

Xinlei Yi, Xiuxian Li, Tao Yang, Lihua Xie, Tianyou Chai, and Karl Henrik Johansson. Distributed bandit online convex optimization with time-varying coupled inequality constraints. *IEEE Transactions on Automatic Control*, 66(10):4620–4635, 2021.

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.

Appendix A. Proof of Theorem 1

According to Algorithm 1, the total T rounds are divided into T/L blocks. For any block $z \in [T/L]$, we define a virtual decision

$$\bar{\mathbf{x}}(z) = \operatorname*{argmin}_{\mathbf{x}\in\mathcal{K}} \langle \mathbf{x}, \bar{\mathbf{z}}(z)\rangle + \frac{1}{\eta}\|\mathbf{x}\|_2^2 \tag{26}$$

where $\bar{\mathbf{z}}(z) = \sum_{\tau=1}^{z-1} \bar{\mathbf{g}}(\tau)$ and $\bar{\mathbf{g}}(\tau) = \frac{1}{n}\sum_{i=1}^{n} \mathbf{g}_i(\tau)$. In the following, we will bound the regret of any learner i by analyzing the regret of the virtual decisions on a sequence of linear losses defined by $\bar{\mathbf{g}}(1), \ldots, \bar{\mathbf{g}}(T/L)$ and the distance $\|\mathbf{x}_i(z) - \bar{\mathbf{x}}(z)\|_2$ for any $z \in [T/L]$.

To this end, we first introduce two useful lemmas.

Lemma 4 *(Lemma 6.6 in Garber and Hazan (2016)) Let $\{\ell_t(\mathbf{x})\}_{t=1}^T$ be a sequence of functions and $\mathbf{x}_t^* \in \operatorname{argmin}_{\mathbf{x}\in\mathcal{K}} \sum_{\tau=1}^{t} \ell_\tau(\mathbf{x})$ for any $t \in [T]$. Then, it holds that*

$$\sum_{t=1}^{T} \ell_t(\mathbf{x}_t^*) - \min_{\mathbf{x}\in\mathcal{K}} \sum_{t=1}^{T} \ell_t(\mathbf{x}) \leq 0.$$

Lemma 5 *(Lemma 5 in Duchi et al. (2011)) Let $\Pi_{\mathcal{K}}(\mathbf{u}, \eta) = \operatorname{argmin}_{\mathbf{x}\in\mathcal{K}}\langle \mathbf{u}, \mathbf{x}\rangle + \frac{1}{\eta}\|\mathbf{x}\|_2^2$. For any $\mathbf{u}, \mathbf{v} \in \mathbb{R}^d$, we have*

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

Let $\ell_1(\mathbf{x}) = \langle \mathbf{x}, \bar{\mathbf{g}}(1)\rangle + \frac{1}{\eta}\|\mathbf{x}\|_2^2$ and $\ell_z(\mathbf{x}) = \langle \mathbf{x}, \bar{\mathbf{g}}(z)\rangle$ for any $z = 2, \ldots, T/L$. Combining Lemma 4 with the definition of $\bar{\mathbf{x}}(z)$ in (26), for any $\mathbf{x} \in \mathcal{K}$, we have

$$\begin{aligned}
&\sum_{z=1}^{T/L} \langle \bar{\mathbf{x}}(z+1) - \mathbf{x}, \bar{\mathbf{g}}(z)\rangle + \frac{\|\bar{\mathbf{x}}(2)\|_2^2 - \|\mathbf{x}\|_2^2}{\eta} \\
=&\sum_{z=1}^{T/L} \left(\ell_z(\bar{\mathbf{x}}(z+1)) - \ell_z(\mathbf{x})\right) \leq 0.
\end{aligned} \tag{27}$$

Then, combining with Assumptions 2 and 3, it is not hard to verify that

$$\begin{aligned}
\sum_{z=1}^{T/L} \langle \bar{\mathbf{x}}(z) - \mathbf{x}, \bar{\mathbf{g}}(z)\rangle =& \sum_{z=1}^{T/L} \langle \bar{\mathbf{x}}(z+1) - \mathbf{x}, \bar{\mathbf{g}}(z)\rangle + \sum_{z=1}^{T/L} \langle \bar{\mathbf{x}}(z) - \bar{\mathbf{x}}(z+1), \bar{\mathbf{g}}(z)\rangle \\
\overset{(27)}{\leq}& \frac{\|\mathbf{x}\|_2^2 - \|\bar{\mathbf{x}}(2)\|_2^2}{\eta} + \sum_{z=1}^{T/L} \langle \bar{\mathbf{x}}(z) - \bar{\mathbf{x}}(z+1), \bar{\mathbf{g}}(z)\rangle \\
\leq& \frac{R^2}{\eta} + \sum_{z=1}^{T/L} \|\bar{\mathbf{x}}(z) - \bar{\mathbf{x}}(z+1)\|_2 \|\bar{\mathbf{g}}(z)\|_2 \\
\leq& \frac{R^2}{\eta} + \sum_{z=1}^{T/L} \frac{\eta}{2}\|\bar{\mathbf{g}}(z)\|_2^2 \overset{(21)}{\leq} \frac{R^2}{\eta} + \frac{\eta T L G^2}{2}
\end{aligned} \tag{28}$$

where the third inequality is due to Lemma 5 and the definition of $\bar{\mathbf{x}}(z)$ in (26).

