COMMUNICATION-EFFICIENT AGNOSTIC FEDERATED LEARNING VIA FASTER CONVERGENCE AND COMPRESSION

Anonymous authors
Paper under double-blind review

ABSTRACT

Agnostic federated learning (AFL) seeks a model that performs reliably across m heterogeneous workers, but communication remains a bottleneck. We improve communication efficiency by reducing the number of synchronization rounds via faster convergence and the communication cost per round via compression. We first propose AFL-BR, which updates the dual weights over workers using online mirror ascent with KL divergence and blockwise restarts. It achieves an $O((\log m)^{1/4}T^{-1/8})$ stationarity rate after T update rounds, reducing the m-dependence of the synchronization rounds required for convergence from polynomial to logarithmic order. Building on AFL-BR, we develop AFL-COM by applying bidirectional compression with error feedback (EF). Instead of compressing local gradients, workers apply EF to their dual-weighted gradients, enabling direct control of the aggregated compression error under time-varying weights. We then establish an $O((\delta^{-1} + (\log m)^{1/4})T^{-1/8})$ stationarity rate for AFL-COM under general δ-approximate compressors and improve the δ-dependence from δ^{-1} to $\delta^{-1/2}$ for additive-and-idempotent compressors with shared randomness (SR). With suitable compression levels, AFL-COM retains the same convergence rate as AFL-BR at a lower per-round communication cost, yielding reductions in total communication complexity by factors of $(\log m)^{1/4}$ with Top-k and $(\log m)^{1/2}$ with Rand-k and SR. Experiments validate the improved synchronization and communication efficiency of our methods.

1 INTRODUCTION

Federated learning (FL) enables multiple workers to collaboratively train a model while keeping data distributed (McMahan et al., 2017), making it attractive for privacy-sensitive applications and large-scale model training (Li et al., 2020). Classical FL methods learn a global model by minimizing the average loss over m workers. In practice, however, data distributions across workers can be highly heterogeneous (Ye et al., 2023), making an average-optimal model vulnerable to distribution shifts and unfair across workers (Mohri et al., 2019). These limitations are particularly concerning in high-stakes applications such as healthcare (Chen et al., 2023) and finance (Long et al., 2020).

To address these issues, Mohri et al. (2019) propose agnostic FL (AFL), which seeks a model that performs robustly across arbitrary mixtures of local data distributions. Let $\Delta_m = \{\mathbf{q} \in \mathbb{R}^m : \mathbf{q} \geq \mathbf{0}_m, \mathbf{1}_m^\top \mathbf{q} = 1\}$ denote the probability simplex. AFL solves the 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 $\mathbf{q}$ denotes the dual weights over m workers and $R_i(\mathbf{w}) = \mathbb{E}_{\mathbf{z}\sim\mathcal{P}_i}[\ell(\mathbf{w}; \mathbf{z})]$ is the local objective of worker i, namely, the expected loss of model $\mathbf{w} \in \mathbb{R}^d$ on samples drawn from the local data distribution $\mathcal{P}_i$. The same objective also appears in group distributionally robust optimization (GDRO), with workers interpreted as groups (Sagawa et al., 2020). We focus on smooth and possibly nonconvex local objectives over $\mathcal{W} = \mathbb{R}^d$, as commonly encountered in neural-network training.

Despite the robustness benefits of AFL, communication overhead remains a major bottleneck to its scalability and efficiency (Kairouz et al., 2021). We measure communication complexity by the

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

Table 1: **Comparison of nonconvex AFL methods using Top-k and Rand-k as examples.** d is the model dimension and m the number of workers. Definitions of compressors are provided in Appendix B.1. $\star, \clubsuit$ Hidden m-dependence is made explicit; see Appendices C.1 and C.2. § Set $k = \Theta(d(\log m)^{-1/4})$ in Theorem 2. $*$ Set $k = \Theta(d(\log m)^{-1/2})$ with SR in Theorem 3. Parfenov et al. (2026) is omitted since it relies on a Minty-inspired condition with a different criterion.

Algorithm	Technique	Synchronization rounds for ε-stationarity	Communication cost per round	Communication complexity
SGDA*	–	$O(m^{7/2}\varepsilon^{-8})$	$O(dm)$	$O(dm^{9/2}\varepsilon^{-8})$
DRFA$^\clubsuit$	Local updates	$O(\varepsilon^{-12} + m^2\varepsilon^{-8})$	$O(dm)$	$O(dm\varepsilon^{-12} + dm^3\varepsilon^{-8})$
	Fully synchronized	$O(m^2\varepsilon^{-8})$	$O(dm)$	$O(dm^3\varepsilon^{-8})$
AFL-BR (Algorithm 1)	–	$O((\log m)^2\varepsilon^{-8})$	$O(dm)$	$O(dm(\log m)^2\varepsilon^{-8})$
AFL-COM (Algorithm 2)	Compression (Top-k)§	$O((\log m)^2\varepsilon^{-8})$	$O(dm(\log m)^{-1/4})$	$O(dm(\log m)^{7/4}\varepsilon^{-8})$
	Compression (Rand-k with SR)*	$O((\log m)^2\varepsilon^{-8})$	$O(dm(\log m)^{-1/2})$	$O(dm(\log m)^{3/2}\varepsilon^{-8})$

total amount of information exchanged between workers and the server to attain an ε-stationary solution. It is determined by the number of synchronization rounds and the communication cost per round, which can be reduced through faster convergence and compression, respectively. Existing nonconvex AFL methods, however, remain limited in both respects:

1. First, existing guarantees exhibit polynomial dependence on m (Lin et al., 2020; Deng et al., 2020), causing the number of synchronization rounds to grow rapidly with the federation size. Although Deng et al. (2020) reduces synchronization frequency via local updates, its slower convergence requires more rounds than its fully synchronized specialization.
2. Second, existing work on compressed AFL only supports uplink communication and is restricted to unbiased compressors (Parfenov et al., 2026). Moreover, its nonconvex guarantee relies on a Minty-inspired global condition, which is generally difficult to verify. Bidirectional compression with general contractive compressors remains unexplored.

In this work, we address both sources of communication overhead in turn. First, we propose AFL-BR, a blockwise-restarted algorithm that updates the dual weights using online mirror ascent with KL divergence (Hazan, 2016). The main challenge in analysis is to control a dual-gap term with a time-varying comparator while exploiting the KL geometry, whose ℓ_1/ℓ_∞ structure avoids the polynomial dimension factors of Euclidean analysis. Since the optimal dual weights vary with the model iterate, static-regret guarantees for a fixed comparator do not apply directly. To handle the changing comparators, we partition the iterations into blocks and use the block-start maximizer as a fixed comparator, decomposing the dual gap into a comparator-mismatch term controlled by model movement and a static-regret term. A further challenge is that the resulting static-regret bound depends on the block-initial KL divergence, which can be arbitrarily large. We therefore restart the dual weights from the uniform distribution at each block boundary, ensuring the initial KL divergence is bounded by $\ln m$. Together, we establish an $O((\log m)^{1/4}T^{-1/8})$ stationarity guarantee, which yields $O((\log m)^2\varepsilon^{-8})$ synchronization rounds for finding an ε-stationary solution, reducing their dependence on m from at least quadratic (Lin et al., 2020; Deng et al., 2020) to logarithmic.

For the second objective, the dominant per-round communication cost arises from transmitting high-dimensional gradients used to update the model, whereas the dual update exchanges only $O(m)$ scalar entries. Building on AFL-BR, we develop AFL-COM, which applies *bidirectional* compression with error feedback (EF) (Karimireddy et al., 2019) to the vectors exchanged for model updates. Since AFL aggregates local gradients using time-varying dual weights, standard EF on local gradients cannot directly control the global compression error. AFL-COM therefore applies EF to dual-weighted local gradients, enabling sharp control of the aggregated error. Under general δ-approximate compressors, we establish an $O((\delta^{-1} + (\log m)^{1/4})T^{-1/8})$ stationarity rate. Choosing $\delta = \Omega((\log m)^{-1/4})$ keeps the compression error no larger than the term already present in AFL-BR, preserving the order of synchronization rounds for convergence while reducing the communication cost per round. For Top-k sparsification (Stich et al., 2018), which retains the k coordinates with largest absolute values, choosing $k = \Theta(d(\log m)^{-1/4})$ reduces the per-round cost

from $O(dm)$ to $O(dm(\log m)^{-1/4})$ and the communication complexity from $O(dm(\log m)^2\varepsilon^{-8})$ of AFL-BR to $O(dm(\log m)^{7/4}\varepsilon^{-8})$.

We further show that, when using additive-and-idempotent compressors with shared randomness (SR), the downlink compression error vanishes, yielding a sharper $O((\delta^{-1/2} + (\log m)^{1/4})T^{-1/8})$ stationarity rate. The corresponding squared-norm bound has the same δ^{-1} dependence as the lower-bound term for fixed-weight distributed nonconvex optimization (Huang et al., 2022). For Rand-k sparsification (Stich et al., 2018) with SR, setting $k = \Theta(d(\log m)^{-1/2})$ reduces the communication complexity to $O(dm(\log m)^{3/2}\varepsilon^{-8})$, improving over AFL-BR by a factor of $(\log m)^{1/2}$. Table 1 summarizes these guarantees. Our contributions are summarized as follows:

1. We propose AFL-BR, which updates the dual weights using KL-based online mirror ascent with blockwise restarts. It achieves a stationarity rate of $O((\log m)^{1/4}T^{-1/8})$, reducing the m-dependence of synchronization rounds from polynomial to logarithmic order. The same approach also yields a new algorithm and convergence analysis for nonconvex GDRO.
2. Building on AFL-BR, we develop AFL-COM, which uses bidirectional compression with EF. Under general δ-approximate compressors, it achieves an $O((\delta^{-1}+(\log m)^{1/4})T^{-1/8})$ stationarity rate. With Top-k and $k = \Theta(d(\log m)^{-1/4})$, it reduces the communication complexity by a factor of $(\log m)^{1/4}$ relative to AFL-BR.
3. For additive-and-idempotent compressors with SR, we show that the downlink compression error vanishes, improving the δ-dependence of the stationarity rate from δ^{-1} to $\delta^{-1/2}$. With Rand-k using $k = \Theta(d(\log m)^{-1/2})$ and SR, AFL-COM further reduces the communication complexity by a factor of $(\log m)^{1/2}$ relative to AFL-BR.
4. Experiments demonstrate the synchronization efficiency of AFL-BR and show that AFL-COM further reduces the total communication cost.

2 RELATED WORK

2.1 AGNOSTIC FEDERATED LEARNING AND GDRO

AFL can be viewed as a worker-level instance of GDRO, with workers serving as predefined groups (Mohri et al., 2019; Sagawa et al., 2020). In the centralized setting, relevant GDRO methods include stochastic mirror descent (Nemirovski et al., 2009; Carmon & Hausler, 2022) and two-player approaches based on online learning (Soma et al., 2022; Haghtalab et al., 2022; Zhang et al., 2023; Bai et al., 2025; Zhang et al., 2026). However, their theoretical guarantees rely on convex losses and do not directly extend to the smooth nonconvex AFL setting considered here.

Nonconvex AFL can be addressed through general nonconvex–concave minimax methods such as SGDA (Lin et al., 2020), which achieves an $O(T^{-1/8})$ stationarity rate. To reduce communication, Deng et al. (2020) propose DRFA, which performs multiple local updates between periodic dual updates. While this design reduces the number of synchronization rounds to $O(T^{3/4})$, it slows the stationarity rate to $O(T^{-1/16})$, resulting in even more synchronization rounds to reach a target accuracy. Faster rates are possible under stronger structures or assumptions, including $O(T^{-1/6})$ using nested inner loops (Rafique et al., 2022) and $O(T^{-1/4})$ under strict-complementarity and bounded-iterate assumptions (Shen et al., 2024). More recently, Parfenov et al. (2026) incorporate communication compression into AFL, but their nonconvex guarantee relies on a Minty-inspired global condition and restricts compression to the worker-to-server direction and unbiased compressors. Additional results for convex AFL and variants are deferred to Appendix A.

2.2 COMMUNICATION COMPRESSION

Communication compression methods can be broadly categorized into quantization (Horváth et al., 2022) and sparsification (Wangni et al., 2018). They have been widely used in distributed optimization (Alistarh et al., 2017; Wang et al., 2022; Li et al., 2022) and federated learning (Haddadpour et al., 2021). A key mechanism for convergence under compression is error feedback (EF) (Seide et al., 2014; Karimireddy et al., 2019), which maintains an accumulated compression error and adds it back to subsequent updates. EF21 (Richtárik et al., 2021; Gruntkowska et al., 2023) is a variant of EF that tracks the transmitted information and compresses the resulting residual. In the parameter-

server setting, compression has been applied to the uplink (from workers to the server) (Reisizadeh et al., 2020; Basu et al., 2020; Richtárik et al., 2021; Parfenov et al., 2026), the downlink (from the server to workers) (Gruntkowska et al., 2024), and in both directions (Sattler et al., 2020; Zheng et al., 2021; Beznosikov et al., 2022; Gruntkowska et al., 2023). The most closely related work is Parfenov et al. (2026), which studies compressed AFL but only for uplink communication with unbiased compressors. Moreover, Zecchin et al. (2023) study compressed AFL under dual regularization in a decentralized setting, which differs from the parameter-server setting considered here.

Regarding fundamental limits, smooth nonconvex distributed stochastic optimization with contractive compression admits a lower bound of $\Omega(\delta^{-1}T^{-1} + m^{-1/2}T^{-1/2})$ on the expected squared gradient norm (Huang et al., 2022). Fixing $\mathbf{q} = m^{-1}\mathbf{1}_m$ reduces AFL to fixed-weight distributed optimization, so this lower bound provides a benchmark for the dependence on δ.

3 Fast Convergence for Nonconvex AFL

In this section, we study AFL with smooth nonconvex local objectives and develop AFL-BR, which reduces the m-dependence of the stationarity guarantee from polynomial to logarithmic order.

3.1 Problem Setup and Existing Guarantees

For problem (1), define the worst-case objective $\Phi(\mathbf{w}) = \max_{\mathbf{q}\in\Delta_m} \phi(\mathbf{w}, \mathbf{q})$, where $\phi(\mathbf{w}, \mathbf{q}) = \sum_{i=1}^m q_i R_i(\mathbf{w})$. Let $\mathcal{F}_t$ contain all randomness revealed before the stochastic oracle queries at round t, and define $\mathbb{E}_t[\cdot] = \mathbb{E}[\cdot \mid \mathcal{F}_t]$. Since Φ is generally nonsmooth even when each $R_i(\cdot)$ is smooth, we measure convergence through the gradient of its Moreau envelope (Davis & Drusvyatskiy, 2019), a standard criterion in nonconvex–concave minimax optimization (Lin et al., 2020; Deng et al., 2020).

Definition 1 (Moreau envelope)**.** For a function $\Phi : \mathbb{R}^d \to \mathbb{R}$ and a parameter $\lambda > 0$, its Moreau envelope is defined as $\Phi_\lambda(\mathbf{x}) = \min_{\mathbf{y}\in\mathbb{R}^d} \left\{\Phi(\mathbf{y}) + \frac{1}{2\lambda}\|\mathbf{y} - \mathbf{x}\|_2^2\right\}$ for $\mathbf{x} \in \mathbb{R}^d$.

We impose the following assumptions for the nonconvex analysis.

Assumption 1. For every $i \in [m]$, the local objective R_i is differentiable and L-smooth, i.e., $\|\nabla R_i(\mathbf{w}) - \nabla R_i(\mathbf{w}')\|_2 \le L\|\mathbf{w} - \mathbf{w}'\|_2$ for all $\mathbf{w}, \mathbf{w}' \in \mathbb{R}^d$.

Assumption 2. The function $\Phi(\mathbf{w})$ is lower bounded, i.e., $\Phi_* = \inf_{\mathbf{w}\in\mathbb{R}^d} \Phi(\mathbf{w}) > -\infty$.

An output $\mathbf{w}$ is called an ε-stationary solution if $\mathbb{E}[\|\nabla\Phi_{1/(2L)}(\mathbf{w})\|_2] \le \varepsilon$. We further impose the following assumptions on the stochastic oracles used for updates.

Assumption 3. For each $i \in [m]$ and any $\mathbf{w} \in \mathbb{R}^d$, a sample $\mathbf{z} \sim \mathcal{P}_i$ produces $\nabla\ell(\mathbf{w}; \mathbf{z})$ satisfying

$$\mathbb{E}_{\mathbf{z}\sim\mathcal{P}_i}[\nabla\ell(\mathbf{w}; \mathbf{z})] = \nabla R_i(\mathbf{w}),\ \|\nabla R_i(\mathbf{w})\|_2 \le G,\ \mathbb{E}_{\mathbf{z}\sim\mathcal{P}_i}\left[\|\nabla\ell(\mathbf{w}; \mathbf{z}) - \nabla R_i(\mathbf{w})\|_2^2\right] \le \sigma_w^2. \quad (2)$$

For convenience, we denote $\Gamma_w^2 := G^2 + \sigma_w^2$, which bounds the stochastic-gradient second moment.

Assumption 4. For every $i \in [m]$ and $\mathbf{w} \in \mathbb{R}^d$, let $\xi_i(\mathbf{w}; \mathbf{z}) = \ell(\mathbf{w}; \mathbf{z}) - R_i(\mathbf{w})$. There exists $\sigma_q > 0$ such that $\mathbb{E}_{\mathbf{z}\sim\mathcal{P}_i}[\xi_i(\mathbf{w}; \mathbf{z})] = 0$ and $\mathbb{E}_{\mathbf{z}\sim\mathcal{P}_i}\left[\exp\left(\lambda\xi_i(\mathbf{w}; \mathbf{z})\right)\right] \le \exp(\lambda^2\sigma_q^2/2)$ for all $\lambda \in \mathbb{R}$.

Assumption 4 constrains only the centered stochastic-loss fluctuation and does not require a uniform bound on either $R_i(\mathbf{w})$ or $\ell(\mathbf{w}; \mathbf{z})$, in contrast to the uniform boundedness assumptions used in prior AFL methods (Mohri et al., 2019; Deng et al., 2020). Moreover, Assumption 4 is satisfied by Gaussian noise with variance at most σ_q^2 and by any centered noise supported on an interval of width at most $2\sigma_q$ via Hoeffding's lemma (Cesa-Bianchi & Lugosi, 2006).

Existing guarantees and motivation. Existing guarantees are stated in terms of global Euclidean quantities whose dependence on m is implicit, potentially obscuring substantial scalability costs. Accounting for this dependence reveals polynomial scaling in m. In particular, SGDA (Lin et al., 2020) specialized to AFL yields an $O(m^{7/16}T^{-1/8})$ stationarity rate, while fully synchronized DRFA (Deng et al., 2020) improves the worker dependence to $O(m^{1/4}T^{-1/8})$, corresponding to $O(m^2\varepsilon^{-8})$ synchronization rounds. Moreover, its local-update variant has a synchronization-round bound of $O(\varepsilon^{-12} + m^2\varepsilon^{-8})$. These polynomial m-factors arise from joint smoothness and oracle variance in SGDA and dual-gradient bounds in DRFA, motivating an analysis that avoids such m-dependent Euclidean quantities. See Appendix C for details.

Algorithm 1 AFL-BR

Initialize $\mathbf{w}_1 \in \mathbb{R}^d$ and $\mathbf{q}_1 = \frac{1}{m}\mathbf{1}_m$
for $t = 1, 2, \ldots, T$ **do**
  On each worker $i \in [m]$**:**
    Draw $\mathbf{z}_t^{(i)} \sim \mathcal{P}_i$ and compute $\mathbf{g}_{w,t}^{(i)} = \nabla \ell(\mathbf{w}_t; \mathbf{z}_t^{(i)})$ and $\mathbf{g}_{q,t}^{(i)} = \ell(\mathbf{w}_t; \mathbf{z}_t^{(i)})$
    Send $\mathbf{g}_{w,t}^{(i)}$ and $\mathbf{g}_{q,t}^{(i)}$ to the server
  On the server:
    Construct $\mathbf{g}_{q,t} = [\mathbf{g}_{q,t}^{(1)}, \ldots, \mathbf{g}_{q,t}^{(m)}]^\top$ and $\mathbf{g}_{w,t} = \sum_{i=1}^m q_{t,i} \mathbf{g}_{w,t}^{(i)}$
    if $t \equiv 0 \pmod{B}$ **then** set $\mathbf{q}_{t+1} = \frac{1}{m}\mathbf{1}_m$ **else** update $\mathbf{q}_{t+1}$ according to (4)
    Send $\mathbf{g}_{w,t}$ to all workers
  On each worker $i \in [m]$**:**
    Receive $\mathbf{g}_{w,t}$ and update $\mathbf{w}_{t+1} = \mathbf{w}_t - \eta_w \mathbf{g}_{w,t}$
end for
Output: $\mathbf{w}_r$, where $r \sim \mathrm{Unif}\{1, \ldots, T\}$

3.2 AFL-BR: A Blockwise-Restarted Algorithm for Nonconvex AFL

The m-dependence of the convergence guarantee is governed by the cumulative dual gap arising in the Moreau-envelope analysis, namely $\sum_{t=1}^T \varepsilon_t^q$, where $\varepsilon_t^q = \phi(\mathbf{w}_t, \mathbf{q}_t^*) - \phi(\mathbf{w}_t, \mathbf{q}_t)$ for $\mathbf{q}_t^* \in \arg\max_{\mathbf{q} \in \Delta_m} \phi(\mathbf{w}_t, \mathbf{q})$. If the comparator $\mathbf{q}_t^*$ is fixed, the term $\sum_{t=1}^T \varepsilon_t^q$ would reduce to static regret, which can be controlled by KL-based online mirror ascent with only logarithmic dependence on m (Hazan, 2016). However, this static-regret guarantee does not apply directly because $\mathbf{q}_t^*$ varies with $\mathbf{w}_t$, which evolves throughout the optimization process. A natural alternative is to track the sequence $\{\mathbf{q}_t^*\}_{t=1}^T$ using dynamic-regret techniques (Herbster & Warmuth, 1998; Lu & Zhang, 2019; Chen et al., 2021). Such guarantees, however, depend on the path variation of $\{\mathbf{q}_t^*\}_{t=1}^T$, which can grow linearly with T and render the resulting convergence bound vacuous.

