GROUP DISTRIBUTIONALLY ROBUST OPTIMIZATION OVER DECENTRALIZED NETWORKS

Anonymous authors
Paper under double-blind review

ABSTRACT

We study group distributionally robust optimization (GDRO) over decentralized networks, where workers collaborate to learn robust models by minimizing the maximum local risk. Existing methods optimize only dual-regularized surrogates and establish guarantees for a globally averaged model that is not available to workers. In this paper, we directly optimize the original GDRO objective and establish guarantees for local models in semi-decentralized (SemiDec) and fully decentralized (Dec) settings. A key difficulty is that applying the primal-dual decomposition from centralized GDRO yields worker-dependent losses whose evaluation requires unavailable cross-worker information. To address this challenge, we develop a novel error decomposition that recasts decentralized GDRO as (i) a decentralized online convex optimization (D-OCO) problem and a prediction-with-expert-advice problem in SemiDec, and as (ii) two coupled D-OCO problems in Dec, with discrepancy terms capturing inconsistencies among local decisions. By isolating consensus errors in analysis, we develop modular frameworks that support different consensus routines. We then establish the *first* per-worker convergence guarantees for optimizing the original GDRO objective over decentralized networks. Instantiations with accelerated gossip and Choco-gossip achieve an $O(\sqrt{(\log m)/T})$ convergence rate over T update rounds, while experiments demonstrate improved worst-worker performance over baselines.

1 INTRODUCTION

Decentralized optimization is a powerful framework for large-scale model training and has attracted considerable attention due to its network robustness, communication efficiency, and scalability (He et al., 2018; Martínez Beltrán et al., 2023). Within this framework, m workers exchange information over a communication network to collaboratively optimize a shared learning objective. Classical methods train models by minimizing the average risk (Wei & Ozdaglar, 2012; Duchi et al., 2012; Yuan et al., 2016). However, in practical deployments, workers' local data distributions are often highly heterogeneous (Ye et al., 2023b), which can lead to performance degradation across workers (Hsieh et al., 2020). As a result, minimizing only the average risk may be unfair to certain workers and lacks robustness under distribution shift (Mohri et al., 2019).

A promising solution is group distributionally robust optimization (GDRO) (Sagawa et al., 2020), which minimizes the maximum risk across workers. For each worker $i \in [m]$, let $\mathcal{P}_i$ denote its local data distribution and define the local risk function as $R_i(\mathbf{w}) = \mathbb{E}_{\mathbf{z}\sim\mathcal{P}_i}[\ell(\mathbf{w};\mathbf{z})]$, where $\mathbf{z}$ denotes a random sample and $\ell(\cdot;\cdot)$ is the loss function. GDRO solves the stochastic minimax problem

$$\min_{\mathbf{w}\in\mathcal{W}} \max_{\mathbf{q}\in\Delta_m} \left\{ \phi(\mathbf{w},\mathbf{q}) = \sum_{i=1}^{m} q_i R_i(\mathbf{w}) \right\}, \tag{1}$$

where $\Delta_m = \{\mathbf{q} \in \mathbb{R}_+^m \mid \mathbf{1}^\top \mathbf{q} = 1\}$ denotes the simplex. We focus on the convex setting and refer to $\mathbf{w}$ as the primal model and $\mathbf{q}$ as the dual vector. Although GDRO has been extensively studied in centralized optimization (Haghtalab et al., 2022; Zhang et al., 2023) and in server-based federated learning (Mohri et al., 2019; Deng et al., 2020), optimizing (1) over *decentralized* networks remains challenging. Existing decentralized GDRO methods only optimize a *dual-regularized* variant of (1) with strongly concave regularization for the dual vector (Issaid et al., 2022; Zecchin et al., 2023).

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.

Consequently, their guarantees do not imply optimality for the worst-worker objective. Moreover, these methods provide guarantees only for a globally averaged model, which requires additional communication to approximate. Since deployment relies on locally maintained models, per-worker guarantees with respect to the global objective are more practically relevant (Duchi et al., 2012).

In this paper, we directly optimize problem (1) over two representative decentralized architectures: (i) a *semi-decentralized* (SemiDec) setting with a lightweight server and worker-to-worker communication, and (ii) a *fully decentralized* (Dec) setting with only worker-to-worker communication. Unlike decentralized optimization with a single variable (Nedic & Ozdaglar, 2009), GDRO couples the primal model and the dual vector, making its decentralized dynamics difficult to analyze. A natural attempt is to adapt the centralized GDRO decomposition (Zhang et al., 2023; Soma et al., 2025), which bounds the optimization error by the sum of primal and dual regret terms. However, applying this decomposition to each worker's local model yields worker-dependent losses whose evaluation requires cross-worker information that is not locally accessible. To overcome this challenge, we develop a novel error decomposition that rewrites each worker's local objectives as shared global objectives plus discrepancy terms arising from inconsistent local decisions. This reformulation turns decentralized GDRO into tractable online learning subproblems. In SemiDec, where a lightweight server maintains a shared dual vector, it yields a decentralized online convex optimization (D-OCO) (Yan et al., 2013) problem for the primal updates and a prediction with expert advice (PEA) (Freund & Schapire, 1997) problem for the dual updates, with model-discrepancy terms among workers. In Dec, where each worker maintains both a local model and a local dual vector, it yields two coupled D-OCO problems together with primal and dual discrepancy terms.

Building on this decomposition, we further isolate communication effects through consensus-error terms in the regret and discrepancy bounds. This separation decouples the optimization analysis from the specific consensus mechanism, yielding two modular frameworks that support different consensus routines. Algorithmically, both employ follow-the-regularized-leader (FTRL) (Orabona, 2019) for the primal and dual updates, while workers track the cumulative global gradient of each decentralized variable via the selected consensus routine. We then establish convergence guarantees for each worker's time-averaged local model under the original GDRO objective. By instantiating the frameworks with accelerated gossip (AG) (Liu & Morse, 2011) and Choco-gossip (CG) (Koloskova et al., 2019), we obtain concrete algorithms with an $O(\sqrt{(\log m)/T})$ convergence rate over T update rounds. Our main contributions are summarized as follows:

- We propose a novel error decomposition that reduces decentralized GDRO to a D-OCO problem and a PEA problem in SemiDec, and to two coupled D-OCO problems in Dec, with discrepancy terms capturing inconsistencies among local decisions.
- Building on the decomposition, we develop modular frameworks for GDRO in the SemiDec and Dec settings that accommodate different consensus routines, and establish per-worker convergence guarantees in terms of the resulting consensus errors.
- By instantiating the consensus routine with AG and CG, we obtain concrete algorithms with a per-worker convergence rate of $O(\sqrt{(\log m)/T})$, and empirically show improved worst-worker performance over prior methods.

2 Related Work

Decentralized minimax optimization. Decentralized learning has been studied in SemiDec (Guo et al., 2022; Wang & Chi, 2025) and Dec architectures (Olfati-Saber et al., 2007; Chen & Sayed, 2012), with applications to energy systems (Kekatos & Giannakis, 2013; Navidi et al., 2021), wireless sensor networks (Avci et al., 2018), and autonomous vehicles (Clancy et al., 2024). Meanwhile, many machine learning problems can be formulated as minimax optimization (Pinto et al., 2017; Goodfellow et al., 2020), motivating its study over decentralized networks. Most existing methods focus on regimes different from the convex-concave setting, including nonconvex-nonconcave (Liu et al., 2020), nonconvex-strongly-concave (Tsaknakis et al., 2020; Xian et al., 2021; You et al., 2025), and strongly-convex-strongly-concave objectives (Metelev et al., 2023). Although Koppel et al. (2015) study convex–concave problems, their dual variables are Lagrange multipliers for consensus constraints rather than distributional weights.

Group distributionally robust optimization. GDRO (Sagawa et al., 2020) minimizes the worst-group risk to learn models that perform uniformly well across multiple data distributions. It has

been studied in centralized settings (Sagawa et al., 2020; Haghtalab et al., 2022; Zhang et al., 2023; Bai et al., 2025; Zhang et al., 2026), server-based federated settings (Mohri et al., 2019; Deng et al., 2020; Guo & Yang, 2024), and decentralized settings (Issaid et al., 2022; Zecchin et al., 2023) with a strongly concave regularizer. Specifically, DR-DSGD (Issaid et al., 2022) adds Kullback–Leibler regularization to the dual vector, which transforms the regularized minimax problem into a single-level minimization problem, thus simplifying the algorithm design and analysis. AD-GDA (Zecchin et al., 2023) further considers general strongly concave regularizers for the dual vector and solves the resulting problem using decentralized gradient descent-ascent with compressed consensus. Despite these advances, the original GDRO objective remains unexplored in decentralized settings.

Decentralized online convex optimization. D-OCO (Yan et al., 2013) considers multiple workers that collaboratively optimize a global objective in an online manner over a decentralized network, and algorithms typically include two key steps: decision updates and consensus. Existing update methods include decentralized online gradient descent (Yan et al., 2013), decentralized follow-the-regularized-leader (FTRL) (Hosseini et al., 2013), and their variants (Zhang et al., 2017; Wan et al., 2022; Wang et al., 2023). For information aggregation, most approaches adopt gossip for neighborhood averaging (Xiao & Boyd, 2004). Wan et al. (2025) introduce online accelerated gossip by combining AG (Liu & Morse, 2011) with a blocking update mechanism, achieving nearly optimal regret bounds. To improve communication efficiency, Tu et al. (2022) combine D-OGD with the difference-compressed consensus mechanism (Koloskova et al., 2019) and Yang et al. (2026) improve the results via a two-level blocking update framework. Subsequently, Zhou et al. (2026) develop FTRL-type algorithms with compressed communication, matching the best known full-information regret bounds and improving both regret and communication costs in the bandit setting.

3 PRELIMINARIES

We equip $\mathcal{W}$ with a closed, 1-strongly convex function $\nu_w : \mathcal{W} \to \mathbb{R}$ with respect to $\|\cdot\|_2$, and Δ_m with negative entropy $\nu_q(\mathbf{q}) = \sum_{i=1}^m q_i \ln q_i$. For any approximate solution $\mathbf{w} \in \mathcal{W}$ of problem (1), its optimization error is defined as $\varepsilon_\phi(\mathbf{w}) = \max_{i\in[m]} R_i(\mathbf{w}) - \min_{\mathbf{w}\in\mathcal{W}} \max_{i\in[m]} R_i(\mathbf{w})$. We adopt standard assumptions as follows (Mohri et al., 2019; Zhang et al., 2023; Zecchin et al., 2023).

Assumption 1. $\mathcal{W}$ is nonempty, compact, convex with $\max_{\mathbf{w}\in\mathcal{W}} \nu_w(\mathbf{w}) - \min_{\mathbf{w}\in\mathcal{W}} \nu_w(\mathbf{w}) \le D^2$.

Assumption 2. For each $i \in [m]$, the risk $R_i(\cdot)$ is convex on $\mathcal{W}$, and for all $\mathbf{w} \in \mathcal{W}$, the stochastic gradient is unbiased, i.e., $\mathbb{E}_{\mathbf{z}\sim\mathcal{P}_i}[\nabla\ell(\mathbf{w};\mathbf{z})] = \nabla R_i(\mathbf{w})$. In addition, for all $\mathbf{w} \in \mathcal{W}$ and all $\mathbf{z}$,

$$0 \le \ell(\mathbf{w};\mathbf{z}) \le 1, \qquad \|\nabla\ell(\mathbf{w};\mathbf{z})\|_2 \le G. \tag{2}$$

Communication network. The worker-to-worker communication network is represented by an undirected graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$, where $\mathcal{V} = [m]$ is the set of workers and $\mathcal{E} \subseteq \mathcal{V} \times \mathcal{V}$ denotes the set of communication links. An edge $(i, j) \in \mathcal{E}$ indicates that workers i and j can exchange information, and we assume $(i, i) \in \mathcal{E}$ for all $i \in \mathcal{V}$. Let $\mathcal{N}_i = \{j \in \mathcal{V} : (i, j) \in \mathcal{E}\}$ denote the set of neighbors of worker i. To model information mixing over $\mathcal{G}$, we associate it with a matrix $P \in \mathbb{R}^{m\times m}$, where $P_{i,j} > 0$ if $(i, j) \in \mathcal{E}$ and $P_{i,j} = 0$ otherwise. Following standard decentralized optimization (Nedic & Ozdaglar, 2009; Duchi et al., 2012), we impose the following assumption on P.

Assumption 3. The communication matrix P satisfies $P \cdot \mathbf{1}_m = \mathbf{1}_m$, $P = P^\top$. Moreover, the graph $\mathcal{G}$ is connected, i.e., there exists a path between any two workers.

Under Assumption 3, the largest eigenvalue of P is $\lambda_1(P) = 1$, while the remaining eigenvalues satisfy $\lambda_\star(P) := \max_{i\ge2} |\lambda_i(P)| < 1$ (Horn & Johnson, 2012). We define the spectral gap as $\rho := 1 - \lambda_\star(P) \in (0, 1]$ and let $\beta := \|I_m - P\|_2 \in [0, 2]$.

Consensus routine. We model the information-exchange mechanism as a black-box consensus routine $\mathcal{A}_c$. It returns updated states $\{\mathbf{x}_{L+1}^{(i)}, \hat{\mathbf{x}}_{L+1}^{(i)}\}_{i\in[m]} = \mathcal{A}_c(\{\mathbf{x}_1^{(i)}, \hat{\mathbf{x}}_1^{(i)}\}_{i\in[m]})$ after L communication rounds, where $\mathbf{x}_1^{(i)}$ is the initial state of worker i, and $\hat{\mathbf{x}}_t^{(i)}$ denotes auxiliary states. We later instantiate $\mathcal{A}_c$ with two representative procedures, AG (Liu & Morse, 2011) and CG (Koloskova et al., 2019), which respectively provide fast mixing and compressed communication. The full procedures and corresponding consensus-error bounds are provided in Appendix A.

4 SEMI-DECENTRALIZED GDRO

We first consider SemiDec with a high-bandwidth worker-to-worker network and a low-bandwidth server–worker network. Such architectures arise naturally in wireless systems, where nearby devices can exchange high-dimensional information through fast D2D links, while a base station provides global coordination over bandwidth-limited device-to-server links (Lin et al., 2021; Guo et al., 2022). In SemiDec, we let the server only maintain the shared m-dimensional dual vector, while high-dimensional model update information is communicated over the worker-to-worker network.

4.1 ERROR DECOMPOSITION

Unlike prior work that provides guarantees only for a globally averaged model (Issaid et al., 2022; Zecchin et al., 2023), we establish guarantees for each worker's *local* model. This is more relevant in decentralized settings, since obtaining the global average requires additional communication and is generally only approximate under finite communication. Let $\mathbf{w}_t^{(i)}$ denote worker i's model at round t, and define its output after T iterations as $\bar{\mathbf{w}}_T^{(i)} = \frac{1}{T}\sum_{t=1}^{T} \mathbf{w}_t^{(i)}$. To decouple the effects of primal and dual updates, our starting point is the error decomposition used in centralized GDRO (Zhang et al., 2023; Soma et al., 2025). In particular, the optimization error of each $\bar{\mathbf{w}}_T^{(i)}$ can be bounded as

$$\begin{aligned}&\varepsilon_\phi(\bar{\mathbf{w}}_T^{(i)})\\ \le&\frac{1}{T}\underbrace{\left(\max_{\mathbf{q}\in\Delta_m}\sum_{t=1}^{T}\phi(\mathbf{w}_t^{(i)},\mathbf{q})-\sum_{t=1}^{T}\phi(\mathbf{w}_t^{(i)},\mathbf{q}_t)\right)}_{\mathcal{R}_q^{(i)}}+\frac{1}{T}\underbrace{\left(\sum_{t=1}^{T}\phi(\mathbf{w}_t^{(i)},\mathbf{q}_t)-\sum_{t=1}^{T}\phi(\mathbf{w}^*,\mathbf{q}_t)\right)}_{\mathcal{R}_w^{(i)}},\end{aligned} \tag{3}$$

where $\mathbf{w}^* \in \arg\min_{\mathbf{w}\in\mathcal{W}} \max_{i\in[m]} R_i(\mathbf{w})$ is a global solution. The term $\mathcal{R}_q^{(i)}$ is the regret of $\{\mathbf{q}_t\}_{t=1}^T$ against the loss sequence $\{\phi(\mathbf{w}_t^{(i)},\cdot)\}_{t=1}^T$ induced by worker i. Similarly, $\mathcal{R}_w^{(i)}$ is the regret of $\{\mathbf{w}_t^{(i)}\}_{t=1}^T$ measured on a shared objective sequence $\{\phi(\cdot,\mathbf{q}_t)\}_{t=1}^T$, with comparator $\mathbf{w}^*$.

Analysis of $\mathcal{R}_q^{(i)}$. Since each worker $i \in [m]$ induces its own regret term $\mathcal{R}_q^{(i)}$, the server needs to select a *single* dual sequence $\{\mathbf{q}_t\}_{t=1}^T$ that controls $\mathcal{R}_q^{(i)}$ *uniformly* for all $i \in [m]$. Constructing such a sequence directly, however, would require access to $R_j(\mathbf{w}_t^{(i)})$ for every $i, j \in [m]$. For $j \neq i$, this risk can only be evaluated at worker j, who does not have access to worker i's model $\mathbf{w}_t^{(i)}$. To overcome this information barrier, we decompose each worker-dependent objective into a shared global term and a discrepancy term. Specifically, we rewrite

$$\phi(\mathbf{w}_t^{(i)},\mathbf{q}) = \sum_{j=1}^{m} q_j R_j(\mathbf{w}_t^{(j)}) + \sum_{j=1}^{m} q_j\big(R_j(\mathbf{w}_t^{(i)}) - R_j(\mathbf{w}_t^{(j)})\big). \tag{4}$$

The first term is independent of i and therefore can serve as the global objective for updating the shared dual vector. Under Assumption 2, the second term can be bounded by $G\max_{j\in[m]}\|\mathbf{w}_t^{(i)} - \mathbf{w}_t^{(j)}\|_2$, which quantifies the maximum model disagreement between worker i and the rest of the network at round t. Consequently, each $\mathcal{R}_q^{(i)}$ can be bounded by the shared dual regret

$$\max_{\mathbf{q}\in\Delta_m}\sum_{t=1}^{T}\sum_{j=1}^{m} q_j R_j(\mathbf{w}_t^{(j)}) - \sum_{t=1}^{T}\sum_{j=1}^{m} q_{t,j} R_j(\mathbf{w}_t^{(j)}), \tag{5}$$

plus a model-discrepancy term. The shared regret in (5) leads to a PEA formulation at the server, where workers act as experts and the server updates $\mathbf{q}_t$ using stochastic estimates of their local risks.

