NEAR-OPTIMAL DECENTRALIZED ONLINE NON-CONVEX OPTIMIZATION

Anonymous authors
Paper under double-blind review

ABSTRACT

This paper studies decentralized online non-convex optimization (D-ONO), where a network of agents collaboratively minimizes a sequence of non-convex global loss functions through local computation and communication. Existing approaches either rely on an offline optimization oracle, which may require computational cost, or analyze window-averaged local regret based on local objectives, which does not accurately characterize the performance of individual agents. How to solve D-ONO under the standard notion of local regret, without requiring computations beyond gradient evaluations, remains largely unexplored. We propose Accelerated Decentralized Gradient Tracking (ADGT), which integrates gradient tracking into a block-wise online accelerated gossip scheme. Under both exact and stochastic gradient oracle feedback, we establish nearly optimal upper bounds on individual local regret and provide matching lower bounds. Extensive experiments on both synthetic and real-world datasets corroborate our theoretical findings and demonstrate the effectiveness of the proposed algorithm.

1 INTRODUCTION

Online optimization is a popular learning framework that models sequential decision-making in a non-stationary environment. Decentralized online optimization extends this framework to a network of agents that learn collaboratively to minimizes a sequence of non-convex global loss functions through local computation and communication (Yan et al., 2012; Hosseini et al., 2013; Zhang et al., 2017; Wan et al., 2024; Sharma & Rajawat, 2024). Specifically, at round t, each agent i selects a decision $\mathbf{x}_{i,t}$, observes feedback on its local loss function $f_{i,t} : \mathbb{R}^d \to \mathbb{R}$, and communicates with its neighbors. The goal of each agents is to minimize a regret defined with respect to the global loss $f_t(\mathbf{x}) := \frac{1}{n}\sum_{i=1}^{n} f_{i,t}(\mathbf{x})$. Existing studies (Hosseini et al., 2013; Zhang et al., 2017; Wan et al., 2024) have developed many efficient decentralized online algorithms with various regret guarantees for convex objective functions. However, the more challenging problem of decentralized online non-convex optimization has not been well explored yet.

Typically, accessing the global minimum of nonconvex functions is infeasible. To address this problem, Jiang et al. (2022) introduced a composite regret that augments the static regret with a network-consensus penalty, they extend FTPL to the decentralized setting and achieves sublinear composite regret, which assume access to an offline optimization oracle. Nazari et al. (2022) proposed a distributed adaptive gradient method and analyzed local regret defined on window-smoothed local objectives, which can not reflect the accurate performance of algorithm. How to solve D-ONO under the standard notion of local regret using only gradient evaluations and local communication remains largely unexplored.

In this work, we develop Accelerated Decentralized Gradient Tracking (ADGT). ADGT adopts a block-wise protocol to jointly performs gradient tracking (Sharma & Rajawat, 2024; Lu et al., 2019; Tang et al., 2018) and accelerated gossip (Liu & Morse, 2011; Wan et al., 2024; Ye et al., 2023). Specifically, the time horizon is partitioned into blocks, with each local decision kept fixed within a block, allowing agents to accumulate gradient feedback while performing one communication step per round. Accelerated gossip rapidly reduces the consensus error, while gradient tracking enables each node to accurately track the average block gradient. We further construct lower bounds for D-ONO under both exact and stochastic gradient oracle feedback using ring-lattice graphs and stochastic zero-chain hard instances, respectively. Our upper bound analysis further shows that ADGT is

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

Table 1: Comparison of upper and lower bounds for D-ONO under exact gradient feedback and stochastic gradient feedback settings. We use the local regret $\mathcal{R}_i(T) = \sum_{t=1}^{T} \|\nabla f_t(\mathbf{x}_{i,t})\|^2$.

Setting	Algorithm	Local Regret Bound	Theorem
deterministic	ADGT	$\tilde{O}(\rho^{-1/2}(1+V_T))$	Theorem 4.1
	Lower Bound	$\Omega(\rho^{-1/2}(1+V_T))$	Theorem 5.1
stochastic	ADGT	$\tilde{O}(\rho^{-1/2}(1+V_T)+\sigma_g\sqrt{(1+V_T)T/n})$	Theorem 4.2
	Lower Bound	$\Omega(\max\{\rho^{-1/2}(1+V_T),\, \sigma_g\sqrt{(1+V_T)T/n}\})$	Theorem 5.2

optimal in both settings up to logarithmic factors, while additionally enjoying linear speedup with respect to n in the stochastic setting. Our contributions are specifically summarized as follows.

- We prove that ADGT achieves individual local regret bounds $\widetilde{O}(\rho^{-1/2}(1+V_T))$ under exact gradient feedback and $\widetilde{O}(\rho^{-1/2}(1+V_T)+\sigma_g\sqrt{(1+V_T)T/n})$ under stochastic gradient feedback. Here ρ is the spectral gap of the communication matrix, V_T measures the variation of functions, and σ_g^2 is the variance of the stochastic gradient feedback.
- For the decentralized algorithm class in Definition 2.5, we establish matching lower bounds under both gradient feedback settings. Under exact gradient feedback, we establish a lower bound $\Omega(\rho^{-1/2}(1+V_T))$. Under stochastic gradient feedback, we establish a lower bound $\Omega(\max\{\rho^{-1/2}(1+V_T), \sigma_g\sqrt{(1+V_T)T/n}\})$, showing the near-optimality of ADGT and the optimality of its linear speedup. Our results are summarized in Table 1.
- We conduct experiments on both synthetic and real-world dataset to demonstrate our theoretical results and the effectiveness of our method. The empirical results show that our algorithm achieves lower individual local regret, improved training loss and test accuracy than prior approaches.

1.1 Related Work

Online non-convex optimization. There are two main lines of work on online non-convex optimization. One line studies *global regret*, which measures the cumulative discrepancy between the loss of the algorithm and the best possible loss in hindsight. Krichene et al. (2015); Yang et al. (2018) developed algorithm based on Hedge algorithm and achieved sublinear regret. Agarwal et al. (2019) extends the bandit algorithm of follow-the-perturbed-leader(FTPL) to the non-convex setting, which is based on a offline optimization oracle. Then Suggala & Netrapalli (2020) proved this algorithm achieves optimal regret.

Another line studies *local regret*, which measures stationarity via gradients of window-smoothed losses. This notion was first introduced by Hazan et al. (2017) and was further developed in Aydore et al. (2019); Hallak et al. (2021); Guan et al. (2023). More recently, Guan et al. (2024) modified this metric by considering the local regret of the original loss functions. Our work follows this local-regret perspective and extends it to the decentralized online non-convex setting with a problem-dependent analysis based on function variation.

Decentralized online convex optimization. Over years of research, various algorithms with optimal regret have been presented for convex and strongly convex functions, respectively (Yan et al., 2012; Hosseini et al., 2013). In particular, Wan et al. (2024) introduced accelerated gossip strategy into D-OCO. Sharma & Rajawat (2024) developed a Generalized Gradient Tracking (GGT) framework in D-OCO. However, these results rely on convexity and their regret analyses do not directly apply to the non-convex setting.

Decentralized online non-convex optimization. Decentralized online non-convex optimization is much less understood. Following the two lines of studies on ONO, existing studies have mainly considered *global regret* and *local regret*. Particularly, Jiang et al. (2022) proposed a composite regret by adding a consensus-disagreement penalty to the standard regret, which is less directly comparable to classical regret measures. Nazari et al. (2022) proposed an Adam-type algorithm for D-ONO, but their window-averaged local regret is based on $f_{i,t}$, does not explicitly capture consensus errors relative to f_t and is less sensitive to temporal variation.

1.2 Paper Organization

The rest of this paper is organized as follows: Section 2 introduces preliminaries and notations we used. Section 3 presents the proposed ADGT algorithm. Section 4 establishes its local regret upper bounds under exact and stochastic gradient oracle feedback, while Section 5 derives the corresponding lower bounds. Section 6 reports the synthetic and real-world experiments. Section 7 presents our conclusions and directions for future work.

2 Preliminaries

In this section, we first introduce notations and assumptions and then formalize the problem setting.

2.1 Notations and Assumptions

In this paper, we consider the decentralized online nonconvex optimization (D-ONO) problem. Specifically, consider n agents connected by an undirected graph $G = ([n], E)$, where $[n] = \{1, 2, \ldots, n\}$ denote the set of agents, $E \subseteq [n] \times [n]$ is the edge set. At each round $t \in [T] = \{1, 2, \ldots, T\}$, each agent $i \in [n]$ first chooses a decision $\mathbf{x}_{i,t} \in \mathbb{R}^d$. Then observes feedback about its local loss function $f_{i,t} : \mathbb{R}^d \to \mathbb{R}$, which is smooth and non-convex. The agents may then exchange messages with their neighbors $\mathcal{N}_i$ according to the graph G. In particular, we consider the following oracle feedback settings:

- **Exact Gradient Oracle (GO) feedback** At each time t, each agent i can access a exact local function gradient $\nabla f_{i,t}(\mathbf{x}_{i,t})$.
- **Stochastic Gradient Oracle (SGO) feedback** At each time t, each agent i can access a stochastic local function gradient $\mathbf{g}_{i,t}(\mathbf{x}_{i,t})$.

Then we introduce the following standard assumptions that are widely used in decentralized online optimization (Yan et al., 2012; Hosseini et al., 2013; Wan et al., 2024) and online non-convex optimization (Guan et al., 2023; 2024; Hallak et al., 2021).

Assumption 2.1. For every $i \in [n]$, $t \in [T]$, and $\mathbf{x} \in \mathbb{R}^d$, the local function $f_{i,t} : \mathbb{R}^d \to \mathbb{R}$ satisfies

$$|f_{i,t}(\mathbf{x})| \leq M$$

for some constant $M \geq 0$.

Assumption 2.2. For every $i \in [n]$ and $t \in [T]$, there exists a constant $L \geq 0$ such that, for all $\mathbf{x}, \mathbf{y} \in \mathbb{R}^d$,

$$\|\nabla f_{i,t}(\mathbf{x}) - \nabla f_{i,t}(\mathbf{y})\| \leq L \|\mathbf{x} - \mathbf{y}\|,$$

where $\|\cdot\|$ denotes the ℓ_2 norm.

Assumption 2.3. The communication matrix $\mathbf{W} \in \mathbb{R}^{n \times n}$ is symmetric, doubly stochastic, positive semidefinite, and supported on the graph $G = ([n], E)$. Specifically,

$$\mathbf{W}\mathbf{1}_n = \mathbf{1}_n, \qquad \mathbf{1}_n^\top \mathbf{W} = \mathbf{1}_n^\top, \qquad \left\| \mathbf{W} - \frac{1}{n}\mathbf{1}_n\mathbf{1}_n^\top \right\|_2 = \sigma_2(\mathbf{W}) < 1.$$

Moreover, $[\mathbf{W}]_{ij} > 0$ if $(i, j) \in E$ or $i = j$, and $[\mathbf{W}]_{ij} = 0$ otherwise. We denote the spectral gap by $\rho := 1 - \sigma_2(\mathbf{W})$.

Assumption 2.4. For all $i \in [n]$, $t \in [T]$, and $\mathbf{x} \in \mathbb{R}^d$, agent i can query a stochastic gradient $\mathbf{g}_{i,t}(\mathbf{x})$ satisfying

$$\mathbb{E}\big[\mathbf{g}_{i,t}(\mathbf{x})\big] = \nabla f_{i,t}(\mathbf{x}), \qquad \mathbb{E}\big[\|\mathbf{g}_{i,t}(\mathbf{x}) - \nabla f_{i,t}(\mathbf{x})\|^2\big] \leq \sigma_g^2.$$

Note that Assumption 2.4 with $\sigma_g = 0$ reduces to the exact gradient oracle.

Algorithm 1 Accelerated Decentralized Gradient Tracking (ADGT)

Input: γ, θ, B, and $\mathbf{W}$
Initialize: set $\mathbf{x}_{i,0} = \mathbf{x}_{i,1} = \mathbf{0}$, $\mathbf{y}_{i,0} = \mathbf{0}$, and $\mathbf{g}_{i,-1} = \mathbf{g}_{i,0} = \mathbf{0}$, $\forall i \in [n]$
for $z = 1, \dots, T/B$ **do**
  for all agents $i \in [n]$ **do**
    Set $\mathbf{u}_{i,z}^{-1} = \mathbf{u}_{i,z}^{0} = \mathbf{x}_{i,z}$, $\mathbf{y}_{i,z}^{-1} = \mathbf{y}_{i,z}^{0} = \mathbf{y}_{i,z-1} + \mathbf{g}_{i,z-1} - \mathbf{g}_{i,z-2}$
    Define the stacked vector $\mathbf{v}_{i,z}^{k} := (\mathbf{u}_{i,z}^{k,\top}, \mathbf{y}_{i,z}^{k,\top})^{\top} \in \mathbb{R}^{2d}$ for $k = -1, 0, \dots, B$
    for $\ell = 0, \dots, B-1$ **do**
      Set $t = (z-1)B + \ell + 1$ and play $\mathbf{x}_{i,z}$
      Query the gradient $\widehat{\nabla} f_{i,t}(\mathbf{x}) := \begin{cases} \nabla f_{i,t}(\mathbf{x}) & \textbf{with GO}, \\ \mathbf{g}_{i,t}(\mathbf{x}) & \textbf{with SGO}. \end{cases}$
      Update $\mathbf{v}_{i,z}^{\ell+1} = (1+\theta)\sum_{j \in \mathcal{N}_i \cup \{i\}} [\mathbf{W}]_{ij} \mathbf{v}_{j,z}^{\ell} - \theta \mathbf{v}_{i,z}^{\ell-1}$
    end for
    Compute $\mathbf{g}_{i,z} = \sum_{t \in \tau_z} \widehat{\nabla} f_{i,t}(\mathbf{x}_{i,z})$, where $\tau_z := \{(z-1)B + 1, \dots, zB\}$
    Set $\mathbf{y}_{i,z} := \mathbf{y}_{i,z}^{B}$ and update $\mathbf{x}_{i,z+1} = \mathbf{u}_{i,z}^{B} - \gamma \mathbf{y}_{i,z}$
  end for
end for

2.2 Individual Local Regret and Algorithm Class

We evaluate the performance of each agent by the following individual local regret definition:

$$\mathcal{R}_i(T) := \sum_{t=1}^{T} \|\nabla f_t(\mathbf{x}_{i,t})\|^2 . \tag{1}$$

Under Assumption 2.1, 2.2, the following average-to-individual reduction holds for all agent $i \in [n]$.

$$\mathcal{R}_i(T) \le 2\overline{\mathcal{R}}(T) + 2L^2 \sum_{t=1}^{T} \left\| \mathbf{X}_t - \mathbf{1}_n \overline{\mathbf{x}}_t^{\top} \right\|_F^2, \quad \text{where} \quad \overline{\mathcal{R}}(T) := \sum_{t=1}^{T} \|\nabla f_t(\overline{\mathbf{x}}_t)\|^2. \tag{2}$$

Here the averaged iterate $\overline{\mathbf{x}}_t := \frac{1}{n}\sum_{i=1}^{n} \mathbf{x}_{i,t}$. Our regret (1) precisely tracks, at each round t, the performance of each individual node's decision on the global problem. We also define the following notation of function variation, which is important to characterize the local regret:

$$V_T := \sum_{t=2}^{T} \max_{i \in [n]} \sup_{\mathbf{x} \in \mathbb{R}^d} |f_{i,t}(\mathbf{x}) - f_{i,t-1}(\mathbf{x})| . \tag{3}$$

Finally, we specify the algorithm class that we consider. Following the classical linear-span oracle model (Nesterov, 2003), we define its decentralized counterpart as follows.

Definition 2.5 (Decentralized linear-span algorithms)**.** For each agent $i \in [n]$, let $\mathcal{H}_{i,t}$ denote the linear span of the information available to agent i before round t, with $\mathcal{H}_{i,1} = \{\mathbf{0}\}$. At round t, agent i plays $\mathbf{x}_{i,t} \in \mathcal{H}_{i,t}$ and queries the current local gradient $\widehat{\nabla} f_{i,t}(\widehat{\mathbf{x}}_{i,t})$ for some $\widehat{\mathbf{x}}_{i,t} \in \mathcal{H}_{i,t}$. Through local computation and neighbor communication, its information space satisfies

$$\mathcal{H}_{i,t+1} \subseteq \operatorname{span}\left(\bigcup_{j \in \mathcal{N}_i \cup \{i\}} \left(\mathcal{H}_{j,t} \cup \{\widehat{\nabla} f_{j,t}(\widehat{\mathbf{x}}_{j,t})\} \right) \right),$$

The above class permits one gradient query and one neighbor-communication step per round. It covers many decentralized gradient-based algorithms, including decentralized gradient descent, decentralized accelerated gradient descent, and our ADGT.

3 Algorithm

In this section, we propose ADGT for D-ONO. The details of the algorithm are presented in Algorithm 1. Since online (stochastic) gradient descent achieves optimal regret for centralized ONO

within the linear-span class (Guan et al., 2024), we retain this backbone in the decentralized setting. However, naive D-SGD cannot achieve optimal regret, since insufficient information mixing leaves consensus errors inadequately controlled. To address this issue, ADGT introduces a block-wise online protocol that coordinates gradient tracking with accelerated gossip strategy. We partitions the time horizon into blocks and keeps each local decision fixed within a block, allowing the agents to accumulate gradient feedback while performing one accelerated gossip step per round, thus control the consensus error.