Following the blockwise analyses (Lin et al., 2020; Deng et al., 2020), we fix the comparator over short intervals. Specifically, we partition $[T]$ into consecutive blocks $\{\mathcal{I}_c\}_{c=1}^N$ of length at most B, and let s_c denote the first index of $\mathcal{I}_c$. Then, for every $t \in \mathcal{I}_c$, we decompose the dual gap as

$$\varepsilon_t^q = \underbrace{\phi(\mathbf{w}_t, \mathbf{q}_t^*) - \phi(\mathbf{w}_t, \mathbf{q}_{s_c}^*)}_{A_t} + \underbrace{\phi(\mathbf{w}_t, \mathbf{q}_{s_c}^*) - \phi(\mathbf{w}_t, \mathbf{q}_t)}_{B_t}. \tag{3}$$

The first term measures the mismatch caused by fixing the comparator as the model evolves, while the second measures the dual error against this fixed comparator.

By the optimality of $\mathbf{q}_{s_c}^*$ at $\mathbf{w}_{s_c}$ and the G-Lipschitzness of $\phi(\cdot, \mathbf{q})$ under Assumption 3, we have

$$\begin{aligned} A_t &= \phi(\mathbf{w}_t, \mathbf{q}_t^*) - \phi(\mathbf{w}_{s_c}, \mathbf{q}_t^*) + \phi(\mathbf{w}_{s_c}, \mathbf{q}_t^*) - \phi(\mathbf{w}_t, \mathbf{q}_{s_c}^*) \\ &\le \phi(\mathbf{w}_t, \mathbf{q}_t^*) - \phi(\mathbf{w}_{s_c}, \mathbf{q}_t^*) + \phi(\mathbf{w}_{s_c}, \mathbf{q}_{s_c}^*) - \phi(\mathbf{w}_t, \mathbf{q}_{s_c}^*) \le 2G\|\mathbf{w}_t - \mathbf{w}_{s_c}\|_2. \end{aligned}$$

Hence, the comparator mismatch is controlled by the model movement within the block. It remains to control B_t. Since $\mathbf{q}_{s_c}^*$ is fixed over $\mathcal{I}_c$, $\sum_{t \in \mathcal{I}_c} B_t$ has the form of static regret. We therefore update the dual weights using online mirror ascent with KL divergence as

$$\mathbf{q}_{t+1} = \arg\max_{\mathbf{q} \in \Delta_m} \left\{ \eta_q \langle \mathbf{g}_{q,t}, \mathbf{q} \rangle - D_{\mathrm{KL}}(\mathbf{q} \| \mathbf{q}_t) \right\}, \tag{4}$$

where $\mathbf{g}_{q,t} = [g_{q,t}^{(1)}, \ldots, g_{q,t}^{(m)}]^\top$ with $g_{q,t}^{(i)} = \ell(\mathbf{w}_t; \mathbf{z}_t^{(i)})$, and $D_{\mathrm{KL}}(\mathbf{q} \| \mathbf{q}_t) = \sum_{i=1}^m q_i \ln(q_i / q_{t,i})$ denotes the KL divergence. The update (4) has the closed form $q_{t+1,i} \propto q_{t,i} \exp(\eta_q g_{q,t}^{(i)}), \forall i \in [m]$.

Moreover, directly running (4) over all T rounds does not provide a uniform blockwise guarantee, since the static-regret bound on each block $\mathcal{I}_c$ depends on the initial divergence $D_{\mathrm{KL}}(\mathbf{q}_{s_c}^* \| \mathbf{q}_{s_c})$, which is not uniformly bounded. Indeed, $\mathbf{q}_{s_c}$ can assign arbitrarily small mass to the support of the comparator $\mathbf{q}_{s_c}^*$. To resolve this issue, we restart the dual weights as $\mathbf{q}_{s_c} = m^{-1}\mathbf{1}_m$ at each block boundary, guaranteeing $D_{\mathrm{KL}}(\mathbf{q}_{s_c}^* \| \mathbf{q}_{s_c}) \le \ln m$ and hence a uniform bound across blocks.

Combining the restarted dual update with stochastic gradient descent on the model yields AFL-BR, summarized in Algorithm 1. At each round, the workers evaluate both stochastic gradients and stochastic losses, while the server aggregates the gradients according to the current weights.

Algorithm 2 AFL-COM

Initialize $\mathbf{w}_1 \in \mathbb{R}^d$, $\mathbf{q}_1 = \frac{1}{m}\mathbf{1}_m$, $\widehat{\mathbf{e}}_1 = \mathbf{0}_d$, and $\mathbf{e}_1^{(i)} = \mathbf{0}_d$ for all $i \in [m]$
for $t = 1, 2, \ldots, T$ **do**
  On each worker $i \in [m]$**:**
    Draw $\mathbf{z}_t^{(i)} \sim \mathcal{P}_i$ and compute $\mathbf{g}_{w,t}^{(i)} = \nabla \ell(\mathbf{w}_t; \mathbf{z}_t^{(i)})$ and $\mathbf{g}_{q,t}^{(i)} = \ell(\mathbf{w}_t; \mathbf{z}_t^{(i)})$
    Compute $\Delta_{w,t}^{(i)} = \mathcal{C}(\mathbf{e}_t^{(i)} + q_{t,i}\mathbf{g}_{w,t}^{(i)})$ and update $\mathbf{e}_{t+1}^{(i)} = \mathbf{e}_t^{(i)} + q_{t,i}\mathbf{g}_{w,t}^{(i)} - \Delta_{w,t}^{(i)}$
    Send $\Delta_{w,t}^{(i)}$ and $\mathbf{g}_{q,t}^{(i)}$ to the server
  On the server:
    Construct $\mathbf{g}_{q,t} = [\mathbf{g}_{q,t}^{(1)}, \ldots, \mathbf{g}_{q,t}^{(m)}]^\top$ and $\widehat{\mathbf{g}}_{w,t} = \sum_{i=1}^m \Delta_{w,t}^{(i)}$
    Compute $\Delta_{w,t} = \mathcal{C}(\widehat{\mathbf{e}}_t + \widehat{\mathbf{g}}_{w,t})$ and update $\widehat{\mathbf{e}}_{t+1} = \widehat{\mathbf{e}}_t + \widehat{\mathbf{g}}_{w,t} - \Delta_{w,t}$
    if $t \equiv 0 \pmod B$ **then** set $\mathbf{q}_{t+1} = \frac{1}{m}\mathbf{1}_m$ **else** update $\mathbf{q}_{t+1}$ according to (4)
    Broadcast $\Delta_{w,t}$ to all workers and send $q_{t+1,i}$ to worker i
  On each worker $i \in [m]$**:**
    Receive $\Delta_{w,t}$ and $q_{t+1,i}$, and update $\mathbf{w}_{t+1} = \mathbf{w}_t - \eta_w \Delta_{w,t}$
end for
Output: $\mathbf{w}_r$, where $r \sim \mathrm{Unif}\{1, \ldots, T\}$

We next establish the convergence guarantee. A further challenge is that standard static-regret bounds depend on the magnitudes of the stochastic losses (Cesa-Bianchi & Lugosi, 2006), which are not uniformly bounded here. To this end, we decompose the stochastic loss into the true loss and noise, and control the noise jointly with the KL stability term through the negative-entropy and log-sum-exp conjugacy. The sub-Gaussian condition then yields a finite bound without requiring uniform boundedness of either $R_i(\mathbf{w})$ or $\ell(\mathbf{w}; \mathbf{z})$, leading to the following guarantee.

Theorem 1. Suppose Assumptions 1–4 hold with $m \geq 2$ and $T \geq 9$. Let $\Delta_\Phi = \Phi_{1/(2L)}(\mathbf{w}_1) - \Phi_*$. For Algorithm 1, set $B = \lceil \sqrt{T} \rceil$, $\eta_w = (8LT^{3/4})^{-1}$, and $\eta_q = 2\sqrt{\ln m}/(\sigma_q T^{1/4})$. Then

$$\mathbb{E}\left[\left\|\nabla \Phi_{1/(2L)}(\mathbf{w}_r)\right\|_2^2\right] \leq \frac{32L\Delta_\Phi}{T^{1/4}} + \frac{4G\Gamma_w}{T^{1/4}} + \frac{16L\sigma_q\sqrt{\ln m}}{T^{1/4}} + \frac{10\Gamma_w^2}{T^{1/2}} + \frac{\Gamma_w^2}{2T^{3/4}}.$$

Remark. Theorem 1 implies $\mathbb{E}[\|\nabla \Phi_{1/(2L)}(\mathbf{w}_r)\|_2] = O((\log m)^{1/4} T^{-1/8})$ by Jensen's inequality. Hence, obtaining an ε-stationary solution requires $O((\log m)^2 \varepsilon^{-8})$ synchronization rounds. Compared with the baselines in Table 1, this reduces the m-dependence of the number of synchronization rounds required for convergence from polynomial to logarithmic order. With $O(dm)$ scalar entries communicated per round, the resulting total communication complexity is $O(dm(\log m)^2 \varepsilon^{-8})$.

Remark. As a byproduct, AFL-BR also yields an $O((\log m)^{1/4} T^{-1/8})$ stationarity guarantee for centralized nonconvex GDRO when all worker-side computations are executed on a single machine.

4 AFL-COM: AFL WITH BIDIRECTIONAL COMPRESSION

AFL-BR reduces the m-dependence of the synchronization-round bound through faster convergence. We further reduce the per-round communication cost via bidirectional compression with δ-approximate compressors, defined below with examples in Appendix B.1.

Definition 2. An operator $\mathcal{C} : \mathbb{R}^d \to \mathbb{R}^d$ is a δ-approximate compressor if, for some $\delta \in (0, 1]$, $\mathbb{E}[\|\mathcal{C}(\mathbf{x}) - \mathbf{x}\|_2^2] \leq (1 - \delta)\|\mathbf{x}\|_2^2$, for all $\mathbf{x} \in \mathbb{R}^d$, where the expectation is over the randomness of $\mathcal{C}$.

4.1 BIDIRECTIONAL COMPRESSION AND GUARANTEES UNDER GENERAL COMPRESSORS

In each round of AFL-BR, workers upload high-dimensional local gradients and scalar losses, while the server broadcasts the aggregated gradient. Since gradient transmission dominates the communication cost, we compress these model-update messages bidirectionally and use EF (Karimireddy et al., 2019) to control the compression errors. A natural approach is to apply EF independently to each worker's raw local gradient, i.e., worker i transmits $\Delta_{w,t}^{(i)} = \mathcal{C}(\mathbf{e}_t^{(i)} + \mathbf{g}_{w,t}^{(i)})$ and accumulates

the residual $\mathbf{e}_{t+1}^{(i)} = \mathbf{e}_t^{(i)} + \mathbf{g}_{w,t}^{(i)} - \Delta_{w,t}^{(i)}$. The server then aggregates the compressed messages using the current dual weights. However, since the weights are applied *outside* the EF recursion, the accumulated uplink error $\sum_{s=1}^{t}\sum_{i=1}^{m} q_{s,i}(\mathbf{g}_{w,s}^{(i)} - \Delta_{w,s}^{(i)})$ is not directly controlled by the worker-side EF guarantee, which only bounds the unweighted residual $\sum_{s=1}^{t}(\mathbf{g}_{w,s}^{(i)} - \Delta_{w,s}^{(i)})$. This mismatch leads to a loose bound on the weighted global error.

We address the mismatch by applying EF to the dual-weighted gradients $q_{t,i}\mathbf{g}_{w,t}^{(i)}$. To this end, the server sends $q_{t,i}$ to worker i at the end of round $t-1$. This incurs a communication cost of m floats per round, which is negligible compared with communicating high-dimensional gradients. Specifically, worker i sends $\Delta_{w,t}^{(i)} = \mathcal{C}(\mathbf{e}_t^{(i)} + q_{t,i}\mathbf{g}_{w,t}^{(i)})$ and updates $\mathbf{e}_{t+1}^{(i)} = \mathbf{e}_t^{(i)} + q_{t,i}\mathbf{g}_{w,t}^{(i)} - \Delta_{w,t}^{(i)}$. The server aggregates $\widehat{\mathbf{g}}_{w,t} = \sum_{i=1}^{m}\Delta_{w,t}^{(i)}$ and applies EF again on the downlink by broadcasting $\Delta_{w,t} = \mathcal{C}(\widehat{\mathbf{e}}_t + \widehat{\mathbf{g}}_{w,t})$ and updating $\widehat{\mathbf{e}}_{t+1} = \widehat{\mathbf{e}}_t + \widehat{\mathbf{g}}_{w,t} - \Delta_{w,t}$. Each worker then updates $\mathbf{w}_{t+1} = \mathbf{w}_t - \eta_w\Delta_{w,t}$, while the dual weights follow the same blockwise-restarted mirror-ascent update as in AFL-BR. We present the resulting procedure in Algorithm 2, termed AFL-COM, and establish its convergence under general δ-approximate compressors as follows.

Theorem 2. Suppose Assumptions 1–4 hold with $m \geq 2$ and $T \geq 9$. Let $\Delta_\Phi = \Phi_{1/(2L)}(\mathbf{w}_1) - \Phi_*$. For Algorithm 2, set $B = \lceil\sqrt{T}\rceil$, $\eta_w = (8LT^{3/4})^{-1}$, and $\eta_q = 2\sqrt{\ln m}/(\sigma_q T^{1/4})$. If $\mathcal{C}$ is a general δ-approximate compressor, then

$$\mathbb{E}\left[\left\|\nabla\Phi_{1/(2L)}(\mathbf{w}_r)\right\|_2^2\right] \leq \frac{32L\Delta_\Phi}{T^{1/4}} + \frac{4G\Gamma_w(16+4\delta-\delta^2)}{\delta^2 T^{1/4}} + \frac{16L\sigma_q\sqrt{\ln m}}{T^{1/4}} + \frac{10\Gamma_w^2}{T^{1/2}} + \frac{3G\Gamma_w(8+2\delta-\delta^2)}{\delta^2 T^{3/4}} + \frac{\Gamma_w^2}{2T^{3/4}}. \tag{5}$$

Remark. Theorem 2 establishes a stationarity bound of $O((\delta^{-1} + (\log m)^{1/4})T^{-1/8})$. Hence, choosing $\delta = \Omega((\log m)^{-1/4})$ preserves the $O((\log m)^{1/4}T^{-1/8})$ rate of AFL-BR while reducing the per-round communication cost. For Top-k with $\delta = k/d$, taking $k = \Theta(d(\log m)^{-1/4})$ reduces the per-round communication cost to $O(dm(\log m)^{-1/4})$, yielding a total communication complexity of $O(dm(\log m)^{7/4}\varepsilon^{-8})$, a factor of $(\log m)^{1/4}$ lower than that of AFL-BR.

4.2 SHARPER GUARANTEE FOR ADDITIVE-AND-IDEMPOTENT COMPRESSORS

For fixed-weight distributed nonconvex optimization, i.e., $\mathbf{q} = m^{-1}\mathbf{1}_m$ in (1), the compression-dependent lower bound on expected squared gradient norm scales as $\Omega(\delta^{-1})$ (Huang et al., 2022). This suggests that the $O(\delta^{-2})$ dependence in Theorem 2 may be improvable and raises the question of whether sharper control of compression errors can improve the δ-dependence. We answer this question affirmatively for additive-and-idempotent compressors with shared randomness (SR).

Definition 3. A randomized compressor satisfying Definition 2 is written as $\mathcal{C}^{(\xi)}(\cdot)$, where ξ denotes its internal randomness. For every realization ξ, we define the following two properties:

1. $\mathcal{C}$ is *additive* if $\mathcal{C}^{(\xi)}\left(\sum_{i=1}^{n}\mathbf{u}_i\right) = \sum_{i=1}^{n}\mathcal{C}^{(\xi)}(\mathbf{u}_i)$ for any $n \in \mathbb{N}^+$ and $\mathbf{u}_1, \ldots, \mathbf{u}_n \in \mathbb{R}^d$.
2. $\mathcal{C}$ is *idempotent* if $\mathcal{C}^{(\xi)}\big(\mathcal{C}^{(\xi)}(\mathbf{x})\big) = \mathcal{C}^{(\xi)}(\mathbf{x})$ for any $\mathbf{x} \in \mathbb{R}^d$.

We denote the class of compressors satisfying both properties by $\mathcal{C}_{\mathrm{AI}}$.

Examples of compressors in $\mathcal{C}_{\mathrm{AI}}$ are detailed in Appendix B.1. In particular, unscaled Rand-k is in $\mathcal{C}_{\mathrm{AI}}$ with $\delta = k/d$. For randomized compressors, SR means that the server and all workers use the same compressor realization in each round. We note that SR can be implemented without per-round coordination or additional per-round communication, with details provided in Appendix B.2.

Key observation. When AFL-COM uses a compressor $\mathcal{C} \in \mathcal{C}_{\mathrm{AI}}$ with SR, the *downlink compression introduces no error*. Indeed, for all $t \in [T]$, the uplink aggregate $\widehat{\mathbf{g}}_{w,t}$ satisfies

$$\begin{aligned}\mathcal{C}^{(\xi_t)}(\widehat{\mathbf{g}}_{w,t}) =& \mathcal{C}^{(\xi_t)}\left(\sum_{i=1}^{m}\mathcal{C}^{(\xi_t)}\left(\mathbf{e}_t^{(i)} + q_{t,i}\mathbf{g}_{w,t}^{(i)}\right)\right) \overset{\text{(additivity)}}{=} \mathcal{C}^{(\xi_t)}\left(\mathcal{C}^{(\xi_t)}\left(\sum_{i=1}^{m}\mathbf{e}_t^{(i)} + \sum_{i=1}^{m}q_{t,i}\mathbf{g}_{w,t}^{(i)}\right)\right) \\ \overset{\text{(idempotence)}}{=}& \mathcal{C}^{(\xi_t)}\left(\sum_{i=1}^{m}\mathbf{e}_t^{(i)} + \sum_{i=1}^{m}q_{t,i}\mathbf{g}_{w,t}^{(i)}\right) \overset{\text{(additivity)}}{=} \sum_{i=1}^{m}\mathcal{C}^{(\xi_t)}\left(\mathbf{e}_t^{(i)} + q_{t,i}\mathbf{g}_{w,t}^{(i)}\right) = \widehat{\mathbf{g}}_{w,t}.\end{aligned}$$

Since $\widehat{\mathbf{e}}_1 = \mathbf{0}_d$, the server-side EF recursion yields $\widehat{\mathbf{e}}_t = \mathbf{0}_d$ for all $t \in [T]$. Hence, the downlink introduces no additional compression residual, as formalized in Lemma 6. This sharper control improves the dependence on δ, leading to the following convergence guarantee.

Theorem 3. Suppose Assumptions 1–4 hold with $m \geq 2$ and $T \geq 9$. Let $\Delta_\Phi = \Phi_{1/(2L)}(\mathbf{w}_1) - \Phi_*$. For Algorithm 2, set $B = \lceil \sqrt{T} \rceil$, $\eta_w = (8LT^{3/4})^{-1}$, and $\eta_q = 2\sqrt{\ln m}/(\sigma_q T^{1/4})$. If $\mathcal{C} \in \mathcal{C}_{\mathrm{AI}}$ and SR is used, then

$$\mathbb{E}\left[\left\|\nabla \Phi_{1/(2L)}(\mathbf{w}_r)\right\|_2^2\right] \leq \frac{32L\Delta_\Phi}{T^{1/4}} + \frac{16G\Gamma_w}{\delta T^{1/4}} + \frac{16L\sigma_q\sqrt{\ln m}}{T^{1/4}} + \frac{10\Gamma_w^2}{T^{1/2}} + \frac{6G\Gamma_w}{\delta T^{3/4}} + \frac{\Gamma_w^2}{2T^{3/4}}.$$

Remark. Theorem 3 implies a stationarity bound of $O((\delta^{-1/2} + (\log m)^{1/4})T^{-1/8})$, improving the dependence on δ from δ^{-1} in Theorem 2 to $\delta^{-1/2}$. Hence, choosing $\delta = \Omega((\log m)^{-1/2})$ preserves the $O((\log m)^{1/4}T^{-1/8})$ rate of AFL-BR. For Rand-k with SR, taking $k = \Theta(d(\log m)^{-1/2})$ reduces the communication cost per round to $O(dm(\log m)^{-1/2})$ and yields a communication complexity of $O(dm(\log m)^{3/2}\varepsilon^{-8})$, a factor of $(\log m)^{1/2}$ lower than that of AFL-BR.

