Journal of Machine Learning Research 26 (2025) 1-43 Submitted 12/24; Published 7/25

Optimal and Efficient Algorithms for Decentralized Online Convex Optimization

Yuanyu Wan WANYY@ZJU.EDU.CN
School of Software Technology, Zhejiang University, Ningbo, China
State Key Laboratory of Blockchain and Data Security, Zhejiang University, Hangzhou, China

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

Bo Xue BOXUE4-C@MY.CITYU.EDU.HK
Department of Computer Science, City University of Hong Kong, Hong Kong, China

Mingli Song BROOKSONG@ZJU.EDU.CN
State Key Laboratory of Blockchain and Data Security, Zhejiang University, Hangzhou, China

Lijun Zhang ZHANGLJ@LAMDA.NJU.EDU.CN
National Key Laboratory for Novel Software Technology, Nanjing University, Nanjing, China
School of Artificial Intelligence, Nanjing University, Nanjing, China

Editor: Lin Xiao

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 a novel D-OCO algorithm 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 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 $\Omega(n\rho^{-1/2}\log T)$, respectively. These results suggest that the regret of our algorithm is nearly optimal in terms of T, n, and ρ for both convex and strongly convex functions. Finally, we propose a projection-free variant of our algorithm to efficiently handle practical applications with complex constraints. Our analysis reveals that the projection-free variant can achieve $O(nT^{3/4})$ and $O(nT^{2/3}(\log T)^{1/3})$ regret bounds for convex and strongly convex functions with nearly optimal $\tilde{O}(\rho^{-1/2}\sqrt{T})$ and $\tilde{O}(\rho^{-1/2}T^{1/3}(\log T)^{2/3})$ communication rounds, respectively.

Keywords: online convex optimization, decentralized optimization, optimal regret, accelerated gossip strategy, efficient algorithms

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.,

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

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 various scenarios (Hosseini et al., 2013; Zhang et al., 2017; Lei et al., 2020; Wan et al., 2020, 2021a, 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 a novel D-OCO algorithm that can 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, respectively.[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

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.

$f_{t,i}(\cdot)$	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	$O(n\rho^{-1/4}\sqrt{T}\log n)$ Corollary 1	$\Omega(n\rho^{-1/4}\sqrt{T})$ Theorem 2	$O(\sqrt{\log n})$
scvx	Previous studies	$O(n^{3/2}\rho^{-1}\log T)$ Yan et al. (2013)	$\Omega(n)$ Wan et al. (2022)	$O(n^{1/2}\rho^{-1}\log T)$
	This work	$O(n\rho^{-1/2}(\log n)\log T)$ Corollary 2	$\Omega(n\rho^{-1/2}\log T)$ Theorem 4	$O(\log n)$

Table 1: Summary of our results and the best previous results on the optimality of D-OCO. Abbreviations: convex $\rightarrow$ cvx, strongly convex $\rightarrow$ scvx.

$f_{t,i}(\cdot)$	Source	Regret	Communication Rounds
cvx	Wan et al. (2022)	$O(n^{5/4}\rho^{-1/2}T^{3/4})$	$O(\sqrt{T})$
	Corollary 3	$O(n\rho^{-1/4}T^{3/4}\sqrt{\log n})$	$O(\sqrt{T})$
		$O(nT^{3/4})$	$O(\rho^{-1/2}\sqrt{T}\log n)$
scvx	Wan et al. (2022)	$O(n^{3/2}\rho^{-1}T^{2/3}(\log T)^{1/3})$	$O(T^{1/3}(\log T)^{2/3})$
	Corollary 4	$O(n\rho^{-1/2}T^{2/3}(\log T)^{1/3}\log n)$	$O(T^{1/3}(\log T)^{2/3})$
		$O(nT^{2/3}(\log T)^{1/3})$	$O(\rho^{-1/2}T^{1/3}(\log T)^{2/3}\log n)$

Table 2: Summary of our projection-free algorithm and the best existing projection-free algorithm for D-OCO. Abbreviations: convex $\rightarrow$ cvx, strongly convex $\rightarrow$ scvx.

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 $\Omega(n\rho^{-1/2}\log T)$ for convex and strongly convex functions, respectively. Compared with existing lower bounds (Wan et al., 2022), our bounds additionally uncover the effect of the spectral gap, by carefully exploiting spectral properties of a specific network topology. Table 1 provides a comparison of our results on the optimality of D-OCO with those of previous studies.

Finally, we notice that each learner of our algorithm needs to perform a projection operation per round to ensure the feasibility of its decision, which could be computationally expensive in applications with complex constraints. For example, in online collaborative filtering (Hazan and Kale, 2012), the decision set consists of all matrices with bounded trace norm, and the corresponding projection needs to compute singular value decomposition of a matrix. To tackle this computational bottleneck, we propose a projection-free variant of our algorithm by replacing the projection operation with a more efficient linear optimization step. Analysis reveals that our projection-free algorithm can achieve a regret bound of $O(nT^{3/4})$ with only $\tilde{O}(\rho^{-1/2}\sqrt{T})$ communication rounds for convex functions, and a better regret bound of $O(nT^{2/3}(\log T)^{1/3})$ with fewer $\tilde{O}(\rho^{-1/2}T^{1/3}(\log T)^{2/3})$ communication rounds for strongly convex functions, respectively. In contrast, the state-of-the-art projection-free D-OCO algorithm (Wan et al., 2022) only achieves worse $O(n^{5/4}\rho^{-1/2}T^{3/4})$ and

$O(n^{3/2}\rho^{-1}T^{2/3}(\log T)^{1/3})$ regret bounds for convex and strongly convex functions respectively, though the number of required communication rounds is less, i.e., $O(\sqrt{T})$ for convex functions and $O(T^{1/3}(\log T)^{2/3})$ for strongly convex functions. Moreover, even if using the same number of communication rounds as Wan et al. (2022), we show that the regret bounds of our projection-free algorithm only degenerate to $\tilde{O}(n\rho^{1/4}T^{3/4})$ and $\tilde{O}(n\rho^{1/2}T^{2/3}(\log T)^{1/3})$, which are still tighter than those of Wan et al. (2022), respectively. It is worth noting that although these regret bounds of our projection-free algorithm no longer match the aforementioned lower bounds, we further extend the latter to demonstrate that the number of required communication rounds is nearly optimal for achieving these regret bounds. Table 2 provides a comparison of our projection-free algorithm and that of Wan et al. (2022).

A preliminary version of this paper was presented at the 37th Annual Conference on Learning Theory in 2024 (Wan et al., 2024). In this paper, we have significantly enriched the preliminary version by adding the following extensions.

- Different from Wan et al. (2024) that propose two algorithms to deal with convex and strongly convex functions respectively, we unify them into a single one, and provide a unified analysis to recover their regret bounds.
- For D-OCO with strongly convex functions, we improve the $\Omega(n\rho^{-1/2})$ lower bound in Wan et al. (2024) to $\Omega(n\rho^{-1/2}\log T)$, which now can recover the classical $\Omega(\log T)$ lower bound for OCO with strongly convex functions (Abernethy et al., 2008; Hazan and Kale, 2014).
- We propose a projection-free variant of our algorithm to efficiently handle complex constraints. It improves the regret of the state-of-the-art projection-free D-OCO algorithm (Wan et al., 2022) while maintaining the same number of communication rounds.
- We establish $\Omega(n\rho^{-1/4}T/\sqrt{C})$ and $\Omega(n\rho^{-1/2}T/C)$ lower bounds for convex and strongly convex functions in a more challenging setting with only C communication rounds, which imply that the communication complexity of our projection-free variant is nearly optimal.
- We demonstrate that the analysis of some existing algorithms (Hosseini et al., 2013; Wan et al., 2022) can be refined to reduce the dependence of their current regret bounds on n to only $\tilde{O}(n)$. Although the refined regret is still not optimal, it may be of independent interest.

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; Hazan and Kale, 2014). 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 tuning η appropriately, 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, Hazan et al. (2007) have proposed a variant of (2) for α-strongly convex functions, which makes the following update

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

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

Nonetheless, as first noticed by Hazan and Kale (2012), these optimal algorithms explicitly or implicitly require a projection operation per round, which could become a computational bottleneck for complex decision sets. To address this issue, Hazan and Kale (2012) propose a projection-free variant of FTRL, and establish an $O(T^{3/4})$ regret bound for convex functions. Their key idea is to approximately solve (2) with only one linear optimization step, which could be much more efficiently carried out than the projection operation. Due to the efficiency, there has been a growing research interest in developing projection-free OCO algorithms (Garber and Hazan, 2016; Levy and Krause, 2019; Hazan and Minasyan, 2020; Garber and Kretzu, 2020, 2021; Wan and Zhang, 2021; Wan et al., 2021b; Garber and Kretzu, 2022, 2023; Wan et al., 2023; Zhang et al., 2024; Wang et al., 2025). Among these works, the most related one to this paper is Wan and Zhang (2021), in which a projection-free variant of FTAL is proposed to achieve an improved regret bound of $O(T^{2/3})$ for strongly convex functions.

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., $\mathrm{R}(T, i)$ 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

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

for each learner i, 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. Note that the cumulative gradients $\sum_{i=1}^{t} \nabla f_i(\mathbf{x}(i))$ used 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 $\nabla f_{t,i}(\mathbf{x}_i(t))$. For convex functions, D-FTRL can also achieve the $O(n^{5/4}\rho^{-1/2}\sqrt{T})$ regret bound.

The first projection-free algorithm for D-OCO is proposed by Zhang et al. (2017), and can be viewed as a combination of D-FTRL and linear optimization steps. For convex functions, analogous to the projection-free variant of FTRL (Hazan and Kale, 2012) in OCO, this algorithm can achieve an $O(n^{5/4}\rho^{-1/2}T^{3/4})$ regret bound. After that, several improvements have been made to projection-free D-OCO (Wan et al., 2020, 2021a, 2022; Wang et al., 2023). First, Wan et al. (2020) demonstrate that by combining the projection-free algorithm in Zhang et al. (2017) with a blocking update mechanism, the number of communication rounds can be reduced from $O(T)$ to $O(\sqrt{T})$ while achieving the same $O(n^{5/4}\rho^{-1/2}T^{3/4})$ regret bound for convex functions. If the functions are strongly convex, Wan et al. (2021a) propose a projection-free and decentralized variant of FTAL (Hazan et al., 2007), which can enjoy an improved regret bound of $O(n^{3/2}\rho^{-1}T^{2/3}(\log T)^{1/3})$ with even fewer $O(T^{1/3}(\log T)^{2/3})$ communication rounds. Then, Wan et al. (2022) unify these two algorithms into a single one that inherits the theoretical guarantees for both convex and strongly convex functions. Moreover, they also provide $\Omega(nT/\sqrt{C})$ and $\Omega(nT/C)$ lower regret bounds for any D-OCO algorithm with C communication rounds by respectively considering convex and strongly convex functions. These lower bounds imply that the number of communication rounds required by their algorithm is (nearly) optimal in terms of T for achieving their current regret bounds. Very recently, Wang et al. (2023) develop a randomized projection-free algorithm for D-OCO with smooth functions, and achieve an expected regret bound of $O(n^{5/4}\rho^{-1/2}T^{2/3})$ with $O(T^{2/3})$ communication rounds.

Additionally, we notice that if the projection operation is allowed, the unified algorithm in Wan et al. (2022) can be simplified as performing the following update

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

for each learner i, where α and h are two parameters. By setting $\alpha = 0$ and $h = 1/\eta$, (5) reduces to D-FTRL and thus can also enjoy the $O(n^{5/4}\rho^{-1/2}\sqrt{T})$ regret bound for convex functions. For α-strongly convex functions, by simply setting $h = 0$, (5) becomes a decentralized variant of FTAL because the local variable $\mathbf{z}_i(t+1)$ now is used to replace the cumulative information $\sum_{i=1}^{t}(\nabla f_i(\mathbf{x}(i)) - \alpha\mathbf{x}(i))$ in (3). This subtle difference allows the algorithm to recover the $O(n^{3/2}\rho^{-1}\log T)$ regret bound for strongly convex functions (see Appendix A for details). Therefore, (5) can be referred to as decentralized follow-the-generalized-leader (D-FTGL). Moreover, by setting $C = O(T)$, the communication-dependent lower bounds in Wan et al. (2022) reduce to $\Omega(n\sqrt{T})$ and $\Omega(n)$ lower bounds for general D-OCO with convex and strongly convex functions, respectively. These results imply that the existing $O(n^{5/4}\rho^{-1/2}\sqrt{T})$ and $O(n^{3/2}\rho^{-1}\log T)$ upper bounds are (nearly) tight in terms of T. However, there still exist gaps in terms of n and ρ between these upper and lower bounds. Note that the value of ρ reflects the connectivity of the network—a larger ρ implies better connectivity, and it 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 in n.

2.3 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; Ye and Chang, 2023; Ye et al., 2023; Song et al., 2024). Among these studies, the one closest 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 $\mathrm{R}(T,i)$ to achieve an approximation error of $O(\mathrm{R}(T,i)/(nT))$ for decentralized offline optimization, but not vice versa. Moreover, one may notice that due to the online-to-batch conversion, it is possible to use 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 reduce to the $\Omega(n\sqrt{T})$ and $\Omega(n)$ lower bounds 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 used in these previous studies on decentralized offline and stochastic optimization, this paper is the first work to apply it in D-OCO.

3. Preliminaries

In this section, we introduce the necessary preliminaries including common assumptions and an algorithmic ingredient. Specifically, similar to previous studies on D-OCO (Yan et al., 2013; Hosseini et al., 2013), we 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., $|f_{t,i}(\mathbf{x}) - f_{t,i}(\mathbf{y})| \leq G\|\mathbf{x}-\mathbf{y}\|_2$ for any $\mathbf{x},\mathbf{y} \in \mathcal{K}$.*

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., $\|\mathbf{x}\|_2 \leq R$ for any $\mathbf{x} \in \mathcal{K}$.*

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

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 used 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 $\bar{\nabla} = \frac{1}{n}\sum_{i=1}^n \nabla_i$ 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 $X^k = \left[(\nabla_1^k)^\top; \ldots; (\nabla_n^k)^\top\right] \in \mathbb{R}^{n\times d}$ for any integer $k \geq -1$. For any integer $k \geq 0$, it is not hard to verify that (7) ensures

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

This process enjoys the following convergence property, where $\bar{X} = \left[\bar{\nabla}^\top; \ldots; \bar{\nabla}^\top\right] \in \mathbb{R}^{n\times d}$.

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

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

Remark 1 As in Lemma 1, our Assumption 1 on the communication matrix P is required mainly for an easy application of the existing result in Ye et al. (2023). Actually, according to the original analysis of Liu and Morse (2011) for the process in (8), P might not need to be symmetric and positive semidefinite. However, they only demonstrate the accelerated convergence by bounding the second largest eigenvalue of an augmented communication matrix $\tilde{P}$ regarding (8). To establish a detailed convergence rate via this result, a complicated analysis similar to Ye et al. (2023) is still required. Thus, we leave the extension to the case without the symmetric and positive semidefinite assumption as a future work.

4. Main Results

In this section, we first present a novel algorithm with improved regret bounds for D-OCO, and establish nearly matching lower bounds. Then, we develop a projection-free variant of our algorithm to efficiently handle complex constraints.