Let $B \in \mathbb{N}$ be the block length, and assume that B divides T without loss of generality. We divide the time horizon into $N = T/B$ blocks. During block z, agent i keeps its decision $\mathbf{x}_{i,z}$ fixed. At each round $t \in \tau_z$, it plays $\mathbf{x}_{i,z}$ and queries the current local gradient $\nabla f_{i,t}(\mathbf{x}_{i,z})$. The sum of gradients of each agent i in block z can be denoted as

$$\mathbf{g}_{i,z} := \sum_{t\in\tau_z} \widehat{\nabla} f_{i,t}(\mathbf{x}_{i,z}), \quad \text{where} \quad \tau_z := \{(z-1)B+1, \dots, zB\}.$$

The block structure enables online accelerated gossip. Instead of performing multiple communication steps between two consecutive online rounds, ADGT distributes B gossip steps over the B rounds of each block. At the beginning of block z, agent i initializes $\mathbf{u}_{i,z}^{-1}=\mathbf{u}_{i,z}^{0}=\mathbf{x}_{i,z}$ and $\mathbf{y}_{i,z}^{-1}=\mathbf{y}_{i,z}^{0}=\mathbf{y}_{i,z-1}+\mathbf{g}_{i,z-1}-\mathbf{g}_{i,z-2}$. The auxiliary variable $\mathbf{u}_{i,z}^{\ell}$ is used to mix the local decisions, whereas $\mathbf{y}_{i,z}^{\ell}$ tracks the network-average block gradient. We stack these two variables into $\mathbf{v}_{i,z}^{\ell}$. At gossip iteration ℓ, each agent exchanges $\mathbf{v}_{i,z}^{\ell}$ with its neighbors and performs

$$\mathbf{v}_{i,z}^{\ell+1} = (1+\theta) \sum_{j\in\mathcal{N}_i\cup\{i\}} [\mathbf{W}]_{ij}\mathbf{v}_{j,z}^{\ell} - \theta\mathbf{v}_{i,z}^{\ell-1}, \quad \text{where} \quad \mathbf{v}_{j,z}^{\ell} := (\mathbf{u}_{i,z}^{\ell\top}, \mathbf{y}_{i,z}^{\ell\top})^{\top} \in \mathbb{R}^{2d} \tag{4}$$

Here, $\mathbf{W}$ is the communication matrix, and $\rho = 1-\sigma_2(\mathbf{W})$ is its spectral gap. $\theta > 0$ is the mixing coefficient in the accelerated gossip recursion(Liu & Morse, 2011). This recursion reduces disagreement among the agents more rapidly than standard gossip and is used to improve the network dependence of ADGT. After B gossip steps, agent i sets $\mathbf{y}_{i,z} := \mathbf{y}_{i,z}^{B}$ and performs

$$\mathbf{x}_{i,z+1} = \mathbf{u}_{i,z}^{B} - \gamma\mathbf{y}_{i,z}. \tag{5}$$

Here $\gamma > 0$ is the stepsize.

4 Local Regret Analysis

In this section, we establish upper bounds on the individual local regret of ADGT under exact gradient feedback and stochastic gradient feedback. These two settings are analyzed in Sections 4.1 and 4.2, respectively.

4.1 ADGT with Exact Gradient Oracle

We first consider the exact gradient feedback, where each agent observes the exact gradient of its local loss function. The following theorem establishes an upper bound of ADGT, explicitly characterizing its dependence on the function variation V_T and the spectral gap ρ.

Theorem 4.1. *Under Assumptions 2.1, 2.2, 2.3, let $B = \Theta(\rho^{-1/2}\log\sqrt{n})$ and $\gamma < O(B^{-1})$, for each node $i \in [n]$, Algorithm 1 can guarantee*

$$\mathcal{R}_i(T) \le \tilde{O}\left(\rho^{-1/2}(1+V_T)\right).$$

Proof sketch. The proof proceeds in three steps.

- First, we reformulate the individual local regret $\mathcal{R}_i(T)$ in terms of the averaged local regret $\overline{\mathcal{R}}(T)$ and the cumulative consensus error due to Eq. (2). Let $\mathbf{X}_t$ denote the matrix whose i-th row is $\mathbf{x}_{i,t}^{\top}$. Due to the block-wise update in ADGT, the consensus error term in Eq. (2) is equivalent to the block-level $B\sum_{z=1}^{N}\|\mathbf{X}_z - \mathbf{1}_n\overline{\mathbf{x}}_z^{\top}\|_F$, where $\overline{\mathbf{x}}_z := \frac{1}{n}\sum_{i=1}^{n}\mathbf{x}_{i,z}$.

- Second, we further exploit the block structure and reformulate $\overline{\mathcal{R}}(T)$ in terms of the block-level averaged local regret $\overline{\mathcal{R}}_{\text{block}}(T) := \sum_{z=1}^{N} \|\nabla F_{\tau_z}(\overline{\mathbf{x}}_z)\|^2$, where $F_{\tau_z}(\mathbf{x}) := \sum_{t\in\tau_z} f_t(\mathbf{x})$ denotes the block function. The following supported reduction holds.

$$\overline{\mathcal{R}}(T) \leq \frac{1}{B}\overline{\mathcal{R}}_{\text{block}}(T) + \sum_{z=1}^{N}\sum_{s=1}^{B}\left\|\nabla f_{(z-1)B+s}(\overline{\mathbf{x}}_z) - \frac{1}{B}\nabla F_{\tau_z}(\overline{\mathbf{x}}_z)\right\|^2.$$

  The second term on the right-hand side of the above inequality highlights the challenge posed by the block-wise update mechanism in time-varying environments, as it characterizes the extent of the variation in the global objective gradients within each block. For smooth functions, the gradient variation can be bounded in terms of the function value variation. Therefore, by the definition of V_T in (3), we show that this term admits an upper bound of order $O(BV_T)$. After bounding $\overline{\mathcal{R}}_{\text{block}}(T)$, we establish a further upper bound on $\overline{\mathcal{R}}(T)$ in terms of the block-level consensus error and V_T, as shown in the following inequality.

$$\overline{\mathcal{R}}(T) \leq O\left(\frac{1+V_T}{\gamma} + BV_T + \frac{B}{n}\sum_{z=1}^{N}\left\|\mathbf{X}_z - \mathbf{1}_n\overline{\mathbf{x}}_z^\top\right\|_F^2\right). \tag{6}$$

- Finally, accelerated gossip contracts the disagreement caused by local communication, while gradient tracking controls the discrepancy between the local update directions and their network average. Their combined effect yields the following consensus error bound.

$$\sum_{z=1}^{N}\left\|\mathbf{X}_z - \mathbf{1}_n\overline{\mathbf{x}}_z^\top\right\|_F^2 \leq O\left(nB^2\lambda_B^2\gamma^2\left(1+V_T+\overline{\mathcal{R}}(T)\right)\right). \tag{7}$$

  Here $\lambda_B = \sqrt{14}(1-(1-\frac{1}{\sqrt{2}})\sqrt{\rho})^B < 1$ is the contraction factor after B accelerated gossip steps. Substituting Eq. (7) into Eq. (6) produces a recursive bound on $\overline{\mathcal{R}}(T)$. Under the prescribed choices of B and γ, the term proportional to $\overline{\mathcal{R}}(T)$ can be absorbed into the left-hand side. This establishes the averaged regret bound, which, together with Eq. (2), yields the claimed individual local regret bound for every agent.

4.2 ADGT WITH STOCHASTIC GRADIENT ORACLE

We consider the stochastic gradient feedback, where each agent receives an stochastic gradient satisfying Assumption 2.4. The following theorem establishes an upper bound on the individual local regret of ADGT, explicitly separating the effects of V_T, ρ, and stochastic gradient noise σ_g.

Theorem 4.2. *Under Assumptions 2.1-2.4, let* $\gamma \leq O(\min\{B^{-1}, \sqrt{\frac{n(1+V_T)}{\sigma_g^2 T}}\})$, $B = \Theta(\rho^{-1/2}\log n)$, *for each node* $i \in [n]$, *Algorithm 1 can gaurantee*

$$\mathbb{E}\left[\mathcal{R}_i(T)\right] \leq \tilde{O}\left(\rho^{-1/2}(1+V_T) + \sigma_g\sqrt{\frac{(1+V_T)T}{n}}\right).$$

Proof sketch. The proof follows the same block-level decomposition and average-to-individual reduction used in the analysis under exact gradient feedback. The main additional difficulty is to control the errors introduced by stochastic gradient noise.

For each block z, define the accumulated noise $\mathbf{e}_{i,z} := \sum_{t\in\tau_z}\boldsymbol{\xi}_{i,t}$, its network average $\overline{\mathbf{e}}_z := \frac{1}{n}\sum_{i=1}^{n}\mathbf{e}_{i,z}$, and the stacked matrix $\mathbf{E}_z := [\mathbf{e}_{1,z}^\top, \ldots, \mathbf{e}_{n,z}^\top]^\top$. We further define $\boldsymbol{\Delta}_{\mathbf{E},z} := \mathbf{E}_z - \mathbf{E}_{z-1}$. The following inequality bounds $\overline{\mathcal{R}}(T)$ in terms of these two noise quantities.

$$\mathbb{E}\left[\overline{\mathcal{R}}(T)\right] \leq O\left(\frac{1+V_T}{\gamma} + BV_T + B^3\lambda_B^2\gamma^4\sum_{z=1}^{N}\mathbb{E}\left\|\overline{\mathbf{e}}_z\right\|^2 + \frac{B\lambda_B^2\gamma^2}{n}\sum_{z=1}^{N}\mathbb{E}\left\|\boldsymbol{\Delta}_{\mathbf{E},z}\right\|_F^2\right). \tag{8}$$

The two noise terms in Eq. (8) arise through different parts of the analysis. The term involving $\overline{\mathbf{e}}_z$ is the noise in the network-average block gradient and corresponds to the stochastic error appearing in centralized SGD. Averaging the independent gradient queries across n agents reduces its cumulative

variance to $\sum_{z=1}^{N}\mathbb{E}[\|\overline{\mathbf{e}}_z\|^2] \le T\sigma_g^2/n$. By contrast, $\mathbf{\Delta}_{\mathbf{E},z}$ enters through the gradient-tracking and consensus recursions. Its cumulative variance satisfies $\sum_{z=1}^{N}\mathbb{E}[\|\mathbf{\Delta}_{\mathbf{E},z}\|_F^2] \le 2nT\sigma_g^2$. The influence of this term on the regret is attenuated by the gossip contraction factor λ_B. In particular, choosing $B = \Theta(\rho^{-1/2}\log n)$ gives $\lambda_B = O(1/n)$.

Substituting these noise bounds into Eq. (8), together with the prescribed choice of γ, yields

$$\mathbb{E}[\overline{\mathcal{R}}(T)] \le \tilde{O}\left(\rho^{-1/2}(1+V_T) + \sigma_g\sqrt{\frac{(1+V_T)T}{n}}\right).$$

Finally, applying the average-to-individual reduction in Eq (2) and the stochastic consensus error bound gives the individual regret guarantee in Theorem 4.2.

Remark 4.3 (Linear speedup). It is worth noting that we improve the stochastic term in Theorem 4.2 from $O(\sigma_g\sqrt{(1+V_T)T})$ in the single-agent setting (Guan et al., 2024) to $O(\sigma_g\sqrt{(1+V_T)T/n})$. After normalization by T, its contribution to the average stationarity measure decreases by a factor of $1/\sqrt{n}$. Therefore, for a fixed variation budget and when the stochastic term dominates, n agents require only $1/n$ as many rounds as a single agent to attain the same stationarity accuracy.

Remark 4.4 (Achieve the near-optimal convergence rate in the offline setting). Based on the proof sketch of our upper bounds, we note that our result also holds for the averaged iterates $\overline{\mathbf{x}}_t$. Specifically, when in static environment, i.e., $V_T = 0$ and $f_t \equiv f$ for all $t \in [T]$, the bound in Theorem 4.2 reduces to

$$\mathbb{E}\left[\overline{\mathcal{R}}(T)\right] \le \tilde{O}\left(\rho^{-1/2} + \sigma_g\sqrt{\frac{T}{n}}\right).$$

Applying the standard online-to-batch conversion, let $\hat{t}$ be sampled uniformly from $[T]$, we have

$$\mathbb{E}\left[\|\nabla f(\overline{\mathbf{x}}_{\hat{t}})\|^2\right] = \frac{1}{T}\mathbb{E}\left[\overline{\mathcal{R}}(T)\right] \le \tilde{O}\left(\frac{1}{T\sqrt{\rho}} + \frac{\sigma_g}{\sqrt{nT}}\right).$$

Hence in the static setting, ADGT recovers the $\tilde{O}(1/(\sqrt{\rho}T) + \sigma_g/\sqrt{nT})$ rate for decentralized stochastic non-convex optimization, achieving both the optimal dependence of ρ and the standard linear speedup in n (Lu & De Sa, 2021; Yuan et al., 2022). This matches the known offline lower bound up to logarithmic factors.

5 Lower Bounds on D-ONO

In this section, we establish lower bounds on the individual local regret of D-ONO under exact gradient feedback and stochastic gradient feedback, respectively. We first establish the following lower bound.

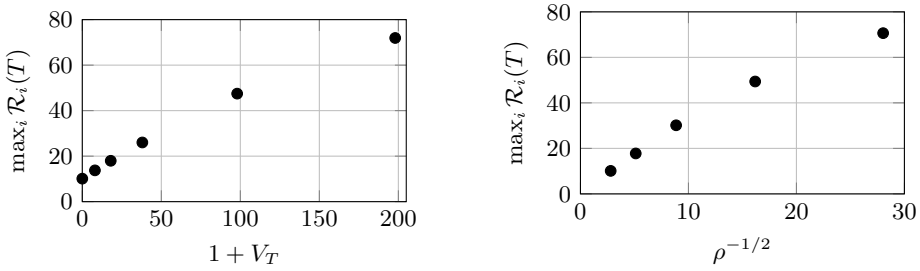
Theorem 5.1. *Consider any algorithm A satisfying Definition 2.5. Under Assumptions 2.1, 2.2, 2.3 and suppose $d \ge \Omega(1+V_T)$. There exist a graph G with diameter $D \le O(T/(1+V_T))$, $\rho \in [1-\cos(\pi/n), 1]$ and $\{f_t\}_{t=1}^T$ with $V_T = o(T)$ such that, for some node $i \in [n]$, $\mathcal{R}_i(T) \ge \Omega(\rho^{-1/2}(1+V_T))$. Moreover, if $D > \Omega(T/(1+V_T))$, then it yields the linear $\mathcal{R}_i(T) = \Theta(T)$.*

We then establish the following lower bound of D-ONO problem under stochastic gradient feedback.

Theorem 5.2. *Consider any algorithm A with stochastic gradient feedback $\mathbf{g}_{i,t}$ satisfying Definition 2.5. Under Assumptions 2.1-2.4 and suppose $d \ge \Omega(T)$. There exist a graph G with $\rho \in (0,1]$ and $\{f_t\}_{t=1}^T$ with $V_T = o(T)$ such that for some node $i \in [n]$, $\mathbb{E}[\mathcal{R}_i(T)] \ge \Omega(\sigma_g\sqrt{(1+V_T)T/n})$.*

Theorems 5.1 and 5.2 reveal the lower bounds from the perspectives of communication difficulty and stochastic noise, respectively. Combining the above two theorems, we naturally obtain the following corollary for stochastic D-ONO:

Corollary 5.3. *Consider any algorithm A with stochastic gradient feedback $\mathbf{g}_{i,t}$ satisfying Definition 2.5. Under Assumptions 2.1-2.4 and suppose $d \ge \Omega(T)$. There exist a graph G with diameter*



(a) Varying $1 + V_T$ with $w = 1/3$.

(b) Varying ρ with $B_{\text{time}} = 1000$.

Figure 1: Cumulative individual local regret of ADGT under different values of V_T and ρ.

$D \le O(T(1+V_T))$, $\rho \in [1-\cos(\pi/n), 1]$ and $\{f_t\}_{t=1}^T$ with $V_T = o(T)$ such that,

$$\mathbb{E}[\mathcal{R}_i(T)] \ge \Omega\left(\max\left\{\rho^{-1/2}(1+V_T),\, \sigma_g\sqrt{\frac{(1+V_T)T}{n}}\right\}\right), \qquad \forall i \in [n].$$

The above lower bound matches the upper bound in Theorem 4.2 up to logarithmic factors, demonstrating that ADGT achieves the near-optimal dependence on the network parameters ρ and n, environmental variation V_T and stochastic noise σ_g. In particular, the two terms characterize the intrinsic difficulties arising from decentralized communication and stochastic gradient noise, respectively.

6 Experiments

In this section, we validate our theoretical findings and demonstrate the effectiveness of our algorithm through both synthetic and real-world experiments.

6.1 Synthetic Experiment

In this section, we evaluate ADGT on a synthetic experiment by varying the functions and network parameters to explicitly control V_T and ρ, respectively. These quantities are typically difficult to compute or observe in real-world scenarios. This controlled experiment allows us to empirically verify the dependence predicted by our theoretical results.

We first consider a random-walk ring network with $n = 10$ nodes and a sequence of local functions $\tilde{f}_{i,k}$ that varies with the time-block size B_{time}, where $k = 1, \ldots, \lfloor T/B_{\text{time}} \rfloor$. We assume that all edges in the ring graph have the same weight $w = 1/3$. In particular, for all $i \in [n]$ we have

$$f_{i,(k-1)B_{\text{time}}+1}(\mathbf{x}) = f_{i,(k-1)B_{\text{time}}+2}(\mathbf{x}) = \cdots = f_{i,kB_{\text{time}}}(\mathbf{x}) = \tilde{f}_{i,k}(\mathbf{x}).$$

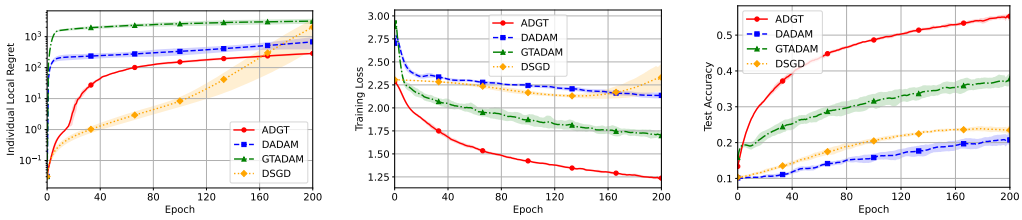
We construct the following local functions $\tilde{f}_{i,k}$, which clearly satisfy Assumption 2.1, 2.2:

$$\tilde{f}_{i,k}(\mathbf{x}) = \tanh([\mathbf{x}]_{i+nc_{i,k}}), \quad \text{where} \quad c_{i,k} = \left\lfloor \frac{k-1}{n} \right\rfloor + \mathbf{1}\{i \le (k-1) \bmod n\}.$$

Intuitively, whenever the process enters a new time block, exactly one node switches its local function $f_{i,t}$ to a previously unused coordinate. Therefore, we have $V_T = 2(\lfloor T/B_{\text{time}} \rfloor - 1)$. As shown in Figure 1a, by varying B_{time} while setting $\gamma = 0.1$ and $B = 5$, the regret of ADGT exhibits an approximately linear dependence on $1 + V_T$. We then fix $B_{\text{time}} = 1000$ and vary the edge weight w of the communication matrix to control the spectral gap ρ. Figure 1b reports the regret of ADGT under different values of ρ. We simultaneously adjust the stepsize γ and block size B according to the parameter choices specified in Theorem 4.1 (see Appendix A). The results show that the regret scales approximately linearly with $\rho^{-1/2}$, further supporting our theoretical results.