Next, we proceed to analyze the distance between $\bar{\mathbf{x}}(z)$ and $\mathbf{x}_i(z)$. Let $\mathbf{z}_i(0) = \mathbf{0}$ and $\bar{\mathbf{z}}(0) = \mathbf{0}$. For any $z \in [T/L]$ and $i \in [n]$, Algorithm 1 ensures

$$\mathbf{x}_i(z) = \operatorname*{argmin}_{\mathbf{x}\in\mathcal{K}} \langle \mathbf{x}, \mathbf{z}_i(z-1)\rangle + \frac{1}{\eta}\|\mathbf{x}\|_2^2.$$

According to Lemma 5, for any $z \in [T/L]$ and $i \in [n]$, we have

$$\begin{aligned}\|\mathbf{x}_i(z) - \bar{\mathbf{x}}(z)\|_2 \leq & \frac{\eta}{2}\left\|\mathbf{z}_i(z-1) - \bar{\mathbf{z}}(z)\right\|_2 \\ \leq & \frac{\eta}{2}\left\|\mathbf{z}_i(z-1) - \bar{\mathbf{z}}(z-1)\right\|_2 + \frac{\eta}{2}\|\bar{\mathbf{g}}(z-1)\|_2 \leq 2\eta LG\end{aligned} \tag{29}$$

where the last inequality is due to Lemma 2 and (21).

Now, we are ready to derive the regret bound of any learner i. Due to Assumption 2, for any $z \in [T/L]$, $t \in \mathcal{T}_z$, $j \in [n]$, and $\mathbf{x} \in \mathcal{K}$, we have

$$\begin{aligned}& f_{t,j}(\mathbf{x}_i(z)) - f_{t,j}(\mathbf{x}) \\ \leq & f_{t,j}(\mathbf{x}_j(z)) - f_{t,j}(\mathbf{x}) + G\|\mathbf{x}_j(z) - \mathbf{x}_i(z)\|_2 \\ \leq & \langle \nabla f_{t,j}(\mathbf{x}_j(z)), \mathbf{x}_j(z) - \mathbf{x}\rangle + G\|\mathbf{x}_j(z) - \bar{\mathbf{x}}(z)\|_2 + G\|\bar{\mathbf{x}}(z) - \mathbf{x}_i(z)\|_2 \\ \overset{(29)}{\leq} & \langle \nabla f_{t,j}(\mathbf{x}_j(z)), \bar{\mathbf{x}}(z) - \mathbf{x} + \mathbf{x}_j(z) - \bar{\mathbf{x}}(z)\rangle + 4\eta LG^2 \\ \leq & \langle \nabla f_{t,j}(\mathbf{x}_j(z)), \bar{\mathbf{x}}(z) - \mathbf{x}\rangle + G\|\mathbf{x}_j(z) - \bar{\mathbf{x}}(z)\|_2 + 4\eta LG^2 \\ \overset{(29)}{\leq} & \langle \nabla f_{t,j}(\mathbf{x}_j(z)), \bar{\mathbf{x}}(z) - \mathbf{x}\rangle + 6\eta LG^2.\end{aligned} \tag{30}$$

Then, for any $\mathbf{x} \in \mathcal{K}$, it is not hard to verify that

$$\begin{aligned}& \sum_{z=1}^{T/L}\sum_{t\in\mathcal{T}_z}\sum_{j=1}^{n} f_{t,j}(\mathbf{x}_i(z)) - \sum_{z=1}^{T/L}\sum_{t\in\mathcal{T}_z}\sum_{j=1}^{n} f_{t,j}(\mathbf{x}) \\ \overset{(30)}{\leq} & \sum_{z=1}^{T/L}\sum_{t\in\mathcal{T}_z}\sum_{j=1}^{n}\langle \nabla f_{t,j}(\mathbf{x}_j(z)), \bar{\mathbf{x}}(z) - \mathbf{x}\rangle + 6n\eta LTG^2 \\ = & \sum_{z=1}^{T/L} n\langle \bar{\mathbf{g}}(z), \bar{\mathbf{x}}(z) - \mathbf{x}\rangle + 6n\eta LTG^2 \overset{(28)}{\leq} \frac{nR^2}{\eta} + \frac{13n\eta LTG^2}{2}\end{aligned}$$

which completes this proof.

Appendix B. Proof of Theorem 2

This proof is similar to that of Theorem 1, but some specific extensions are required for utilizing the strong convexity of functions. First, for $z = 2, \ldots, T/L + 1$, we define a virtual decision

$$\bar{\mathbf{x}}(z) = \operatorname*{argmin}_{\mathbf{x}\in\mathcal{K}} \langle \mathbf{x}, \bar{\mathbf{z}}(z)\rangle + \frac{(z-1)L\alpha}{2}\|\mathbf{x}\|_2^2 \tag{31}$$

where $\bar{\mathbf{z}}(z) = \sum_{\tau=1}^{z-1}\bar{\mathbf{d}}(\tau)$ and $\bar{\mathbf{d}}(\tau) = \frac{1}{n}\sum_{i=1}^{n}\mathbf{d}_i(\tau)$. In the following, we will bound the regret of any learner i by analyzing the regret of $\bar{\mathbf{x}}(2), \ldots, \bar{\mathbf{x}}(T/L+1)$ on a sequence of loss functions defined by $\bar{\mathbf{d}}(1), \ldots, \bar{\mathbf{d}}(T/L)$ and the distance $\|\mathbf{x}_i(z) - \bar{\mathbf{x}}(z+1)\|_2$ for any $z \in [T/L]$.

Specifically, for any $z \in [T/L]$, we define $\ell_z(\mathbf{x}) = \langle \mathbf{x}, \bar{\mathbf{d}}(z) \rangle + \frac{L\alpha}{2}\|\mathbf{x}\|_2^2$. Combining Lemma 4 with the definition in (31), for any $\mathbf{x} \in \mathcal{K}$, we have

$$\begin{aligned}&\sum_{z=1}^{T/L}\left(\langle \bar{\mathbf{x}}(z+1) - \mathbf{x}, \bar{\mathbf{d}}(z)\rangle + \frac{L\alpha}{2}(\|\bar{\mathbf{x}}(z+1)\|_2^2 - \|\mathbf{x}\|_2^2)\right)\\ =&\sum_{z=1}^{T/L}(\ell_z(\bar{\mathbf{x}}(z+1)) - \ell_z(\mathbf{x})) \leq 0.\end{aligned} \tag{32}$$