4.1 A Novel Algorithm with Improved Regret Bounds

Before introducing our algorithms, we first compare the regret of D-OCO and OCO, which provides insights into our improvements. Specifically, compared with the $O(\sqrt{T})$ 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 $n^{5/4}\rho^{-1/2}$. We notice 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.,

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

By minimizing the bound in (10) with $\eta = \Theta(\sqrt{\rho/(\sqrt{n}T)})$, 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. Moreover, to handle convex and strongly convex functions with a unified algorithm, there are two possible options: refining D-OGD or D-FTGL, i.e., the generalized variant of D-FTRL in (5). However, the projection operation in D-OGD makes the analysis of the approximation error more complex. To this end, we propose to improve D-FTGL via the accelerated gossip strategy in (7). Let $\mathbf{d}_i(t) = \nabla f_{t,i}(\mathbf{x}_i(t)) - \alpha\mathbf{x}_i(t)$ and $\bar{\mathbf{d}}(t) = \frac{1}{n}\sum_{i=1}^n \mathbf{d}_i(t)$. According to D-FTGL, we now need to maintain a better $\mathbf{z}_i(t)$ to approximate $\bar{\mathbf{z}}(t) = \sum_{\tau=1}^{t-1} \bar{\mathbf{d}}(\tau)$, which is more general than the above definition. A natural idea is to replace the standard gossip step in (5) with multiple accelerated gossip steps, i.e., setting $\mathbf{z}_i(1) = \mathbf{z}_i^{L-1}(1) = \mathbf{0}$ and computing $\mathbf{z}_i(t) = \mathbf{z}_i^L(t)$ for any $t \geq 2$ via the following iterations

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

where $\mathbf{z}_i^0(t) = \mathbf{z}_i(t-1) + \mathbf{d}_i(t-1)$, $\mathbf{z}_i^{-1}(t) = \mathbf{z}_i^{L-1}(t-1) + \mathbf{d}_i(t-1)$. One can prove that (11) ensures (see Equation 23 in Section 5.1 for details)

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

where $\mathbf{d}_i^{(t-\tau)L}(\tau)$ denotes the output of virtually performing (7) with $\nabla_i = \mathbf{d}_i(\tau)$ and $L = (t-\tau)L$. Due to the convergence behavior of the accelerated gossip strategy, we can control the error of approximating $\bar{\mathbf{z}}(t)$ under any desired level by using a large enough L. Unfortunately, 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). To be precise, we divide the total T rounds into T/L blocks, and

Algorithm 1 AD-FTGL

Input: α, h, θ, L
Initialization: set $\mathbf{x}_i(1) = \mathbf{z}_i(1) = \mathbf{z}_i^{L-1}(1) = \mathbf{0}, \forall i \in [n]$
for $z = 1, \dots, T/L$ **do**
  If $2 \leq z$, set $\mathbf{z}_i^0(z) = \mathbf{z}_i(z-1) + \mathbf{d}_i(z-1)$, $\mathbf{z}_i^{-1}(z) = \mathbf{z}_i^{L-1}(z-1) + \mathbf{d}_i(z-1), \forall i \in [n]$
  for $t = (z-1)L+1, \dots, zL$ **do**
    for each local learner $i \in [n]$ **do**
      Play $\mathbf{x}_i(z)$, query $\nabla f_{t,i}(\mathbf{x}_i(z))$, and set $k = t-(z-1)L-1$
      If $2 \leq z$, update $\mathbf{z}_i^{k+1}(z) = (1+\theta)\sum_{j\in N_i} P_{ij}\mathbf{z}_j^k(z) - \theta\mathbf{z}_i^{k-1}(z)$
    end for
  end for
  Set $\mathbf{d}_i(z) = \sum_{t\in\mathcal{T}_z}(\nabla f_{t,i}(\mathbf{x}_i(z)) - \alpha\mathbf{x}_i(z)), \forall i \in [n]$, where $\mathcal{T}_z = \{(z-1)L+1, \dots, zL\}$
  If $2 \leq z$, set $\mathbf{z}_i(z) = \mathbf{z}_i^L(z), \forall i \in [n]$
  Compute $\mathbf{x}_i(z+1) = \operatorname{argmin}_{\mathbf{x}\in\mathcal{K}} \langle \mathbf{z}_i(z), \mathbf{x}\rangle + \frac{(z-1)L\alpha}{2}\|\mathbf{x}\|_2^2 + h\|\mathbf{x}\|_2^2, \forall i \in [n]$
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. Let $\mathcal{T}_z = \{(z-1)L+1, \dots, zL\}$ denote all rounds contained in each block z. With some abuse of notations, we redefine $\mathbf{d}_i(z) = \sum_{t\in\mathcal{T}_z}(\nabla f_{t,i}(\mathbf{x}_i(z)) - \alpha\mathbf{x}_i(z))$ and $\bar{\mathbf{d}}(z) = \frac{1}{n}\sum_{i=1}^n \mathbf{d}_i(z)$ for each block z. In this way, we only need to maintain a local variable $\mathbf{z}_i(z)$ to approximate $\bar{\mathbf{z}}(z) = \sum_{\tau=1}^{z-1}\bar{\mathbf{d}}(\tau)$ for each learner i in block z. The good news is that now L communications can be used to update $\mathbf{z}_i(z)$ per block by uniformly allocating them to every round in the block. As a result, we set $\mathbf{z}_i(1) = \mathbf{z}_i^{L-1}(1) = \mathbf{0}$, and compute $\mathbf{z}_i(z) = \mathbf{z}_i^L(z)$ for any $z \geq 2$ in a way similar to (11), i.e., performing the following iterations during block z

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

where $\mathbf{z}_i^0(z) = \mathbf{z}_i(z-1) + \mathbf{d}_i(z-1)$ and $\mathbf{z}_i^{-1}(z) = \mathbf{z}_i^{L-1}(z-1) + \mathbf{d}_i(z-1)$. Then, inspired by D-FTGL in (5), 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

$$\mathbf{x}_i(z+1) = \operatorname*{argmin}_{\mathbf{x}\in\mathcal{K}} \langle \mathbf{z}_i(z), \mathbf{x}\rangle + \frac{(z-1)L\alpha}{2}\|\mathbf{x}\|_2^2 + h\|\mathbf{x}\|_2^2. \tag{13}$$

We name the proposed algorithm as accelerated decentralized follow-the-generalized-leader (AD-FTGL), 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-FTGL, which demonstrates the advantage of using the accelerated gossip strategy.

Lemma 2 *Let $\bar{\mathbf{z}}(z) = \sum_{\tau=1}^{z-1}\bar{\mathbf{d}}(\tau)$, where $\bar{\mathbf{d}}(\tau) = \frac{1}{n}\sum_{i=1}^n \mathbf{d}_i(\tau)$, 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, 2, 3, and 4, for any $i \in [n]$ and $z \in [T/L]$, Algorithm 1 with θ, L defined in (14) *ensures*

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

From Lemma 2, our AD-FTGL can enjoy an error bound of $O(\rho^{-1/2}\log n)$ for approximating $\bar{\mathbf{z}}(z)$, which is tighter than the $O(\rho^{-1}\sqrt{n})$ error bound in (9). By exploiting this improvement, we establish the following guarantee on the regret bound of AD-FTGL.

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

$$\mathrm{R}(T,i) \leq 3nLG\left(\sum_{z=2}^{T/L}\frac{3L(G+\alpha R)}{(z-2)L\alpha+2h}+\sum_{z=1}^{T/L}\frac{4L(G+2\alpha R)}{zL\alpha+2h}\right)+nhR^2. \tag{15}$$

Then, by combining Theorem 1 with suitable α and h, we can achieve specific regret bounds for convex and strongly convex functions, respectively.

Corollary 1 *Under Assumptions 1, 2, 3, and 4 with $\alpha = 0$, for any $i \in [n]$, Algorithm 1 with $\alpha = 0$, $h = \sqrt{11LT}G/R$, and θ, L defined in* (14) *ensures*

$$\mathrm{R}(T,i) \leq 2nGR\sqrt{11LT}.$$

Corollary 2 *Under Assumptions 1, 2, 3, and 4 with $\alpha > 0$, for any $i \in [n]$, Algorithm 1 with $\alpha > 0$, $h = \alpha L$, and θ, L defined in* (14) *ensures*

$$\mathrm{R}(T,i) \leq \frac{3nLG(7G+11\alpha R)(1+\ln(T/L))}{\alpha}+n\alpha LR^2.$$

Corollary 1 shows that all the learners of AD-FTGL enjoys 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. From Corollary 2, all the learners of AD-FTGL can exploit the strong convexity of functions to achieve an $O(n\rho^{-1/2}(\log n)\log T)$ regret bound, which has a much tighter dependence on T than the regret bound established by only using the convexity condition. 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) in terms of both n and ρ.

Remark 2 One may notice that the values of θ and L are carefully selected to establish the theoretical guarantees of our AD-FTGL. To emphasize their significance, we further consider two extreme cases: one with $\theta = 0$ and the other with $L = 1$. First, by setting $\theta = 0$, our AD-FTGL is equivalent to improving D-FTGL by only using multiple standard gossip steps. 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 $\bar{\mathbf{z}}(z)$. Correspondingly, the regret bounds of AD-FTGL for convex and strongly convex functions will degenerate to $O(n\rho^{-1/2}\sqrt{T\log n})$ and $O(n\rho^{-1}(\log n)\log T)$. Although these bounds are still tighter than the existing $O(n^{5/4}\rho^{-1/2}\sqrt{T})$ and $O(n^{3/2}\rho^{-1}\log T)$ regret bounds respectively, their dependence on ρ is worse than the regret bounds achieved in Corollaries 1 and 2. Moreover, by revisiting the analysis of D-FTGL (Wan et al., 2022), we show that the approximation error of its standard gossip step can be improved to $O(\rho^{-1}\log n)$, instead of $O(\rho^{-1}\sqrt{n})$ in (9) (see Appendix A for details). This result allows us to establish the $O(n\rho^{-1/2}\sqrt{T\log n})$ and $O(n\rho^{-1}(\log n)\log T)$ regret bounds for convex and strongly convex functions via the original D-FTGL. In other words, it is unnecessary to combine the blocking update mechanism with the standard gossip step. In contrast, the blocking update mechanism is critical for exploiting the accelerated gossip strategy. Note that if $L = 1$, our AD-FTGL becomes a non-blocked combination of D-FTGL with the accelerated gossip strategy. 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., $\mathbf{d}_i^{z-\tau}(\tau)$ in (12) for τ close to $z-1$, we can only modify the analysis of Lemma 2 to derive a worse error bound of $O(\rho^{-1/2}\sqrt{n})$. Correspondingly, the regret bounds of AD-FTGL for convex and strongly convex functions will degenerate to $O(n^{5/4}\rho^{-1/4}\sqrt{T})$ and $O(n^{3/2}\rho^{-1/2}\log T)$, whose dependence on n is much worse than the regret bounds achieved in Corollaries 1 and 2.

4.2 Lower Bounds

Although there still exist gaps between our improved regret bounds and the $\Omega(n\sqrt{T})$ and $\Omega(n)$ 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 $\Omega(n)$ communication delays for receiving the information of the global function $f_t(\mathbf{x})$. Because of the $\Omega(n)$ communication delays, we can establish $\Omega(n\sqrt{nT})$ and $\Omega(n^2)$ lower bounds for convex and strongly convex functions, respectively. Then, by exploiting the dependence of spectral properties on the network size n, we obtain lower bounds involving the spectral gap. Moreover, inspired by Wan et al. (2022), we also make an extension to obtain the following lower bounds for the more challenging setting with only C communication rounds.

Theorem 2 *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 communicating C rounds before round T, 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*

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

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 communicating C rounds before round T, there exists a sequence of 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*

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

Note that in 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 2 and 3 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, by combining Theorem 2 with $C = O(T)$, we can establish 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-FTGL up to polylogarithmic factors in n. For D-OCO with strongly convex functions, a lower bound of $\Omega(n\rho^{-1/2})$ can be established by combining Theorem 3 with $C = O(T)$, which matches the $O(n\rho^{-1/2}(\log n)\log T)$ regret of our AD-FTGL up to polylogarithmic factors in both n and T.

Now, our AD-FTGL have been shown to be nearly optimal for D-OCO with both convex and strongly convex functions. Nonetheless, there still exists an unsatisfactory point in the lower bound for strongly convex functions—it cannot recover the well-known $\Omega(\log T)$ lower bound for OCO with strongly convex functions (Abernethy et al., 2008; Hazan and Kale, 2014). To address this issue, we establish the following result by extending the analysis of Hazan and Kale (2014) from OCO into D-OCO.

Theorem 4 *Suppose $\mathcal{K} = [0, R/\sqrt{d}]^d$, which satisfies Assumption 3 and $n = 2(m+1)$ for some positive integer m. For any D-OCO algorithm, if $16n + 1 \leq T$, there exists a sequence of loss functions satisfying Assumption 4 and Assumption 2 with $G = \alpha R$, a graph $\mathcal{G} = ([n], E)$, and a matrix P satisfying Assumption 1 such that*

$$\mathrm{R}(T,1) \geq \frac{16^{-5}\alpha\pi(\log_{16}(30(T-1)/n) - 2)(n-2)R^2}{4\sqrt{1-\sigma_2(P)}}.$$

Compared with Theorem 3, Theorem 4 establishes an improved lower bound of $\Omega(n\rho^{-1/2}\log T)$ for D-OCO with strongly convex functions, which matches the $O(n\rho^{-1/2}(\log n)\log T)$ regret of our AD-FTGL up to polylogarithmic factors in only n.

Remark 3 One may also wonder whether it is possible to extend the result in Theorem 4 into the setting with only C communication rounds. However, there do exist some technical challenges for this extension (see discussions at the end of the proof of Theorem 4 for details), and thus we leave it as a future work.

4.3 A Projection-free Variant of Our Algorithm

Furthermore, to efficiently handle applications with complex constraints, we propose a projection-free variant of our AD-FTGL. Following the existing projection-free D-OCO algorithm in Wan et al. (2022), our main idea is to combine AD-FTGL with conditional gradient (CG)—a classical projection-free algorithm for offline optimization (Frank and Wolfe, 1956; Jaggi, 2013). Specifically, the detailed procedure of CG is outlined in Algorithm 2. Given a function $F(\mathbf{x}) : \mathcal{K} \mapsto \mathbb{R}$ and an initial point $\mathbf{y}_0 = \mathbf{x}_{\text{in}} \in \mathcal{K}$, it iteratively performs K linear optimization steps as shown from steps 3 to 7, and finally outputs $\mathbf{x}_{\text{out}} = \mathbf{y}_K$. To make AD-FTGL projection-free, it is natural to approximately solve (13) via CG.

However, there are still some technical details that require careful attention. First, for every invocation of CG, the number of iterations must equal to the block size, i.e., $K = L$, which ensures that each learner at most requires T linear optimization steps in total. Otherwise, even only using linear optimization steps, the time complexity could be equivalent to that of projection-based algorithms. Second, a straightforward combination of (13) and CG requires the algorithm to stop at the end of each block and wait until L linear optimization steps are completed. To avoid this issue, we

Algorithm 2 CG

Input: $\mathcal{K}$, K, $F(\mathbf{x})$, $\mathbf{x}_{\text{in}}$
Initialization: $\mathbf{y}_0 = \mathbf{x}_{\text{in}}$
for $k = 0, \ldots, K-1$ **do**
  $\mathbf{v}_k = \operatorname{argmin}_{\mathbf{x}\in\mathcal{K}} \langle \nabla F(\mathbf{y}_k), \mathbf{x}\rangle$
  $s_k = \operatorname{argmin}_{s\in[0,1]} F(\mathbf{y}_k + s(\mathbf{v}_k - \mathbf{y}_k))$
  $\mathbf{y}_{k+1} = \mathbf{y}_k + s_k(\mathbf{v}_k - \mathbf{y}_k)$
end for
return $\mathbf{x}_{\text{out}} = \mathbf{y}_K$

Algorithm 3 Projection-free Variant of AD-FTGL

Input: α, h, θ, L, 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**
  If $2 \leq z$, define $F_{z,i}(\mathbf{x}) = \langle \mathbf{z}_i(z-1), \mathbf{x}\rangle + \frac{(z-2)L\alpha}{2}\|\mathbf{x}\|_2^2 + h\|\mathbf{x}\|_2^2, \forall i \in [n]$
  If $2 \leq z$, set $\mathbf{z}_i^0(z) = \mathbf{z}_i(z-1) + \mathbf{d}_i(z-1)$, $\mathbf{z}_i^{-1}(z) = \mathbf{z}_i^{L'-1}(z-1) + \mathbf{d}_i(z-1), \forall i \in [n]$
  for $t = (z-1)L+1, \ldots, zL$ **do**
    for each local learner $i \in [n]$ **do**
      Play $\mathbf{x}_i(z)$, query $\nabla f_{t,i}(\mathbf{x}_i(z))$, and set $k = t - (z-1)L - 1$
      If $2 \leq z$ and $k < L'$, update $\mathbf{z}_i^{k+1}(z) = (1+\theta)\sum_{j\in N_i} P_{ij}\mathbf{z}_j^k(z) - \theta\mathbf{z}_i^{k-1}(z)$
    end for
  end for
  Set $\mathbf{d}_i(z) = \sum_{t\in\mathcal{T}_z}(\nabla f_{t,i}(\mathbf{x}_i(z)) - \alpha\mathbf{x}_i(z)), \forall i \in [n]$, where $\mathcal{T}_z = \{(z-1)L+1, \ldots, zL\}$
  If $2 \leq z$, set $\mathbf{z}_i(z) = \mathbf{z}_i^{L'}(z)$ and compute $\mathbf{x}_i(z+1) = \text{CG}(\mathcal{K}, L, F_{z,i}(\mathbf{x}), \mathbf{x}_i(z)), \forall i \in [n]$
end for

simply set $\mathbf{x}_i(1) = \mathbf{x}_i(2) = \mathbf{0}$, and then compute $\mathbf{x}_i(z+1)$ based on $\mathbf{z}_i(z-1)$, rather than $\mathbf{z}_i(z)$ used in (13), i.e., computing

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

where $F_{z,i}(\mathbf{x})$ denotes the intermediate objective function based on $\mathbf{z}_i(z-1)$, i.e.,

$$F_{z,i}(\mathbf{x}) = \langle \mathbf{z}_i(z-1), \mathbf{x}\rangle + \frac{(z-2)L\alpha}{2}\|\mathbf{x}\|_2^2 + h\|\mathbf{x}\|_2^2.$$

In this way, the L linear optimization steps required by CG can be uniformly allocated to every round in block z, since $\mathbf{z}_i(z-1)$ is available at the beginning of this block. Last, inspired by Wan et al. (2022), to control the approximation error of CG, L now should be much larger than that defined in (14). Note that the latter is sufficient for generating a good approximation of $\bar{\mathbf{z}}(z)$. Thus, we keep the number of accelerated gossip steps used in each block unchanged, and denote the specific value by L'. This enables us to achieve sublinear communication complexity.