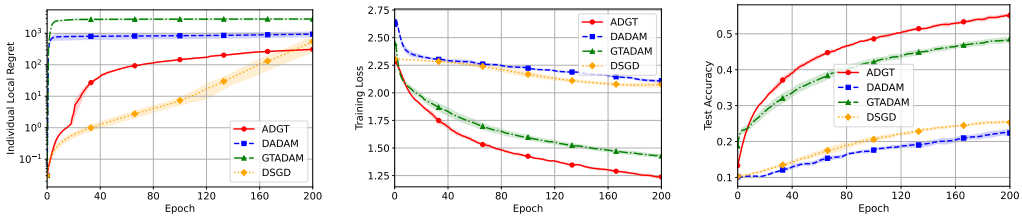
6.2 Real-world Experiment

In this section, we evaluate our ADGT against D-PSGD (Lian et al., 2017), DADAM (Nazari et al., 2022) and GTADAM (Carnevale et al., 2022) on CIFAR10 dataset (Alex, 2009).



(a) regret (b) training loss (c) test accuracy

Figure 2: Performance of different algorithms on CIFAR10 over the ring network. In subfigure (a), we plot the cumulative regret of the worst-performing node, shown as $\max_{i\in[n]} \sum_{t=1}^{T} \|\nabla f_t(\mathbf{x}_{i,t})\|^2$.



(a) regret (b) training loss (c) test accuracy

Figure 3: Performance of different algorithms on CIFAR10 over the Erdös-Rényi network.

We train a lightweight CNN model, consisting of two 3×3 convolutional layers with 32 channels, each followed by ReLU activation and 2×2 max pooling. The resulting $32 \times 5 \times 5$ feature map is flattened and fed into a fully connected layer with 128 hidden units, followed by a 10-class output layer. Dropout with rates 0.25 and 0.5 is applied after the convolutional feature extractor and the hidden layer, respectively. We employ two network topologies, i.e., the random-walk ring graph and the Erdös-Rényi graph (ERDdS & R&wi, 1959) with the connectivity probability of $p = 0.5$.

We construct an online data stream with a batch size of 128 and consider a fully heterogeneous data distribution. We set $n = 10$ nodes, where all data samples at each node share the same label. At fixed time intervals, we simulate distribution shifts in the data stream by changing all data at each node with samples from a new label, i.e., node i contains data with label i. After once change, node i receives new data with label $(i+1) \bmod n$. We perform a total of 9 distribution shifts. The stepsize is tuned over {5e-2, 1e-2, 5e-3, 1e-3, 5e-4, 1e-4, 5e-5, 1e-5}. For ADGT, we additionally tune the block size B over $\{1, 2, 3, 4\}$. For DADAM, we follow the setting in Nazari et al. (2022), with $\beta_1 = \beta_3 = 0.9$, $\beta_2 = 0.999$, and $\alpha_t = (\rho/t)^{1/2}$. For GTADAM, G is tuned over $\{1, 5, 10, 100\}$.

We present the results on the ring graph in Figure 2, and the results on the Erdös-Rényi graph in Figure 3. All results are averaged over five independent runs. As shown, even under a fully heterogeneous data stream, ADGT effectively controls the cumulative regret along the individual sequence of every node, while achieving the best training loss and validation accuracy. These results empirically support the effectiveness of our algorithm and are consistent with our theoretical guarantees.

7 Conclusion

In this paper, we study D-ONO under exact and stochastic gradient feedback and develop ADGT, which jointly performs gradient tracking and accelerated gossip through a block-wise online protocol. We establish near-optimal individual local regret upper bounds together with matching lower bounds up to logarithmic factors, characterizing the effects of function variation, network connectivity, stochastic gradient noise, and the number of agents. Our results also demonstrate an optimal linear speedup under stochastic feedback. Synthetic and real-world experiments further support our theoretical findings and the effectiveness of ADGT.

Future work. A natural direction for future work is to relax the requirement imposed by V_T on the temporal variation of each local loss function. Another interesting question is whether our lower bounds can be extended beyond the Decentralized linear-span algorithms class 2.5.

AI USE STATEMENT

In this work, We used generative AI tools solely to proofread the manuscript and polish language. All AI-assisted edits were carefully reviewed and revised by the authors. We take responsibility for the final content of this work.

REPRODUCIBILITY STATEMENT

We provide complete proofs for all theoretical claims in the appendix. To facilitate reproducibility, we include the source code in the supplementary material. We use the CIFAR10 [1] dataset in our experiments, which is publicly available.

REFERENCES

Naman Agarwal, Alon Gonen, and Elad Hazan. Learning in non-convex games with an optimization oracle. In *Conference on Learning Theory*, pp. 18–29. PMLR, 2019.

Krizhevsky Alex. Learning multiple layers of features from tiny images. *https://www. cs. toronto. edu/kriz/learning-features-2009-TR. pdf*, 2009.

Yossi Arjevani, Yair Carmon, John C. Duchi, Dylan J. Foster, Nathan Srebro, and Blake Woodworth. Lower bounds for non-convex stochastic optimization. *Mathematical Programming*, 199(1–2): 165–214, 2023. doi: 10.1007/s10107-022-01822-7.

Sergul Aydore, Tianhao Zhu, and Dean P Foster. Dynamic local regret for non-convex online forecasting. *Advances in neural information processing systems*, 32, 2019.

Yair Carmon, John C Duchi, Oliver Hinder, and Aaron Sidford. Lower bounds for finding stationary points i. *Mathematical Programming*, 184(1):71–120, 2020.

Guido Carnevale, Francesco Farina, Ivano Notarnicola, and Giuseppe Notarstefano. Gtadam: Gradient tracking with adaptive momentum for distributed online optimization. *IEEE Transactions on Control of Network Systems*, 10(3):1436–1448, 2022.

PARW ERDdS and A R&wi. On random graphs i. *Publ. math. debrecen*, 6(290-297):18, 1959.

Ziwei Guan, Yi Zhou, and Yingbin Liang. Online nonconvex optimization with limited instantaneous oracle feedback. In *The Thirty Sixth Annual Conference on Learning Theory*, pp. 3328–3355. PMLR, 2023.

Ziwei Guan, Yi Zhou, and Yingbin Liang. On the hardness of online nonconvex optimization with single oracle feedback. In *International Conference on Learning Representations*, volume 2024, pp. 34606–34629, 2024.

Nadav Hallak, Panayotis Mertikopoulos, and Volkan Cevher. Regret minimization in stochastic non-convex learning via a proximal-gradient approach. In *International Conference on Machine Learning*, pp. 4008–4017. PMLR, 2021.

Elad Hazan, Karan Singh, and Cyril Zhang. Efficient regret minimization in non-convex games. In *International Conference on Machine Learning*, pp. 1433–1441. PMLR, 2017.

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

Zhanhong Jiang, Aditya Balu, Xian Yeow Lee, Young M Lee, Chinmay Hegde, and Soumik Sarkar. Distributed online non-convex optimization with composite regret. In *2022 58th Annual Allerton Conference on Communication, Control, and Computing (Allerton)*, pp. 1–8. IEEE, 2022.

Walid Krichene, Maximilian Balandat, Claire Tomlin, and Alexandre Bayen. The hedge algorithm on a continuum. In *International Conference on Machine Learning*, pp. 824–832. PMLR, 2015.

[1] https://cave.cs.toronto.edu/kriz/cifar.html

Xiangru Lian, Ce Zhang, Huan Zhang, Cho-Jui Hsieh, Wei Zhang, and Ji Liu. Can decentralized algorithms outperform centralized algorithms? a case study for decentralized parallel stochastic gradient descent. *Advances in neural information processing systems*, 30, 2017.

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

Songtao Lu, Xinwei Zhang, Haoran Sun, and Mingyi Hong. GNSD: A gradient-tracking based non-convex stochastic algorithm for decentralized optimization. In *2019 IEEE Data Science Workshop (DSW)*, pp. 315–321. IEEE, 2019.

Yucheng Lu and Christopher De Sa. Optimal complexity in decentralized training. In *International conference on machine learning*, pp. 7111–7123. PMLR, 2021.

Parvin Nazari, Davoud Ataee Tarzanagh, and George Michailidis. Dadam: A consensus-based distributed adaptive gradient method for online optimization. *IEEE Transactions on Signal Processing*, 70:6065–6079, 2022.

Yurii Nesterov. *Introductory Lectures on Convex Optimization: A Basic Course*, volume 87. Springer Science & Business Media, 2003.

Shivangi Dubey Sharma and Ketan Rajawat. Optimized gradient tracking for decentralized online learning. *IEEE Transactions on Signal Processing*, 72:1443–1459, 2024.

Arun Sai Suggala and Praneeth Netrapalli. Online non-convex learning: Following the perturbed leader is optimal. In *Algorithmic Learning Theory*, pp. 845–861. PMLR, 2020.

Hanlin Tang, Xiangru Lian, Ming Yan, Ce Zhang, and Ji Liu. D^2: Decentralized training over decentralized data. In *International Conference on Machine Learning*, pp. 4848–4856. PMLR, 2018.

Yuanyu Wan, Tong Wei, Mingli Song, and Lijun Zhang. Nearly optimal regret for decentralized online convex optimization. In *The Thirty Seventh Annual Conference on Learning Theory*, pp. 4862–4888. PMLR, 2024.

Feng Yan, Shreyas Sundaram, SVN 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, 2012.

Lin Yang, Lei Deng, Mohammad H Hajiesmaili, Cheng Tan, and Wing Shing Wong. An optimal algorithm for online non-convex learning. *Proceedings of the ACM on Measurement and Analysis of Computing Systems*, 2(2):1–25, 2018.

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.

Kun Yuan, Xinmeng Huang, Yiming Chen, Xiaohan Zhang, Yingya Zhang, and Pan Pan. Revisiting optimal convergence rate for smooth and non-convex stochastic decentralized optimization. *Advances in Neural Information Processing Systems*, 35:36382–36395, 2022.

Wenpeng Zhang, Peilin Zhao, Wenwu Zhu, Steven CH Hoi, and Tong Zhang. Projection-free distributed online learning in networks. In *International conference on machine learning*, pp. 4054–4062. PMLR, 2017.

A Experimental Details

In Section 6.1, we consider a ring network consisting of 10 nodes and assign the same edge weight w to every pair of connected nodes. In Figure 1a, we fix the initial edge weight as $w_0 = 1/3$ and set the initial parameters $\gamma_0 = 0.1$ and $B_0 = 5$. By varying B_{time}, we verify the approximately linear dependence of the regret of ADGT on $1 + V_T$.

In Figure 1b, we vary w to control the spectral quantity ρ and investigate the dependence of the regret bound on $\rho^{-1/2}$. Specifically, we consider

$$w \in \left\{ \frac{1}{3}, \frac{1}{10}, \frac{1}{30}, \frac{1}{100}, \frac{1}{300} \right\}.$$

According to Theorem 4.1, the parameter choices should satisfy $\gamma = O(B^{-1})$ and $B = O(\rho^{-1/2})$. Therefore, we scale the parameters relative to the initial configuration as

$$B = \left\lfloor \frac{B_0 \rho^{-1/2}}{\rho_0^{-1/2}} \right\rfloor \qquad \text{and} \qquad \gamma = \frac{\gamma_0 B_0}{B}.$$

For completeness, we report the exact data points used in Figures 1a and 1b in Tables 2a and 2b, respectively.

Table 2: Cumulative individual local regret of ADGT under different values of V_T and ρ.

(a) Varying V_T with $w = 1/3$.

B_{time}	V_T	$\max_i \mathcal{R}_i(T)$
10	198.00	71.93
20	98.00	47.47
50	38.00	26.01
100	18.00	17.96
200	8.00	13.75
1000	0.00	10.11

(b) Varying ρ with $B_{\text{time}} = 1000$.

w	$\rho^{-1/2}$	B	γ	$\max_i \mathcal{R}_i(T)$
$1/3$	2.80	5	0.1000	10.11
$1/10$	5.12	9	0.0556	17.77
$1/30$	8.86	16	0.0313	30.15
$1/100$	16.18	29	0.0172	49.39
$1/300$	28.03	50	0.0100	70.64

B Local Regret Analysis of ADGT

In this section, we prove the local regret upper bound for ADGT (Algorithm 1) under exact gradient feedback. In Appendices B.2.1 and B.2.2, we prove Lemmas B.2 and B.3, which bound the averaged local regret and the cumulative consensus error, respectively. Finally, Appendix B.2.3 provides the proof of Theorem 4.1.

B.1 Average-Tracking Property

Lemma B.1. *Let* $\overline{\mathbf{g}}_z := \frac{1}{n}\sum_{i=1}^n \mathbf{g}_{i,z}$ *and* $\overline{\mathbf{y}}_z := \frac{1}{n}\sum_{i=1}^n \mathbf{y}_{i,z}$. *Under Assumption 2.3, Algorithm 1 holds that*

$$\overline{\mathbf{y}}_z = \overline{\mathbf{g}}_{z-1}, \qquad z \geq 1.$$

We first introduce the notation used in the proof. For each block τ_z, define

$$\overline{\mathbf{x}}_z := \frac{1}{n}\sum_{i=1}^n \mathbf{x}_{i,z}, \qquad F_{\tau_z}(\mathbf{x}) := \sum_{t\in\tau_z} f_t(\mathbf{x}), \qquad f_t(\mathbf{x}) := \frac{1}{n}\sum_{i=1}^n f_{i,t}(\mathbf{x}).$$

The corresponding block gradient is

$$\nabla F_{\tau_z}(\mathbf{x}) := \sum_{t\in\tau_z} \nabla f_t(\mathbf{x}).$$

Moreover, define the averaged block gradient and the averaged tracking variable by

$$\overline{\mathbf{g}}_z := \frac{1}{n}\sum_{i=1}^n \mathbf{g}_{i,z}(\mathbf{x}_{i,z}), \qquad \overline{\mathbf{y}}_z := \frac{1}{n}\sum_{i=1}^n \mathbf{y}_{i,z}.$$

In algorithm 1, the tracker used in block z tracks the previous block gradient in average. More precisely, we prove

$$\overline{\mathbf{y}}_z = \overline{\mathbf{g}}_{z-1}, \qquad z \geq 1,$$

where we use the dummy initialization $\mathbf{G}_0 = \mathbf{G}_{-1} = \mathbf{0}$ and $\overline{\mathbf{g}}_0 = \overline{\mathbf{g}}_{-1} = \mathbf{0}$.

Let $\mathbf{Y}_z \in \mathbb{R}^{n\times d}$ and $\mathbf{G}_z \in \mathbb{R}^{n\times d}$ be the matrices whose i-th rows are $\mathbf{y}_{i,z}^\top$ and $\mathbf{g}_{i,z}(\mathbf{x}_{i,z})^\top$, respectively. Let $\mathbf{Y}_z^k \in \mathbb{R}^{n\times d}$ be the inner accelerated gossip iterates. Define $\mathbf{J} := \frac{1}{n}\mathbf{1}\mathbf{1}^\top$. Since the gossip matrix $\mathbf{W}$ is doubly stochastic, we have $\mathbf{J}\mathbf{W} = \mathbf{W}\mathbf{J} = \mathbf{J}$.

In the delayed block algorithm, the gradient-tracking correction from block $z-1$ is injected before the accelerated gossip in block z. Define the pre-gossip tracker input

$$\mathbf{Q}_z := \mathbf{Y}_{z-1} + \mathbf{G}_{z-1} - \mathbf{G}_{z-2}, \qquad z \geq 1,$$

with the initialization $\mathbf{Y}_0 = \mathbf{0}$, $\mathbf{G}_0 = \mathbf{G}_{-1} = \mathbf{0}$. The inner accelerated gossip variables are initialized as $\mathbf{Y}_z^0 = \mathbf{Y}_z^{-1} = \mathbf{Q}_z$. Within block z, the accelerated gossip recursion is

$$\mathbf{Y}_z^{k+1} = (1+\theta)\mathbf{W}\mathbf{Y}_z^k - \theta\mathbf{Y}_z^{k-1}, \qquad k = 0, \ldots, L-1.$$

Multiplying both sides by $\mathbf{J}$ gives

$$\begin{aligned}\mathbf{J}\mathbf{Y}_z^{k+1} &= (1+\theta)\mathbf{J}\mathbf{W}\mathbf{Y}_z^k - \theta\mathbf{J}\mathbf{Y}_z^{k-1} \\ &= (1+\theta)\mathbf{J}\mathbf{Y}_z^k - \theta\mathbf{J}\mathbf{Y}_z^{k-1}. \end{aligned} \tag{9}$$

Since $\mathbf{J}\mathbf{Y}_z^0 = \mathbf{J}\mathbf{Y}_z^{-1} = \mathbf{J}\mathbf{Q}_z$, an induction on k in (9) gives

$$\mathbf{J}\mathbf{Y}_z^k = \mathbf{J}\mathbf{Q}_z, \qquad k = 0, 1, \ldots, L.$$

In particular, after L accelerated gossip steps,

$$\mathbf{J}\mathbf{Y}_z = \mathbf{J}\mathbf{Y}_z^B = \mathbf{J}\mathbf{Q}_z. \tag{10}$$

We now prove the delayed average-tracking identity by induction. For $z = 1$, by the initialization $\mathbf{Q}_1 = \mathbf{Y}_0 + \mathbf{G}_0 - \mathbf{G}_{-1} = \mathbf{0}$ and (10), it holds that

$$\mathbf{J}\mathbf{Y}_1 = \mathbf{J}\mathbf{Q}_1 = \mathbf{0} = \mathbf{J}\mathbf{G}_0.$$

Equivalently, $\overline{\mathbf{y}}_1 = \overline{\mathbf{g}}_0 = \mathbf{0}$. Suppose that for some $z \geq 2$, $\mathbf{J}\mathbf{Y}_{z-1} = \mathbf{J}\mathbf{G}_{z-2}$. Then, by the definition of $\mathbf{Q}_z$,

$$\begin{aligned}\mathbf{J}\mathbf{Q}_z &= \mathbf{J}\mathbf{Y}_{z-1} + \mathbf{J}\mathbf{G}_{z-1} - \mathbf{J}\mathbf{G}_{z-2} \\ &= \mathbf{J}\mathbf{G}_{z-2} + \mathbf{J}\mathbf{G}_{z-1} - \mathbf{J}\mathbf{G}_{z-2} \\ &= \mathbf{J}\mathbf{G}_{z-1}. \end{aligned} \tag{11}$$

Combining (10) and (11), we obtain $\mathbf{J}\mathbf{Y}_z = \mathbf{J}\mathbf{G}_{z-1}$. Since $\mathbf{J}\mathbf{Y}_z = \mathbf{1}\overline{\mathbf{y}}_z^\top$ and $\mathbf{J}\mathbf{G}_{z-1} = \mathbf{1}\overline{\mathbf{g}}_{z-1}^\top$, this is equivalent to $\overline{\mathbf{y}}_z = \overline{\mathbf{g}}_{z-1}$. Thus, our conclusion holds for all blocks $z \geq 1$.

