Byzantine-Resilient Zero-Order Optimization for Scalable Federated Fine-Tuning of Large Language Models

Maximilian Egger [1] **Mayank Bakshi** [2] **Rawad Bitar** [1]

Abstract

We introduce FEDBYZO, a Byzantine-resilient federated zero-order optimization method that is robust under Byzantine attacks and provides significant savings in uplink and downlink communication costs. We introduce transformed robust aggregation to give convergence guarantees for general non-convex objectives under client data heterogeneity. Empirical evaluations for standard learning tasks and fine-tuning large language models show that FEDBYZO exhibits stable performance with only a few scalars per-round communication cost and reduced memory requirements.

1 Introduction

Federated Learning (FL) (McMahan et al., 2017) enables model training across distributed clients without sharing raw data. However, it suffers from high communication costs since each client sends high dimensional updates to a central *federator*, which returns a global model. These challenges intensify with large language models (LLMs), increasing communication, privacy, and security concerns (Daly et al., 2024), prompting extensive research.

Communication cost and ZO. Reducing communication overhead has received significant attention (Wen et al., 2017; Karimireddy et al., 2019; Makkuva et al., 2024; Tang et al., 2024b; Qin et al., 2023). Zero order (ZO) optimization (Kiefer & Wolfowitz, 1952; Spall, 1992), which estimates gradients using random perturbations and loss evaluations, is increasingly popular. ZO methods: *(i)* avoid explicit gradients; *(ii)* enable training using forward passes only (Salimans et al., 2017; Ilyas et al., 2018; Liu et al., 2020; Malladi et al., 2023); and *(iii)* allow clients to send only a few scalars in FL (Fang et al., 2022; Qiu et al., 2023; Chen et al., 2023; Li et al., 2024). These advantages are especially useful for fine tuning tasks with low intrinsic dimensionality (Salimans et al., 2017; Malladi et al., 2023).

[1]School of Computation, Information and Technology, Technical University of Munich, Munich, Germany [2]School of Electrical, Computer and Energy Engineering, Arizona State University, Tempe, USA. Correspondence to: Maximilian Egger <maximilian.egger@tum.de>.

Byzantine clients. Robustness to adversarial (Byzantine) clients, those sending malicious updates, is essential for secure FL (Qi et al., 2024). Although privacy in ZO FL has been studied (Zhang et al., 2023; Tang et al., 2024a; Zhang et al., 2025), robustness to Byzantine clients has not. Even one Byzantine client can prevent convergence (Blanchard et al., 2017), making robust aggregation vital (Blanchard et al., 2017; Li et al., 2020; Yin et al., 2018; Allouah et al., 2023; Guerraoui et al., 2024). These defenses often reduce convergence speed, making communication efficiency even more important.

Data heterogeneity. FL clients often have data from different distributions, which hinders convergence (Zhao et al., 2018; Zhu et al., 2021) and may cause privacy leakage (Schlegel et al., 2023; Egger et al., 2023; Jahani-Nezhad et al., 2023; Tang et al., 2024a). This heterogeneity worsens robustness, as aggregation rules may downweight legitimate but outlying updates (El-Mhamdi et al., 2021; Karimireddy et al., 2022; Charikar et al., 2017; Liu et al., 2021). Nearest Neighbor Mixing (NNM) (Allouah et al., 2023) mitigates this by averaging each client's update with similar ones before aggregation.

Our contribution. We propose FEDBYZO, the first communication-efficient and Byzantine resilient FL framework using ZO optimization with robust aggregation. FEDBYZO supports any aggregation rule (e.g., trimmed mean (Yin et al., 2018), Krum (Blanchard et al., 2017)) and integrates preprocessing like NNM (Allouah et al., 2023). We prove convergence under general nonconvex losses and bounded heterogeneity (Wang et al., 2024), and validate FEDBYZO on MNIST and RoBERTa large (Liu, 2019) fine-tuning tasks. Experiments show FEDBYZO matches gradient based Byzantine resilient FL in accuracy while requiring only scalar communication and using less memory and computation.

2 System Model and Preliminaries

The L_2 norm and inner product are denoted by $\|\mathbf{x}\|$ and $\langle \mathbf{x}, \mathbf{y} \rangle$. Let $\mathbb{S}^d \triangleq \{\mathbf{x} \in \mathbb{R}^d : \|\mathbf{x}\|^2 = 1\}$ be the unit sphere and $\mathcal{U}(\mathbb{S}^d)$ the uniform distribution over $\mathbb{S}^d$. For $a \in \mathbb{N}$, define $[a] \triangleq 1, \ldots, a$. The horizontal stacking of vectors $\mathbf{v}_i$ is denoted by $(\mathbf{v}_i)_{i=1}^a = (\mathbf{v}_1, \ldots, \mathbf{v}_a)$.

Algorithm 1 FEDBYZO: Robust Efficient Zero-Order FL

Require: Shared seed for PRNG, $\mu \geq 0$, $\eta > 0$, $\nu > 0$, R.

Initialize and broadcast global model $\mathbf{w}^{(1)}$.
for $t = 1$ to T **do**
  for each client $i \in [n]$ **in parallel do**
    Initialize local model $\mathbf{w}_{t,1}^i = \mathbf{w}^{(t)}$.
    for $\ell = 1$ to K **do**
      Draw $\mathbf{z}_{t,\ell}^1, \cdots, \mathbf{z}_{t,\ell}^\nu \sim \mathcal{U}(\mathbb{S}^d)$, let $\mathbf{Z}_{t,\ell} \triangleq (\mathbf{z}_{t,\ell}^r)_{r\in[\nu]}$
      Compute $g_i(\mathbf{w}_{t,\ell}^i, \mathbf{z}_{t,\ell}^r) \triangleq g(\mathbf{w}_{t,\ell}^i, \mathbf{z}_{t,\ell}^r, \mu, \mathcal{D}_i), r \in [\nu]$ (cf. Definition 2.1)
      Let $\mathbf{g}_i(\mathbf{w}_{t,\ell}^i, \mathbf{Z}_{t,\ell}) \triangleq \frac{1}{\nu}((g_i(\mathbf{w}_{t,\ell}^i, \mathbf{z}_{t,\ell}^r))_{r=1}^\nu)^\top$
      Update $\mathbf{w}_{t,\ell+1}^i = \mathbf{w}_{t,\ell}^i - \eta \mathbf{Z}_{t,\ell} \mathbf{g}_i(\mathbf{w}_{t,\ell}^i, \mathbf{Z}_{t,\ell})$
    end for
    Send $\{\mathbf{g}_i(\mathbf{w}_{t,\ell}^i, \mathbf{Z}_{t,\ell})\}_{\ell=1}^K$ to federator.
  end for
  Aggregate $R_{t,\ell} = R(\{\mathbf{g}_i(\mathbf{w}_{t,\ell}^i, \mathbf{Z}_{t,\ell})\}_{i=1}^n), \ell \in [K]$
  Update $\mathbf{w}^{(t+1)} = \mathbf{w}^{(t)} - \eta \sum_{\ell=1}^K \mathbf{Z}_{t,\ell} R_{t,\ell}$.
  Broadcast $R_{t,\ell}$.
  Clients recover $\mathbf{w}^{(t+1)}$ using $R_{t,\ell}$ and the known $\mathbf{Z}_{t,\ell}$.
end for

In an FL setup with n clients and a federator, each client $i \in [n]$ holds a dataset $\mathcal{D}_i$, and the global dataset is $\mathcal{D} = \cup_i \mathcal{D}_i$. Let $F : \mathbb{R}^d \times \mathcal{D} \to \mathbb{R}^+$ be the loss. Define $F(\mathbf{w}, \tilde{\mathcal{D}}) \triangleq \sum_{D\in\tilde{\mathcal{D}}} F(\mathbf{w}, D)/|\tilde{\mathcal{D}}|$, and $F_i(\mathbf{w}) \triangleq F(\mathbf{w}, \mathcal{D}_i)$. For $\mathcal{A} \subseteq [n]$, let $F_\mathcal{A}(\mathbf{w}) \triangleq \frac{1}{|\mathcal{A}|}\sum_{i\in\mathcal{A}} F_i(\mathbf{w})$ and $F_\mathcal{A}^\star \triangleq \min_{\mathbf{w}\in\mathbb{R}^d} F_\mathcal{A}(\mathbf{w})$. The goal is to minimize $F_{[n]}(\mathbf{w})$. All clients and the federator initialize with a model $\mathbf{w}^{(1)} \in \mathbb{R}^d$ and iteratively generate models $\{\mathbf{w}^{(t)}\}_{t\in[T]}$ over T global epochs. In each epoch t, client i receives $\mathbf{w}^{(t)}$, computes an update using one or more mini batches $\tilde{\mathcal{D}}_i \subseteq \mathcal{D}_i$, and sends it to the federator. The federator aggregates all updates into $\mathbf{w}^{(t+1)}$ and broadcasts it. For a mini batch $\tilde{\mathcal{D}}_i$, define $\mathbf{g}_i \triangleq \nabla F(\mathbf{w}^{(t)}, \tilde{\mathcal{D}}_i)$.

We assume fewer than half the clients $b < n/2$ are Byzantine, and may fully coordinate, knowing the algorithm, defenses, and honest outputs. Let $\mathcal{H}$ be the set of honest clients with $|\mathcal{H}| = n - b$. The goal becomes minimizing $F_\mathcal{H}(\mathbf{w})$.