We also notice that for any $z = 3, \ldots, T/L$, Algorithm 2 ensures

$$\mathbf{x}_i(z) = \operatorname*{argmin}_{\mathbf{x}\in\mathcal{K}} \langle \mathbf{x}, \mathbf{z}_i(z-1)\rangle + \frac{(z-2)L\alpha}{2}\|\mathbf{x}\|_2^2. \tag{33}$$

Combining Lemma 5 with (31) and (33), for any $z = 3, \ldots, T/L$, we have

$$\begin{aligned}\|\mathbf{x}_i(z) - \bar{\mathbf{x}}(z-1)\|_2 \leq &\frac{1}{(z-2)L\alpha}\|\mathbf{z}_i(z-1) - \bar{\mathbf{z}}(z-1)\|_2\\ \leq &\frac{3(G+\alpha R)}{(z-2)\alpha}\end{aligned} \tag{34}$$

where the last inequality is due to Lemma 3.

To bound $\|\mathbf{x}_i(z) - \bar{\mathbf{x}}(z+1)\|_2$, we still need to analyze the term $\|\bar{\mathbf{x}}(z-1) - \bar{\mathbf{x}}(z+1)\|_2$. Let $F_z(\mathbf{x}) = \sum_{\tau=1}^{z}\ell_\tau(\mathbf{x})$ for any $z \in [T/L]$. It is easy to verify that $F_z(\mathbf{x})$ is $(zL\alpha)$-strongly convex over $\mathcal{K}$, and $\bar{\mathbf{x}}(z+1) = \operatorname{argmin}_{\mathbf{x}\in\mathcal{K}} F_z(\mathbf{x})$. Note that as proved by Hazan and Kale (2012), for any α-strongly convex function $f(\mathbf{x}) : \mathcal{K} \mapsto \mathbb{R}$ and $\mathbf{x} \in \mathcal{K}$, it holds that

$$\frac{\alpha}{2}\|\mathbf{x} - \mathbf{x}^*\|_2^2 \leq f(\mathbf{x}) - f(\mathbf{x}^*) \tag{35}$$

where $\mathbf{x}^* = \operatorname{argmin}_{\mathbf{x}\in\mathcal{K}} f(\mathbf{x})$. Moreover, for any $\mathbf{x}, \mathbf{y} \in \mathcal{K}$ and $z \in [T/L]$, we have

$$\begin{aligned}|\ell_z(\mathbf{x}) - \ell_z(\mathbf{y})| \leq &|\langle \nabla\ell_z(\mathbf{x}), \mathbf{x} - \mathbf{y}\rangle| \leq \|\nabla\ell_z(\mathbf{x})\|_2\|\mathbf{x} - \mathbf{y}\|_2\\ =&\|\bar{\mathbf{d}}(z) + \alpha L\mathbf{x}\|_2\|\mathbf{x} - \mathbf{y}\|_2 \overset{(25)}{\leq} L(G + 2\alpha R)\|\mathbf{x} - \mathbf{y}\|_2.\end{aligned} \tag{36}$$

Then, for any $z = 3, \ldots, T/L$, it is not hard to verify that

$$\begin{aligned}&\|\bar{\mathbf{x}}(z-1) - \bar{\mathbf{x}}(z+1)\|_2^2\\ \overset{(35)}{\leq}&\frac{2}{zL\alpha}(F_z(\bar{\mathbf{x}}(z-1)) - F_z(\bar{\mathbf{x}}(z+1)))\\ =&\frac{2}{zL\alpha}\left(F_{z-2}(\bar{\mathbf{x}}(z-1)) - F_{z-2}(\bar{\mathbf{x}}(z+1)) + \sum_{\tau=z-1}^{z}(\ell_\tau(\bar{\mathbf{x}}(z-1)) - \ell_\tau(\bar{\mathbf{x}}(z+1)))\right)\\ \overset{(31)}{\leq}&\frac{2}{zL\alpha}\sum_{\tau=z-1}^{z}(\ell_\tau(\bar{\mathbf{x}}(z-1)) - \ell_\tau(\bar{\mathbf{x}}(z+1))) \overset{(36)}{\leq} \frac{4(G+2\alpha R)}{z\alpha}\|\bar{\mathbf{x}}(z-1) - \bar{\mathbf{x}}(z+1)\|_2\end{aligned}$$

which implies that

$$\|\bar{\mathbf{x}}(z-1) - \bar{\mathbf{x}}(z+1)\|_2 \leq \frac{4(G+2\alpha R)}{z\alpha}. \tag{37}$$

Combining (34) and (37), for any $z = 3, \ldots, T/L$, we have

$$\begin{aligned}\|\mathbf{x}_i(z) - \bar{\mathbf{x}}(z+1)\|_2 \leq & \|\mathbf{x}_i(z) - \bar{\mathbf{x}}(z-1)\|_2 + \|\bar{\mathbf{x}}(z-1) - \bar{\mathbf{x}}(z+1)\|_2 \\ \leq & \frac{3(G+\alpha R)}{(z-2)\alpha} + \frac{4(G+2\alpha R)}{z\alpha} \leq \frac{7G+11\alpha R}{(z-2)\alpha}.\end{aligned} \tag{38}$$

For $z = 1$ and $z = 2$, because of $\mathbf{x}_i(1) = \mathbf{x}_i(2) = \mathbf{0}$ and Assumption 3, we have

$$\|\mathbf{x}_i(z) - \bar{\mathbf{x}}(z+1)\|_2 = \|\bar{\mathbf{x}}(z+1)\|_2 \leq R. \tag{39}$$