Analysis of $\mathcal{R}_w^{(i)}$. In contrast to $\mathcal{R}_q^{(i)}$, which requires further decomposition, each primal regret term $\mathcal{R}_w^{(i)}$ in (3) is already evaluated on the same objective sequence $\{\phi(\cdot,\mathbf{q}_t)\}_{t=1}^T$, i.e.,

$$\mathcal{R}_w^{(i)} = \sum_{t=1}^{T}\left(\sum_{j=1}^{m} q_{t,j} R_j(\mathbf{w}_t^{(i)}) - \sum_{j=1}^{m} q_{t,j} R_j(\mathbf{w}^*)\right). \tag{6}$$

Algorithm 1 GDRO-SEMIDEC

Initialize $\mathbf{w}_1^{(i)} = \arg\min_{\mathbf{w}\in\mathcal{W}} \nu_w(\mathbf{w})$, $\mathbf{L}_0 = \mathbf{0}_m$, $\mathbf{q}_1 = \frac{1}{m}\mathbf{1}_m$, $\mathbf{F}_0^{(i)} = \mathbf{0}_d$, $\hat{\mathbf{F}}_0^{(i)} = \emptyset$ (AG) **or** $\mathbf{0}_d$ (CG) $\forall i \in [m]$.
for $t = 1, 2, \ldots, T$ **do**
  In parallel for all workers $i \in [m]$**:**
    Draw a sample $\mathbf{z}_t^{(i)}$ and send $\ell(\mathbf{w}_t^{(i)}; \mathbf{z}_t^{(i)})$ to the server
    Compute $\mathbf{g}_{w,t}^{(i)} = q_{t,i} \cdot \nabla\ell(\mathbf{w}_t^{(i)}; \mathbf{z}_t^{(i)})$
    $\{\mathbf{F}_t^{(i)}, \hat{\mathbf{F}}_t^{(i)}\}_{i\in[m]} = \mathcal{A}_c(\{\mathbf{F}_{t-1}^{(i)} + \mathbf{g}_{w,t}^{(i)}, \hat{\mathbf{F}}_{t-1}^{(i)}\}_{i\in[m]})$ *// communication among workers*
    Compute $\mathbf{w}_{t+1}^{(i)}$ via (9)
  On server:
    Construct $\mathbf{g}_{q,t}$ in (7) and update $\mathbf{L}_t = \mathbf{L}_{t-1} + \mathbf{g}_{q,t}$
    Compute $\mathbf{q}_{t+1}$ via (8) and send $q_{t+1,i}$ to worker i for all $i \in [m]$
end for
Return $\bar{\mathbf{w}}_T^{(i)} = T^{-1}\sum_{t=1}^T \mathbf{w}_t^{(i)}$ at each worker $i \in [m]$

To obtain uniform guarantees for all $\{\mathcal{R}_w^{(i)}\}_{i=1}^m$, we let workers jointly solve a D-OCO problem (Yan et al., 2013), with each worker computing its local gradient and using consensus to aggregate information across the network.

In summary, the above decomposition reduces the optimization error of each worker in SemiDec to a shared dual regret handled by PEA, a primal regret handled by D-OCO, and a model-discrepancy term. We next develop the corresponding algorithmic framework.

Remark. Different from standard uniform-weight D-OCO (Yan et al., 2013), our formulation involves time-varying dual weights. A direct reduction would require rescaling and leads to loose bounds. Instead, we exploit $\mathbf{q} \in \Delta_m$ to avoid the extra dependence on m; see Appendix D.2.1.

4.2 FRAMEWORK DESIGN AND INSTANTIATION

We first introduce the server-side update of the dual vector $\mathbf{q}_t$. To control the shared dual regret (5), the server updates $\mathbf{q}_t$ to maximize the shared term $\sum_{j=1}^m q_j R_j(\mathbf{w}_t^{(j)})$, equivalently minimizing $\sum_{j=1}^m q_j(1 - R_j(\mathbf{w}_t^{(j)}))$ under Assumption 2. Accordingly, in each round t, each worker i uploads only the scalar stochastic loss $\ell(\mathbf{w}_t^{(i)}; \mathbf{z}_t^{(i)})$ to the server. The server then forms gradient estimator

$$\mathbf{g}_{q,t} = [\mathbf{g}_{q,t}^{(1)}, \ldots, \mathbf{g}_{q,t}^{(m)}]^\top, \quad \text{with} \quad \mathbf{g}_{q,t}^{(i)} = 1 - \ell(\mathbf{w}_t^{(i)}; \mathbf{z}_t^{(i)}) \tag{7}$$

and computes $\mathbf{q}_{t+1}$ via FTRL (Orabona, 2019) as

$$\mathbf{q}_{t+1} = \arg\min_{\mathbf{q}\in\Delta_m} \{\langle \eta_q \mathbf{L}_t, \mathbf{q}\rangle + \nu_q(\mathbf{q})\}, \quad \text{with} \quad \mathbf{L}_t = \sum_{s=1}^t \mathbf{g}_{q,s} \tag{8}$$

The server then broadcasts $q_{t+1,i}$ to worker i, who uses it to scale the local stochastic gradient.

To update the local models, each worker performs FTRL using an estimate of the *global* cumulative gradient. Concretely, at round t, worker i maintains $\mathbf{F}_{t-1}^{(i)} \in \mathbb{R}^d$ and auxiliary state variables $\hat{\mathbf{F}}_{t-1}^{(j)}$ for $j \in \mathcal{N}_i$ when required by the consensus routine $\mathcal{A}_c$. Each worker i first computes the weighted stochastic gradient $\mathbf{g}_{w,t}^{(i)} = q_{t,i}\nabla\ell(\mathbf{w}_t^{(i)}; \mathbf{z}_t^{(i)})$, and then applies $\mathcal{A}_c$ to the message $\mathbf{F}_{t-1}^{(i)} + \mathbf{g}_{w,t}^{(i)}$, yielding an updated estimate $\mathbf{F}_t^{(i)}$. Finally, worker i computes its local model via

$$\mathbf{w}_{t+1}^{(i)} = \arg\min_{\mathbf{w}\in\mathcal{W}} \left\{ \left\langle \eta_w \mathbf{F}_t^{(i)}, \mathbf{w} \right\rangle + \nu_w(\mathbf{w}) \right\}. \tag{9}$$

Since consensus errors affect both the primal regret $\mathcal{R}_w^{(i)}$ and the model-discrepancy term arising from $\mathcal{R}_q^{(i)}$, we isolate its contribution in the corresponding bounds and treat the consensus routine $\mathcal{A}_c$ as a black box. We present the resulting framework termed GDRO-SEMIDEC in Algorithm 1, and establish its theoretical guarantees as follows.

Theorem 1. *Under Assumptions 1–3, set the step sizes as* $\eta_w = \sqrt{2}mD/(\sqrt{T}G)$ *and* $\eta_q = \sqrt{2(\ln m)/T}$. *Suppose the consensus routine* $\mathcal{A}_c$ *ensures that* $\sum_{i=1}^m \mathbf{F}_t^{(i)} = \sum_{i=1}^m \sum_{\tau=1}^t \mathbf{g}_{w,\tau}^{(i)}$ *and* $\mathbb{E}[\max_{i\in[m]} \|\mathbf{F}_t^{(i)} - \frac{1}{m}\sum_{r=1}^m \mathbf{F}_t^{(r)}\|_2] \le \varepsilon_c$ *for all* $t \in [T]$. *Then, Algorithm 1 guarantees*

$$\mathbb{E}\left[\varepsilon_\phi(\bar{\mathbf{w}}_T^{(i)})\right] \le 4\sqrt{\frac{\ln m}{T}} + \sqrt{\frac{2}{T}}DG + \frac{7\sqrt{2}mD}{\sqrt{T}}\varepsilon_c, \qquad \forall i \in [m].$$

Remark. Theorem 1 establishes a convergence guarantee for each worker's time-averaged local model that only depends on the consensus error ε_c. This modularity allows $\mathcal{A}_c$ to be instantiated by different consensus routines, with different update rules or communication rounds per iteration.

We next instantiate $\mathcal{A}_c$ with AG and CG, and present the corresponding convergence guarantees. While the analysis is formulated through an online learning framework, the optimization problem itself is offline, and thus we allow multiple communication steps per iteration.

Corollary 1. *Under the conditions of Theorem 1, let* $\mathcal{A}_c$ *be AG in Algorithm 3, with* $P \succeq 0$. *With the communication-round choice* $L = O((\log m)/\sqrt{\rho})$ *specified in Appendix D.3, Algorithm 1 guarantees, for all* $i \in [m]$, $\mathbb{E}[\varepsilon_\phi(\bar{\mathbf{w}}_T^{(i)})] \le (8\sqrt{2}DG + 4)\sqrt{(\ln m)/T}$.

Corollary 2. *Under the conditions of Theorem 1, let* $\mathcal{A}_c$ *be CG in Algorithm 4. With the communication-round choice* $L = O((\log m)/(\delta\rho^2))$ *specified in Appendix D.4, Algorithm 1 guarantees, for all* $i \in [m]$, $\mathbb{E}[\varepsilon_\phi(\bar{\mathbf{w}}_T^{(i)})] \le (15\sqrt{2}DG + 4)\sqrt{(\ln m)/T}$.

5 Decentralized GDRO

We now consider the Dec setting, where the lightweight server in SemiDec is removed and each worker i maintains both a local dual vector $\mathbf{q}_t^{(i)}$ and local model $\mathbf{w}_t^{(i)}$. This introduces dual inconsistency in addition to the model inconsistency already present in SemiDec.

5.1 Error Decomposition

Using again the centralized error decomposition, we bound the optimization error of $\bar{\mathbf{w}}_T^{(i)}$ as

$$\begin{aligned}&\varepsilon_\phi(\bar{\mathbf{w}}_T^{(i)})\\ \le&\frac{1}{T}\underbrace{\left(\max_{\mathbf{q}\in\Delta_m}\sum_{t=1}^T \phi(\mathbf{w}_t^{(i)}, \mathbf{q}) - \sum_{t=1}^T \phi(\mathbf{w}_t^{(i)}, \mathbf{q}_t^{(i)})\right)}_{\mathcal{R}_q^{(i)}} + \frac{1}{T}\underbrace{\left(\sum_{t=1}^T \phi(\mathbf{w}_t^{(i)}, \mathbf{q}_t^{(i)}) - \sum_{t=1}^T \phi(\mathbf{w}^*, \mathbf{q}_t^{(i)})\right)}_{\mathcal{R}_w^{(i)}}.\end{aligned}$$

The key difference to (3) is that the dual sequence is no longer shared across workers. In SemiDec, the shared dual sequence $\{\mathbf{q}_t\}_{t=1}^T$ makes $\mathcal{R}_w^{(i)}$ a regret on a shared objective sequence, so only $\mathcal{R}_q^{(i)}$ requires reformulation. In Dec, however, the local dual sequence $\{\mathbf{q}_t^{(i)}\}_{t=1}^T$ also makes $\mathcal{R}_w^{(i)}$ worker-dependent. Meanwhile, $\mathcal{R}_q^{(i)}$ faces the same cross-worker information barrier as in SemiDec, but must now be controlled for worker-specific dual sequences. Accordingly, we reuse the SemiDec reformulation for $\mathcal{R}_q^{(i)}$ and derive an additional reformulation for $\mathcal{R}_w^{(i)}$.

Analysis of $\mathcal{R}_q^{(i)}$. As in the SemiDec setting, we apply the decomposition in (4) to $\phi(\mathbf{w}_t^{(i)}, \mathbf{q})$. This decomposes $\mathcal{R}_q^{(i)}$ into two parts. The first is a local dual regret term:

$$\max_{\mathbf{q}\in\Delta_m}\sum_{t=1}^T\sum_{j=1}^m q_j R_j(\mathbf{w}_t^{(j)}) - \sum_{t=1}^T\sum_{j=1}^m q_{t,j}^{(i)} R_j(\mathbf{w}_t^{(j)}), \tag{10}$$

which evaluates the local dual sequence $\{\mathbf{q}_t^{(i)}\}_{t=1}^T$ on the shared global objective $\sum_{j=1}^m q_j R_j(\mathbf{w}_t^{(j)})$, leading to a D-OCO formulation. The second is a discrepancy term bounded by $2G\sum_{t=1}^T \max_{j\in[m]} \|\mathbf{w}_t^{(i)} - \mathbf{w}_t^{(j)}\|_2$, which measures cumulative disagreement among local models.

Analysis of $\mathcal{R}_w^{(i)}$. This term introduces an additional difficulty specific to Dec. Since $\mathbf{q}_t^{(i)}$ varies across workers, the objective $\phi(\mathbf{w}, \mathbf{q}_t^{(i)})$ is no longer shared across workers. To remove this worker dependence, we introduce the proxy vector $\hat{\mathbf{q}}_t = (q_{t,1}^{(1)}, \ldots, q_{t,m}^{(m)})$ and rewrite

$$\phi(\mathbf{w}, \mathbf{q}_t^{(i)}) = \sum_{j=1}^{m} q_{t,j}^{(j)} R_j(\mathbf{w}) + \sum_{j=1}^{m} \big(q_{t,j}^{(i)} - q_{t,j}^{(j)}\big) R_j(\mathbf{w}). \tag{11}$$

The first term on the right-hand side of (11) is shared across workers and induces the primal regret

$$\sum_{t=1}^{T}\sum_{j=1}^{m} q_{t,j}^{(j)} R_j(\mathbf{w}_t^{(i)}) - \sum_{t=1}^{T}\sum_{j=1}^{m} q_{t,j}^{(j)} R_j(\mathbf{w}^*). \tag{12}$$

We control this regret uniformly across workers by letting them jointly solve a D-OCO problem.

A key challenge is controlling the second term in (11), which measures inconsistency among local dual vectors and can be bounded by $\|\mathbf{q}_t^{(i)} - \hat{\mathbf{q}}_t\|_1$ under Assumption 2. Unlike the shared dual vector in SemiDec, the proxy $\hat{\mathbf{q}}_t$ need not lie in Δ_m, because its coordinates come from different workers' local dual vectors. Consequently, the simplex-based argument used in SemiDec no longer applies, while the direct bound $\|\mathbf{q}_t^{(i)} - \hat{\mathbf{q}}_t\|_1 \le m$ incurs an undesirable factor of m. To overcome this difficulty, we introduce a virtual simplex reference sequence $\{\tilde{\mathbf{q}}_t\}_{t\ge1}$ driven by the averaged dual gradient and bound $\|\mathbf{q}_t^{(i)} - \hat{\mathbf{q}}_t\|_1 \le \|\mathbf{q}_t^{(i)} - \tilde{\mathbf{q}}_t\|_1 + \|\hat{\mathbf{q}}_t - \tilde{\mathbf{q}}_t\|_1$. This construction allows both deviations to be controlled through the dual consensus error, avoiding the additional factor of m.

In summary, the optimization error of each worker's output is bounded by two D-OCO regret terms over shared global objectives, together with two discrepancy terms.

5.2 Framework Design and Instantiation

To control the two coupled D-OCO regrets, we design a consensus-based FTRL scheme in which each worker uses $\mathcal{A}_{c,w}$ and $\mathcal{A}_{c,q}$ to maintain and update local estimates $\mathbf{F}_t^{(i)}$ and $\mathbf{L}_t^{(i)}$ of the cumulative global gradients for the primal and dual updates, respectively.

We first introduce the update of the local dual vectors. To optimize the shared dual objective in (10), we equivalently minimize $\sum_{j=1}^{m} q_j(1 - R_j(\mathbf{w}_t^{(j)}))$ under Assumption 2. At round t, each worker i forms a sparse stochastic contribution to the global dual gradient $\mathbf{g}_{q,t}^{(i)} = \mathbf{e}_i\big(1 - \ell(\mathbf{w}_t^{(i)}; \mathbf{z}_t^{(i)})\big)$, where $\mathbf{e}_i \in \mathbb{R}^m$ denotes the i-th standard basis vector. The cumulative messages $\mathbf{L}_{t-1}^{(i)} + \mathbf{g}_{q,t}^{(i)}$ are then aggregated through $\mathcal{A}_{c,q}$ to update $\mathbf{L}_t^{(i)}$. Finally, each worker i performs the FTRL update

$$\mathbf{q}_{t+1}^{(i)} = \arg\min_{\mathbf{q}\in\Delta_m} \Big\{ \big\langle \eta_q \mathbf{L}_t^{(i)}, \mathbf{q} \big\rangle + \nu_q(\mathbf{q}) \Big\}. \tag{13}$$

The model update follows the same consensus-based FTRL scheme as in SemiDec, but now uses locally maintained dual weights. To optimize the shared primal objective $\sum_{j=1}^{m} q_{t,j}^{(j)} R_j(\mathbf{w})$ in (12), each worker i forms the weighted stochastic gradient $\mathbf{g}_{w,t}^{(i)} = q_{t,i}^{(i)} \nabla \ell(\mathbf{w}_t^{(i)}; \mathbf{z}_t^{(i)})$. The workers then aggregate the cumulative messages $\mathbf{F}_{t-1}^{(i)} + \mathbf{g}_{w,t}^{(i)}$ through $\mathcal{A}_{c,w}$, yielding updated estimates $\mathbf{F}_t^{(i)}$. Finally, each worker i updates its local model via the FTRL step in (9).

We summarize the resulting framework, termed GDRO-Dec, in Algorithm 2 and establish its convergence guarantee in the following theorem.