ZO gradients are estimated by querying the loss along perturbation directions. We use the two point ZO estimator:

Definition 2.1 (Two-Point Zero-Order Estimate)**.** Let $\mathbf{z} \in \mathbb{S}^d$, $\tilde{\mathcal{D}} \subseteq \mathcal{D} \setminus \emptyset$ and $\mathbf{w} \in \mathbb{R}^d$. The two-point ZO estimate of the gradient $\mathbf{g} \triangleq \nabla F(\mathbf{w}, \tilde{\mathcal{D}})$ in direction $\mathbf{z}$ is defined as $\mathbf{z}g(\mathbf{w}, \mathbf{z}, \mu, \tilde{\mathcal{D}})$, where

$$g(\mathbf{w}, \mathbf{z}, \mu, \tilde{\mathcal{D}}) \triangleq \begin{cases} d\frac{F(\mathbf{w}+\mu\mathbf{z}, \tilde{\mathcal{D}}) - F(\mathbf{w}-\mu\mathbf{z}, \tilde{\mathcal{D}})}{2\mu} & \mu > 0 \\ d\left\langle \nabla F(\mathbf{w}, \tilde{\mathcal{D}}), \mathbf{z} \right\rangle & \mu = 0 \end{cases}$$

In Byzantine resilient FL, the federator uses a *robust aggregation* rule $R(\cdot)$, which defaults to averaging when all clients are honest. We adopt the robustness notion from (Allouah et al., 2023), where the parameter κ quantifies resilience to up to b Byzantine clients.

Definition 2.2 ((b, κ)-Robust Aggregation)**.** Let $\kappa \geq 0$ and $b < n/2$. For vectors $\mathbf{v}_1, \cdots, \mathbf{v}_n$ and any set $\mathcal{H} \subset [n]$ of size $|\mathcal{H}| = n - b$, letting $\bar{\mathbf{v}}_\mathcal{H} = \frac{1}{|\mathcal{H}|}\sum_{i\in\mathcal{H}} \mathbf{v}_i$, an aggregation rule $R(\{\mathbf{v}_i\}_{i=1}^n)$ is (b, κ)-robust if

$$\|R(\{\mathbf{v}_i\}_{i=1}^n) - \bar{\mathbf{v}}_\mathcal{H}\|^2 \leq \frac{\kappa}{|\mathcal{H}|}\sum_{i\in\mathcal{H}} \|\mathbf{v}_i - \bar{\mathbf{v}}_\mathcal{H}\|^2.$$

3 Overview of FEDBYZO

FEDBYZO (Algorithm 1) begins with a shared seed that enables the federator and all clients to generate identical random perturbation vectors $\mathbf{z} \sim \mathcal{U}(\mathbb{S}^d)$ via a common pseudorandom number generator (PRNG). The initial model is set to $\mathbf{w}^{(1)}$ on all parties. The algorithm proceeds for T global epochs, each comprising K local epochs. Let $\mu \geq 0$ be the ZO estimator scale and η the learning rate. At the start of global epoch t, each client sets $\mathbf{w}_{t,1}^i = \mathbf{w}^{(t)}$.

During each local epoch $\ell \in [K]$, the parties generate $\nu \geq 1$ pseudorandom perturbation vectors $\mathbf{z}_{t,\ell}^1, \ldots, \mathbf{z}_{t,\ell}^\nu$. In the unbiased variant, new vectors are sampled at each ℓ. In the biased variant, the same set of perturbations is reused across local epochs, i.e., $\mathbf{z}_{t,\ell}^r = \mathbf{z}_{t,1}^r$ for $\ell > 1$.

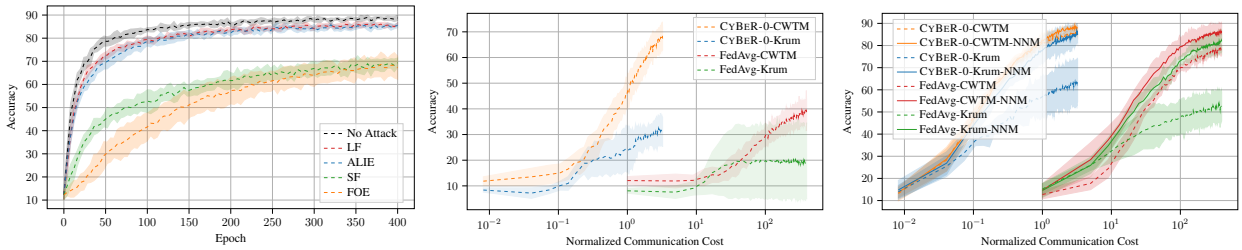
Each client samples a mini-batch $\tilde{\mathcal{D}}_i$ and computes two-point ZO gradient estimates $\mathbf{z}_{t,\ell}^r g(\mathbf{w}_{t,\ell}^i, \mathbf{z}_{t,\ell}^r, \mu, \tilde{\mathcal{D}}_i)$ for all $r \in [\nu]$, where $\mathbf{w}_{t,\ell}^i$ is the local model. Defining $g_i(\mathbf{w}_{t,\ell}^i, \mathbf{z}_{t,\ell}^r) \triangleq g(\mathbf{w}_{t,\ell}^i, \mathbf{z}_{t,\ell}^r, \mu, \tilde{\mathcal{D}}_i)$ and stacking the perturbations and scalar projections into $\mathbf{Z}_{t,\ell} \in \mathbb{R}^{d\times\nu}$ and $\mathbf{g}_i(\mathbf{w}_{t,\ell}^i, \mathbf{Z}_{t,\ell}) \in \mathbb{R}^\nu$, the model is updated as $\mathbf{w}_{t,\ell+1}^i = \mathbf{w}_{t,\ell}^i - \eta \mathbf{Z}_{t,\ell} \mathbf{g}_i(\mathbf{w}_{t,\ell}^i, \mathbf{Z}_{t,\ell})$. At the end of local training, each client sends scalar projections of the cumulative update to the federator. This entails a communication cost of $K\nu$ scalars in the unbiased case (transmitting each $\mathbf{g}_i(\mathbf{w}_{t,\ell}^i, \mathbf{Z}_{t,\ell})$) or ν scalars in the biased case (transmitting $\sum_{\ell=1}^K \mathbf{g}_i(\mathbf{w}_{t,\ell}^i, \mathbf{Z}_{t,\ell})$).

The federator performs robust aggregation in the projection space $\mathbb{R}^\nu$ using $R : (\mathbb{R}^\nu)^n \to \mathbb{R}^\nu$. In the unbiased case, aggregation is performed separately per ℓ, yielding $\mathbf{w}^{(t+1)} = \mathbf{w}^{(t)} - \eta \sum_{\ell=1}^K \mathbf{Z}_{t,\ell} R\left(\{\mathbf{g}_i(\mathbf{w}_{t,\ell}^i, \mathbf{Z}_{t,\ell})\} i \in [n]\right)$. In the biased case, we aggregate over the summed projections, i.e., $\mathbf{w}^{(t+1)} = \mathbf{w}^{(t)} - \eta \mathbf{Z}_{t,1} R\big(\big\{\sum_{\ell=1}^K \mathbf{g}_i(\mathbf{w}_{t,\ell}^i, \mathbf{Z}_{t,\ell})\big\}_{i\in[n]}\big)$.

Finally, the federator broadcasts $\mathbf{w}^{(t+1)}$ via its projection along the shared perturbation vectors $\mathbf{z}_{t,\ell}^r$. Since clients know both $\mathbf{w}^{(t)}$ and the perturbations, they can reconstruct $\mathbf{w}^{(t+1)}$ locally. The update always lies in the span of the perturbation directions, so this projection suffices.

3.1 Choice of Robust Aggregation Rule

FEDBYZO supports any robust aggregation rule. Its theoretical guarantees depend on b and κ (see Definition 2.2).



(a) Attacks on FEDBYZO-CWTM. (b) Worst case attack accuracies. (c) Label flipping attack.

Figure 1. Performance of different robust aggregation rules against different attacks for logistic regression on MNIST.

We use the following standard methods in our experiments:

- **Coordinate-wise trimmed mean (CWTM)** (Yin et al., 2018): Removes the smallest and largest $\lfloor \beta n \rfloor$ values per coordinate from n vectors, then averages the rest. For an input $\mathcal{X} = \{\mathbf{v}_1, \ldots, \mathbf{v}_n\} \subset \mathbb{R}^\nu$, CWTM returns a vector where each entry is the coordinate-wise trimmed mean.
- **Krum** (Blanchard et al., 2017): Selects a single vector whose sum of distances to its $n - b - 2$ closest neighbors is minimal. For an input $\mathcal{X} = \{\mathbf{v}_1, \ldots, \mathbf{v}_n\}$, the selected vector is $\mathbf{v}_{i^\star}$, where $i^\star = \arg\min_i \sum_{j \in \mathcal{C}_i} \|\mathbf{v}_i - \mathbf{v}_j\|$.
- **Nearest neighbor mixing (NNM)** (Allouah et al., 2023): A pre-processing step that enhances robustness under data heterogeneity. Each client vector $\mathbf{v}_i$ is replaced by the average of its $n - b$ nearest neighbors in Euclidean distance.

3.2 Innovations in FEDBYZO

FEDBYZO introduces several innovations for efficient and resilient ZO-based FL: (1) Transformed robust aggregation aggregates updates directly in the perturbation space $\mathbb{R}^\nu$, avoiding projection errors from nonlinear aggregation and preserving robustness via Johnson–Lindenstrauss embeddings (Johnson & Lindenstrauss, 1984); (2) Communication efficiency is achieved as clients send only scalar projections and the federator broadcasts updates in the same subspace; (3) a shared seed protocol ensures synchronized perturbations without transmitting high-dimensional vectors; (4) support for multiple local epochs reduces communication frequency, with a tradeoff between unbiased (accurate, costly) and biased (compact, biased) ZO estimators; (5) FEDBYZO inherits ZO memory savings, reducing inference memory up to $12\times$ (Malladi et al., 2023), enabling deployment on constrained devices (cf. Appendix B for details).