Now, we are ready to derive the regret bound of any learner i. For brevity, let $\epsilon_z = \frac{7G+11\alpha R}{(z-2)\alpha}$ for any $z = 3, \ldots, T/L$ and $\epsilon_z = R$ for $z \in \{1, 2\}$. For any $z \in [T/L]$, $t \in \mathcal{T}_z$, $j \in [n]$, and $\mathbf{x} \in \mathcal{K}$, because of Assumptions 2 and 4, we have

$$\begin{aligned}& f_{t,j}(\mathbf{x}_i(z)) - f_{t,j}(\mathbf{x}) \\ \leq & f_{t,j}(\mathbf{x}_j(z)) - f_{t,j}(\mathbf{x}) + G\|\mathbf{x}_j(z) - \mathbf{x}_i(z)\|_2 \\ \leq & \langle \nabla f_{t,j}(\mathbf{x}_j(z)), \mathbf{x}_j(z) - \mathbf{x}\rangle - \frac{\alpha}{2}\|\mathbf{x}_j(z) - \mathbf{x}\|_2^2 + G\|\mathbf{x}_j(z) - \bar{\mathbf{x}}(z+1) + \bar{\mathbf{x}}(z+1) - \mathbf{x}_i(z)\|_2 \\ \leq & \langle \nabla f_{t,j}(\mathbf{x}_j(z)), \bar{\mathbf{x}}(z+1) - \mathbf{x}\rangle - \frac{\alpha}{2}\|\mathbf{x}_j(z) - \mathbf{x}\|_2^2 + \langle \nabla f_{t,j}(\mathbf{x}_j(z)), \mathbf{x}_j(z) - \bar{\mathbf{x}}(z+1)\rangle + 2G\epsilon_z \\ \leq & \langle \nabla f_{t,j}(\mathbf{x}_j(z)), \bar{\mathbf{x}}(z+1) - \mathbf{x}\rangle - \frac{\alpha}{2}\|\mathbf{x}_j(z) - \mathbf{x}\|_2^2 + G\|\mathbf{x}_j(z) - \bar{\mathbf{x}}(z+1)\|_2 + 2G\epsilon_z \\ \leq & \langle \nabla f_{t,j}(\mathbf{x}_j(z)), \bar{\mathbf{x}}(z+1) - \mathbf{x}\rangle - \frac{\alpha}{2}\|\mathbf{x}_j(z) - \mathbf{x}\|_2^2 + 3G\epsilon_z\end{aligned}$$

where the third and last inequalities are due to (38) and (39).

Moreover, we also have

$$\begin{aligned}\|\mathbf{x}_j(z) - \mathbf{x}\|_2^2 = & \|\mathbf{x}_j(z) - \bar{\mathbf{x}}(z+1)\|_2^2 + 2\langle \mathbf{x}_j(z), \bar{\mathbf{x}}(z+1) - \mathbf{x}\rangle + \|\mathbf{x}\|_2^2 - \|\bar{\mathbf{x}}(z+1)\|_2^2 \\ \geq & 2\langle \mathbf{x}_j(z), \bar{\mathbf{x}}(z+1) - \mathbf{x}\rangle + \|\mathbf{x}\|_2^2 - \|\bar{\mathbf{x}}(z+1)\|_2^2.\end{aligned}$$

Combining the above two inequalities, for any $z \in [T/L]$, $t \in \mathcal{T}_z$, $j \in [n]$, and $\mathbf{x} \in \mathcal{K}$, we have

$$\begin{aligned}& f_{t,j}(\mathbf{x}_i(z)) - f_{t,j}(\mathbf{x}) \\ \leq & \langle \nabla f_{t,j}(\mathbf{x}_j(z)), \bar{\mathbf{x}}(z+1) - \mathbf{x}\rangle - \frac{\alpha}{2}\left(2\langle \mathbf{x}_j(z), \bar{\mathbf{x}}(z+1) - \mathbf{x}\rangle + \|\mathbf{x}\|_2^2 - \|\bar{\mathbf{x}}(z+1)\|_2^2\right) + 3G\epsilon_z \\ = & \langle \nabla f_{t,j}(\mathbf{x}_j(z)) - \alpha\mathbf{x}_j(z), \bar{\mathbf{x}}(z+1) - \mathbf{x}\rangle + \frac{\alpha}{2}\left(\|\bar{\mathbf{x}}(z+1)\|_2^2 - \|\mathbf{x}\|_2^2\right) + 3G\epsilon_z.\end{aligned}$$

From the above inequality, it is not hard to verify that

$$\begin{aligned}& \sum_{z=1}^{T/L}\sum_{t\in\mathcal{T}_z}\sum_{j=1}^{n} f_{t,j}(\mathbf{x}_i(z)) - \sum_{z=1}^{T/L}\sum_{t\in\mathcal{T}_z}\sum_{j=1}^{n} f_{t,j}(\mathbf{x}) \\ \leq & \sum_{z=1}^{T/L}\sum_{t\in\mathcal{T}_z}\sum_{j=1}^{n}\left(\langle \nabla f_{t,j}(\mathbf{x}_j(z)) - \alpha\mathbf{x}_j(z), \bar{\mathbf{x}}(z+1) - \mathbf{x}\rangle + \frac{\alpha}{2}\left(\|\bar{\mathbf{x}}(z+1)\|_2^2 - \|\mathbf{x}\|_2^2\right) + 3G\epsilon_z\right) \\ = & n\sum_{z=1}^{T/L}\left(\langle \bar{\mathbf{d}}(z), \bar{\mathbf{x}}(z+1) - \mathbf{x}\rangle + \frac{L\alpha}{2}(\|\bar{\mathbf{x}}(z+1)\|_2^2 - \|\mathbf{x}\|_2^2)\right) + 3nLG\sum_{z=1}^{T/L}\epsilon_z \\ \overset{(32)}{\leq} & 3nLG\sum_{z=1}^{T/L}\epsilon_z.\end{aligned}$$

Finally, combining the above inequality with the definition of ϵ_z, we have

$$\begin{aligned}\sum_{z=1}^{T/L}\sum_{t\in\mathcal{T}_z}\sum_{j=1}^{n} f_{t,j}(\mathbf{x}_i(z)) - \sum_{z=1}^{T/L}\sum_{t\in\mathcal{T}_z}\sum_{j=1}^{n} f_{t,j}(\mathbf{x}) \leq & 6nLGR + \frac{3nLG(7G+11\alpha R)}{\alpha}\sum_{z=1}^{T/L-2}\frac{1}{z} \\ \leq & \frac{3nLG(7G+13\alpha R)(\ln(T/L)+1)}{\alpha}\end{aligned}$$

where the last inequality is due to $\sum_{z=1}^{T/L-2}\frac{1}{z} \leq \ln(T/L)+1$.