5 Experiments

We evaluate the two components of our communication-efficiency gains, namely, reducing the number of synchronization rounds and reducing the per-round communication cost through compression.

5.1 Scaling with the Number of Workers

We first examine how the number of workers affects optimization efficiency. We use FEMNIST (Caldas et al., 2018), which provides naturally heterogeneous writer-level partitions with 62 output classes, and consider $m \in \{20, 50, 100, 500, 1000\}$. For each m, we rank writers by their number of local samples and select the top-m writers, with each writer's data randomly split 80/20 into training and test sets.

We compare AFL-BR with SGDA and DRFA under full participation, with DRFA performing three local updates per synchronization. All methods train a two-layer CNN from scratch, with more details provided in Appendix D.1. For each m, we choose a shared target worst-worker test accuracy for all methods and record the first checkpoint reaching that target. We report the mean over five runs, with error bars showing one standard deviation. As shown in Figure 1, AFL-BR and DRFA are comparable for $m = 20$ and $m = 50$, while AFL-BR reaches the target in fewer synchronization rounds than both baselines for $m = 100$, 500, and 1000. These large-m results are consistent with the improved m-dependence in the number of synchronization rounds achieved by AFL-BR, demonstrating its efficiency as the federation scales.

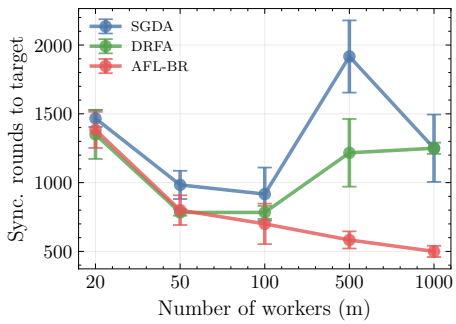


Figure 1: Synchronization rounds versus m on FEMNIST.

5.2 Communication Efficiency with Compression

We evaluate AFL-BR and AFL-Com on Fashion-MNIST (Xiao et al., 2017) and CIFAR-10 (Krizhevsky & Hinton, 2009) with $m = 10$ workers and label-skewed data. We instantiate Algorithm 2 with Top-k and Rand-k, denoted by AFL-Com(T) and AFL-Com(R), retaining 30% and 10% of the coordinates, respectively. The Rand-k variant uses shared randomness across all workers and the server. The baselines include AFL (Mohri et al., 2019), equivalent here to applying SGDA to problem (1), and DRFA (Deng et al., 2020). We also include FedAvg (McMahan et al., 2017) as a standard FL baseline that minimizes the average local loss. FedAvg and DRFA perform three local updates per synchronization, whereas AFL, AFL-BR, and AFL-Com perform one model update. Further details and results are provided in Appendix D.2. Figure 2 reports worst-worker test accuracy against update rounds, synchronization rounds, and cumulative communication cost. Results are averaged over five runs, with shaded regions indicating one standard deviation.

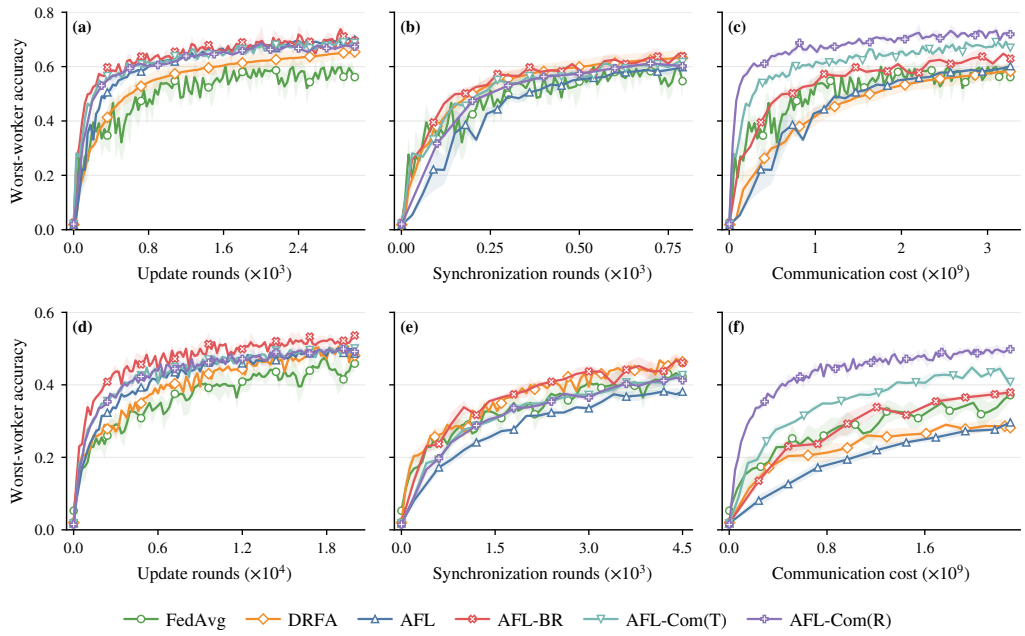


Figure 2: **Worst-worker test accuracy.** The top and bottom rows correspond to Fashion-MNIST and CIFAR-10, respectively. From left to right, the columns report performance against update rounds, synchronization rounds, and cumulative communication cost.

As shown in Figure 2(a) and (d), the AFL methods generally achieve higher worst-worker test accuracy than FEDAVG, consistent with their objective of minimizing the worst-worker risk. Among the AFL methods, AFL-BR reaches a given worst-worker accuracy in fewer update rounds than AFL and DRFA, indicating faster convergence, while AFL-COM(T) and AFL-COM(R) closely track AFL-BR despite compression. Panels (b) and (e) of Figure 2 show that AFL-BR reaches a given accuracy in fewer synchronization rounds than AFL and remains competitive with or better than DRFA. Despite a less favorable synchronization-round bound than its fully synchronized specialization, DRFA improves over AFL in practice through multiple local updates. Moreover, AFL-BR achieves comparable or better efficiency through faster convergence with only one update per synchronization.

We finally compare the performance against cumulative communication cost. Each Top-k message costs $k + k\lceil \log_2 d \rceil / 32$ float-equivalents, accounting for values and indices, while Rand-k with SR costs only k, as the coordinate indices are reconstructed locally using SR. Panels (c) and (f) of Figure 2 show that both AFL-COM variants attain comparable worst-worker accuracy at substantially lower communication cost than the uncompressed AFL baselines. AFL-COM(R) is the most communication-efficient on both datasets, consistent with our sharper analysis for additive-and-idempotent compressors with SR, where the downlink compression residual vanishes.

Overall, panels (a), (b), (d), and (e) of Figure 2 illustrate the benefit of faster convergence, with AFL-BR requiring fewer update and synchronization rounds to attain a given accuracy, while panels (c) and (f) show that AFL-COM further reduces the communication cost through compression.

6 CONCLUSION

We improve the communication efficiency of AFL by reducing both the number of synchronization rounds required for convergence and the communication cost per round. We first propose AFL-BR, which reduces the worker dependence of existing guarantees from polynomial to logarithmic order. Building on it, we develop AFL-COM by introducing bidirectional compression with EF for dual-weighted local gradients. We establish its convergence under general δ-approximate compressors and further obtain a sharper dependence on δ for additive-and-idempotent compressors with SR. Experiments suggest substantial communication savings with competitive worst-worker performance.

AI USE STATEMENT

Generative AI tools were used to assist with language refinement and manuscript organization. All AI-assisted edits were reviewed and verified by the authors, who take full responsibility for the final content.

REFERENCES

Jayadev Acharya, Clément L. Canonne, and Himanshu Tyagi. Inference under information constraints ii: Communication constraints and shared randomness. *IEEE Transactions on Information Theory*, 66(12):7856–7877, 2020.

Dan Alistarh, Demjan Grubic, Jerry Li, Ryota Tomioka, and Milan Vojnovic. Qsgd: Communication-efficient sgd via gradient quantization and encoding. In *Advances in Neural Information Processing Systems 30*, pp. 1707–1718, 2017.

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.

Debraj Basu, Deepesh Data, Can Karakus, and Suhas N. Diggavi. Qsparse-local-sgd: Distributed sgd with quantization, sparsification, and local computations. *IEEE Journal on Selected Areas in Information Theory*, 1(1):217–226, 2020.

Aleksandr Beznosikov, Peter Richtárik, Michael Diskin, Max Ryabinin, and Alexander Gasnikov. Distributed methods with compressed communication for solving variational inequalities, with theoretical guarantees. In *Advances in Neural Information Processing Systems 35*, pp. 14013–14029, 2022.

Aleksandr Beznosikov, Samuel Horváth, Peter Richtárik, and Mher Safaryan. On biased compression for distributed learning. *Journal of Machine Learning Research*, 24(276):1–50, 2023.

Sebastian Caldas, Sai Meher Karthik Duddu, Peter Wu, Tian Li, Jakub Konečný, H. Brendan McMahan, Virginia Smith, and Ameet Talwalkar. Leaf: A benchmark for federated settings. *ArXiv e-prints*, arXiv: 1812.01097, 2018.

Yair Carmon and Danielle Hausler. Distributionally robust optimization via ball oracle acceleration. In *Advances in Neural Information Processing Systems 35*, pp. 35866–35879, 2022.

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

Chuyan Chen, Yutong He, Pengrui Li, Weichen Jia, and Kun Yuan. Greedy low-rank gradient compression for distributed learning with convergence guarantees. *IEEE Transactions on Signal Processing*, 74:937–951, 2026.

Liyu Chen, Haipeng Luo, and Chen-Yu Wei. Impossible tuning made possible: A new expert algorithm and its applications. In *Proceedings of Thirty Fourth Conference on Learning Theory*, volume 134, pp. 1216–1259, 2021.

Richard J. Chen, Judy J. Wang, Drew F. K. Williamson, Tiffany Y. Chen, Jana Lipkova, Ming Y. Lu, Sharifa Sahai, and Faisal Mahmood. Algorithmic fairness in artificial intelligence for medicine and healthcare. *Nature Biomedical Engineering*, 7(6):719–742, 2023.

Damek Davis and Dmitriy Drusvyatskiy. Stochastic model-based minimization of weakly convex functions. *SIAM Journal on Optimization*, 29(1):207–239, 2019.

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

Kaja Gruntkowska, Alexander Tyurin, and Peter Richtárik. EF21-p and friends: Improved theoretical communication complexity for distributed optimization with bidirectional compression. In *Proceedings of the 40th International Conference on Machine Learning*, pp. 11761–11807, 2023.

Kaja Gruntkowska, Alexander Tyurin, and Peter Richtárik. Improving the worst-case bidirectional communication complexity for nonconvex distributed optimization under function similarity. In *Neural Information Processing Systems 37*, 2024.

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

Farzin Haddadpour, Mohammad Mahdi Kamani, Aryan Mokhtari, and Mehrdad Mahdavi. Federated learning with compression: Unified analysis and sharp guarantees. In *Proceedings of The 24th International Conference on Artificial Intelligence and Statistics*, pp. 2350–2358, 2021.

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.

Jenny Hamer, Mehryar Mohri, and Ananda Theertha Suresh. FedBoost: A communication-efficient algorithm for federated learning. In *Proceedings of the 37th International Conference on Machine Learning*, pp. 3973–3983, 2020.

Elad Hazan. Introduction to online convex optimization. *Found. Trends Optim.*, 2(3–4):157–325, 2016.

Mark Herbster and Manfred K. Warmuth. Tracking the best expert. *Machine Learning*, 32(2): 151–178, 1998.

Samuel Horváth, Chen-Yu Ho, Ludovit Horvath, Atal Narayan Sahu, Marco Canini, and Peter Richtárik. Natural compression for distributed deep learning. In *Proceedings of Mathematical and Scientific Machine Learning*, pp. 129–141, 2022.

Xinmeng Huang, Yiming Chen, Wotao Yin, and Kun Yuan. Lower bounds and nearly optimal algorithms in distributed learning with communication compression. In *Advances in Neural Information Processing Systems 35*, pp. 18955–18969, 2022.

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

Peter Kairouz, H. Brendan McMahan, Brendan Avent, Aurélien Bellet, Mehdi Bennis, Arjun Nitin Bhagoji, Kallista Bonawitz, Zachary Charles, Graham Cormode, Rachel Cummings, Rafael G. L. D'Oliveira, Hubert Eichner, Salim El Rouayheb, David Evans, Josh Gardner, Zachary Garrett, Adrià Gascón, Badih Ghazi, Phillip B. Gibbons, Marco Gruteser, Zaid Harchaoui, Chaoyang He, Lie He, Zhouyuan Huo, Ben Hutchinson, Justin Hsu, Martin Jaggi, Tara Javidi, Gauri Joshi, Mikhail Khodak, Jakub Konečný, Aleksandra Korolova, Farinaz Koushanfar, Sanmi Koyejo, Tancrède Lepoint, Yang Liu, Prateek Mittal, Mehryar Mohri, Richard Nock, Ayfer Özgür, Rasmus Pagh, Hang Qi, Daniel Ramage, Ramesh Raskar, Mariana Raykova, Dawn Song, Weikang Song, Sebastian U. Stich, Ziteng Sun, Ananda Theertha Suresh, Florian Tramèr, Praneeth Vepakomma, Jianyu Wang, Li Xiong, Zheng Xu, Qiang Yang, Felix X. Yu, Han Yu, and Sen Zhao. Advances and open problems in federated learning. *Foundations and Trends® in Machine Learning*, 14: 1–210, 2021.

Sai Praneeth Karimireddy, Quentin Rebjock, Sebastian Stich, and Martin Jaggi. Error feedback fixes SignSGD and other gradient compression schemes. In *Proceedings of the 36th International Conference on Machine Learning*, pp. 3252–3261, 2019.

Alex Krizhevsky and Geoffrey Hinton. Learning multiple layers of features from tiny images. Technical Report 0, University of Toronto, Toronto, Ontario, 2009.

Mounssif Krouka, Chaouki Ben Issaid, and Mehdi Bennis. Distributionally robust federated learning with client drift minimization. *IEEE Transactions on Machine Learning in Communications and Networking*, 4:438–456, 2026.

Li Li, Yuxi Fan, Mike Tse, and Kuo-Yi Lin. A review of applications in federated learning. *Computers & Industrial Engineering*, 149:106854, 2020.

Zhize Li, Haoyu Zhao, Boyue Li, and Yuejie Chi. Soteriafl: A unified framework for private federated learning with communication compression. In *Advances in Neural Information Processing Systems 35*, pp. 4285–4300, 2022.

Tianyi Lin, Chi Jin, and Michael Jordan. On gradient descent ascent for nonconvex-concave minimax problems. In *Proceedings of the 37th International Conference on Machine Learning*, pp. 6083–6093, 2020.

Guodong Long, Yue Tan, Jing Jiang, and Chengqi Zhang. Federated learning for open banking. In *Federated Learning: Privacy and Incentive*, pp. 240–254. Springer International Publishing, 2020.

Shiyin Lu and Lijun Zhang. Adaptive and efficient algorithms for tracking the best expert. *ArXiv e-prints*, arXiv: 1909.02187, 2019.

Brendan McMahan, Eider Moore, Daniel Ramage, Seth Hampson, and Blaise Aguera y Arcas. Communication-Efficient Learning of Deep Networks from Decentralized Data. In *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics*, pp. 1273–1282, 2017.

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.

A. Nemirovski, A. Juditsky, G. Lan, and A. Shapiro. Robust stochastic approximation approach to stochastic programming. *SIAM Journal on Optimization*, 19(4):1574–1609, 2009.

Valery Parfenov, Nail Bashirov, Daniil Medyakov, Dmitry Bylinkin, and Aleksandr Beznosikov. Unlocking the potential of weighting methods in federated learning through communication compression. In *The Fourteenth International Conference on Learning Representations*, 2026.

Hassan Rafique, Mingrui Liu, Qihang Lin, and Tianbao Yang. Weakly-convex–concave min–max optimization: provable algorithms and applications in machine learning. *Optimization Methods and Software*, 37(3):1087–1121, 2022.

Amirhossein Reisizadeh, Aryan Mokhtari, Hamed Hassani, Ali Jadbabaie, and Ramtin Pedarsani. Fedpaq: A communication-efficient federated learning method with periodic averaging and quantization. In *Proceedings of the Twenty Third International Conference on Artificial Intelligence and Statistics*, pp. 2021–2031, 2020.

Peter Richtárik, Igor Sokolov, and Ilyas Fatkhullin. EF21: A new, simpler, theoretically better, and practically faster error feedback. In *Advances in Neural Information Processing Systems 34*, pp. 4384–4396, 2021.

Jae Ro, Mingqing Chen, Rajiv Mathews, Mehryar Mohri, and Ananda Theertha Suresh. Communication-efficient agnostic federated averaging. In *Interspeech 2021*, pp. 871–875, 2021.

Mher Safaryan, Egor Shulgin, and Peter Richtárik. Uncertainty principle for communication compression in distributed and federated learning and the search for an optimal compressor. *Information and Inference: A Journal of the IMA*, 11(2):557–580, 2022.

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.

Felix Sattler, Simon Wiedemann, Klaus-Robert Müller, and Wojciech Samek. Robust and communication-efficient federated learning from non-i.i.d. data. *IEEE Transactions on Neural Networks and Learning Systems*, 31(9):3400–3413, 2020.

Frank Seide, Hao Fu, Jasha Droppo, Gang Li, and Dong Yu. 1-bit stochastic gradient descent and its application to data-parallel distributed training of speech dnns. In *Interspeech 2014*, pp. 1058–1062, 2014.

Wei Shen, Minhui Huang, Jiawei Zhang, and Cong Shen. Stochastic smoothed gradient descent ascent for federated minimax optimization. In *Proceedings of The 27th International Conference on Artificial Intelligence and Statistics*, pp. 3988–3996, 2024.

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

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.

Shay Vargaftik, Ran Ben Basat, Amit Portnoy, Gal Mendelson, Yaniv Ben Itzhak, and Michael Mitzenmacher. EDEN: Communication-efficient and robust distributed mean estimation for federated learning. In *Proceedings of the 39th International Conference on Machine Learning*, pp. 21984–22014, 2022.

Roman Vershynin. *High-dimensional probability: An introduction with applications in data science*. Cambridge university press, 2018.

Yujia Wang, Lu Lin, and Jinghui Chen. Communication-efficient adaptive federated learning. In *Proceedings of the 39th International Conference on Machine Learning*, pp. 22802–22838, 2022.

Jianqiao Wangni, Jialei Wang, Ji Liu, and Tong Zhang. Gradient sparsification for communication-efficient distributed optimization. In *Advances in Neural Information Processing Systems 31*, pp. 1306–1316, 2018.

Han Xiao, Kashif Rasul, and Roland Vollgraf. Fashion-mnist: a novel image dataset for benchmarking machine learning algorithms. *ArXiv e-prints*, arXiv:1708.07747, 2017.

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), 2023.

Yaodong Yu, Sai Praneeth Karimireddy, Yi Ma, and Michael I. Jordan. Scaff-pd: Communication efficient fair and robust federated learning. *ArXiv e-prints*, arXiv:2307.13381, 2023.

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.

Sihui Zheng, Cong Shen, and Xiang Chen. Design and analysis of uplink and downlink communications for federated learning. *IEEE Journal on Selected Areas in Communications*, 39(7): 2150–2167, 2021.

A Additional Related Work

For convex AFL, Mohri et al. (2019) achieve an $O(T^{-1/2})$ optimization-error rate. DRFA (Deng et al., 2020) uses local updates to reduce the number of synchronization rounds to $O(T^{3/4})$, but slows the optimization-error rate to $O(T^{-3/8})$. Consequently, both methods require $O(\varepsilon^{-2})$ synchronization rounds to reach an ε-accurate solution. Other convex variants incorporate worker-drift correction (Krouka et al., 2026), use local updates to improve empirical communication efficiency without convergence guarantees (Ro et al., 2021), or adopt alternative communication models based on pretrained predictor ensembles (Hamer et al., 2020). A complementary line of work studies constrained, regularized, or decentralized variants of AFL and federated GDRO (Issaid et al., 2022; Yu et al., 2023; Guo & Yang, 2024; Zecchin et al., 2023).

B Additional Discussions

B.1 Examples of Compressors

We provide several representative compressors considered in this paper. More extensive collections of biased and unbiased compressors, together with their contraction or variance parameters, can be found in Safaryan et al. (2022); Beznosikov et al. (2023) and the references therein.

We first give standard examples satisfying Definition 2.