The complete procedure of our projection-free algorithm is summarized in Algorithm 3. Despite its simplicity, we will demonstrate that this algorithm can improve the regret bounds of the existing projection-free D-OCO algorithm (Wan et al., 2022), while still ensuring the nearly optimal communication complexity.

By combining our original analysis of AD-FTGL with the convergence property of CG, we first establish the following guarantee for the regret bound of our projection-free algorithm.

Theorem 5 *Under Assumptions 1, 2, 3, and 4, for any $i \in [n]$, Algorithm 3 with*

$$\theta = \frac{1}{1+\sqrt{1-\sigma_2^2(P)}},\ L' = \left\lceil \frac{\sqrt{2}\ln(\sqrt{14n})}{(\sqrt{2}-1)\sqrt{1-\sigma_2(P)}} \right\rceil \tag{16}$$

and $L \geq L'$ ensures

$$\mathrm{R}(T,i) \leq 3nLG\left(\sum_{z=3}^{T/L} \frac{3L(G+\alpha R)}{(z-3)L\alpha+2h} + \sum_{z=1}^{T/L} \frac{6L(G+2\alpha R)}{(z-1)L\alpha+2h}\right) + nhR^2 + \frac{12nGRT}{\sqrt{L+2}}. \tag{17}$$

Compared with Theorem 1, the key difference of Theorem 5 is the term $12nGRT/\sqrt{L+2}$ in (17), which is caused by the approximation error of CG. On the one hand, to make this term sublinear in T, the final regret bound derived from Theorem 5 must be worse than that previously derived from Theorem 1. This is a common price for achieving the projection-free property. On the other hand, since the number of communication rounds used in Algorithm 3 is TL'/L, the choice of L for the trade-off between this term and other terms in (17) is also critical for reducing the communication complexity. Moreover, we notice that to reduce the communication complexity, the existing projection-free D-OCO algorithm in Wan et al. (2022) only communicates once per block. In contrast, our Algorithm 3 communicates L' times per block. This subtle difference allows us to further make different trade-offs between the regret and communication complexity.

Specifically, by combining Theorem 5 with suitable α, h, and L, we can establish specific regret bounds for our projection-free algorithm.

Corollary 3 *Suppose Assumptions 1, 2, 3, and 4 with $\alpha = 0$ hold. For any $i \in [n]$, Algorithm 3 with $\alpha = 0$, $h = \sqrt{14LT}G/R$, $L = \sqrt{T}L'$, and θ, L' defined in* (16) *ensures*

$$\mathrm{R}(T,i) \leq 2\sqrt{14}nGR\sqrt{L'}T^{3/4} + \frac{12nGRT^{3/4}}{\sqrt{L'}}. \tag{18}$$

Moreover, if (16) *satisfies $L' \leq \sqrt{T}$, for any $i \in [n]$, Algorithm 3 with $\alpha = 0$, $h = \sqrt{14LT}G/R$, $L = \sqrt{T}$, and θ, L' defined in* (16) *ensures*

$$\mathrm{R}(T,i) \leq (2\sqrt{14}+12)nGRT^{3/4}. \tag{19}$$

Corollary 4 *Suppose Assumptions 1, 2, 3, and 4 with $\alpha > 0$ hold. For any $i \in [n]$, Algorithm 3 with $\alpha > 0$, $h = \alpha L$, $L = T^{2/3}(\ln T)^{-2/3}L'$, and θ, L' defined in* (16) *ensures*

$$\begin{aligned}\mathrm{R}(T,i) \leq & \frac{3nG(9G+15\alpha R)T^{2/3}L'((\ln T)^{-2/3}+(\ln T)^{1/3})}{\alpha} \\ & + n\alpha T^{2/3}L'R^2(\ln T)^{-2/3} + \frac{12nGRT^{2/3}(\ln T)^{1/3}}{\sqrt{L'}}.\end{aligned} \tag{20}$$

Moreover, if (16) *satisfies $L' \leq T^{2/3}(\ln T)^{-2/3}$, for any $i \in [n]$, Algorithm 3 with $\alpha > 0$, $h = \alpha L$, $L = T^{2/3}(\ln T)^{-2/3}$, and θ, L' defined in* (16) *ensures*

$$\begin{aligned}\mathrm{R}(T,i) \leq & \frac{3nG(9G+15\alpha R)T^{2/3}((\ln T)^{-2/3}+(\ln T)^{1/3})}{\alpha} \\ & + n\alpha T^{2/3}R^2(\ln T)^{-2/3} + 12nGRT^{2/3}(\ln T)^{1/3}.\end{aligned} \tag{21}$$

Note that both Corollaries 3 and 4 present two choices for parameters of our Algorithm 3. From (18) and (20), our projection-free algorithm can enjoy an $O(n\rho^{-1/4}\sqrt{\log n}T^{3/4})$ regret bound for convex functions with $O(\sqrt{T})$ communication rounds, and an $O(n\rho^{-1/2}T^{2/3}(\log T)^{1/3}\log n)$ regret bound for strongly convex functions with $O(T^{1/3}(\log T)^{2/3})$ communication rounds. In contrast, with the same number of communication rounds, the existing projection-free algorithm in Wan et al. (2022) only achieves the worse $O(n^{5/4}\rho^{-1/2}T^{3/4})$ and $O(n^{3/2}\rho^{-1}T^{2/3}(\log T)^{1/3})$ regret bounds for convex functions and strongly convex functions, respectively. Since their projection-free algorithm is a variant of D-FTGL, this comparison implies that our projection-free algorithm can simply inherit the improvement of AD-FTGL over D-FTGL. Besides that, from (19) and (21), our projection-free algorithm can further reduce regret bounds for convex and strongly convex functions to $O(nT^{3/4})$ and $O(nT^{2/3}(\log T)^{1/3})$ by increasing the number of communication rounds to $O(\rho^{-1/2}\sqrt{T}\log n)$ and $O(\rho^{-1/2}T^{1/3}(\log T)^{2/3}\log n)$, respectively. This is somewhat surprising since without other additional assumptions, even running existing centralized projection-free OCO algorithms (Hazan and Kale, 2012; Wan and Zhang, 2021) over the global function $f_t(\mathbf{x})$ can only achieve the same regret bounds (up to polylogarithmic factors in T for strongly convex functions). Finally, according to the lower bounds in Theorems 2 and 3, the number of communication rounds required by our projection-free algorithm to achieve the above regret bounds is optimal up to polylogarithmic factors in n for convex functions, and polylogarithmic factors in n and T for strongly convex functions, respectively.

5. Theoretical Analysis

Here, we provide the proofs of our theoretical guarantees on AD-FTGL, lower bounds, and the projection-free variant of AD-FTGL. The refined analysis for D-FTGL can be found in the appendix.

5.1 Proof of Lemma 2

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

$$\mathbf{d}_i^{k+1}(z) = (1+\theta)\sum_{j\in N_i} P_{ij}\mathbf{d}_j^k(z) - \theta\mathbf{d}_i^{k-1}(z). \tag{22}$$

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

$$\mathbf{z}_i^k(z) = \sum_{\tau=1}^{z-1}\mathbf{d}_i^{(z-\tau-1)L+k}(\tau),\ \forall k = 1, \ldots, L \tag{23}$$

by the induction method.

For $z = 2$, it is easy to verify that (23) holds due to $\mathbf{z}_i^0(2) = \mathbf{z}_i^{-1}(2) = \mathbf{d}_i(1)$ and (22). Then, we assume that (23) holds for some $z \geq 2$, and prove it also holds for $z+1$. From step 4 of our Algorithm 1, we have

$$\begin{aligned}
\mathbf{z}_i^0(z+1) &= \mathbf{z}_i(z) + \mathbf{d}_i(z) = \mathbf{z}_i^L(z) + \mathbf{d}_i^0(z) \overset{(23)}{=} \sum_{\tau=1}^{z}\mathbf{d}_i^{(z-\tau)L}(\tau),\\
\mathbf{z}_i^{-1}(z+1) &= \mathbf{z}_i^{L-1}(z) + \mathbf{d}_i(z) = \mathbf{z}_i^{L-1}(z) + \mathbf{d}_i^{-1}(z) \overset{(23)}{=} \sum_{\tau=1}^{z}\mathbf{d}_i^{(z-\tau)L-1}(\tau).
\end{aligned} \tag{24}$$

By combining (24) with step 8 of Algorithm 1, for $k = 1$, we have

$$\begin{aligned}\mathbf{z}_i^k(z+1) =&(1+\theta)\sum_{j\in N_i}P_{ij}\mathbf{z}_j^{k-1}(z+1)-\theta\mathbf{z}_i^{k-2}(z+1)\\ =&(1+\theta)\sum_{j\in N_i}P_{ij}\sum_{\tau=1}^{z}\mathbf{d}_j^{(z-\tau)L+k-1}(\tau)-\theta\sum_{\tau=1}^{z}\mathbf{d}_i^{(z-\tau)L-1+k-1}(\tau)\\ =&\sum_{\tau=1}^{z}\left((1+\theta)\sum_{j\in N_i}P_{ij}\mathbf{d}_j^{(z-\tau)L+k-1}(\tau)-\theta\mathbf{d}_i^{(z-\tau)L-1+k-1}(\tau)\right)\\ \overset{(22)}{=}&\sum_{\tau=1}^{z}\mathbf{d}_i^{(z-\tau)L+k}(\tau).\end{aligned} \tag{25}$$

By repeating (25) for $k = 2, \dots, L$, the proof of (23) for $z + 1$ is completed.

Then, from (23), for any $i \in [n]$ and $z = 2, \dots, T/L$, we have

$$\|\mathbf{z}_i(z)-\bar{\mathbf{z}}(z)\|_2=\left\|\sum_{\tau=1}^{z-1}\mathbf{d}_i^{(z-\tau)L}(\tau)-\sum_{\tau=1}^{z-1}\bar{\mathbf{d}}(\tau)\right\|_2\leq\sum_{\tau=1}^{z-1}\left\|\mathbf{d}_i^{(z-\tau)L}(\tau)-\bar{\mathbf{d}}(\tau)\right\|_2. \tag{26}$$

To further analyze the right side of (26), we define

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

for any integer $k \geq -1$ and $\bar{X} = \left[\bar{\mathbf{d}}(\tau)^\top; \dots; \bar{\mathbf{d}}(\tau)^\top\right] \in \mathbb{R}^{n\times d}$. According to (22), it is not hard to verify that the sequence of $X^1, \dots, X^L$ follows the update rule in (8).

Let $c = 1 - 1/\sqrt{2}$. Due to Lemma 1, for any $\tau < z$, we have

$$\begin{aligned}&\left\|X^{(z-\tau)L}-\bar{X}\right\|_F\leq\sqrt{14}\left(1-c\sqrt{1-\sigma_2(P)}\right)^{(z-\tau)L}\left\|X^0-\bar{X}\right\|_F\\ \leq&\sqrt{14}\left(1-c\sqrt{1-\sigma_2(P)}\right)^{(z-\tau)L}\left(\left\|X^0\right\|_F+\left\|\bar{X}\right\|_F\right)\\ =&\sqrt{14}\left(1-c\sqrt{1-\sigma_2(P)}\right)^{(z-\tau)L}\left(\sqrt{\sum_{i=1}^{n}\|\mathbf{d}_i(\tau)\|_2^2}+\sqrt{n\|\bar{\mathbf{d}}(\tau)\|_2^2}\right).\end{aligned} \tag{27}$$

Because of Assumptions 2 and 3, for any $z \in [T/L]$ and $i \in [n]$, it is easy to verify that

$$\begin{aligned}\|\mathbf{d}_i(z)\|_2&=\left\|\sum_{t\in\mathcal{T}_z}(\nabla f_{t,i}(\mathbf{x}_i(z))-\alpha\mathbf{x}_i(z))\right\|_2\leq L(G+\alpha R),\\ \|\bar{\mathbf{d}}(z)\|_2&=\left\|\frac{1}{n}\sum_{i=1}^{n}\mathbf{d}_i(z)\right\|_2\leq L(G+\alpha R).\end{aligned} \tag{28}$$

By combining (27) with (28), for any $i \in [n]$ and $\tau < z$, we have

$$\begin{aligned}\left\|\mathbf{d}_i^{(z-\tau)L}(\tau)-\bar{\mathbf{d}}(\tau)\right\|_2\leq&\left\|X^{(z-\tau)L}-\bar{X}\right\|_F\\ \leq&2\sqrt{14n}\left(1-c\sqrt{1-\sigma_2(P)}\right)^{(z-\tau)L}L(G+\alpha R).\end{aligned} \tag{29}$$

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

$$\begin{aligned}\epsilon=&\left(1-c\sqrt{1-\sigma_2(P)}\right)^L\leq\left(1-c\sqrt{1-\sigma_2(P)}\right)^{\frac{\ln(\sqrt{14n})}{c\sqrt{1-\sigma_2(P)}}}\\\leq&\left(1-c\sqrt{1-\sigma_2(P)}\right)^{\frac{\ln(\sqrt{14n})}{\ln\left(1/\left(1-c\sqrt{1-\sigma_2(P)}\right)\right)}}=\frac{1}{\sqrt{14n}}\end{aligned}\tag{30}$$

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

By combining (26) with (29) and (30), for any $i\in[n]$ and $z=2,\dots,T/L$, we have

$$\begin{aligned}\|\mathbf{z}_i(z)-\bar{\mathbf{z}}(z)\|_2\leq&2L(G+\alpha R)\sqrt{14n}\sum_{\tau=1}^{z-1}\epsilon^{(z-\tau)}\overset{(30)}{\leq}2L(G+\alpha R)\sum_{\tau=1}^{z-1}\epsilon^{(z-\tau-1)}\\\leq&\frac{2L(G+\alpha R)}{1-\epsilon}\overset{(30)}{\leq}2L(G+\alpha R)+\frac{2L(G+\alpha R)}{\sqrt{14n}-1}\leq 3L(G+\alpha R)\end{aligned}\tag{31}$$

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 (31) with $\|\mathbf{z}_i(1)-\bar{\mathbf{z}}(1)\|_2=0$.

5.2 Proof of Theorem 1

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

$$\bar{\mathbf{x}}(z)=\operatorname*{argmin}_{\mathbf{x}\in\mathcal{K}}\langle\mathbf{x},\bar{\mathbf{z}}(z)\rangle+\frac{(z-1)L\alpha}{2}\|\mathbf{x}\|_2^2+h\|\mathbf{x}\|_2^2\tag{32}$$

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 $\bar{\mathbf{x}}(2),\dots,\bar{\mathbf{x}}(T/L+1)$ on a sequence of loss functions defined by $\bar{\mathbf{d}}(1),\dots,\bar{\mathbf{d}}(T/L)$ and the distance $\|\mathbf{x}_i(z)-\bar{\mathbf{x}}(z+1)\|_2$ for any $z\in[T/L]$. To this end, we first introduce two useful lemmas.

Lemma 3 *(Garber and Hazan, 2016, Lemma 6.6) 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 4 *(Duchi et al., 2011, Lemma 5) 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 $\ell_z(\mathbf{x})=\langle\mathbf{x},\bar{\mathbf{d}}(z)\rangle+\frac{L\alpha}{2}\|\mathbf{x}\|_2^2$ for any $z\in[T/L]$. By combining Lemma 3 with (32), for any $\mathbf{x}\in\mathcal{K}$, it is easy to verify that

$$\sum_{z=1}^{T/L}\ell_z(\bar{\mathbf{x}}(z+1))-\sum_{z=1}^{T/L}\ell_z(\mathbf{x})\leq h\left(\|\mathbf{x}\|_2^2-\|\bar{\mathbf{x}}(2)\|_2^2\right)\leq hR^2\tag{33}$$

where the last inequality is due to Assumption 3 and $\|\bar{\mathbf{x}}(2)\|_2^2 \geq 0$.

Then, we also notice that for any $z = 2, \dots, T/L$, Algorithm 1 ensures

$$\mathbf{x}_i(z) = \operatorname*{argmin}_{\mathbf{x}\in\mathcal{K}} \langle \mathbf{x}, \mathbf{z}_i(z-1)\rangle + \frac{(z-2)L\alpha}{2}\|\mathbf{x}\|_2^2 + h\|\mathbf{x}\|_2^2. \tag{34}$$

By combining Lemma 4 with (32) and (34), for any $z = 2, \dots, T/L$, we have

$$\|\mathbf{x}_i(z) - \bar{\mathbf{x}}(z-1)\|_2 \leq \frac{\|\mathbf{z}_i(z-1) - \bar{\mathbf{z}}(z-1)\|_2}{(z-2)L\alpha + 2h} \leq \frac{3L(G+\alpha R)}{(z-2)L\alpha + 2h} \tag{35}$$

where the last inequality is due to Lemma 2.