Theorem 2. *Under Assumptions 1–3, set $\eta_w = \sqrt{2}mD/(G\sqrt{T})$ and $\eta_q = m\sqrt{2(\ln m)/T}$. Suppose the consensus routines satisfy $\sum_i \mathbf{F}_t^{(i)} = \sum_i \sum_{\tau=1}^{t} \mathbf{g}_{w,\tau}^{(i)}$ and $\sum_i \mathbf{L}_t^{(i)} = \sum_i \sum_{\tau=1}^{t} \mathbf{g}_{q,\tau}^{(i)}$. Writing $\bar{\mathbf{F}}_t = m^{-1}\sum_i \mathbf{F}_t^{(i)}$ and $\bar{\mathbf{L}}_t = m^{-1}\sum_i \mathbf{L}_t^{(i)}$, assume $\big(\mathbb{E}[\max_i \|\mathbf{F}_t^{(i)} - \bar{\mathbf{F}}_t\|_2^2]\big)^{1/2} \le \varepsilon_{c,w}$ and $\big(\mathbb{E}[\max_i \|\mathbf{L}_t^{(i)} - \bar{\mathbf{L}}_t\|_2^2]\big)^{1/2} \le \varepsilon_{c,q}$ for all $t \in [T]$. Then, Algorithm 2 guarantees*

$$\mathbb{E}\left[\varepsilon_\phi(\bar{\mathbf{w}}_T^{(i)})\right] \le O\left(\frac{C_1 + C_2 + C_3 + \sqrt{\log m}}{\sqrt{T}}\right), \qquad \forall i \in [m], \tag{14}$$

Algorithm 2 GDRO-DEC

1: Initialize $\mathbf{w}_1^{(i)} = \arg\min_{\mathbf{w}\in\mathcal{W}} \nu_w(\mathbf{w})$, $\mathbf{F}_0^{(i)} = \mathbf{0}_d$, $\mathbf{L}_0^{(i)} = \mathbf{0}_m$, $\mathbf{q}_1^{(i)} = \frac{1}{m}\mathbf{1}_m$, $\hat{\mathbf{F}}_0^{(i)} = \hat{\mathbf{L}}_0^{(i)} = \emptyset$ (AG) **or** $\hat{\mathbf{F}}_0^{(i)} = \mathbf{0}_d$, $\hat{\mathbf{L}}_0^{(i)} = \mathbf{0}_m$ (CG), $\forall i \in [m]$.
2: **for** $t = 1, 2, \ldots, T$ **do**
3: **In parallel for all workers** $i \in [m]$**:**
4: Draw a sample $\mathbf{z}_t^{(i)}$ and compute $\mathbf{g}_{w,t}^{(i)} = q_{t,i}^{(i)} \nabla \ell(\mathbf{w}_t^{(i)}; \mathbf{z}_t^{(i)})$, $\mathbf{g}_{q,t}^{(i)} = \mathbf{e}_i(1 - \ell(\mathbf{w}_t^{(i)}; \mathbf{z}_t^{(i)}))$
5: $\{\mathbf{F}_t^{(i)}, \hat{\mathbf{F}}_t^{(i)}\}_{i\in[m]} = \mathcal{A}_{c,w}(\{\mathbf{F}_{t-1}^{(i)} + \mathbf{g}_{w,t}^{(i)}, \hat{\mathbf{F}}_{t-1}^{(i)}\}_{i\in[m]})$ *// primal consensus*
6: $\{\mathbf{L}_t^{(i)}, \hat{\mathbf{L}}_t^{(i)}\}_{i\in[m]} = \mathcal{A}_{c,q}(\{\mathbf{L}_{t-1}^{(i)} + \mathbf{g}_{q,t}^{(i)}, \hat{\mathbf{L}}_{t-1}^{(i)}\}_{i\in[m]})$ *// dual consensus*
7: Update $\mathbf{q}_{t+1}^{(i)}$ in (13) and $\mathbf{w}_{t+1}^{(i)}$ in (9)
8: **end for**
9: **Return** $\bar{\mathbf{w}}_T^{(i)} = T^{-1}\sum_{t=1}^{T} \mathbf{w}_t^{(i)}$ at each worker $i \in [m]$

where C_1, C_2, and C_3 are consensus-dependent terms defined in Appendix D.5. Under exact consensus, $C_1 = 1$ and $C_2 = C_3 = 0$.

Similar to Theorem 1, Theorem 2 isolates the effect of communication through the consensus errors $\varepsilon_{c,w}$ and $\varepsilon_{c,q}$, allowing $\mathcal{A}_{c,w}$ and $\mathcal{A}_{c,q}$ to be instantiated independently. For example, one may use AG for the low-dimensional dual updates and CG for the high-dimensional model updates. For simplicity, we next present results for setting both $\mathcal{A}_{c,w}$ and $\mathcal{A}_{c,q}$ to either AG or CG.

Corollary 3. *Under the conditions of Theorem 2, let $\mathcal{A}_{c,w}$ and $\mathcal{A}_{c,q}$ be AG in Algorithm 3, with $P \succeq 0$. With the communication-round choices $L_w, L_q = O\left((\log m)/\sqrt{\rho}\right)$ specified in Appendix D.6, Algorithm 2 guarantees, for all $i \in [m]$, $\mathbb{E}[\varepsilon_\phi(\bar{\mathbf{w}}_T^{(i)})] \le (32DG + 9\sqrt{2})\sqrt{(\ln m)/T}$.*

Corollary 4. *Under the conditions of Theorem 2, let $\mathcal{A}_{c,w}$ and $\mathcal{A}_{c,q}$ be CG in Algorithm 4. With the communication-round choices $L_w, L_q = O((\log m)/(\delta\rho^2))$ specified in Appendix D.7, Algorithm 2 guarantees, for all $i \in [m]$, $\mathbb{E}[\varepsilon_\phi(\bar{\mathbf{w}}_T^{(i)})] \le (61DG + 9\sqrt{2})\sqrt{(\ln m)/T}$.*

6 DISCUSSION OF THEORETICAL GUARANTEES

To the best of our knowledge, Theorems 1 and 2 provide the *first* per-worker convergence guarantees for the original GDRO objective over decentralized networks. In both settings, the per-worker optimization error bound consists of an intrinsic optimization term and communication-induced consensus errors, which we discuss below.

First, the optimization term is tight in sample complexity up to a logarithmic factor in m. Under exact consensus, i.e., $\varepsilon_c = 0$ in SemiDec and $\varepsilon_{c,w} = \varepsilon_{c,q} = 0$ in Dec, the error reduces to $O(\sqrt{(\log m)/T})$. Thus, finding an ε-optimal solution requires $O((\log m)/\varepsilon^2)$ updates and $O((m \log m)/\varepsilon^2)$ samples, matching the best-known centralized GDRO results (Zhang et al., 2023) and the information-theoretic lower bound $\Omega(m/\varepsilon^2)$ up to a factor of $O(\log m)$ (Soma et al., 2022).

Second, to preserve the $O(\sqrt{(\log m)/T})$ rate, we choose the number of communication rounds in Corollaries 1–4 according to the best-known scaling of the D-OCO subproblems. Specifically, the AG instantiations match the communication-round scaling of the near-optimal D-OCO method in Wan et al. (2025), while the CG instantiations inherit the state-of-the-art dependence on network connectivity and compression (Yang et al., 2026). Importantly, these choices are sufficient rather than necessary, and the main theorems remain valid for different numbers of consensus rounds.

7 EXPERIMENTS

We evaluate our SemiDec and Dec methods on three datasets. Additional experimental details and worst-worker loss results are provided in Appendix B.

Experimental setup. We consider Adult (Becker & Kohavi, 1996), MNIST (Lecun et al., 1998), and CIFAR-10 (Krizhevsky, 2009). We instantiate GDRO-SEMIDEC and GDRO-DEC with AG and

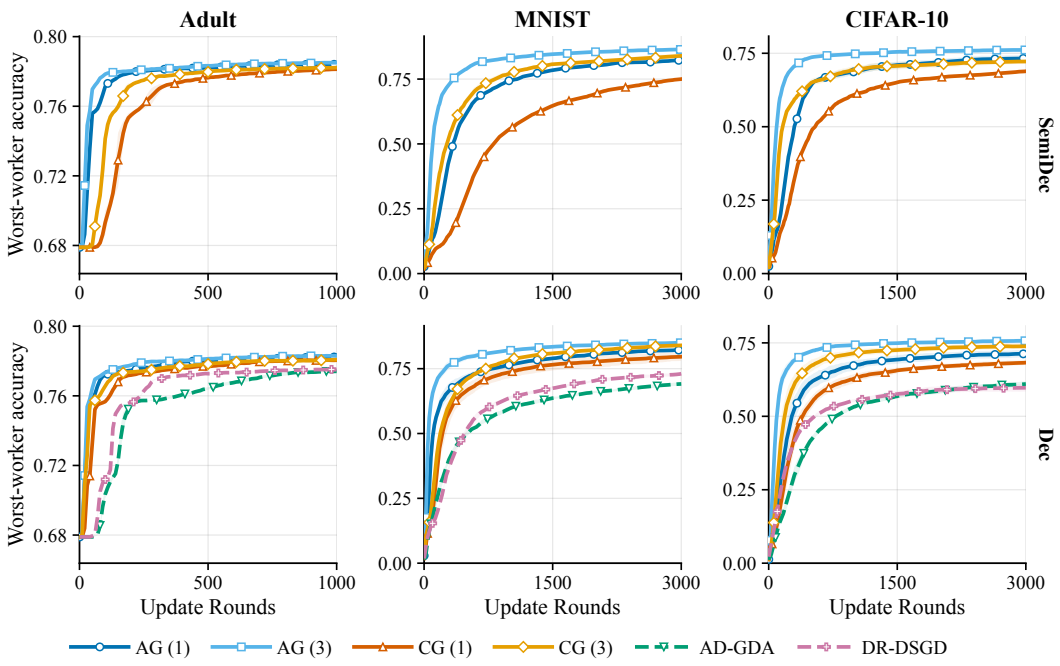


Figure 1: Worst-worker accuracy over update rounds on Adult, MNIST, and CIFAR-10. The top and bottom rows correspond to SemiDec and Dec, respectively. Curves report the mean over five random seeds, with shaded regions showing one standard deviation.

CG, denoting the resulting variants by AG(L) and CG(L) within each setting, where L is the number of consensus rounds per update. For Dec, we set $L_w = L_q = L$ and compare with DR-DSGD (Issaid et al., 2022) and AD-GDA (Zecchin et al., 2023). AD-GDA and the CG-based variants use Top-k compression (Stich et al., 2018) with a 50% compression ratio. Let $\mathrm{Acc}_j(\bar{\mathbf{w}}_t^{(i)})$ and $R_j(\bar{\mathbf{w}}_t^{(i)})$ denote the test accuracy and loss of worker i's local model on worker j's test data, respectively. We report worst-worker accuracy $\min_{i,j\in[m]} \mathrm{Acc}_j(\bar{\mathbf{w}}_t^{(i)})$ in the main text and worst-worker loss $\max_{i,j\in[m]} R_j(\bar{\mathbf{w}}_t^{(i)})$ in Appendix B.2. Workers communicate over an undirected ring graph. We train linear classifiers on Adult and MNIST. On CIFAR-10, all methods use the same pretrained ResNet-20 (Polivin, 2025) with a frozen backbone and a randomly initialized linear head.

Results. Figure 1 reports the worst-worker accuracy over update rounds in both settings. First, in Dec, our variants generally converge faster and attain higher accuracy than the baselines. These results support the practical advantage of directly optimizing the unregularized GDRO objective, whereas the baselines optimize dual-regularized surrogates. Second, across both settings, the $L = 3$ variants generally improve faster than their $L = 1$ counterparts. This behavior is consistent with the consensus-error dependence in Theorems 1 and 2, as additional consensus rounds reduce the consensus errors. Finally, both AG and CG exhibit stable convergence across the datasets, demonstrating the flexibility of our framework in accommodating different consensus routines.

8 Conclusion

In this paper, we study GDRO over decentralized networks in both SemiDec and Dec settings. Our key technical contribution is an error decomposition that bounds each worker's local optimization error by global regret terms plus disagreement-induced discrepancies, leading to a D-OCO problem and a PEA problem in SemiDec, and two coupled D-OCO problems in Dec. Based on this decomposition, we develop modular frameworks supporting different consensus routines and bound the expected optimization error of each worker's time-averaged local model for the original GDRO objective. We then instantiate the frameworks with AG and CG, yielding an $O(\sqrt{(\log m)/T})$ convergence rate. Experiments demonstrate improved worst-worker performance over baselines.

AI USE STATEMENT

In this work, we used generative AI tools to assist with language editing, manuscript organization, and the presentation of experimental figures. All AI-assisted content was reviewed by the authors, who take full responsibility for the final manuscript.

REFERENCES

Onur Avci, Osama Abdeljaber, Serkan Kiranyaz, Mohammed Hussein, and Daniel J. Inman. Wireless and real-time structural damage detection: A novel decentralized method for wireless sensor networks. *Journal of Sound and Vibration*, 424:158–172, 2018.

Haomin Bai, Dingzhi Yu, Shuai Li, Haipeng Luo, and Lijun Zhang. Group distributionally robust optimization with flexible sample queries. *ArXiv e-prints*, arXiv:2505.15212, 2025.

Barry Becker and Ronny Kohavi. Adult. UCI Machine Learning Repository, 1996.

Sébastien Bubeck and Nicolò Cesa-Bianchi. Regret analysis of stochastic and nonstochastic multi-armed bandit problems. *Foundations and Trends in Machine Learning*, 5(1):1–122, 2012.

N. Cesa-Bianchi and G. Lugosi. *Prediction, Learning, and Games*. Cambridge University Press, 2006.

Jianshu Chen and Ali H. Sayed. Diffusion adaptation strategies for distributed optimization and learning over networks. *IEEE Transactions on Signal Processing*, 60(8):4289–4305, 2012.

Joseph Clancy, Darragh Mullins, Brian Deegan, Jonathan Horgan, Enda Ward, Ciarán Eising, Patrick Denny, Edward Jones, and Martin Glavin. Wireless access for v2x communications: Research, challenges and opportunities. *IEEE Communications Surveys & Tutorials*, 26(3):2082–2119, 2024.

Yuyang Deng, Mohammad Mahdi Kamani, and Mehrdad Mahdavi. Distributionally robust federated averaging. In *Advances in Neural Information Processing Systems 33*, pp. 15111–15122, 2020.

John C. Duchi, Alekh Agarwal, and Martin J. Wainwright. Dual averaging for distributed optimization: Convergence analysis and network scaling. *IEEE Transactions on Automatic Control*, 57 (3):592–606, 2012.

Yoav Freund and Robert E Schapire. A decision-theoretic generalization of on-line learning and an application to boosting. *Journal of Computer and System Sciences*, 55(1):119–139, 1997.

Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial networks. *Commun. ACM*, 63(11): 139–144, 2020.

Yuanxiong Guo, Ying Sun, Rui Hu, and Yanmin Gong. Hybrid local SGD for federated learning with heterogeneous communications. In *International Conference on Learning Representations*, 2022.

Zhishuai Guo and Tianbao Yang. Communication-efficient federated group distributionally robust optimization. In *Advances in Neural Information Processing Systems 37*, pp. 23040–23077, 2024.

Nika Haghtalab, Michael I. Jordan, and Eric Zhao. On-demand sampling: Learning optimally from multiple distributions. In *Advances in Neural Information Processing Systems 35*, pp. 406–419, 2022.

Lie He, An Bian, and Martin Jaggi. Cola: Decentralized linear learning. In *Advances in Neural Information Processing Systems 31*, pp. 4541–4551, 2018.

Roger A Horn and Charles R Johnson. *Matrix analysis*. Cambridge university press, 2012.

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

Kevin Hsieh, Amar Phanishayee, Onur Mutlu, and Phillip Gibbons. The non-IID data quagmire of decentralized machine learning. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119, pp. 4387–4398, 2020.

Chaouki Ben Issaid, Anis Elgabli, and Mehdi Bennis. DR-DSGD: A distributionally robust decentralized learning algorithm over graphs. *Transactions on Machine Learning Research*, 2022.

Vassilis Kekatos and Georgios B. Giannakis. Distributed robust power system state estimation. *IEEE Transactions on Power Systems*, 28(2):1617–1626, 2013.

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

Alec Koppel, Felicia Y. Jakubiec, and Alejandro Ribeiro. A saddle point algorithm for networked online convex optimization. *IEEE Transactions on Signal Processing*, 63(19):5149–5164, 2015.

Alex Krizhevsky. Learning multiple layers of features from tiny images. Technical report, University of Toronto, 2009.

Y. Lecun, L. Bottou, Y. Bengio, and P. Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11), 1998.

Frank Po-Chen Lin, Seyyedali Hosseinalipour, Sheikh Shams Azam, Christopher G. Brinton, and Nicolò Michelusi. Semi-decentralized federated learning with cooperative d2d local model aggregations. *IEEE Journal on Selected Areas in Communications*, 39(12):3851–3869, 2021.

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

Mingrui Liu, Wei Zhang, Youssef Mroueh, Xiaodong Cui, Jarret Ross, Tianbao Yang, and Payel Das. A decentralized parallel algorithm for training generative adversarial nets. In *Advances in Neural Information Processing Systems 33*, pp. 11056–11070, 2020.

Enrique Tomás Martínez Beltrán, Mario Quiles Pérez, Pedro Miguel Sánchez Sánchez, Sergio López Bernal, Gérôme Bovet, Manuel Gil Pérez, Gregorio Martínez Pérez, and Alberto Huertas Celdrán. Decentralized federated learning: Fundamentals, state of the art, frameworks, trends, and challenges. *IEEE Communications Surveys & Tutorials*, 25(4):2983–3013, 2023.

Dmitry Metelev, Alexander Rogozin, Alexander Gasnikov, and Dmitry Kovalev. Decentralized saddle-point problems with different constants of strong convexity and strong concavity. *Computational Management Science*, 21(1), 2023.

Mehryar Mohri, Gary Sivek, and Ananda Theertha Suresh. Agnostic federated learning. In *Proceedings of the 36th International Conference on Machine Learning*, pp. 4615–4625, 2019.

Mohammad Navidi, Seyed Masoud Moghaddas-Tafreshi, and Amir Mohammad Alishvandi. A semi-decentralized framework for simultaneous expansion planning of privately owned multi-regional energy systems and sub-transmission grid. *International Journal of Electrical Power & Energy Systems*, 128:106795, 2021.

Angelia Nedic and Asuman Ozdaglar. Distributed subgradient methods for multi-agent optimization. *IEEE Transactions on Automatic Control*, 54(1):48–61, 2009.

Reza Olfati-Saber, J. Alex Fax, and Richard M. Murray. Consensus and cooperation in networked multi-agent systems. *Proceedings of the IEEE*, 95(1):215–233, 2007.

Francesco Orabona. A modern introduction to online learning. *ArXiv e-prints*, arXiv:1912.13213, 2019.

Lerrel Pinto, James Davidson, Rahul Sukthankar, and Abhinav Gupta. Robust adversarial reinforcement learning. In *Proceedings of the 34th International Conference on Machine Learning*, volume 70, pp. 2817–2826, 2017.

Sergey Polivin. Resnet-20 model for cifar-10. Hugging Face, 2025. URL https://huggingface.co/spolivin/cnn-cifar10.

Shiori Sagawa, Pang Wei Koh, Tatsunori B. Hashimoto, and Percy Liang. Distributionally robust neural networks for group shifts: On the importance of regularization for worst-case generalization. In *International Conference on Learning Representations*, 2020.

Tasuku Soma, Khashayar Gatmiry, and Stefanie Jegelka. Optimal algorithms for group distributionally robust optimization and beyond. *ArXiv e-prints*, arXiv:2212.13669, 2022.

Tasuku Soma, Khashayar Gatmiry, and Stefanie Jegelka. Optimal algorithms for group distributionally robust optimization and beyond. *ArXiv e-prints*, arXiv:2212.13669, 2025.

Sebastian U. Stich, Jean-Baptiste Cordonnier, and Martin Jaggi. Sparsified sgd with memory. In *Advances in Neural Information Processing Systems 31*, pp. 4452–4463, 2018.

Ioannis Tsaknakis, Mingyi Hong, and Sijia Liu. Decentralized min-max optimization: Formulations, algorithms and applications in network poisoning attack. In *IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 5755–5759, 2020.

Zhipeng Tu, Xi Wang, Yiguang Hong, Lei Wang, Deming Yuan, and Guodong Shi. Distributed online convex optimization with compressed communication. In *Advances in Neural Information Processing Systems 35*, pp. 34492–34504, 2022.

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

Yuanyu Wan, Tong Wei, Bo Xue, Mingli Song, and Lijun Zhang. Optimal and efficient algorithms for decentralized online convex optimization. *Journal of Machine Learning Research*, 26(135): 1–43, 2025.