- *Top-k sparsification* (Stich et al., 2018) is defined by $\mathcal{C}_{\text{top-}k}(\mathbf{x}) = \sum_{j \in S_k(\mathbf{x})} x_j \mathbf{e}_j$, where $S_k(\mathbf{x})$ indexes the k coordinates of $\mathbf{x}$ with the largest absolute values. Since the retained coordinates contain at least a k/d fraction of the squared Euclidean norm, $\|\mathcal{C}_{\text{top-}k}(\mathbf{x}) - \mathbf{x}\|_2^2 \le (1 - k/d)\|\mathbf{x}\|_2^2$. Hence, Definition 2 holds with $\delta = k/d$. The support $S_k(\mathbf{x})$ depends on the input, so Top-k is nonlinear and is not additive in general.
- *Scaled-sign compression* (Karimireddy et al., 2019) is defined by $\mathcal{C}_{\text{sign}}(\mathbf{x}) = \frac{\|\mathbf{x}\|_1}{d} \operatorname{sign}(\mathbf{x})$, with $\operatorname{sign}(0) = 0$. A direct calculation gives $\|\mathcal{C}_{\text{sign}}(\mathbf{x}) - \mathbf{x}\|_2^2 \le \|\mathbf{x}\|_2^2 - \|\mathbf{x}\|_1^2/d \le (1 - 1/d)\,\|\mathbf{x}\|_2^2$, where the second inequality follows from $\|\mathbf{x}\|_1 \ge \|\mathbf{x}\|_2$. Thus, it is a δ-approximate compressor with $\delta = 1/d$.

The class $\mathcal{C}_{\text{AI}}$ additionally requires additivity and idempotence for every fixed realization of the compressor randomness. Two useful examples are randomized projection operators.

- *Unscaled Rand-k sparsification* (Stich et al., 2018). Let $S \subseteq [d]$ be sampled uniformly from all subsets of cardinality k, and define $\mathcal{C}^{(S)}_{\text{rand-}k}(\mathbf{x}) = P_S\mathbf{x} = \sum_{j \in S} x_j \mathbf{e}_j$. For every fixed S, P_S is linear and satisfies $P_S^2 = P_S$. Hence, the compressor is additive and idempotent pathwise. Moreover, $\mathbb{E}_S[\|P_S\mathbf{x} - \mathbf{x}\|_2^2] = (1 - k/d)\|\mathbf{x}\|_2^2$, so Definition 2 holds with $\delta = k/d$.
  We emphasize that this is the *unscaled* contractive version of Rand-k. The commonly used unbiased scaling $(d/k)P_S\mathbf{x}$ is not idempotent unless $k = d$. Under SR, all workers and the server use the same subset S within a communication round, so the coordinate indices need not be transmitted.
- *Random rank-r orthogonal projection* (Vershynin, 2018). Let $U \in \mathbb{R}^{d \times r}$ have orthonormal columns spanning a uniformly random r-dimensional subspace, and define $\mathcal{C}^{(U)}_{\text{proj}}(\mathbf{x}) = UU^\top\mathbf{x}$. For every fixed U, the matrix $P_U = UU^\top$ is linear and satisfies $P_U^2 = P_U$. Therefore, the compressor is additive and idempotent. By rotational symmetry, $\mathbb{E}_U[P_U] = \frac{r}{d} I_d$, and consequently $\mathbb{E}_U[\|UU^\top\mathbf{x} - \mathbf{x}\|_2^2] = (1 - r/d)\|\mathbf{x}\|_2^2$. Thus, Definition 2 holds with $\delta = r/d$. With SR, the subspace U can be reconstructed locally from the common random seed, so only the r coefficients $U^\top\mathbf{x}$ need to be transmitted. The receiver can then reconstruct $UU^\top\mathbf{x}$ locally.

B.2 Implementation of Shared Randomness

Shared randomness (SR) requires the server and all workers to use the same compressor realization in each communication round. In practice, SR can be implemented without per-round coordination. Before training, the server and all workers synchronize a single random seed. Then, in each round, they locally generate a common realization using a fixed deterministic rule (Acharya et al., 2020) or synchronized pseudorandom number generators (Vargaftik et al., 2022). Similar SR mechanisms have been used in communication-efficient distributed learning, for example, by Chen et al. (2026). Therefore, SR introduces no per-round coordination or additional per-round communication. The only overhead is the initial transmission of a scalar seed, which is negligible compared with transmitting high-dimensional model updates.

C Details for Table 1

This section explains two technical details used in Table 1.

C.1 SPECIALIZATION OF SGDA TO AFL

The SGDA guarantee in (Lin et al., 2020, Theorem 4.9) applies to general nonconvex–concave minimax problems. We specialize this result to the centralized GDRO objective $\phi(\mathbf{w}, \mathbf{q}) = \sum_{i=1}^{m} q_i R_i(\mathbf{w})$ with $\mathbf{q} \in \Delta_m$ and make explicit the dependence hidden in its Euclidean joint smoothness and stochastic-oracle variance parameters. Unlike the DRFA analysis discussed in Appendix C.2, this specialization does not require a uniform bound on $\|\mathbf{R}(\mathbf{w})\|_2$. Its dependence on m instead arises from the Euclidean geometry of the m-dimensional dual weights.

Primal Lipschitzness and joint smoothness. Let G_ϕ denote the uniform Lipschitz constant of $\phi(\cdot, \mathbf{q})$ and let L_{joint} denote the Euclidean joint smoothness constant of ϕ, corresponding to the parameter denoted by ℓ in (Lin et al., 2020, Assumption 4.6). By Assumption 3, for every $\mathbf{q} \in \Delta_m$,

$$\|\nabla_{\mathbf{w}} \phi(\mathbf{w}, \mathbf{q})\|_2 = \left\| \sum_{i=1}^{m} q_i \nabla R_i(\mathbf{w}) \right\|_2 \leq \sum_{i=1}^{m} q_i \|\nabla R_i(\mathbf{w})\|_2 \leq G.$$

Thus, $\phi(\cdot, \mathbf{q})$ is G-Lipschitz uniformly over $\mathbf{q} \in \Delta_m$, and one may take $G_\phi = G$.

Let $\mathbf{R}(\mathbf{w}) = (R_1(\mathbf{w}), \ldots, R_m(\mathbf{w}))^\top$. The joint gradient is

$$\nabla \phi(\mathbf{w}, \mathbf{q}) = \begin{bmatrix} \sum_{i=1}^{m} q_i \nabla R_i(\mathbf{w}) \\ \mathbf{R}(\mathbf{w}) \end{bmatrix}. \tag{6}$$

Since $\|\nabla R_i(\mathbf{w})\|_2 \leq G$ under Assumption 3, each R_i is G-Lipschitz. Therefore,

$$\|\mathbf{R}(\mathbf{w}) - \mathbf{R}(\mathbf{w}')\|_2 = \left(\sum_{i=1}^{m} |R_i(\mathbf{w}) - R_i(\mathbf{w}')|^2 \right)^{1/2} \leq G\sqrt{m}\, \|\mathbf{w} - \mathbf{w}'\|_2.$$

For the primal component of (6), Assumptions 1 and 3 imply

$$\begin{aligned} &\left\| \sum_{i=1}^{m} q_i \nabla R_i(\mathbf{w}) - \sum_{i=1}^{m} q_i' \nabla R_i(\mathbf{w}') \right\|_2 \\ &\leq \left\| \sum_{i=1}^{m} q_i \big(\nabla R_i(\mathbf{w}) - \nabla R_i(\mathbf{w}') \big) \right\|_2 + \left\| \sum_{i=1}^{m} (q_i - q_i') \nabla R_i(\mathbf{w}') \right\|_2 \\ &\leq L \|\mathbf{w} - \mathbf{w}'\|_2 + G \|\mathbf{q} - \mathbf{q}'\|_1 \\ &\leq L \|\mathbf{w} - \mathbf{w}'\|_2 + G\sqrt{m}\, \|\mathbf{q} - \mathbf{q}'\|_2. \end{aligned}$$

Combining the two inequalities gives

$$\|\nabla \phi(\mathbf{w}, \mathbf{q}) - \nabla \phi(\mathbf{w}', \mathbf{q}')\|_2^2 \leq 2\big(L^2 + mG^2\big) \left(\|\mathbf{w} - \mathbf{w}'\|_2^2 + \|\mathbf{q} - \mathbf{q}'\|_2^2 \right).$$

Hence, one may take $L_{\text{joint}} = \sqrt{2(L^2 + mG^2)} = O(\sqrt{m})$.

We next show that the $\sqrt{m}$ dependence of L_{joint} cannot generally be removed under the Euclidean geometry used by SGDA. Since $\nabla_{\mathbf{q}} \phi(\mathbf{w}, \mathbf{q}) = \mathbf{R}(\mathbf{w})$, joint L_{joint}-smoothness implies

$$\begin{aligned} \|\mathbf{R}(\mathbf{w}) - \mathbf{R}(\mathbf{w}')\|_2 &= \|\nabla_{\mathbf{q}} \phi(\mathbf{w}, \mathbf{q}) - \nabla_{\mathbf{q}} \phi(\mathbf{w}', \mathbf{q})\|_2 \\ &\leq \|\nabla \phi(\mathbf{w}, \mathbf{q}) - \nabla \phi(\mathbf{w}', \mathbf{q})\|_2 \leq L_{\text{joint}} \|\mathbf{w} - \mathbf{w}'\|_2, \end{aligned}$$

and therefore

$$L_{\text{joint}} \geq \sup_{\mathbf{w} \neq \mathbf{w}'} \frac{\|\mathbf{R}(\mathbf{w}) - \mathbf{R}(\mathbf{w}')\|_2}{\|\mathbf{w} - \mathbf{w}'\|_2}.$$

To see that the right-hand side can be of order $\sqrt{m}$, let $\mathbf{a} \in \mathbb{R}^d$ satisfy $\|\mathbf{a}\|_2 = 1$ and define $R_i(\mathbf{w}) = \frac{1}{2} + \frac{1}{2} \tanh\left(2G\langle \mathbf{a}, \mathbf{w} \rangle\right)$ for $i \in [m]$. Then $R_i(\mathbf{w}) \in (0, 1)$, $\nabla R_i(\mathbf{w}) = G \operatorname{sech}^2\left(2G\langle \mathbf{a}, \mathbf{w} \rangle\right) \mathbf{a}$, and therefore $\|\nabla R_i(\mathbf{w})\|_2 \leq G$. Moreover, $\|\nabla^2 R_i(\mathbf{w})\|_2 \leq 4G^2$, so the smoothness constant of each local objective is independent of m. Taking $\mathbf{w}' = \mathbf{0}_d$ and $\mathbf{w} = h\mathbf{a}$ gives

$$\frac{\|\mathbf{R}(h\mathbf{a}) - \mathbf{R}(\mathbf{0}_d)\|_2}{\|h\mathbf{a}\|_2} = \sqrt{m}\, \frac{|\tanh(2Gh)|}{2|h|} \longrightarrow G\sqrt{m} \qquad \text{as } h \to 0.$$

Thus, $L_{\text{joint}} = \Omega(\sqrt{m})$ in the worst case. Together with the upper bound $L_{\text{joint}} = O(\sqrt{m})$, this shows that $L_{\text{joint}} = \Theta(\sqrt{m})$ even when all local smoothness and gradient bounds are independent of m.

Variance of the joint stochastic oracle. At a centralized SGDA iteration, consider the unbiased joint stochastic oracle

$$\widehat{\nabla\phi}(\mathbf{w},\mathbf{q};\mathbf{z}) = \begin{bmatrix} \sum_{i=1}^m q_i \nabla \ell(\mathbf{w};\mathbf{z}^{(i)}) \\ \widehat{\mathbf{R}}(\mathbf{w};\mathbf{z}) \end{bmatrix},$$

where $\widehat{\mathbf{R}}(\mathbf{w};\mathbf{z}) = [\ell(\mathbf{w};\mathbf{z}^{(1)}),\dots,\ell(\mathbf{w};\mathbf{z}^{(m)})]^\top$. By convexity of the squared norm and Assumption 3,

$$\mathbb{E}_{\mathbf{z}}\left[\left\|\sum_{i=1}^m q_i\big(\nabla\ell(\mathbf{w};\mathbf{z}^{(i)}) - \nabla R_i(\mathbf{w})\big)\right\|_2^2\right] \le \sum_{i=1}^m q_i \mathbb{E}_{\mathbf{z}^{(i)}}\left[\|\nabla\ell(\mathbf{w};\mathbf{z}^{(i)}) - \nabla R_i(\mathbf{w})\|_2^2\right] \le \sigma_w^2.$$

No independence across workers is required for this inequality. Moreover, Assumption 4 implies $\mathbb{E}_{\mathbf{z}^{(i)}\sim\mathcal{P}_i}[(\ell(\mathbf{w};\mathbf{z}^{(i)}) - R_i(\mathbf{w}))^2] \le \sigma_q^2$ for every $i \in [m]$. Hence,

$$\mathbb{E}_{\mathbf{z}}\left[\left\|\widehat{\mathbf{R}}(\mathbf{w};\mathbf{z}) - \mathbf{R}(\mathbf{w})\right\|_2^2\right] = \sum_{i=1}^m \mathbb{E}_{\mathbf{z}^{(i)}\sim\mathcal{P}_i}\left[(\ell(\mathbf{w};\mathbf{z}^{(i)}) - R_i(\mathbf{w}))^2\right] \le m\sigma_q^2.$$

Consequently, the variance parameter σ_{joint}^2 of SGDA can be chosen as $\sigma_{\text{joint}}^2 \le \sigma_w^2 + m\sigma_q^2$.

Centralized iteration complexity. Using the notation above, the stochastic joint-oracle complexity in (Lin et al., 2020, Theorem 4.9) can be written as

$$N_{\text{SGDA}}(\varepsilon) = O\left(\left(\frac{L_{\text{joint}}^3(G_\phi^2 + \sigma_{\text{joint}}^2)D_\Delta^2\widehat{\Delta}_\Phi}{\varepsilon^6} + \frac{L_{\text{joint}}^3 D_\Delta^2 \widehat{\Delta}_0}{\varepsilon^4}\right)\max\left\{1, \frac{\sigma_{\text{joint}}^2}{\varepsilon^2}\right\}\right), \tag{7}$$

where D_Δ is the Euclidean diameter of Δ_m, $\widehat{\Delta}_\Phi = \Phi_{1/(2L_{\text{joint}})}(\mathbf{w}_0) - \inf_{\mathbf{w}} \Phi_{1/(2L_{\text{joint}})}(\mathbf{w})$, and $\widehat{\Delta}_0 = \Phi(\mathbf{w}_0) - \phi(\mathbf{w}_0,\mathbf{q}_0)$ is the initial dual gap. Since $D_\Delta^2 \le 2$, $L_{\text{joint}}^3 = O(m^{3/2})$ and $\sigma_{\text{joint}}^2 = O(m)$, (7) therefore gives $N_{\text{SGDA}}(\varepsilon) = O\left(m^{7/2}\varepsilon^{-8}\right)$.

The above stationarity criterion is expressed using $\nabla\Phi_{1/(2L_{\text{joint}})}$, whereas the nonconvex results in this paper use $\nabla\Phi_{1/(2L)}$. The following comparison shows that the SGDA stationarity guarantee implies our stationarity criterion up to an absolute constant.

Lemma 1. Let $f:\mathbb{R}^d \to \mathbb{R}$ be ρ-weakly convex. For any $0 < \mu \le \lambda < 1/\rho$, the gradients of its Moreau envelopes satisfy $\|\nabla f_\lambda(\mathbf{x})\|_2 \le \frac{1-\rho\mu}{1-\rho\lambda}\|\nabla f_\mu(\mathbf{x})\|_2$ for all $\mathbf{x}\in\mathbb{R}^d$.

By Lemma 2, Φ is L-weakly convex. Since $L_{\text{joint}} = \sqrt{2(L^2 + mG^2)} \ge L$, we may apply Lemma 1 with $\rho = L$, $\mu = 1/(2L_{\text{joint}})$, and $\lambda = 1/(2L)$. This gives

$$\begin{aligned}\left\|\nabla\Phi_{1/(2L)}(\mathbf{w})\right\|_2 &\le \frac{1 - L/(2L_{\text{joint}})}{1 - L/(2L)}\left\|\nabla\Phi_{1/(2L_{\text{joint}})}(\mathbf{w})\right\|_2 \\ &= \left(2 - \frac{L}{L_{\text{joint}}}\right)\left\|\nabla\Phi_{1/(2L_{\text{joint}})}(\mathbf{w})\right\|_2 \\ &\le 2\left\|\nabla\Phi_{1/(2L_{\text{joint}})}(\mathbf{w})\right\|_2.\end{aligned}$$

Therefore, running SGDA until $\mathbb{E}[\|\nabla\Phi_{1/(2L_{\text{joint}})}(\mathbf{w}_r)\|_2] \le \varepsilon/2$ is sufficient to guarantee $\mathbb{E}[\|\nabla\Phi_{1/(2L)}(\mathbf{w}_r)\|_2] \le \varepsilon$. Replacing ε by $\varepsilon/2$ changes only an absolute constant.

Equivalently, after T centralized SGDA iterations, $\mathbb{E}\left[\left\|\nabla\Phi_{1/(2L)}(\mathbf{w}_r)\right\|_2\right] = O\left(m^{7/16}T^{-1/8}\right)$. When SGDA is implemented in the parameter-server setting, each iteration requires one communication round to aggregate the m local primal gradients and construct the m-dimensional stochastic loss vector. Since each round communicates $O(dm)$ entries, the resulting communication complexity is $O\left(dm^{9/2}\varepsilon^{-8}\right)$. The above rate and the communication complexity are the SGDA entries reported in Table 1.

C.2 Refined DRFA Bounds with Partial Participation

We make explicit the worker and participation dependence of the nonconvex DRFA analysis in Deng et al. (2020). Their notation uses N for the total number of workers and m for the sampling size. Here, m denotes the total number of workers, b the sampling size, and $p = b/m$. We additionally assume $|R_i(\mathbf{w})| \le G_R$, with G_R, σ_q, L, G, and σ_w independent of m and p.

Dual oracle under partial participation. Let $\mathbf{R}(\mathbf{w}) = (R_1(\mathbf{w}), \dots, R_m(\mathbf{w}))^\top = \nabla_{\mathbf{q}}\phi(\mathbf{w}, \mathbf{q})$. At each dual update, DRFA uniformly samples $U_s \subseteq [m]$ with $|U_s| = b$ and constructs

$$\widehat{\mathbf{R}}_p(\mathbf{w}) = \frac{1}{p} \sum_{i \in U_s} \ell(\mathbf{w}; \mathbf{z}_i)\mathbf{e}_i.$$

Since $\Pr(i \in U_s) = p$, we obtain $\mathbb{E}[\widehat{\mathbf{R}}_p(\mathbf{w})] = \mathbf{R}(\mathbf{w})$. Moreover, $G_\lambda^2 := \sup_{\mathbf{w}} \|\mathbf{R}(\mathbf{w})\|_2^2 \le mG_R^2$, and, writing $\xi_i(\mathbf{w}; \mathbf{z}_i) = \ell(\mathbf{w}; \mathbf{z}_i) - R_i(\mathbf{w})$,

$$\begin{aligned} V_p = \mathbb{E}\|\widehat{\mathbf{R}}_p(\mathbf{w}) - \mathbf{R}(\mathbf{w})\|_2^2 &= \left(\frac{1}{p} - 1\right) \|\mathbf{R}(\mathbf{w})\|_2^2 + \frac{1}{p} \sum_{i=1}^{m} \mathbb{E}[\xi_i(\mathbf{w}; \mathbf{z}_i)^2] \\ &\le m \left[\left(\frac{1}{p} - 1\right) G_R^2 + \frac{\sigma_q^2}{p} \right]. \end{aligned}$$

We then define $A_p := G_\lambda^2 + V_p = O\left(m/p\right)$.

Let $\mathbf{w}^{(t)}$ be the virtual averaged model and sample t' uniformly from $\{s\tau+1, \dots, (s+1)\tau\}$. Conditioning on the local trajectories, $\mathbb{E}[\tau\widehat{\mathbf{R}}_p(\mathbf{w}^{(t')})] = \sum_{t=s\tau+1}^{(s+1)\tau} \mathbf{R}(\mathbf{w}^{(t)})$, so the snapshot dual estimator remains unbiased. Applying the projection argument and the grouping argument of Lemmas 8–9 in Deng et al. (2020), the dual contribution is

$$O\left(\gamma\tau A_p + \frac{D_\Lambda^2}{\gamma\tau\sqrt{S}}\right), \qquad S = \frac{T}{\tau}.$$

The remaining primal terms introduce no polynomial dependence on m or p. Indeed, for fixed $\mathbf{q}$, $\|\nabla_{\mathbf{w}}\phi(\mathbf{w}, \mathbf{q}) - \nabla_{\mathbf{w}}\phi(\mathbf{w}', \mathbf{q})\|_2 \le L\|\mathbf{w} - \mathbf{w}'\|_2$, the weighted gradient dissimilarity satisfies $\Gamma \le 4G^2$, and for $\widetilde{\mathbf{w}} \in \arg\min_{\mathbf{u}}\{\Phi(\mathbf{u}) + L\|\mathbf{u} - \mathbf{w}\|_2^2\}$, $\|\widetilde{\mathbf{w}} - \mathbf{w}\|_2 \le G/L$. Thus, up to problem-dependent constants independent of m and p, the proof of Theorem 2 yields

$$\frac{1}{T} \sum_{t=1}^{T} \mathbb{E}\|\nabla\Phi_{1/(2L)}(\mathbf{w}^{(t)})\|_2^2 = O\left(\frac{1}{\eta T} + \eta\tau\sqrt{S} + \eta\tau + \eta + \gamma\tau A_p + \frac{D_\Lambda^2}{\gamma\tau\sqrt{S}}\right). \tag{8}$$

Local updates. Set $\tau = T^{1/4}$, $S = T^{3/4}$, and $\eta = \Theta(T^{-3/4})$. Balancing the two dual terms in (8) gives $\gamma = \Theta(A_p^{-1/2}T^{-7/16})$ and hence

$$\frac{1}{T} \sum_{t=1}^{T} \mathbb{E}\|\nabla\Phi_{1/(2L)}(\mathbf{w}^{(t)})\|_2^2 = O\left(T^{-1/8} + \sqrt{A_p}\,T^{-3/16}\right).$$

For $r \sim \mathrm{Unif}[T]$, Jensen's inequality and $A_p = O(m/p)$ give

$$\mathbb{E}\|\nabla\Phi_{1/(2L)}(\mathbf{w}^{(r)})\|_2 = O\left(T^{-1/16} + \left(\frac{m}{p}\right)^{1/4} T^{-3/32}\right).$$

Therefore,

$$T = O\left(\varepsilon^{-16} + \left(\frac{m}{p}\right)^{8/3} \varepsilon^{-32/3}\right),$$

Since each synchronization stage communicates $O(db) = O(dpm)$ entries, achieving an ε-stationary solution requires $O\left(\varepsilon^{-12} + m^2p^{-2}\varepsilon^{-8}\right)$ synchronization rounds and $O\left(dpm\varepsilon^{-12} + dm^3p^{-1}\varepsilon^{-8}\right)$ total communication.