To bound $\|\mathbf{x}_i(z) - \bar{\mathbf{x}}(z+1)\|_2$, we still need to analyze the term $\|\bar{\mathbf{x}}(z) - \bar{\mathbf{x}}(z+1)\|_2$ for any $z \in [T/L]$. Let $F_z(\mathbf{x}) = \sum_{\tau=1}^{z} \ell_\tau(\mathbf{x}) + h\|\mathbf{x}\|_2^2$ for any $z \in [T/L]$. It is easy to verify that $F_z(\mathbf{x})$ is $(zL\alpha + 2h)$-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

$$\frac{\alpha}{2}\|\mathbf{x} - \mathbf{x}^*\|_2^2 \leq f(\mathbf{x}) - f(\mathbf{x}^*) \tag{36}$$

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

$$\begin{aligned}|\ell_z(\mathbf{x}) - \ell_z(\mathbf{y})| &\leq |\langle \nabla \ell_z(\mathbf{x}), \mathbf{x} - \mathbf{y}\rangle| \leq \|\nabla \ell_z(\mathbf{x})\|_2 \|\mathbf{x} - \mathbf{y}\|_2 \\ &= \|\bar{\mathbf{d}}(z) + \alpha L \mathbf{x}\|_2 \|\mathbf{x} - \mathbf{y}\|_2 \overset{(28)}{\leq} L(G + 2\alpha R)\|\mathbf{x} - \mathbf{y}\|_2.\end{aligned} \tag{37}$$

Then, for any $z' \leq z \in [T/L]$, it is not hard to verify that

$$\begin{aligned}&\|\bar{\mathbf{x}}(z') - \bar{\mathbf{x}}(z+1)\|_2^2 \overset{(36)}{\leq} \frac{2}{zL\alpha + 2h}\left(F_z(\bar{\mathbf{x}}(z')) - F_z(\bar{\mathbf{x}}(z+1))\right) \\ =&\frac{2}{zL\alpha + 2h}\left(F_{z'-1}(\bar{\mathbf{x}}(z')) - F_{z'-1}(\bar{\mathbf{x}}(z+1)) + \sum_{\tau=z'}^{z}(\ell_\tau(\bar{\mathbf{x}}(z')) - \ell_\tau(\bar{\mathbf{x}}(z+1)))\right) \\ \overset{(32)}{\leq}&\frac{2\sum_{\tau=z'}^{z}(\ell_\tau(\bar{\mathbf{x}}(z')) - \ell_\tau(\bar{\mathbf{x}}(z+1)))}{zL\alpha + 2h} \\ \overset{(37)}{\leq}&\frac{2(z - z' + 1)L(G + 2\alpha R)\|\bar{\mathbf{x}}(z') - \bar{\mathbf{x}}(z+1)\|_2}{zL\alpha + 2h}\end{aligned}$$

which implies that

$$\|\bar{\mathbf{x}}(z') - \bar{\mathbf{x}}(z+1)\|_2 \leq \frac{2(z - z' + 1)L(G + 2\alpha R)}{zL\alpha + 2h}. \tag{38}$$

By combining (35) and (38), for any $z = 2, \dots, T/L$, we have

$$\begin{aligned}\|\mathbf{x}_i(z) - \bar{\mathbf{x}}(z+1)\|_2 \leq& \|\mathbf{x}_i(z) - \bar{\mathbf{x}}(z-1)\|_2 + \|\bar{\mathbf{x}}(z-1) - \bar{\mathbf{x}}(z+1)\|_2 \\ \leq& \frac{3L(G+\alpha R)}{(z-2)L\alpha + 2h} + \frac{4L(G + 2\alpha R)}{zL\alpha + 2h}.\end{aligned} \tag{39}$$

For $z = 1$, we notice that $\mathbf{x}_i(1) = \bar{\mathbf{x}}(1) = \mathbf{0}$, and it is easy to verify that

$$\|\mathbf{x}_i(z) - \bar{\mathbf{x}}(z+1)\|_2 = \|\bar{\mathbf{x}}(z) - \bar{\mathbf{x}}(z+1)\|_2 \overset{(38)}{\leq} \frac{2L(G + 2\alpha R)}{zL\alpha + 2h}. \tag{40}$$

For brevity, let ϵ_z denote the upper bound of $\|\mathbf{x}_i(z)-\bar{\mathbf{x}}(z+1)\|_2$ derived in (39) and (40).

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

$$\begin{aligned}
f_{t,j}(\mathbf{x}_i(z))-f_{t,j}(\mathbf{x}) \leq& f_{t,j}(\mathbf{x}_j(z))-f_{t,j}(\mathbf{x})+G\|\mathbf{x}_j(z)-\mathbf{x}_i(z)\|_2\\
\leq&\langle\nabla f_{t,j}(\mathbf{x}_j(z)),\mathbf{x}_j(z)-\mathbf{x}\rangle-\frac{\alpha}{2}\|\mathbf{x}_j(z)-\mathbf{x}\|_2^2\\
&+G\|\mathbf{x}_j(z)-\bar{\mathbf{x}}(z+1)+\bar{\mathbf{x}}(z+1)-\mathbf{x}_i(z)\|_2\\
\leq&\langle\nabla f_{t,j}(\mathbf{x}_j(z)),\bar{\mathbf{x}}(z+1)-\mathbf{x}\rangle-\frac{\alpha}{2}\|\mathbf{x}_j(z)-\mathbf{x}\|_2^2\\
&+\langle\nabla f_{t,j}(\mathbf{x}_j(z)),\mathbf{x}_j(z)-\bar{\mathbf{x}}(z+1)\rangle+2G\epsilon_z\\
\leq&\langle\nabla f_{t,j}(\mathbf{x}_j(z)),\bar{\mathbf{x}}(z+1)-\mathbf{x}\rangle-\frac{\alpha}{2}\|\mathbf{x}_j(z)-\mathbf{x}\|_2^2\\
&+G\|\mathbf{x}_j(z)-\bar{\mathbf{x}}(z+1)\|_2+2G\epsilon_z\\
\leq&\langle\nabla f_{t,j}(\mathbf{x}_j(z)),\bar{\mathbf{x}}(z+1)-\mathbf{x}\rangle-\frac{\alpha}{2}\|\mathbf{x}_j(z)-\mathbf{x}\|_2^2+3G\epsilon_z
\end{aligned}\tag{41}$$

where the third and last inequalities are due to (39) and (40). Moreover, for any $\mathbf{x},\mathbf{y},\mathbf{y}'$, we have

$$\|\mathbf{y}-\mathbf{x}\|_2^2=\|\mathbf{y}-\mathbf{y}'\|_2^2+2\langle\mathbf{y},\mathbf{y}'-\mathbf{x}\rangle+\|\mathbf{x}\|_2^2-\|\mathbf{y}'\|_2^2\geq2\langle\mathbf{y},\mathbf{y}'-\mathbf{x}\rangle+\|\mathbf{x}\|_2^2-\|\mathbf{y}'\|_2^2. \tag{42}$$

By combining (41) with (42), for any $z\in[T/L]$, $t\in\mathcal{T}_z$, $j\in[n]$, and $\mathbf{x}\in\mathcal{K}$, we have

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

Finally, from the above inequality and the definition of ϵ_z, for any $\mathbf{x}\in\mathcal{K}$, it is not hard to verify that

$$\begin{aligned}
&\sum_{z=1}^{T/L}\sum_{t\in\mathcal{T}_z}\sum_{j=1}^{n}f_{t,j}(\mathbf{x}_i(z))-\sum_{z=1}^{T/L}\sum_{t\in\mathcal{T}_z}\sum_{j=1}^{n}f_{t,j}(\mathbf{x})\\
\leq&\sum_{z=1}^{T/L}\sum_{t\in\mathcal{T}_z}\sum_{j=1}^{n}\left(\langle\nabla f_{t,j}(\mathbf{x}_j(z))-\alpha\mathbf{x}_j(z),\bar{\mathbf{x}}(z+1)-\mathbf{x}\rangle+\frac{\alpha}{2}\left(\|\bar{\mathbf{x}}(z+1)\|_2^2-\|\mathbf{x}\|_2^2\right)+3G\epsilon_z\right)\\
=&n\sum_{z=1}^{T/L}\left(\langle\bar{\mathbf{d}}(z),\bar{\mathbf{x}}(z+1)-\mathbf{x}\rangle+\frac{L\alpha}{2}(\|\bar{\mathbf{x}}(z+1)\|_2^2-\|\mathbf{x}\|_2^2)\right)+3nLG\sum_{z=1}^{T/L}\epsilon_z\\
\overset{(33)}{\leq}&nhR^2+3nLG\left(\sum_{z=2}^{T/L}\frac{3L(G+\alpha R)}{(z-2)L\alpha+2h}+\sum_{z=1}^{T/L}\frac{4L(G+2\alpha R)}{zL\alpha+2h}\right).
\end{aligned}\tag{43}$$

5.3 Proof of Corollaries 1 and 2

By substituting $\alpha=0$ and $h=\sqrt{11LT}G/R$ into (15), we have

$$\mathrm{R}(T,i)\leq3nLG\left(\sum_{z=2}^{T/L}\frac{3LG}{2h}+\sum_{z=1}^{T/L}\frac{2LG}{h}\right)+nhR^2\leq\frac{11nLG^2T}{h}+nhR^2=2nGR\sqrt{11LT}$$

which completes the proof of Corollary 1.

Similarly, by substituting $h = \alpha L$ into (15), we have

$$\begin{aligned}
\mathrm{R}(T,i) \leq & 3nLG\left(\sum_{z=2}^{T/L}\frac{3(G+\alpha R)}{z\alpha}+\sum_{z=1}^{T/L}\frac{4(G+2\alpha R)}{(z+2)\alpha}\right)+n\alpha LR^2\\
\leq & \frac{3nLG(7G+11\alpha R)}{\alpha}\sum_{z=1}^{T/L}\frac{1}{z}+n\alpha LR^2\\
\leq & \frac{3nLG(7G+11\alpha R)(1+\ln(T/L))}{\alpha}+n\alpha LR^2
\end{aligned}$$

which completes the proof of Corollary 2.

5.4 Proof of Theorem 2

Recall that Wan et al. (2022) have established an $\Omega(nT/\sqrt{C})$ lower bound by extending the classical randomized lower bound for OCO (Abernethy et al., 2008) into D-OCO with limited communications. 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 $\delta_i = |N_i| - 1$ 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 gossip matrix P satisfying Assumption 1, i.e.,

$$P = I_n - \frac{1}{\delta_{\max}+1}(D-A) \tag{44}$$

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 use 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 5 *For the 1-connected cycle graph with $n = 2(m+1)$, where m denotes a positive integer, the gossip matrix defined in (44) satisfies*

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

Then, we only need to derive a lower bound of $\Omega(n\sqrt{n}T/\sqrt{C})$ since combining it with Lemma 5 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 $f_{t,\lceil m/2\rceil+1}(\mathbf{x}),\ldots,f_{t,n-\lceil m/2\rceil+1}(\mathbf{x})$.

Without loss of generality, we denote the set of communication rounds by $\mathcal{C} = \{c_1,\ldots,c_C\}$, where $1 \leq c_1 < \cdots < c_C < T$. 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 $K = \lceil m/2 \rceil$, $Z = \lfloor C/K \rfloor$, $c_0 = 0$, and $c_{(Z+1)K} = T$. The total T rounds can be divided into the following $Z + 1$ intervals

$$[c_0 + 1, c_K], [c_K + 1, c_{2K}], \ldots, [c_{ZK} + 1, c_{(Z+1)K}]. \tag{45}$$

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

$$f_t(\mathbf{x}) = (n - 2K + 1)h_i(\mathbf{x}). \tag{46}$$

Moreover, according to the above discussion, the decisions $\mathbf{x}_1(c_{iK} + 1), \ldots, \mathbf{x}_1(c_{(i+1)K})$ for any $i \in \{0, \ldots, Z\}$ are made before the function $h_i(\mathbf{x})$ can be revealed to learner 1. As a result, we can use the classical randomized strategy to select $h_i(\mathbf{x})$ for any $i \in \{0, \ldots, Z\}$, and derive an expected lower bound for $\mathrm{R}(T, 1)$.

To be precise, we independently select $h_i(\mathbf{x}) = \langle \mathbf{w}_i, \mathbf{x} \rangle$ for any $i \in \{0, \ldots, Z\}$, 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

$$\begin{aligned}
&\mathbb{E}_{\mathbf{w}_0,\ldots,\mathbf{w}_Z}[\mathrm{R}(T,1)] \\
\overset{(46)}{=}&\mathbb{E}_{\mathbf{w}_0,\ldots,\mathbf{w}_Z}\left[\sum_{i=0}^{Z}\sum_{t=c_{iK}+1}^{c_{(i+1)K}}(n-2K+1)h_i(\mathbf{x}_1(t)) - \min_{\mathbf{x}\in\mathcal{K}}\sum_{i=0}^{Z}\sum_{t=c_{iK}+1}^{c_{(i+1)K}}(n-2K+1)h_i(\mathbf{x})\right] \\
=&(n-2K+1)\mathbb{E}_{\mathbf{w}_0,\ldots,\mathbf{w}_Z}\left[\sum_{i=0}^{Z}\sum_{t=c_{iK}+1}^{c_{(i+1)K}}\langle \mathbf{w}_i, \mathbf{x}_1(t)\rangle - \min_{\mathbf{x}\in\mathcal{K}}\sum_{i=0}^{Z}(c_{(i+1)K}-c_{iK})\langle \mathbf{w}_i, \mathbf{x}\rangle\right] \\
=&-(n-2K+1)\mathbb{E}_{\mathbf{w}_0,\ldots,\mathbf{w}_Z}\left[\min_{\mathbf{x}\in\mathcal{K}}\sum_{i=0}^{Z}(c_{(i+1)K}-c_{iK})\langle \mathbf{w}_i, \mathbf{x}\rangle\right] \\
=&-(n-2K+1)\mathbb{E}_{\mathbf{w}_0,\ldots,\mathbf{w}_Z}\left[\min_{\mathbf{x}\in\left\{-R/\sqrt{d},R/\sqrt{d}\right\}^d}\left\langle \mathbf{x}, \sum_{i=0}^{Z}(c_{(i+1)K}-c_{iK})\mathbf{w}_i\right\rangle\right]
\end{aligned} \tag{47}$$

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

Then, let $\epsilon_{01}, \ldots, \epsilon_{0d}, \ldots, \epsilon_{Z1}, \ldots, \epsilon_{Zd}$ be independent and identically distributed variables with $\Pr(\epsilon_{ij} = \pm 1) = 1/2$ for any $i \in \{0, \ldots, Z\}$ and $j \in \{1, \ldots, d\}$. By combining these notations with (47), we have

$$\begin{aligned}
\mathbb{E}_{\mathbf{w}_0,\ldots,\mathbf{w}_Z}[\mathrm{R}(T,1)] =&-(n-2K+1)\mathbb{E}_{\epsilon_{01},\ldots,\epsilon_{Zd}}\left[\sum_{j=1}^{d}-\frac{R}{\sqrt{d}}\left|\sum_{i=0}^{Z}(c_{(i+1)K}-c_{iK})\frac{\epsilon_{ij}G}{\sqrt{d}}\right|\right] \\
=&(n-2K+1)RG\mathbb{E}_{\epsilon_{01},\ldots,\epsilon_{Z1}}\left[\left|\sum_{i=0}^{Z}(c_{(i+1)K}-c_{iK})\epsilon_{i1}\right|\right].
\end{aligned} \tag{48}$$

Moreover, by combining (48) with the Khintchine inequality, we have

$$\begin{aligned}\mathbb{E}_{\mathbf{w}_0,\ldots,\mathbf{w}_Z}[\mathrm{R}(T,1)] \geq&\frac{(n-2K+1)RG}{\sqrt{2}}\sqrt{\sum_{i=0}^{Z}(c_{(i+1)K}-c_{iK})^2}\\ \geq&\frac{(n-2K+1)RG}{\sqrt{2}}\sqrt{\frac{(c_{(Z+1)K}-c_0)^2}{Z+1}}=\frac{(n-2K+1)RGT}{\sqrt{2(Z+1)}}\end{aligned}\tag{49}$$

where the second inequality is due to the Cauchy-Schwarz inequality.

Note that the expected lower bound in (49) implies that for any D-OCO algorithm with communication rounds $\mathcal{C}=\{c_1,\ldots,c_C\}$, there exists a particular choice of $\mathbf{w}_0,\ldots,\mathbf{w}_Z$ such that

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

where the last inequality is due to

$$\begin{aligned}\frac{n-2K+1}{\sqrt{Z+1}}=&\frac{n-2\lceil m/2\rceil+1}{\sqrt{\lfloor C/\lceil m/2\rceil\rfloor+1}}\geq\frac{n-m-1}{\sqrt{C/\lceil m/2\rceil+1}}=\frac{(m+1)\sqrt{m+1}}{\sqrt{(C/\lceil m/2\rceil+1)(m+1)}}\\ \geq&\frac{(m+1)\sqrt{m+1}}{\sqrt{4C+m+1}}=\frac{n\sqrt{n}}{2\sqrt{8C+n}}.\end{aligned}\tag{50}$$

If $n\leq 8C+16$, by combining the above result on $\mathrm{R}(T,1)$ with Lemma 5, we have

$$\mathrm{R}(T,1)\geq\frac{n\sqrt{\pi}RGT}{16(1-\sigma_2(P))^{1/4}\sqrt{C+1}}.$$

Otherwise, we have $8C<n-16<n$, and thus $\mathrm{R}(T,1)\geq\frac{nRGT}{4}$.