He Wang and Yuejie Chi. Communication-efficient federated optimization over semi-decentralized networks. *IEEE Transactions on Signal and Information Processing over Networks*, 11:147–160, 2025.

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

Ermin Wei and Asuman Ozdaglar. Distributed alternating direction method of multipliers. In *2012 IEEE 51st IEEE Conference on Decision and Control*, pp. 5445–5450, 2012.

Wenhan Xian, Feihu Huang, Yanfu Zhang, and Heng Huang. A faster decentralized algorithm for nonconvex minimax problems. In *Advances in Neural Information Processing Systems 34*, pp. 25865–25877, 2021.

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

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

Sifan Yang, Wenhao Yang, Wei Jiang, and Lijun Zhang. Distributed online convex optimization with efficient communication: Improved algorithm and lower bounds. *ArXiv e-prints*, 2601.04907, 2026.

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

Mang Ye, Xiuwen Fang, Bo Du, Pong C. Yuen, and Dacheng Tao. Heterogeneous federated learning: State-of-the-art and research challenges. *ACM Comput. Surv.*, 56(3), 2023b.

Runze You, Kun Huang, and Shi Pu. Decentralized min-max optimization with gradient tracking. *ArXiv e-prints*, 2505.10631, 2025.

Kun Yuan, Qing Ling, and Wotao Yin. On the convergence of decentralized gradient descent. *SIAM Journal on Optimization*, 26(3):1835–1854, 2016.

Matteo Zecchin, Marios Kountouris, and David Gesbert. Communication-efficient distributionally robust decentralized learning. *Transactions on Machine Learning Research*, 2023.

Lijun Zhang, Peng Zhao, Zhenhua Zhuang, Tianbao Yang, and Zhi-Hua Zhou. Stochastic approximation approaches to group distributionally robust optimization. In *Advances in Neural Information Processing Systems 36*, 2023.

Lijun Zhang, Haomin Bai, Peng Zhao, and Zhi-Hua Zhou. Stochastic approximation approaches to group distributionally robust optimization and beyond. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pp. 1–12, 2026.

Wenpeng Zhang, Peilin Zhao, Wenwu Zhu, Steven C. H. Hoi, and Tong Zhang. Projection-free distributed online learning in networks. In *Proceedings of the 34th International Conference on Machine Learning*, volume 70, pp. 4054–4062, 2017.

Hao Zhou, Xiaoyu Wang, Chang Yao, Mingli Song, and Yuanyu Wan. Revisiting decentralized online convex optimization with compressed communication. *ArXiv e-prints*, 2607.01665, 2026.

A Consensus Routine Instantiations

In this appendix, we present two representative instantiations of the black-box consensus routine $\mathcal{A}_c$: accelerated gossip (AG) and Choco-gossip (CG). We provide their procedures and the standard contraction guarantees used in our analysis.

A.1 Accelerated Gossip

AG (Liu & Morse, 2011) is an accelerated linear consensus scheme. Starting from $\mathbf{x}_0^{(i)} = \mathbf{x}_1^{(i)}$ for all $i \in [m]$, it performs L communication rounds. At each round $k \in [L]$, worker i exchanges $\mathbf{x}_k^{(i)}$ with its neighbors and updates

$$\mathbf{x}_{k+1}^{(i)} = (1+\theta) \sum_{j \in \mathcal{N}_i} P_{i,j} \mathbf{x}_k^{(j)} - \theta \mathbf{x}_{k-1}^{(i)}.$$

where $\theta > 0$ is the stepsize. For the AG instantiation, we additionally assume that $P \succeq 0$. The AG procedure is summarized in Algorithm 3 and enjoys the following guarantee.

Lemma 1. *(Ye et al., 2023a) Let the step size $\theta = \frac{1}{1+\sqrt{1-\lambda_\star^2(P)}}$. Define $X_k = \left[\mathbf{x}_k^{(1)}, \dots, \mathbf{x}_k^{(m)}\right] \in \mathbb{R}^{d \times m}$ and the average $\bar{\mathbf{x}} = \frac{1}{m} X_0 \mathbf{1} \in \mathbb{R}^d$. Then $\bar{\mathbf{x}} = \frac{1}{m} X_{L+1} \mathbf{1}$ and*

$$\left\| X_{L+1} - \bar{\mathbf{x}} \mathbf{1}^\top \right\|_F \leq \sqrt{14} c^L \left\| X_0 - \bar{\mathbf{x}} \mathbf{1}^\top \right\|_F,$$

where $c = 1 - (1 - 1/\sqrt{2})\sqrt{1 - \lambda_\star(P)}$ and $\lambda_\star(P)$ denotes the second largest eigenvalue of P.

A.2 Choco-gossip

CG is a communication-efficient consensus procedure that reduces bandwidth by exchanging *compressed* messages. It relies on a compressor $\mathcal{C}$ defined as follows.

Definition 1. Let $\mathcal{Q} \subseteq \mathbb{R}^d$ and $\delta \in (0, 1]$. An operator $\mathcal{C} : \mathbb{R}^d \to \mathbb{R}^d$ is a δ-approximate compressor over $\mathcal{Q}$ if

$$\mathbb{E}\left[\|\mathcal{C}(\mathbf{x}) - \mathbf{x}\|_2^2 \right] \leq (1-\delta) \|\mathbf{x}\|_2^2, \qquad \forall \mathbf{x} \in \mathcal{Q}. \tag{15}$$

where the expectation is taken with respect to the internal randomness of $\mathcal{C}$.

Algorithm 3 AG($\{\mathbf{x}_1^{(i)}\}_{i\in[m]}$): Accelerated Gossip

Input: Initial states $\{\mathbf{x}_1^{(i)}\}_{i\in[m]}$, rounds L, communication matrix $P \in \mathbb{R}^{m\times m}$, mixing coefficient $\theta > 0$.

1: Initialize $\mathbf{x}_0^{(i)} = \mathbf{x}_1^{(i)}$ for all $i \in [m]$
2: **for** $k = 1, \dots, L$ **do**
3: **In parallel for all workers** $i \in [m]$**:**
4: Send $\mathbf{x}_k^{(i)}$ to all neighbors $j \in \mathcal{N}_i$ and receive $\{\mathbf{x}_k^{(j)}\}_{j\in\mathcal{N}_i}$
5: Update $\mathbf{x}_{k+1}^{(i)} = (1+\theta)\sum_{j\in\mathcal{N}_i} P_{i,j}\,\mathbf{x}_k^{(j)} - \theta\mathbf{x}_{k-1}^{(i)}$
6: **end for**
7: **Return** $\{\mathbf{x}_{L+1}^{(i)}\}_{i\in[m]}$

Algorithm 4 CG($\{\mathbf{x}_1^{(i)}, \hat{\mathbf{x}}_1^{(i)}\}_{i\in[m]}$): Choco-gossip

Input: States $\mathbf{x}_1^{(i)}, \hat{\mathbf{x}}_1^{(i)}$ for all $i \in [m]$, step size γ, rounds L, compressor $\mathcal{C}$, communication matrix P

1: **for** $k = 1, 2, \dots, L$ **do**
2: **In parallel for all workers** $i \in [m]$**:**
3: Send $\Delta_k^{(i)} = \mathcal{C}(\mathbf{x}_k^{(i)} - \hat{\mathbf{x}}_k^{(i)})$ to all neighbors $j \in \mathcal{N}_i$
4: Update $\hat{\mathbf{x}}_{k+1}^{(j)} = \hat{\mathbf{x}}_k^{(j)} + \Delta_k^{(j)}$ for $j \in \mathcal{N}_i$
5: Update $\mathbf{x}_{k+1}^{(i)} = \mathbf{x}_k^{(i)} + \gamma\sum_{j\in\mathcal{N}_i} P_{i,j}(\hat{\mathbf{x}}_{k+1}^{(j)} - \hat{\mathbf{x}}_{k+1}^{(i)})$
6: **end for**
7: **Return** $\{\mathbf{x}_{L+1}^{(i)}, \hat{\mathbf{x}}_{L+1}^{(i)}\}_{i\in[m]}$

At each communication round $k \in [L]$, CG maintains both the current states $\{\mathbf{x}_k^{(i)}\}_{i\in[m]}$ and the transmitted/received information $\{\hat{\mathbf{x}}_k^{(i)}\}_{i\in[m]}$ (i.e., worker i stores $\{\hat{\mathbf{x}}_k^{(j)}\}_{j\in\mathcal{N}_i}$). Each worker i sends the compressed difference $\Delta_k^{(i)} = \mathcal{C}(\mathbf{x}_k^{(i)} - \hat{\mathbf{x}}_k^{(i)})$ to its neighbors. Upon receiving $\{\Delta_k^{(j)}\}_{j\in\mathcal{N}_i}$, worker i updates $\{\hat{\mathbf{x}}_{k+1}^{(j)}\}_{j\in\mathcal{N}_i}$ and $\mathbf{x}_{k+1}^{(i)}$. CG is summarized in Algorithm 4 and satisfies the following guarantee.

Lemma 2. *(Koloskova et al., 2019) Let P satisfy Assumption 3, and assume $\mathcal{C}$ satisfies (15) on $\mathbb{R}^d$ with fresh internal randomness at each call. Define $X_t = [\mathbf{x}_t^{(1)}, \dots, \mathbf{x}_t^{(m)}]$, $\hat{X}_t = [\hat{\mathbf{x}}_t^{(1)}, \dots, \hat{\mathbf{x}}_t^{(m)}]$, and $\bar{\mathbf{x}}_t = \frac{1}{m}\sum_{i=1}^m \mathbf{x}_t^{(i)}$. Algorithm 4 ensures $\bar{\mathbf{x}}_t = \bar{\mathbf{x}}$ for all $t \in [L]$. Define $\bar{X} = \bar{\mathbf{x}}\mathbf{1}^\top$, and $e_{t+1} = \mathbb{E}[\|X_{t+1} - \hat{X}_{t+1}\|_F^2 + \|X_{t+1} - \bar{X}\|_F^2]$. With step size $\gamma = \frac{\rho\delta}{16\rho+\rho^2+4\beta^2+2\rho\beta^2-8\rho\delta}$, it holds that*

$$e_{L+1} \le \left(1 - \frac{\gamma\rho}{2}\right) e_L \le \left(1 - \frac{\gamma\rho}{2}\right)^L e_1. \tag{16}$$

B More Experiment Results

B.1 Experimental Details

We run all experiments in Python on an Ubuntu 20.04.4 LTS server with two Intel(R) Xeon(R) Platinum 8358P processors at 2.60 GHz.

For Adult, we use the official training and test files and form $m = 6$ workers using the Cartesian product of race $\in \{\text{White}, \text{Black}, \text{Other}\}$ and sex $\in \{\text{Male}, \text{Female}\}$. For MNIST and CIFAR-10, we use the official training and test sets with $m = 10$ workers. Each split is partitioned independently: 80% of samples from class i are assigned to worker i, and the remaining 20% are distributed approximately uniformly among the other workers.

Workers communicate over an undirected ring with mixing matrix $P = I - \frac{L_G}{2d_{\max}}$. For the degree-two ring used in our experiments, $P_{ii} = 1/2$ and $P_{i,i\pm1} = 1/4$, with indices taken modulo m.

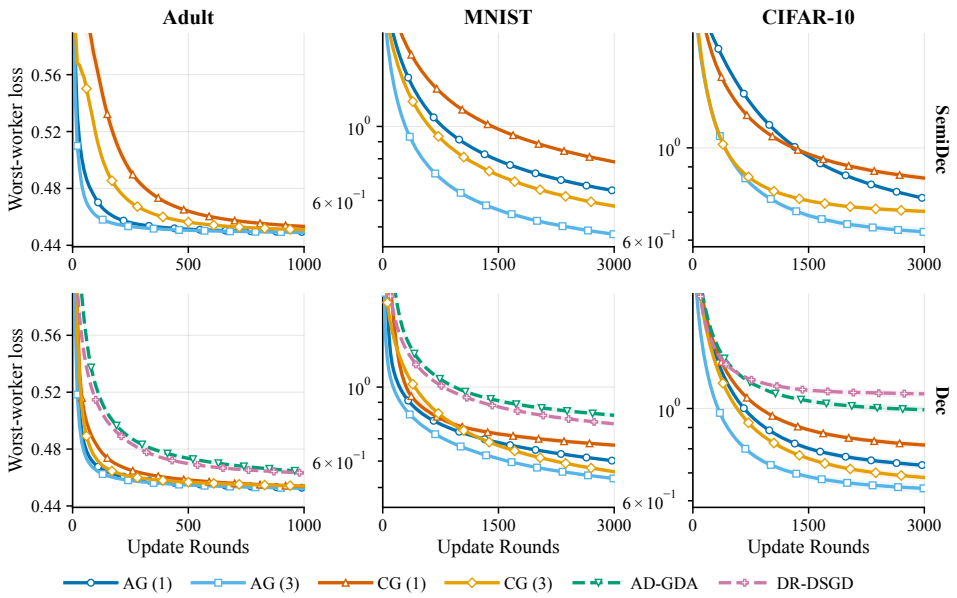


Figure 2: Worst-worker loss over update rounds on Adult, MNIST, and CIFAR-10. The top and bottom rows correspond to SemiDec and Dec, respectively. Parenthesized numbers denote the consensus rounds per update.

The resulting matrix is symmetric, doubly stochastic, and positive semidefinite for both $m = 6$ and $m = 10$, and therefore satisfies the matrix condition in our AG corollaries.

We use a linear logistic classifier on Adult and a linear 784-to-10 classifier on MNIST. On Adult, the trainable parameter vector is projected onto the Euclidean ball of radius 3 after each update. For CIFAR-10, we use a pretrained ResNet-20 as a fixed feature extractor. All backbone parameters and BatchNorm statistics are frozen, and only a randomly initialized 64-to-10 linear head is trained.

For the proposed methods, the primal update uses the original batch loss, while the dual feedback is clipped and normalized. Let $\ell_{\mathrm{ce}}(\mathbf{w}; z)$ denote the cross-entropy loss and define

$$\ell_C(\mathbf{w}; z) = \frac{\min\{\ell_{\mathrm{ce}}(\mathbf{w}; z), C\}}{C}, \qquad C = 5.$$

The dual update uses

$$\hat{s}_{t,i} = 1 - \ell_C(\mathbf{w}_t^{(i)}; z_t^{(i)}),$$

so that $\hat{s}_{t,i} \in [0, 1]$. The primal update and evaluation use the original unclipped loss.

Top-k compression retains the $\lceil 0.5d \rceil$ coordinates with largest absolute magnitude in each parameter tensor. We use a mini-batch size of 128 for Adult and MNIST and 32 for CIFAR-10. For AD-GDA, we use its chi-square dual regularization with $\alpha = 0.1$, 0.1, and 0.05 on Adult, MNIST, and CIFAR-10, respectively; for DR-DSGD, we set $\mu = 3$, 3, and 1, respectively.

B.2 Worst-Worker Loss

Figure 2 reports the loss curves corresponding to the accuracy results in Section 7. In SemiDec, worst-worker loss decreases for all variants. AG decreases more rapidly than CG on MNIST, while the $L = 3$ variants generally improve earlier, particularly on CIFAR-10. In Dec, our variants generally finish below AD-GDA and DR-DSGD, with the clearest separation on CIFAR-10 and smaller differences on Adult and MNIST.

C Further Discussion of Theoretical Guarantees

Let $|\mathcal{E}_{\mathrm{ns}}|$ denote the number of non-self undirected edges and let N count bidirectional neighbor exchanges. Over T updates, SemiDec uses $N_{\mathrm{SemiDec}} = |\mathcal{E}_{\mathrm{ns}}|TL$ exchanges, whereas Dec

uses $N_{\mathrm{Dec}} = |\mathcal{E}_{\mathrm{ns}}|T(L_w + L_q)$. For AG, $L, L_w, L_q = O((\log m)/\sqrt{\rho})$, giving $N_{\mathrm{AG}} = O(|\mathcal{E}_{\mathrm{ns}}|(\log m)^2/(\varepsilon^2\sqrt{\rho}))$ to reach expected error ε, up to a constant factor between SemiDec and Dec. For CG, $L, L_w, L_q = O((\log m)/(\delta\rho^2))$, yielding $N_{\mathrm{CG}} = O(|\mathcal{E}_{\mathrm{ns}}|(\log m)^2/(\varepsilon^2\delta\rho^2))$.

Theoretically, the need for multiple consensus rounds is inherited from the underlying D-OCO subproblems. In particular, Wan et al. (2025) show that using a single AG step per round leads to a worse regret bound, whereas their blocking-based method uses $O((\log m)/\sqrt{\rho})$ communication rounds between consecutive decision updates and achieves a near-optimal regret bound, matching the corresponding lower bound up to a logarithmic factor. Under the correspondence between one online block and one update in our offline framework, this gives exactly the same $O((\log m)/\sqrt{\rho})$ communication-round order as our AG instantiations. The CG instantiations similarly inherit the state-of-the-art network and compression dependence of the corresponding compressed D-OCO method (Yang et al., 2026).

From a communication perspective, SemiDec and Dec require the same order of peer-to-peer consensus rounds for the d-dimensional primal information. Dec additionally communicates m-dimensional dual information through worker-to-worker consensus, whereas SemiDec maintains the shared dual vector through scalar server–worker communication.

D Analysis

D.1 Useful Lemmas

We first state two lemmas for FTRL that will be used in our analysis. The first lemma controls the regret in terms of the norm of the gradients.

Lemma 3. *(Orabona, 2019, Corollary 7.9.) Let V be convex and let $\psi : V \to \mathbb{R}$ be a closed and μ-strongly convex function w. r. t. $\|\cdot\|$. Set the sequence of regularizers as $\psi_t(\boldsymbol{x}) = \frac{1}{\eta_{t-1}}(\psi(\boldsymbol{x}) - \min_{\boldsymbol{z}\in V}\psi(\boldsymbol{z}))$, where $\eta_0 \geq \eta_1 \geq \cdots \geq \eta_T > 0$ and $\boldsymbol{x}_t \in \operatorname{argmin}_{\boldsymbol{x}\in V}\psi_t(\boldsymbol{x}) + \sum_{i=1}^{t-1}\ell_i(\boldsymbol{x})$. Assume ℓ_t is convex and closed, and let $\boldsymbol{g}_t \in \partial\ell_t(\boldsymbol{x}_t)$. Then, for every $\boldsymbol{u} \in V$, FTRL guarantees*

$$\sum_{t=1}^{T}\ell_t(\boldsymbol{x}_t) - \sum_{t=1}^{T}\ell_t(\boldsymbol{u}) \leq \frac{\psi(\boldsymbol{u}) - \min_{\boldsymbol{x}\in V}\psi(\boldsymbol{x})}{\eta_{T-1}} + \frac{1}{2\mu}\sum_{t=1}^{T}\eta_{t-1}\|\boldsymbol{g}_t\|_*^2.$$

The second lemma is a local-norm instantiation of Lemma 7.14 by Orabona (2019), specialized to the simplex and the negative-entropy regularizer (see § 7.5 in Orabona (2019) for further details).