Synchronization at every step. For $\tau = 1$, we have $S = T$. Choosing $\eta = \Theta(T^{-3/4})$ and $\gamma = \Theta(A_p^{-1/2}T^{-1/4})$ gives

$$\mathbb{E}\|\nabla\Phi_{1/(2L)}(\mathbf{w}^{(r)})\|_2 = O\left(\left(\frac{m}{p}\right)^{1/4} T^{-1/8}\right).$$

Therefore, $T = O\left(m^2 p^{-2} \varepsilon^{-8}\right)$. Since every iteration is a synchronization stage and each stage communicates $O(db) = O(dpm)$ entries, achieving an ε-stationary solution requires $O\left(m^2 p^{-2} \varepsilon^{-8}\right)$ synchronization rounds and $O\left(dm^3 p^{-1} \varepsilon^{-8}\right)$ total communication.

In Table 1, we report the dependence on m and ε for a fixed sampling fraction $p = \Theta(1)$ independent of both quantities. Absorbing the resulting p-dependent constants yields $O(\varepsilon^{-12} + m^2 \varepsilon^{-8})$ synchronization rounds and $O(dm\varepsilon^{-12} + dm^3 \varepsilon^{-8})$ communication for local-update DRFA, and $O(m^2 \varepsilon^{-8})$ synchronization rounds and $O(dm^3 \varepsilon^{-8})$ communication for $\tau = 1$.

D Experimental Details

D.1 Scaling with the Number of Workers

In all experiments, we use a full-participation variant of DRFA (Deng et al., 2020) to compare optimization progress per synchronization stage under the same participation model as the other methods. We let every worker perform the local updates and use $\mathbf{q}^{(s)}$ to aggregate both the end-of-block models and the randomly selected snapshot models:

$$\bar{\mathbf{w}}^{(s+1)} = \sum_{i=1}^{m} q_i^{(s)} \mathbf{w}_i^{((s+1)\tau)}, \qquad \mathbf{w}^{(t')} = \sum_{i=1}^{m} q_i^{(s)} \mathbf{w}_i^{(t')}.$$

The weighted aggregation preserves the weighting represented by the adaptive-sampling expectation. All methods are trained for $T = 5000$ model updates, and for AFL-BR we set the block length to $B = \lceil \sqrt{T} \rceil = 71$. The target worst-worker test accuracies for $m = 20, 50, 100, 500$, and 1000 are $0.72, 0.71, 0.70, 0.48$, and 0.46, respectively.

D.2 Communication Efficiency with Compression

For Fashion-MNIST (Xiao et al., 2017) and CIFAR-10 (Krizhevsky & Hinton, 2009), we construct label-skewed local datasets as follows. For each class $i \in [m]$, worker i receives 80% of the samples with label i, while the remaining 20% are distributed uniformly among the other $m - 1$ workers. Each worker's local data are then split into training and test sets using an $80/20$ split. For the model architecture, we use a three-layer MLP with layer normalization and ReLU activations on Fashion-MNIST, and a two-block CNN followed by a linear classifier on CIFAR-10. Both models are trained using the cross-entropy loss.

Update rounds count every model update, whereas synchronization rounds count worker–server synchronization blocks. FedAvg and DRFA perform three local updates per synchronization, while AFL, AFL-BR, and both AFL-Com variants perform one update. Communication cost follows the accounting in Section 5.2, and the communication accounting for DRFA includes the additional model exchanges required by snapshotting.

Figure 3 reports the maximum test loss across workers, corresponding to the robust objective evaluated on test data. Panels (a) and (d) show that AFL-BR generally achieves the lowest loss for a given number of updates, while both AFL-Com variants remain close despite compression.

Panels (b) and (e) show that AFL-BR remains competitive with DRFA in synchronization efficiency while using only one update per synchronization. At comparable communication budgets, panels (c) and (f) show that AFL-Com(R) achieves the lowest maximum loss on both datasets, while AFL-Com(T) also substantially outperforms the uncompressed AFL baselines. These results corroborate the accuracy-based communication-efficiency comparison in Section 5.2.

E Nonconvex Analysis

E.1 Common Lemmas for Nonconvex Analysis

The following lemmas are shared by the analyses of AFL-BR and AFL-Com.

Definition 4. A function $f(\cdot)$ is called L-weakly convex if $f(\cdot) + \frac{L}{2} \|\cdot\|_2^2$ is convex.

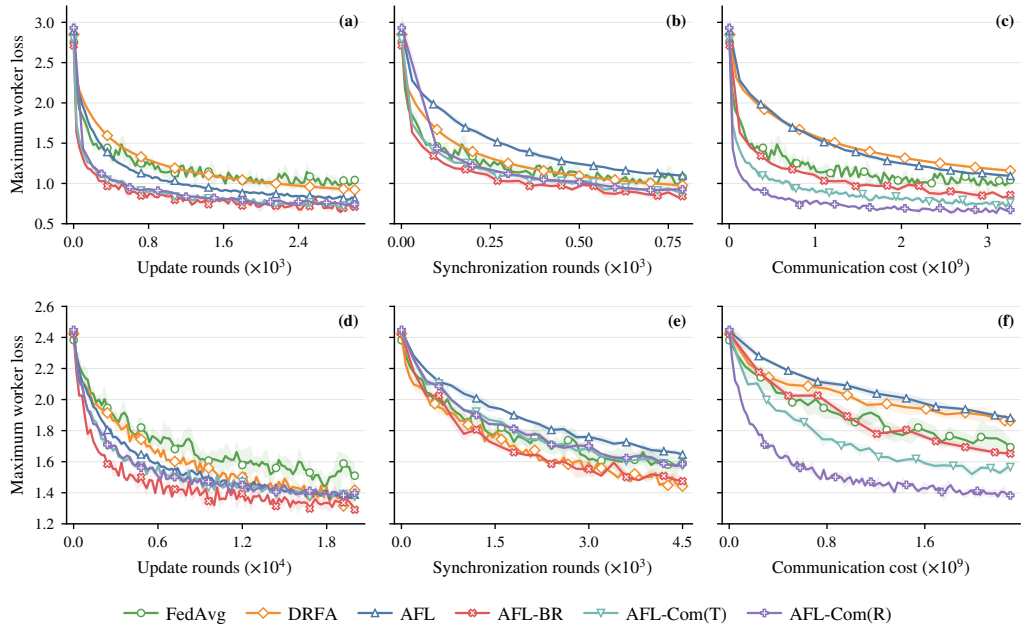


Figure 3: **Maximum worker test loss.** The top and bottom rows correspond to Fashion-MNIST and CIFAR-10, respectively. From left to right, the columns report performance against update rounds, synchronization rounds, and cumulative communication cost.

Lemma 2. (Lin et al., 2020, Lemma 4.7) Under Assumption 1, for every fixed $\mathbf{q} \in \Delta_m$, the function $\phi(\cdot, \mathbf{q})$ is L-weakly convex. Moreover, $\Phi(\cdot)$ is L-weakly convex.

Lemma 3. Suppose $\Phi : \mathbb{R}^d \to \mathbb{R}$ is L-weakly convex and G-Lipschitz. Then $\Phi_{1/(2L)}$ is differentiable, and its gradient satisfies

$$\|\nabla\Phi_{1/(2L)}(\mathbf{x}) - \nabla\Phi_{1/(2L)}(\mathbf{x}')\|_2 \le 2L\|\mathbf{x} - \mathbf{x}'\|_2, \quad \|\nabla\Phi_{1/(2L)}(\mathbf{x})\|_2 \le G, \quad \forall\, \mathbf{x}, \mathbf{x}' \in \mathbb{R}^d.$$

Lemma 4 (Perturbed Moreau-envelope descent)**.** Suppose Assumptions 1 and 2 hold. Let $\mathbf{u}_t$, $\mathbf{w}_t$, and $\mathbf{q}_t \in \Delta_m$ be $\mathcal{F}_t$-measurable for every $t \in [T+1]$, and suppose that, for some $\eta_w > 0$,

$$\mathbf{u}_{t+1} = \mathbf{u}_t - \eta_w \mathbf{g}_{w,t}, \quad \mathbb{E}_t[\mathbf{g}_{w,t}] = \nabla_{\mathbf{w}}\phi(\mathbf{w}_t, \mathbf{q}_t), \quad \mathbb{E}_t\left[\|\mathbf{g}_{w,t} - \nabla_{\mathbf{w}}\phi(\mathbf{w}_t, \mathbf{q}_t)\|_2^2\right] \le \sigma_w^2. \tag{9}$$

Assume further that $\|\nabla_{\mathbf{w}}\phi(\mathbf{w}, \mathbf{q})\|_2 \le G$ for every $\mathbf{w} \in \mathbb{R}^d$ and $\mathbf{q} \in \Delta_m$. Let $\mathcal{J} \subseteq [T]$ be a deterministic index set and define $\mathcal{S} = [T] \setminus \mathcal{J}$. For every $t \in [T]$, let $d_t = \|\mathbf{u}_t - \mathbf{w}_t\|_2$ and $\varepsilon_t^u = \Phi(\mathbf{u}_t) - \phi(\mathbf{u}_t, \mathbf{q}_t)$, and define $\Delta_u = \Phi_{1/(2L)}(\mathbf{u}_1) - \Phi_*$ and $\Gamma_w^2 = G^2 + \sigma_w^2$. Then

$$\begin{aligned}\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}\left[\left\|\nabla\Phi_{1/(2L)}(\mathbf{u}_t)\right\|_2^2\right] \le \frac{4\Delta_u}{\eta_w T} + \frac{8L}{T}\sum_{t\in\mathcal{J}}\mathbb{E}[\varepsilon_t^u] + \frac{4LG}{T}\sum_{t\in\mathcal{J}}\mathbb{E}[d_t] \\ + 4L\eta_w\Gamma_w^2 + \frac{5\Gamma_w^2|\mathcal{S}|}{T}.\end{aligned} \tag{10}$$

Lemma 5 (Blockwise dual-gap bound)**.** Let $1 \le B \le T$ and partition $[T]$ into $N = \lceil T/B \rceil$ consecutive blocks $\{\mathcal{I}_c\}_{c=1}^N$, where $\mathcal{I}_c = \{s_c, \dots, e_c\}$ has length B for every $c < N$ and length at most B for $c = N$. Define the block-start and non-start index sets as $\mathcal{S} = \{s_1, \dots, s_N\}$ and $\mathcal{J} = [T] \setminus \mathcal{S}$. Suppose that $\mathbf{q}_{s_c} = m^{-1}\mathbf{1}_m$ for every $c \in [N]$. For every $t \in \mathcal{I}_c$, define the pre-restart mirror-ascent iterate

$$\bar{\mathbf{q}}_{t+1} = \arg\max_{\mathbf{q}\in\Delta_m}\left\{\eta_q\langle\widehat{\mathbf{R}}_t, \mathbf{q}\rangle - D_{\mathrm{KL}}(\mathbf{q}\|\mathbf{q}_t)\right\}. \tag{11}$$

Assume that $\bar{\mathbf{q}}_{t+1} = \mathbf{q}_{t+1}$ for every $t < e_c$, while at the end of each nonfinal block the algorithm restarts and sets $\mathbf{q}_{e_c+1} = m^{-1}\mathbf{1}_m$. Suppose Assumption 4 holds, the round-t oracle $\widehat{\mathbf{R}}_t$ is evaluated

at $\mathbf{w}_t$, and every R_i is G-Lipschitz. Define $\varepsilon_t^q = \Phi(\mathbf{w}_t) - \phi(\mathbf{w}_t, \mathbf{q}_t)$. Then

$$\frac{1}{T}\sum_{t\in\mathcal{J}} \mathbb{E}[\varepsilon_t^q] \le \frac{2G(B+2)}{T}\sum_{t=1}^{T-1}\mathbb{E}[\|\mathbf{w}_{t+1}-\mathbf{w}_t\|_2] + \frac{2\ln m}{\eta_q B} + \frac{\eta_q \sigma_q^2}{2}. \tag{12}$$

Lemma 6. Under Assumption 3, consider Algorithm 2 and define $\mathbf{e}_t := \sum_{i=1}^m \mathbf{e}_t^{(i)}$, $\mathbf{g}_{w,t} := \sum_{i=1}^m q_{t,i}\mathbf{g}_{w,t}^{(i)}$, and $\mathrm{b}_t := \mathbf{e}_t + \widehat{\mathbf{e}}_t$. For every $t \le T+1$, a general δ-approximate compressor satisfies

$$\mathbb{E}\left[\|\mathbf{e}_t\|_2\right] \le \frac{2-\delta}{\delta}\Gamma_w, \qquad \mathbb{E}\left[\|\widehat{\mathbf{e}}_t\|_2\right] \le \frac{8}{\delta^2}\Gamma_w.$$

Moreover, since $\Delta_{w,t} = \mathbf{g}_{w,t} + \mathrm{b}_t - \mathrm{b}_{t+1}$,

$$\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}\left[\|\mathrm{b}_t\|_2\right] \le \frac{8+2\delta-\delta^2}{\delta^2}\Gamma_w, \qquad \frac{1}{T}\sum_{t=1}^{T}\mathbb{E}\left[\|\Delta_{w,t}\|_2\right] \le \frac{16+4\delta-\delta^2}{\delta^2}\Gamma_w. \tag{13}$$

If $\mathcal{C} \in \mathcal{C}_{\mathrm{AI}}$ and Algorithm 2 uses SR, then $\widehat{\mathbf{e}}_t = \mathbf{0}_d$ pathwise for every $t \le T+1$, and the bounds (13) improve to

$$\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}\left[\|\mathrm{b}_t\|_2\right] \le \frac{2\Gamma_w}{\delta}, \qquad \frac{1}{T}\sum_{t=1}^{T}\mathbb{E}\left[\|\Delta_{w,t}\|_2\right] \le \frac{4\Gamma_w}{\delta}.$$

E.2 Proof of Theorem 1

Let $\mathcal{S}$ be the set of block-start indices and let $\mathcal{J} = [T] \setminus \mathcal{S}$. For Algorithm 1, set $\mathbf{u}_t = \mathbf{w}_t$. The aggregated stochastic gradient is $\mathbf{g}_{w,t} = \sum_{i=1}^m q_{t,i}\mathbf{g}_{w,t}^{(i)}$. By Assumption 3,

$$\mathbb{E}_t[\mathbf{g}_{w,t}] = \sum_{i=1}^m q_{t,i}\nabla R_i(\mathbf{w}_t) = \nabla_{\mathbf{w}}\phi(\mathbf{w}_t, \mathbf{q}_t). \tag{14}$$

Let $\boldsymbol{\zeta}_{t,i} = \mathbf{g}_{w,t}^{(i)} - \nabla R_i(\mathbf{w}_t)$. Since $\|\sum_i q_{t,i}\boldsymbol{\zeta}_{t,i}\|_2^2 \le \sum_i q_{t,i}\|\boldsymbol{\zeta}_{t,i}\|_2^2$ pointwise,

$$\mathbb{E}_t\left[\|\mathbf{g}_{w,t} - \nabla_{\mathbf{w}}\phi(\mathbf{w}_t, \mathbf{q}_t)\|_2^2\right] \le \sum_{i=1}^m q_{t,i}\mathbb{E}_t[\|\boldsymbol{\zeta}_{t,i}\|_2^2] \le \sigma_w^2. \tag{15}$$

Moreover, $\|\nabla_{\mathbf{w}}\phi(\mathbf{w}_t, \mathbf{q}_t)\|_2 \le \sum_i q_{t,i}\|\nabla R_i(\mathbf{w}_t)\|_2 \le G$. Hence, $\mathbb{E}_t[\|\mathbf{g}_{w,t}\|_2^2] \le \Gamma_w^2$ and $\mathbb{E}_t[\|\mathbf{g}_{w,t}\|_2] \le \Gamma_w$.

Since $\mathbf{u}_t = \mathbf{w}_t$, we have $d_t = 0$, $\varepsilon_t^u = \varepsilon_t^q$, and $\Delta_u = \Delta_\Phi$. Applying Lemma 4 gives

$$\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}\left[\left\|\nabla\Phi_{1/(2L)}(\mathbf{w}_t)\right\|_2^2\right] \le \frac{4\Delta_\Phi}{\eta_w T} + \frac{8L}{T}\sum_{t\in\mathcal{J}}\mathbb{E}[\varepsilon_t^q] + 4L\eta_w\Gamma_w^2 + \frac{5\Gamma_w^2|\mathcal{S}|}{T}. \tag{16}$$

The primal update satisfies $\mathbf{w}_{t+1} - \mathbf{w}_t = -\eta_w\mathbf{g}_{w,t}$, and therefore

$$\sum_{t=1}^{T-1}\mathbb{E}[\|\mathbf{w}_{t+1}-\mathbf{w}_t\|_2] \le \eta_w\Gamma_w T. \tag{17}$$

Applying Lemma 5 with $\widehat{\mathbf{R}}_t = \mathbf{g}_{q,t}$ and then (17) gives

$$\frac{1}{T}\sum_{t\in\mathcal{J}}\mathbb{E}[\varepsilon_t^q] \le 2G\Gamma_w\eta_w(B+2) + \frac{2\ln m}{\eta_q B} + \frac{\eta_q\sigma_q^2}{2}. \tag{18}$$

Substituting (18) into (16) and using $|\mathcal{S}| = N \le 2T/B$ yields

$$\begin{aligned}\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}\left[\left\|\nabla\Phi_{1/(2L)}(\mathbf{w}_t)\right\|_2^2\right] \le \frac{4\Delta_\Phi}{\eta_w T} + 16LG\Gamma_w\eta_w(B+2) + \frac{16L\ln m}{\eta_q B} + 4L\eta_q\sigma_q^2 \\ + 4L\eta_w\Gamma_w^2 + \frac{10\Gamma_w^2}{B}.\end{aligned} \tag{19}$$

Because $T \geq 9$ and $B = \lceil \sqrt{T} \rceil$, we have $B \geq \sqrt{T}$ and $B + 2 \leq 2\sqrt{T}$. Substituting $\eta_w = (8LT^{3/4})^{-1}$ gives

$$\frac{4\Delta_\Phi}{\eta_w T} = \frac{32L\Delta_\Phi}{T^{1/4}}, \qquad 16LG\Gamma_w\eta_w(B+2) \leq \frac{4G\Gamma_w}{T^{1/4}}, \qquad 4L\eta_w\Gamma_w^2 = \frac{\Gamma_w^2}{2T^{3/4}}. \tag{20}$$

Substituting $\eta_q = 2\sqrt{\ln m}/(\sigma_q T^{1/4})$ and using $B \geq \sqrt{T}$ gives

$$\frac{16L\ln m}{\eta_q B} + 4L\eta_q\sigma_q^2 \leq \frac{8L\sigma_q\sqrt{\ln m}}{T^{1/4}} + \frac{8L\sigma_q\sqrt{\ln m}}{T^{1/4}} = \frac{16L\sigma_q\sqrt{\ln m}}{T^{1/4}}. \tag{21}$$

Finally, $10\Gamma_w^2/B \leq 10\Gamma_w^2/T^{1/2}$. Combining (19)–(21) yields

$$\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}\left[\left\|\nabla\Phi_{1/(2L)}(\mathbf{w}_t)\right\|_2^2\right] \leq \frac{32L\Delta_\Phi}{T^{1/4}} + \frac{4G\Gamma_w}{T^{1/4}} + \frac{16L\sigma_q\sqrt{\ln m}}{T^{1/4}} + \frac{10\Gamma_w^2}{T^{1/2}} + \frac{\Gamma_w^2}{2T^{3/4}}. \tag{22}$$

Since r is sampled uniformly and independently from $[T]$, the left-hand side of (22) equals $\mathbb{E}[\|\nabla\Phi_{1/(2L)}(\mathbf{w}_r)\|_2^2]$. This completes the proof.