5.5 Proof of Lemma 5

We start this proof by introducing a general lemma regarding the spectral gap of the communication matrix P defined in (44).

Lemma 6 *(Duchi et al., 2011, Lemma 4) 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 (44) 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 $\mathcal{L}$.

As discussed in Duchi et al. (2011), $\mathcal{L}$ has the following eigenvalues

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

for the 1-connected cycle graph.

Then, 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 6, 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 $\lambda_{n-1}(\mathcal{L}) \leq 1 - \cos(\pi/2) = 1$.

Finally, it is easy to verify that

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

5.6 Proof of Theorem 3

The proof of Theorem 3 is similar to the proof of Theorem 2. 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 $K = \lceil m/2 \rceil$, $Z = \lfloor C/K \rfloor$, $c_0 = 0$, and $c_{(Z+1)K} = T$. 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 divide the total T rounds into $Z+1$ intervals defined in (45). 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 Assumption 4. Moreover, for any $i \in \{0, \ldots, Z\}$ and $t \in [c_{iK}+1, c_{(i+1)K}]$, 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 $1/2$. It is easy to verify that $h_i(\mathbf{x})$ also satisfies Assumption 2 with $G = 2\alpha R$ and Assumption 4. Following the proof of Theorem 2, we set $\mathcal{G}$ as the 1-connected cycle graph, which ensures that the decisions $\mathbf{x}_1(c_{iK}+1), \ldots, \mathbf{x}_1(c_{(i+1)K})$ are independent of $\mathbf{w}_i$.

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

$$\begin{aligned}
\sum_{t=1}^{T} f_t(\mathbf{x}) =& \sum_{i=0}^{Z}(c_{(i+1)K} - c_{iK})\left((n - 2K + 1)\langle \mathbf{w}_i, \mathbf{x} \rangle + \frac{\alpha n}{2}\|\mathbf{x}\|_2^2\right) \\
=& \alpha(n - 2K + 1)T\langle \bar{\mathbf{w}}, \mathbf{x} \rangle + \frac{\alpha n T}{2}\|\mathbf{x}\|_2^2 \\
=& \frac{\alpha T}{2}\left(\left\|\sqrt{n}\mathbf{x} + \frac{(n-2K+1)}{\sqrt{n}}\bar{\mathbf{w}}\right\|_2^2 - \left\|\frac{(n-2K+1)}{\sqrt{n}}\bar{\mathbf{w}}\right\|_2^2\right).
\end{aligned} \tag{51}$$

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

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

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 (51), 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

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

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

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

$$\begin{aligned}
\mathbb{E}_{\mathbf{w}_0,\dots,\mathbf{w}_Z}\left[\mathrm{R}(T,1)\right]\geq&\frac{(n-2K+1)^2}{2\alpha nT}\mathbb{E}_{\mathbf{w}_0,\dots,\mathbf{w}_Z}\left[\left\|\sum_{i=0}^{Z}(c_{(i+1)K}-c_{iK})\mathbf{w}_i\right\|_2^2\right]\\
=&\frac{(n-2K+1)^2}{2\alpha nT}\mathbb{E}_{\epsilon_{01},\dots,\epsilon_{Zd}}\left[\sum_{j=1}^{d}\left|\sum_{i=0}^{Z}(c_{(i+1)K}-c_{iK})\frac{\epsilon_{ij}\alpha R}{\sqrt{d}}\right|^2\right]\\
=&\frac{\alpha(n-2K+1)^2R^2}{2nT}\mathbb{E}_{\epsilon_{01},\dots,\epsilon_{Z1}}\left[\left|\sum_{i=0}^{Z}(c_{(i+1)K}-c_{iK})\epsilon_{i1}\right|^2\right]\\
=&\frac{\alpha(n-2K+1)^2R^2}{2nT}\sum_{i=0}^{Z}(c_{(i+1)K}-c_{iK})^2\geq\frac{\alpha(n-2K+1)^2R^2T}{2n(Z+1)}
\end{aligned}\tag{52}$$

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

The expected lower bound in (52) implies that for any D-OCO algorithm with communication rounds $\mathcal{C}=\{c_1,\dots,c_C\}$, there exists a particular choice of $\mathbf{w}_0,\dots,\mathbf{w}_Z$ such that

$$\mathrm{R}(T,1)\geq\frac{\alpha(n-2K+1)^2R^2T}{2n(Z+1)}\overset{(50)}{\geq}\frac{\alpha n^2R^2T}{8(8C+n)}.$$

If $n \leq 8C + 16$, according to Lemma 5, by using the gossip matrix P defined in (44), we have

$$\mathrm{R}(T,1) \geq \frac{\alpha \pi n R^2 T}{256(C+1)\sqrt{1-\sigma_2(P)}}.$$

Otherwise, we have $8C < n - 16 < n$, and thus $\mathrm{R}(T,1) \geq \frac{\alpha n R^2 T}{16}$.

5.7 Proof of Theorem 4

Compared with the proof of Theorem 3, the main difference of this proof is to focus on the case with $C = T - 1$ and redefine the loss functions as well as the decision set. Specifically, for any D-OCO algorithm, we still denote the sequence of decisions made by the local learner 1 as $\mathbf{x}_1(1), \ldots, \mathbf{x}_1(T)$, but divide the total T rounds into the following $Z + 1$ intervals

$$[c_0 + 1, c_1], [c_1 + 1, c_2], \ldots, [c_Z + 1, c_{Z+1}] \tag{53}$$

where $Z = \lfloor (T-1)/K \rfloor$, $K = \lceil m/2 \rceil$, $c_{Z+1} = T$, and $c_i = iK$ for $i = 0, \ldots, Z$. At each round t, we first 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 is α-strongly convex and satisfies Assumption 2 with $G = \alpha R$ over the set $\mathcal{K} = [0, R/\sqrt{d}]^d$. Then, let $\mathcal{B}_p$ denote the Bernoulli distribution with probability of obtaining 1 equal to p, and let $\mathbf{1}$ denote the all-ones vector in $\mathbb{R}^d$. For any $i \in \{0, \ldots, Z\}$ and $t \in [c_i + 1, c_{i+1}]$, we set

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

where $\mathbf{w}_i$ is sampled from the vector set $\{\mathbf{0}, \mathbf{1}\}$ according to $\mathcal{B}_p$, i.e., $\Pr(\mathbf{w}_i = \mathbf{1}) = p$. It is easy to verify that $h_i(\mathbf{x})$ also satisfies the definition of α-strongly convex functions and Assumption 2 with $G = \alpha R$ over the set $\mathcal{K} = [0, R/\sqrt{d}]^d$. Then, for any $i \in \{0, \ldots, Z\}$ and $t \in [c_i + 1, c_{i+1}]$, the global loss function in each round can be written as

$$\begin{aligned} f_t(\mathbf{x}) = &\sum_{j=1}^{n} f_{t,j}(\mathbf{x}) = \frac{\alpha(n-2K+1)}{2}\left\|\mathbf{x} - \frac{R\mathbf{w}_i}{\sqrt{d}}\right\|_2^2 + \frac{\alpha(2K-1)}{2}\|\mathbf{x}\|_2^2 \\ = &\frac{\alpha n}{2}\|\mathbf{x}\|_2^2 - \frac{\alpha(n-2K+1)R}{\sqrt{d}}\langle \mathbf{x}, \mathbf{w}_i\rangle + \frac{\alpha(n-2K+1)R^2}{2d}\|\mathbf{w}_i\|_2^2 \end{aligned}$$

whose expectation is

$$\begin{aligned} \mathbb{E}_{\mathbf{w}_i}[f_t(\mathbf{x})] = &\frac{\alpha n}{2}\|\mathbf{x}\|_2^2 - \frac{\alpha(n-2K+1)R}{\sqrt{d}}\langle \mathbf{x}, \mathbf{p}\rangle + \frac{\alpha(n-2K+1)R^2}{2d}\langle \mathbf{1}, \mathbf{p}\rangle \\ = &\frac{\alpha n}{2}\left\|\mathbf{x} - \frac{(n-2K+1)R\mathbf{p}}{n\sqrt{d}}\right\|_2^2 + \frac{\alpha(n-2K+1)R^2}{2d}\left\langle \mathbf{1} - \frac{(n-2K+1)\mathbf{p}}{n}, \mathbf{p}\right\rangle \end{aligned}$$

where $\mathbf{p}$ denotes the vector with the same value of p in each dimension. For brevity, let $F(\mathbf{x}) = \mathbb{E}_{\mathbf{w}_i}[f_t(\mathbf{x})]$. Because of $p \in [0,1]$ and $\mathcal{K} = [0, R/\sqrt{d}]^d$, this function can be simply minimized by

$$\mathbf{x}^* = \frac{(n-2K+1)R\mathbf{p}}{n\sqrt{d}} \in \mathcal{K}$$

which implies that any $\mathbf{x} \in \mathcal{K}$ has

$$F(\mathbf{x}) - F(\mathbf{x}^*) = \frac{\alpha n}{2} \left\| \mathbf{x} - \frac{(n-2K+1)R\mathbf{p}}{n\sqrt{d}} \right\|_2^2 \geq 0. \tag{54}$$

Moreover, it is not hard to verify that

$$\mathbb{E}_{\mathbf{w}_0,\ldots,\mathbf{w}_Z}\left[\min_{\mathbf{x}\in\mathcal{K}} \sum_{i=0}^{Z} \sum_{t=c_i+1}^{c_{i+1}} f_t(\mathbf{x})\right] \leq \mathbb{E}_{\mathbf{w}_0,\ldots,\mathbf{w}_Z}\left[\sum_{i=0}^{Z} \sum_{t=c_i+1}^{c_{i+1}} f_t(\mathbf{x}^*)\right] = \sum_{i=0}^{Z} \sum_{t=c_i+1}^{c_{i+1}} F(\mathbf{x}^*). \tag{55}$$

Following the proof of Theorem 3, we continue to 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 $\mathbf{w}_i$. Therefore, we have

$$\begin{aligned}
\mathbb{E}_{\mathbf{w}_0,\ldots,\mathbf{w}_Z}[\mathrm{R}(T,1)] =& \mathbb{E}_{\mathbf{w}_0,\ldots,\mathbf{w}_Z}\left[\sum_{i=0}^{Z} \sum_{t=c_i+1}^{c_{i+1}} f_t(\mathbf{x}_1(t)) - \min_{\mathbf{x}\in\mathcal{K}} \sum_{i=0}^{Z} \sum_{t=c_i+1}^{c_{i+1}} f_t(\mathbf{x})\right] \\
=& \mathbb{E}_{\mathbf{w}_0,\ldots,\mathbf{w}_Z}\left[\sum_{i=0}^{Z} \sum_{t=c_i+1}^{c_{i+1}} F(\mathbf{x}_1(t))\right] - \mathbb{E}_{\mathbf{w}_0,\ldots,\mathbf{w}_Z}\left[\min_{\mathbf{x}\in\mathcal{K}} \sum_{i=0}^{Z} \sum_{t=c_i+1}^{c_{i+1}} f_t(\mathbf{x})\right] \\
\overset{(55)}{\geq}& \mathbb{E}_{\mathbf{w}_0,\ldots,\mathbf{w}_Z}\left[\sum_{i=0}^{Z} \sum_{t=c_i+1}^{c_{i+1}} F(\mathbf{x}_1(t)) - \sum_{i=0}^{Z} \sum_{t=c_i+1}^{c_{i+1}} F(\mathbf{x}^*)\right].
\end{aligned} \tag{56}$$

To lower bound the right side of (56), we assume that the D-OCO algorithm is deterministic without loss of generality.[3] Note that all local functions $\{f_{t,1}(\mathbf{x}), \ldots, f_{t,n}(\mathbf{x})\}$ for any $t \in [c_i+1, c_{i+1}]$ are either deterministic or parameterized by the same random vector $\mathbf{w}_i$ drawn according to $\mathcal{B}_p$. Therefore, for any $t \in [c_i+1, c_{i+1}]$, $\mathbf{x}_1(t)$ actually can be specified by a bit string $X \in \{0,1\}^i$ drawn from $\mathcal{B}_p^i$, i.e., the product measure on $\{0,1\}^i$ induced by taking i independent trials from $\mathcal{B}_p$. To be precise, the local learner 1 of the D-OCO algorithm at any round $t \in [c_i+1, c_{i+1}]$ can be denoted as a mapping function $\mathcal{A}_t(\cdot) : \{0,1\}^i \mapsto \mathcal{K}$ such that $\mathbf{x}_1(t) = \mathcal{A}_t(X)$. Moreover, one should notice that the value p in the above procedures can be replaced by another value p', and the corresponding random vectors can be rewritten as $\mathbf{w}_0', \ldots, \mathbf{w}_Z'$ and $\mathbf{x}_1'(1), \ldots, \mathbf{x}_1'(T)$. Similarly, $\mathbf{x}_1'(t)$ for any round $t \in [c_i+1, c_{i+1}]$ can be specified by a bit string $X' \in \{0,1\}^i$ drawn from $\mathcal{B}_{p'}^i$, i.e., $\mathbf{x}_1'(t) = \mathcal{A}_t(X')$. Interestingly, following Hazan and Kale (2014), we can show that the expected instantaneous regret of the local learner 1 on at least one of the two distributions parameterized by appropriate p and p' must be large.

Lemma 7 *Fix an interval i and let $\epsilon \leq \frac{1}{32\sqrt{i+1}}$ be a parameter. Assume that $p, p' \in \left[\frac{1}{4}, \frac{3}{4}\right]$ such that $2\epsilon \leq |p - p'| \leq 4\epsilon$. Following the above notations, for any $t \in [c_i+1, c_{i+1}]$, we have*

$$\mathbb{E}_X\left[\|\mathcal{A}_t(X) - \xi\mathbf{p}\|_2^2\right] + \mathbb{E}_{X'}\left[\|\mathcal{A}_t(X') - \xi\mathbf{p}'\|_2^2\right] \geq \frac{d(\xi\epsilon)^2}{4}$$

where $\xi = (n-2K+1)R/(n\sqrt{d})$ and $\mathbf{p}'$ denotes the vector with the same value of p' in each dimension.

3. As in the lower bound analysis of Hazan and Kale (2014), even if the algorithm is randomized, it can be viewed as a deterministic one by fixing its random seed.

Let $M = \lfloor \log_{16}(15Z+16) - 1 \rfloor$, and it is not hard to verify that $M \geq 1$ due to $16n+1 \leq T$. To exploit the above lemma, we further divide the first $Z' = \frac{1}{15}(16^{M+1} - 16) < Z+1$ intervals into M epochs with the length $16, 16^2, \ldots, 16^M$. More specifically, the m-th epoch E_m consists of the intervals $\frac{1}{15}(16^m - 16), \ldots, \frac{1}{15}(16^{m+1} - 16) - 1$. Then, for these M epochs, we can prove the following lemma based on Lemma 7.

Lemma 8 *Following the notations used in Lemma 7, there exists a collection of nested intervals, $\left[\frac{1}{4}, \frac{3}{4}\right] \supseteq I_1 \supseteq I_2 \supseteq \cdots \supseteq I_M$, such that interval I_m corresponds to epoch m, with the property that I_m has length $4^{-(m+3)}$, and for every $p \in I_m$, we have*

$$\mathbb{E}_X \left[\|\mathcal{A}_t(X) - \xi \mathbf{p}\|_2^2 \right] \geq \frac{16^{-(m+3)} d\xi^2}{8}$$

over at least half the rounds t in intervals of epoch m.