Appendix C. Proof of Theorem 3

Recall that Wan et al. (2022) have established an $\Omega(n\sqrt{T})$ lower bound by extending the classical randomized lower bound for OCO (Abernethy et al., 2008) to D-OCO. The main limitation of their analysis is that they ignore the topology of the graph $\mathcal{G}$ and the spectral properties of the matrix P. To address this limitation, our main idea is to refine their analysis by carefully choosing $\mathcal{G}$ and P.

Specifically, let $A \in \mathbb{R}^{n\times n}$ denote the adjacency matrix of $\mathcal{G}$, and let δ_i denote the degree of node i. As presented in (8) of Duchi et al. (2011), for any connected undirected graph, there exists a specific choice of the communication matrix P satisfying Assumption 1, i.e.,

$$P = I_n - \frac{1}{\delta_{\max}+1}(D-A) \tag{40}$$

where I_n is the identity matrix, $\delta_{\max} = \max\{\delta_1,\ldots,\delta_n\}$, and $D = \mathrm{diag}\{\delta_1,\ldots,\delta_n\}$. Moreover, Duchi et al. (2011) have discussed the connection of the spectral gap $1-\sigma_2(P)$ and the network size n for several classes of interesting networks. Here, we need to utilize the 1-connected cycle graph, i.e., constructing the graph $\mathcal{G}$ by placing the n nodes on a circle and only connecting each node to one neighbor on its right and left. We can derive the following lemma for the 1-connected cycle graph.

Lemma 6 *For the 1-connected cycle graph with $n = 2(m+1)$ where m denotes a positive integer, the communication matrix defined in (40) satisfies*

$$\frac{\pi^2}{1-\sigma_2(P)} \leq 4n^2.$$

Then, we only need to derive a lower bound of $\Omega(n\sqrt{nT})$ because combining it with Lemma 6 will complete this proof. To this end, we set

$$f_{t,n-\lceil m/2\rceil+2}(\mathbf{x}) = \cdots = f_{t,n}(\mathbf{x}) = f_{t,1}(\mathbf{x}) = f_{t,2}(\mathbf{x}) = \cdots = f_{t,\lceil m/2\rceil}(\mathbf{x}) = 0$$

and carefully choose other local functions $f_{t,\lceil m/2\rceil+1}(\mathbf{x}),\ldots,f_{t,n-\lceil m/2\rceil+1}(\mathbf{x})$. According to the topology of the 1-connected cycle graph, it is easy to verify that the learner 1 cannot receive the information generated by learners $\lceil m/2\rceil+1,\ldots,n-\lceil m/2\rceil+1$ at round t unless there exist $\lceil m/2\rceil$ communication rounds since round t.

Let $K = \lceil m/2\rceil$, $Z = \lfloor (T-1)/K\rfloor$, $c_{Z+1} = T$, and $c_i = iK$ for $i = 0,\ldots,Z$. The total T rounds can be divided into the following $Z+1$ intervals

$$[c_0+1, c_1], [c_1+1, c_2], \ldots, [c_Z+1, c_{Z+1}].$$

To maximize the impact of the communication and the topology on the regret of learner 1, for any $i \in \{0,\dots,Z\}$ and $t \in [c_i+1, c_{i+1}]$, we will set $f_{t,\lceil m/2\rceil+1}(\mathbf{x}) = \cdots = f_{t,n-\lceil m/2\rceil+1}(\mathbf{x}) = h_i(\mathbf{x})$. In this way, the global loss function can be written as

$$f_t(\mathbf{x}) = (n-2K+1)h_i(\mathbf{x}) \tag{41}$$

for any $i \in \{0,\dots,Z\}$ and $t \in [c_i+1, c_{i+1}]$. Moreover, according to the above discussion, the decisions $\mathbf{x}_1(c_i+1),\dots,\mathbf{x}_1(c_{i+1})$ for any $i \in \{0,\dots,Z\}$ are made before the function $h_i(\mathbf{x})$ can be revealed to learner 1. As a result, we can utilize the classical randomized strategy to select $h_i(\mathbf{x})$ for any $i \in \{0,\dots,Z\}$, and derive an expected lower bound for $R_{T,1}$.

To be precise, we independently select $h_i(\mathbf{x}) = \langle \mathbf{w}_i, \mathbf{x}\rangle$ for any $i \in \{0,\dots,Z\}$, where the coordinates of $\mathbf{w}_i$ are $\pm G/\sqrt{d}$ with probability $1/2$ and $h_i(\mathbf{x})$ satisfies Assumption 2. It is not hard to verify that