B.2 D-ONO WITH GO

In this section, we prove the regret upper bound of Algorithm 1, ADGT, under exact gradient feedback. In the Appendix B.2.1 and B.2.2, we provide the proofs of Lemmas B.2 and B.3, which establish the upper bound on the averaged local regret and the consensus error, respectively. Finally, in the Appendix B.2.3, we prove Theorem 4.1.

B.2.1 UPPER BOUND ON THE AVERAGED LOCAL REGRET

Lemma B.2. *Under Assumptions 2.1, 2.2, 2.3, with $\gamma \leq \frac{1}{2\sqrt{30}BL}$, Algorithm 1 can achieve*

$$\overline{\mathcal{R}}(T) \leq \frac{24M + 12V_T}{\gamma} + 248BLV_T + \frac{30BL^2}{n}\sum_{z=1}^{N}\|\mathbf{X}_z - \mathbf{1}_n\overline{\mathbf{x}}_z^\top\|_F^2.$$

Proof. Recall that

$$F_{\tau_z}(\mathbf{x}) = \sum_{t\in\tau_z} f_t(\mathbf{x}), \quad f_t(\mathbf{x}) = \frac{1}{n}\sum_{i=1}^{n} f_{i,t}(\mathbf{x}), \quad \overline{\mathbf{x}}_z = \frac{1}{n}\sum_{i=1}^{n}\mathbf{x}_{i,z},$$

$$\overline{\mathbf{g}}_z = \frac{1}{n}\sum_{i=1}^{n}\mathbf{g}_{i,z}(\mathbf{x}_{i,z}) \quad \text{and} \quad \overline{\mathbf{y}}_z = \frac{1}{n}\sum_{i=1}^{n}\mathbf{y}_{i,z}.$$

It begins with the fact that F_{τ_z} is BL-smooth and $\overline{\mathbf{x}}_{z+1} = \overline{\mathbf{x}}_z - \gamma\overline{\mathbf{y}}_z$, we have

$$\begin{aligned} F_{\tau_z}(\overline{\mathbf{x}}_{z+1}) &\le F_{\tau_z}(\overline{\mathbf{x}}_z) - \gamma\left\langle \nabla F_{\tau_z}(\overline{\mathbf{x}}_z), \overline{\mathbf{y}}_z\right\rangle + \frac{BL\gamma^2}{2}\|\overline{\mathbf{y}}_z\|^2 \\ &\le F_{\tau_z}(\overline{\mathbf{x}}_z) - \left(\frac{\gamma}{2} - BL\gamma^2\right)\|\nabla F_{\tau_z}(\overline{\mathbf{x}}_z)\|^2 + \left(\frac{\gamma}{2} + BL\gamma^2\right)\|\overline{\mathbf{y}}_z - \nabla F_{\tau_z}(\overline{\mathbf{x}}_z)\|^2. \quad (12) \end{aligned}$$

Under the condition $\gamma \le \frac{1}{3BL}$, rearranging the terms in (12) yields

$$\|\nabla F_{\tau_z}(\overline{\mathbf{x}}_z)\|^2 \le \frac{6}{\gamma}\left(F_{\tau_z}(\overline{\mathbf{x}}_z) - F_{\tau_z}(\overline{\mathbf{x}}_{z+1})\right) + 5\|\overline{\mathbf{y}}_z - \nabla F_{\tau_z}(\overline{\mathbf{x}}_z)\|^2. \quad (13)$$

Summing (13) over $z = 1, \dots, N$ gives

$$\sum_{z=1}^{N}\|\nabla F_{\tau_z}(\overline{\mathbf{x}}_z)\|^2 \le \frac{6}{\gamma}\sum_{z=1}^{N}\left(F_{\tau_z}(\overline{\mathbf{x}}_z) - F_{\tau_z}(\overline{\mathbf{x}}_{z+1})\right) + 5\sum_{z=1}^{N}\|\overline{\mathbf{y}}_z - \nabla F_{\tau_z}(\overline{\mathbf{x}}_z)\|^2. \quad (14)$$

We next convert the block-level quantity into the time-level regret. Note that

$$\sum_{t\in\tau_z}\|\nabla f_t(\overline{\mathbf{x}}_z)\|^2 \le \frac{1}{B^2}\sum_{t\in\tau_z}\|\nabla F_{\tau_z}(\overline{\mathbf{x}}_z)\|^2 + \sum_{s=1}^{B}\left\|\nabla f_{(z-1)B+s}(\overline{\mathbf{x}}_z) - \frac{1}{B}\nabla F_{\tau_z}(\overline{\mathbf{x}}_z)\right\|^2.$$

Summing the above inequality over $z = 1, \dots, N$ gives

$$\overline{\mathcal{R}}(T) \le \frac{1}{B}\sum_{z=1}^{N}\|\nabla F_{\tau_z}(\overline{\mathbf{x}}_z)\|^2 + \sum_{z=1}^{N}\sum_{s=1}^{B}\left\|\nabla f_{(z-1)B+s}(\overline{\mathbf{x}}_z) - \frac{1}{B}\nabla F_{\tau_z}(\overline{\mathbf{x}}_z)\right\|^2. \quad (15)$$

The second term on the right-hand side captures the temporal dissimilarity of the gradients within each block. Define the per-round function variation as

$$\nu_r := \sup_{\mathbf{x}\in\mathbb{R}^d}\left|f_r(\mathbf{x}) - f_{r-1}(\mathbf{x})\right|, \qquad r = 2, \dots, T.$$

For each block τ_z, let

$$\mathbf{g}_t := \nabla f_t(\overline{\mathbf{x}}_z), \qquad \overline{\mathbf{g}}_z := \frac{1}{B}\nabla F_{\tau_z}(\overline{\mathbf{x}}_z) = \frac{1}{B}\sum_{t\in\tau_z}\mathbf{g}_t.$$

Using the variance–pairwise distance identity, we have

$$\sum_{t\in\tau_z}\|\mathbf{g}_t - \overline{\mathbf{g}}_z\|^2 = \frac{1}{2B}\sum_{s\in\tau_z}\sum_{t\in\tau_z}\|\mathbf{g}_t - \mathbf{g}_s\|^2 = \frac{1}{B}\sum_{\substack{s,t\in\tau_z\\ s<t}}\|\mathbf{g}_t - \mathbf{g}_s\|^2. \quad (16)$$

For any $s < t$, define $h_{s,t}(\mathbf{x}) := f_t(\mathbf{x}) - f_s(\mathbf{x})$. Since both f_t and f_s are L-smooth, $h_{s,t}$ is $2L$-smooth. Moreover,

$$\sup_{\mathbf{x}\in\mathbb{R}^d}|h_{s,t}(\mathbf{x})| = \sup_{\mathbf{x}\in\mathbb{R}^d}|f_t(\mathbf{x}) - f_s(\mathbf{x})| \le \sum_{r=s+1}^{t}\sup_{\mathbf{x}\in\mathbb{R}^d}|f_r(\mathbf{x}) - f_{r-1}(\mathbf{x})| = \sum_{r=s+1}^{t}\nu_r. \quad (17)$$

By the standard self-bounding property of smooth functions,

$$\|\nabla h(\mathbf{x})\|^2 \le 4M\sup_{\mathbf{y}}|h(\mathbf{y})|$$

for any M-smooth function h. Applying this inequality to $h_{s,t}$ with $M = 2L$ yields

$$\|\mathbf{g}_t - \mathbf{g}_s\|^2 = \|\nabla h_{s,t}(\overline{\mathbf{x}}_z)\|^2 \le 8L\sum_{r=s+1}^{t}\nu_r. \quad (18)$$

Suppose that $\tau_z = \{a_z, a_z + 1, \ldots, a_z + B - 1\}$. Combining (16) and (18), we obtain

$$\sum_{t\in\tau_z} \|\mathbf{g}_t - \overline{\mathbf{g}}_z\|^2 \le \frac{8L}{B} \sum_{\substack{s,t\in\tau_z \\ s<t}} \sum_{r=s+1}^{t} \nu_r \le \frac{8L}{B} \cdot \frac{B^2}{2} \sum_{r\in\tau_z} \nu_r = 4BL \sum_{r\in\tau_z} \nu_r, \tag{19}$$

Summing (19) over all blocks, we conclude that

$$\sum_{z=1}^{B} \sum_{t\in\tau_z} \|\mathbf{g}_t - \overline{\mathbf{g}}_z\|^2 \le 4BL \sum_{z=1}^{N} \sum_{t\in\tau_z} \nu_t \le 4BLV_T. \tag{20}$$

Combining (14), (15) with (20) yields

$$\overline{\mathcal{R}}(T) \le \frac{6}{\gamma B} \sum_{z=1}^{N} \left(F_{\tau_z}(\overline{\mathbf{x}}_z) - F_{\tau_z}(\overline{\mathbf{x}}_{z+1})\right) + \frac{5}{B} \sum_{z=1}^{N} \|\overline{\mathbf{y}}_z - \nabla F_{\tau_z}(\overline{\mathbf{x}}_z)\|^2 + 4BLV_T. \tag{21}$$

For the first term, we use the telescoping decomposition

$$\sum_{z=1}^{N} \left(F_{\tau_z}(\overline{\mathbf{x}}_z) - F_{\tau_z}(\overline{\mathbf{x}}_{z+1})\right) = F_{\tau_1}(\overline{\mathbf{x}}_1) - F_{\tau_N}(\overline{\mathbf{x}}_{N+1}) + \sum_{z=1}^{B-1} \left(F_{\tau_{z+1}}(\overline{\mathbf{x}}_{z+1}) - F_{\tau_z}(\overline{\mathbf{x}}_{z+1})\right). \tag{22}$$

By boundedness and temporal variation, the right-hand side of (22) is bounded by $B(2M + V_T)$. Therefore, (21) gives

$$\overline{\mathcal{R}}(T) \le \frac{12M + 6V_T}{\gamma} + \frac{5}{B} \sum_{z=1}^{N} \|\overline{\mathbf{y}}_z - \nabla F_{\tau_z}(\overline{\mathbf{x}}_z)\|^2 + 4BLV_T. \tag{23}$$

It remains to control the tracking error. Since $\overline{\mathbf{y}}_z = \overline{\mathbf{g}}_{z-1}$ as shown in Lemma B.1, by smoothness and the definition of $\mathbf{\Delta}_z^{\mathbf{x}} := \mathbf{X}_z - \mathbf{1}\overline{\mathbf{x}}_z^\top$,

$$\begin{aligned}\|\nabla F_{\tau_z}(\overline{\mathbf{x}}_z) - \overline{\mathbf{y}}_z\|^2 \le{}& 3\|\nabla F_{\tau_z}(\overline{\mathbf{x}}_z) - \nabla F_{\tau_z}(\overline{\mathbf{x}}_{z-1})\|^2 + 3\|\nabla F_{\tau_z}(\overline{\mathbf{x}}_{z-1}) - \nabla F_{\tau_{z-1}}(\overline{\mathbf{x}}_{z-1})\|^2 \\ &+ 3\|F_{\tau_{z-1}}(\overline{\mathbf{x}}_{z-1}) - \overline{\mathbf{y}}_z\|^2.\end{aligned} \tag{24}$$

For the second term of the right-hand side of (24), define $\nu_t := \max_{i\in[n]} \sup_{\mathbf{x}\in\mathbb{R}^d} |f_{i,t}(\mathbf{x}) - f_{i,t-1}(\mathbf{x})|$, we have

$$\begin{aligned}\|\nabla F_{\tau_z}(\overline{\mathbf{x}}_{z-1}) - \nabla F_{\tau_{z-1}}(\overline{\mathbf{x}}_{z-1})\|^2 &= \left\|\sum_{s=1}^{B} \left(\nabla f_{(z-1)B+s}(\overline{\mathbf{x}}_{z-1}) - \nabla f_{(z-2)B+s}(\overline{\mathbf{x}}_{z-1})\right)\right\|^2 \\ &\le B \sum_{s=1}^{B} \left\|\nabla f_{(z-1)B+s}(\overline{\mathbf{x}}_{z-1}) - \nabla f_{(z-2)B+s}(\overline{\mathbf{x}}_{z-1})\right\|^2 \\ &\overset{(i)}{\le} 8BL \sum_{s=1}^{B} \sum_{r=1}^{B} \nu_{(z-2)B+s+r} \le 8B^2L \sum_{t=(z-2)B+1}^{zB} \nu_t,\end{aligned} \tag{25}$$

where (i) comes from the same fact as (18). For the Third term of the right-hand side of (24), it holds that

$$\begin{aligned}\|\nabla F_{\tau_z}(\overline{\mathbf{x}}_z) - \overline{\mathbf{g}}_z\|^2 &= \left\|\frac{1}{n} \sum_{i=1}^{n} \sum_{t\in\tau_z} \left(\nabla f_{i,t}(\overline{\mathbf{x}}_z) - \nabla f_{i,t}(\mathbf{x}_{i,z})\right)\right\|^2 \\ &\le \frac{BL^2}{n} \sum_{i=1}^{n} \sum_{t\in\tau_z} \|\overline{\mathbf{x}}_z - \mathbf{x}_{i,z}\|^2 \le \frac{B^2L^2}{n} \|\mathbf{\Delta}_z^{\mathbf{x}}\|_F^2\end{aligned} \tag{26}$$

for any $z = 1, \ldots, N$. Finally, substituting (25) and (26) into (24) and summing over $z = 1, \ldots, N$ yields

$$\sum_{z=1}^{N} \|\nabla F_{\tau_z}(\overline{\mathbf{x}}_z) - \overline{\mathbf{y}}_z\|^2 \le 3B^2L^2\gamma^2 \sum_{z=1}^{N} \|\overline{\mathbf{y}}_z\|^2 + 24B^2LV_T + \frac{3B^2L^2}{n} \|\mathbf{\Delta}_z^{\mathbf{x}}\|_F^2$$

$$\leq 6B^2L^2\gamma^2\sum_{z=1}^{N}\|\nabla F_{\tau_z}(\overline{\mathbf{x}}_z)\|^2 + 6B^2L^2\gamma^2\sum_{z=1}^{N}\|\nabla F_{\tau_z}(\overline{\mathbf{x}}_z)-\overline{\mathbf{y}}_z\|^2 + 24B^2LV_T + \frac{3B^2L^2}{n}\|\mathbf{\Delta}_z^{\mathbf{x}}\|_F^2. \tag{27}$$

Let $1-6B^2L^2\gamma^2 \geq \frac{1}{2}$, which means $\gamma \leq \frac{1}{2\sqrt{3}BL}$, it follows that

$$\sum_{z=1}^{N}\|\nabla F_{\tau_z}(\overline{\mathbf{x}}_z)-\overline{\mathbf{y}}_z\|^2 \leq 12B^2L^2\gamma^2\sum_{z=1}^{N}\|\nabla F_{\tau_z}(\overline{\mathbf{x}}_z)\|^2 + 24B^2LV_T + \frac{3B^2L^2}{n}\sum_{z=1}^{N}\|\mathbf{\Delta}_z^{\mathbf{x}}\|_F^2$$
$$\leq 12B^3L^2\gamma^2\overline{\mathcal{R}}(T) + 24B^2LV_T + \frac{3B^2L^2}{n}\sum_{z=1}^{N}\|\mathbf{\Delta}_z^{\mathbf{x}}\|_F^2. \tag{28}$$

Substituing (28) into (21), we obtain

$$\overline{\mathcal{R}}(T) \leq \frac{12M+6V_T}{\gamma} + 60B^2L^2\gamma^2\overline{\mathcal{R}}(T) + 120BLV_T + \frac{15BL^2}{n}\sum_{z=1}^{N}\|\mathbf{\Delta}_z^{\mathbf{x}}\|_F^2 + 4BLV_T$$
$$\leq \frac{24M+12V_T}{\gamma} + 248BLV_T + \frac{30BL^2}{n}\sum_{z=1}^{N}\|\mathbf{\Delta}_z^{\mathbf{x}}\|_F^2,$$

where in the last inequality, we set $\gamma \leq \frac{1}{2\sqrt{30}BL}$ to keep $1-60B^2L^2\gamma^2 \geq \frac{1}{2}$. This completes the proof.

□

B.2.2 Upper Bound on the Cumulative Consensus Error

Lemma B.3. *Under Assumptions 2.1, 2.2, 2.3, with* $\gamma < \frac{(1-\lambda_B)^2}{2BL}\sqrt{\frac{C_1}{13}}$, *Algorithm 1 can achieve*

$$\sum_{z=1}^{B}\|\mathbf{X}_z - \mathbf{1}_n\overline{\mathbf{x}}_z^\top\|_F^2 \leq \frac{24nB^3L^2\gamma^4\lambda_B^2}{C_1C_2(1-\lambda_B)^2}\overline{\mathcal{R}}(T) + \frac{32B^2L\lambda_B^2\gamma^2}{C_1C_2(1-\lambda_B)^2}\sum_{i=1}^{n}V_{i,T} + \frac{16nB^2LM\lambda_B^2\gamma^2}{C_1C_2(1-\lambda_B)^2},$$

where $\lambda_B = \sqrt{14}(1-(1-\frac{1}{\sqrt{2}})\sqrt{\rho})^B < 1$ *for some chosen* B *and positive constants* $C_1 = 1-\frac{6B^2L^2\gamma^2\lambda_B^2}{(1-\lambda_B)^2}$, $C_2 = 1-\frac{52B^2L^2\lambda_B^2\gamma^2}{C_1(1-\lambda_B)^4}$.