Lemma 4. *(Orabona, 2019, Lemma 7.14) Denote by $F_t(\boldsymbol{x}) = \psi_t(\boldsymbol{x}) + \sum_{i=1}^{t-1}\langle \boldsymbol{g}_i, \boldsymbol{x}\rangle$ with $\psi_t(\boldsymbol{x}) = \frac{1}{\eta_t}(\sum_{i=1}^{m} x_i \ln x_i - \min_{\boldsymbol{x}\in\Delta_m}\sum_{i=1}^{m} x_i \ln x_i)$, where $\eta_{t+1} \leq \eta_t, \forall t = 1, \ldots, T$. Set $\boldsymbol{x}_t \in \operatorname{argmin}_{\boldsymbol{x}\in\Delta_m} F_t(\boldsymbol{x})$. Assume $\boldsymbol{g}_t \in \mathbb{R}_+^m$ and $\eta_t > 0$ for all t. The following inequality holds for any $\boldsymbol{u} \in \Delta_m$,*

$$\sum_{t=1}^{T}\langle \boldsymbol{g}_t, \boldsymbol{x}_t - \boldsymbol{u}\rangle \leq \psi_T(\boldsymbol{u}) - \min_{\boldsymbol{x}\in\Delta_m}\psi_1(\boldsymbol{x}) + \frac{1}{2}\sum_{t=1}^{T}\eta_t\sum_{i=1}^{m} g_{t,i}^2 x_{t,i}.$$

Lemma 5. *(Duchi et al., 2012, Lemma 5) Let $\mathcal{X}$ be nonempty, closed, and convex, let ψ be closed and 1-strongly convex with respect to $\|\cdot\|$, and let $\alpha > 0$. Define $\Pi_{\mathcal{X}}^{\psi}(z, \alpha) = \operatorname{argmin}_{x\in\mathcal{X}}\left\{\langle z, x\rangle + \frac{1}{\alpha}\psi(x)\right\}$. For arbitrary $u, v \in \mathbb{R}^d$, we have*

$$\left\|\Pi_{\mathcal{X}}^{\psi}(u, \alpha) - \Pi_{\mathcal{X}}^{\psi}(v, \alpha)\right\| \leq \alpha\|u - v\|_*. \tag{17}$$

Lemma 6. *(Cesa-Bianchi & Lugosi, 2006, Lemma A.7) Let $V_1, V_2, \ldots$ be a martingale difference sequence with respect to some sequence $X_1, X_2, \ldots$ such that $V_i \in [A_i, A_i + c_i]$ for some random variable A_i, measurable with respect to $X_1, \ldots, X_{i-1}$ and a positive constant c_i. If $S_n = \sum_{i=1}^{n} V_i$, then for any $t > 0$,*

$$\Pr[S_n > t] \leq \exp\left(-\frac{2t^2}{\sum_{i=1}^{n} c_i^2}\right).$$

Lemma 7. *(Bubeck & Cesa-Bianchi, 2012, § 3.2) For any real-valued random variable X, $\mathbb{E}[X] \le \int_0^1 \frac{1}{\delta} \Pr\left[X > \ln \frac{1}{\delta}\right] \, \mathrm{d}\delta$.*

Based on Lemmas 1 and 2, we obtain the following bound on the communication error.

Lemma 8 (Accelerated Gossip Error Bound)**.** *Let $P \succeq 0$ satisfy Assumption 3, let $\rho = 1 - \lambda_\star(P)$, and initialize $\mathbf{X}_0^{(i)} = \mathbf{0}$ for all $i \in [m]$. Consider the Accelerated Gossip operator*

$$\{\mathbf{X}_t^{(i)}\}_{i\in[m]} = \mathrm{AG}\big(\{\mathbf{X}_{t-1}^{(i)} + \mathbf{g}_t^{(i)}\}_{i\in[m]}, L, P\big), \tag{18}$$

i.e., run Algorithm 3 for L communication rounds with $\theta = \frac{1}{1+\sqrt{1-\lambda_\star^2(P)}}$. Let $\mathbf{X}_t = [\mathbf{X}_t^{(1)}, \dots, \mathbf{X}_t^{(m)}]$ and $\bar{\mathbf{X}}_t = \frac{1}{m}\mathbf{X}_t\mathbf{1}$ be the column-wise average and denote $c = 1 - \left(1 - \frac{1}{\sqrt{2}}\right)\sqrt{\rho}$. Fix a constant $k > 2$ and choose the number of gossip rounds as

$$L = \left\lceil \frac{\ln(14k)}{2\left(1 - \frac{1}{\sqrt{2}}\right)\sqrt{\rho}} \right\rceil,$$

so that $14c^{2L} \le \frac{1}{k}$. Assume additionally that $\sum_{i=1}^m \|\mathbf{g}_t^{(i)}\|_2^2 \le Q$ for all $t \in [T]$. Then the per-round communication error satisfies

$$\|\mathbf{X}_t - \bar{\mathbf{X}}_t\mathbf{1}^\top\|_F \le \frac{\sqrt{Q}}{\sqrt{k}-1}, \qquad \forall t \in [T]. \tag{19}$$

Moreover, the average is preserved in the sense that $\bar{\mathbf{X}}_t = \bar{\mathbf{X}}_{t-1} + \bar{\mathbf{g}}_t$, where $\mathbf{g}_t = [\mathbf{g}_t^{(1)}, \dots, \mathbf{g}_t^{(m)}]$ and $\bar{\mathbf{g}}_t = \frac{1}{m}\mathbf{g}_t\mathbf{1}$.

Lemma 9 (Choco-gossip Error Bound)**.** *Let $\mathcal{C}$ satisfy (15) on $\mathbb{R}^d$ with fresh internal randomness at each call, and let P satisfy Assumption 3. Initialize $\mathbf{X}_0^{(i)} = \hat{\mathbf{X}}_0^{(i)} = \mathbf{0}$ for all $i \in [m]$. Consider the Choco-gossip operator*

$$\{\mathbf{X}_t^{(i)}, \hat{\mathbf{X}}_t^{(i)}\}_{i\in[m]} = \mathrm{CG}(\{\mathbf{X}_{t-1}^{(i)} + \mathbf{g}_t^{(i)}, \hat{\mathbf{X}}_{t-1}^{(i)}\}_{i\in[m]}, \gamma, L, \mathcal{C}, P) \tag{20}$$

and let the gossip stepsize $\gamma = \frac{\delta\rho}{16\rho+\rho^2+4\beta^2+2\rho\beta^2-8\rho\delta}$ and the number of gossip iterations $L = \left\lceil \frac{2\ln k}{\gamma\rho} \right\rceil$ for a constant $k > 2$. Denote by $\mathbf{X}_t = [\mathbf{X}_t^{(1)}, \dots, \mathbf{X}_t^{(m)}]$ and let $\bar{\mathbf{X}}_t = \frac{1}{m}\mathbf{X}_t\mathbf{1}$ be the column-wise average. If $\mathbb{E}[\sum_{i=1}^m \|\mathbf{g}_t^{(i)}\|_2^2] \le Q$ for all $t \in [T]$, with the inputs formed before the current consensus call, then the per-round communication error satisfies

$$\mathbb{E}\left[\|\mathbf{X}_t - \bar{\mathbf{X}}_t\mathbf{1}^\top\|_F^2 + \|\mathbf{X}_t - \hat{\mathbf{X}}_t\|_F^2\right] \le \frac{4Q}{k-2}, \qquad \forall t \in [T]. \tag{21}$$

D.2 The Proof of Theorem 1

Recall the time-averaged local output $\bar{\mathbf{w}}_T^{(i)} = \frac{1}{T}\sum_{t=1}^T \mathbf{w}_t^{(i)}$, and define the time-averaged shared dual vector $\bar{\mathbf{q}}_T = \frac{1}{T}\sum_{t=1}^T \mathbf{q}_t$. For any worker $i \in [m]$, we decompose the optimization error at $\bar{\mathbf{w}}_T^{(i)}$ as

$$\begin{aligned}
&\varepsilon_\phi(\bar{\mathbf{w}}_T^{(i)}) = \max_{j\in[m]} R_j(\bar{\mathbf{w}}_T^{(i)}) - \min_{\mathbf{w}\in\mathcal{W}} \max_{j\in[m]} R_j(\mathbf{w}) = \max_{j\in[m]} R_j(\bar{\mathbf{w}}_T^{(i)}) - \max_{j\in[m]} R_j(\mathbf{w}^*) \\
=& \max_{\mathbf{q}\in\Delta_m} \phi(\bar{\mathbf{w}}_T^{(i)}, \mathbf{q}) - \max_{\mathbf{q}\in\Delta_m} \phi(\mathbf{w}^*, \mathbf{q}) \le \max_{\mathbf{q}\in\Delta_m} \phi(\bar{\mathbf{w}}_T^{(i)}, \mathbf{q}) - \phi\left(\mathbf{w}^*, \bar{\mathbf{q}}_T\right) \\
\le& \frac{1}{T}\max_{\mathbf{q}\in\Delta_m} \sum_{t=1}^T \phi(\mathbf{w}_t^{(i)}, \mathbf{q}) - \frac{1}{T}\sum_{t=1}^T \phi(\mathbf{w}^*, \mathbf{q}_t) \\
\le& \frac{1}{T}\left(\max_{\mathbf{q}\in\Delta_m} \sum_{t=1}^T \phi(\mathbf{w}_t^{(i)}, \mathbf{q}) - \sum_{t=1}^T \phi(\mathbf{w}_t^{(i)}, \mathbf{q}_t)\right) + \frac{1}{T}\left(\sum_{t=1}^T \phi(\mathbf{w}_t^{(i)}, \mathbf{q}_t) - \sum_{t=1}^T \phi(\mathbf{w}^*, \mathbf{q}_t)\right).
\end{aligned} \tag{22}$$

As introduced in Section 4, we rewrite

$$\phi(\mathbf{w}_t^{(i)}, \mathbf{q}) = \sum_{j=1}^{m} q_j R_j(\mathbf{w}_t^{(i)}) = \sum_{j=1}^{m} q_j R_j(\mathbf{w}_t^{(j)}) + \sum_{j=1}^{m} q_j (R_j(\mathbf{w}_t^{(i)}) - R_j(\mathbf{w}_t^{(j)})).$$

We denote $j(i) = \arg\max_{j\in[m]} \sum_{t=1}^{T} R_j(\mathbf{w}_t^{(i)})$ and obtain

$$\begin{aligned}
&\max_{\mathbf{q}\in\Delta_m} \sum_{t=1}^{T} \phi(\mathbf{w}_t^{(i)}, \mathbf{q}) - \sum_{t=1}^{T} \phi(\mathbf{w}_t^{(i)}, \mathbf{q}_t) \\
=&\max_{\mathbf{q}\in\Delta_m} \sum_{t=1}^{T}\sum_{j=1}^{m} q_j R_j(\mathbf{w}_t^{(i)}) - \sum_{t=1}^{T}\sum_{j=1}^{m} q_{t,j} R_j(\mathbf{w}_t^{(i)}) \\
=&\sum_{t=1}^{T} R_{j(i)}(\mathbf{w}_t^{(i)}) - \sum_{t=1}^{T}\sum_{j=1}^{m} q_{t,j} R_j(\mathbf{w}_t^{(i)}) \\
=&\sum_{t=1}^{T} \left(R_{j(i)}(\mathbf{w}_t^{(i)}) - R_{j(i)}(\mathbf{w}_t^{(j(i))})\right) + \sum_{t=1}^{T} \left(R_{j(i)}(\mathbf{w}_t^{(j(i))}) - \sum_{j=1}^{m} q_{t,j} R_j(\mathbf{w}_t^{(j)})\right) \\
&+ \sum_{t=1}^{T}\sum_{j=1}^{m} q_{t,j} \left(R_j(\mathbf{w}_t^{(j)}) - R_j(\mathbf{w}_t^{(i)})\right) \\
\overset{(2)}{\leq}& G \sum_{t=1}^{T} \left(\|\mathbf{w}_t^{(i)} - \mathbf{w}_t^{(j(i))}\|_2 + \sum_{j=1}^{m} q_{t,j} \|\mathbf{w}_t^{(i)} - \mathbf{w}_t^{(j)}\|_2\right) \\
&+ \sum_{t=1}^{T} \left(R_{j(i)}(\mathbf{w}_t^{(j(i))}) - \sum_{j=1}^{m} q_{t,j} R_j(\mathbf{w}_t^{(j)})\right).
\end{aligned} \tag{23}$$

where $s_{t,j}^{(i)} = 1 - R_j(\mathbf{w}_t^{(i)}) = 1 - \mathbb{E}_{\mathbf{z}\sim\mathcal{P}_j}[\ell(\mathbf{w}_t^{(i)}; \mathbf{z})] \overset{(2)}{\in} [0, 1]$. Substituting (23) into (22), we obtain

$$\begin{aligned}
\varepsilon_\phi(\bar{\mathbf{w}}_T^{(i)}) \leq& \frac{1}{T} \underbrace{\max_{k\in[m]} \sum_{t=1}^{T} \left(\sum_{j=1}^{m} q_{t,j} s_{t,j}^{(j)} - s_{t,k}^{(k)}\right)}_{\mathrm{Reg}_{q,T}} + \underbrace{\frac{1}{T} \sum_{t=1}^{T}\sum_{j=1}^{m} \left(q_{t,j} R_j(\mathbf{w}_t^{(i)}) - q_{t,j} R_j(\mathbf{w}^*)\right)}_{\mathrm{Reg}_{w,T}^{(i)}(\mathbf{w}^*)} \\
&+ \underbrace{\frac{1}{T} G \sum_{t=1}^{T} \left(\|\mathbf{w}_t^{(i)} - \mathbf{w}_t^{(j(i))}\|_2 + \sum_{j=1}^{m} q_{t,j} \|\mathbf{w}_t^{(i)} - \mathbf{w}_t^{(j)}\|_2\right)}_{\texttt{term(dis-semi)}}
\end{aligned} \tag{24}$$

D.2.1 REGRET $\mathrm{Reg}_{w,T}^{(i)}$

For later reuse of the analysis, in SemiDec we write $q_{t,i}^{(i)} = q_{t,i}$, the weight sent by the server to worker i, for all $i \in [m]$ and $t \in [T]$.

Let $\bar{\mathbf{g}}_{w,t} = \frac{1}{m}\sum_{i=1}^{m} \mathbf{g}_{w,t}^{(i)}$. We introduce a virtual decision sequence $\{\tilde{\mathbf{w}}_t\}_{t\geq 1}$ initialized by $\tilde{\mathbf{w}}_1 = \arg\min_{\mathbf{w}\in\mathcal{W}} \nu_w(\mathbf{w})$ and updated by

$$\tilde{\mathbf{w}}_{t+1} = \arg\min_{\mathbf{w}\in\mathcal{W}} \left\{ \left\langle \eta_w \sum_{\tau=1}^{t} \bar{\mathbf{g}}_{w,\tau}, \mathbf{w} \right\rangle + \nu_w(\mathbf{w}) \right\} \quad \forall t \in [T]. \tag{25}$$

We characterize the deviation between each worker's local decision and the virtual decision as $\varepsilon_{w,t} = \max_{i\in[m]} \left\|\mathbf{w}_t^{(i)} - \tilde{\mathbf{w}}_t\right\|_2$.

For a fixed worker $i \in [m]$, any distribution index $j \in [m]$ and any fixed point $\mathbf{u} \in \mathcal{W}$, we have

$$\begin{aligned}
&R_j(\mathbf{w}_t^{(i)}) - R_j(\mathbf{u}) = R_j(\mathbf{w}_t^{(j)}) - R_j(\mathbf{u}) + R_j(\mathbf{w}_t^{(i)}) - R_j(\mathbf{w}_t^{(j)}) \\
=&R_j(\mathbf{w}_t^{(j)}) - R_j(\mathbf{u}) + \mathbb{E}_{\mathbf{z}\sim\mathcal{P}_j}[\ell(\mathbf{w}_t^{(i)};\mathbf{z})] - \mathbb{E}_{\mathbf{z}\sim\mathcal{P}_j}[\ell(\mathbf{w}_t^{(j)};\mathbf{z})] \\
\overset{(2)}{\leq}&R_j(\mathbf{w}_t^{(j)}) - R_j(\mathbf{u}) + G\|\mathbf{w}_t^{(i)} - \mathbf{w}_t^{(j)}\| \leq \langle \nabla R_j(\mathbf{w}_t^{(j)}), \mathbf{w}_t^{(j)} - \mathbf{u}\rangle + G\|\mathbf{w}_t^{(i)} - \mathbf{w}_t^{(j)}\| \\
\leq&\langle \nabla \ell(\mathbf{w}_t^{(j)};\mathbf{z}_t^{(j)}), \tilde{\mathbf{w}}_t - \mathbf{u}\rangle + \langle \nabla \ell(\mathbf{w}_t^{(j)};\mathbf{z}_t^{(j)}), \mathbf{w}_t^{(j)} - \tilde{\mathbf{w}}_t\rangle \\
&+ \langle \nabla R_j(\mathbf{w}_t^{(j)}) - \nabla \ell(\mathbf{w}_t^{(j)};\mathbf{z}_t^{(j)}), \mathbf{w}_t^{(j)} - \mathbf{u}\rangle + G\|\mathbf{w}_t^{(i)} - \mathbf{w}_t^{(j)}\| \\
\leq&\langle \nabla \ell(\mathbf{w}_t^{(j)};\mathbf{z}_t^{(j)}), \tilde{\mathbf{w}}_t - \mathbf{u}\rangle + \langle \nabla R_j(\mathbf{w}_t^{(j)}) - \nabla \ell(\mathbf{w}_t^{(j)};\mathbf{z}_t^{(j)}), \mathbf{w}_t^{(j)} - \mathbf{u}\rangle \\
&+ G\left(\|\tilde{\mathbf{w}}_t - \mathbf{w}_t^{(j)}\| + \|\mathbf{w}_t^{(i)} - \mathbf{w}_t^{(j)}\|\right) \\
\leq&\langle \nabla \ell(\mathbf{w}_t^{(j)};\mathbf{z}_t^{(j)}), \tilde{\mathbf{w}}_t - \mathbf{u}\rangle + \langle \nabla R_j(\mathbf{w}_t^{(j)}) - \nabla \ell(\mathbf{w}_t^{(j)};\mathbf{z}_t^{(j)}), \mathbf{w}_t^{(j)} - \mathbf{u}\rangle + 3G\varepsilon_{w,t}.
\end{aligned} \tag{26}$$

Multiplying both sides of (26) by $q_{t,j}^{(j)}$, using $q_{t,j}^{(j)} \nabla \ell(\mathbf{w}_t^{(j)};\mathbf{z}_t^{(j)}) = \mathbf{g}_{w,t}^{(j)}$, summing over $j \in [m]$ and $t \in [T]$, and taking expectations, we obtain

$$\begin{aligned}
\mathbb{E}\left[\mathrm{Reg}_{w,T}^{(i)}(\mathbf{u})\right] =&\mathbb{E}\left[\sum_{t=1}^{T}\sum_{j=1}^{m} q_{t,j}^{(j)} R_j(\mathbf{w}_t^{(i)}) - \sum_{t=1}^{T}\sum_{j=1}^{m} q_{t,j}^{(j)} R_j(\mathbf{u})\right] \\
\leq& m\sum_{t=1}^{T}\mathbb{E}\left[\langle \bar{\mathbf{g}}_{w,t}, \tilde{\mathbf{w}}_t - \mathbf{u}\rangle\right] + 3G\sum_{t=1}^{T}\mathbb{E}[\varepsilon_{w,t}].
\end{aligned} \tag{27}$$

For the first term on the right-hand side of (27), using Lemma 3 with the update rule of $\tilde{\mathbf{w}}_{t+1}$ in (25), we have

$$\begin{aligned}
\sum_{t=1}^{T}\langle \bar{\mathbf{g}}_{w,t}, \tilde{\mathbf{w}}_t - \mathbf{u}\rangle \leq& \frac{\nu_w(\mathbf{u}) - \min_{\mathbf{w}\in\mathcal{W}} \nu_w(\mathbf{w})}{\eta_w} + \frac{\eta_w}{2}\sum_{t=1}^{T}\|\bar{\mathbf{g}}_{w,t}\|_2^2 \\
\leq& \frac{D^2}{\eta_w} + \frac{\eta_w}{2}\sum_{t=1}^{T}\left\|\frac{1}{m}\sum_{i=1}^{m} q_{t,i}^{(i)} \cdot \nabla \ell(\mathbf{w}_t^{(i)};\mathbf{z}_t^{(i)})\right\|_2^2 \\
\overset{(2)}{\leq}& \frac{D^2}{\eta_w} + \frac{\eta_w T G^2}{2m^2},
\end{aligned} \tag{28}$$

where the second inequality follows from Assumption 1, and the last inequality uses the fact that, in SemiDec, the centralized update of $\mathbf{q}_t$ guarantees that $\sum_{i=1}^{m} q_{t,i}^{(i)} = \sum_{i=1}^{m} q_{t,i} = 1$.