E.3 Proofs of Theorems 2 and 3

We first derive a master inequality common to both compressor classes.

Let $\mathcal{S}$ be the set of block-start indices and let $\mathcal{J} = [T]\backslash\mathcal{S}$. Define the aggregated uplink residual $\mathbf{e}_t = \sum_{i=1}^m \mathbf{e}_t^{(i)}$ and the total residual $\mathrm{b}_t = \mathbf{e}_t + \widehat{\mathbf{e}}_t$. The uplink EF recursion gives $\widehat{\mathbf{g}}_{w,t} = \mathbf{g}_{w,t} + \mathbf{e}_t - \mathbf{e}_{t+1}$. Combining this identity with the downlink EF recursion yields

$$\Delta_{w,t} = \widehat{\mathbf{g}}_{w,t} + \widehat{\mathbf{e}}_t - \widehat{\mathbf{e}}_{t+1} = \mathbf{g}_{w,t} + \mathrm{b}_t - \mathrm{b}_{t+1}. \tag{23}$$

Introduce the virtual iterate $\mathbf{u}_t = \mathbf{w}_t - \eta_w \mathrm{b}_t$. Using $\mathbf{w}_{t+1} = \mathbf{w}_t - \eta_w\Delta_{w,t}$ and (23), we obtain

$$\mathbf{u}_{t+1} = \mathbf{w}_{t+1} - \eta_w\mathrm{b}_{t+1} = \mathbf{w}_t - \eta_w\Delta_{w,t} - \eta_w\mathrm{b}_{t+1} = \mathbf{u}_t - \eta_w\mathbf{g}_{w,t}. \tag{24}$$

The same argument as in (14)–(15) gives

$$\mathbb{E}_t[\mathbf{g}_{w,t}] = \nabla_{\mathbf{w}}\phi(\mathbf{w}_t, \mathbf{q}_t), \qquad \mathbb{E}_t\left[\|\mathbf{g}_{w,t} - \nabla_{\mathbf{w}}\phi(\mathbf{w}_t, \mathbf{q}_t)\|_2^2\right] \leq \sigma_w^2, \qquad \mathbb{E}_t[\|\mathbf{g}_{w,t}\|_2] \leq \Gamma_w.$$

All residuals are initialized at zero, so $\mathbf{u}_1 = \mathbf{w}_1$ and $\Delta_u = \Delta_\Phi$. Moreover, $d_t = \|\mathbf{u}_t - \mathbf{w}_t\|_2 = \eta_w\|\mathrm{b}_t\|_2$. Applying Lemma 4 to (24) gives

$$\begin{aligned}\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}\left[\left\|\nabla\Phi_{1/(2L)}(\mathbf{u}_t)\right\|_2^2\right] \leq \frac{4\Delta_\Phi}{\eta_w T} + \frac{8L}{T}\sum_{t\in\mathcal{J}}\mathbb{E}[\varepsilon_t^u] + \frac{4LG\eta_w}{T}\sum_{t\in\mathcal{J}}\mathbb{E}[\|\mathrm{b}_t\|_2] \\ + 4L\eta_w\Gamma_w^2 + \frac{5\Gamma_w^2|\mathcal{S}|}{T}.\end{aligned} \tag{25}$$

Because Φ and $\phi(\cdot, \mathbf{q}_t)$ are both G-Lipschitz,

$$\varepsilon_t^u = \Phi(\mathbf{u}_t) - \phi(\mathbf{u}_t, \mathbf{q}_t) \leq \Phi(\mathbf{w}_t) - \phi(\mathbf{w}_t, \mathbf{q}_t) + 2G\|\mathbf{u}_t - \mathbf{w}_t\|_2 = \varepsilon_t^q + 2G\eta_w\|\mathrm{b}_t\|_2. \tag{26}$$

Substituting (26) into (25), enlarging the residual sum from $\mathcal{J}$ to $[T]$, and using $|\mathcal{S}| \leq 2T/B$ yield

$$\begin{aligned}\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}\left[\left\|\nabla\Phi_{1/(2L)}(\mathbf{u}_t)\right\|_2^2\right] \leq \frac{4\Delta_\Phi}{\eta_w T} + \frac{8L}{T}\sum_{t\in\mathcal{J}}\mathbb{E}[\varepsilon_t^q] + \frac{20LG\eta_w}{T}\sum_{t=1}^{T}\mathbb{E}[\|\mathrm{b}_t\|_2] \\ + 4L\eta_w\Gamma_w^2 + \frac{10\Gamma_w^2}{B}.\end{aligned} \tag{27}$$

The actual update satisfies $\|\mathbf{w}_{t+1} - \mathbf{w}_t\|_2 = \eta_w\|\Delta_{w,t}\|_2$. Applying Lemma 5 gives

$$\frac{1}{T}\sum_{t\in\mathcal{J}}\mathbb{E}[\varepsilon_t^q] \leq \frac{2G\eta_w(B+2)}{T}\sum_{t=1}^{T}\mathbb{E}[\|\Delta_{w,t}\|_2] + \frac{2\ln m}{\eta_q B} + \frac{\eta_q\sigma_q^2}{2}. \tag{28}$$

Combining (27) and (28) yields

$$\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}\left[\left\|\nabla\Phi_{1/(2L)}(\mathbf{u}_t)\right\|_2^2\right] \le \frac{4\Delta_\Phi}{\eta_w T} + \frac{16LG\eta_w(B+2)}{T}\sum_{t=1}^{T}\mathbb{E}[\|\Delta_{w,t}\|_2] + \frac{16L\ln m}{\eta_q B} + 4L\eta_q\sigma_q^2 + \frac{20LG\eta_w}{T}\sum_{t=1}^{T}\mathbb{E}[\|\mathbf{b}_t\|_2] + 4L\eta_w\Gamma_w^2 + \frac{10\Gamma_w^2}{B}. \tag{29}$$

Let $\mathcal{G}_t^w = \nabla\Phi_{1/(2L)}(\mathbf{w}_t)$ and $\mathcal{G}_t^u = \nabla\Phi_{1/(2L)}(\mathbf{u}_t)$. Since $\nabla\Phi_{1/(2L)}$ is $2L$-Lipschitz and both gradients have norm at most G by Lemma 3, we have

$$\|\mathcal{G}_t^w\|_2^2 - \|\mathcal{G}_t^u\|_2^2 \le \|\mathcal{G}_t^w - \mathcal{G}_t^u\|_2(\|\mathcal{G}_t^w\|_2 + \|\mathcal{G}_t^u\|_2) \le 4LG\|\mathbf{w}_t - \mathbf{u}_t\|_2 = 4LG\eta_w\|\mathbf{b}_t\|_2. \tag{30}$$

Taking expectations, averaging (30), and applying (29) give

$$\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}\left[\left\|\nabla\Phi_{1/(2L)}(\mathbf{w}_t)\right\|_2^2\right] \le \frac{4\Delta_\Phi}{\eta_w T} + \frac{16LG\eta_w(B+2)}{T}\sum_{t=1}^{T}\mathbb{E}[\|\Delta_{w,t}\|_2] + \frac{16L\ln m}{\eta_q B} + 4L\eta_q\sigma_q^2 + \frac{24LG\eta_w}{T}\sum_{t=1}^{T}\mathbb{E}[\|\mathbf{b}_t\|_2] + 4L\eta_w\Gamma_w^2 + \frac{10\Gamma_w^2}{B}. \tag{31}$$

Since r is sampled uniformly from $[T]$ independently of the algorithmic randomness,

$$\mathbb{E}\left[\left\|\nabla\Phi_{1/(2L)}(\mathbf{w}_r)\right\|_2^2\right] = \frac{1}{T}\sum_{t=1}^{T}\mathbb{E}\left[\left\|\nabla\Phi_{1/(2L)}(\mathbf{w}_t)\right\|_2^2\right]. \tag{32}$$

Proof of Theorem 2. For a general δ-approximate compressor, Lemma 6 gives

$$\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}\left[\|\mathbf{b}_t\|_2\right] \le \frac{8+2\delta-\delta^2}{\delta^2}\Gamma_w, \qquad \frac{1}{T}\sum_{t=1}^{T}\mathbb{E}\left[\|\Delta_{w,t}\|_2\right] \le \frac{16+4\delta-\delta^2}{\delta^2}\Gamma_w.$$

Substituting these bounds into (31) yields

$$\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}\left[\left\|\nabla\Phi_{1/(2L)}(\mathbf{w}_t)\right\|_2^2\right] \le \frac{4\Delta_\Phi}{\eta_w T} + \frac{16LG\Gamma_w\eta_w(B+2)(16+4\delta-\delta^2)}{\delta^2} + \frac{16L\ln m}{\eta_q B} + 4L\eta_q\sigma_q^2 + \frac{24LG\Gamma_w\eta_w(8+2\delta-\delta^2)}{\delta^2} + 4L\eta_w\Gamma_w^2 + \frac{10\Gamma_w^2}{B}.$$

Using $B = \lceil\sqrt{T}\rceil$, $B \ge \sqrt{T}$, $B+2 \le 2\sqrt{T}$, $\eta_w = (8LT^{3/4})^{-1}$, and $\eta_q = 2\sqrt{\ln m}/(\sigma_q T^{1/4})$ gives

$$\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}\left[\left\|\nabla\Phi_{1/(2L)}(\mathbf{w}_t)\right\|_2^2\right] \le \frac{32L\Delta_\Phi}{T^{1/4}} + \frac{4G\Gamma_w(16+4\delta-\delta^2)}{\delta^2 T^{1/4}} + \frac{16L\sigma_q\sqrt{\ln m}}{T^{1/4}} + \frac{10\Gamma_w^2}{T^{1/2}} + \frac{3G\Gamma_w(8+2\delta-\delta^2)}{\delta^2 T^{3/4}} + \frac{\Gamma_w^2}{2T^{3/4}}. \tag{33}$$

Combining (32) and (33) proves (5).

Proof of Theorem 3. If $\mathcal{C} \in \mathcal{C}_{\mathrm{AI}}$ and SR is used, Lemma 6 gives

$$\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}\left[\|\mathbf{b}_t\|_2\right] \le \frac{2\Gamma_w}{\delta}, \qquad \frac{1}{T}\sum_{t=1}^{T}\mathbb{E}\left[\|\Delta_{w,t}\|_2\right] \le \frac{4\Gamma_w}{\delta}.$$

Substituting these bounds into (31) gives

$$\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}\left[\left\|\nabla\Phi_{1/(2L)}(\mathbf{w}_t)\right\|_2^2\right] \le \frac{4\Delta_\Phi}{\eta_w T} + \frac{64LG\Gamma_w\eta_w(B+2)}{\delta} + \frac{16L\ln m}{\eta_q B} + 4L\eta_q\sigma_q^2 + \frac{48LG\Gamma_w\eta_w}{\delta} + 4L\eta_w\Gamma_w^2 + \frac{10\Gamma_w^2}{B}.$$

Using the same parameter choices yields

$$\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}\left[\left\|\nabla\Phi_{1/(2L)}(\mathbf{w}_t)\right\|_2^2\right] \le \frac{32L\Delta_\Phi}{T^{1/4}} + \frac{16G\Gamma_w}{\delta T^{1/4}} + \frac{16L\sigma_q\sqrt{\ln m}}{T^{1/4}} + \frac{10\Gamma_w^2}{T^{1/2}} + \frac{6G\Gamma_w}{\delta T^{3/4}} + \frac{\Gamma_w^2}{2T^{3/4}}. \tag{34}$$

Combining (32) and (34) proves Theorem 3.

F Proof of Lemmas

F.1 Proof of Lemma 1

Because f is ρ-weakly convex, the function $h(\mathbf{y}) := f(\mathbf{y}) + \frac{\rho}{2}\|\mathbf{y}\|_2^2$ is convex. For any $\lambda < 1/\rho$, the proximal objective can be written as

$$f(\mathbf{y}) + \frac{1}{2\lambda}\|\mathbf{y}-\mathbf{x}\|_2^2 = h(\mathbf{y}) + \frac{1}{2}\left(\frac{1}{\lambda}-\rho\right)\|\mathbf{y}\|_2^2 - \frac{1}{\lambda}\langle\mathbf{x},\mathbf{y}\rangle + \frac{1}{2\lambda}\|\mathbf{x}\|_2^2.$$

Since $1/\lambda-\rho > 0$, this objective is $(1/\lambda-\rho)$-strongly convex and therefore has a unique minimizer. The same argument applies to $\mu < 1/\rho$. Define

$$\mathbf{y}_\lambda = \arg\min_{\mathbf{y}\in\mathbb{R}^d}\left\{f(\mathbf{y}) + \frac{1}{2\lambda}\|\mathbf{y}-\mathbf{x}\|_2^2\right\}, \qquad \mathbf{y}_\mu = \arg\min_{\mathbf{y}\in\mathbb{R}^d}\left\{f(\mathbf{y}) + \frac{1}{2\mu}\|\mathbf{y}-\mathbf{x}\|_2^2\right\}.$$

The uniqueness of these minimizers and Danskin's theorem imply that the corresponding Moreau envelopes are differentiable at $\mathbf{x}$, with

$$\mathbf{g}_\lambda := \nabla f_\lambda(\mathbf{x}) = \frac{\mathbf{x}-\mathbf{y}_\lambda}{\lambda}, \qquad \mathbf{g}_\mu := \nabla f_\mu(\mathbf{x}) = \frac{\mathbf{x}-\mathbf{y}_\mu}{\mu}.$$

The first-order optimality conditions for the proximal subproblems give

$$\mathbf{0} \in \partial f(\mathbf{y}_\lambda) + \frac{1}{\lambda}(\mathbf{y}_\lambda - \mathbf{x}), \qquad \mathbf{0} \in \partial f(\mathbf{y}_\mu) + \frac{1}{\mu}(\mathbf{y}_\mu - \mathbf{x}).$$

Hence, $\mathbf{g}_\lambda \in \partial f(\mathbf{y}_\lambda)$, and $\mathbf{g}_\mu \in \partial f(\mathbf{y}_\mu)$. Since f is ρ-weakly convex, its subdifferential is ρ-hypomonotone. Therefore,

$$\langle \mathbf{g}_\lambda - \mathbf{g}_\mu, \mathbf{y}_\lambda - \mathbf{y}_\mu\rangle \ge -\rho\|\mathbf{y}_\lambda - \mathbf{y}_\mu\|_2^2.$$

Substituting $\mathbf{y}_\lambda = \mathbf{x} - \lambda\mathbf{g}_\lambda$ and $\mathbf{y}_\mu = \mathbf{x} - \mu\mathbf{g}_\mu$ into the hypomonotonicity inequality, its left-hand side becomes

$$\langle \mathbf{g}_\lambda - \mathbf{g}_\mu, -\lambda\mathbf{g}_\lambda + \mu\mathbf{g}_\mu\rangle = -\lambda\|\mathbf{g}_\lambda\|_2^2 - \mu\|\mathbf{g}_\mu\|_2^2 + (\lambda+\mu)\langle\mathbf{g}_\lambda,\mathbf{g}_\mu\rangle,$$

whereas the squared norm on the right-hand side satisfies

$$\|-\lambda\mathbf{g}_\lambda + \mu\mathbf{g}_\mu\|_2^2 = \lambda^2\|\mathbf{g}_\lambda\|_2^2 + \mu^2\|\mathbf{g}_\mu\|_2^2 - 2\lambda\mu\langle\mathbf{g}_\lambda,\mathbf{g}_\mu\rangle.$$

Combining these two expansions and rearranging gives

$$\lambda(1-\rho\lambda)\|\mathbf{g}_\lambda\|_2^2 + \mu(1-\rho\mu)\|\mathbf{g}_\mu\|_2^2 \le (\lambda+\mu-2\rho\lambda\mu)\langle\mathbf{g}_\lambda,\mathbf{g}_\mu\rangle \le (\lambda+\mu-2\rho\lambda\mu)\|\mathbf{g}_\lambda\|_2\|\mathbf{g}_\mu\|_2. \tag{35}$$

We first consider the case $\mathbf{g}_\mu = \mathbf{0}$. The preceding inequality (35) then reduces to

$$\lambda(1-\rho\lambda)\|\mathbf{g}_\lambda\|_2^2 \le 0.$$

Since $\lambda > 0$ and $1-\rho\lambda > 0$, it follows that $\mathbf{g}_\lambda = \mathbf{0}$, and the desired inequality holds.

Now suppose that $\mathbf{g}_\mu \neq \mathbf{0}$, and define $r := \|\mathbf{g}_\lambda\|_2/\|\mathbf{g}_\mu\|_2$. Dividing both sides of (35) by $\|\mathbf{g}_\mu\|_2^2$ yields

$$\lambda(1-\rho\lambda)r^2 - (\lambda+\mu-2\rho\lambda\mu)r + \mu(1-\rho\mu) \le 0.$$

The quadratic polynomial on the left factors as

$$\lambda(1-\rho\lambda)\left(r-\frac{\mu}{\lambda}\right)\left(r-\frac{1-\rho\mu}{1-\rho\lambda}\right).$$

Because $\lambda < 1/\rho$, its leading coefficient $\lambda(1-\rho\lambda)$ is positive. Moreover, $\mu \le \lambda$ implies

$$\frac{1-\rho\mu}{1-\rho\lambda} - \frac{\mu}{\lambda} = \frac{\lambda-\mu}{\lambda(1-\rho\lambda)} \ge 0,$$

so $(1-\rho\mu)/(1-\rho\lambda)$ is the larger of the two roots. Since the quadratic is nonpositive, r lies between its two roots, and in particular,

$$r \le \frac{1-\rho\mu}{1-\rho\lambda}.$$

Recalling the definition of r, we conclude that

$$\|\nabla f_\lambda(\mathbf{x})\|_2 = \|\mathbf{g}_\lambda\|_2 \le \frac{1-\rho\mu}{1-\rho\lambda}\|\mathbf{g}_\mu\|_2 = \frac{1-\rho\mu}{1-\rho\lambda}\|\nabla f_\mu(\mathbf{x})\|_2.$$

F.2 Proof of Lemma 2

By Assumption 1, each R_i is L-smooth. Hence, for any $\mathbf{x}, \mathbf{y} \in \mathbb{R}^d$,

$$R_i(\mathbf{y}) \ge R_i(\mathbf{x}) + \langle \nabla R_i(\mathbf{x}), \mathbf{y}-\mathbf{x}\rangle - \frac{L}{2}\|\mathbf{y}-\mathbf{x}\|_2^2.$$

Define $f_i(\mathbf{x}) = R_i(\mathbf{x}) + \frac{L}{2}\|\mathbf{x}\|_2^2$. Then $\nabla f_i(\mathbf{x}) = \nabla R_i(\mathbf{x}) + L\mathbf{x}$. Using the above lower bound, we have

$$\begin{aligned} f_i(\mathbf{y}) &= R_i(\mathbf{y}) + \frac{L}{2}\|\mathbf{y}\|_2^2 \\ &\ge R_i(\mathbf{x}) + \langle \nabla R_i(\mathbf{x}), \mathbf{y}-\mathbf{x}\rangle - \frac{L}{2}\|\mathbf{y}-\mathbf{x}\|_2^2 + \frac{L}{2}\|\mathbf{y}\|_2^2 \\ &= R_i(\mathbf{x}) + \frac{L}{2}\|\mathbf{x}\|_2^2 + \langle \nabla R_i(\mathbf{x}) + L\mathbf{x}, \mathbf{y}-\mathbf{x}\rangle \\ &= f_i(\mathbf{x}) + \langle \nabla f_i(\mathbf{x}), \mathbf{y}-\mathbf{x}\rangle. \end{aligned}$$

Therefore f_i is convex, and hence R_i is L-weakly convex.

For any fixed $\mathbf{q} \in \Delta_m$, since $q_i \ge 0$ and $\sum_{i=1}^m q_i = 1$, we have

$$\phi(\cdot,\mathbf{q}) + \frac{L}{2}\|\cdot\|_2^2 = \sum_{i=1}^m q_i R_i(\cdot) + \frac{L}{2}\|\cdot\|_2^2 = \sum_{i=1}^m q_i\left(R_i(\cdot) + \frac{L}{2}\|\cdot\|_2^2\right).$$

The right-hand side is a convex combination of convex functions, and is therefore convex. Hence $\phi(\cdot,\mathbf{q})$ is L-weakly convex for every $\mathbf{q} \in \Delta_m$.

Moreover,

$$\Phi(\cdot) + \frac{L}{2}\|\cdot\|_2^2 = \max_{\mathbf{q}\in\Delta_m} \phi(\cdot,\mathbf{q}) + \frac{L}{2}\|\cdot\|_2^2 = \max_{\mathbf{q}\in\Delta_m}\left\{\phi(\cdot,\mathbf{q}) + \frac{L}{2}\|\cdot\|_2^2\right\}.$$

Since the pointwise maximum of convex functions is convex, $\Phi(\cdot) + \frac{L}{2}\|\cdot\|_2^2$ is convex. Therefore, Φ is L-weakly convex.