$$\begin{aligned}
&\mathbb{E}_{\mathbf{w}_0,\dots,\mathbf{w}_Z}[R_{T,1}]\\
\overset{(41)}{=}&\mathbb{E}_{\mathbf{w}_0,\dots,\mathbf{w}_Z}\left[\sum_{i=0}^{Z}\sum_{t=c_i+1}^{c_{i+1}}(n-2K+1)h_i(\mathbf{x}_1(t)) - \min_{\mathbf{x}\in\mathcal{K}}\sum_{i=0}^{Z}\sum_{t=c_i+1}^{c_{i+1}}(n-2K+1)h_i(\mathbf{x})\right]\\
=&(n-2K+1)\mathbb{E}_{\mathbf{w}_0,\dots,\mathbf{w}_Z}\left[\sum_{i=0}^{Z}\sum_{t=c_i+1}^{c_{i+1}}\langle \mathbf{w}_i, \mathbf{x}_1(t)\rangle - \min_{\mathbf{x}\in\mathcal{K}}\sum_{i=0}^{Z}(c_{i+1}-c_i)\langle \mathbf{w}_i, \mathbf{x}\rangle\right]\\
=&-(n-2K+1)\mathbb{E}_{\mathbf{w}_0,\dots,\mathbf{w}_Z}\left[\min_{\mathbf{x}\in\mathcal{K}}\sum_{i=0}^{Z}(c_{i+1}-c_i)\langle \mathbf{w}_i, \mathbf{x}\rangle\right]\\
=&-(n-2K+1)\mathbb{E}_{\mathbf{w}_0,\dots,\mathbf{w}_Z}\left[\min_{\mathbf{x}\in\left\{-R/\sqrt{d},R/\sqrt{d}\right\}^d}\left\langle \mathbf{x}, \sum_{i=0}^{Z}(c_{i+1}-c_i)\mathbf{w}_i\right\rangle\right]
\end{aligned} \tag{42}$$

where the third equality is due to $\mathbb{E}_{\mathbf{w}_0,\dots,\mathbf{w}_Z}[\mathbf{w}_i^\top \mathbf{x}_1(t)] = 0$ for any $t \in [c_i+1, c_{i+1}]$, and the last equality is because a linear function is minimized at the vertices of the cube.

Then, let $\epsilon_{01},\dots,\epsilon_{0d},\dots,\epsilon_{Z1},\dots,\epsilon_{Zd}$ be independent and identically distributed variables with $\Pr(\epsilon_{ij} = \pm 1) = 1/2$ for $i \in \{0,\dots,Z\}$ and $j \in \{1,\dots,d\}$. Combining with (42), we further have

$$\begin{aligned}
\mathbb{E}_{\mathbf{w}_0,\dots,\mathbf{w}_Z}[R_{T,1}] =&-(n-2K+1)\mathbb{E}_{\epsilon_{01},\dots,\epsilon_{Zd}}\left[\sum_{j=1}^{d}-\frac{R}{\sqrt{d}}\left|\sum_{i=0}^{Z}(c_{i+1}-c_i)\frac{\epsilon_{ij}G}{\sqrt{d}}\right|\right]\\
=&(n-2K+1)RG\mathbb{E}_{\epsilon_{01},\dots,\epsilon_{Z1}}\left[\left|\sum_{i=0}^{Z}(c_{i+1}-c_i)\epsilon_{i1}\right|\right]\\
\geq&\frac{(n-2K+1)RG}{\sqrt{2}}\sqrt{\sum_{i=0}^{Z}(c_{i+1}-c_i)^2}\\
\geq&\frac{(n-2K+1)RG}{\sqrt{2}}\sqrt{\frac{(c_{Z+1}-c_0)^2}{Z+1}} = \frac{(n-2K+1)RGT}{\sqrt{2(Z+1)}}
\end{aligned} \tag{43}$$

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

Note that the expected lower bound in (43) implies that for any D-OCO algorithm, there exists a particular choice of $\mathbf{w}_0, \ldots, \mathbf{w}_Z$ such that

$$R_{T,1} \geq \frac{(n-2K+1)RGT}{\sqrt{2(Z+1)}} \geq \frac{n\sqrt{2n}RGT}{4\sqrt{8T-8+n}}$$

where the last inequality is due to

$$\begin{aligned}\frac{n-2K+1}{\sqrt{Z+1}} =& \frac{n-2\lceil m/2\rceil+1}{\sqrt{\lfloor (T-1)/\lceil m/2\rceil\rfloor+1}} \geq \frac{n-m-1}{\sqrt{(T-1)/\lceil m/2\rceil+1}} \\ =& \frac{(m+1)\sqrt{m+1}}{\sqrt{((T-1)/\lceil m/2\rceil+1)(m+1)}} \\ \geq& \frac{(m+1)\sqrt{m+1}}{\sqrt{4T-4+m+1}} = \frac{n\sqrt{n}}{2\sqrt{8T-8+n}}.\end{aligned} \tag{44}$$

If $n \leq 8T+8$, combining the above result on $R_{T,1}$ with Lemma 6, we have

$$R_{T,1} \geq \frac{n\sqrt{\pi}RG\sqrt{T}}{16(1-\sigma_2(P))^{1/4}}.$$

Otherwise, we have $8T-8+n < 2n$, and thus $R_{T,1} \geq \frac{nRGT}{4}$.

Appendix D. Proof of Lemma 6

We start this proof by introducing a general lemma regarding the spectral gap of the communication matrix P defined in (40).

Lemma 7 *(Lemma 4 of Duchi et al. (2011)) Let δ_i denote the degree of each node i in a connected undirected graph $\mathcal{G}$. For the graph $\mathcal{G}$, the matrix P defined in (40) satisfies*

$$\sigma_2(P) \leq \max\left\{1 - \frac{\delta_{\min}}{\delta_{\max}+1}\lambda_{n-1}(\mathcal{L}), \frac{\delta_{\max}}{\delta_{\max}+1}\lambda_1(\mathcal{L}) - 1\right\}$$

where $\delta_{\min} = \min\{\delta_1, \ldots, \delta_n\}$, $\delta_{\max} = \max\{\delta_1, \ldots, \delta_n\}$, $\mathcal{L}$ denotes the normalized graph Laplacian of $\mathcal{G}$, and $\lambda_i(\mathcal{L})$ denotes the i-th largest real eigenvalue of $\mathcal{L}$.