4 Experimental Evaluation

Logistic Regression on MNIST. We evaluate FEDBYZO on MNIST under various Byzantine attacks, comparing it to FedAvg and FedZO with standard robust aggregation (CWTM with $\beta = b/n$, Krum). FEDBYZO consistently achieves higher worst-case accuracy while reducing communication by several orders of magnitude.

With transformed CWTM and Krum (without NNM), FEDBYZO reaches 69.9% accuracy under worst-case attacks, compared to 61.8% for the best non-transformed FedZO variant and 58.5% for FedAvg with NNM, an improvement of over 8%. Additionally, uplink and downlink costs are drastically reduced by operating in the perturbation space.

Figure 1 shows accuracy vs. global epochs and communication cost. Figure 1(a) summarizes performance across attacks, with FOE causing the most degradation. Figure 1(b) shows that FEDBYZO improves worst-case performance by over 10% with Krum and over 25% with CWTM. Figure 1(c) shows robustness gains against LF when using NNM. Further results are in Appendix A.6. We also compare local update strategies in Appendix A.4, highlighting a bias-variance-efficiency trade-off. The effect of ν on performance is studied in Appendix A.5.

Fine-Tuning Large Language Models. Following (Malladi et al., 2023; Li et al., 2024), we fine-tune RoBERTa-large (main paper) and OPT-125M (appendix) on SST-2, SNLI, TREC, RTE, and MNLI, using fixed CWTM aggregation. The results are averaged over 3 runs. Table 1 reports performance under ALIE (Baruch et al., 2019), FOE (Xie et al., 2020), SF (Allen-Zhu et al., 2021), and TMA (Algorithm 4) in a non-i.i.d. setting ($\alpha = 1$), with $n = 12$, $b = 3$, and $\nu = 1$. Worst-case accuracy drops were modest: $\sim 1\%$ on SST-2 and 5–8% on SNLI, TREC, RTE, and MNLI. FOE remains the most effective attack (Figure 2). Results under i.i.d. data ($\alpha = \infty$) show similar stability. Full results are in Appendix A.8.

We perform an extensive hyperparameter study (FOE on

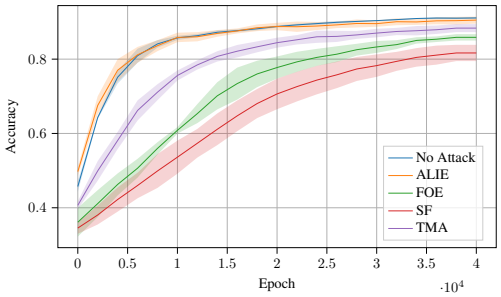


Figure 2. Accuracy over epochs for fine-tuning RoBERTa-large on TREC under different attack scenarios for non-i.i.d. data.

Table 1. Mean and standard deviation of maximum accuracies across seeds for fine-tuning RoBERTa-large with non-i.i.d. data distribution

Dataset	ALIE	FOE	SF	TMA	No Attack	Worst Case
SST-2	93.3 ± 0.4	91.7 ± 1.5	91.6 ± 0.5	92.1 ± 1.2	92.9 ± 0.1	91.6
TREC	93.3 ± 1.2	91.7 ± 1.0	87.7 ± 2.9	92.2 ± 1.1	94.7 ± 0.7	87.7
SNLI	80.8 ± 1.4	78.4 ± 2.1	76.1 ± 1.0	80.0 ± 0.7	84.3 ± 0.2	76.1
RTE	78.4 ± 0.3	76.9 ± 0.3	75.6 ± 2.2	77.7 ± 1.1	80.0 ± 0.8	75.6
MNLI	72.8 ± 1.6	69.2 ± 2.1	68.0 ± 1.1	71.2 ± 1.1	76.4 ± 1.1	68.0

SST-2). Under extreme heterogeneity ($\alpha = 0.1$), robustness improves with larger n. Varying ν (with $\nu T = 20000$ fixed) confirms a near-inverse tradeoff between ν and T. Similar trends hold when varying local epochs K. Performance remains stable up to $b = 6$ (out of 16 clients). Even in the extreme case ($\alpha \to 0$), FEDBYZO achieves 90.9% on SST-2 vs. 91.6% in i.i.d.. See Appendix A.9 for details.

5 Convergence Analysis

We establish convergence guarantees for FEDBYZO under arbitrary (b, κ)-robust aggregation rules, adapted to ZO updates, with heterogeneous data and non-convex losses. Proofs are in Appendix C. Extending (Allouah et al., 2023) to the ZO setting is challenging due to the bias and high variance of ZO estimates, which are aggregated in a transformed space where robust aggregation outputs may not align with the honest mean. Prior work on ZO FL (Fang et al., 2022) does not consider Byzantine clients or the practical heterogeneity model of (Wang et al., 2024), which itself does not apply to ZO methods due to estimator-specific challenges.

We prove that FEDBYZO converges in this setting using projection theorems. The combination of ZO updates, transformed aggregation, adversaries, heterogeneity, and non-convexity introduces novel theoretical complexity. Our analysis assumes Lipschitz smoothness and bounded gradient variance (Assumptions 5.1–5.2), and bounded heterogeneity (Assumptions 5.3–5.4) per (Wang et al., 2024). Assumption 5.5 captures fine-tuning regimes, where the initial model is close to a local optimum.

Assumption 5.1 (Lipschitz gradient)**.** For all $\mathbf{w}, \omega \in \mathbb{R}^d$ and $i \in [n]$, $\|\nabla F_i(\mathbf{w}) - \nabla F_i(\omega)\| \leq L \|\mathbf{w} - \omega\|$.

Assumption 5.2 (Bounded gradient variance)**.** The variance of the clients' gradient estimate is uniformly bounded by σ^2, *i.e.*, $\mathbb{E}\left[\|\mathbf{g}_i(\mathbf{w}) - \nabla F_i(\mathbf{w})\|^2\right] \leq \sigma^2 \forall \mathbf{w}, i \in [n]$.

Assumption 5.3 (Bounded gradient divergence)**.** $\|\nabla F_i(\mathbf{w}) - \nabla F_{\mathcal{H}}(\mathbf{w})\|^2 \leq \zeta^2, \forall i \in [n]$.

Assumption 5.4 (Pseudo-Lipschitz on averaged gradients)**.** $\left\|\frac{1}{n}\sum_{i=1}^n \nabla F_i(\mathbf{w}_i) - \nabla F_{\mathcal{H}}(\bar{\mathbf{w}})\right\|^2 \leq \frac{D^2}{n}\sum_{i=1}^n \|\mathbf{w}_i - \bar{\mathbf{w}}\|^2$.

Assumption 5.5 (Lipschitz Objective function)**.** $\forall i \in [n]$, $\|F_i(\mathbf{w}) - F_i(\omega)\| \leq G \|\mathbf{w} - \omega\|, \forall \mathbf{w}, \omega \in \mathbb{R}^d$.

Our first result establishes the convergence rate of FEDBYZO in the setting of fine-tuning. Let $\epsilon = \sqrt{\frac{64}{\nu}\log(\frac{2(|\mathcal{H}|-1)}{\delta})}$, and $\epsilon' \triangleq \frac{(1+\epsilon)}{(1-\epsilon)}$. In this setting, we show that as long as the number of perturbation directions is logarithmic in the number of optimization steps, FEDBYZO achieves a convergence rate that is consistent with the non-Byzantine literature (Li et al., 2024).

Theorem 5.6 (Lipschitz objective functions)**.** *Let $0 < \Delta < 1$, and suppose that Assumptions 5.1 to 5.5 hold. Consider* FEDBYZO *with $\mu > 0$ and a $(|\mathcal{H}|, \kappa)$-robust aggregation rule. If* (a) $\eta \leq \min\left\{\frac{1}{26KL}, \frac{1}{4K\sqrt{D^2+\zeta^2}}\right\}$, *and* (b) $\nu \geq 64\log(2(|\mathcal{H}|-1)TK/\Delta)$, *the following convergence guarantee holds with probability at least $1-\Delta$ and for a suitable numerical constant $\varphi > 0$:*

$$\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}\left[\|\nabla F_{\mathcal{H}}(\mathbf{w}_t)\|^2\right] \leq \frac{4(F_{\mathcal{H}}(\mathbf{w}_1) - F_{\mathcal{H}}^\star)}{\eta TK}$$
$$+12KL\eta\left(\kappa\epsilon'+\mu\right)+2\eta\frac{\varphi^2G^2d}{\nu}\left(\frac{1}{5|\mathcal{H}|}+\frac{8}{|\mathcal{H}|^2}+2K\epsilon'\kappa\right)$$
$$+2\eta KD^2\left(\frac{\varphi^2G^2d}{L^2\nu}+\frac{3}{L}\right)\left(13K+24\epsilon'\kappa(1+\frac{\zeta^2}{D^2})+4\right).$$

We now consider general non-convex loss landscapes without the fine-tuning assumption and establish convergence rates matching those in the non-Byzantine ZO literature (Li et al., 2024). Specifically, with $K = O(1)$, $\nu = \mathcal{O}(d)$, and a sufficiently small learning rate, we obtain a convergence rate of $O(1/T) + O(1)$. See Theorem C.3 in Appendix C for the full statement and parameter dependencies.