F.3 Proof of Lemma 3

For any $\mathbf{x} \in \mathbb{R}^d$, define

$$\mathbf{y}(\mathbf{x}) = \arg\min_{\mathbf{y}\in\mathbb{R}^d}\left\{\Phi(\mathbf{y}) + L\|\mathbf{y}-\mathbf{x}\|_2^2\right\}. \tag{36}$$

Since Φ is L-weakly convex, the function $\mathbf{y} \mapsto \Phi(\mathbf{y}) + \frac{L}{2}\|\mathbf{y}\|_2^2$ is convex. Hence, for every fixed $\mathbf{x}$, the objective in (36) is strongly convex, and $\mathbf{y}(\mathbf{x})$ is uniquely defined.

The first-order optimality condition for (36) gives $\mathbf{0} \in \partial\Phi(\mathbf{y}(\mathbf{x})) + 2L(\mathbf{y}(\mathbf{x}) - \mathbf{x})$. Then, we define $\mathbf{G}(\mathbf{x}) = 2L(\mathbf{x} - \mathbf{y}(\mathbf{x})) \in \partial\Phi(\mathbf{y}(\mathbf{x}))$. Since the minimizer in (36) is unique, Danskin's theorem implies that the Moreau envelope is differentiable and

$$\nabla\Phi_{1/(2L)}(\mathbf{x}) = 2L(\mathbf{x} - \mathbf{y}(\mathbf{x})) = \mathbf{G}(\mathbf{x}). \tag{37}$$

We first establish the uniform bound on the Moreau-envelope gradient. Since Φ is G-Lipschitz, every subgradient $\mathbf{v} \in \partial\Phi(\mathbf{y})$ satisfies $\|\mathbf{v}\|_2 \le G$. Combining this property with $\mathbf{G}(\mathbf{x}) \in \partial\Phi(\mathbf{y}(\mathbf{x}))$ and (37) yields

$$\|\nabla\Phi_{1/(2L)}(\mathbf{x})\|_2 = \|\mathbf{G}(\mathbf{x})\|_2 \le G. \tag{38}$$

It remains to prove that $\mathbf{G}(\cdot)$ is $2L$-Lipschitz. For any $\mathbf{x}, \mathbf{x}' \in \mathbb{R}^d$, the optimality condition gives $\mathbf{G}(\mathbf{x}) \in \partial\Phi(\mathbf{y}(\mathbf{x}))$ and $\mathbf{G}(\mathbf{x}') \in \partial\Phi(\mathbf{y}(\mathbf{x}'))$. Since Φ is L-weakly convex, its subdifferential is L-hypomonotone. Therefore,

$$\langle \mathbf{G}(\mathbf{x}) - \mathbf{G}(\mathbf{x}'), \mathbf{y}(\mathbf{x}) - \mathbf{y}(\mathbf{x}')\rangle \ge -L\|\mathbf{y}(\mathbf{x}) - \mathbf{y}(\mathbf{x}')\|_2^2. \tag{39}$$

By the definition of $\mathbf{G}$,

$$\mathbf{G}(\mathbf{x}) - \mathbf{G}(\mathbf{x}') = 2L\left(\mathbf{x} - \mathbf{x}' - \big(\mathbf{y}(\mathbf{x}) - \mathbf{y}(\mathbf{x}')\big)\right). \tag{40}$$

Substituting (40) into (39) and rearranging gives $2\langle \mathbf{x} - \mathbf{x}', \mathbf{y}(\mathbf{x}) - \mathbf{y}(\mathbf{x}')\rangle \ge \|\mathbf{y}(\mathbf{x}) - \mathbf{y}(\mathbf{x}')\|_2^2$. Consequently,

$$\begin{aligned}\big\|\mathbf{x} - \mathbf{x}' - \big(\mathbf{y}(\mathbf{x}) - \mathbf{y}(\mathbf{x}')\big)\big\|_2^2 &= \|\mathbf{x} - \mathbf{x}'\|_2^2 - 2\langle \mathbf{x} - \mathbf{x}', \mathbf{y}(\mathbf{x}) - \mathbf{y}(\mathbf{x}')\rangle + \|\mathbf{y}(\mathbf{x}) - \mathbf{y}(\mathbf{x}')\|_2^2 \\ &\le \|\mathbf{x} - \mathbf{x}'\|_2^2.\end{aligned} \tag{41}$$

Using (40) and (41), we obtain

$$\|\mathbf{G}(\mathbf{x}) - \mathbf{G}(\mathbf{x}')\|_2 = 2L\big\|\mathbf{x} - \mathbf{x}' - \big(\mathbf{y}(\mathbf{x}) - \mathbf{y}(\mathbf{x}')\big)\big\|_2 \le 2L\|\mathbf{x} - \mathbf{x}'\|_2. \tag{42}$$

Combining (37), (38), and (42) proves the result.

F.4 Proof of Lemma 4

Since $\|\nabla_{\mathbf{w}}\phi(\mathbf{w}, \mathbf{q})\|_2 \le G$, the function $\phi(\cdot, \mathbf{q})$ is G-Lipschitz for every $\mathbf{q} \in \Delta_m$. As the pointwise maximum of G-Lipschitz functions, Φ is also G-Lipschitz. Moreover, Lemma 2 shows that Φ is L-weakly convex. Hence, Lemma 3 applies.

For every $t \in [T]$, define the proximal point

$$\mathbf{y}_t = \operatorname*{arg\,min}_{\mathbf{y}\in\mathbb{R}^d}\left\{\Phi(\mathbf{y}) + L\|\mathbf{y} - \mathbf{u}_t\|_2^2\right\}, \qquad \mathcal{G}_t^{\Phi} := \nabla\Phi_{1/(2L)}(\mathbf{u}_t) = 2L(\mathbf{u}_t - \mathbf{y}_t). \tag{43}$$

By the L-smoothness of $\phi(\cdot, \mathbf{q}_t)$,

$$\langle \nabla_{\mathbf{w}}\phi(\mathbf{u}_t, \mathbf{q}_t), \mathbf{u}_t - \mathbf{y}_t\rangle \ge \phi(\mathbf{u}_t, \mathbf{q}_t) - \phi(\mathbf{y}_t, \mathbf{q}_t) - \frac{L}{2}\|\mathbf{u}_t - \mathbf{y}_t\|_2^2. \tag{44}$$

By definition, $\phi(\mathbf{u}_t, \mathbf{q}_t) = \Phi(\mathbf{u}_t) - \varepsilon_t^u$, while $\phi(\mathbf{y}_t, \mathbf{q}_t) \le \Phi(\mathbf{y}_t)$. In addition, the optimality of $\mathbf{y}_t$ in (43) gives $\Phi(\mathbf{y}_t) + L\|\mathbf{y}_t - \mathbf{u}_t\|_2^2 \le \Phi(\mathbf{u}_t)$, and consequently $\Phi(\mathbf{u}_t) - \Phi(\mathbf{y}_t) \ge L\|\mathbf{u}_t - \mathbf{y}_t\|_2^2$. Multiplying (44) by $2L$ and using these relations yields

$$\begin{aligned}\langle \nabla_{\mathbf{w}}\phi(\mathbf{u}_t, \mathbf{q}_t), \mathcal{G}_t^{\Phi}\rangle &= 2L\langle \nabla_{\mathbf{w}}\phi(\mathbf{u}_t, \mathbf{q}_t), \mathbf{u}_t - \mathbf{y}_t\rangle \\ &\ge 2L\phi(\mathbf{u}_t, \mathbf{q}_t) - 2L\phi(\mathbf{y}_t, \mathbf{q}_t) - L^2\|\mathbf{u}_t - \mathbf{y}_t\|_2^2 \\ &\ge 2L\phi(\mathbf{u}_t, \mathbf{q}_t) - 2L\Phi(\mathbf{y}_t) - L^2\|\mathbf{u}_t - \mathbf{y}_t\|_2^2 \\ &= 2L\Phi(\mathbf{u}_t) - 2L\varepsilon_t^u - 2L\Phi(\mathbf{y}_t) - L^2\|\mathbf{u}_t - \mathbf{y}_t\|_2^2 \\ &= 2L\big(\Phi(\mathbf{u}_t) - \Phi(\mathbf{y}_t)\big) - L^2\|\mathbf{u}_t - \mathbf{y}_t\|_2^2 - 2L\varepsilon_t^u \\ &\ge 2L^2\|\mathbf{u}_t - \mathbf{y}_t\|_2^2 - L^2\|\mathbf{u}_t - \mathbf{y}_t\|_2^2 - 2L\varepsilon_t^u \\ &= L^2\|\mathbf{u}_t - \mathbf{y}_t\|_2^2 - 2L\varepsilon_t^u \\ &= \frac{1}{4}\|\mathcal{G}_t^{\Phi}\|_2^2 - 2L\varepsilon_t^u,\end{aligned} \tag{45}$$

The L-smoothness of $\phi(\cdot, \mathbf{q}_t)$ also gives

$$\|\nabla_{\mathbf{w}}\phi(\mathbf{w}_t, \mathbf{q}_t) - \nabla_{\mathbf{w}}\phi(\mathbf{u}_t, \mathbf{q}_t)\|_2 \leq L\|\mathbf{w}_t - \mathbf{u}_t\|_2 = Ld_t.$$

Combining this inequality with (45) and the bound $\|\mathcal{G}_t^{\Phi}\|_2 \leq G$ from Lemma 3, we obtain

$$\begin{aligned}\left\langle \nabla_{\mathbf{w}}\phi(\mathbf{w}_t, \mathbf{q}_t), \mathcal{G}_t^{\Phi} \right\rangle &= \left\langle \nabla_{\mathbf{w}}\phi(\mathbf{u}_t, \mathbf{q}_t), \mathcal{G}_t^{\Phi} \right\rangle + \left\langle \nabla_{\mathbf{w}}\phi(\mathbf{w}_t, \mathbf{q}_t) - \nabla_{\mathbf{w}}\phi(\mathbf{u}_t, \mathbf{q}_t), \mathcal{G}_t^{\Phi} \right\rangle \\ &\geq \frac{1}{4}\|\mathcal{G}_t^{\Phi}\|_2^2 - 2L\varepsilon_t^u - Ld_t\|\mathcal{G}_t^{\Phi}\|_2 \\ &\geq \frac{1}{4}\|\mathcal{G}_t^{\Phi}\|_2^2 - 2L\varepsilon_t^u - LGd_t.\end{aligned} \tag{46}$$

Using the definition of the Moreau envelope and the update in (9), we have

$$\begin{aligned}\Phi_{1/(2L)}(\mathbf{u}_{t+1}) &= \min_{\mathbf{y}\in\mathbb{R}^d}\{\Phi(\mathbf{y}) + L\|\mathbf{y} - \mathbf{u}_{t+1}\|_2^2\} \\ &\leq \Phi(\mathbf{y}_t) + L\|\mathbf{y}_t - \mathbf{u}_{t+1}\|_2^2 \\ &= \Phi(\mathbf{y}_t) + L\|\mathbf{y}_t - \mathbf{u}_t + \eta_w\mathbf{g}_{w,t}\|_2^2 \\ &= \Phi(\mathbf{y}_t) + L\|\mathbf{y}_t - \mathbf{u}_t\|_2^2 + 2L\eta_w\langle \mathbf{y}_t - \mathbf{u}_t, \mathbf{g}_{w,t}\rangle + L\eta_w^2\|\mathbf{g}_{w,t}\|_2^2 \\ &= \Phi_{1/(2L)}(\mathbf{u}_t) - \eta_w\langle \mathbf{g}_{w,t}, \mathcal{G}_t^{\Phi}\rangle + L\eta_w^2\|\mathbf{g}_{w,t}\|_2^2.\end{aligned} \tag{47}$$

Using $\mathbb{E}_t[\mathbf{g}_{w,t}] = \nabla_{\mathbf{w}}\phi(\mathbf{w}_t, \mathbf{q}_t)$ and $\mathbb{E}_t[\|\mathbf{g}_{w,t}\|_2^2] \leq \Gamma_w^2$, taking the conditional expectation of (47) yields

$$\begin{aligned}\mathbb{E}_t[\Phi_{1/(2L)}(\mathbf{u}_{t+1})] &\leq \Phi_{1/(2L)}(\mathbf{u}_t) - \eta_w\left\langle \nabla_{\mathbf{w}}\phi(\mathbf{w}_t, \mathbf{q}_t), \mathcal{G}_t^{\Phi}\right\rangle + L\eta_w^2\mathbb{E}_t[\|\mathbf{g}_{w,t}\|_2^2] \\ &\leq \Phi_{1/(2L)}(\mathbf{u}_t) - \eta_w\left\langle \nabla_{\mathbf{w}}\phi(\mathbf{w}_t, \mathbf{q}_t), \mathcal{G}_t^{\Phi}\right\rangle + L\eta_w^2\Gamma_w^2 \\ &\overset{(46)}{\leq} \Phi_{1/(2L)}(\mathbf{u}_t) - \frac{\eta_w}{4}\|\mathcal{G}_t^{\Phi}\|_2^2 + 2L\eta_w\varepsilon_t^u + LG\eta_w d_t + L\eta_w^2\Gamma_w^2,\end{aligned} \tag{48}$$

where we use

$$\mathbb{E}_t[\|\mathbf{g}_{w,t}\|_2^2] = \|\nabla_{\mathbf{w}}\phi(\mathbf{w}_t, \mathbf{q}_t)\|_2^2 + \mathbb{E}_t\left[\|\mathbf{g}_{w,t} - \nabla_{\mathbf{w}}\phi(\mathbf{w}_t, \mathbf{q}_t)\|_2^2\right] \leq G^2 + \sigma_w^2 = \Gamma_w^2.$$

Taking the total expectation on both sides of (48) and rearranging the stationarity term gives

$$\frac{\eta_w}{4}\mathbb{E}[\|\mathcal{G}_t^{\Phi}\|_2^2] \leq \mathbb{E}[\Phi_{1/(2L)}(\mathbf{u}_t)] - \mathbb{E}[\Phi_{1/(2L)}(\mathbf{u}_{t+1})] + 2L\eta_w\mathbb{E}[\varepsilon_t^u] + LG\eta_w\mathbb{E}[d_t] + L\eta_w^2\Gamma_w^2. \tag{49}$$

We next sum (49) only over $t \in \mathcal{J}$. Let $F_t = \Phi_{1/(2L)}(\mathbf{u}_t)$. By Lemma 3, $\|\nabla\Phi_{1/(2L)}(\mathbf{x})\|_2 \leq G$, and hence $\Phi_{1/(2L)}$ is G-Lipschitz. Therefore, for every $t \in \mathcal{S}$,

$$\mathbb{E}[F_{t+1} - F_t] \leq G\mathbb{E}[\|\mathbf{u}_{t+1} - \mathbf{u}_t\|_2] = G\eta_w\mathbb{E}[\|\mathbf{g}_{w,t}\|_2] \leq G\eta_w\sqrt{\mathbb{E}[\|\mathbf{g}_{w,t}\|_2^2]} \leq G\eta_w\Gamma_w. \tag{50}$$

Since $\sum_{t\in\mathcal{J}}(F_t - F_{t+1}) = F_1 - F_{T+1} + \sum_{t\in\mathcal{S}}(F_{t+1} - F_t)$ and $F_{T+1} \geq \Phi_*$, it follows from (50) that

$$\sum_{t\in\mathcal{J}}(\mathbb{E}[F_t] - \mathbb{E}[F_{t+1}]) \leq \Delta_u + G\eta_w\Gamma_w|\mathcal{S}|. \tag{51}$$

Summing (49) over $t \in \mathcal{J}$, applying (51), and using $|\mathcal{J}| \leq T$ yields

$$\frac{1}{T}\sum_{t\in\mathcal{J}}\mathbb{E}[\|\mathcal{G}_t^{\Phi}\|_2^2] \leq \frac{4\Delta_u}{\eta_w T} + \frac{8L}{T}\sum_{t\in\mathcal{J}}\mathbb{E}[\varepsilon_t^u] + \frac{4LG}{T}\sum_{t\in\mathcal{J}}\mathbb{E}[d_t] + 4L\eta_w\Gamma_w^2 + \frac{4G\Gamma_w|\mathcal{S}|}{T}. \tag{52}$$

Finally, Lemma 3 and $G \leq \Gamma_w$ imply $\sum_{t\in\mathcal{S}}\mathbb{E}[\|\mathcal{G}_t^{\Phi}\|_2^2] \leq G^2|\mathcal{S}| \leq \Gamma_w^2|\mathcal{S}|$. Adding this bound to (52) and using $4G\Gamma_w + \Gamma_w^2 \leq 5\Gamma_w^2$ proves (10).

F.5 Proof of Lemma 5

Let $\mathbf{q}^{\mathrm{unif}} = m^{-1}\mathbf{1}_m$. For every block $\mathcal{I}_c$, choose an anchor $\mathbf{q}^*_{s_c} \in \arg\max_{\mathbf{q}\in\Delta_m} \phi(\mathbf{w}_{s_c}, \mathbf{q})$. For every $t \in \mathcal{I}_c$, decompose the instantaneous dual gap as

$$\varepsilon^q_t = \underbrace{\Phi(\mathbf{w}_t) - \phi(\mathbf{w}_t, \mathbf{q}^*_{s_c})}_{A_t} + \underbrace{\phi(\mathbf{w}_t, \mathbf{q}^*_{s_c}) - \phi(\mathbf{w}_t, \mathbf{q}_t)}_{B_t}. \tag{53}$$

We first control the error induced by fixing the anchor within each block. Let $\mathbf{q}^*_t \in \arg\max_{\mathbf{q}\in\Delta_m} \phi(\mathbf{w}_t, \mathbf{q})$. By the optimality of $\mathbf{q}^*_{s_c}$ at $\mathbf{w}_{s_c}$, we obtain $\phi(\mathbf{w}_{s_c}, \mathbf{q}^*_t) \le \phi(\mathbf{w}_{s_c}, \mathbf{q}^*_{s_c})$. Since every R_i is G-Lipschitz, $\phi(\cdot, \mathbf{q})$ is G-Lipschitz uniformly over $\mathbf{q} \in \Delta_m$. Hence,

$$\begin{aligned}
A_t &= \phi(\mathbf{w}_t, \mathbf{q}^*_t) - \phi(\mathbf{w}_t, \mathbf{q}^*_{s_c}) \\
&= \phi(\mathbf{w}_t, \mathbf{q}^*_t) - \phi(\mathbf{w}_{s_c}, \mathbf{q}^*_t) + \phi(\mathbf{w}_{s_c}, \mathbf{q}^*_t) - \phi(\mathbf{w}_{s_c}, \mathbf{q}^*_{s_c}) + \phi(\mathbf{w}_{s_c}, \mathbf{q}^*_{s_c}) - \phi(\mathbf{w}_t, \mathbf{q}^*_{s_c}) \\
&\le \phi(\mathbf{w}_t, \mathbf{q}^*_t) - \phi(\mathbf{w}_{s_c}, \mathbf{q}^*_t) + \phi(\mathbf{w}_{s_c}, \mathbf{q}^*_{s_c}) - \phi(\mathbf{w}_t, \mathbf{q}^*_{s_c}) \\
&\le 2G\|\mathbf{w}_t - \mathbf{w}_{s_c}\|_2.
\end{aligned}$$

For $t \in \mathcal{I}_c$, the triangle inequality gives $\|\mathbf{w}_t - \mathbf{w}_{s_c}\|_2 \le \sum_{\tau=s_c}^{t-1} \|\mathbf{w}_{\tau+1} - \mathbf{w}_\tau\|_2$. Since $|\mathcal{I}_c| \le B$, summing over each block yields

$$\sum_{c=1}^{N} \sum_{t\in\mathcal{I}_c} \mathbb{E}[A_t] \le 2GB \sum_{c=1}^{N} \sum_{\tau=s_c}^{e_c-1} \mathbb{E}[\|\mathbf{w}_{\tau+1} - \mathbf{w}_\tau\|_2] \le 2GB \sum_{t=1}^{T-1} \mathbb{E}[\|\mathbf{w}_{t+1} - \mathbf{w}_t\|_2]. \tag{54}$$

We next control B_t using KL-based online mirror ascent. Let $\boldsymbol{\xi}_t = \widehat{\mathbf{R}}_t - \mathbf{R}(\mathbf{w}_t)$, where $\xi_{t,i} = \ell(\mathbf{w}_t; \mathbf{z}^{(i)}_t) - R_i(\mathbf{w}_t)$. Since $\mathbf{w}_t$ is $\mathcal{F}_t$-measurable, Assumption 4 implies

$$\mathbb{E}_t[\xi_{t,i}] = 0, \qquad \mathbb{E}_t\left[\exp(\lambda\xi_{t,i})\right] \le \exp\left(\frac{\lambda^2\sigma_q^2}{2}\right), \qquad \forall i \in [m],\ \lambda \in \mathbb{R}. \tag{55}$$

The first-order optimality condition for (11), together with the Bregman three-point identity, implies that for every $\mathbf{q} \in \Delta_m$,

$$\eta_q\langle \widehat{\mathbf{R}}_t, \mathbf{q} - \bar{\mathbf{q}}_{t+1}\rangle \le D_{\mathrm{KL}}(\mathbf{q}\|\mathbf{q}_t) - D_{\mathrm{KL}}(\mathbf{q}\|\bar{\mathbf{q}}_{t+1}) - D_{\mathrm{KL}}(\bar{\mathbf{q}}_{t+1}\|\mathbf{q}_t). \tag{56}$$