Combining (27) and (28) and setting $\eta_w = \frac{\sqrt{2}mD}{\sqrt{T}G}$, we obtain

$$\begin{aligned}
\mathbb{E}\left[\mathrm{Reg}_{w,T}^{(i)}(\mathbf{u})\right] =&\mathbb{E}\left[\sum_{t=1}^{T}\sum_{j=1}^{m} q_{t,j}^{(j)} R_j(\mathbf{w}_t^{(i)}) - \sum_{t=1}^{T}\sum_{j=1}^{m} q_{t,j}^{(j)} R_j(\mathbf{u})\right] \\
\leq& \frac{mD^2}{\eta_w} + \frac{\eta_w T G^2}{2m} + 3G\sum_{t=1}^{T}\mathbb{E}[\varepsilon_{w,t}] = \sqrt{2T}DG + 3G\sum_{t=1}^{T}\mathbb{E}[\varepsilon_{w,t}],
\end{aligned} \tag{29}$$

D.2.2 REGRET $\mathrm{Reg}_{q,T}$

In this part, we derive a bound on the regret $\mathrm{Reg}_{q,T}$ in the semi-decentralized setting. Recall that the server solves a *prediction with expert advice* (PEA) problem and adopts the Hedge algorithm (Freund & Schapire, 1997). Following Bai et al. (2025, Lemma 1), let $\hat{s}_{t,j}^{(i)} = 1 - \ell(\mathbf{w}_t^{(i)};\mathbf{z}_t^{(i)})$, $\tilde{\mathbf{s}}_t = [\hat{s}_{t,1}^{(1)};\ldots;\hat{s}_{t,m}^{(m)}]$, and $\mathbf{s}_t = (s_{t,1}^{(1)},\ldots,s_{t,m}^{(m)})^\top$. We decompose $\mathrm{Reg}_{q,T}$ as

$$\mathrm{Reg}_{q,T} \leq \underbrace{\sum_{t=1}^{T}\langle \mathbf{q}_t, \mathbf{s}_t - \tilde{\mathbf{s}}_t\rangle}_{\mathtt{term}(\mathtt{q}_1)} + \underbrace{\sum_{t=1}^{T}\langle \mathbf{q}_t, \tilde{\mathbf{s}}_t\rangle - \min_{i\in[m]}\sum_{t=1}^{T}\hat{s}_{t,i}^{(i)}}_{\mathtt{term}(\mathtt{q}_2)} + \underbrace{\sum_{t=1}^{T}\hat{s}_{t,k^*}^{(k^*)} - \sum_{t=1}^{T}s_{t,k^*}^{(k^*)}}_{\mathtt{term}(\mathtt{q}_3)}. \tag{30}$$

where $k^* \in \arg\min_{k\in[m]} \sum_{t=1}^{T} s_{t,k}^{(k)}$.

Let $\mathcal{F}_{t-1}$ denote the history before round t. For $\texttt{term}(\mathtt{q}_1)$, since $\mathbf{q}_t$ is $\mathcal{F}_{t-1}$-measurable and $\mathbb{E}[\tilde{\mathbf{s}}_t \mid \mathcal{F}_{t-1}] = \mathbf{s}_t$, we have $\mathbb{E}[\texttt{term}(\mathtt{q}_1)] = 0$.

For $\texttt{term}(\mathtt{q}_2)$, by Lemma 4 and setting $\eta_q = \sqrt{\frac{2\ln m}{T}}$, we can derive for each $k \in [m]$,

$$\begin{aligned}\sum_{t=1}^{T}\left(\langle \mathbf{q}_t, \tilde{\mathbf{s}}_t\rangle - \hat{s}_{t,k}^{(k)}\right) &\le \frac{\ln m}{\eta_q} + \frac{\eta_q}{2}\sum_{t=1}^{T}\sum_{k=1}^{m} q_{t,k}(\hat{s}_{t,k}^{(k)})^2 \\ &\le \frac{\ln m}{\eta_q} + \frac{T\eta_q}{2} = \sqrt{2T\ln m}.\end{aligned} \tag{31}$$

Thus, we obtain $\texttt{term}(\mathtt{q}_2) \le \sqrt{2T\ln m}$.

For $\texttt{term}(\mathtt{q}_3)$, define $X_{t,k} = \hat{s}_{t,k}^{(k)} - s_{t,k}^{(k)}$. Then $\mathbb{E}[X_{t,k} \mid \mathcal{F}_{t-1}] = 0$. Moreover, by Assumption 2, $\hat{s}_{t,k}^{(k)} \in [0,1]$, we obtain $|X_{t,k}| \le 1$. By Lemma 6 and a union bound over $i \in [m]$, with probability at least $1-\delta$,

$$\texttt{term}(\mathbf{q}_3) \le \max_{k\in[m]} \sum_{t=1}^{T} X_{t,k} \le \sqrt{2T\ln\frac{m}{\delta}}.$$

We now convert this high-probability bound into an expectation bound. Let $Z = \frac{(\texttt{term}(\mathbf{q}_3))_+^2}{2T} - \ln m$. The above inequality implies that, for any $\delta \in (0,1)$,

$$\Pr\left[Z > \ln\frac{1}{\delta}\right] = \Pr\left[\texttt{term}(\mathbf{q}_3) > \sqrt{2T\ln\frac{m}{\delta}}\right] \le \delta.$$

Applying Lemma 7 to Z, we get

$$\mathbb{E}[Z] \le \int_0^1 \frac{1}{\delta}\Pr\left[Z > \ln\frac{1}{\delta}\right]\mathrm{d}\delta \le 1.$$

Therefore,

$$\mathbb{E}[\texttt{term}(\mathbf{q}_3)] \le \mathbb{E}[(\texttt{term}(\mathbf{q}_3))_+] \le \sqrt{\mathbb{E}[(\texttt{term}(\mathbf{q}_3))_+^2]} \le \sqrt{2T(\ln(m)+1)}. \tag{32}$$

Taking expectations on (30), combining it with the three bounds and using $\ln m \ge \ln 2 > 1/2$ and $\sqrt{2}+\sqrt{6} < 4$, we obtain

$$\mathbb{E}\left[\mathrm{Reg}_{q,T}\right] \le 4\sqrt{T\ln m}. \tag{33}$$

D.2.3 Discrepancy error term and results

To control the discrepancy term, recalling that $\varepsilon_{w,t} = \max_{i\in[m]} \|\mathbf{w}_t^{(i)} - \tilde{\mathbf{w}}_t\|_2$, we obtain

$$\begin{aligned}\texttt{term}(\texttt{dis-semi}) &= G\sum_{t=1}^{T}\left(\|\mathbf{w}_t^{(i)} - \mathbf{w}_t^{(j(i))}\|_2 + \sum_{j=1}^{m} q_{t,j}\|\mathbf{w}_t^{(i)} - \mathbf{w}_t^{(j)}\|_2\right) \\ &\le G\sum_{t=1}^{T}\left(2\varepsilon_{w,t} + \sum_{j=1}^{m} q_{t,j}\cdot 2\varepsilon_{w,t}\right) = 4G\sum_{t=1}^{T}\varepsilon_{w,t}.\end{aligned} \tag{34}$$

Substituting (29), (33), (34) into (24), setting $\eta_w = \frac{\sqrt{2}mD}{\sqrt{T}G}$, $\eta_q = \sqrt{\frac{2\ln m}{T}}$, and taking expectations yields

$$\begin{aligned}\mathbb{E}\left[\varepsilon_\phi(\bar{\mathbf{w}}_T^{(i)})\right] &\le \frac{1}{T}\mathbb{E}[\mathrm{Reg}_{q,T}] + \frac{1}{T}\mathbb{E}\left[\mathrm{Reg}_{w,T}^{(i)}(\mathbf{w}^*)\right] + \frac{1}{T}\mathbb{E}\left[\texttt{term}(\texttt{dis-semi})\right] \\ &\le 4\sqrt{\frac{\ln m}{T}} + \sqrt{\frac{2}{T}}DG + \frac{3G}{T}\sum_{t=1}^{T}\mathbb{E}[\varepsilon_{w,t}] + \frac{4G}{T}\sum_{t=1}^{T}\mathbb{E}[\varepsilon_{w,t}] \\ &= 4\sqrt{\frac{\ln m}{T}} + \sqrt{\frac{2}{T}}DG + \frac{7G}{T}\sum_{t=1}^{T}\mathbb{E}[\varepsilon_{w,t}].\end{aligned}$$

For any $i \in [m]$ and $t > 1$, by Lemma 5 and the update rule (9) and (25), we have

$$\|\mathbf{w}_t^{(i)} - \tilde{\mathbf{w}}_t\|_2 \overset{(17)}{\leq} \eta_w \left\| \sum_{\tau=1}^{t-1} \bar{\mathbf{g}}_{w,\tau} - \mathbf{F}_{t-1}^{(i)} \right\|_2 \leq \eta_w \left(\left\| \mathbf{F}_{t-1}^{(i)} - \bar{\mathbf{F}}_{t-1} \right\|_2 + \left\| \bar{\mathbf{F}}_{t-1} - \sum_{\tau=1}^{t-1} \bar{\mathbf{g}}_{w,\tau} \right\|_2 \right), \tag{35}$$

If the consensus routine guarantees $\bar{\mathbf{F}}_{t-1} = \sum_{\tau=1}^{t-1} \bar{\mathbf{g}}_{w,\tau}$ and $\mathbb{E}[\max_{r\in[m]} \|\mathbf{F}_{t-1}^{(r)} - \bar{\mathbf{F}}_{t-1}\|_2] \leq \varepsilon_c$ for all $t \in [T]$, then for any worker $i \in [m]$ we obtain

$$\begin{aligned} \mathbb{E}\left[\varepsilon_\phi(\bar{\mathbf{w}}_T^{(i)})\right] \leq & 4\sqrt{\frac{\ln m}{T}} + \sqrt{\frac{2}{T}} DG + \frac{7\sqrt{2} m D}{T\sqrt{T}} \sum_{t=1}^{T} \mathbb{E}\left[\max_{r\in[m]} \left\| \mathbf{F}_{t-1}^{(r)} - \bar{\mathbf{F}}_{t-1} \right\|_2 \right] \\ \leq & 4\sqrt{\frac{\ln m}{T}} + \sqrt{\frac{2}{T}} DG + \frac{7\sqrt{2} m D}{\sqrt{T}} \varepsilon_c \end{aligned} \tag{36}$$

D.3 The Proof of Corollary 1

When adopting the accelerated gossip update (18), we set $\mathbf{F}_t^{(i)} = \mathbf{X}_t^{(i)}$ and $\bar{\mathbf{F}}_t = \bar{\mathbf{X}}_t = \frac{1}{m}\mathbf{X}_t\mathbf{1}$. By Lemma 8, the average is preserved, i.e., $\bar{\mathbf{F}}_t = \bar{\mathbf{F}}_{t-1} + \bar{\mathbf{g}}_t$, where $\bar{\mathbf{g}}_t = \frac{1}{m}\sum_{i=1}^m \mathbf{g}_t^{(i)}$. Since $\bar{\mathbf{F}}_0 = \mathbf{0}$, we obtain by induction that $\bar{\mathbf{F}}_{t-1} = \sum_{\tau=1}^{t-1} \bar{\mathbf{g}}_{w,\tau}$, and hence the second term in (35) vanishes. Let $a_m = \sqrt{(m-1)/m}$. For centered columns $\sum_i v_i = 0$, Cauchy–Schwarz gives $\|v_i\|_2^2 \leq (m-1)\sum_{j\neq i} \|v_j\|_2^2$, and thus $\max_i \|v_i\|_2 \leq a_m(\sum_i \|v_i\|_2^2)^{1/2}$. Applying this inequality and the per-round communication error bound (19) in Lemma 8, we obtain

$$\|\mathbf{F}_{t-1}^{(i)} - \bar{\mathbf{F}}_{t-1}\|_2 \leq \frac{a_m G}{\sqrt{k}-1}, \qquad \forall t \in [T], \tag{37}$$

since $\sum_{i=1}^m \|\mathbf{g}_t^{(i)}\|_2^2 = \sum_{i=1}^m \|q_{t,i} \cdot \nabla\ell(\mathbf{w}_t^{(i)}; \mathbf{z}_t^{(i)})\|_2^2 \leq G^2$ for all $t \in [T]$. Combining (37) with (36), and setting $k = (\frac{m}{\sqrt{\ln m}} + 1)^2$, we obtain

$$\begin{aligned} \mathbb{E}\left[\varepsilon_\phi(\bar{\mathbf{w}}_T^{(i)})\right] &\leq 4\sqrt{\frac{\ln m}{T}} + \sqrt{\frac{2}{T}} DG + 7\sqrt{2} a_m DG \sqrt{\frac{\ln m}{T}} \\ &\leq (8\sqrt{2}DG + 4)\sqrt{\frac{\ln m}{T}} = O\left(\sqrt{\frac{\log m}{T}}\right). \end{aligned} \tag{38}$$

The last inequality uses $(\ln m)^{-1/2} + 7a_m \leq 8$: for $m \geq 3$, use $\ln m > 1$ and $a_m \leq 1$; for $m = 2$, use $\ln 2 > 1/2$ and $a_2 = 1/\sqrt{2}$.

D.4 The Proof of Corollary 2

We first show that the second term in (35) vanishes, i.e., $\bar{\mathbf{F}}_t - \sum_{\tau=1}^t \bar{\mathbf{g}}_{w,\tau} = \frac{1}{m}\sum_{i=1}^m \mathbf{F}_t^{(i)} - \frac{1}{m}\sum_{\tau=1}^t \sum_{i=1}^m \mathbf{g}_{w,\tau}^{(i)} = \mathbf{0}_d$. This follows from the average-preserving property of Algorithm 4. In particular, for the inner gossip iterates, we have

$$\begin{aligned} \frac{1}{m}\sum_{i=1}^m \mathbf{x}_{L+1}^{(i)} &= \frac{1}{m}\sum_{i=1}^m \left(\mathbf{x}_L^{(i)} + \gamma \sum_{j\in\mathcal{N}_i} P_{i,j}\big(\hat{\mathbf{x}}_{L+1}^{(j)} - \hat{\mathbf{x}}_{L+1}^{(i)}\big) \right) \\ &= \frac{1}{m}\sum_{i=1}^m \mathbf{x}_L^{(i)} = \cdots = \frac{1}{m}\sum_{i=1}^m \mathbf{x}_1^{(i)}, \end{aligned} \tag{39}$$

where the second equality uses Assumption 3. Applying this property to the outer update of $\mathbf{F}_t^{(i)}$ yields

$$\bar{\mathbf{F}}_t = \frac{1}{m}\sum_{i=1}^m \mathbf{F}_t^{(i)} = \frac{1}{m}\sum_{i=1}^m \left(\mathbf{F}_{t-1}^{(i)} + \mathbf{g}_{w,t}^{(i)} \right) = \bar{\mathbf{F}}_{t-1} + \bar{\mathbf{g}}_{w,t} = \sum_{\tau=1}^t \bar{\mathbf{g}}_{w,\tau}, \tag{40}$$

where the last equality uses $\bar{\mathbf{F}}_0 = \mathbf{0}_d$. Consequently, we obtain $\bar{\mathbf{F}}_{t-1} = \sum_{\tau=1}^{t-1} \bar{\mathbf{g}}_{w,\tau}$. Applying the per-round communication error bound (21) in Lemma 9 and Jensen's inequality yields for all $t \in [T]$

$$\mathbb{E}\left[\|\mathbf{F}_{t-1} - \bar{\mathbf{F}}_{t-1}\mathbf{1}^\top\|_F\right] \le \sqrt{\mathbb{E}\left[\|\mathbf{F}_{t-1} - \bar{\mathbf{F}}_{t-1}\mathbf{1}^\top\|_F^2\right]} \le \sqrt{\frac{4Q}{k-2}} = 2\frac{G}{\sqrt{k-2}}. \tag{41}$$

since $\sum_{i=1}^m \|\mathbf{g}_t^{(i)}\|_2^2 = \sum_{i=1}^m \|q_{t,i}\nabla\ell(\mathbf{w}_t^{(i)}; \mathbf{z}_t^{(i)})\|_2^2 \le G^2$ for all $t \in [T]$.

Using the centered-column inequality from Appendix D.3, with $a_m = \sqrt{(m-1)/m}$, we may take $\varepsilon_c = 2a_m G/\sqrt{k-2}$. Combining (41) with (36) and setting $k = \frac{m^2}{\ln m} + 2$, we obtain

$$\begin{aligned}
&\mathbb{E}\left[\varepsilon_\phi(\bar{\mathbf{w}}_T^{(i)})\right] \\
\le& 4\sqrt{\frac{\ln m}{T}} + \sqrt{\frac{2}{T}}DG + 14\sqrt{2}a_m DG\sqrt{\frac{\ln m}{T}} \le (15\sqrt{2}DG + 4)\sqrt{\frac{\ln m}{T}} = O\left(\sqrt{\frac{\log m}{T}}\right)
\end{aligned} \tag{42}$$

The last inequality uses $(\ln m)^{-1/2} + 14a_m \le 15$, which follows from $\ln m > 1$ for $m \ge 3$, and from $\ln 2 > 1/2$ and $a_2 = 1/\sqrt{2}$ for $m = 2$.