Moreover, for the 1-connected cycle graph, Duchi et al. (2011) have proved that $\mathcal{L}$ has the following eigenvalues

$$\left\{1 - \cos\left(\frac{2\pi i}{n}\right) \middle| i = 1, \cdots, n\right\}.$$

Therefore, because of $n = 2(m+1)$, it is easy to verify that

$$\lambda_1(\mathcal{L}) = 1 - \cos\left(\frac{2(m+1)\pi}{n}\right) = 1 - \cos(\pi) = 2.$$

Moreover, because of $n = 2(m+1)$ and $\cos(x) = \cos(2\pi - x)$ for any x, we have

$$\lambda_{n-1}(\mathcal{L}) = \min\left\{1 - \cos\left(\frac{2\pi}{n}\right), 1 - \cos\left(\frac{2\pi(n-1)}{n}\right)\right\} = 1 - \cos\left(\frac{\pi}{m+1}\right) \geq \frac{\pi^2}{4(m+1)^2}.$$

Since the 1-connected cycle graph also satisfies that $\delta_{\max} = \delta_{\min} = 2$, by using Lemma 7, we have

$$\sigma_2(P) \leq \max\left\{1 - \frac{2}{3}\lambda_{n-1}(\mathcal{L}), \frac{1}{3}\right\} = 1 - \frac{2}{3}\lambda_{n-1}(\mathcal{L}) \leq 1 - \frac{\pi^2}{6(m+1)^2}$$

where the equality is due to $\lambda_{n-1}(\mathcal{L}) \leq 1 - \cos(\pi/2) \leq 1$. Finally, it is easy to verify that

$$\frac{\pi^2}{1 - \sigma_2(P)} \leq 6(m+1)^2 \leq 4n^2.$$

Appendix E. Proof of Theorem 4

This proof is similar to the proof of Theorem 3. The main modification is to make the previous local functions α-strongly convex by adding a term $\frac{\alpha}{2}\|\mathbf{x}\|_2^2$.

To be precise, let $K = \lceil m/2 \rceil$, $Z = \lfloor (T-1)/K \rfloor$, $c_{Z+1} = T$, and $c_i = iK$ for $i = 0, \ldots, Z$. We still divide the total T rounds into the following $Z+1$ intervals

$$[c_0 + 1, c_1], [c_1 + 1, c_2], \ldots, [c_Z + 1, c_{Z+1}].$$

At each round t, we simply set

$$f_{t,n-\lceil m/2 \rceil+2}(\mathbf{x}) = \cdots = f_{t,n}(\mathbf{x}) = f_{t,1}(\mathbf{x}) = f_{t,2}(\mathbf{x}) = \cdots = f_{t,\lceil m/2 \rceil}(\mathbf{x}) = \frac{\alpha}{2}\|\mathbf{x}\|_2^2$$

which satisfies Assumption 2 with $G = 2\alpha R$ and the definition of α-strongly convex functions. Moreover, for any $i \in \{0, \ldots, Z\}$ and $t \in [c_i + 1, c_{i+1}]$, we set

$$f_{t,\lceil m/2 \rceil+1}(\mathbf{x}) = \cdots = f_{t,n-\lceil m/2 \rceil+1}(\mathbf{x}) = h_i(\mathbf{x}) = \langle \mathbf{w}_i, \mathbf{x} \rangle + \frac{\alpha}{2}\|\mathbf{x}\|_2^2$$

where the coordinates of $\mathbf{w}_i$ are $\pm\alpha R/\sqrt{d}$ with probability $1/2$. Note that $h_i(\mathbf{x})$ satisfies Assumption 2 with $G = 2\alpha R$ and Assumption 4. Following the proof of Theorem 3, we set $\mathcal{G}$ as the 1-connected cycle graph, which ensures that the decisions $\mathbf{x}_1(c_i+1), \ldots, \mathbf{x}_1(c_{i+1})$ are independent of $\mathbf{w}_i$.

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

$$\begin{aligned}
\sum_{t=1}^{T} f_t(\mathbf{x}) =& \sum_{i=0}^{Z}(c_{i+1} - c_i)\left((n - 2K + 1)\langle \mathbf{w}_i, \mathbf{x} \rangle + \frac{\alpha n}{2}\|\mathbf{x}\|_2^2\right) \\
=& \alpha(n - 2K + 1)T\langle \bar{\mathbf{w}}, \mathbf{x} \rangle + \frac{\alpha n T}{2}\|\mathbf{x}\|_2^2 \\
=& \frac{\alpha T}{2}\left(\left\|\sqrt{n}\mathbf{x} + \frac{(n - 2K + 1)}{\sqrt{n}}\bar{\mathbf{w}}\right\|_2^2 - \left\|\frac{(n - 2K + 1)}{\sqrt{n}}\bar{\mathbf{w}}\right\|_2^2\right).
\end{aligned} \tag{45}$$

According to the definition of $\mathbf{w}_i$, the absolute value of each element in $-\frac{n-2K+1}{n}\bar{\mathbf{w}}$ is bounded by

$$\frac{n - 2K + 1}{n\alpha T}\sum_{i=0}^{Z}\frac{(c_{i+1} - c_i)\alpha R}{\sqrt{d}} = \frac{(n - 2K + 1)R}{n\sqrt{d}} \leq \frac{R}{\sqrt{d}}$$

which implies that $-\frac{n-2K+1}{n}\bar{\mathbf{w}}$ belongs to $\mathcal{K} = [-R/\sqrt{d}, R/\sqrt{d}]^d$.