Theorem 5.7 (General non-convex landscapes)**.** *Let $0 < \Delta < 1$ and suppose Assumptions 5.1 to 5.4 hold. Consider* FEDBYZO *with a $(|\mathcal{H}|, \kappa)$-robust aggregation rule. If* (a) $\eta^2 \leq \min\{\frac{1}{72K^2L}, \frac{|\mathcal{H}|\nu}{96KdL^2}, 6\frac{D^2}{K^2}+6\frac{\zeta^2}{K^2}+\frac{32d}{\nu K^2}L^2\}$, (b) $\nu \geq 64\log(\frac{2(|\mathcal{H}|-1)TK}{\Delta})$, *and* (c) $\nu = \mathcal{O}(d)$, *then the following convergence guarantee holds with probability at least $1-\Delta$:*

$$\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}\left[\|\nabla F_{\mathcal{H}}(\mathbf{w}_t)\|^2\right] \leq \frac{4(F_{\mathcal{H}}(\mathbf{w}_1) - F_{\mathcal{H}}^\star)}{\eta KT} + O(1).$$

6 Conclusion

We proposed FEDBYZO, a communication-efficient, Byzantine-resilient FL framework using ZO optimization and transformed robust aggregation. By working in the perturbation space and using shared seeds, it achieves strong robustness with minimal communication, matching state-of-the-art performance while reducing communication by up to seven orders of magnitude.

Acknowledgments

This material is based upon work supported by the National Science Foundation under Grant Nos. CCF-1908725 and CCF-2107526, and the German Research Foundation (DFG) under Grant Agreement Nos. BI 2492/1-1 and WA 3907/7-1.

References

Allen-Zhu, Z., Ebrahimianghazani, F., Li, J., and Alistarh, D. Byzantine-resilient non-convex stochastic gradient descent. In *International Conference on Learning Representations*, 2021.

Allouah, Y., Farhadkhani, S., Guerraoui, R., Gupta, N., Pinot, R., and Stephan, J. Fixing by mixing: A recipe for optimal byzantine ml under heterogeneity. In *International Conference on Artificial Intelligence and Statistics*, pp. 1232–1300, 2023.

Baruch, G., Baruch, M., and Goldberg, Y. A little is enough: Circumventing defenses for distributed learning. *Advances in Neural Information Processing Systems*, 32, 2019.

Blanchard, P., El Mhamdi, E. M., Guerraoui, R., and Stainer, J. Machine learning with adversaries: Byzantine tolerant gradient descent. *Advances in neural information processing systems*, 30, 2017.

Charikar, M., Steinhardt, J., and Valiant, G. Learning from untrusted data. In *ACM SIGACT Symposium on Theory of Computing*, pp. 47–60, 2017.

Chen, J., Chen, H., Gu, B., and Deng, H. Fine-grained theoretical analysis of federated zeroth-order optimization. In *Neural Information Processing Systems*, 2023.

Daly, K., Eichner, H., Kairouz, P., McMahan, H. B., Ramage, D., and Xu, Z. Federated learning in practice: reflections and projections. In *IEEE International Conference on Trust, Privacy and Security in Intelligent Systems, and Applications*, pp. 148–156. IEEE, 2024.

Egger, M., Hofmeister, C., Wachter-Zeh, A., and Bitar, R. Private aggregation in wireless federated learning with heterogeneous clusters. In *IEEE International Symposium on Information Theory (ISIT)*, pp. 54–59, 2023.

El-Mhamdi, E. M., Farhadkhani, S., Guerraoui, R., Guirguis, A., Hoang, L.-N., and Rouault, S. Collaborative learning in the jungle (decentralized, byzantine, heterogeneous, asynchronous and nonconvex learning). *Advances in neural information processing systems*, 34:25044–25057, 2021.

Fang, W., Yu, Z., Jiang, Y., Shi, Y., Jones, C. N., and Zhou, Y. Communication-efficient stochastic zeroth-order optimization for federated learning. *IEEE Transactions on Signal Processing*, 70:5058–5073, 2022.

Flaxman, A. D., Kalai, A. T., and McMahan, H. B. Online convex optimization in the bandit setting: gradient descent without a gradient. In *ACM-SIAM Symposium on Discrete Algorithms*, pp. 385–394, 2005.

Gao, X., Jiang, B., and Zhang, S. On the information-adaptive variants of the admm: an iteration complexity perspective. *Journal of Scientific Computing*, 76:327–363, 2018.

Guerraoui, R., Gupta, N., and Pinot, R. *Robust Machine Learning*. Springer, 2024.

Ilyas, A., Engstrom, L., Athalye, A., and Lin, J. Black-box adversarial attacks with limited queries and information. In *International Conference on Machine Learning*, pp. 2137–2146, 2018.

Jahani-Nezhad, T., Maddah-Ali, M. A., Li, S., and Caire, G. SwiftAgg+: Achieving asymptotically optimal communication loads in secure aggregation for federated learning. *IEEE Journal on Selected Areas in Communications*, 41 (4):977–989, 2023.

Johnson, W. and Lindenstrauss, J. Extensions of lipschitz maps into a hilbert space. *Contemporary Mathematics*, 26:189–206, 01 1984.

Karimireddy, S. P., Rebjock, Q., Stich, S., and Jaggi, M. Error feedback fixes SignSGD and other gradient compression schemes. In *International Conference on Machine Learning*, pp. 3252–3261, 2019.

Karimireddy, S. P., He, L., and Jaggi, M. Byzantine-robust learning on heterogeneous datasets via bucketing. In *International Conference on Learning Representations*, 2022.

Kiefer, J. and Wolfowitz, J. Stochastic Estimation of the Maximum of a Regression Function. *The Annals of Mathematical Statistics*, 23(3):462 – 466, 1952.

Li, S., Cheng, Y., Wang, W., Liu, Y., and Chen, T. Learning to detect malicious clients for robust federated learning. *arXiv preprint arXiv:2002.00211*, 2020.

Li, Y. Simple, unified analysis of johnson-lindenstrauss with applications. *arXiv preprint arXiv:2402.10232*, 2024.

Li, Z., Ying, B., Liu, Z., Dong, C., and Yang, H. Achieving dimension-free communication in federated learning via zeroth-order optimization. *arXiv preprint arXiv:2405.15861*, 2024.

Liu, S., Chen, P.-Y., Kailkhura, B., Zhang, G., Hero III, A. O., and Varshney, P. K. A primer on zeroth-order optimization in signal processing and machine learning: Principals, recent advances, and applications. *IEEE Signal Processing Magazine*, 37(5):43–54, 2020.

Liu, S., Gupta, N., and Vaidya, N. H. Approximate byzantine fault-tolerance in distributed optimization. In *ACM Symposium on Principles of Distributed Computing*, pp. 379–389, 2021.

Liu, Y. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*, 364, 2019.

Makkuva, A. V., Bondaschi, M., Vogels, T., Jaggi, M., Kim, H., and Gastpar, M. LASER: Linear compression in wireless distributed optimization. In *International Conference on Machine Learning*, 2024.

Malladi, S., Gao, T., Nichani, E., Damian, A., Lee, J. D., Chen, D., and Arora, S. Fine-tuning language models with just forward passes. *Advances in Neural Information Processing Systems*, 36:53038–53075, 2023.

McMahan, B., Moore, E., Ramage, D., Hampson, S., and y Arcas, B. A. Communication-efficient learning of deep networks from decentralized data. In *Artificial intelligence and statistics*, pp. 1273–1282, 2017.

Qi, X., Huang, Y., Zeng, Y., Debenedetti, E., Geiping, J., He, L., Huang, K., Madhushani, U., Sehwag, V., Shi, W., et al. AI risk management should incorporate both safety and security. *arXiv preprint arXiv:2405.19524*, 2024.

Qin, Z., Chen, D., Qian, B., Ding, B., Li, Y., and Deng, S. Federated full-parameter tuning of billion-sized language models with communication cost under 18 kilobytes. *arXiv preprint arXiv:2312.06353*, 2023.

Qiu, Y., Shanbhag, U., and Yousefian, F. Zeroth-order methods for nondifferentiable, nonconvex, and hierarchical federated optimization. *Advances in Neural Information Processing Systems*, 36, 2023.

Salimans, T., Ho, J., Chen, X., Sidor, S., and Sutskever, I. Evolution strategies as a scalable alternative to reinforcement learning. *arXiv preprint arXiv:1703.03864*, 2017.

Salmon, J. K., Moraes, M. A., Dror, R. O., and Shaw, D. E. Parallel random numbers: as easy as 1, 2, 3. In *International conference for high performance computing, networking, storage and analysis*, pp. 1–12, 2011.

Schlegel, R., Kumar, S., Rosnes, E., and i Amat, A. G. CodedPaddedFL and CodedSecAgg: Straggler mitigation and secure aggregation in federated learning. *IEEE Transactions on Communications*, 2023.

Spall, J. Multivariate stochastic approximation using a simultaneous perturbation gradient approximation. *IEEE Transactions on Automatic Control*, 37(3):332–341, 1992.

Tang, X., Panda, A., Nasr, M., Mahloujifar, S., and Mittal, P. Private fine-tuning of large language models with zeroth-order optimization. *CoRR*, 2024a.

Tang, Y., Zhang, J., and Li, N. Distributed zero-order algorithms for nonconvex multiagent optimization. *IEEE Transactions on Control of Network Systems*, 8(1):269–281, 2020.

Tang, Z., Wang, Y., and Chang, T.-H. z-SignFedAvg: A unified stochastic sign-based compression for federated learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pp. 15301–15309, 2024b.