D.5 The Proof of Theorem 2

Let $\bar{\mathbf{q}}_T^{(i)} = T^{-1}\sum_{t=1}^T \mathbf{q}_t^{(i)}$. We decompose the optimization error at $\bar{\mathbf{w}}_T^{(i)}$ as

$$\begin{aligned}
&\varepsilon_\phi(\bar{\mathbf{w}}_T^{(i)}) = \max_{j\in[m]} R_j(\bar{\mathbf{w}}_T^{(i)}) - \min_{\mathbf{w}\in\mathcal{W}}\max_{j\in[m]} R_j(\mathbf{w}) = \max_{j\in[m]} R_j(\bar{\mathbf{w}}_T^{(i)}) - \max_{j\in[m]} R_j(\mathbf{w}^*) \\
=& \max_{\mathbf{q}\in\Delta_m} \phi(\bar{\mathbf{w}}_T^{(i)}, \mathbf{q}) - \max_{\mathbf{q}\in\Delta_m} \phi(\mathbf{w}^*, \mathbf{q}) \le \max_{\mathbf{q}\in\Delta_m} \phi(\bar{\mathbf{w}}_T^{(i)}, \mathbf{q}) - \phi\left(\mathbf{w}^*, \bar{\mathbf{q}}_T^{(i)}\right) \\
\le& \frac{1}{T}\max_{\mathbf{q}\in\Delta_m}\sum_{t=1}^T \phi(\mathbf{w}_t^{(i)}, \mathbf{q}) - \frac{1}{T}\sum_{t=1}^T \phi(\mathbf{w}^*, \mathbf{q}_t^{(i)}) \\
\le& \frac{1}{T}\left(\max_{\mathbf{q}\in\Delta_m}\sum_{t=1}^T \phi(\mathbf{w}_t^{(i)}, \mathbf{q}) - \sum_{t=1}^T \phi(\mathbf{w}_t^{(i)}, \mathbf{q}_t^{(i)})\right) + \frac{1}{T}\left(\sum_{t=1}^T \phi(\mathbf{w}_t^{(i)}, \mathbf{q}_t^{(i)}) - \sum_{t=1}^T \phi(\mathbf{w}^*, \mathbf{q}_t^{(i)})\right)
\end{aligned} \tag{43}$$

In the decentralized setting, there is no central server to compute a shared dual sequence $\{\mathbf{q}_t\}$. Consequently, the above decomposition is expressed in terms of regret defined with respect to each worker's local dual iterates $\{\mathbf{q}_t^{(i)}\}$.

Recall that $j(i) = \arg\max_{j\in[m]} \sum_{t=1}^T R_j(\mathbf{w}_t^{(i)})$ and $s_{t,j}^{(i)} = 1 - R_j(\mathbf{w}_t^{(i)})$. For the first term, following (23), we have for all $i \in [m]$,

$$\begin{aligned}
&\max_{\mathbf{q}\in\Delta_m}\sum_{t=1}^T \phi(\mathbf{w}_t^{(i)}, \mathbf{q}) - \sum_{t=1}^T \phi(\mathbf{w}_t^{(i)}, \mathbf{q}_t^{(i)}) \\
\le& G\sum_{t=1}^T\left(\|\mathbf{w}_t^{(i)} - \mathbf{w}_t^{(j(i))}\|_2 + \sum_{j=1}^m q_{t,j}^{(i)}\|\mathbf{w}_t^{(i)} - \mathbf{w}_t^{(j)}\|_2\right) + \sum_{t=1}^T\sum_{j=1}^m q_{t,j}^{(i)} s_{t,j}^{(j)} - \sum_{t=1}^T s_{t,j(i)}^{(j(i))}
\end{aligned} \tag{44}$$

For the second term, we define $\hat{\mathbf{q}}_t = [q_{t,1}^{(1)}, \ldots, q_{t,m}^{(m)}]^\top$ and have

$$\begin{aligned}
&\sum_{t=1}^T \phi(\mathbf{w}_t^{(i)}, \mathbf{q}_t^{(i)}) - \sum_{t=1}^T \phi(\mathbf{w}^*, \mathbf{q}_t^{(i)}) \\
=& \sum_{t=1}^T\sum_{j=1}^m \left(q_{t,j}^{(i)} R_j(\mathbf{w}_t^{(i)}) - q_{t,j}^{(j)} R_j(\mathbf{w}_t^{(i)})\right) + \sum_{t=1}^T\sum_{j=1}^m \left(q_{t,j}^{(j)} R_j(\mathbf{w}_t^{(i)}) - q_{t,j}^{(j)} R_j(\mathbf{w}^*)\right) \\
&+ \sum_{t=1}^T\sum_{j=1}^m \left(q_{t,j}^{(j)} R_j(\mathbf{w}^*) - q_{t,j}^{(i)} R_j(\mathbf{w}^*)\right) \\
\overset{(2)}{\le}& 2\sum_{t=1}^T \|\hat{\mathbf{q}}_t - \mathbf{q}_t^{(i)}\|_1 + \sum_{t=1}^T\sum_{j=1}^m \left(q_{t,j}^{(j)} R_j(\mathbf{w}_t^{(i)}) - q_{t,j}^{(j)} R_j(\mathbf{w}^*)\right)
\end{aligned} \tag{45}$$

Combining (44) and (45) with the decomposition in (43), we obtain the following bound:

$$\begin{aligned}\varepsilon_\phi(\bar{\mathbf{w}}_T^{(i)}) \leq & \underbrace{\frac{1}{T}\sum_{t=1}^{T}\sum_{j=1}^{m}\left(q_{t,j}^{(j)}R_j(\mathbf{w}_t^{(i)}) - q_{t,j}^{(j)}R_j(\mathbf{w}^*)\right)}_{\mathrm{Reg}_{w,T}^{(i)}(\mathbf{w}^*)} + \underbrace{\frac{1}{T}\sum_{t=1}^{T}\left(\sum_{j=1}^{m} q_{t,j}^{(i)}s_{t,j}^{(j)} - s_{t,j(i)}^{(j(i))}\right)}_{\mathrm{Reg}_{q,T}^{(i)}} \\ & + \underbrace{\frac{1}{T}\sum_{t=1}^{T}\left(2\|\hat{\mathbf{q}}_t - \mathbf{q}_t^{(i)}\|_1 + G\|\mathbf{w}_t^{(i)} - \mathbf{w}_t^{(j(i))}\|_2 + G\sum_{j=1}^{m} q_{t,j}^{(i)}\|\mathbf{w}_t^{(i)} - \mathbf{w}_t^{(j)}\|_2\right)}_{\texttt{term(dis-dec)}}\end{aligned} \tag{46}$$

Let $\bar{\mathbf{g}}_{q,t} = m^{-1}\sum_{i=1}^{m}\mathbf{g}_{q,t}^{(i)}$. We define the virtual iterate $\tilde{\mathbf{w}}_{t+1}$ as in (25) and

$$\tilde{\mathbf{q}}_{t+1} = \operatorname*{arg\,min}_{\mathbf{q}\in\Delta_m}\left\{\left\langle \eta_q\sum_{\tau=1}^{t}\bar{\mathbf{g}}_{q,\tau}, \mathbf{q}\right\rangle + \nu_q(\mathbf{q})\right\}, \tag{47}$$

with the initialization $\tilde{\mathbf{q}}_1 = \arg\min_{\mathbf{q}\in\Delta_m}\nu_q(\mathbf{q})$. To streamline the subsequent bounds, we introduce the following discrepancy measures

$$\varepsilon_{w,t} = \max_{i\in[m]}\left\|\mathbf{w}_t^{(i)} - \tilde{\mathbf{w}}_t\right\|_2, \qquad \varepsilon_{q,t} = \max_{i\in[m]}\left\|\mathbf{q}_t^{(i)} - \tilde{\mathbf{q}}_t\right\|_2, \qquad t\in[T].$$

These quantities capture, respectively, the disagreement of the primal and dual iterates around the corresponding virtual sequences. In the following, we analyze each of these components in (46).

D.5.1 Regret $\mathrm{Reg}_{w,T}^{(i)}$

The analysis of the local regret $\mathrm{Reg}_{w,T}^{(i)}$ in the decentralized setting closely follows the SemiDec analysis in Section D.2.1. The main difference is that the regret uses the locally maintained dual weights $q_{t,i}^{(i)}$ for all $i\in[m]$, rather than the corresponding weights from the shared dual vector $\mathbf{q}_t$.

Note that, in the decentralized setting, the averaged gradient is written as $\bar{\mathbf{g}}_{w,t} = \frac{1}{m}\sum_{i=1}^{m}\mathbf{g}_{w,t}^{(i)} = \frac{1}{m}\sum_{i=1}^{m} q_{t,i}^{(i)}\cdot\nabla\ell(\mathbf{w}_t^{(i)};\mathbf{z}_t^{(i)})$. Following (27), we obtain

$$\begin{aligned}\mathbb{E}[\mathrm{Reg}_{w,T}^{(i)}(\mathbf{u})] \leq & \, m\sum_{t=1}^{T}\mathbb{E}[\langle\bar{\mathbf{g}}_{w,t}, \tilde{\mathbf{w}}_t - \mathbf{u}\rangle] \\ & + G\sum_{t=1}^{T}\mathbb{E}\left[\sum_{j=1}^{m} q_{t,j}^{(j)}\left(\|\tilde{\mathbf{w}}_t - \mathbf{w}_t^{(j)}\|_2 + \|\mathbf{w}_t^{(i)} - \mathbf{w}_t^{(j)}\|_2\right)\right].\end{aligned} \tag{48}$$

For the first term on the right-hand side of (48), we proceed as in the proof of (28). In the decentralized setting, the averaged gradient can be rewritten as

$$\bar{\mathbf{g}}_{w,t} = \frac{1}{m}\sum_{i=1}^{m} q_{t,i}^{(i)}\nabla\ell(\mathbf{w}_t^{(i)};\mathbf{z}_t^{(i)}) = \frac{1}{m}\sum_{i=1}^{m}\big(q_{t,i}^{(i)} - \tilde{q}_{t,i}\big)\nabla\ell(\mathbf{w}_t^{(i)};\mathbf{z}_t^{(i)}) + \frac{1}{m}\sum_{i=1}^{m}\tilde{q}_{t,i}\nabla\ell(\mathbf{w}_t^{(i)};\mathbf{z}_t^{(i)}).$$

Under Assumption 2, we have

$$\begin{aligned}\|\bar{\mathbf{g}}_{w,t}\|_2 &\leq \frac{1}{m}\sum_{i=1}^{m}|q_{t,i}^{(i)} - \tilde{q}_{t,i}|\cdot\|\nabla\ell(\mathbf{w}_t^{(i)};\mathbf{z}_t^{(i)})\|_2 + \left\|\frac{1}{m}\sum_{i=1}^{m}\tilde{q}_{t,i}\nabla\ell(\mathbf{w}_t^{(i)};\mathbf{z}_t^{(i)})\right\|_2 \\ &\leq \frac{G}{m}\|\hat{\mathbf{q}}_t - \tilde{\mathbf{q}}_t\|_1 + \frac{G}{m}\sum_{i=1}^{m}\tilde{q}_{t,i} = (1 + \|\hat{\mathbf{q}}_t - \tilde{\mathbf{q}}_t\|_1)\frac{G}{m} \leq (1 + m\varepsilon_{q,t})\frac{G}{m},\end{aligned} \tag{49}$$

where the last inequality follows

$$\|\hat{\mathbf{q}}_t - \tilde{\mathbf{q}}_t\|_1 = \sum_{i=1}^{m}|q_{t,i}^{(i)} - \tilde{q}_{t,i}| \leq \sum_{i=1}^{m}\|\mathbf{q}_t^{(i)} - \tilde{\mathbf{q}}_t\|_2 \leq m\varepsilon_{q,t}. \tag{50}$$

Consequently,

$$
\begin{aligned}
m\sum_{t=1}^{T}\mathbb{E}\left[\langle\bar{\mathbf{g}}_{w,t},\tilde{\mathbf{w}}_t-\mathbf{u}\rangle\right] &\le m\left(\frac{D^2}{\eta_w}+\frac{\eta_w}{2}\sum_{t=1}^{T}\mathbb{E}\left[\|\bar{\mathbf{g}}_{w,t}\|_2^2\right]\right)\\
&\le \frac{mD^2}{\eta_w}+\frac{\eta_w G^2}{2m}\sum_{t=1}^{T}\mathbb{E}\left[\left(1+m\varepsilon_{q,t}\right)^2\right].
\end{aligned}
\tag{51}
$$

We now bound the last term on the right-hand side of (48). By the definition of $\varepsilon_{w,t}$, we have

$$
\begin{aligned}
&\sum_{t=1}^{T}\sum_{j=1}^{m}q_{t,j}^{(j)}\left(\|\tilde{\mathbf{w}}_t-\mathbf{w}_t^{(j)}\|+\|\mathbf{w}_t^{(i)}-\mathbf{w}_t^{(j)}\|\right)\le 3\sum_{t=1}^{T}\sum_{j=1}^{m}q_{t,j}^{(j)}\varepsilon_{w,t}\\
=&3\sum_{t=1}^{T}\sum_{j=1}^{m}\tilde{q}_{t,j}\varepsilon_{w,t}+3\sum_{t=1}^{T}\sum_{j=1}^{m}\left(q_{t,j}^{(j)}-\tilde{q}_{t,j}\right)\varepsilon_{w,t}\\
\le&3\sum_{t=1}^{T}\left(1+\|\hat{\mathbf{q}}_t-\tilde{\mathbf{q}}_t\|_1\right)\varepsilon_{w,t}\le 3\sum_{t=1}^{T}\left(1+m\varepsilon_{q,t}\right)\varepsilon_{w,t},
\end{aligned}
\tag{52}
$$

where the last inequality follows (50). Substituting (51) and (52) into (48) and setting $\eta_w=\frac{\sqrt{2}mD}{\sqrt{T}G}$ yields

$$
\mathbb{E}\left[\mathrm{Reg}_{w,T}^{(i)}(\mathbf{u})\right]\le\sqrt{\frac{T}{2}}DG+\frac{DG}{\sqrt{2T}}\sum_{t=1}^{T}\mathbb{E}\left[\left(1+m\varepsilon_{q,t}\right)^2\right]+3G\sum_{t=1}^{T}\mathbb{E}\left[\left(1+m\varepsilon_{q,t}\right)\varepsilon_{w,t}\right].
\tag{53}
$$

D.5.2 Regret $\mathrm{Reg}_{q,T}^{(i)}$

Recall $s_{t,j}^{(i)}=1-R_j(\mathbf{w}_t^{(i)})$, $\hat{s}_{t,j}^{(i)}=1-\ell(\mathbf{w}_t^{(i)};\mathbf{z}_t^{(i)})$, $\tilde{\mathbf{s}}_t=[\hat{s}_{t,1}^{(1)};\ldots;\hat{s}_{t,m}^{(m)}]$. We then decompose the regret $\mathrm{Reg}_{q,T}^{(i)}$ as

$$
\begin{aligned}
\mathbb{E}\left[\mathrm{Reg}_{q,T}^{(i)}\right]&=\mathbb{E}\left[\sum_{t=1}^{T}\left(\sum_{j=1}^{m}q_{t,j}^{(i)}s_{t,j}^{(j)}-s_{t,j(i)}^{(j(i))}\right)\right]=\mathbb{E}\left[\sum_{t=1}^{T}\langle\mathbf{q}_t^{(i)},\tilde{\mathbf{s}}_t\rangle-\sum_{t=1}^{T}s_{t,j(i)}^{(j(i))}\right]\\
&=\underbrace{\mathbb{E}\left[\sum_{t=1}^{T}\left(\langle\mathbf{q}_t^{(i)},\tilde{\mathbf{s}}_t\rangle-\langle\tilde{\mathbf{q}}_t,\tilde{\mathbf{s}}_t\rangle\right)\right]}_{\texttt{term}(\mathtt{q}_1)}+\underbrace{\mathbb{E}\left[\sum_{t=1}^{T}\left(\langle\tilde{\mathbf{q}}_t,\tilde{\mathbf{s}}_t\rangle-\hat{s}_{t,j(i)}^{(j(i))}\right)\right]}_{\texttt{term}(\mathtt{q}_2)}\\
&\quad+\underbrace{\mathbb{E}\left[\sum_{t=1}^{T}\left(\hat{s}_{t,j(i)}^{(j(i))}-s_{t,j(i)}^{(j(i))}\right)\right]}_{\texttt{term}(\mathtt{q}_3)}
\end{aligned}
\tag{54}
$$

For $\texttt{term}(\mathtt{q}_1)$, we have

$$
\sum_{t=1}^{T}\left(\langle\mathbf{q}_t^{(i)},\tilde{\mathbf{s}}_t\rangle-\langle\tilde{\mathbf{q}}_t,\tilde{\mathbf{s}}_t\rangle\right)\le\sum_{t=1}^{T}\|\mathbf{q}_t^{(i)}-\tilde{\mathbf{q}}_t\|_2\|\tilde{\mathbf{s}}_t\|_2\overset{(2)}{\le}\sqrt{m}\sum_{t=1}^{T}\|\mathbf{q}_t^{(i)}-\tilde{\mathbf{q}}_t\|_2\le\sqrt{m}\sum_{t=1}^{T}\varepsilon_{q,t}.
\tag{55}
$$

Then, we analyze $\texttt{term}(\mathtt{q}_2)$. By Lemma 4 and the update rule (47), we obtain for all $i\in[m]$,

$$
\sum_{t=1}^{T}\left(\langle\tilde{\mathbf{q}}_t,\bar{\mathbf{g}}_{q,t}\rangle-\frac{1}{m}\hat{s}_{t,i}^{(i)}\right)\le\frac{\ln m}{\eta_q}+\frac{\eta_q}{2}\sum_{t=1}^{T}\sum_{i=1}^{m}\tilde{q}_{t,i}\left(\frac{1}{m}\hat{s}_{t,i}^{(i)}\right)^2\overset{(2)}{\le}\frac{\ln m}{\eta_q}+\frac{\eta_q T}{2m^2},
\tag{56}
$$

Noting that $\tilde{\mathbf{s}}_t=m\bar{\mathbf{g}}_{q,t}$, we have

$$
\mathbb{E}\left[\sum_{t=1}^{T}\left(\langle\tilde{\mathbf{q}}_t,\tilde{\mathbf{s}}_t\rangle-\hat{s}_{t,j(i)}^{(j(i))}\right)\right]\overset{(56)}{\le}\frac{m\ln m}{\eta_q}+\frac{\eta_q T}{2m}.
\tag{57}
$$

For $\texttt{term}(\mathtt{q}_3)$, we apply the argument for (32) to the data-dependent index $j(i)$ and obtain

$$\texttt{term}(q_3) \le \sqrt{2T(\ln(m)+1)} \le \sqrt{6}\,\sqrt{T\ln m}. \tag{58}$$