From Lemma 8, there exists a value of $p \in \cap_{m \in [M]} I_m$ such that

$$\begin{aligned}
\mathbb{E}_{\mathbf{w}_0,\ldots,\mathbf{w}_Z}[\mathrm{R}(T,1)] \geq & \mathbb{E}_{\mathbf{w}_0,\ldots,\mathbf{w}_Z} \left[\sum_{i=0}^{Z} \sum_{t=c_i+1}^{c_{i+1}} \frac{\alpha n}{2} \left\| \mathbf{x}_1(t) - \frac{(n-2K+1)R\mathbf{p}}{n\sqrt{d}} \right\|_2^2 \right] \\
\geq & \mathbb{E}_{\mathbf{w}_0,\ldots,\mathbf{w}_Z} \left[\sum_{i=0}^{Z'-1} \sum_{t=c_i+1}^{c_{i+1}} \frac{\alpha n}{2} \left\| \mathbf{x}_1(t) - \frac{(n-2K+1)R\mathbf{p}}{n\sqrt{d}} \right\|_2^2 \right] \\
= & \mathbb{E}_{\mathbf{w}_0,\ldots,\mathbf{w}_Z} \left[\sum_{m=1}^{M} \sum_{i \in E_m} \sum_{t=c_i+1}^{c_{i+1}} \frac{\alpha n}{2} \left\| \mathbf{x}_1(t) - \frac{(n-2K+1)R\mathbf{p}}{n\sqrt{d}} \right\|_2^2 \right] \\
= & \sum_{m=1}^{M} \sum_{i \in E_m} \sum_{t=c_i+1}^{c_{i+1}} \mathbb{E}_X \left[\frac{\alpha n}{2} \left\| \mathcal{A}_t(X) - \frac{(n-2K+1)R\mathbf{p}}{n\sqrt{d}} \right\|_2^2 \right] \\
\geq & \sum_{m=1}^{M} \frac{\left(c_{\frac{1}{15}(16^{m+1}-16)} - c_{\frac{1}{15}(16^m-16)} \right) 16^{-(m+3)} \alpha (n-2K+1)^2 R^2}{32n} \\
= & \sum_{m=1}^{M} \frac{16^{-4} \alpha K (n-2K+1)^2 R^2}{2n} = \frac{16^{-4} \alpha M K (n-2K+1)^2 R^2}{2n}
\end{aligned} \tag{57}$$

where the first inequality is due to (54) and (56), and the third equality is due to $c_i = iK$ for any $i \leq Z$. Moreover, because of the definitions of M, Z, K, we have

$$\begin{aligned}
\frac{MK(n-2K+1)^2}{2n} \geq & \frac{(\log_{16}(15Z+16) - 2)m(n-m-1)^2}{4n} \\
= & \frac{(\log_{16}(15\lfloor (T-1)/K \rfloor + 16) - 2)(n-2)n}{32} \\
\geq & \frac{(\log_{16}(30(T-1)/n) - 2)(n-2)n}{32}.
\end{aligned} \tag{58}$$

By combining (57) and (58) with Lemma 5, there exists a particular choice of $\mathbf{w}_0, \ldots, \mathbf{w}_Z$ such that

$$\mathrm{R}(T,1) \geq \frac{16^{-5} \alpha \pi (\log_{16}(30(T-1)/n) - 2)(n-2)R^2}{4\sqrt{1-\sigma_2(P)}} \tag{59}$$

which completes the proof.

Additionally, we notice that it is also appealing to extend the lower bound in (59) into the setting with only C communication rounds. Following the proof of Theorem 3, a natural idea is to divide the total T rounds into $Z+1$ intervals defined in (45), rather than (53), and then repeat the proof steps outlined above. However, we want to emphasize that the last two equalities in (57) require the number of rounds in the first Z' intervals to be the same, which is not necessarily satisfied by (45). This is because the algorithm can allocate their C communication rounds arbitrarily. Therefore, instead of simply using this natural idea, new analytical tools may be required for the extension, which will be investigated in the future.

5.8 Proof of Lemma 7

This proof is heavily inspired by the proof of Lemma 16 in Hazan and Kale (2014), but requires specific modifications to generalize their result from one dimension to high dimensions. Specifically, we suppose that there exists an integer j such that

$$\mathbb{E}_X\left[(\mathcal{A}_{t,j}(X)-\xi p)^2\right]+\mathbb{E}_{X'}\left[\left(\mathcal{A}_{t,j}(X')-\xi p'\right)^2\right]<\frac{(\xi\epsilon)^2}{4} \tag{60}$$

where $\mathcal{A}_{t,j}(X)$ denotes the j-th element in the vector $\mathcal{A}_t(X)$. In the following, we will complete the proof by contradiction.

We first consider the case with $p'\geq p+2\epsilon$. Let $\Pr_p[\cdot]$ and $\Pr_{p'}[\cdot]$ denote the probability of an event under the distribution $\mathcal{B}_p^i$ and $\mathcal{B}_{p'}^i$, respectively. By combining (60) with Markov's inequality, we have

$$\Pr_p\left[(\mathcal{A}_{t,j}(X)-\xi p)^2<(\xi\epsilon)^2\right]\geq 3/4$$

which implies that

$$\Pr_p\left[\mathcal{A}_{t,j}(X)<\xi(p+\epsilon)\right]\geq 3/4. \tag{61}$$

Similarly, we can show that

$$\Pr_{p'}\left[\mathcal{A}_{t,j}(X')>\xi(p+\epsilon)\right]\geq\Pr_{p'}\left[\mathcal{A}_{t,j}(X')>\xi(p'-\epsilon)\right]\geq 3/4 \tag{62}$$

where the first inequality is due to $p'\geq p+2\epsilon$.

Now, we define an event

$$\mathcal{E}=\left\{Y\in\{0,1\}^i : A_{t,j}(Y)>\xi(p+\epsilon)\right\}.$$

It is easy to verify that

$$\left|\Pr_p[\mathcal{E}]-\Pr_{p'}[\mathcal{E}]\right|=\Pr_{p'}[\mathcal{E}]-\Pr_p[\mathcal{E}]>\frac{1}{2} \tag{63}$$

where both the equality and inequality are due to $\Pr_p[\mathcal{E}]<1/4$ derived from (61) and $\Pr_{p'}[\mathcal{E}]\geq 3/4$ derived from (62).

Moreover, let $\Delta_{\text{TV}}(\mathcal{B}_p^i,\mathcal{B}_{p'}^i)$ denote the total variation distance between the two distributions $\mathcal{B}_p^i$ and $\mathcal{B}_{p'}^i$ on the same probability space, i.e.,

$$\Delta_{\text{TV}}(\mathcal{B}_p^i,\mathcal{B}_{p'}^i)=\sup_{\mathcal{E}}\left|\Pr_p[\mathcal{E}]-\Pr_{p'}[\mathcal{E}]\right|.$$

Note that Hazan and Kale (2014) have provided an upper bound on $\Delta_{\text{TV}}(\mathcal{B}_p^i,\mathcal{B}_{p'}^i)$.

Lemma 9 *(Hazan and Kale, 2014, Lemma 15) Let $p, p' \in \left[\frac{1}{4}, \frac{3}{4}\right]$ such that $|p' - p| \leq 1/8$. Then, it holds that*

$$\Delta_{\mathrm{TV}}(\mathcal{B}_p^i, \mathcal{B}_{p'}^i) \leq 2\sqrt{(p'-p)^2 i}$$

for any integer $i \geq 0$.[4]

Recall that $|p - p'| \leq 4\epsilon \leq 1/8$. Then, from Lemma 9, we have

$$\left|\Pr_p[\mathcal{E}] - \Pr_{p'}[\mathcal{E}]\right| \leq \Delta_{\mathrm{TV}}(\mathcal{B}_p^i, \mathcal{B}_{p'}^i) \leq 2\sqrt{(p'-p)^2 i} \leq 2\sqrt{16\epsilon^2 i} \leq \frac{1}{4}$$

which contradicts (63), and thus implies that our assumption about (60) is unavailable. Additionally, it is easy to construct the same contradiction in the case with $p \geq p' + 2\epsilon$. Therefore, we have

$$\mathbb{E}_X\left[\left(\mathcal{A}_{t,j}(X) - \xi p\right)^2\right] + \mathbb{E}_{X'}\left[\left(\mathcal{A}_{t,j}(X') - \xi p'\right)^2\right] \geq \frac{(\xi\epsilon)^2}{4} \tag{64}$$

for any $j \in [d]$. Finally, by summing both sides of (64) over $j \in [d]$, we complete this proof.

5.9 Proof of Lemma 8

This lemma can be proved by slightly modifying the proof of Lemma 19 in Hazan and Kale (2014). Here, we include the detailed proof for the completeness. Following Hazan and Kale (2014), we will iteratively build the required interval I_m for $m = 1, \ldots, M$. Specifically, we first select an arbitrary interval $I_0 = \left[a, a + 4^{-4}\right]$ of length 4^{-3} inside $\left[\frac{1}{4}, \frac{3}{4}\right]$. To find the required interval for $m = 1$, we divide I_{m-1} into four equal quarters of length $4^{-(m+3)}$, and show that either the first quarter $Q_1 = \left[a, a + 4^{-(m+3)}\right]$ or the last quarter $Q_4 = \left[a + 3 \cdot 4^{-(m+3)}, a + 4 \cdot 4^{-(m+3)}\right]$ is a valid choice for I_m.

To this end, we suppose that Q_1 is not a valid choice for I_m, i.e., there exist some $p \in Q_1$ such that it holds

$$\mathbb{E}_X\left[\|\mathcal{A}_t(X) - \xi\mathbf{p}\|_2^2\right] < \frac{16^{-(m+3)} d\xi^2}{8}$$

for more than half the rounds t in intervals of epoch m. Then, we define the following set

$$H = \bigcup_{i \in E_m}\left\{t \in [c_i + 1, c_{i+1}] : \mathbb{E}_X\left[\|\mathcal{A}_t(X) - \xi\mathbf{p}\|_2^2\right] < \frac{16^{-(m+3)} d\xi^2}{8}\right\} \tag{65}$$

where $E_m = \left\{\frac{1}{15}(16^m - 16), \ldots, \frac{1}{15}(16^{m+1} - 16) - 1\right\}$. In the following, we proceed to prove that for all $p' \in Q_4$ and $t \in H$, the following inequality must hold

$$\mathbb{E}_{X'}\left[\left\|\mathcal{A}_t(X') - \xi\mathbf{p}'\right\|_2^2\right] \geq \frac{16^{-(m+3)} d\xi^2}{8} \tag{66}$$

which implies that Q_4 is a valid choice for I_m because the set H contains more than half the rounds in intervals of epoch m.

To be precise, we fix any $p' \in Q_4$ and $t \in H$, where t must belong to $[c_i + 1, c_{i+1}]$ for some $i \in E_m$. Let $\epsilon = 4^{-(m+3)}$. It is easy to verify that $4(i + 1) \leq 16^{m+1}$ due to the definition of E_m,

4. Following the errata of Hazan and Kale (2014), we have used the correct constant 2 in the upper bound.

and we thus have $\epsilon \leq 1/(32\sqrt{i+1})$. Additionally, we have $2\epsilon \leq |p-p'| \leq 4\epsilon$ due to $p \in Q_1$ and $p' \in Q_4$. Therefore, from Lemma 7, we have

$$\mathbb{E}_X\left[\|\mathcal{A}_t(X)-\xi\mathbf{p}\|_2^2\right]+\mathbb{E}_{X'}\left[\|\mathcal{A}_t(X')-\xi\mathbf{p}'\|_2^2\right] \geq \frac{16^{-(m+3)}d\xi^2}{4}. \tag{67}$$

Then, the previously mentioned (66) can be simply derived by combining (65) and (67). Finally, it is worth noting that for any $m=2,\ldots,M$, the required interval I_m can be built one by one by starting the division from the valid I_{m-1} and repeating the above procedures.

5.10 Proof of Theorem 5

Following the definitions of $\bar{\mathbf{x}}(z)$, $\bar{\mathbf{z}}(z)$, $\bar{\mathbf{d}}(z)$, and $\ell_z(\mathbf{x})$ in the proof of Theorem 1, we only need to analyze the distance $\|\mathbf{x}_i(z)-\bar{\mathbf{x}}(z+1)\|_2$. To this end, we first define

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

for any $z=2,\ldots,T/L$, which is exactly the same as $\mathbf{x}_i(z)$ generated by Algorithm 1. Note that the distance $\|\mathbf{x}_i^*(z)-\bar{\mathbf{x}}(z+1)\|_2$ has been analyzed in the proof of Theorem 1. However, due to the use of CG, $\mathbf{x}_i(z)$ generated by Algorithm 3 is only an approximation of $\mathbf{x}_i^*(z)$. Moreover, according to Algorithm 3, now $\mathbf{x}_i(z)$ is computed based on $\mathbf{z}_i(z-2)$, rather than $\mathbf{z}_i(z-1)$. As a result, we first upper bound $\|\mathbf{x}_i(z)-\bar{\mathbf{x}}(z+1)\|_2$ as

$$\begin{aligned}\|\mathbf{x}_i(z)-\bar{\mathbf{x}}(z+1)\|_2 \leq&\|\mathbf{x}_i(z)-\mathbf{x}_i^*(z-1)\|_2+\|\mathbf{x}_i^*(z-1)-\bar{\mathbf{x}}(z+1)\|_2\\ \leq&\|\mathbf{x}_i(z)-\mathbf{x}_i^*(z-1)\|_2+\|\mathbf{x}_i^*(z-1)-\bar{\mathbf{x}}(z)\|_2\\ &+\|\bar{\mathbf{x}}(z)-\bar{\mathbf{x}}(z+1)\|_2\end{aligned} \tag{68}$$

for any $z=3,\ldots,T/L$.

To bound the first term in the right side of (68), we introduce the following lemma regarding the convergence property of CG.

Lemma 10 *(Jaggi, 2013, Theorem 1) If $F(\mathbf{x}):\mathcal{K}\mapsto\mathbb{R}$ is a convex and β-smooth function, and $\|\mathbf{x}\|_2\leq R$ for any $\mathbf{x}\in\mathcal{K}$, Algorithm 2 with $K\geq 1$ ensures*

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

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

It is easy to verify that $F_{z-1,i}(\mathbf{x})$ defined in Algorithm 3 is $((z-3)L\alpha+2h)$-strongly convex and $((z-3)L\alpha+2h)$-smooth over $\mathcal{K}$, for any $z=3,\ldots,T/L$. Moreover, due to the above definition, we have $\mathbf{x}_i^*(z-1)=\operatorname{argmin}_{\mathbf{x}\in\mathcal{K}}F_{z-1,i}(\mathbf{x})$. Then, because of $\mathbf{x}_i(z)=\text{CG}(\mathcal{K},L,F_{z-1,i}(\mathbf{x}),\mathbf{x}_i(z-1))$ and Lemma 10, for any $z=3,\ldots,T/L$, we have

$$F_{z-1,i}(\mathbf{x}_i(z))-F_{z-1,i}(\mathbf{x}_i^*(z-1))\leq\frac{8((z-3)L\alpha+2h)R^2}{L+2}. \tag{69}$$

By combining (36) with (69), for any $z=3,\ldots,T/L$, we further have

$$\|\mathbf{x}_i(z)-\mathbf{x}_i^*(z-1)\|_2\leq\sqrt{\frac{2\left(F_{z-1,i}(\mathbf{x}_i(z))-F_{z-1,i}(\mathbf{x}_i^*(z-1))\right)}{(z-3)L\alpha+2h}}\leq\frac{4R}{\sqrt{L+2}}. \tag{70}$$

Then, to bound the second term in the right side of (68), we recall that θ and L' used in Algorithm 3 equal to θ and L used in Algorithm 1, respectively. Following the proof of Lemma 2, it is easy to verify that that Algorithm 3 also enjoys the error bound presented in Lemma 2, i.e.,

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

for any $i \in [n]$ and $z \in [T/L]$, though L in Algorithm 3 is different from L'. Therefore, we can directly use (39) in the proof of Theorem 1 to derive the following upper bound

$$\|\mathbf{x}_i^*(z-1) - \bar{\mathbf{x}}(z)\|_2 \leq \frac{3L(G+\alpha R)}{(z-3)L\alpha + 2h} + \frac{4L(G+2\alpha R)}{(z-1)L\alpha + 2h}. \tag{71}$$

for any $z = 3, \ldots, T/L$. Moreover, it is not hard to verify that the third term in the right side of (68) can be bounded by (38) in the proof of Theorem 1.

Substituting (70), (71), and (38) into (68), for any $z = 3, \ldots, T/L$, we have

$$\begin{aligned}
&\|\mathbf{x}_i(z) - \bar{\mathbf{x}}(z+1)\|_2 \\
\leq&\frac{4R}{\sqrt{L+2}} + \frac{3L(G+\alpha R)}{(z-3)L\alpha + 2h} + \frac{4L(G+2\alpha R)}{(z-1)L\alpha + 2h} + \frac{2L(G+2\alpha R)}{zL\alpha + 2h} \\
\leq&\frac{4R}{\sqrt{L+2}} + \frac{3L(G+\alpha R)}{(z-3)L\alpha + 2h} + \frac{6L(G+2\alpha R)}{(z-1)L\alpha + 2h}.
\end{aligned}$$

Moreover, because of $\mathbf{x}_i(1) = \mathbf{x}_i(2) = \bar{\mathbf{x}}(1) = \mathbf{0}$, it is easy to verify that

$$\|\mathbf{x}_i(z) - \bar{\mathbf{x}}(z+1)\|_2 = \|\bar{\mathbf{x}}(z) - \bar{\mathbf{x}}(z+1)\|_2 \overset{(38)}{\leq} \frac{2L(G+2\alpha R)}{zL\alpha + 2h}$$

for $z = 1$, and

$$\|\mathbf{x}_i(z) - \bar{\mathbf{x}}(z+1)\|_2 = \|\bar{\mathbf{x}}(z-1) - \bar{\mathbf{x}}(z+1)\|_2 \overset{(38)}{\leq} \frac{4L(G+2\alpha R)}{zL\alpha + 2h}.$$

for $z = 2$. Finally, following (43) in the proof of Theorem 1, for any $\mathbf{x} \in \mathcal{K}$, it is easy to verify that

$$\begin{aligned}
&\sum_{z=1}^{T/L}\sum_{t\in\mathcal{T}_z}\sum_{j=1}^{n} f_{t,j}(\mathbf{x}_i(z)) - \sum_{z=1}^{T/L}\sum_{t\in\mathcal{T}_z}\sum_{j=1}^{n} f_{t,j}(\mathbf{x}) \\
\leq& nhR^2 + \frac{12nGRT}{\sqrt{L+2}} + 3nLG\left(\sum_{z=3}^{T/L}\frac{3L(G+\alpha R)}{(z-3)L\alpha + 2h} + \sum_{z=1}^{T/L}\frac{6L(G+2\alpha R)}{(z-1)L\alpha + 2h}\right).
\end{aligned}$$

5.11 Proof of Corollaries 3 and 4

By substituting $\alpha = 0$ into (17), we have

$$\begin{aligned}
\mathrm{R}(T,i) \leq& 3nLG\left(\sum_{z=3}^{T/L}\frac{3LG}{2h} + \sum_{z=1}^{T/L}\frac{3LG}{h}\right) + nhR^2 + \frac{12nGRT}{\sqrt{L+2}} \\
\leq&\frac{14nLG^2T}{h} + nhR^2 + \frac{12nGRT}{\sqrt{L+2}}.
\end{aligned} \tag{72}$$

Then, by substituting $h = \sqrt{14LT}G/R$ and $L = \sqrt{T}L'$ into (72), we have

$$\mathrm{R}(T,i) \leq 2\sqrt{14}nGR\sqrt{L'}T^{3/4} + \frac{12nGRT^{3/4}}{\sqrt{L'}}. \tag{73}$$

If $L' \leq \sqrt{T}$, by substituting $h = \sqrt{14LT}G/R$ and $L = \sqrt{T}$ into (72), we have

$$\mathrm{R}(T,i) \leq (2\sqrt{14} + 12)nGRT^{3/4}. \tag{74}$$

From (73) and (74), the proof of Corollary 3 is completed.