Proof. Recall that $\mathbf{\Delta}_z^{\mathbf{x}} = (\mathbf{I}-\mathbf{J})\mathbf{X}_z$ and $\mathbf{\Delta}_z^{\mathbf{y}} = (\mathbf{I}-\mathbf{J})\mathbf{Y}_z$, where $\mathbf{J} = \frac{1}{n}\mathbf{1}\mathbf{1}^\top$. $\rho = 1-\sigma_2(\mathbf{W})$ denotes the spectral gap of the communication matrix. From the update rule of $\mathbf{X}_z$, we have

$$\mathbf{X}_{z+1} = \mathbf{U}_z^B - \gamma\mathbf{Y}_z,$$

where $\mathbf{U}_z^B = \mathbf{P}_B\mathbf{X}_z$, and the accelerated mixing matrix $\mathbf{P}_B$ satisfies

$$\|\mathbf{P}_B - \mathbf{J}\|_2 \leq \lambda_B := \sqrt{14}\left(1-\left(1-\frac{1}{\sqrt{2}}\right)\sqrt{\rho}\right)^B < 1.$$

Since $\mathbf{P}_B$ is doubly stochastic, it holds that $\mathbf{P}_B\mathbf{J} = \mathbf{J}\mathbf{P}_B = \mathbf{J}$. Therefore,

$$\begin{aligned}\mathbf{\Delta}_{z+1}^{\mathbf{x}} &= (\mathbf{I}-\mathbf{J})\mathbf{X}_{z+1}\\ &= (\mathbf{I}-\mathbf{J})\mathbf{P}_B\mathbf{X}_z - \gamma(\mathbf{I}-\mathbf{J})\mathbf{Y}_z\\ &= (\mathbf{P}_B-\mathbf{J})\mathbf{\Delta}_z^{\mathbf{x}} - \gamma\mathbf{\Delta}_z^{\mathbf{y}}.\end{aligned}$$

Consequently,

$$\|\mathbf{\Delta}_{z+1}^{\mathbf{x}}\|_F \leq \lambda_B\|\mathbf{\Delta}_z^{\mathbf{x}}\|_F + \gamma\|\mathbf{\Delta}_z^{\mathbf{y}}\|_F.$$

For any $a, b \geq 0$, we have

$$(\lambda_B a + \gamma b)^2 \leq \lambda_B a^2 + \frac{\gamma^2}{1-\lambda_B}b^2,$$

since

$$\lambda_B a^2 + \frac{\gamma^2}{1-\lambda_B} b^2 - (\lambda_B a + \gamma b)^2 = \lambda_B \left(\sqrt{1-\lambda_B}\, a - \frac{\gamma}{\sqrt{1-\lambda_B}} b \right)^2 \geq 0.$$

Thus it follows that

$$\|\mathbf{\Delta}^{\mathbf{x}}_{z+1}\|_F^2 \leq \lambda_B \|\mathbf{\Delta}^{\mathbf{x}}_{z}\|_F^2 + \frac{\gamma^2}{1-\lambda_B} \|\mathbf{\Delta}^{\mathbf{y}}_{z}\|_F^2.$$

Summing the above inequality over $z = 1, \ldots, N-1$ and using the consensual initialization $\mathbf{\Delta}^{\mathbf{x}}_1 = \mathbf{\Delta}^{\mathbf{x}}_0 = \mathbf{0}$, we obtain

$$\sum_{z=1}^{B} \|\mathbf{\Delta}^{\mathbf{x}}_{z}\|_F^2 \leq \lambda_B \sum_{z=1}^{B} \|\mathbf{\Delta}^{\mathbf{x}}_{z}\|_F^2 + \frac{\gamma^2}{1-\lambda_B} \sum_{z=1}^{B} \|\mathbf{\Delta}^{\mathbf{y}}_{z}\|_F^2.$$

Rearranging the terms yields

$$\sum_{z=1}^{B} \|\mathbf{\Delta}^{\mathbf{x}}_{z}\|_F^2 \leq \frac{\gamma^2}{(1-\lambda_B)^2} \sum_{z=1}^{B} \|\mathbf{\Delta}^{\mathbf{y}}_{z}\|_F^2. \tag{29}$$

We next bound the tracking consensus error. By the accelerated gossip contraction, it satisfies

$$\begin{aligned} \|\mathbf{\Delta}^{\mathbf{y}}_{z+1}\|_F &= \|(\mathbf{I}-\mathbf{J})\mathbf{Y}_{z+1}\|_F \\ &\leq \lambda_B \left\|(\mathbf{I}-\mathbf{J})\left(\mathbf{Y}_z + \mathbf{G}_z - \mathbf{G}_{z-1}\right)\right\|_F \\ &\leq \lambda_B \|\mathbf{\Delta}^{\mathbf{y}}_{z}\|_F + \lambda_B \|(\mathbf{I}-\mathbf{J})(\mathbf{G}_z - \mathbf{G}_{z-1})\|_F. \end{aligned} \tag{30}$$

Under the consensual initialization $\mathbf{\Delta}^{\mathbf{y}}_1 = 0$, recursively applying (30) and using the discrete Young inequality for convolutions and $\sum_{r=1}^{\infty} \lambda_B^r = \frac{\lambda_B}{1-\lambda_B}$, we obtain

$$\begin{aligned} \sum_{z=1}^{B} \|\mathbf{\Delta}^{\mathbf{y}}_{z}\|_F^2 &\leq \frac{\lambda_B^2}{(1-\lambda_B)^2} \sum_{z=1}^{B-1} \|(\mathbf{I}-\mathbf{J})(\mathbf{G}_z - \mathbf{G}_{z-1})\|_F^2 \\ &\leq \frac{\lambda_B^2}{(1-\lambda_B)^2} \sum_{z=1}^{B} \|\mathbf{G}_z - \mathbf{G}_{z-1}\|_F^2, \end{aligned} \tag{31}$$

where the last inequality follows from $\|\mathbf{I}-\mathbf{J}\|_2 \leq 1$. For each $z \geq 2$, define

$$\mathbf{A}_{i,z} = \mathbf{g}_{i,z}(x_{i,z}) - \mathbf{g}_{i,z}(\mathbf{x}_{i,z-1}), \qquad \mathbf{B}_{i,z} = \mathbf{g}_{i,z}(\mathbf{x}_{i,z-1}) - \mathbf{g}_{i,z-1}(\mathbf{x}_{i,z-1}).$$

Consequently,

$$\|\mathbf{G}_z - \mathbf{G}_{z-1}\|_F^2 \leq 2\sum_{i=1}^{n} \|\mathbf{A}_{i,z}\|^2 + 2\sum_{i=1}^{n} \|\mathbf{B}_{i,z}\|^2.$$

Since each block contains L loss functions and every $f_{i,t}$ is L-smooth,

$$\sum_{z=2}^{N}\sum_{i=1}^{n} \|\mathbf{A}_{i,z}\|^2 \leq B^2 L^2 \sum_{z=2}^{N} \|\mathbf{X}_z - \mathbf{X}_{z-1}\|_F^2.$$

Moreover, recall (25), similarily we have

$$\sum_{z=2}^{N}\sum_{i=1}^{n} \|\mathbf{B}_{i,z}\|^2 \leq 8B^2 L \sum_{i=1}^{n} V_{i,T}, \qquad \sum_{i=1}^{n} \|\mathbf{B}_{i,1}\|^2 \leq 4nB^2 LM.$$

Combining the above estimates yields

$$\sum_{z=1}^{B} \|\mathbf{G}_z - \mathbf{G}_{z-1}\|_F^2 \leq 2B^2 L^2 \sum_{z=2}^{N} \|\mathbf{X}_z - \mathbf{X}_{z-1}\|_F^2 + 16B^2 L \sum_{i=1}^{n} V_{i,T} + 8nB^2 LM. \tag{32}$$

Substituting (32) into (31), we finally obtain

$$\sum_{z=1}^{B} \|\mathbf{\Delta}^{\mathbf{y}}_{z}\|_F^2 \leq \frac{2B^2 L^2 \lambda_B^2}{(1-\lambda_B)^2} \sum_{z=2}^{N} \|\mathbf{X}_z - \mathbf{X}_{z-1}\|_F^2 + \frac{16B^2 L \lambda_B^2}{(1-\lambda_B)^2} \sum_{i=1}^{n} V_{i,T} + \frac{8nB^2 LM\lambda_B^2}{(1-\lambda_B)^2}. \tag{33}$$

We next bound the movement of the local iterates. From the model update $\mathbf{X}_{z+1} = \mathbf{U}_z^B - \gamma \mathbf{Y}_z$, we have

$$\mathbf{X}_{z+1} - \mathbf{X}_z = \mathbf{U}_z^B - \mathbf{X}_z - \gamma \mathbf{Y}_z.$$

Recall that the L-step accelerated gossip operator preserves the network average and contracts the disagreement, i.e.,

$$\mathbf{J}\mathbf{U}_z^B = \mathbf{J}\mathbf{X}_z, \qquad \|(\mathbf{I}-\mathbf{J})\mathbf{U}_z^B\|_F \le \lambda_B \|(\mathbf{I}-\mathbf{J})\mathbf{X}_z\|_F.$$

Therefore,

$$\begin{aligned}\|\mathbf{U}_z^B - \mathbf{X}_z\|_F &= \|\mathbf{U}_z^B - \mathbf{J}\mathbf{U}_z^B + \mathbf{J}\mathbf{X}_z - \mathbf{X}_z\|_F \\ &\le \|(\mathbf{I}-\mathbf{J})\mathbf{U}_z^B\|_F + \|(\mathbf{I}-\mathbf{J})\mathbf{X}_z\|_F \\ &\le (1+\lambda_B)\|\mathbf{\Delta}_z^{\mathbf{x}}\|_F.\end{aligned}$$

Moreover, by the average-preserving property of the gradient-tracking update, $\mathbf{J}\mathbf{Y}_z = \mathbf{1}\overline{\mathbf{g}}_{z-1}^\top$. Hence, $\mathbf{Y}_z = \mathbf{\Delta}_z^y + \mathbf{1}\overline{\mathbf{g}}_{z-1}^\top$. Substituting this decomposition into the model difference gives

$$\mathbf{X}_{z+1} - \mathbf{X}_z = \mathbf{U}_z^B - \mathbf{X}_z - \gamma \mathbf{\Delta}_z^y - \gamma \mathbf{1}\overline{\mathbf{g}}_{z-1}^\top.$$

Thus we obtain

$$\begin{aligned}\|\mathbf{X}_{z+1} - \mathbf{X}_z\|_F^2 &\le 3(1+\lambda_B)^2\|\mathbf{\Delta}_z^{\mathbf{x}}\|_F^2 + 3\gamma^2\|\mathbf{\Delta}_z^y\|_F^2 + 3\gamma^2\|\mathbf{1}\overline{\mathbf{g}}_{z-1}^\top\|_F^2 \\ &\le 12\|\mathbf{\Delta}_z^{\mathbf{x}}\|_F^2 + 3\gamma^2\|\mathbf{\Delta}_z^y\|_F^2 + 3n\gamma^2\|\overline{\mathbf{g}}_{z-1}\|^2, \end{aligned} \tag{34}$$

where the last inequality comes from $\|\mathbf{1}\overline{\mathbf{g}}_{z-1}^\top\|_F^2 = n\|\overline{\mathbf{g}}_{z-1}\|^2$ and $\lambda_B \le 1$. With

$$\begin{aligned}\sum_{z=1}^N \|\overline{\mathbf{g}}_{z-1}\|^2 \le \sum_{z=1}^N \|\overline{\mathbf{g}}_z\|^2 &\le 2\sum_{z=1}^N \|\nabla F_{\tau_z}(\overline{\mathbf{x}}_z)\|^2 + 2\sum_{z=1}^N \|\overline{\mathbf{g}}_z - \nabla F_{\tau_z}(\overline{\mathbf{x}}_z)\|^2 \\ &\le 2B\overline{\mathcal{R}}(T) + \frac{2B^2L^2}{n}\sum_{z=1}^N \|\mathbf{\Delta}_z^{\mathbf{x}}\|_F^2 \end{aligned} \tag{35}$$

and summing (34) over z, we obtain

$$\sum_{z=1}^N \|\mathbf{X}_{z+1} - \mathbf{X}_z\|_F^2 \le (12 + 6B^2L^2\gamma^2)\|\mathbf{\Delta}_z^{\mathbf{x}}\|_F^2 + 3\gamma^2\|\mathbf{\Delta}_z^y\|_F^2 + 6nB\gamma^2\overline{\mathcal{R}}(T). \tag{36}$$

Substituting (36) into (33) yields

$$\begin{aligned}\sum_{z=1}^B \|\mathbf{\Delta}_z^y\|_F^2 &\le \frac{26B^2L^2\lambda_B^2}{(1-\lambda_B)^2}\sum_{z=1}^B \|\mathbf{\Delta}_z^x\|_F^2 + \frac{6B^2L^2\gamma^2\lambda_B^2}{(1-\lambda_B)^2}\sum_{z=1}^B \|\mathbf{\Delta}_z^y\|_F^2 + \frac{12nB^3L^2\gamma^2\lambda_B^2}{(1-\lambda_B)^2}\overline{\mathcal{R}}(T) \\ &\quad + \frac{16B^2L\lambda_B^2}{(1-\lambda_B)^2}\sum_{i=1}^n V_{i,T} + \frac{4nB^2LM\lambda_B^2}{(1-\lambda_B)^2} \\ &\le \frac{26B^2L^2\lambda_B^2}{C_1(1-\lambda_B)^2}\sum_{z=1}^B \|\mathbf{\Delta}_z^x\|_F^2 + \frac{12nB^3L^2\gamma^2\lambda_B^2}{C_1(1-\lambda_B)^2}\overline{\mathcal{R}}(T) + \frac{16B^2L\lambda_B^2}{C_1(1-\lambda_B)^2}\sum_{i=1}^n V_{i,T} \\ &\quad + \frac{8nB^2LM\lambda_B^2}{C_1(1-\lambda_B)^2}, \end{aligned} \tag{37}$$

where in the last inequality, we set $C_1 := 1 - \frac{6B^2L^2\gamma^2\lambda_B^2}{(1-\lambda_B)^2} > 1 - \frac{6B^2L^2\gamma^2}{(1-\lambda_B)^2}$ and $\gamma < \frac{1-\lambda_B}{\sqrt{6}BL}$ to keep it positive. Finally, combining (37) and (29), we obtain

$$\begin{aligned}\sum_{z=1}^B \|\mathbf{\Delta}_z^x\|_F^2 &\le \frac{52B^2L^2\lambda_B^2\gamma^2}{C_1(1-\lambda_B)^4}\sum_{z=1}^B \|\mathbf{\Delta}_z^x\|_F^2 + \frac{24nB^3L^2\gamma^4\lambda_B^2}{C_1(1-\lambda_B)^2}\overline{\mathcal{R}}(T) + \frac{32B^2L\lambda_B^2\gamma^2}{C_1(1-\lambda_B)^2}\sum_{i=1}^n V_{i,T} \\ &\le \frac{24nB^3L^2\gamma^4\lambda_B^2}{C_1C_2(1-\lambda_B)^2}\overline{\mathcal{R}}(T) + \frac{32B^2L\lambda_B^2\gamma^2}{C_1C_2(1-\lambda_B)^2}\sum_{i=1}^n V_{i,T} + \frac{16nB^2LM\lambda_B^2\gamma^2}{C_1C_2(1-\lambda_B)^2}, \end{aligned}$$

where $C_2 := 1 - \frac{52B^2L^2\lambda_B^2\gamma^2}{C_1(1-\lambda_B)^4} > 0$ if we set $\gamma < \frac{(1-\lambda_B)^2}{2BL}\sqrt{\frac{C_1}{13}}$. Thus we finish the proof.

□

B.2.3 Proof of Theorem 4.1

Theorem B.4 (Restatement of Theorem 4.1)**.** *Under Assumptions 2.1, 2.2, 2.3, let* $L = \Theta(\rho^{-1/2}\log\sqrt{n})$ *and* $\gamma < \frac{\sqrt{(1-\lambda_B)}}{2BL}\sqrt[4]{\frac{C_1C_2}{45}}$ *with some constants* $C_1, C_2 > 0$*, for each node* $i \in [n]$*, Algorithm 1 can gaurantee*

$$\mathcal{R}_i(T) \le 2\overline{\mathcal{R}}(T) + \frac{48nB^4L^4\gamma^4\lambda_B^2}{C_1C_2(1-\lambda_B)^4}\overline{\mathcal{R}}(T) + \frac{64nB^3L^3\gamma^2\lambda_B^2}{C_1C_2(1-\lambda_B)^4}V_T \le \tilde{O}\left(\rho^{-1/2}(1+V_T)\right).$$

Proof. We first give the regret upper bound on the averaged node, which comes from the combination of Lemma B.2 and Lemma B.3, shown as

$$\begin{aligned}\overline{\mathcal{R}}(T) &\le \frac{24M+12V_T}{\gamma} + 248BLV_T + \frac{30BL^2}{n}\sum_{z=1}^{N}\|\boldsymbol{\Delta}_z^x\|_F^2 \\ &\le \frac{24M+12V_T}{\gamma} + 248BLV_T + \frac{720B^4L^4\gamma^4\lambda_B^2}{C_1C_2(1-\lambda_B)^2}\overline{\mathcal{R}}(T) + \frac{960B^3L^3\lambda_B^2\gamma^2}{C_1C_2(1-\lambda_B)^2}V_T + O(1).\end{aligned}$$

After rearranging terms and let $\gamma < \frac{\sqrt{(1-\lambda_B)}}{2BL}\sqrt[4]{\frac{C_1C_2}{45}} = O(B^{-1})$ to keep the left-hand side positive, $B = \Theta(\rho^{-1/2}\log\sqrt{n})$ yields

$$\overline{\mathcal{R}}(T) \le \tilde{O}\left(\rho^{-1/2}(1+V_T)\right).$$

We next convert the average-node regret into the individual regret of each node $i \in [n]$. By the L-smoothness of f_t, for any node i and any time $t \in [T]$,

$$\|\nabla f_t(\mathbf{x}_{i,t})\|^2 \le 2\|\nabla f_t(\overline{\mathbf{x}}_t)\|^2 + 2L^2\|\mathbf{x}_{i,t} - \overline{\mathbf{x}}_t\|^2.$$

Summing over $t = 1, \dots, T$ yields

$$\mathcal{R}_i(T) \le 2\overline{\mathcal{R}}(T) + 2L^2\sum_{t=1}^{T}\|\mathbf{x}_{i,t} - \overline{\mathbf{x}}_t\|^2 \le 2\overline{\mathcal{R}}(T) + 2BL^2\sum_{z=1}^{B}\|\boldsymbol{\Delta}_z^x\|_F^2. \tag{38}$$

Finally, recall that $\gamma \le O(B^{-1})$, $B = \Theta(\rho^{-1/2}\log\sqrt{n})$, which means $\lambda_B^2 = \Theta(n^{-1})$, substituting Lemma B.3 into 38 yields

$$\mathcal{R}_i(T) \le 2\overline{\mathcal{R}}(T) + \frac{48nB^4L^4\gamma^4\lambda_B^2}{C_1C_2(1-\lambda_B)^4}\overline{\mathcal{R}}(T) + \frac{64nB^3L^3\gamma^2\lambda_B^2}{C_1C_2(1-\lambda_B)^4}V_T + O(1) \le \tilde{O}\left(\rho^{-1/2}(1+V_T)\right).$$

Thus we finish the proof. □

B.3 D-ONO with SGO

In this section, we prove the local regret upper bound for ADGT(Algorithm 1) under stochastic gradient feedback. Appendix B.3.1 establishes an upper bound on the averaged local regret in Lemma B.5, and Appendix B.3.2 provides the proof of Theorem 4.2.