Wang, J., Liu, Q., Liang, H., Joshi, G., and Poor, H. V. A novel framework for the analysis and design of heterogeneous federated learning. *IEEE Transactions on Signal Processing*, 69:5234–5249, 2021.

Wang, J., Wang, S., Chen, R.-R., and Ji, M. A new theoretical perspective on data heterogeneity in federated optimization. In *International Conference on Machine Learning*, 2024.

Wen, W., Xu, C., Yan, F., Wu, C., Wang, Y., Chen, Y., and Li, H. TernGrad: Ternary gradients to reduce communication in distributed deep learning. *Advances in neural information processing systems*, 30, 2017.

Xie, C., Koyejo, O., and Gupta, I. Fall of empires: Breaking byzantine-tolerant sgd by inner product manipulation. In *Uncertainty in Artificial Intelligence*, pp. 261–270, 2020.

Yin, D., Chen, Y., Ramchandran, K., and Bartlett, P. Byzantine-robust distributed learning: Towards optimal statistical rates. In *International conference on machine learning*, pp. 5650–5659, 2018.

Zhang, J., Chen, E., Liu, C., and Brinton, C. G. DPZV: Resource efficient zo optimization for differentially private vfl. *arXiv preprint arXiv:2502.20565*, 2025.

Zhang, L., Li, B., Thekumparampil, K. K., Oh, S., and He, N. DPZero: Private fine-tuning of language models without backpropagation. *arXiv preprint arXiv:2310.09639*, 2023.

Zhao, Y., Li, M., Lai, L., Suda, N., Civin, D., and Chandra, V. Federated learning with non-iid data. *arXiv preprint arXiv:1806.00582*, 2018.

Zhu, H., Xu, J., Liu, S., and Jin, Y. Federated learning on non-iid data: A survey. *Neurocomputing*, 465:371–390, 2021.

A Numerical Experiments

A.1 Experimental Details

A.1.1 Sampling of Perturbation Vectors

To sample the directions $\mathbf{z}$, for fine-tuning large language models we use a practical approach similar to (Salimans et al., 2017; Malladi et al., 2023) that draws each coordinate independently from a standard Gaussian distribution. This minor modification has substantial practical implications by alleviating the allocation of the entire vector, and instead iteratively samples each model coordinate. Thereby, considerably reducing the memory footprint of our method.

A.1.2 Reconstruction of the Seed

Let s be a seed initially broadcast by the federator to all clients. Then, at each global and local iteration t and ℓ, the r-th random perturbation is sampled by setting the seed of the PRNG to $s' \triangleq (s, t, \ell, r)$. In this way, the perturbations $\mathbf{z}_{t,\ell}^r$ sampled by all clients will be equivalent. The client then compute the estimate according to Definition 2.1.

A.1.3 In-Place Model Perturbation

Similar to (Salimans et al., 2017; Malladi et al., 2023), we use in-place perturbations of the model for memory efficient zero-order optimization throughout the training phase. In particular, client i employs Algorithms 2 and 3 to compute the zero-order estimate as $g_i(\mathbf{w}_{t,\ell}^i, \mathbf{z}_{t,\ell}^r) = \textsc{ZeroOrderEstimate}(\mathbf{w}_{t,\ell}^i, s', \mu, \mathcal{D}_i)$. Note that the function, instead of $\mathbf{z}_{t,\ell}^r$, takes as input the seed $s' = (s, t, \ell, r)$ used to reconstruct the projection $\mathbf{z}_{t,\ell}^r$.

Algorithm 2 PERTURB: Perturbing Model Parameters

Input: Model parameters $\mathbf{w}$, scaling factor μ, seed s'
Output: Perturbed model $\mathbf{w}$
Initialize PRNG with seed s'
for $p = 1$ to d **do**
  Sample $z \sim \mathcal{N}(0, 1)$
  Perturb parameter $\mathbf{w}^{(p)} = \mathbf{w}^{(p)} + \mu \cdot z$
end for
Return: Perturbed model $\mathbf{w}$

Algorithm 3 ZEROORDERESTIMATE: Compute $g(\mathbf{w}, s', \mu, \mathcal{D})$ via Model Perturbation

Input: Model $\mathbf{w}$, seed s', scaling factor μ, data $\mathcal{D}$
Output: Zero estimate $g(\mathbf{w}, s', \mu, \mathcal{D})$
Step 1: $\texttt{Perturb}(\mathbf{w}, \mu, s')$ (cf. Algorithm 2)
Compute $F_1 = F(\mathbf{w}^{(p)}, \mathcal{D})$
Step 2: $\texttt{Perturb}(\mathbf{w}, -2\mu, s')$
Compute $F_2 = F(\mathbf{w}^{(p)}, \mathcal{D})$
Step 3: $\texttt{Perturb}(\mathbf{w}, \mu, s')$ {Reset the model}
Return: $g(\mathbf{w}, s', \mu, \mathcal{D}) = \frac{F_1 - F_2}{2\mu}$

A.2 Byzantine Attacks

We test and compare our algorithm using several state-of-the-art gradient attacks, i.e., *A little is enough* (ALIE) (Baruch et al., 2019), *Fall of Empires* (FOE) (Xie et al., 2020), *Sign Flipping* (SF) (Allen-Zhu et al., 2021), *Label Flipping* (LF) (Allen-Zhu et al., 2021), and a tailored trimmed mean attack (TMA) (cf. Algorithm 4). For all non-zero-order experiments, we conduct the attacks on the gradients $\mathbf{g}_i(\mathbf{w}_{t,\ell}^i)$. For the zero-order experiments, the attacks are conducted on the projected gradients, i.e., on $\mathbf{g}_i(\mathbf{w}_{t,\ell}^i, \mathbf{Z}_{t,\ell}) \triangleq \frac{1}{\nu}((g_i(\mathbf{w}_{t,\ell}^i, \mathbf{z}_{t,\ell}^r))_{r=1}^{\nu})^{\mathrm{T}}$, which we believe is the strongest attack scenario. Let in the following $\mathbf{g}_{t,\ell}^i$ denote the contribution of client i at global epoch t and local epoch ℓ. The attacks are summarized as follows. Let $\bar{\mathbf{g}}_{t,\ell} \triangleq \frac{1}{|\mathcal{H}|}\sum_{i\in\mathcal{H}} \mathbf{g}_{t,\ell}^i$ be the average of the honest clients gradients. For ALIE, FOE, and SF, the Byzantine clients

Algorithm 4 Transformed Trimmed-Mean Attack (TMA)

Require: β, n, $g_i(\mathbf{w}_{t,\ell}^i, \mathbf{z}_{t,\ell}^r) \forall i \in [n]$, honest clients $\mathcal{H}$
1: Compute $\bar{g}(\mathbf{w}_{t,\ell}^i, \mathbf{z}_{t,\ell}^r) = \frac{1}{n}\sum_{i=1}^n g_i(\mathbf{w}_{t,\ell}^i, \mathbf{z}_{t,\ell}^r)$
2: **for** Byzantine client $i \in [n] \setminus \mathcal{H}$ **do**
3: **if** $\bar{g}(\mathbf{w}_{t,\ell}^i, \mathbf{z}_{t,\ell}^r) > 0$ **then**
4: return $\lfloor \beta n \rfloor$ smallest value in $\{g_i(\mathbf{w}_{t,\ell}^i, \mathbf{z}_{t,\ell}^r)\}_{i\in\mathcal{H}}$
5: **else**
6: return $\lfloor \beta n \rfloor$ largest value in $\{g_i(\mathbf{w}_{t,\ell}^i, \mathbf{z}_{t,\ell}^r)\}_{i\in\mathcal{H}}$
7: **end if**
8: **end for**

$i \in \mathcal{B} \triangleq [n] \setminus \mathcal{H}$ compute their corrupted gradient as $\mathbf{g}_{t,\ell}^i = \bar{\mathbf{g}}_{t,\ell} + \omega \mathbf{a}_{t,\ell}$ for some optimized ω, where

- for ALIE, we have $\mathbf{a}_{t,\ell} = \sigma_{t,\ell}$, where $\sigma_{t,\ell}$ is the coordinate standard deviation of $\bar{\mathbf{g}}_{t,\ell}$,
- for FOE, we have $\mathbf{a}_{t,\ell} = -\bar{\mathbf{g}}_{t,\ell}$, and hence $\mathbf{g}_{t,\ell}^i = (1-\omega)\bar{\mathbf{g}}_{t,\ell}$,
- for SF, we have $\mathbf{a}_{t,\ell} = -\bar{\mathbf{g}}_{t,\ell}$ for fixed $\omega = 2$, s.t. $\mathbf{g}_{t,\ell}^i = -\bar{\mathbf{g}}_{t,\ell}$.

For ALIE and FOE, similar to (Allouah et al., 2023), we linearly optimize of potential choices of ω such that the L2 distance of the final aggregation $R_{t,\ell}$ to the honest clients' average $\bar{\mathbf{g}}_{t,\ell}^i$ is maximized. For LF, each Byzantine worker manipulates the labels of its local dataset. In particular, if for a Byzantine client $i \in \mathcal{B}$ a sample in $\mathcal{D}_i$ is labeled ℓ, they instead train on the label $\ell' = 9 - \ell$ for a 10-class classification task.

The details of the tailored trimmed mean attack can be found in Algorithm 4.

A.3 Hyperparameters

We detail in the following Tables 2 and 3 the hyperparameters used through the experiments in Section 4.