To prove Corollary 4, we first substitute $h = \alpha L$ into (17) to derive the following regret bound

$$\begin{aligned}
\mathrm{R}(T,i) \leq & 3nLG\left(\sum_{z=3}^{T/L}\frac{3(G+\alpha R)}{(z-1)\alpha} + \sum_{z=1}^{T/L}\frac{6(G+2\alpha R)}{(z+1)\alpha}\right) + n\alpha LR^2 + \frac{12nGRT}{\sqrt{L+2}} \\
\leq & \frac{3nLG(9G+15\alpha R)}{\alpha}\sum_{z=1}^{T/L}\frac{1}{z} + n\alpha LR^2 + \frac{12nGRT}{\sqrt{L+2}} \\
\leq & \frac{3nLG(9G+15\alpha R)(1+\ln(T/L))}{\alpha} + n\alpha LR^2 + \frac{12nGRT}{\sqrt{L+2}}.
\end{aligned} \tag{75}$$

Then, by substituting $L = T^{2/3}(\ln T)^{-2/3}L'$ into (75), we have

$$\begin{aligned}
\mathrm{R}(T,i) \leq & \frac{3nG(9G+15\alpha R)T^{2/3}L'((\ln T)^{-2/3} + (\ln T)^{1/3})}{\alpha} \\
& + n\alpha T^{2/3}L'R^2(\ln T)^{-2/3} + \frac{12nGRT^{2/3}(\ln T)^{1/3}}{\sqrt{L'}}.
\end{aligned} \tag{76}$$

Moreover, if $L' \leq T^{2/3}(\ln T)^{-2/3}$, by substituting $L = T^{2/3}(\ln T)^{-2/3}$ into (75), we have

$$\begin{aligned}
\mathrm{R}(T,i) \leq & \frac{3nG(9G+15\alpha R)T^{2/3}((\ln T)^{-2/3} + (\ln T)^{1/3})}{\alpha} \\
& + n\alpha T^{2/3}R^2(\ln T)^{-2/3} + 12nGRT^{2/3}(\ln T)^{1/3}.
\end{aligned} \tag{77}$$

From (76) and (77), the proof of Corollary 4 is completed.

6. Conclusion

This paper investigates D-OCO with convex and strongly convex functions, and aims to develop optimal and efficient algorithms. To this end, we first propose a novel D-OCO algorithm, namely AD-FTGL, which reduces 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 to $\tilde{O}(n\rho^{-1/4}\sqrt{T})$ and $\tilde{O}(n\rho^{-1/2}\log T)$, respectively. Furthermore, we demonstrate its optimality for D-OCO by deriving $\Omega(n\rho^{-1/4}\sqrt{T})$ and $\Omega(n\rho^{-1/2}\log T)$ lower bounds for convex and strongly convex functions, respectively. Finally, to efficiently handle complex constraints, we propose a projection-free variant of AD-FTGL, which can respectively achieve $O(nT^{3/4})$ and $O(nT^{2/3}(\log T)^{1/3})$ regret bounds for convex and strongly convex functions with only $\tilde{O}(\rho^{-1/2}\sqrt{T})$ and $\tilde{O}(\rho^{-1/2}T^{1/3}(\log T)^{2/3})$ communication rounds. Although these regret bounds cannot match the aforementioned lower bounds, they are much tighter than those of existing projection-free algorithms for D-OCO. Moreover, we provide communication-dependent lower bounds to demonstrate that the number of communication rounds required by our projection-free algorithm is nearly optimal for achieving these regret bounds.

Acknowledgments

This work was partially supported by National Natural Science Foundation of China (62306275), Zhejiang Provincial Natural Science Foundation of China (LMS25F030002), Yongjiang Talent Introduction Programme (2023A-193-G), and Ningbo Natural Science Foundation (2024J206). The authors are grateful for the anonymous reviewers and the editor for their helpful comments.

Appendix A. Revisiting D-FTGL

Wan et al. (2022) originally develop a projection-free version of D-FTGL for α-strongly convex functions, which also adopts the blocking update mechanism to select the decision of each learner. In this section, we first discuss how to simplify the projection-free version into D-FTGL, and then provide a refined analysis of its regret.

A.1 The Algorithm

Following the notations in our Algorithm 1, at the end of each block $z = 1, \ldots, T/L$, each learner i of their algorithm updates as

$$\begin{aligned} \mathbf{z}_i(z+1) &= \sum_{j \in N_i} P_{ij}\mathbf{z}_j(z) + \mathbf{d}_i(z) \\ \mathbf{x}_i(z+1) &= \mathrm{CG}(\mathcal{K}, K, F_{z,i}(\mathbf{x}), \mathbf{x}_i(z)) \end{aligned} \tag{78}$$

where $F_{z,i}(\mathbf{x}) = \langle \mathbf{z}_i(z), \mathbf{x}\rangle + \frac{(z-1)L\alpha}{2}\|\mathbf{x}\|_2^2 + h\|\mathbf{x}\|_2^2$, and $\mathbf{x}_i(z+1)$ is computed by using CG (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 2 of Wan et al. (2022), under Assumptions 1, 2, 3, and 4, their algorithm can achieve the following regret bound

$$\mathrm{R}(T,i) \leq \frac{12nGRT}{\sqrt{K+2}} + \sum_{z=2}^{T/L} \frac{3nGL^2(G+\alpha R)\sqrt{n}}{((z-2)\alpha L + 2h)(1-\sigma_2(P))} + \sum_{z=1}^{T/L} \frac{4nL^2(G+2\alpha R)^2}{z\alpha L + 2h} + 4nhR^2. \tag{79}$$

By substituting $K = L = \sqrt{T}$, $\alpha = 0$, and $h = n^{1/4}T^{3/4}G/(\sqrt{\rho}R)$ into (79), they derive a regret bound of $O(n^{5/4}\rho^{-1/2}T^{3/4})$ for convex functions. Moreover, by substituting $K = L = T^{2/3}(\ln T)^{-2/3}$ and $h = \alpha L$ into (79), they derive a regret bound of $O(n^{3/2}\rho^{-1}T^{2/3}(\log T)^{1/3})$ for strongly convex functions.

However, these choices of K and L are used for achieving the projection-free property, i.e., only one linear optimization step is used per round on average. Actually, it is easy to derive the $O(n^{5/4}\rho^{-1/2}\sqrt{T})$ regret bound for convex functions by substituting $K = \infty$, $L = 1$, and $h = n^{1/4}\sqrt{T}G/(\sqrt{\rho}R)$, and the $O(n^{3/2}\rho^{-1}\log T)$ regret bound for strongly convex functions by substituting $K = \infty$, $L = 1$, and $h = \alpha$ into (79). With the new choice of K and L, the algorithm of Wan et al. (2022) reduces to performing the following update

$$\begin{aligned} \mathbf{z}_i(t+1) &= \sum_{j \in N_i} P_{ij}\mathbf{z}_j(t) + (\nabla f_{t,i}(\mathbf{x}_i(t)) - \alpha\mathbf{x}_i(t)) \\ \mathbf{x}_i(t+1) &= \operatorname*{argmin}_{\mathbf{x} \in \mathcal{K}} \langle \mathbf{z}_i(t), \mathbf{x}\rangle + \frac{(t-1)\alpha}{2}\|\mathbf{x}\|_2^2 + h\|\mathbf{x}\|_2^2 \end{aligned} \tag{80}$$

Algorithm 4 D-FTGL

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

for each learner i at round t. Additionally, it is worth noting that the main reason for computing $\mathbf{x}_i(z+1)$ in (78) based on $\mathbf{z}_i(z)$ is to allocate K iterations of CG to every round in block z, rather than only the last round in the block. In contrast, here computing $\mathbf{x}_i(t+1)$ based on $\mathbf{z}_i(t)$ provides no benefit because of $L = 1$. Thus, (80) can be further simplified to D-FTGL, as given in (5). For the sake of completeness, we summarize the detailed procedure in Algorithm 4.

A.2 Theoretical Guarantees

Although D-FTGL is slightly different from (80), it is easy to verify that D-FTGL can also achieve the aforementioned regret bounds for convex and strongly convex functions. To be precise, we first establish the following guarantee for D-FTGL.

Theorem 6 *Let C denote an upper bound of the approximate error of the standard gossip step in Algorithm 4, i.e., $\|\mathbf{z}_i(t) - \bar{\mathbf{z}}(t)\|_2 \leq C$ for any $t \in [T]$ and $i \in [n]$, where $\bar{\mathbf{z}}(t) = \frac{1}{n}\sum_{i=1}^n \mathbf{z}_i(t)$. Under Assumptions 1, 2, 3, and 4, for any $i \in [n]$, Algorithm 4 ensures*

$$\mathrm{R}(T,i) \leq 3nG\left(\sum_{t=2}^{T} \frac{C}{(t-1)\alpha + 2h} + \sum_{t=1}^{T} \frac{2(G+2\alpha R)}{t\alpha + 2h}\right) + nhR^2. \tag{81}$$

Note that Lemma 3 of Wan et al. (2022) has provided an error bound of $C = \sqrt{n}(G + \alpha R)\rho^{-1}$ for the standard gossip step. By substituting this bound into (81), we can set $h = n^{1/4}\sqrt{T}G/(\sqrt{\rho}R)$ and $\alpha = 0$ to achieve the $O(n^{5/4}\rho^{-1/2}\sqrt{T})$ regret bound for convex functions, and simply set $h = 0$ to achieve the $O(n^{3/2}\rho^{-1}\log T)$ regret bound for strongly convex functions.

More importantly, we want to emphasize that the power of D-FTGL extends beyond recovering existing results. In the following, we provide a novel and improved error bound for the standard gossip step in D-FTGL, and further derive tighter regret bounds.

Lemma 11 *For any $i \in [n]$, let $\nabla_i(1), \ldots, \nabla_i(T) \in \mathbb{R}^d$ be a sequence of vectors. Let $\mathbf{z}_i(1) = \mathbf{0}$, $\mathbf{z}_i(t+1) = \sum_{j \in N_i} P_{ij}\mathbf{z}_j(t) + \nabla_i(t)$, and $\bar{\mathbf{z}}(t) = \frac{1}{n}\sum_{i=1}^n \mathbf{z}_i(t)$ for any $t \in [T]$, where P satisfies Assumption 1. For any $i \in [n]$ and $t \in [T]$, assuming $\|\nabla_i(t)\|_2 \leq \xi$ where $\xi > 0$ is a constant, we have*

$$\|\mathbf{z}_i(t) - \bar{\mathbf{z}}(t)\|_2 \leq 2\xi\left(\frac{1 + \ln(\sqrt{n})}{1 - \sigma_2(P)} + 1\right).$$

This lemma is inspired by an existing error bound of $O(\rho^{-1}\log(nT))$ for the standard gossip step in decentralized offline optimization (Duchi et al., 2011). However, possibly because of the additional dependence on $\log T$, it is overlooked in D-OCO, where T could be very large and even the $\log T$ factor is unacceptable. In contrast, Lemma 11 provides an $O(\rho^{-1}\log n)$ error bound without the $\log T$ factor. Moreover, compared with the $O(\sqrt{n}\rho^{-1})$ error bound in previous studies on D-OCO (Hosseini et al., 2013; Wan et al., 2022), our error bound has a much tighter dependence on n. Then, by substituting $C = O(\rho^{-1}\log n)$, $h = G\sqrt{T\ln n}/(\sqrt{\rho}R)$, and $\alpha = 0$ into (81), we can achieve an $O(n\rho^{-1/2}\sqrt{T\log n})$ regret bound for convex functions. Recall that D-FTGL with $\alpha = 0$ reduces to D-FTRL (Hosseini et al., 2013). Here, the improved error bound allows us to tune the parameter h better, and thus improve the existing $O(n^{5/4}\rho^{-1/2}\sqrt{T})$ regret bound. If functions are strongly convex, we can substitute $C = O(\rho^{-1}\log n)$ and $h = 0$ into (81) to derive an $O(n\rho^{-1}(\log n)\log T)$ regret bound, which is also tighter than the existing bound.

Remark 4 It is worth noting that besides D-FTGL, other existing D-OCO algorithms with a similar use of the standard gossip step can also benefit from our improved error bound in Lemma 11. For example, it is easy to verify that the $O(n^{5/4}\rho^{-1/2}T^{3/4})$ and $O(n^{3/2}\rho^{-1}T^{2/3}(\log T)^{1/3})$ regret bounds of the projection-free algorithm in Wan et al. (2022) can be reduced to $O(n\rho^{-1/2}T^{3/4}\sqrt{\log n})$ and $O(n\rho^{-1}T^{2/3}(\log T)^{1/3}\log n)$, respectively.

A.3 Analysis

In the following, we provide the detailed proofs of Theorem 6 and Lemma 11.

A.3.1 Proof of Theorem 6

We start this proof by defining a virtual decision for any $t \in [T+1]$ as

$$\bar{\mathbf{x}}(t) = \operatorname*{argmin}_{\mathbf{x}\in\mathcal{K}} \langle \mathbf{x}, \bar{\mathbf{z}}(t)\rangle + \frac{(t-1)\alpha}{2}\|\mathbf{x}\|_2^2 + h\|\mathbf{x}\|_2^2. \tag{82}$$

For brevity, let $\mathbf{d}_i(t) = \nabla f_{t,i}(\mathbf{x}_i(t)) - \alpha\mathbf{x}_i(t)$ and $\bar{\mathbf{d}}(t) = \frac{1}{n}\sum_{i=1}^n \mathbf{d}_i(t)$. In the following, we will bound the regret of any learner i by analyzing the regret of $\bar{\mathbf{x}}(2), \ldots, \bar{\mathbf{x}}(T+1)$ on a sequence of loss functions defined by $\bar{\mathbf{d}}(1), \ldots, \bar{\mathbf{d}}(T)$ and the distance $\|\mathbf{x}_i(t) - \bar{\mathbf{x}}(t+1)\|_2$ for any $t \in [T]$. To this end, we notice that

$$\begin{aligned}
\bar{\mathbf{z}}(t+1) =& \frac{1}{n}\sum_{i=1}^n \mathbf{z}_i(t+1) = \frac{1}{n}\sum_{i=1}^n \left(\sum_{j\in N_i} P_{ij}\mathbf{z}_j(t) + \mathbf{d}_i(t)\right) \\
=& \frac{1}{n}\sum_{i=1}^n\sum_{j=1}^n P_{ij}\mathbf{z}_j(t) + \bar{\mathbf{d}}(t) = \frac{1}{n}\sum_{j=1}^n\sum_{i=1}^n P_{ij}\mathbf{z}_j(t) + \bar{\mathbf{d}}(t) \\
=& \bar{\mathbf{z}}(t) + \bar{\mathbf{d}}(t) = \sum_{i=1}^t \bar{\mathbf{d}}(i)
\end{aligned} \tag{83}$$

where the third and fifth equalities are due to Assumption 1.

Then, let $\ell_t(\mathbf{x}) = \langle \mathbf{x}, \bar{\mathbf{d}}(t)\rangle + \frac{\alpha}{2}\|\mathbf{x}\|_2^2$ for any $t \in [T]$. By combining Lemma 3 with (82) and (83), for any $\mathbf{x} \in \mathcal{K}$, it is easy to verify that

$$\sum_{t=1}^T \ell_t(\bar{\mathbf{x}}(t+1)) - \sum_{t=1}^T \ell_t(\mathbf{x}) \le h\left(\|\mathbf{x}\|_2^2 - \|\bar{\mathbf{x}}(2)\|_2^2\right) \le hR^2 \tag{84}$$

where the last inequality is due to Assumption 3 and $\|\bar{\mathbf{x}}(2)\|_2^2 \geq 0$.