By further combining with (45), we have

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

As a result, it is not hard to verify that

$$\begin{aligned}
&\mathbb{E}_{\mathbf{w}_0,\ldots,\mathbf{w}_Z}\left[R_{T,1}\right]\\
=&\mathbb{E}_{\mathbf{w}_0,\ldots,\mathbf{w}_Z}\left[\sum_{i=0}^{Z}\sum_{t=c_i+1}^{c_{i+1}}\left((n-2K+1)\langle\mathbf{w}_i,\mathbf{x}_1(t)\rangle+\frac{\alpha n}{2}\|\mathbf{x}_1(t)\|_2^2\right)+\frac{\alpha T}{2}\left\|\frac{(n-2K+1)}{\sqrt{n}}\bar{\mathbf{w}}\right\|_2^2\right]\\
\geq&\mathbb{E}_{\mathbf{w}_0,\ldots,\mathbf{w}_Z}\left[\sum_{i=0}^{Z}\sum_{t=c_i+1}^{c_{i+1}}(n-2K+1)\langle\mathbf{w}_i,\mathbf{x}_1(t)\rangle+\frac{\alpha(n-2K+1)^2T}{2n}\|\bar{\mathbf{w}}\|_2^2\right]\\
=&\mathbb{E}_{\mathbf{w}_0,\ldots,\mathbf{w}_Z}\left[\frac{\alpha(n-2K+1)^2T}{2n}\|\bar{\mathbf{w}}\|_2^2\right]
\end{aligned}$$

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

Next, let $\epsilon_{01},\ldots,\epsilon_{0d},\ldots,\epsilon_{Z1},\ldots,\epsilon_{Zd}$ be independent and identically distributed variables with $\Pr(\epsilon_{ij}=\pm1)=1/2$ for $i\in\{0,\ldots,Z\}$ and $j\in\{1,\ldots,d\}$. Combining the definition of $\bar{\mathbf{w}}$ with the above equality, we further have

$$\begin{aligned}
\mathbb{E}_{\mathbf{w}_0,\ldots,\mathbf{w}_Z}\left[R_{T,1}\right]=&\frac{(n-2K+1)^2}{2\alpha nT}\mathbb{E}_{\mathbf{w}_0,\ldots,\mathbf{w}_Z}\left[\left\|\sum_{i=0}^{Z}(c_{i+1}-c_i)\mathbf{w}_i\right\|_2^2\right]\\
=&\frac{(n-2K+1)^2}{2\alpha nT}\mathbb{E}_{\epsilon_{01},\ldots,\epsilon_{Zd}}\left[\sum_{j=1}^{d}\left|\sum_{i=0}^{Z}(c_{i+1}-c_i)\frac{\epsilon_{ij}\alpha R}{\sqrt{d}}\right|^2\right]\\
=&\frac{\alpha(n-2K+1)^2R^2}{2nT}\mathbb{E}_{\epsilon_{01},\ldots,\epsilon_{Z1}}\left[\left|\sum_{i=0}^{Z}(c_{i+1}-c_i)\epsilon_{i1}\right|^2\right]\\
=&\frac{\alpha(n-2K+1)^2R^2}{2nT}\sum_{i=0}^{Z}(c_{i+1}-c_i)^2\geq\frac{\alpha(n-2K+1)^2R^2T}{2n(Z+1)}
\end{aligned}\tag{46}$$

where the inequality is due to the Cauchy-Schwarz inequality and $(c_{Z+1}-c_0)^2=T^2$.

The expected lower bound in (46) implies that for any D-OCO algorithm, there exists a particular choice of $\mathbf{w}_0,\ldots,\mathbf{w}_Z$ such that

$$R_{T,1}\geq\frac{\alpha(n-2K+1)^2R^2T}{2n(Z+1)}\overset{(44)}{\geq}\frac{\alpha n^2R^2T}{8(8T-8+n)}.$$

If $n\leq 8T+8$, according to Lemma 6, by using the communication matrix P defined in (40), we have

$$R_{T,1}\geq\frac{\alpha\pi nR^2}{256\sqrt{1-\sigma_2(P)}}.$$

Otherwise, we have $8T-8+n<2n$, and thus $R_{T,1}\geq\frac{\alpha nR^2T}{16}$.

Appendix F. Regret Bound of D-FTAL

Note that Wan et al. (2021) originally develop a projection-free version of D-FTAL for α-strongly convex functions, which also adopts the blocking update mechanism to update the decision of each learner. Following the notations in our Algorithm 2, at the end of each block $z = 1, \dots, T/L$, each learner i of their algorithm updates as

$$\begin{aligned}\mathbf{z}_i(z+1) &= \sum_{j \in N_i} P_{ij}\mathbf{z}_j(z) + \mathbf{d}_i(z) \\ \mathbf{x}_i(z+1) &= \text{CG}(\mathcal{K}, K, F_{z,i}(\mathbf{x}), \mathbf{x}_i(z))\end{aligned} \tag{47}$$

where $F_{z,i}(\mathbf{x}) = \langle \mathbf{z}_i(z+1), \mathbf{x}\rangle + \frac{zL\alpha}{2}\|\mathbf{x}\|_2^2$, and $\mathbf{x}_i(z+1)$ is computed by utilizing the classical conditional gradient (CG) method (Frank and Wolfe, 1956; Jaggi, 2013) with the initialization $\mathbf{x}_i(z)$ and K iterations to minimize the function $F_{z,i}(\mathbf{x})$ over the decision set $\mathcal{K}$.

According to Theorem 1 of Wan et al. (2021), under Assumptions 1, 2, 3, and 4, their algorithm can achieve the following regret bound

$$R_{T,i} \leq \left(\frac{2nL(G+2\alpha R)^2}{\alpha} + \frac{3GL(G+\alpha R)n^{3/2}}{\alpha\rho}\right)(1+\ln(T/L)) + \frac{12nGRT}{\sqrt{K+2}}. \tag{48}$$

By substituting $K = L = T^{2/3}$ into (48), they derive an $O(n^{3/2}\rho^{-1}T^{2/3}\log T)$ regret bound for strongly convex functions. However, this choice of K and L is for achieving the projection-free property, i.e., only one CG iteration based on a linear optimization step is utilized per round on average. It is easy to derive an improved regret bound of $O(n^{3/2}\rho^{-1}\log T)$ for strongly convex functions by substituting $K = \infty$ and $L = 1$ into (48). Moreover, one can verify that by setting $K = \infty$ and $L = 1$, the algorithm of Wan et al. (2021) reduces to D-FTAL in (5) because the second step in (47) now is equivalent to computing $\mathbf{x}_i(z+1) = \operatorname{argmin}_{\mathbf{x}\in\mathcal{K}} F_{z,i}(\mathbf{x})$ exactly via the projection operation.