Table 2. Simulation Parameters and Hyperparameters for MNIST

MNIST	
Global Train Samples	60000
Number of Clients	40
Number of Byzantine Clients	10
Scaling Factor μ	0.001
Learning Rate η	0.01
Batch Size	64
Global Epochs T	400

Table 3. Simulation Parameters and Hyperparameters for NLP

	SST-2	SNLI	TREC	MNLI	RTE
Global Train Samples	512				
Scaling Factor μ	0.001				
Learning Rate η	10^{-6}				
Batch Size	64				
Global Epochs T	20,000	20,000	40,000	40,000	40,000

The numerical experiments were conducted on the following cluster of simulation servers.

Algorithm 5 FEDBYZO: Robust Efficient Zero-Order FL with Biased ZO Estimator

Require: Shared seed for PRNG, $\mu \geq 0$, $\eta > 0$, $\nu > 0$, R.
Initialize and broadcast global model $\mathbf{w}^{(1)}$.
for $t = 1$ to T **do**
  for each client $i \in [n]$ **in parallel do**
    Initialize local model $\mathbf{w}_{t,1}^i = \mathbf{w}^{(t)}$.
    Draw $\mathbf{z}_{t,1}^1, \cdots, \mathbf{z}_{t,1}^\nu \sim \mathcal{U}(\mathbb{S}^d)$, let $\mathbf{Z}_{t,1} \triangleq (\mathbf{z}_{t,1}^r)_{r\in[\nu]}$
    for $\ell = 1$ to K **do**
      Compute $g_i(\mathbf{w}_{t,\ell}^i, \mathbf{z}_{t,1}^r) \triangleq g(\mathbf{w}_{t,\ell}^i, \mathbf{z}_{t,1}^r, \mu, \mathcal{D}_i), r \in [\nu]$ (cf. Definition 2.1)
      Let $\mathbf{g}_i(\mathbf{w}_{t,\ell}^i, \mathbf{Z}_{t,1}) \triangleq \frac{1}{\nu}((g_i(\mathbf{w}_{t,\ell}^i, \mathbf{z}_{t,1}^r))_{r=1}^\nu)^\top$
      Update $\mathbf{w}_{t,\ell+1}^i = \mathbf{w}_{t,\ell}^i - \eta \mathbf{Z}_{t,1}\mathbf{g}_i(\mathbf{w}_{t,\ell}^i, \mathbf{Z}_{t,1})$
    end for
    Send $\{\sum_{\ell=1}^K \mathbf{g}_i(\mathbf{w}_{t,\ell}^i, \mathbf{Z}_{t,1})\}$ to federator.
  end for
  Aggregate $R_t = R(\{\sum_{\ell=1}^K \mathbf{g}_i(\mathbf{w}_{t,\ell}^i, \mathbf{Z}_{t,\ell})\}_{i=1}^n)$
  Update $\mathbf{w}^{(t+1)} = \mathbf{w}^{(t)} - \eta \sum_{\ell=1}^K \mathbf{Z}_{t,1} R_t$.
  Broadcast R_t; clients accordingly update using $\mathbf{Z}_{t,1}$
end for
Clients recover the updated global model $\mathbf{w}^{(t+1)}$ using known perturbations $\mathbf{Z}_{t,1}$.

Table 4. System specifications of our simulation cluster.

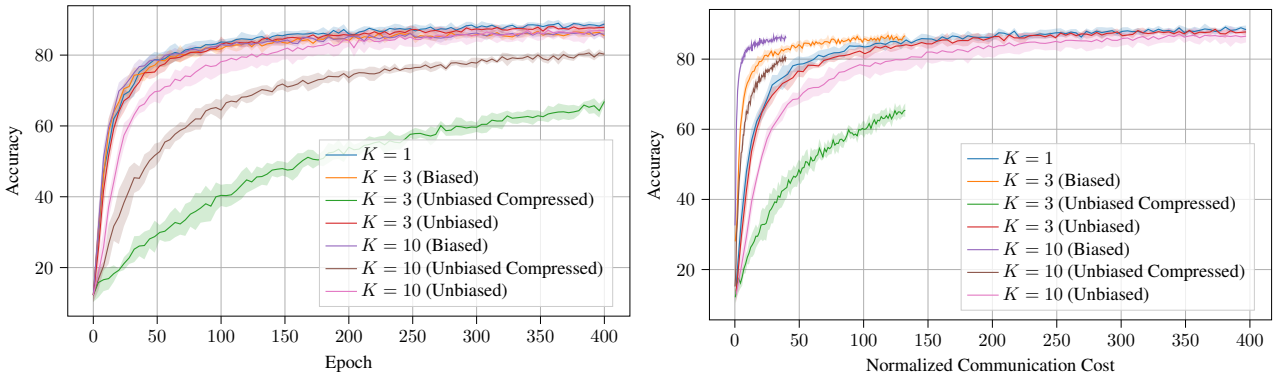
CPU(s)	RAM	GPU(s)	VRAM
2x Intel Xeon Platinum 8176 (56 cores)	256 GB	2x NVIDIA GeForce GTX 1080 Ti	11 GB
2x AMD EPYC 7282 (32 cores)	512 GB	NVIDIA GeForce RTX 4090	24 GB
2x AMD EPYC 7282 (32 cores)	640 GB	NVIDIA GeForce RTX 4090	24 GB
2x AMD EPYC 7282 (32 cores)	448 GB	NVIDIA GeForce RTX 4080	16 GB
2x AMD EPYC 7282 (32 cores)	256 GB	NVIDIA GeForce RTX 4080	16 GB
HGX-A100 (96 cores)	1 TB	4x NVIDIA A100	80 GB
DGX-A100 (252 cores)	2 TB	8x NVIDIA Tesla A100	80 GB
DGX-1-V100 (76 cores)	512 GB	8x NVIDIA Tesla V100	16 GB
DGX-1-P100 (76 cores)	512 GB	8x NVIDIA Tesla P100	16 GB
HPE-P100 (28 cores)	256 GB	4x NVIDIA Tesla P100	16 GB

A.4 Local Iterations

FEDBYZO offers different option to conduct local epochs at the clients. We term the approach for local epochs introduced in Algorithm 1 "Unbiased".

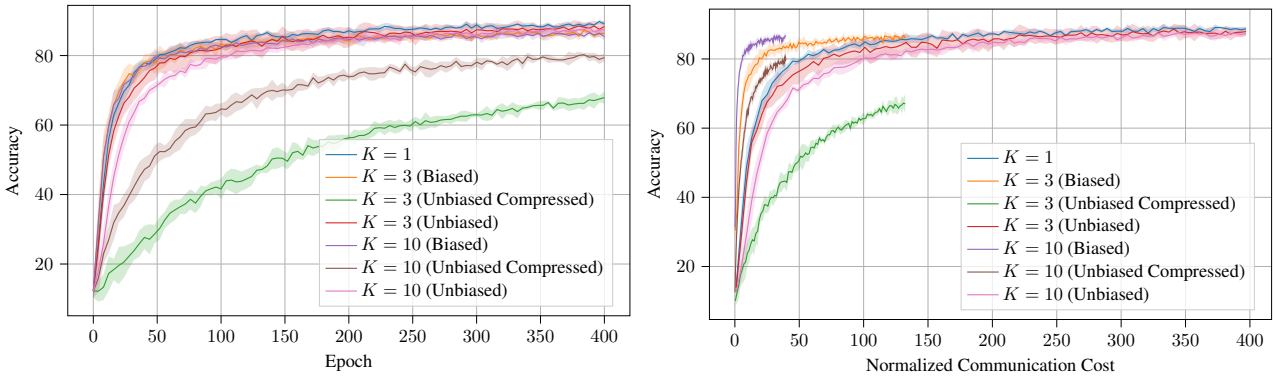
A second approach is to follow Algorithm 1, but to replace the sending of $\{\mathbf{g}_i(\mathbf{w}_{t,\ell}^i, \mathbf{Z}_{t,\ell})\}_{\ell=1}^K$ from clients to the federator, followed by $R_{t,\ell} = R(\{\mathbf{g}_i(\mathbf{w}_{t,\ell}^i, \mathbf{Z}_{t,\ell})\}_{i=1}^n), \ell \in [K]$ and $\mathbf{w}^{(t+1)} = \mathbf{w}^{(t)} - \eta \sum_{\ell=1}^{\mathcal{L}} \mathbf{Z}_{t,\ell} R_{t,\ell}$. Instead of transmitting the results $\{\mathbf{g}_i(\mathbf{w}_{t,\ell}^i, \mathbf{Z}_{t,\ell})\}_{\ell=1}^K$ for all local epochs, the clients can instead reconstruct the aggregated local gradient updates as $\mathbf{g}_i = \sum_{\ell=1}^K \mathbf{Z}_{t,\ell}\mathbf{g}_i(\mathbf{w}_{t,\ell}^i, \mathbf{Z}_{t,\ell})$ and project this gradient approximation onto random directions $\mathbf{Z}_t$ (known to the federator and all clients) according to Definition 2.1, and only transmit the result of $\mathbf{Z}_t^{\mathrm{T}}\mathbf{g}_i$ to the federator. The federator conducts the transformed aggregation on $R_{t,\ell} = R(\{\mathbf{Z}_t^{\mathrm{T}}\mathbf{g}_i\}_{i=1}^n)$ and updates the global model as $\mathbf{w}^{(t+1)} = \mathbf{w}^{(t)} - \eta \mathbf{Z}_t R_{t,\ell}$. This approach is by a factor of K more communication efficient. However, it can introduce significant additional variance, especially for small values of ν. This is because two randomly drawn vectors in high dimensions are likely almost orthogonal, and hence the subspaces resulting from $\mathbf{Z}_t$ and $\{\mathbf{Z}_{t,\ell}\}_{\ell=1}^K$ might only be weakly dependent. However, for large values of ν, this approach might be beneficial due to the drastic savings in the cost of communication. We term this approach "Unbiased Compressed".