We also notice that for any $t \in [T]$, Algorithm 4 ensures

$$\mathbf{x}_i(t) = \operatorname*{argmin}_{\mathbf{x}\in\mathcal{K}} \langle \mathbf{x}, \mathbf{z}_i(t)\rangle + \frac{(t-1)\alpha}{2}\|\mathbf{x}\|_2^2 + h\|\mathbf{x}\|_2^2. \tag{85}$$

By combining Lemma 4 with (82) and (85), for any $t \in [T]$, we have

$$\|\mathbf{x}_i(t) - \bar{\mathbf{x}}(t)\|_2 \leq \frac{\|\mathbf{z}_i(t) - \bar{\mathbf{z}}(t)\|_2}{(t-1)\alpha + 2h} \leq \frac{C}{(t-1)\alpha + 2h} \tag{86}$$

where the last inequality is due to the definition of C.

To bound $\|\mathbf{x}_i(t) - \bar{\mathbf{x}}(t+1)\|_2$, we still need to analyze the term $\|\bar{\mathbf{x}}(t) - \bar{\mathbf{x}}(t+1)\|_2$ for any $t \in [T]$. Let $F_t(\mathbf{x}) = \sum_{\tau=1}^t \ell_\tau(\mathbf{x}) + h\|\mathbf{x}\|_2^2$ for any $t \in [T]$. It is easy to verify that $F_t(\mathbf{x})$ is $(t\alpha + 2h)$-strongly convex over $\mathcal{K}$, and $\bar{\mathbf{x}}(t+1) = \operatorname{argmin}_{\mathbf{x}\in\mathcal{K}} F_t(\mathbf{x})$. Moreover, for any $\mathbf{x}, \mathbf{y} \in \mathcal{K}$, we have

$$\begin{aligned}
|\ell_t(\mathbf{x}) - \ell_t(\mathbf{y})| \leq& |\langle \nabla \ell_t(\mathbf{x}), \mathbf{x} - \mathbf{y}\rangle| \leq \|\nabla \ell_t(\mathbf{x})\|_2 \|\mathbf{x} - \mathbf{y}\|_2 \\
=& \|\bar{\mathbf{d}}(t) + \alpha\mathbf{x}\|_2 \|\mathbf{x} - \mathbf{y}\|_2 \leq \left(\left\| \frac{1}{n}\sum_{i=1}^n \mathbf{d}_i(t) \right\|_2 + \|\alpha\mathbf{x}\|_2 \right) \|\mathbf{x} - \mathbf{y}\|_2 \\
\leq& (G + 2\alpha R)\|\mathbf{x} - \mathbf{y}\|_2.
\end{aligned}$$

By combining the strong convexity of $F_t(\mathbf{x})$ with the above inequality, for any $t \in [T]$, we have

$$\begin{aligned}
\|\bar{\mathbf{x}}(t) - \bar{\mathbf{x}}(t+1)\|_2^2 \overset{(36)}{\leq}& \frac{2}{t\alpha + 2h}\left(F_t(\bar{\mathbf{x}}(t)) - F_t(\bar{\mathbf{x}}(t+1))\right) \\
=& \frac{2}{t\alpha + 2h}\left(F_{t-1}(\bar{\mathbf{x}}(t)) - F_{t-1}(\bar{\mathbf{x}}(t+1)) + \ell_t(\bar{\mathbf{x}}(t)) - \ell_t(\bar{\mathbf{x}}(t+1))\right) \\
\overset{(82)}{\leq}& \frac{2(\ell_t(\bar{\mathbf{x}}(t)) - \ell_t(\bar{\mathbf{x}}(t+1)))}{t\alpha + 2h} \leq \frac{2(G + 2\alpha R)\|\bar{\mathbf{x}}(t) - \bar{\mathbf{x}}(t+1)\|_2}{t\alpha + 2h}
\end{aligned}$$

which can be further simplified to

$$\|\bar{\mathbf{x}}(t) - \bar{\mathbf{x}}(t+1)\|_2 \leq \frac{2(G + 2\alpha R)}{t\alpha + 2h}. \tag{87}$$

By combining (86) and (87), for any $t = 2, \ldots, T$, we have

$$\begin{aligned}
\|\mathbf{x}_i(t) - \bar{\mathbf{x}}(t+1)\|_2 \leq& \|\mathbf{x}_i(t) - \bar{\mathbf{x}}(t)\|_2 + \|\bar{\mathbf{x}}(t) - \bar{\mathbf{x}}(t+1)\|_2 \\
\leq& \frac{C}{(t-1)\alpha + 2h} + \frac{2(G + 2\alpha R)}{t\alpha + 2h}.
\end{aligned} \tag{88}$$

For $t = 1$, it is worth noting that $\mathbf{x}_i(1) = \bar{\mathbf{x}}(1) = \mathbf{0}$, which ensures

$$\|\mathbf{x}_i(1) - \bar{\mathbf{x}}(2)\|_2 = \|\bar{\mathbf{x}}(1) - \bar{\mathbf{x}}(2)\|_2 \overset{(87)}{\leq} \frac{2(G + 2\alpha R)}{\alpha + 2h}. \tag{89}$$

Now, we are ready to derive the regret bound of any learner i. For brevity, let ϵ_t denote the upper bound of $\|\mathbf{x}_i(t) - \bar{\mathbf{x}}(t+1)\|_2$ derived in (88) and (89). Similar to (41), for any $t \in [T]$, $j \in [n]$, and $\mathbf{x} \in \mathcal{K}$, it is easy to verify that

$$f_{t,j}(\mathbf{x}_i(t)) - f_{t,j}(\mathbf{x}) \leq \langle \nabla f_{t,j}(\mathbf{x}_j(t)), \bar{\mathbf{x}}(t+1) - \mathbf{x}\rangle - \frac{\alpha}{2}\|\mathbf{x}_j(t) - \mathbf{x}\|_2^2 + 3G\epsilon_t.$$

By combining the above inequality with (42), for any $t \in [T]$, $j \in [n]$, and $\mathbf{x} \in \mathcal{K}$, we have

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

Finally, from the above inequality and the definition of ϵ_t, for any $\mathbf{x} \in \mathcal{K}$, it is not hard to verify that

$$
\begin{aligned}
&\sum_{t=1}^{T}\sum_{j=1}^{n} f_{t,j}(\mathbf{x}_i(t)) - \sum_{t=1}^{T}\sum_{j=1}^{n} f_{t,j}(\mathbf{x}) \\
\leq&\sum_{t=1}^{T}\sum_{j=1}^{n}\left(\langle \nabla f_{t,j}(\mathbf{x}_j(t)) - \alpha\mathbf{x}_j(t), \bar{\mathbf{x}}(t+1) - \mathbf{x}\rangle + \frac{\alpha}{2}\left(\|\bar{\mathbf{x}}(t+1)\|_2^2 - \|\mathbf{x}\|_2^2\right) + 3G\epsilon_t\right) \\
=&n\sum_{t=1}^{T}\left(\langle \bar{\mathbf{d}}(t), \bar{\mathbf{x}}(t+1) - \mathbf{x}\rangle + \frac{\alpha}{2}\left(\|\bar{\mathbf{x}}(t+1)\|_2^2 - \|\mathbf{x}\|_2^2\right)\right) + 3nG\sum_{t=1}^{T}\epsilon_t \\
\overset{(84)}{\leq}&nhR^2 + 3nG\left(\sum_{t=2}^{T}\frac{C}{(t-1)\alpha + 2h} + \sum_{t=1}^{T}\frac{2(G+2\alpha R)}{t\alpha + 2h}\right).
\end{aligned}
$$

A.3.2 PROOF OF LEMMA 11

This lemma is derived by refining the existing analysis for the standard gossip step (Duchi et al., 2011; Hosseini et al., 2013; Zhang et al., 2017; Wan et al., 2022). Let P^s denote the s-th power of P and P_{ij}^s denote the j-th entry of the i-th row in P^s for any $s \geq 0$. Note that P^0 denotes the identity matrix I_n. For $t = 1$, it is easy to verify that

$$\|\mathbf{z}_i(t) - \bar{\mathbf{z}}(t)\|_2 = 0. \tag{90}$$

To analyze the case with $T \geq t \geq 2$, 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), under Assumption 1, for any $T \geq t \geq 2$, we have

$$\|\mathbf{z}_i(t) - \bar{\mathbf{z}}(t)\|_2 \leq \sum_{\tau=1}^{t-1}\sum_{j=1}^{n}\left|P_{ij}^{t-1-\tau} - \frac{1}{n}\right|\|\nabla_j(\tau)\|_2. \tag{91}$$

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

$$\|P^s\mathbf{x} - \mathbf{1}/n\|_1 \leq \sigma_2(P)^s\sqrt{n} \tag{92}$$

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 (92), for any positive integer s, we have

$$\|P^s\mathbf{e}_i - \mathbf{1}/n\|_1 \leq \sigma_2(P)^s\sqrt{n}. \tag{93}$$

If $s=0$, we also have

$$\|P^0\mathbf{e}_i-\mathbf{1}/n\|_1=\frac{2(n-1)}{n}\leq\sqrt{n}=\sigma_2(P)^0\sqrt{n} \tag{94}$$

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

Moreover, by combining (91) with $\|\nabla_i(t)\|_2\leq\xi$, for any $T\geq t\geq 2$, we have

$$\begin{aligned}\|\mathbf{z}_i(t)-\bar{\mathbf{z}}(t)\|_2\leq&\xi\sum_{\tau=1}^{t-1}\sum_{j=1}^{n}\left|P_{ij}^{t-1-\tau}-\frac{1}{n}\right|=\xi\sum_{\tau=1}^{t-1}\sum_{j=1}^{n}\left|P_{ji}^{t-1-\tau}-\frac{1}{n}\right|\\=&\xi\sum_{\tau=1}^{t-1}\left\|P^{t-1-\tau}\mathbf{e}_i-\frac{\mathbf{1}}{n}\right\|_1\end{aligned} \tag{95}$$

where the first equality is due to the symmetry of P. To further bound the right side of (95), previous studies have provided two different choices.

First, as in Duchi et al. (2011), one can divide $\tau\in[1,t-1]$ into two parts by

$$\tau'=t-1-\left\lceil\frac{\ln(T\sqrt{n})}{\ln(\sigma_2(P)^{-1})}\right\rceil. \tag{96}$$

For any $\tau\in[1,\tau']$, it is not hard to verify that

$$\left\|P^{t-1-\tau}\mathbf{e}_i-\frac{\mathbf{1}}{n}\right\|_1\overset{(93)}{\leq}\sigma_2(P)^{t-1-\tau}\sqrt{n}\overset{(96)}{\leq}\sigma_2(P)^{\frac{\ln(T\sqrt{n})}{\ln(\sigma_2(P)^{-1})}}\sqrt{n}=\frac{1}{T}. \tag{97}$$

Then, by combining (95) and (97), for any $T\geq t\geq 2$, it is easy to verify that

$$\begin{aligned}\|\mathbf{z}_i(t)-\bar{\mathbf{z}}(t)\|_2\leq&\frac{\xi\tau'}{T}+\xi\sum_{\tau=\tau'+1}^{t-1}\left\|P^{t-1-\tau}\mathbf{e}_i-\frac{\mathbf{1}}{n}\right\|_1\\\leq&\xi+2\xi\left\lceil\frac{\ln(T\sqrt{n})}{\ln(\sigma_2(P)^{-1})}\right\rceil\\\leq&\xi+2\xi\left(\frac{\ln(T\sqrt{n})}{1-\sigma_2(P)}+1\right)\end{aligned} \tag{98}$$

where the second inequality is due to $\tau'\leq T$ and $\left\|P^{t-1-\tau}\mathbf{e}_i-\frac{\mathbf{1}}{n}\right\|_1\leq 2$ for any $\tau\leq t-1$, and the last inequality is due to $\ln(x^{-1})\geq 1-x$ for any $x>0$.

The second choice is to simply combine (95) with (93) and (94) as in Hosseini et al. (2013), Zhang et al. (2017), and Wan et al. (2022), which provides the following upper bound

$$\|\mathbf{z}_i(t)-\bar{\mathbf{z}}(t)\|_2\leq\xi\sum_{\tau=1}^{t-1}\sigma_2(P)^{t-1-\tau}\sqrt{n}=\frac{(1-\sigma_2(P)^{t-1})\xi\sqrt{n}}{1-\sigma_2(P)}\leq\frac{\xi\sqrt{n}}{1-\sigma_2(P)} \tag{99}$$

for any $T\geq t\geq 2$.

However, both bounds in (98) and (99) are unsatisfactory, since the former has a factor of $\ln T$ and the latter has a factor of $\sqrt{n}$. To address these issues, we incorporate the above two ideas. Specifically, we redefine τ' as

$$\tau'=t-1-\left\lceil\frac{\ln(\sqrt{n})}{\ln(\sigma_2(P)^{-1})}\right\rceil. \tag{100}$$

Then, from (95), for any $T \geq t \geq 2$, we have

$$\begin{aligned}
\|\mathbf{z}_i(t) - \bar{\mathbf{z}}(t)\|_2 \leq & \xi \sum_{\tau=1}^{\tau'} \left\| P^{t-1-\tau}\mathbf{e}_i - \frac{\mathbf{1}}{n} \right\|_1 + \xi \sum_{\tau=\tau'+1}^{t-1} \left\| P^{t-1-\tau}\mathbf{e}_i - \frac{\mathbf{1}}{n} \right\|_1 \\
\overset{(93)}{\leq} & \xi \sum_{\tau=1}^{\tau'} \sigma_2(P)^{t-1-\tau}\sqrt{n} + \xi \sum_{\tau=\tau'+1}^{t-1} \left\| P^{t-1-\tau}\mathbf{e}_i - \frac{\mathbf{1}}{n} \right\|_1 \\
\leq & \xi \sum_{\tau=1}^{\tau'} \sigma_2(P)^{\tau'-\tau} + \xi \sum_{\tau=\tau'+1}^{t-1} \left\| P^{t-1-\tau}\mathbf{e}_i - \frac{\mathbf{1}}{n} \right\|_1 \\
= & \frac{(1-\sigma_2(P)^{\tau'})\xi}{1-\sigma_2(P)} + \xi \sum_{\tau=\tau'+1}^{t-1} \left\| P^{t-1-\tau}\mathbf{e}_i - \frac{\mathbf{1}}{n} \right\|_1 \\
\leq & \frac{\xi}{1-\sigma_2(P)} + 2\xi \left(\frac{\ln(\sqrt{n})}{1-\sigma_2(P)} + 1 \right)
\end{aligned} \tag{101}$$

where the third inequality is due to $\sigma_2(P)^{t-1-\tau'} \leq 1/\sqrt{n}$ for τ' in (100), and the last inequality is due to $\left\|P^{t-1-\tau}\mathbf{e}_i - \frac{\mathbf{1}}{n}\right\|_1 \leq 2$ for any $\tau \leq t-1$ and (100). Finally, this proof can be completed by combining (90) and (101).

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. *arXiv:1005.2012v3*, 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.

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.

Dan Garber and Ben Kretzu. New projection-free algorithms for online convex optimization with adaptive regret guarantees. In *Proceedings of 35th Conference on Learning Theory*, pages 326–2359, 2022.

Dan Garber and Ben Kretzu. Projection-free online exp-concave optimization. In *Proceedings of 36th Conference on Learning Theory*, pages 1259–1284, 2023.

Eduard Gorbunov, Darina Dvinskikh, and Alexander Gasnikov. Optimal decentralized distributed algorithms for stochastic convex optimization. *arXiv:1911.07363v6*, 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 and Satyen Kale. Beyond the regret minimization barrier: Optimal algorithms for stochastic strongly-convex optimization. *Journal of Machine Learning Research*, 15(71):2489–2512, 2014.

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

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.

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.

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.

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*, 207:1–53, 2024.

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

Yuanyu Wan, Bo Xue, and Lijun Zhang. Projection-free online learning in dynamic environments. In *Proceedings of the 35th AAAI Conference on Artificial Intelligence*, pages 10067–10075, 2021b.

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.

Yuanyu Wan, Lijun Zhang, and Mingli Song. Improved dynamic regret for online Frank-Wolfe. In *Proceedings of the 36th Annual Conference on Learning Theory*, pages 3304–3327, 2023.

Yuanyu Wan, Tong Wei, Mingli Song, and Lijun Zhang. Nearly optimal regret for decentralized online convex optimization. In *Proceedings of the 37th Annual Conference on Learning Theory*, pages 4862–4888, 2024.

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.

Yibo Wang, Yuanyu Wan, and Lijun Zhang. Revisiting projection-free online learning with time-varying constraints. In *Proceedings of the 39th Annual AAAI Conference on Artificial Intelligence*, pages 21339–21347, 2025.

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.

Chenxu Zhang, Yibo Wang, Peng Tian, Xiao Cheng, Yuanyu Wan, and Mingli Song. Projection-free bandit convex optimization over strongly convex sets. In *Proceedings of the 28th Pacific-Asia Conference on Knowledge Discovery and Data Mining*, pages 118–129, 2024.

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.