Here we use $\ln m \ge \ln 2 > 1/2$. Combining (54), (55), (57)–(58), using $\sqrt{2}+\sqrt{6}<4$, and setting $\eta_q = m\sqrt{\frac{2\ln m}{T}}$, we obtain

$$\mathbb{E}\left[\mathrm{Reg}_{q,T}^{(i)}\right] \le \sqrt{m}\sum_{t=1}^{T}\mathbb{E}[\varepsilon_{q,t}] + 4\sqrt{T\ln m}. \tag{59}$$

D.5.3 Discrepancy error term and results

We now bound the error caused by the disagreement among local models and between each local dual vector and the proxy vector, captured by

$$\begin{aligned}
\texttt{term}(\texttt{dis}-\texttt{dec}) &= \sum_{t=1}^{T}\left(2\|\hat{\mathbf{q}}_t - \mathbf{q}_t^{(i)}\|_1 + G\|\mathbf{w}_t^{(i)} - \mathbf{w}_t^{(j(i))}\|_2 + G\sum_{j=1}^{m} q_{t,j}^{(i)}\|\mathbf{w}_t^{(i)} - \mathbf{w}_t^{(j)}\|_2\right)\\
&\le \sum_{t=1}^{T}\left(2\|\hat{\mathbf{q}}_t - \tilde{\mathbf{q}}_t\|_1 + 2\|\tilde{\mathbf{q}}_t - \mathbf{q}_t^{(i)}\|_1 + 4G\varepsilon_{w,t}\right)\\
&\le \sum_{t=1}^{T}\left(2\|\hat{\mathbf{q}}_t - \tilde{\mathbf{q}}_t\|_1 + 2\sqrt{m}\|\tilde{\mathbf{q}}_t - \mathbf{q}_t^{(i)}\|_2 + 4G\varepsilon_{w,t}\right)\\
&\overset{(50)}{\le} 4\sum_{t=1}^{T}\left(m\varepsilon_{q,t} + G\varepsilon_{w,t}\right)
\end{aligned} \tag{60}$$

Substituting the regret bounds (53) and (59), together with the discrepancy estimate $\mathbb{E}[\texttt{term}(\texttt{dis-dec})] \le 4\sum_{t=1}^{T}\mathbb{E}\left[m\varepsilon_{q,t} + G\varepsilon_{w,t}\right]$, into the decomposition (46) yields, for all $i \in [m]$,

$$\begin{aligned}
\mathbb{E}\left[\varepsilon_\phi(\bar{\mathbf{w}}_T^{(i)})\right] \le \frac{1}{T}\Bigg[&\sqrt{\frac{T}{2}}DG + \frac{DG}{\sqrt{2T}}\sum_{t=1}^{T}\mathbb{E}\left[(1+m\varepsilon_{q,t})^2\right] + 3G\sum_{t=1}^{T}\mathbb{E}\left[(1+m\varepsilon_{q,t})\,\varepsilon_{w,t}\right]\\
&+ \sqrt{m}\sum_{t=1}^{T}\mathbb{E}[\varepsilon_{q,t}] + 4\sqrt{T\ln m} + 4\sum_{t=1}^{T}\mathbb{E}\left[m\varepsilon_{q,t} + G\varepsilon_{w,t}\right]\Bigg].
\end{aligned} \tag{61}$$

For any $i \in [m]$ and $t > 1$, Lemma 5 and average preservation give $\|\mathbf{w}_t^{(i)} - \tilde{\mathbf{w}}_t\|_2 \le \eta_w\|\mathbf{F}_{t-1}^{(i)} - \bar{\mathbf{F}}_{t-1}\|_2$. The same argument for the dual update yields $\|\mathbf{q}_t^{(i)} - \tilde{\mathbf{q}}_t\|_2 \le \eta_q\|\mathbf{L}_{t-1}^{(i)} - \bar{\mathbf{L}}_{t-1}\|_2$. Thus the second-moment assumptions imply $(\mathbb{E}[\varepsilon_{w,t}^2])^{1/2} \le \eta_w\varepsilon_{c,w}$, and $(\mathbb{E}[\varepsilon_{q,t}^2])^{1/2} \le \eta_q\varepsilon_{c,q}$.

Minkowski's and Cauchy–Schwarz inequalities then give

$$\mathbb{E}[(1+m\varepsilon_{q,t})^2] \le (1+m\eta_q\varepsilon_{c,q})^2, \quad \mathbb{E}[(1+m\varepsilon_{q,t})\varepsilon_{w,t}] \le (1+m\eta_q\varepsilon_{c,q})\eta_w\varepsilon_{c,w}.$$

Substituting these bounds into (61), together with $\mathbb{E}[\varepsilon_{w,t}] \le \eta_w\varepsilon_{c,w}$ and $\mathbb{E}[\varepsilon_{q,t}] \le \eta_q\varepsilon_{c,q}$, yields, for all $i \in [m]$,

$$\begin{aligned}
\mathbb{E}\left[\varepsilon_\phi(\bar{\mathbf{w}}_T^{(i)})\right] \le &\frac{DG}{\sqrt{2T}} + \frac{DG}{\sqrt{2T}}\left(1 + m^2\sqrt{\frac{2\ln m}{T}}\varepsilon_{c,q}\right)^2 + \frac{3\sqrt{2}mD}{\sqrt{T}}\left(1 + m^2\sqrt{\frac{2\ln m}{T}}\varepsilon_{c,q}\right)\varepsilon_{c,w}\\
&+ m^{3/2}\sqrt{\frac{2\ln m}{T}}\varepsilon_{c,q} + 4\sqrt{\frac{\ln m}{T}} + 4\left(m^2\sqrt{\frac{2\ln m}{T}}\varepsilon_{c,q} + \frac{\sqrt{2}mD}{\sqrt{T}}\varepsilon_{c,w}\right)
\end{aligned} \tag{62}$$

For the compact notation used in Theorem 2, define

$$C_1 = \left(1 + m^2\sqrt{\frac{2\ln m}{T}}\,\varepsilon_{c,q}\right)^2,\ C_2 = m\left(1 + m^2\sqrt{\frac{2\ln m}{T}}\,\varepsilon_{c,q}\right)\varepsilon_{c,w},\ C_3 = m^2\sqrt{2\ln m}\,\varepsilon_{c,q}.$$

Equation (62) then yields the bound in Theorem 2.

D.6 THE PROOF OF COROLLARY 3

When adopting the accelerated gossip update (18), we set $\mathbf{F}_t^{(i)} = \mathbf{X}_t^{(i)}$ and $\bar{\mathbf{F}}_t = \bar{\mathbf{X}}_t = \frac{1}{m}\mathbf{X}_t\mathbf{1}$. By Lemma 8, the average is preserved, i.e., $\bar{\mathbf{F}}_t = \bar{\mathbf{F}}_{t-1} + \bar{\mathbf{g}}_{w,t}$, and hence the conservation condition in Theorem 2 is satisfied.

First consider $T \geq 2\ln m$. Set $k_q = \lceil (m^{5/2}+1)^2 \rceil$ and $k_w = \lceil (m/\sqrt{\ln m}+1)^2 \rceil$. Since $\sum_i \|\mathbf{g}_{q,t}^{(i)}\|_2^2 \leq m$, Lemma 8 gives the uniform pathwise bound

$$\max_{t\in[T]} \|\mathbf{L}_t - \bar{\mathbf{L}}_t\mathbf{1}^\top\|_F \leq \frac{\sqrt{m}}{\sqrt{k_q}-1} \leq \frac{1}{m^2}. \tag{63}$$

Consequently, $\varepsilon_{q,t} \leq \eta_q/m^2$ for all t, and

$$\sum_i \|\mathbf{g}_{w,t}^{(i)}\|_2^2 \leq (2+2m\varepsilon_{q,t}^2)G^2 \leq \left(2+\frac{4\ln m}{mT}\right)G^2 \leq 4G^2.$$

Applying Lemma 8 with the uniform input bound $Q = 4G^2$ yields

$$\max_{t\in[T]} \|\mathbf{F}_t - \bar{\mathbf{F}}_t\mathbf{1}^\top\|_F \leq \frac{2G}{\sqrt{k_w}-1} \leq \frac{2G\sqrt{\ln m}}{m}. \tag{64}$$

Thus we may take $\varepsilon_{c,q} = 1/m^2$ and $\varepsilon_{c,w} = 2G\sqrt{\ln m}/m$. Substituting these relations into the general optimization bound (62) yields, for all $i \in [m]$,

$$\begin{aligned}
\mathbb{E}\left[\varepsilon_\phi(\bar{\mathbf{w}}_T^{(i)})\right] \leq & \frac{DG}{\sqrt{2T}} + \frac{DG}{\sqrt{2T}}\left(1+\sqrt{\frac{2\ln m}{T}}\right)^2 + \frac{6\sqrt{2\ln m}DG}{\sqrt{T}}\left(1+\sqrt{\frac{2\ln m}{T}}\right) \\
& + \frac{1}{\sqrt{m}}\sqrt{\frac{2\ln m}{T}} + 4\sqrt{\frac{\ln m}{T}} + 4\sqrt{\frac{2\ln m}{T}} + \frac{8\sqrt{2\ln m}DG}{\sqrt{T}} \\
\leq & (32DG+9\sqrt{2})\sqrt{\frac{\ln m}{T}},
\end{aligned} \tag{65}$$

For the last inequality, set $a = \sqrt{2\ln m/T} \leq 1$. After factoring out $\sqrt{(\ln m)/T}$, the coefficient of DG is

$$A(m,a) = \frac{1+(1+a)^2}{\sqrt{2\ln m}} + \sqrt{2}(14+6a).$$

For $m \geq 3$, $A(m,a) \leq 20\sqrt{2} + 5/\sqrt{2\ln 3} < 32$. For $m = 2$, the integer horizon satisfies $T \geq 2$, so $a \leq \sqrt{\ln 2}$ and

$$A(2,a) \leq \frac{1+(1+\sqrt{\ln 2})^2}{\sqrt{2\ln 2}} + \sqrt{2}(14+6\sqrt{\ln 2}) < 32.$$

The remaining coefficient is at most $9\sqrt{2}$, since $m^{-1/2} + 2\sqrt{2} + 4 \leq 9$. For $T < 2\ln m$, the claimed bound follows from $\varepsilon_\phi(\bar{\mathbf{w}}_T^{(i)}) \leq 1$.

D.7 THE PROOF OF COROLLARY 4

We now analyze Algorithm 2 under Choco-gossip. By Lemma 2, the average-preservation conditions hold automatically, namely $\sum_{i=1}^m \mathbf{F}_t^{(i)} = \sum_{i=1}^m \sum_{\tau=1}^t \mathbf{g}_{w,\tau}^{(i)}$ and $\sum_{i=1}^m \mathbf{L}_t^{(i)} = \sum_{i=1}^m \sum_{\tau=1}^t \mathbf{g}_{q,\tau}^{(i)}$.

First consider $T \geq 2\ln m$, and set $k_q = 4m^5 + 2$ and $k_w = m^2/(\ln m) + 2$. Lemma 9, with $\sum_i \|\mathbf{g}_{q,t}^{(i)}\|_2^2 \leq m$, gives

$$\left(\mathbb{E}[\|\mathbf{L}_t - \bar{\mathbf{L}}_t\mathbf{1}^\top\|_F^2]\right)^{1/2} \leq 2\sqrt{\frac{m}{k_q-2}} = \frac{1}{m^2}.$$

Hence $\mathbb{E}[\varepsilon_{q,t}^2] \le \eta_q^2/m^4$, and

$$\mathbb{E}\left[\sum_i \|\mathbf{g}_{w,t}^{(i)}\|_2^2\right] \le (2 + 2m\mathbb{E}[\varepsilon_{q,t}^2])G^2 \le \left(2 + \frac{4\ln m}{mT}\right) G^2 \le 4G^2.$$

Applying Lemma 9 with $Q = 4G^2$ therefore yields

$$\left(\mathbb{E}[\|\mathbf{F}_t - \bar{\mathbf{F}}_t \mathbf{1}^\top\|_F^2]\right)^{1/2} \le \frac{4G}{\sqrt{k_w - 2}} = \frac{4G\sqrt{\ln m}}{m}.$$

We may thus take $\varepsilon_{c,q} = 1/m^2$ and $\varepsilon_{c,w} = 4G\sqrt{\ln m}/m$ in (62), yielding

$$\begin{aligned}
\mathbb{E}\left[\varepsilon_\phi(\bar{\mathbf{w}}_T^{(i)})\right] \le & \frac{DG}{\sqrt{2T}} + \frac{DG}{\sqrt{2T}}\left(1 + \sqrt{\frac{2\ln m}{T}}\right)^2 + \frac{12\sqrt{2\ln m}DG}{\sqrt{T}}\left(1 + \sqrt{\frac{2\ln m}{T}}\right) \\
& + \frac{1}{\sqrt{m}}\sqrt{\frac{2\ln m}{T}} + 4\sqrt{\frac{\ln m}{T}} + 4\sqrt{\frac{2\ln m}{T}} + \frac{16\sqrt{2\ln m}DG}{\sqrt{T}} \\
\le & \frac{5DG}{\sqrt{2T}} + \frac{40\sqrt{2\ln m}DG}{\sqrt{T}} + 9\sqrt{\frac{2\ln m}{T}} \\
\le & (61DG + 9\sqrt{2})\sqrt{\frac{\ln m}{T}}.
\end{aligned} \tag{66}$$

where we use $T \ge 2\ln m$, $m^{-1/2} + 2\sqrt{2} + 4 \le 9$, and $40\sqrt{2} + 5/\sqrt{2\ln m} \le 40\sqrt{2} + 5/\sqrt{2\ln 2} < 61$. For $T < 2\ln m$, the claimed bound follows from $\varepsilon_\phi(\bar{\mathbf{w}}_T^{(i)}) \le 1$.

D.8 The Proof of Lemma 8

Denote $\mathbf{g}_t = [\mathbf{g}_t^{(1)}, \ldots, \mathbf{g}_t^{(m)}]$ and $\bar{\mathbf{g}}_t = \frac{1}{m}\mathbf{g}_t\mathbf{1}$. By the choice of L, we have $14c^{2L} \le \frac{1}{k}$. By Lemma 1 and (18), Algorithm 3 preserves the average, i.e.,

$$\bar{\mathbf{X}}_t = \frac{1}{m}\mathbf{X}_t\mathbf{1} = \frac{1}{m}(\mathbf{X}_{t-1} + \mathbf{g}_t)\mathbf{1} = \bar{\mathbf{X}}_{t-1} + \bar{\mathbf{g}}_t.$$

Therefore,

$$\begin{aligned}
\|\mathbf{X}_t - \bar{\mathbf{X}}_t\mathbf{1}^\top\|_F &\le \sqrt{14}c^L\|\mathbf{X}_{t-1} + \mathbf{g}_t - (\bar{\mathbf{X}}_{t-1} + \bar{\mathbf{g}}_t)\mathbf{1}^\top\|_F \\
&\le \frac{1}{\sqrt{k}}\|(\mathbf{X}_{t-1} - \bar{\mathbf{X}}_{t-1}\mathbf{1}^\top) + (\mathbf{g}_t - \bar{\mathbf{g}}_t\mathbf{1}^\top)\|_F \\
&\le \frac{1}{\sqrt{k}}\|\mathbf{X}_{t-1} - \bar{\mathbf{X}}_{t-1}\mathbf{1}^\top\|_F + \frac{1}{\sqrt{k}}\|\mathbf{g}_t - \bar{\mathbf{g}}_t\mathbf{1}^\top\|_F,
\end{aligned}$$

Since $\bar{\mathbf{g}}_t$ is the minimizer of $\sum_i \|\mathbf{g}_t^{(i)} - \mathbf{u}\|_2^2$ over $\mathbf{u}$, we have the standard bound

$$\|\mathbf{g}_t - \bar{\mathbf{g}}_t\mathbf{1}^\top\|_F^2 = \sum_{i=1}^m \|\mathbf{g}_t^{(i)} - \bar{\mathbf{g}}_t\|_2^2 \le \sum_{i=1}^m \|\mathbf{g}_t^{(i)}\|_2^2 \le Q.$$

Define $\mathbf{e}_t = \|\mathbf{X}_t - \bar{\mathbf{X}}_t\mathbf{1}^\top\|_F$. Then we obtain $\mathbf{e}_t \le \frac{1}{\sqrt{k}}\mathbf{e}_{t-1} + \frac{\sqrt{Q}}{\sqrt{k}}$. By recursion and $\mathbf{e}_0 = 0$, we have

$$\mathbf{e}_t \le \frac{\sqrt{Q}}{\sqrt{k}}\sum_{s=0}^{t-1}\left(\frac{1}{\sqrt{k}}\right)^s \le \frac{\sqrt{Q}}{\sqrt{k} - 1},$$

which proves (19).

D.9 THE PROOF OF LEMMA 9

Recall that the number of communication rounds is defined as $L = \left\lceil \frac{2\ln k}{\gamma\rho} \right\rceil$ for a constant $k > 2$. Then denote $\mathbf{g}_t = [\mathbf{g}_t^{(1)}, \ldots, \mathbf{g}_t^{(m)}]$ and $\bar{\mathbf{g}}_t = \frac{1}{m}\mathbf{g}_t\mathbf{1}$ we have $\left(1 - \frac{\gamma\rho}{2}\right)^L \leq \frac{1}{k}$. Let $a_t = \mathbb{E}[\|\mathbf{X}_t - \bar{\mathbf{X}}_t\mathbf{1}^\top\|_F^2 + \|\mathbf{X}_t - \hat{\mathbf{X}}_t\|_F^2]$. Conditioning on the inputs to the current consensus call, applying Lemma 2, and then taking total expectations gives

$$\begin{aligned} a_t &\leq \frac{1}{k}\mathbb{E}\left[\|\mathbf{X}_{t-1} + \mathbf{g}_t - (\bar{\mathbf{X}}_{t-1} + \bar{\mathbf{g}}_t)\mathbf{1}^\top\|_F^2 + \|\mathbf{X}_{t-1} + \mathbf{g}_t - \hat{\mathbf{X}}_{t-1}\|_F^2\right] \\ &\leq \frac{2}{k}a_{t-1} + \frac{2}{k}\mathbb{E}\left[\sum_{i=1}^{m}\left(\|\mathbf{g}_t^{(i)} - \bar{\mathbf{g}}_t\|_2^2 + \|\mathbf{g}_t^{(i)}\|_2^2\right)\right] \\ &\leq \frac{2}{k}a_{t-1} + \frac{4Q}{k}, \end{aligned}$$

where the last inequality uses $\sum_i \|\mathbf{g}_t^{(i)} - \bar{\mathbf{g}}_t\|_2^2 \leq \sum_i \|\mathbf{g}_t^{(i)}\|_2^2$. Since $a_0 = 0$, unrolling the recursion gives

$$a_t \leq \frac{4Q}{k}\sum_{s=0}^{t-1}(2/k)^s \leq \frac{4Q}{k-2},$$

which proves (21).