B.3.1 Upper Bound on the Averaged Local Regret

Lemma B.5. *Under Assumptions 2.1-2.4, with* $\gamma < \frac{1}{4\sqrt{15}BL}$*, Algorithm 1 ca achieve*

$$\mathbb{E}\left[\overline{\mathcal{R}}(T)\right] \le \frac{24M+12V_T}{\gamma} + \frac{60BL^2}{n}\sum_{z=1}^{N}\mathbb{E}\left[\|\boldsymbol{\Delta}_z^{\mathbf{x}}\|_F^2\right] + 126L^2\gamma\sum_{z=1}^{N}\mathbb{E}\left[\|\overline{\mathbf{e}}_z\|^2\right] + 968BLV_T.$$

Proof. We prove the averaged-node regret bound by following the deterministic proof and tracking the additional stochastic errors. For each block z, decompose the stochastic block gradient into the true-gradient part and the noise part:

$$\mathbf{g}_{i,z}(\mathbf{x}_{i,z}) = \mathbf{g}_{i,z}^{\mathrm{true}}(\mathbf{x}_{i,z}) + \mathbf{e}_{i,z}, \qquad \mathbf{e}_{i,z} := \sum_{t\in\tau_z}\boldsymbol{\xi}_{i,t}.$$

Correspondingly, define

$$\overline{\mathbf{g}}_z^{\text{true}} := \frac{1}{n}\sum_{i=1}^{n} \mathbf{g}_{i,z}^{\text{true}}(\mathbf{x}_{i,z}), \qquad \overline{\mathbf{e}}_z := \frac{1}{n}\sum_{i=1}^{n} \mathbf{e}_{i,z},$$

so that $\overline{\mathbf{g}}_z = \overline{\mathbf{g}}_z^{\text{true}} + \overline{\mathbf{e}}_z$. Moreover, let

$$\mathbf{E}_z := \begin{bmatrix} \mathbf{e}_{1,z}^\top \\ \vdots \\ \mathbf{e}_{n,z}^\top \end{bmatrix}, \qquad \boldsymbol{\Delta}_{\mathbf{E},z} := \mathbf{E}_z - \mathbf{E}_{z-1}.$$

By Lemma B.1, the stochastic tracking variable satisfies

$$\overline{\mathbf{y}}_z = \overline{\mathbf{g}}_{z-1} = \overline{\mathbf{g}}_{z-1}^{\text{true}} + \overline{\mathbf{e}}_{z-1},$$

where we use the convention $\overline{\mathbf{g}}_0 = \overline{\mathbf{g}}_0^{\text{true}} = \overline{\mathbf{e}}_0 = \mathbf{0}$. Thus, the averaged update is

$$\overline{\mathbf{x}}_{z+1} = \overline{\mathbf{x}}_z - \gamma \overline{\mathbf{y}}_z,$$

Following the deterministic case, since F_{τ_z} is BL-smooth, we have

$$F_{\tau_z}(\overline{\mathbf{x}}_{z+1}) \le F_{\tau_z}(\overline{\mathbf{x}}_z) - \gamma \left\langle \nabla F_{\tau_z}(\overline{\mathbf{x}}_z), \overline{\mathbf{y}}_z \right\rangle + \frac{BL\gamma^2}{2}\|\overline{\mathbf{y}}_z\|^2. \tag{39}$$

Taking conditional expectation with respect to a history that contains all randomness before block $z-1$ but excludes the oracle noise in block $z-1$, and using $\mathbb{E}\left[\overline{\mathbf{e}}_{z-1} \mid \mathcal{H}_z^-\right] = \mathbf{0}$, we obtain

$$\begin{aligned}\mathbb{E}\left[F_{\tau_z}(\overline{\mathbf{x}}_{z+1}) \mid \mathcal{H}_z^-\right] \le{}& F_{\tau_z}(\overline{\mathbf{x}}_z) - \gamma \left\langle \nabla F_{\tau_z}(\overline{\mathbf{x}}_z), \overline{\mathbf{g}}_{z-1}^{\text{true}} \right\rangle \\ &+ \frac{BL\gamma^2}{2}\|\overline{\mathbf{g}}_{z-1}^{\text{true}}\|^2 + \frac{BL\gamma^2}{2}\mathbb{E}\left[\|\overline{\mathbf{e}}_{z-1}\|^2 \mid \mathcal{H}_z^-\right].\end{aligned} \tag{40}$$

Applying the same quadratic decomposition as in the deterministic case gives

$$\begin{aligned}\|\nabla F_{\tau_z}(\overline{\mathbf{x}}_z)\|^2 \le{}& \frac{6}{\gamma}\left(F_{\tau_z}(\overline{\mathbf{x}}_z) - \mathbb{E}\left[F_{\tau_z}(\overline{\mathbf{x}}_{z+1}) \mid \mathcal{H}_z^-\right]\right) \\ &+ 5\left\|\overline{\mathbf{g}}_{z-1}^{\text{true}} - \nabla F_{\tau_z}(\overline{\mathbf{x}}_z)\right\|^2 + 3BL\gamma\mathbb{E}\left[\|\overline{\mathbf{e}}_{z-1}\|^2 \mid \mathcal{H}_z^-\right],\end{aligned} \tag{41}$$

where we used $\gamma \le 1/(3BL)$. Next, we convert the block-level quantity into the time-level regret. The block-to-time conversion itself is unchanged, because $\overline{\mathbf{x}}_z$ is fixed inside block τ_z. Combining the block-to-time conversion with (41), summing over z, taking total expectation, and using the same telescoping argument gives

$$\begin{aligned}\mathbb{E}\left[\overline{\mathcal{R}}(T)\right] \le{}& \frac{12M + 6V_T}{\gamma} + \frac{5}{B}\sum_{z=1}^{B}\mathbb{E}\left[\left\|\overline{\mathbf{g}}_{z-1}^{\text{true}} - \nabla F_{\tau_z}(\overline{\mathbf{x}}_z)\right\|^2\right] \\ &+ 3L\gamma\sum_{z=1}^{B}\mathbb{E}\left[\|\overline{\mathbf{e}}_{z-1}\|^2\right] + 4BLV_T.\end{aligned} \tag{42}$$

Since $\overline{\mathbf{e}}_0 = \mathbf{0}$, the noise term satisfies

$$\sum_{z=1}^{B}\mathbb{E}\left[\|\overline{\mathbf{e}}_{z-1}\|^2\right] \le \sum_{z=1}^{B}\mathbb{E}\left[\|\overline{\mathbf{e}}_z\|^2\right].$$

It remains to control the delayed true-gradient error

$$\overline{\mathbf{g}}_{z-1}^{\text{true}} - \nabla F_{\tau_z}(\overline{\mathbf{x}}_z).$$

For $z \ge 2$, decompose it as

$$\overline{\mathbf{g}}_{z-1}^{\text{true}} - \nabla F_{\tau_z}(\overline{\mathbf{x}}_z) = \underbrace{\overline{\mathbf{g}}_{z-1}^{\text{true}} - \nabla F_{\tau_{z-1}}(\overline{\mathbf{x}}_{z-1})}_{\mathbf{A}_z} + \underbrace{\nabla F_{\tau_{z-1}}(\overline{\mathbf{x}}_{z-1}) - \nabla F_{\tau_{z-1}}(\overline{\mathbf{x}}_z)}_{\mathbf{B}_z}$$

$$+\underbrace{\nabla F_{\tau_{z-1}}(\overline{\mathbf{x}}_z) - \nabla F_{\tau_z}(\overline{\mathbf{x}}_z)}_{\mathbf{C}_z}. \tag{43}$$

We bound the three terms separately. First, by L-smoothness,

$$\|\mathbf{A}_z\|^2 = \left\| \frac{1}{n} \sum_{i=1}^{n} \sum_{t \in \tau_{z-1}} (\nabla f_{i,t}(\mathbf{x}_{i,z-1}) - \nabla f_{i,t}(\overline{\mathbf{x}}_{z-1})) \right\|^2 \leq \frac{B^2 L^2}{n} \|\Delta_{z-1}^x\|_F^2. \tag{44}$$

Second, since $F_{\tau_{z-1}}$ is BL-smooth,

$$\|\mathbf{B}_z\|^2 \leq B^2 L^2 \|\overline{\mathbf{x}}_z - \overline{\mathbf{x}}_{z-1}\|^2 = B^2 L^2 \gamma^2 \|\overline{\mathbf{y}}_{z-1}\|^2 \leq 2B^2 L^2 \gamma^2 \|\overline{\mathbf{g}}_{z-2}^{\text{true}}\|^2 + 2B^2 L^2 \gamma^2 \|\overline{\mathbf{e}}_{z-2}\|^2 \tag{45}$$

This is the movement error caused by the one-block delay. Summing (45) over z yields

$$\begin{aligned} \sum_{z=1}^{N} \|\mathbf{B}_z\|^2 &\leq 2B^2 L^2 \gamma^2 \sum_{z=1}^{N} \|\overline{\mathbf{g}}_{z-1}^{\text{true}}\|^2 + 2B^2 L^2 \gamma^2 \sum_{z=1}^{N} \|\overline{\mathbf{e}}_z\|^2 \\ &\leq 4B^2 L^2 \gamma^2 \sum_{z=1}^{N} \|\overline{\mathbf{g}}_{z-1}^{\text{true}} - \nabla F_{\tau_z}(\overline{\mathbf{x}}_z)\|^2 + 4B^3 L^2 \gamma^2 \overline{\mathcal{R}}(T) + 2B^2 L^2 \gamma^2 \sum_{z=1}^{N} \|\overline{\mathbf{e}}_z\|^2 \end{aligned} \tag{46}$$

Third, $\mathbf{C}_z$ is the temporal block-variation error. Let

$$H_z(\mathbf{x}) := F_{\tau_z}(\mathbf{x}) - F_{\tau_{z-1}}(\mathbf{x}).$$

Then $\mathbf{C}_z = -\nabla H_z(\overline{\mathbf{x}}_z)$. The function H_z is $2BL$-smooth. Moreover, using the adjacent variation budget,

$$\sup_{\mathbf{x}} |H_z(\mathbf{x})| \leq B \sum_{t=(z-2)B+2}^{zB} v_t.$$

Therefore, by the similar process as in 25, it holds that

$$\|\mathbf{C}_z\|^2 \leq 8BL \sup_{\mathbf{x}} |H_z(\mathbf{x})| \leq 8B^2 L \sum_{t=(z-2)B+2}^{zB} v_t.$$

Summing over z gives

$$\sum_{z=2}^{N} \|\mathbf{C}_z\|^2 \leq 16B^2 L V_T. \tag{47}$$

The boundary term $z = 1$ is absorbed into the initial term $(12M + 6V_T)/\gamma$ under the standard stepsize condition $\gamma \leq 1/(BL)$. Combining (43), (44), (46), and (47), we obtain

$$\begin{aligned} &\sum_{z=1}^{B} \mathbb{E}\left[\left\| \overline{\mathbf{g}}_{z-1}^{\text{true}} - \nabla F_{\tau_z}(\overline{\mathbf{x}}_z) \right\|^2 \right] \\ &\leq \frac{3B^2 L^2}{n} \sum_{z=1}^{N} \mathbb{E}\left[\|\mathbf{\Delta}_z^{\mathbf{x}}\|_F^2 \right] + 12B^2 L^2 \gamma^2 \sum_{z=1}^{B} \mathbb{E}\left[\left\| \overline{\mathbf{g}}_{z-1}^{\text{true}} - \nabla F_{\tau_z}(\overline{\mathbf{x}}_z) \right\|^2 \right] \\ &\quad + 12B^3 L^2 \gamma^2 \mathbb{E}\left[\overline{\mathcal{R}}(T) \right] + 6B^2 L^2 \gamma^2 \sum_{z=1}^{N} \mathbb{E}\left[\|\overline{\mathbf{e}}_z\|^2 \right] + 48B^2 L V_T \\ &\leq \frac{6B^2 L^2}{n} \sum_{z=1}^{N} \|\mathbf{\Delta}_z^{\mathbf{x}}\|_F^2 + 24B^3 L^2 \gamma^2 \mathbb{E}\left[\overline{\mathcal{R}}(T) \right] + 12B^2 L^2 \gamma^2 \sum_{z=1}^{N} \mathbb{E}\left[\|\overline{\mathbf{e}}_z\|^2 \right] + 96B^2 L V_T, \end{aligned} \tag{48}$$

where in the last inequality, we let $\gamma \leq \frac{1}{2\sqrt{6}BL}$ to keep $1 - 12B^2 L^2 \gamma^2 \geq \frac{1}{2}$. Substituting (48) into (42) yields

$$\mathbb{E}\left[\overline{\mathcal{R}}(T) \right] \leq \frac{12M + 6V_T}{\gamma} + \frac{30BL^2}{n} \sum_{z=1}^{N} \mathbb{E}\left[\|\mathbf{\Delta}_z^{\mathbf{x}}\|_F^2 \right] + 120B^2 L^2 \gamma^2 \mathbb{E}\left[\overline{\mathcal{R}}(T) \right]$$

$$+ 63L^2\gamma \sum_{z=1}^{N} \mathbb{E}\left[\|\overline{\mathbf{e}}_z\|^2\right] + 484BLV_T. \tag{49}$$

Let $\gamma < \frac{1}{4\sqrt{15}BL}$ and rearrange the terms, we finish the proof by obtaining

$$\mathbb{E}\left[\overline{\mathcal{R}}(T)\right] \leq \frac{24M + 12V_T}{\gamma} + \frac{60BL^2}{n} \sum_{z=1}^{N} \mathbb{E}\left[\|\mathbf{\Delta}_z^{\mathbf{x}}\|_F^2\right] + 126L^2\gamma \sum_{z=1}^{N} \mathbb{E}\left[\|\overline{\mathbf{e}}_z\|^2\right] + 968BLV_T.$$

Thus we finish the proof. □

B.3.2 Proof of Theorem 4.2

We first establish the upper bound on the consensus error of Algorithm 1 under the stochastic gradient feedback in the following lemma.

Lemma B.6. *Under Assumptions 2.1-2.4, with $\gamma < \frac{1}{2BL}\sqrt{\frac{C_1}{13}}$, Algorithm 1 can achieve*

$$\begin{aligned}\sum_{z=1}^{B} \mathbb{E}\left[\|\mathbf{\Delta}_z^{\mathbf{x}}\|_F^2\right] \leq{}& \frac{24nB^3L^2\gamma^4\lambda_B^2}{C_1C_2(1-\lambda_B)^4}\mathbb{E}[\overline{\mathcal{R}}(T)] + \frac{12nB^2L^2\gamma^4\lambda_B^2}{C_1C_2(1-\lambda_B)^4} \sum_{z=1}^{B} \mathbb{E}\left[\|\overline{\mathbf{e}}_z\|^2\right] \\ &+ \frac{32B^2L\lambda_B^2\gamma^2}{C_1C_2(1-\lambda_B)^4} \sum_{i=1}^{n} V_{i,T} + \frac{8nB^2LM\lambda_B^2\gamma^2}{C_1C_2(1-\lambda_B)^4} \\ &+ \frac{2\gamma^2\lambda_B^2}{C_1C_2(1-\lambda_B)^4} \sum_{z=1}^{B} \mathbb{E}\left[\|\mathbf{\Delta}_{\mathbf{E},z}\|_F^2\right],\end{aligned}$$

where $\lambda_B = \sqrt{14}(1-(1-\frac{1}{\sqrt{2}})\sqrt{\rho})^B < 1$ for some chosen L and positive constants $C_1 = 1 - \frac{12B^2L^2\gamma^2\lambda_B^2}{(1-\lambda_B)^2}$, $C_2 = 1 - \frac{52B^2L^2\lambda_B^2\gamma^2}{C_1(1-\lambda_B)^4}$.