Since ϕ is linear in $\mathbf{q}$, and both $\mathbf{q}^*_{s_c}$ and $\mathbf{q}_t$ are $\mathcal{F}_t$-measurable, the inequality (56) with $\mathbf{q} = \mathbf{q}^*_{s_c}$ yields

$$\begin{aligned}
B_t &= \left\langle \mathbf{R}(\mathbf{w}_t), \mathbf{q}^*_{s_c} - \mathbf{q}_t \right\rangle \\
&= \mathbb{E}_t\left[\left\langle \widehat{\mathbf{R}}_t, \mathbf{q}^*_{s_c} - \bar{\mathbf{q}}_{t+1}\right\rangle + \left\langle \widehat{\mathbf{R}}_t, \bar{\mathbf{q}}_{t+1} - \mathbf{q}_t\right\rangle\right] \\
&\le \frac{D_{\mathrm{KL}}(\mathbf{q}^*_{s_c}\|\mathbf{q}_t) - \mathbb{E}_t[D_{\mathrm{KL}}(\mathbf{q}^*_{s_c}\|\bar{\mathbf{q}}_{t+1})]}{\eta_q} - \frac{1}{\eta_q}\mathbb{E}_t[D_{\mathrm{KL}}(\bar{\mathbf{q}}_{t+1}\|\mathbf{q}_t)] \\
&\quad + \mathbb{E}_t\left[\langle \mathbf{R}(\mathbf{w}_t), \bar{\mathbf{q}}_{t+1} - \mathbf{q}_t\rangle + \langle \boldsymbol{\xi}_t, \bar{\mathbf{q}}_{t+1} - \mathbf{q}_t\rangle\right] \\
&= \frac{D_{\mathrm{KL}}(\mathbf{q}^*_{s_c}\|\mathbf{q}_t) - \mathbb{E}_t[D_{\mathrm{KL}}(\mathbf{q}^*_{s_c}\|\bar{\mathbf{q}}_{t+1})]}{\eta_q} + \mathbb{E}_t\left[\phi(\mathbf{w}_t, \bar{\mathbf{q}}_{t+1}) - \phi(\mathbf{w}_t, \mathbf{q}_t)\right] \\
&\quad + \mathbb{E}_t\left[\langle \boldsymbol{\xi}_t, \bar{\mathbf{q}}_{t+1} - \mathbf{q}_t\rangle - \frac{1}{\eta_q} D_{\mathrm{KL}}(\bar{\mathbf{q}}_{t+1}\|\mathbf{q}_t)\right].
\end{aligned} \tag{57}$$

We next control the final term in (57) through the conjugacy between negative entropy and log-sum-exp. Since the mirror-ascent iterates remain in the relative interior of Δ_m, for every $\boldsymbol{\xi} \in \mathbb{R}^m$,

$$\sup_{\mathbf{q}\in\Delta_m}\left\{\langle \boldsymbol{\xi}, \mathbf{q} - \mathbf{q}_t\rangle - \frac{1}{\eta_q} D_{\mathrm{KL}}(\mathbf{q}\|\mathbf{q}_t)\right\} = \frac{1}{\eta_q}\ln\left(\sum_{i=1}^{m} q_{t,i}\exp(\eta_q\xi_i)\right) - \langle \boldsymbol{\xi}, \mathbf{q}_t\rangle. \tag{58}$$

Since $\bar{\mathbf{q}}_{t+1} \in \Delta_m$, applying (58) with $\boldsymbol{\xi} = \boldsymbol{\xi}_t$ and taking the conditional expectation gives

$$\begin{aligned}
&\mathbb{E}_t\left[\langle \boldsymbol{\xi}_t, \bar{\mathbf{q}}_{t+1} - \mathbf{q}_t\rangle - \frac{1}{\eta_q} D_{\mathrm{KL}}(\bar{\mathbf{q}}_{t+1} \| \mathbf{q}_t)\right] \\
&\le \frac{1}{\eta_q}\mathbb{E}_t\left[\ln\left(\sum_{i=1}^m q_{t,i}\exp(\eta_q \xi_{t,i})\right)\right] - \mathbb{E}_t[\langle \boldsymbol{\xi}_t, \mathbf{q}_t\rangle] \\
&\le \frac{1}{\eta_q}\ln\left(\sum_{i=1}^m q_{t,i}\mathbb{E}_t[\exp(\eta_q \xi_{t,i})]\right) \le \frac{\eta_q \sigma_q^2}{2}.
\end{aligned} \tag{59}$$

The second inequality uses conditional Jensen's inequality, the $\mathcal{F}_t$-measurability of $\mathbf{q}_t$, and $\mathbb{E}_t[\boldsymbol{\xi}_t] = \mathbf{0}_m$. The last inequality follows from (55) and $\sum_{i=1}^m q_{t,i} = 1$. Combining (57) and (59) gives

$$B_t \le \frac{D_{\mathrm{KL}}(\mathbf{q}^*_{s_c} \| \mathbf{q}_t) - \mathbb{E}_t[D_{\mathrm{KL}}(\mathbf{q}^*_{s_c} \| \bar{\mathbf{q}}_{t+1})]}{\eta_q} + \mathbb{E}_t\left[\phi(\mathbf{w}_t, \bar{\mathbf{q}}_{t+1}) - \phi(\mathbf{w}_t, \mathbf{q}_t)\right] + \frac{\eta_q \sigma_q^2}{2}. \tag{60}$$

Taking the total expectation and summing (60) over $t \in \mathcal{I}_c$, the KL-divergence terms telescope because $\bar{\mathbf{q}}_{t+1} = \mathbf{q}_{t+1}$ for every $t < e_c$. Dropping the final nonnegative divergence gives

$$\sum_{t \in \mathcal{I}_c} \mathbb{E}[B_t] \le \frac{\mathbb{E}[D_{\mathrm{KL}}(\mathbf{q}^*_{s_c} \| \mathbf{q}_{s_c})]}{\eta_q} + \mathbb{E}[J_c] + \frac{\eta_q \sigma_q^2}{2}|\mathcal{I}_c|, \tag{61}$$

where $J_c := \sum_{t=s_c}^{e_c} \left(\phi(\mathbf{w}_t, \bar{\mathbf{q}}_{t+1}) - \phi(\mathbf{w}_t, \mathbf{q}_t)\right)$. Since $\mathbf{q}_{s_c} = m^{-1}\mathbf{1}_m$, we obtain

$$D_{\mathrm{KL}}(\mathbf{q}^*_{s_c} \| \mathbf{q}_{s_c}) = \sum_{i=1}^m q^*_{s_c,i} \ln(m q^*_{s_c,i}) = \ln m + \sum_{i=1}^m q^*_{s_c,i} \ln q^*_{s_c,i} \le \ln m.$$

It remains to control the increments J_c. Since $\bar{\mathbf{q}}_{t+1} = \mathbf{q}_{t+1}$ for every $t < e_c$ and $\mathbf{q}_{s_c} = \mathbf{q}^{\mathrm{unif}}$, rearranging the sum gives

$$J_c = \phi(\mathbf{w}_{e_c}, \bar{\mathbf{q}}_{e_c+1}) - \phi(\mathbf{w}_{s_c}, \mathbf{q}^{\mathrm{unif}}) + \sum_{t=s_c}^{e_c-1} \left(\phi(\mathbf{w}_t, \mathbf{q}_{t+1}) - \phi(\mathbf{w}_{t+1}, \mathbf{q}_{t+1})\right). \tag{62}$$

For every nonfinal block $c < N$, we have $s_{c+1} = e_c + 1$ and $\mathbf{q}_{s_{c+1}} = \mathbf{q}^{\mathrm{unif}}$. Since $\phi(\mathbf{w}_{e_c}, \bar{\mathbf{q}}_{e_c+1}) \le \Phi(\mathbf{w}_{e_c})$ and both Φ and $\phi(\cdot, \mathbf{q})$ are G-Lipschitz,

$$\begin{aligned}
J_c &\le \Phi(\mathbf{w}_{e_c}) - \phi(\mathbf{w}_{s_c}, \mathbf{q}^{\mathrm{unif}}) + G\sum_{t=s_c}^{e_c-1} \|\mathbf{w}_{t+1} - \mathbf{w}_t\|_2 \\
&\le \Phi(\mathbf{w}_{e_c+1}) - \phi(\mathbf{w}_{e_c+1}, \mathbf{q}^{\mathrm{unif}}) + G\|\mathbf{w}_{e_c+1} - \mathbf{w}_{e_c}\|_2 + G\|\mathbf{w}_{e_c+1} - \mathbf{w}_{s_c}\|_2 \\
&\quad + G\sum_{t=s_c}^{e_c-1} \|\mathbf{w}_{t+1} - \mathbf{w}_t\|_2 \\
&\le \varepsilon^q_{s_{c+1}} + 2G\sum_{t=s_c}^{e_c} \|\mathbf{w}_{t+1} - \mathbf{w}_t\|_2.
\end{aligned} \tag{63}$$

For the final block, $e_N = T$. Applying (62), $\phi(\mathbf{w}_T, \bar{\mathbf{q}}_{T+1}) \le \Phi(\mathbf{w}_T)$, and the same Lipschitz bounds gives

$$\begin{aligned}
J_N &\le \Phi(\mathbf{w}_T) - \phi(\mathbf{w}_{s_N}, \mathbf{q}^{\mathrm{unif}}) + G\sum_{t=s_N}^{T-1} \|\mathbf{w}_{t+1} - \mathbf{w}_t\|_2 \\
&\le \Phi(\mathbf{w}_1) - \phi(\mathbf{w}_1, \mathbf{q}^{\mathrm{unif}}) + G\|\mathbf{w}_T - \mathbf{w}_1\|_2 + G\|\mathbf{w}_{s_N} - \mathbf{w}_1\|_2 + G\sum_{t=s_N}^{T-1} \|\mathbf{w}_{t+1} - \mathbf{w}_t\|_2 \\
&\le \varepsilon^q_1 + 2G\sum_{t=1}^{T-1} \|\mathbf{w}_{t+1} - \mathbf{w}_t\|_2.
\end{aligned} \tag{64}$$

Summing (63) over $c < N$, adding (64), and taking expectations yields

$$\sum_{c=1}^{N} \mathbb{E}[J_c] \le \sum_{t\in\mathcal{S}} \mathbb{E}[\varepsilon_t^q] + 4G \sum_{t=1}^{T-1} \mathbb{E}[\|\mathbf{w}_{t+1} - \mathbf{w}_t\|_2]. \tag{65}$$

Combining (53), (54), (61), and (65), and using $\sum_{c=1}^{N} |\mathcal{I}_c| = T$, gives

$$\sum_{t=1}^{T} \mathbb{E}[\varepsilon_t^q] \le 2G(B+2) \sum_{t=1}^{T-1} \mathbb{E}[\|\mathbf{w}_{t+1} - \mathbf{w}_t\|_2] + \frac{N \ln m}{\eta_q} + \frac{\eta_q \sigma_q^2 T}{2} + \sum_{t\in\mathcal{S}} \mathbb{E}[\varepsilon_t^q].$$

Subtracting the block-start gaps from both sides gives

$$\sum_{t\in\mathcal{J}} \mathbb{E}[\varepsilon_t^q] \le 2G(B+2) \sum_{t=1}^{T-1} \mathbb{E}[\|\mathbf{w}_{t+1} - \mathbf{w}_t\|_2] + \frac{N \ln m}{\eta_q} + \frac{\eta_q \sigma_q^2 T}{2}.$$

Finally, $N = \lceil T/B \rceil \le T/B + 1 \le 2T/B$ because $B \le T$. Dividing the preceding inequality by T proves (12).

F.6 Proof of Lemma 6

Let $\rho = \sqrt{1-\delta} \in [0, 1)$. By the definition of a δ-approximate compressor, for every deterministic $\mathbf{x} \in \mathbb{R}^d$,

$$\mathbb{E}_{\mathcal{C}}\left[\|\mathcal{C}(\mathbf{x}) - \mathbf{x}\|_2\right] \le \sqrt{\mathbb{E}_{\mathcal{C}}\left[\|\mathcal{C}(\mathbf{x}) - \mathbf{x}\|_2^2\right]} \le \rho \|\mathbf{x}\|_2. \tag{66}$$

Moreover, the same inequality holds conditionally for any random input independent of the fresh compressor realization.

General compressors. We first control the aggregated uplink residual. Define $S_t = \sum_{i=1}^{m} \mathbb{E}[\|\mathbf{e}_t^{(i)}\|_2]$. The uplink error-feedback recursion in Algorithm 2 and (66) give

$$\begin{aligned} \mathbb{E}\left[\|\mathbf{e}_{t+1}^{(i)}\|_2\right] &= \mathbb{E}\left[\left\|\mathbf{e}_t^{(i)} + q_{t,i}\mathbf{g}_{w,t}^{(i)} - \mathcal{C}\left(\mathbf{e}_t^{(i)} + q_{t,i}\mathbf{g}_{w,t}^{(i)}\right)\right\|_2\right] \\ &\le \rho \mathbb{E}\left[\left\|\mathbf{e}_t^{(i)} + q_{t,i}\mathbf{g}_{w,t}^{(i)}\right\|_2\right] \\ &\le \rho \mathbb{E}\left[\|\mathbf{e}_t^{(i)}\|_2\right] + \rho \mathbb{E}\left[q_{t,i}\|\mathbf{g}_{w,t}^{(i)}\|_2\right]. \end{aligned} \tag{67}$$

Assumption 3 and conditional Jensen's inequality imply $\mathbb{E}_t[\|\mathbf{g}_{w,t}^{(i)}\|_2] \le \Gamma_w$. Summing (67) over $i \in [m]$ and applying $\sum_{i=1}^{m} \mathbb{E}[q_{t,i}\|\mathbf{g}_{w,t}^{(i)}\|_2] \le \Gamma_w$ yields

$$S_{t+1} \le \rho S_t + \rho \Gamma_w. \tag{68}$$

Because $\mathbf{e}_1^{(i)} = \mathbf{0}_d$ for all $i \in [m]$, iterating (68) gives

$$S_{t+1} \le \sum_{k=1}^{t} \rho^k \Gamma_w \le \frac{\rho}{1-\rho}\Gamma_w, \qquad t \le T.$$

Recall that $\mathbf{e}_t = \sum_{i=1}^{m} \mathbf{e}_t^{(i)}$. By the triangle inequality,

$$\mathbb{E}\left[\|\mathbf{e}_t\|_2\right] \le S_t \le \frac{\rho}{1-\rho}\Gamma_w \le \frac{2-\delta}{\delta}\Gamma_w, \qquad t \le T+1, \tag{69}$$

where the last inequality follows from

$$\frac{\rho}{1-\rho} = \frac{\rho(1+\rho)}{\delta} = \frac{\sqrt{1-\delta}+1-\delta}{\delta} \le \frac{2-\delta}{\delta}.$$

We next control the downlink residual. Define the uncompressed weighted stochastic gradient $\mathbf{g}_{w,t} = \sum_{i=1}^{m} q_{t,i}\mathbf{g}_{w,t}^{(i)}$. Summing the uplink error-feedback recursions over the workers gives

$$\widehat{\mathbf{g}}_{w,t} = \sum_{i=1}^{m} \Delta_{w,t}^{(i)} = \mathbf{g}_{w,t} + \mathbf{e}_t - \mathbf{e}_{t+1}.$$

Combining $\mathbb{E}[\|\mathbf{g}_{w,t}\|_2] \le \Gamma_w$ with (69) yields

$$\begin{aligned}\mathbb{E}\left[\|\widehat{\mathbf{g}}_{w,t}\|_2\right] &\le \mathbb{E}\left[\|\mathbf{g}_{w,t}\|_2\right] + \mathbb{E}\left[\|\mathbf{e}_t\|_2\right] + \mathbb{E}\left[\|\mathbf{e}_{t+1}\|_2\right] \\ &\le \left(1 + \frac{2\rho}{1-\rho}\right)\Gamma_w = \frac{1+\rho}{1-\rho}\Gamma_w.\end{aligned} \tag{70}$$

The downlink error-feedback recursion and (66) imply

$$\begin{aligned}\mathbb{E}\left[\|\widehat{\mathbf{e}}_{t+1}\|_2\right] &= \mathbb{E}\left[\left\|\widehat{\mathbf{e}}_t + \widehat{\mathbf{g}}_{w,t} - \mathcal{C}\left(\widehat{\mathbf{e}}_t + \widehat{\mathbf{g}}_{w,t}\right)\right\|_2\right] \\ &\le \rho\mathbb{E}\left[\|\widehat{\mathbf{e}}_t + \widehat{\mathbf{g}}_{w,t}\|_2\right] \\ &\le \rho\mathbb{E}\left[\|\widehat{\mathbf{e}}_t\|_2\right] + \rho\mathbb{E}\left[\|\widehat{\mathbf{g}}_{w,t}\|_2\right].\end{aligned} \tag{71}$$

Using $\widehat{\mathbf{e}}_1 = \mathbf{0}_d$, substituting (70), and iterating (71), we obtain

$$\mathbb{E}\left[\|\widehat{\mathbf{e}}_t\|_2\right] \le \frac{\rho}{1-\rho}\frac{1+\rho}{1-\rho}\Gamma_w = \frac{\rho(1+\rho)}{(1-\rho)^2}\Gamma_w = \frac{\rho(1+\rho)^3}{\delta^2}\Gamma_w \le \frac{8}{\delta^2}\Gamma_w, \qquad t \le T+1.$$

Since $\mathrm{b}_t = \mathbf{e}_t + \widehat{\mathbf{e}}_t$, the two uniform residual bounds imply

$$\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}\left[\|\mathrm{b}_t\|_2\right] \le \frac{2-\delta}{\delta}\Gamma_w + \frac{8}{\delta^2}\Gamma_w = \frac{8+2\delta-\delta^2}{\delta^2}\Gamma_w.$$

Moreover, the error-feedback identity $\Delta_{w,t} = \mathbf{g}_{w,t} + \mathrm{b}_t - \mathrm{b}_{t+1}$ and $\mathbb{E}[\|\mathbf{g}_{w,t}\|_2] \le \Gamma_w$ give

$$\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}\left[\|\Delta_{w,t}\|_2\right] \le \Gamma_w + 2\left(\frac{2-\delta}{\delta} + \frac{8}{\delta^2}\right)\Gamma_w = \frac{16+4\delta-\delta^2}{\delta^2}\Gamma_w.$$

Additive-and-idempotent compressors with shared randomness. Suppose that $\mathcal{C} \in \mathcal{C}_{\mathrm{AI}}$ and that the server and workers use the same compressor realization at round t. Denote this realization by $\mathcal{C}^{(\xi_t)}$. By additivity,

$$\widehat{\mathbf{g}}_{w,t} = \sum_{i=1}^{m}\mathcal{C}^{(\xi_t)}\left(\mathbf{e}_t^{(i)} + q_{t,i}\mathbf{g}_{w,t}^{(i)}\right) = \mathcal{C}^{(\xi_t)}\left(\sum_{i=1}^{m}\mathbf{e}_t^{(i)} + \sum_{i=1}^{m}q_{t,i}\mathbf{g}_{w,t}^{(i)}\right). \tag{72}$$

Applying idempotence to (72) gives the pathwise identity

$$\mathcal{C}^{(\xi_t)}\left(\widehat{\mathbf{g}}_{w,t}\right) = \widehat{\mathbf{g}}_{w,t}. \tag{73}$$

We now prove by induction that $\widehat{\mathbf{e}}_t = \mathbf{0}_d$ for every $t \le T+1$. The claim holds at $t = 1$ by initialization. Suppose that $\widehat{\mathbf{e}}_t = \mathbf{0}_d$. Using (73),

$$\Delta_{w,t} = \mathcal{C}^{(\xi_t)}\left(\widehat{\mathbf{e}}_t + \widehat{\mathbf{g}}_{w,t}\right) = \mathcal{C}^{(\xi_t)}\left(\widehat{\mathbf{g}}_{w,t}\right) = \widehat{\mathbf{g}}_{w,t}.$$

Consequently, $\widehat{\mathbf{e}}_{t+1} = \widehat{\mathbf{e}}_t + \widehat{\mathbf{g}}_{w,t} - \Delta_{w,t} = \mathbf{0}_d$. Thus, we obtain $\widehat{\mathbf{e}}_t = \mathbf{0}_d$ for all $t \le T+1$.

Since $\widehat{\mathbf{e}}_t = \mathbf{0}_d$ pathwise, we have $\mathrm{b}_t = \mathbf{e}_t$ and $\Delta_{w,t} = \mathbf{g}_{w,t} + \mathbf{e}_t - \mathbf{e}_{t+1}$. Therefore,

$$\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}\left[\|\mathrm{b}_t\|_2\right] \le \frac{2-\delta}{\delta}\Gamma_w \le \frac{2\Gamma_w}{\delta},$$

and

$$\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}\left[\|\Delta_{w,t}\|_2\right] \le \Gamma_w + \frac{2(2-\delta)}{\delta}\Gamma_w = \frac{4-\delta}{\delta}\Gamma_w \le \frac{4\Gamma_w}{\delta}. \tag{74}$$