A third approach, termed "Biased", is to make the clients reuse the directions $\mathbf{Z}_{t,\ell}$ at each iteration, i.e., $\mathbf{Z}_{t,\ell} = \mathbf{Z}_{t,m}$



(a) Accuracy over Epochs (b) Accuracy over Normalized Communication

Figure 3. Comparisons of Local Epoch Strategies for $\mu = 0.001$



(a) Accuracy over Epochs (b) Accuracy over Normalized Communication

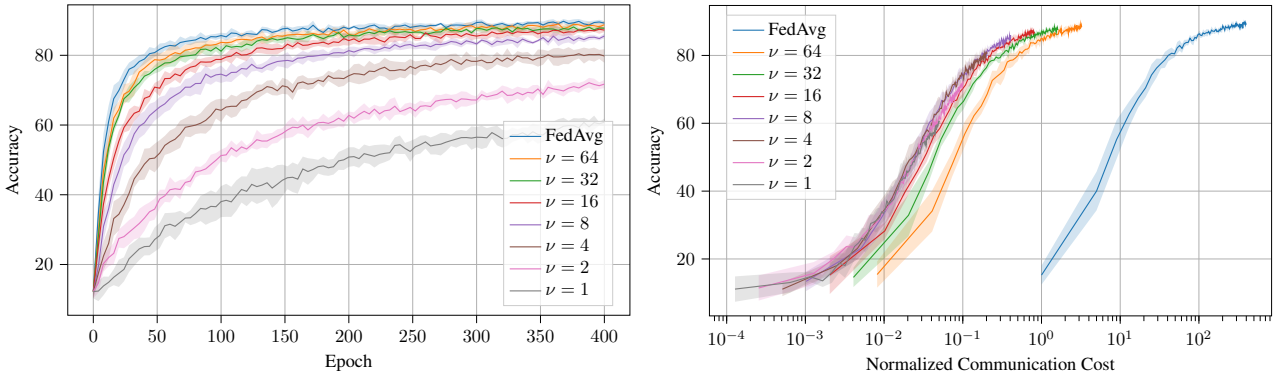
Figure 4. Comparisons of Local Epoch Strategies for $\mu = 0$

for $\ell \neq m \in [K]$. It suffices for the clients to communicate to the federator $\sum_{\ell=1}^{K} \mathbf{g}_i(\mathbf{w}_{t,\ell}^i, \mathbf{Z}_{t,\ell})$, thus reducing the communication cost by the same factor of K as for the second approach above. However, this strategy incurs bias in the local training process, since the gradient updates are not uniformly and independently chosen at each local iteration $\ell \in [K]$. We summarize this approach in Algorithm 5.

The above mentioned approaches expose an efficiency-bias-variance trade-off that we will examine in the following. In Figure 3(a), we provide a study for $\mu = 0.001$ in terms of accuracies over epochs. It can be observed that Unbiased Compressed local epochs are harmful, expecially for smalle K. The larger K, the larger the space covered during the local training process, and the smaller the loss incurred by projection the approximated overall local gradient onto an independent subspace. Looking at the accuracies over the normlized communication cost in Figure 3, we can observe that Biased local epochs with reasonably large values of Kcan indeed significantly improve the performance when normalized by communication cost. The Unbiased approach, although reducing the number of packets to be transmitted, does not significantly improve the factual communication cost. Results for $\mu = 0$ in Figure 3(b) and Figure 4 exhibit the same trade offs.

A.5 Effect of Number of Perturbations

To highlight the effect of the number of perturbations ν on the convergence of zero-order optimization in standard learning tasks, we show in Figure 5 the performance of FEDBYZO compared to FedAvg (McMahan et al., 2017) for different values of perturbations ν. While $\nu = 1$ exhibits a substantial performance gap to FedAvg, this gap decreases with increasing ν, until nearly vanishing with $\nu = 64$.

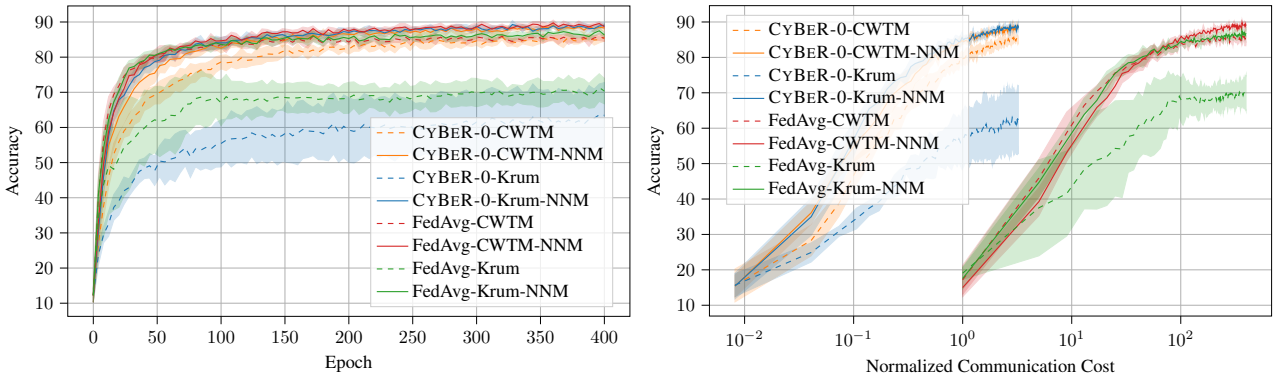


(a) Accuracy over Epochs (b) Accuracy over Normalized Communication

Figure 5. Comparison of Zero-Order Optimization for Different Values of ν Compared to the Baseline FedAvg.

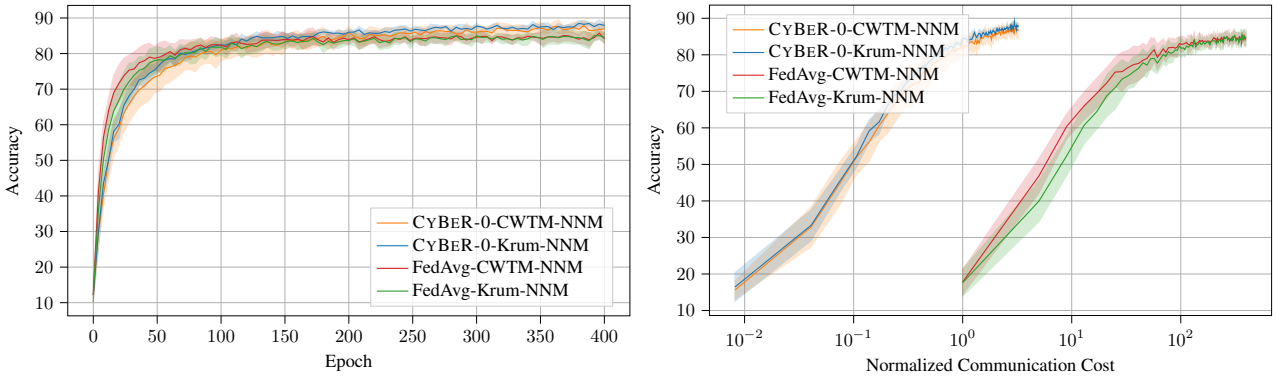
A.6 Accuracies over Epochs for all Attacks on MNIST

We present in the following, extending the case of LF (cf. Figure 1(c)), the comparison of all countermeasures for ALIE, ALIE-NNM, FOE, FOE-NNM, and SF. We present the results for accuracies over epochs, and accuracies over normalized communication cost.



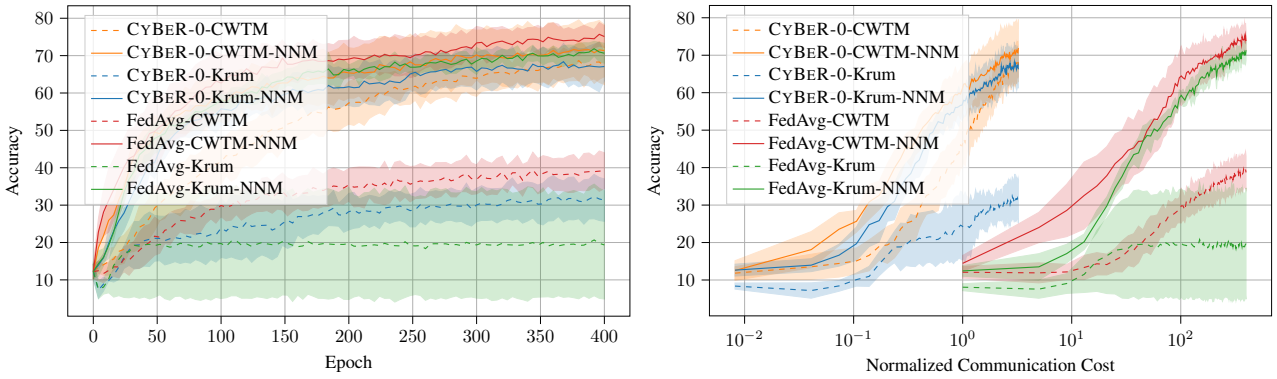
(a) Accuracy over Epochs (b) Accuracy over Normalized Communication

Figure 6. ALIE attack on logistic regression on MNIST.