Proof. Similar in deterministic case, recall that (29), we have

$$\sum_{z=1}^{B} \mathbb{E}\left[\|\mathbf{\Delta}_z^{\mathbf{x}}\|_F^2\right] \leq \frac{\gamma^2}{(1-\lambda_B)^2} \sum_{z=1}^{B} \mathbb{E}\left[\|\mathbf{\Delta}_z^{\mathbf{y}}\|_F^2\right]. \tag{50}$$

Since the gradient increment is included inside the accelerated-gossip operator, the tracking recursion gives

$$\sum_{z=1}^{B} \mathbb{E}\left[\|\mathbf{\Delta}_z^{\mathbf{y}}\|_F^2\right] \leq \frac{\lambda_B^2}{(1-\lambda_B)^2} \sum_{z=1}^{B} \mathbb{E}\left[\|\mathbf{\Delta}_{\mathbf{G},z}\|_F^2\right], \tag{51}$$

where $\mathbf{\Delta}_{\mathbf{G},z} := \mathbf{G}_z - \mathbf{G}_{z-1}$. Writing

$$\mathbf{G}_z = \mathbf{G}_z^{\text{true}} + \mathbf{E}_z, \qquad \mathbf{\Delta}_{\mathbf{E},z} = \mathbf{E}_z - \mathbf{E}_{z-1},$$

we have

$$\mathbf{\Delta}_{\mathbf{G},z} = \mathbf{\Delta}_{\mathbf{G},z}^{\text{true}} + \mathbf{\Delta}_{\mathbf{E},z}.$$

Consequently, as similar in (32), it follows that

$$\begin{aligned}\sum_{z=1}^{B} \mathbb{E}\left[\|\mathbf{\Delta}_{\mathbf{G},z}\|_F^2\right] &\leq 2\sum_{z=1}^{B} \mathbb{E}\left[\|\mathbf{\Delta}_{\mathbf{G},z}^{\text{true}}\|_F^2\right] + 2\sum_{z=1}^{B} \mathbb{E}\left[\|\mathbf{\Delta}_{\mathbf{E},z}\|_F^2\right] \\ &\leq 4B^2L^2 \sum_{z=2}^{N} \mathbb{E}\left[\|\mathbf{X}_z - \mathbf{X}_{z-1}\|_F^2\right] + 32B^2L \sum_{i=1}^{n} V_{i,T} + 8nB^2LM \\ &\quad + 2\sum_{z=1}^{B} \mathbb{E}\left[\|\mathbf{\Delta}_{\mathbf{E},z}\|_F^2\right],\end{aligned} \tag{52}$$

Next, from

$$\mathbf{X}_{z+1} - \mathbf{X}_z = \mathbf{U}_z^B - \mathbf{X}_z - \gamma \boldsymbol{\Delta}_z^{\mathbf{y}} - \gamma \mathbf{J}\mathbf{Y}_z,$$

the average-preserving and contraction properties of accelerated gossip imply

$$\|\mathbf{U}_z^B - \mathbf{X}_z\|_F \leq (1+\lambda_B)\|\boldsymbol{\Delta}_z^{\mathbf{x}}\|_F.$$

Moreover, recalling $\overline{\mathbf{y}}_z = \overline{\mathbf{g}}_{z-1}^{\text{true}} + \overline{\mathbf{e}}_{z-1}$ and (35), also using the conditional zero-mean property of $\overline{\mathbf{e}}_{z-1}$, we obtain

$$\begin{aligned}
\sum_{z=1}^{B} \mathbb{E}\left[\|\mathbf{X}_{z+1} - \mathbf{X}_z\|_F^2\right] &\leq 12 \sum_{z=1}^{N} \mathbb{E}\left[\|\boldsymbol{\Delta}_z^{\mathbf{x}}\|_F^2\right] + 3\gamma^2 \sum_{z=1}^{B} \mathbb{E}\left[\|\boldsymbol{\Delta}_z^{\mathbf{y}}\|_F^2\right] + 3n\gamma^2 \sum_{z=1}^{N} \mathbb{E}\|\overline{\mathbf{g}}_z^{\text{true}} + \overline{\mathbf{e}}_z\|^2 \\
&\leq 13 \sum_{z=1}^{B} \mathbb{E}\left[\|\boldsymbol{\Delta}_z^{\mathbf{x}}\|_F^2\right] + 3\gamma^2 \sum_{z=1}^{B} \mathbb{E}\left[\|\boldsymbol{\Delta}_z^{\mathbf{y}}\|_F^2\right] + 6nB\gamma^2 \mathbb{E}[\overline{\mathcal{R}}(T)] \\
&\quad + 3n\gamma^2 \sum_{z=1}^{B} \mathbb{E}\left[\|\overline{\mathbf{e}}_z\|^2\right], \qquad (53)
\end{aligned}$$

where we used $\lambda_B \leq 1$ and $\gamma \leq 1/(3BL)$ for rescaling. Combining (51), (52), and (53), we obtain

$$\begin{aligned}
\sum_{z=1}^{B} \mathbb{E}\left[\|\boldsymbol{\Delta}_z^{\mathbf{y}}\|_F^2\right] &\leq \frac{52B^2L^2\lambda_B^2}{(1-\lambda_B)^2} \sum_{z=1}^{B} \mathbb{E}\left[\|\boldsymbol{\Delta}_z^{\mathbf{x}}\|_F^2\right] + \frac{12B^2L^2\gamma^2\lambda_B^2}{(1-\lambda_B)^2} \sum_{z=1}^{B} \mathbb{E}\left[\|\boldsymbol{\Delta}_z^{\mathbf{y}}\|_F^2\right] \\
&\quad + \frac{24nB^3L^2\gamma^2\lambda_B^2}{(1-\lambda_B)^2} \mathbb{E}[\overline{\mathcal{R}}(T)] + \frac{12nB^2L^2\gamma^2\lambda_B^2}{(1-\lambda_B)^2} \sum_{z=1}^{B} \mathbb{E}\left[\|\overline{\mathbf{e}}_z\|^2\right] \\
&\quad + \frac{32B^2L\lambda_B^2}{(1-\lambda_B)^2} \sum_{i=1}^{n} V_{i,T} + \frac{8nB^2LM\lambda_B^2}{(1-\lambda_B)^2} + \frac{2\lambda_B^2}{(1-\lambda_B)^2} \sum_{z=1}^{B} \mathbb{E}\left[\|\boldsymbol{\Delta}_{\mathbf{E},z}\|_F^2\right].
\end{aligned}$$

Define

$$C_1 := 1 - \frac{12B^2L^2\gamma^2\lambda_B^2}{(1-\lambda_B)^2} > 0$$

for some chosen $\gamma < \frac{1-\lambda_B}{2\sqrt{3}BL}$. Rearranging the terms gives

$$\begin{aligned}
\sum_{z=1}^{B} \mathbb{E}\left[\|\boldsymbol{\Delta}_z^{\mathbf{y}}\|_F^2\right] &\leq \frac{52B^2L^2\lambda_B^2}{C_1(1-\lambda_B)^2} \sum_{z=1}^{B} \mathbb{E}\left[\|\boldsymbol{\Delta}_z^{\mathbf{x}}\|_F^2\right] + \frac{24nB^3L^2\gamma^2\lambda_B^2}{C_1(1-\lambda_B)^2} \mathbb{E}[\overline{\mathcal{R}}(T)] \\
&\quad + \frac{32B^2L\lambda_B^2}{C_1(1-\lambda_B)^2} \sum_{i=1}^{n} V_{i,T} + \frac{8nB^2LM\lambda_B^2}{C_1(1-\lambda_B)^2} \\
&\quad + \frac{12nB^2L^2\gamma^2\lambda_B^2}{C_1(1-\lambda_B)^2} \sum_{z=1}^{B} \mathbb{E}\left[\|\overline{\mathbf{e}}_z\|^2\right] + \frac{2\lambda_B^2}{C_1(1-\lambda_B)^2} \sum_{z=1}^{B} \mathbb{E}\left[\|\boldsymbol{\Delta}_{\mathbf{E},z}\|_F^2\right]. \qquad (54)
\end{aligned}$$

Finally substituting (54) into (50) yields

$$\begin{aligned}
\sum_{z=1}^{B} \mathbb{E}\left[\|\boldsymbol{\Delta}_z^{\mathbf{x}}\|_F^2\right] &\leq \frac{52B^2L^2\lambda_B^2\gamma^2}{C_1(1-\lambda_B)^4} \sum_{z=1}^{B} \mathbb{E}\left[\|\boldsymbol{\Delta}_z^{\mathbf{x}}\|_F^2\right] + \frac{24nB^3L^2\gamma^4\lambda_B^2}{C_1(1-\lambda_B)^4} \mathbb{E}[\overline{\mathcal{R}}(T)] \\
&\quad + \frac{32B^2L\lambda_B^2\gamma^2}{C_1(1-\lambda_B)^4} \sum_{i=1}^{n} V_{i,T} + \frac{8nB^2LM\lambda_B^2\gamma^2}{C_1(1-\lambda_B)^4} \\
&\quad + \frac{12nB^2L^2\gamma^4\lambda_B^2}{C_1(1-\lambda_B)^4} \sum_{z=1}^{B} \mathbb{E}\left[\|\overline{\mathbf{e}}_z\|^2\right] + \frac{2\gamma^2\lambda_B^2}{C_1(1-\lambda_B)^4} \sum_{z=1}^{B} \mathbb{E}\left[\|\boldsymbol{\Delta}_{\mathbf{E},z}\|_F^2\right] \\
&\leq \frac{24nB^3L^2\gamma^4\lambda_B^2}{C_1C_2(1-\lambda_B)^4} \mathbb{E}[\overline{\mathcal{R}}(T)] + \frac{12nB^2L^2\gamma^4\lambda_B^2}{C_1C_2(1-\lambda_B)^4} \sum_{z=1}^{B} \mathbb{E}\left[\|\overline{\mathbf{e}}_z\|^2\right]
\end{aligned}$$

$$+\frac{32B^2L\lambda_B^2\gamma^2}{C_1C_2(1-\lambda_B)^4}\sum_{i=1}^{n}V_{i,T}+\frac{8nB^2LM\lambda_B^2\gamma^2}{C_1C_2(1-\lambda_B)^4}$$
$$+\frac{2\gamma^2\lambda_B^2}{C_1C_2(1-\lambda_B)^4}\sum_{z=1}^{B}\mathbb{E}\left[\|\mathbf{\Delta}_{\mathbf{E},z}\|_F^2\right],$$

where in the last inequality, we define

$$C_2 := 1-\frac{52B^2L^2\lambda_B^2\gamma^2}{C_1(1-\lambda_B)^4}>0,$$

for some chosen $\gamma < \frac{1}{2BL}\sqrt{\frac{C_1}{13}}$. Thus we finish the proof.

□

Theorem B.7 (Restatement of Theorem 4.2)**.** *Under Assumptions 2.1-2.4, let* $B=\Theta(\rho^{-1/2}\log n)$, $\gamma \le \min\left\{\frac{1-\lambda_B}{2BL}\sqrt[4]{\frac{C_1C_2}{90}}, \sqrt{\frac{n(1+V_T)}{\sigma_g^2T}}\right\}$*, for each node* $i\in[n]$*, Algorithm 1 can gaurantee*

$$\mathbb{E}\left[\mathcal{R}_i(T)\right] \le O\left(\max\left\{\rho^{-1/2}(1+V_T),\, \sigma_g\sqrt{\frac{(1+V_T)T}{n}}\right\}\right).$$

Proof. We first establish an upper bound on the averaged-node regret of the proposed algorithm. Combining Lemma B.5 and Lemma B.6 yields

$$\begin{aligned}\mathbb{E}\left[\overline{\mathcal{R}}(T)\right] \le{}& \frac{24M+12V_T}{\gamma}+\frac{1440B^4L^4\gamma^4\lambda_B^2}{C_1C_2(1-\lambda_B)^4}\mathbb{E}\left[\overline{\mathcal{R}}(T)\right]+\frac{720B^3L^4\gamma^4\lambda_B^2}{C_1C_2(1-\lambda_B)^4}\sum_{z=1}^{N}\mathbb{E}\left[\|\overline{\mathbf{e}}_z\|^2\right]\\
&+968BLV_T+\frac{1920B^3L^3\gamma^2\lambda_B^2}{C_1C_2(1-\lambda_B)^2}V_T+\frac{120BL^2\gamma^2\lambda_B^2}{nC_1C_2(1-\lambda_B)^4}\sum_{z=1}^{N}\mathbb{E}\left[\|\mathbf{\Delta}_{\mathbf{E},z}\|_F^2\right]\\
&+126L^2\gamma\sum_{z=1}^{N}\mathbb{E}\left[\|\overline{\mathbf{e}}_z\|^2\right]+O(1). \end{aligned} \tag{55}$$

We then handle the two noise terms separately. For the noise vector $\|\overline{\mathbf{e}}_z\|$, we have

$$\sum_{z=1}^{N}\mathbb{E}\left[\|\overline{\mathbf{e}}_z\|^2\right]=\frac{1}{n^2}\sum_{z=1}^{N}\mathbb{E}\left[\left\|\sum_{i=1}^{n}\mathbf{e}_{i,z}\right\|^2\right]\le\frac{1}{n^2}\sum_{z=1}^{N}\sum_{i=1}^{n}\mathbb{E}\left[\|\mathbf{e}_{i,z}\|^2\right]\le\frac{LB\sigma_g^2}{n}=\frac{T\sigma_g^2}{n}. \tag{56}$$

Notably, this noise term preserves the linear speedup with respect to the number of agents. For the noise-difference matrix $\|\mathbf{\Delta}_{\mathbf{E},z}\|_F$, we have

$$\sum_{z=1}^{B}\mathbb{E}\left[\|\mathbf{\Delta}_{\mathbf{E},z}\|_F^2\right]\le 2\sum_{z=1}^{N}\mathbb{E}\left[\|\mathbf{E}_z\|_F^2\right]\le 2\sum_{z=1}^{N}\sum_{i=1}^{n}\mathbb{E}\left[\|\mathbf{e}_{i,z}\|^2\right]\le 2nT\sigma_g^2. \tag{57}$$

Substituting (56) and (57) into (55) yields

$$\begin{aligned}\mathbb{E}\left[\overline{\mathcal{R}}(T)\right] \le{}& \frac{24M+12V_T}{\gamma C_3}+\frac{720B^3L^4\gamma^4\lambda_B^2\sigma_g^2T}{nC_1C_2C_3(1-\lambda_B)^4}+\frac{1920B^3L^3\gamma^2\lambda_B^2}{C_1C_2C_3(1-\lambda_B)^2}V_T\\
&+\frac{240BL^2\gamma^2\lambda_B^2\sigma_g^2T}{C_1C_2C_3(1-\lambda_B)^4}+\frac{126L^2\gamma\sigma_g^2T}{nC_3}+\frac{968BL}{C_3}V_T+O(1)\\
={}& O\left(\frac{1+V_T}{\gamma}+\frac{B^3\gamma^4\sigma_g^2T}{n^2}+\frac{B^3\gamma^2}{n}V_T+\frac{B\gamma^2\sigma_g^2T}{n}+\frac{\gamma\sigma_g^2T}{n}+BV_T\right). \end{aligned} \tag{58}$$

Recall that $B=\Theta(\rho^{-1/2}\log\sqrt{n})$ and $\lambda_B^2=\Theta(n^{-1})$, with

$$\gamma \le \min\left\{\frac{1-\lambda_B}{2BL}\sqrt[4]{\frac{C_1C_2}{90}}, \sqrt{\frac{n(1+V_T)}{\sigma_g^2T}}\right\}, \tag{59}$$

we can gaurantee that

$$\mathbb{E}\left[\overline{\mathcal{R}}(T)\right] \leq O\left(\max\left\{\rho^{-1/2}(1+V_T),\, \sigma_g\sqrt{\frac{(1+V_T)T}{n}}\right\}\right). \tag{60}$$

Finally, we transfer the regret bound of the virtual averaged node to each individual agent. As in the deterministic case, the only additional term comes from the disagreement between the local decision $\mathbf{x}_{i,z}$ and the averaged decision $\overline{\mathbf{x}}_z$. For any block z and any $t \in \tau_z$, by adding and subtracting $\nabla f_t(\overline{\mathbf{x}}_z)$, we have

$$\begin{aligned}\|\nabla f_t(\mathbf{x}_{i,z})\|^2 &= \|\nabla f_t(\mathbf{x}_{i,z}) - \nabla f_t(\overline{\mathbf{x}}_z) + \nabla f_t(\overline{\mathbf{x}}_z)\|^2 \\ &\leq 2\|\nabla f_t(\mathbf{x}_{i,z}) - \nabla f_t(\overline{\mathbf{x}}_z)\|^2 + 2\|\nabla f_t(\overline{\mathbf{x}}_z)\|^2 \\ &\leq 2L^2\|\mathbf{x}_{i,z} - \overline{\mathbf{x}}_z\|^2 + 2\|\nabla f_t(\overline{\mathbf{x}}_z)\|^2. \end{aligned} \tag{61}$$

Summing (61) over $t \in \tau_z$ and then over all blocks, and taking expectation, gives

$$\begin{aligned}\mathbb{E}\left[\mathcal{R}_i(T)\right] &\leq 2\mathbb{E}\left[\overline{\mathcal{R}}(T)\right] + 2BL^2\sum_{z=1}^{B}\mathbb{E}\left[\|\mathbf{x}_{i,z} - \overline{\mathbf{x}}_z\|^2\right] \\ &\leq 2\mathbb{E}\left[\overline{\mathcal{R}}(T)\right] + 2BL^2\sum_{z=1}^{B}\mathbb{E}\left[\|\mathbf{\Delta}_z^{\mathbf{x}}\|_F^2\right]. \end{aligned} \tag{62}$$

Substituting Lemma B.6 and (58) into (62), With $B = \Theta(\rho^{-1/2}\log n)$, $\lambda_B^2 = \Theta(1/n^2)$, we have

$$\begin{aligned}\mathbb{E}\left[\mathcal{R}_i(T)\right] &\leq O\left(\left(1 + nL^4\gamma^4\lambda_B^2\right)\mathbb{E}\left[\overline{\mathcal{R}}(T)\right] + L^3\gamma^4\lambda_B^2\sigma_g^2 T + nL^3\gamma^2\lambda_B^2 V_T + nL\gamma^2\lambda_B^2\sigma_g^2 T\right) \\ &= O\left(\frac{1+V_T}{\gamma} + LV_T + \frac{\gamma\sigma_g^2 T}{n} + \frac{L\gamma^2\sigma_g^2 T}{n}\right) \leq O\left(\frac{1+V_T}{\gamma} + LV_T + \frac{\gamma\sigma_g^2 T}{n}\right).\end{aligned}$$

Under the same stepsize setting as shown in (59), we obtain the same upper bound in (60). Thus we finish the proof. □

C Lower bounds of D-ONO

In this section, we establish lower bounds for D-ONO over the decentralized algorithm class in Definition 2.5. In Appendix C.1, we derive the lower bound $\Omega\big(\rho^{-1/2}(1+V_T)\big)$ under deterministic gradient feedback. In Appendix C.2, we establish lower bound $\Omega\big(\sigma_g\sqrt{(1+V_T)T/n}\big)$ under stochastic gradient feedback.

C.1 Lower Bound with GO

This subsection establishes a deterministic lower bound for D-ONO. We first introduce a class of ring-lattice graphs in Appendix C.1.1, which will be used to construct hard graph instances for the lower bound proof in Appendix C.1.2. Our motivation for adopting this graph structure is consistent with that in Yuan et al. (2022): its greater flexibility allows us to preserve the required communication delay while extending the construction to as broad a range of ρ as possible.

C.1.1 Ring-Lattice Graph

A ring-lattice graph is a graph structure lying between a ring graph and a complete graph. A ring-lattice graph $G_{n,k}$ is determined by the number of nodes n and an even integer $k \in [2, n-1]$, where k denotes the number of neighboring nodes on the ring to which each node is connected. When $k = 2$, $G_{n,k}$ reduces to a standard ring graph, when $k = n-1$, $G_{n,k}$ becomes a complete graph.

The following two lemmas establish an important property of ring-lattice graphs.

Lemma C.1 (Lemma 1 of Yuan et al. (2022)). *The diameter of the ring-lattice graph $G_{n,k}$ is $D = \Theta(n/k)$.*

Lemma C.2 (Theorem 1 of Yuan et al. (2022)). *Given a fixed $n > 0$ and any $\rho \in [1-\cos(\pi/n), 1]$, we can always construct a ring-lattice graph $G_{n,k}$ with chosen $k = \Theta(n\rho^{1/2})$, so that its associated communication matrix $\mathbf{W}$ satisfies $\|\mathbf{W} - \frac{1}{n}\mathbf{1}_n\mathbf{1}_n^\top\| = 1-\rho$, and its diameter D satisfies $D = \Theta(\rho^{-1/2})$.*

Lemma C.2 will be used in the communication lower bound proof in the next subsection. It relates the communication delay D of the graph to the connectivity measure ρ over a certain range of ρ.

C.1.2 Proof of Theorem 5.1

Theorem C.3 (Restatement of Theorem 5.1). *Consider any algorithm A satisfying Definition 2.5. Under Assumptions 2.1, 2.2, 2.3 and suppose $d \geq \Omega(1+V_T)$. There exist a graph G with diameter $D \leq O(T/(1+V_T))$, $\rho \in [1-\cos(\pi/n), 1]$ and $\{f_t\}_{t=1}^T$ with $V_T = o(T)$ such that, for some node $i \in [n]$, $\mathcal{R}_i(T) \geq \Omega(\rho^{-1/2}(1+V_T))$. Moreover, if $D > \Omega(T/(1+V_T))$, then it yields the linear $\mathcal{R}_i(T) = \Theta(T)$.*

Proof. For any vector $\mathbf{v} \in \mathbb{R}^d$, we use $[\mathbf{v}]_i$ to denote its i-th coordinate. Our proof is presented as follows.

Step 1: Constructing ring-lattice graph $G_{n,k}$ and $f_{i,t}$ that satisfying Assumptions 2.1, 2.2, 2.3.

For any node $i \in [n]$, we fixedly devide $\{f_{i,t}\}_{t=1}^T$ into N blocks, with block length $B = \lceil T/N \rceil$. For $k \in [N]$, we set $f_{i,(k-1)B+1}(\mathbf{x}) = f_{i,(k-1)B+2}(\mathbf{x}) = \cdots = f_{i,kB}(\mathbf{x})$. Consider a single-coordinate construction function $\widehat{f}_k(\mathbf{x})$, defined as

$$\widehat{f}_k(\mathbf{x}) = c\sigma([\mathbf{x}]_k), \quad \text{where} \quad \sigma(z) = \frac{\exp(z)}{1+\exp(z)}.$$

Now, consider a ring-lattice graph $G_{n,k}$ with $\rho \in [1-\cos(\pi/n), 1]$ and communication matrix $\mathbf{W}$ satisfying Assumption 2.3. We assign each local function $f_{i,t}$ to be either the above-defined $\widehat{f}_k$ or the identically zero function, according to the following arrangement:

$$f_{i,t}(\mathbf{x}) = \begin{cases} \widehat{f}_k(\mathbf{x}) & i \in \mathcal{N}_{\text{good}} \\ 0 & \text{otherwise} \end{cases}, \quad \text{where} \quad \mathcal{N}_{\text{good}} = \left\{ \left\lfloor \frac{n}{4} \right\rfloor, \left\lfloor \frac{n}{4} \right\rfloor + 1, \ldots, \left\lceil \frac{3n}{4} \right\rceil \right\}.$$

Here $\mathcal{N}_{\text{good}}$ denotes the "good neighborhood", whose nodes have nonzero local functions and can therefore directly learn the k-th coordinate. The global objective function is then shown as

$$f_t(\mathbf{x}) = \frac{|\mathcal{N}_{\text{good}}|}{n} c\sigma([\mathbf{x}]_k) = \frac{c}{n}\left\lceil \frac{n}{2} \right\rceil \sigma([\mathbf{x}]_k).$$

Clearly, node 0 is located far from $\mathcal{N}_{\text{good}}$, and we therefore refer to it as the isolated node. It naturally holds that, for any $i \in [n]$ and $\mathbf{x}, \mathbf{y} \in \mathbb{R}^d$,

$$|f_{i,t}(\mathbf{x})| \leq c \leq M, \quad \|\nabla f_{i,t}(\mathbf{x}) - \nabla f_{i,t}(\mathbf{y})\| \leq \frac{\sqrt{3}c}{18}\|\mathbf{x}-\mathbf{y}\| \leq L_{f,1}\|\mathbf{x}-\mathbf{y}\|,$$

if $c = \min\{M, 6\sqrt{3}L_{f,1}\}$. Thus, our construction satisfies Assumption 2.1, 2.2. Consider the exact variation V_T^f we constructed, defined in (3), it holds that

$$\begin{aligned} V_T^f =& \sum_{t=2}^{T} \max_{i\in[n]} \sup_{\mathbf{x}\in\mathbb{R}^d} |f_{i,t}(\mathbf{x}) - f_{i,t-1}(\mathbf{x})| \\ =& c\sum_{k=2}^{N} \max_{i\in[a]} \sup_{\mathbf{x}\in\mathbb{R}^d} |\sigma([\mathbf{x}]_k) - \sigma([\mathbf{x}]_{k-1})| \leq c(N-1) \leq V_T, \end{aligned}$$

if set $N = \lceil \frac{1+V_T}{c} \rceil$, and thus $B = \left\lfloor \frac{cT}{1+V_T} \right\rfloor$.

Step 2: Characterizing the lower bound of $\mathcal{R}_0(T)$ on isolated node.

The key observation is that, for any algorithm A satisfying Definition 2.5, performing only one communication step per round implies that the isolated node 0 must wait for a number of rounds equal to the shortest-path distance between itself and the good neighbor $\mathcal{N}_{\text{good}}$, represented by a, before it can receive the information needed to activate the coordinate revealed by f_t. According to our construction, we have

$$a = \min\left\{\operatorname{dist}\left(0, \left\lfloor \frac{n}{4} \right\rfloor\right), \operatorname{dist}\left(0, \left\lceil \frac{3n}{4} \right\rceil\right)\right\} = \Theta(D).$$

This implies that, during the first a steps of each time block B, we have

$$\mathcal{H}_{0,(k-1)B+1} = \mathcal{H}_{0,(k-1)B+2} = \cdots = \mathcal{H}_{0,(k-1)B+a} = \{\mathbf{e}_1, \mathbf{e}_2, \ldots, \mathbf{e}_{k-1}\}.$$

Therefore, the regret suffered by node 0 admits the following lower bound:

$$\sum_{j=1}^{B} \left\|\nabla f_{(k-1)B+j}(\mathbf{x}_{0,(k-1)B+j})\right\|^2 \geq \sum_{j=1}^{\min\{a,B\}} \left\|\frac{c}{n}\left\lceil \frac{n}{2} \right\rceil \sigma'(0)\right\|^2 \geq \frac{c^2}{64}\min\{a, B\}.$$

Finally, under $a < B$, which is equivalent to $D < O(T/(1+V_T))$, summing over all blocks yields

$$\begin{aligned}\mathcal{R}_0(T) =& \sum_{k=1}^{N}\sum_{j=1}^{B} \left\|\nabla f_{(k-1)B+j}(\mathbf{x}_{0,(k-1)B+j})\right\|^2 \\ \geq& \frac{ac^2N}{64} \geq \Omega\left(D(1+V_T)\right) = \Omega\left(\rho^{-1/2}(1+V_T)\right),\end{aligned}$$

where the last equality follows from Lemma C.2 by choosing a $\mathbf{W}$ such that $D = \Theta(\rho^{-1/2})$. It is then clear that if $D \geq \Omega(T/(1+V_T))$, node 0 cannot receive any information about the revealed coordinates before the end of a block, thereby incurring linear regret:

$$\mathcal{R}_0(T) = \sum_{k=1}^{N}\sum_{j=1}^{B} \left\|\nabla f_{(k-1)B+j}(\mathbf{x}_{0,(k-1)B+j})\right\|^2 \geq \frac{c^2BN}{64} = \frac{c^2T}{64}.$$

Thus we finish the proof. □

C.2 Lower Bound with SGO

We first introduce in Appendix C.2.1 a zero-chain function construction that is commonly used in lower bound analyses for non-convex optimization Carmon et al. (2020); Arjevani et al. (2023). Based on this hard-to-learn function, we construct the corresponding online function sequence for the lower bound proof in Appendix C.2.2.

C.2.1 Hard-to-Learn Non-Convex Function

In Carmon et al. (2020), the following hard-to-learn non-convex function $\widehat{f}_K : \mathbb{R}^K \to \mathbb{R}$ is constructed:

$$\widehat{f}_K(\mathbf{v}) = -\Psi(1)\Phi([\mathbf{v}]_1) + \sum_{j=2}^{K}\left(\Psi(-[\mathbf{v}]_{j-1})\Phi(-[\mathbf{v}]_j) - \Psi([\mathbf{v}]_{j-1})\Phi([\mathbf{v}]_j)\right), \quad \text{where}$$

$$\Psi(z) = \begin{cases}\exp\left(1 - \frac{1}{(2z-1)^2}\right) & z > \frac{1}{2} \\ 0 & z \leq \frac{1}{2}\end{cases} \quad \text{and} \quad \Phi(z) = \sqrt{e}\int_{-\infty}^{z} \exp(-\frac{x^2}{2})\,\mathrm{d}x.$$

This construction ensures that, for gradient-based algorithms initialized at the origin, each gradient query can activate at most the first currently 0 coordinate, changing it from 0 to a nonzero value. Consequently, revealing the complete coordinate information necessarily requires at least K gradient queries. The following lemma summarizes several useful properties of this function.

Lemma C.4 (Lemma 2 of Arjevani et al. (2023)). *For any vector $\mathbf{v} \in \mathbb{R}^d$ and $K \in \mathbb{Z}$, $\widehat{f}_K$ satisfies:*

- *For any $\mathbf{v} \in \mathbb{R}^d$, $\widehat{f}_K(\mathbf{v}) - \inf_{\mathbf{v}} \widehat{f}_K(\mathbf{v}) \leq 12K$;*

- *The gradient of $\widehat{f}_K$ is 152-Lipschitz continuous;*
- *For any $\mathbf{v} \in \mathbb{R}^d$, $\|\nabla \widehat{f}_K(\mathbf{v})\|_\infty \leq 23$;*
- *For any $\mathbf{v} \in \mathbb{R}^d$, $\|\nabla \widehat{f}_K(\mathbf{v})\|_\infty \geq 1$ with $[\mathbf{v}]_K = 0$.*

C.2.2 Proof of Theorem 5.2

Theorem C.5 (Restatement of Theorem 5.2)**.** *Consider any algorithm A with stochastic gradient feedback $\mathbf{g}_{i,t}$ satisfying Definition 2.5. Under Assumptions 2.1-2.4 and suppose $d \geq \Omega(T)$. There exist a graph G with $\rho \in (0,1]$ and $\{f_t\}_{t=1}^T$ with $V_T = o(T)$ such that for some node $i \in [n]$, $\mathcal{R}_i(T) \geq \Omega(\sigma_g \sqrt{(1+V_T)T/n})$.*

Proof. Our proof is presented as follows.

Step 1: Constructing graph $\mathcal{G}$ and $f_{i,t}$ that satisfying Assumptions 2.1, 2.2, 2.3.

Consider a complete graph G with $\rho \in (0,1]$ and $\mathbf{W}$ satisfying Assumption 2.3. Similar to the function construction in Theorem 5.1, we fixedly devide $\{f_{i,t}\}_{t=1}^T$ into N blocks for all $i \in [n]$, with block length $B = \lceil T/N \rceil$. For $k \in [N]$, we have $f_{i,(k-1)B+1}(\mathbf{x}) = f_{i,(k-1)B+2}(\mathbf{x}) = \cdots = f_{i,kB}(\mathbf{x})$. Remind of the hard-to-learn function introduced in Section C.2.1, for some $c > 0$, $K \in \mathbb{Z}$ and $\lambda = 1/\sqrt{K}$, our local function is defined as follows:

$$f_t(\mathbf{x}) = f_{i,t}(\mathbf{x}) = c\lambda^2 \widehat{f}_K\left(\frac{[\mathbf{x}]_{(k-1)B+1:(k-1)B+K}}{\lambda}\right).$$

The property of $f_{i,t}$ can be given by Lemma C.4 as follows

$$|f_{i,t}(\mathbf{x})| \leq 12c, \quad \|\nabla f_{i,t}(\mathbf{x})\| \leq \frac{23c}{\sqrt{K}} \leq 23c, \quad \|\nabla f_{i,t}(\mathbf{x}) - \nabla f_{i,t}(\mathbf{y})\| \leq 152c\|\mathbf{x}-\mathbf{y}\|.$$

With $c = \min\{M/12, L/152\}$, it follows directly that $f_{i,t}$ satisfies Assumption 2.1, 2.2 for any $i \in [n]$. Consider the exact variation V_T^f we constructed, it holds that

$$V_T^f = \sum_{t=2}^T \max_{i\in[n]} \sup_{\mathbf{x}\in\mathbb{R}^d} |f_{i,t}(\mathbf{x}) - f_{i,t-1}(\mathbf{x})| \leq 2M(N-1) \leq V_T,$$

if set $N = \left\lfloor \frac{1+V_T}{2M} \right\rfloor$, and thus $B = \left\lceil \frac{2MT}{1+V_T} \right\rceil$.

Step 2: Characterizing the local regret lower bound for some node $i \in [n]$.

Here, we specifically define the stochastic gradient in Definition 2.5 as

$$[\mathbf{g}_{t,i}(\mathbf{x})]_j = [\nabla f_{t,i}(\mathbf{x})]_j \cdot \left(1 + \mathbb{1}_{j > \mathrm{proj}_{\frac{1}{4}}(\mathbf{x})}\left(\frac{z}{p} - 1\right)\right), \quad j = 1, \ldots, d,$$

where $z \sim \mathrm{Bernoulli}(p)$ and $p = \min\{1, \frac{(23c)^2}{K\sigma_g^2}\} \leq 1$. This stochastic gradient setting was introduced in Arjevani et al. (2023), which is unbiased and satisfies Assumption 2.4, shown as

$$\mathbb{E}\left[\|\mathbf{g}_{i,t}(\mathbf{x}) - \nabla f_{i,t}(\mathbf{x})\|^2\right] = \mathbb{E}\left[\left(\frac{z}{p} - 1\right)^2\right] \sum_{j > \mathrm{proj}_{\frac{1}{4}}(\mathbf{x})} |[\nabla f_{i,t}(\mathbf{x})]_j|^2 \leq \frac{1-p}{p}\|\nabla f_{i,t}(\mathbf{x})\|^2 \leq \sigma_g^2.$$

By the probabilistic zero-chain property of $\widehat{f}_K$, a stochastic gradient query can reveal at most one new coordinate beyond the coordinates that have already been explored, and this happens with probability at most p. For simplicity of the proof, we ignore the one round delay in transmitting information from the other nodes to a particular node i in the complete graph G. Consequently, we may regard node i as having n stochastic gradient queries, and hence n opportunities to reveal a new coordinate, at each round.

Consider block k and let γ_s denote the largest coordinate of the corresponding zero-chain that has been revealed to any node after the first s rounds of this block. The chain can advance at round s only if at least one of the n stochastic gradient queries reveals the next coordinate. Therefore,

$$\mathbb{P}(\gamma_s - \gamma_{s-1} = 1 \mid \mathcal{F}_{s-1}) \leq 1 - (1-p)^n \leq np. \tag{63}$$

Following the standard probabilistic zero-chain argument, for any $s \geq 1$,

$$\mathbb{P}(\gamma_s \geq K) \leq \exp\left(2\left(1-(1-p)^n\right)s - K\right). \tag{64}$$

Consequently, according to Lemma 1 in Arjevani et al. (2023), if

$$B \leq \frac{K - \log 2}{2np}, \tag{65}$$

then, with probability at least $1/2$, the K-th coordinate is not revealed throughout the whole block.

We next choose the parameters so that (65) holds. Recall that $\lambda = 1/\sqrt{K}$ and $p = \min\{1, \frac{(23c)^2}{K\sigma_g^2}\} \leq 1$. When $p < 1$, it is sufficient to require

$$B \leq \frac{K}{4np} = \frac{K^2\sigma_g^2}{4(23c)^2 n}. \tag{66}$$

Hence, choosing

$$K = \Theta\left(\frac{c\sqrt{nB}}{\sigma_g}\right) \tag{67}$$

with a sufficiently large numerical constant guarantees (65).

Now consider any node $i \in [n]$. On the event that the chain has not been completely revealed, $\mathbf{x}_{i,t}$ cannot have nonzero entries beyond the revealed coordinates. By Lemma C.4, we have

$$\|\nabla f_t(\mathbf{x}_{i,t})\| = c\lambda \left\|\nabla \widehat{f}_K\left(\frac{[\mathbf{x}_{i,t}]_{(k-1)B+1:(k-1)B+K}}{\lambda}\right)\right\| \geq \frac{c}{\sqrt{K}},$$

Since the event in (65) occurs with probability at least $1/2$, the expected regret incurred by node i in every block satisfies

$$\mathbb{E}\left[\sum_{t=(k-1)B+1}^{(k-1)B+K} \|\nabla f_t(\mathbf{x}_{i,t})\|^2\right] \geq \frac{c^2 B}{2K}. \tag{68}$$

Summing (68) over all N blocks with the choice of K in (67) and $V_T = \Theta(1 + V_T)$ yields

$$\mathbb{E}\left[\mathcal{R}_i(T)\right] \geq \frac{c^2 NB}{2K} = \Omega\left(\sigma_g\sqrt{\frac{NT}{n}}\right) = \Omega\left(\sigma_g\sqrt{\frac{(1+V_T)T}{n}}\right).$$

In fact, the above argument holds for every node $i \in [n]$, and hence in particular establishes the claimed lower bound for some $i \in [n]$. Thus we finish the proof.

□

C.3 Some Discussion of Lower Bound

It is worth noting that we do not establish the lower bound over the entire range $\rho \in (0, 1]$ in Theorem 5.1. This limitation reflects a fundamental mismatch between the quantities underlying the upper and lower bound analyses: lower bound construction exploits the communication delay induced by graph distance, whereas the upper bound is characterized by the spectral gap governing gossip mixing. The relation $D = \Theta(\rho^{-1/2})$ provides the key bridge between these two notions of difficulty. However, since $D \leq n-1$, such a relation can hold only when $\rho = \Omega(n^{-2})$. In particular, $1 - \cos(\pi/n) = \Theta(n^{-2})$. Therefore, subject to this intrinsic limitation, our construction establishes the lower bound over the largest range of ρ that can be covered by this distance-based argument.