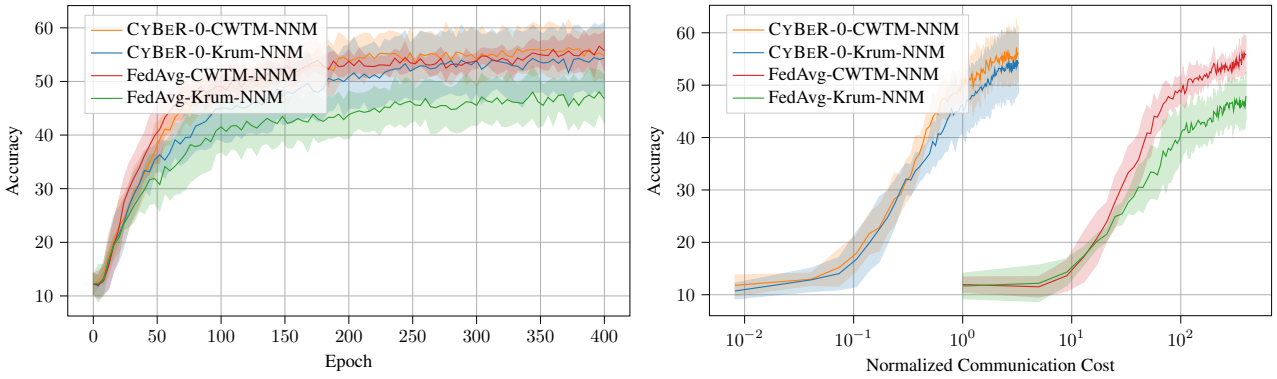
(a) Accuracy over Epochs (b) Accuracy over Normalized Communication

Figure 7. ALIE-NNM attack on logistic regression on MNIST.



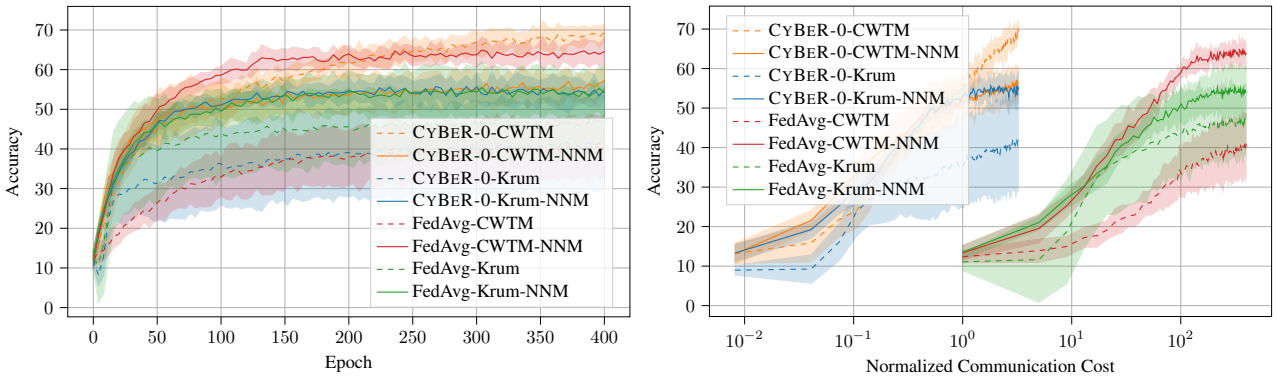
(a) Accuracy over Epochs (b) Accuracy over Normalized Communication

Figure 8. FOE attack on logistic regression on MNIST.



(a) Accuracy over Epochs (b) Accuracy over Normalized Communication

Figure 9. FOE-NNM attack on logistic regression on MNIST.

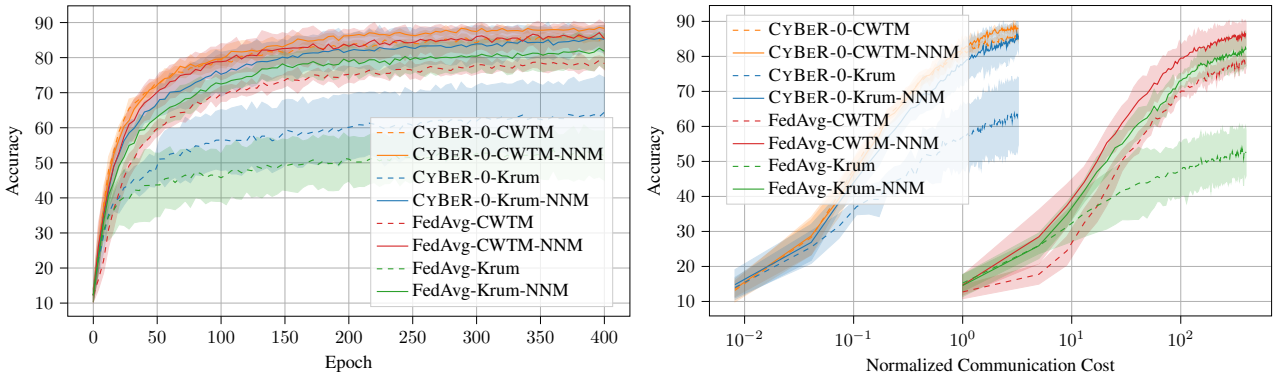


(a) Accuracy over Epochs (b) Accuracy over Normalized Communication

Figure 10. SF attack on logistic regression on MNIST.

Table 5. Mean and standard deviations of maximum accuracies for non-i.i.d. data with $\alpha = 0.1$. The baseline FedAvg without robust aggregation nor Byzantine attacks achieves 91.0 ± 0.2

Algorithm	$R(\cdot)$	NNM	ALIE	ALIE-NNM	FOE	FOE-NNM	SF	LF
FEDBYZO	CWTM	No	87.4 ± 0.6	-	$\mathbf{69.9 \pm 4.8}$	-	71.2 ± 2.0	87.6 ± 0.6
FEDBYZO	CWTM	Yes	90.3 ± 0.3	88.9 ± 1.3	74.0 ± 7.3	$\mathbf{58.9 \pm 4.9}$	59.2 ± 3.5	90.1 ± 0.6
FEDBYZO	Krum	No	65.1 ± 8.9	-	$\mathbf{34.7 \pm 5.4}$	-	44.1 ± 11.9	66.3 ± 9.9
FEDBYZO	Krum	Yes	90.6 ± 0.4	89.9 ± 0.4	70.4 ± 4.4	$\mathbf{56.9 \pm 6.0}$	58.2 ± 2.7	87.1 ± 3.5
FEDZO	CWTM	No	86.9 ± 1.1	-	$\mathbf{61.8 \pm 4.0}$	-	69.6 ± 2.1	87.3 ± 0.9
FEDZO	CWTM	Yes	90.1 ± 0.9	89.4 ± 0.7	74.3 ± 4.3	$\mathbf{59.2 \pm 3.5}$	62.7 ± 1.8	90.6 ± 0.2
FEDZO	KRUM	No	72.2 ± 8.4	-	$\mathbf{32.2 \pm 11.3}$	-	41.7 ± 4.1	67.2 ± 2.3
FEDZO	KRUM	Yes	90.4 ± 0.4	89.9 ± 0.7	67.0 ± 2.7	$\mathbf{56.4 \pm 5.5}$	65.2 ± 6.8	90.2 ± 0.6
FedAvg	CWTM	No	87.6 ± 0.7	-	$\mathbf{41.7 \pm 4.8}$	-	42.5 ± 7.9	80.8 ± 1.4
FedAvg	CWTM	Yes	90.6 ± 0.3	86.9 ± 1.2	76.8 ± 3.2	$\mathbf{58.5 \pm 2.8}$	67.3 ± 2.6	87.9 ± 3.5
FedAvg	Krum	No	75.7 ± 3.5	-	$\mathbf{23.9 \pm 13.5}$	-	50.0 ± 8.9	55.9 ± 6.9
FedAvg	Krum	Yes	88.6 ± 0.6	86.9 ± 1.1	73.0 ± 2.0	$\mathbf{50.0 \pm 4.1}$	57.7 ± 5.8	83.9 ± 4.6



(a) Accuracy over Epochs

(b) Accuracy over Normalized Communication

Figure 11. LF attack on logistic regression on MNIST.

A.7 Comparison to Non-Transformed Zero-Order FL

We provide extensive experiments comparing our method to a natural extension of the non-Byzantine resilient FEDZO (Fang et al., 2022) to robust aggregation for various aggregation functions. In particular, the robust aggregation is performed on the reconstructed gradients, i.e., in $\mathbb{R}^d$. While uplink communication efficiency can be achieved similarly to FEDBYZO by leveraging a shared seed concept (similar to (Li et al., 2024)), it is impossible to achieve downlink communication efficiency after robust aggregation on the full gradients. The reasoning follows the same arguments as described in Appendix B. That is, when performing robust aggregation on the reconstructed approximate gradients, it is not guaranteed that the aggregated vector lies in the subspace spanned by the perturbations. Hence, the aggregate cannot be compressed by projection onto this subspace without information loss. Consequently, communication efficiency can only be achieved on the uplink. FEDBYZO achieves both uplink and downlink communication efficiency while improving on the best-performing aggregation rule's worst-case performance by more than 10%. This is shown in Table 5, where we compare the performance of FEDBYZO against robust variants of FedAvg (McMahan et al., 2017) and FEDZO (Li et al., 2024). FEDBYZO-CWTM outperforms all baselines by at least 8% in terms of worst-case accuracy.

A.8 Accuracies over Epochs for all Attacks on Fine-Tuning Tasks

We provide in Table 6 extensive results on an i.i.d. data distribution for fine-tuning RoBERTa-large, analog to the non-i.i.d. results in Table 1. It can be found that our algorithm exhibits stable performance for both i.i.d. and non-i.i.d. distributions, and is not significantly affected by heterogeneity. We further show in Table 8 and Table 7 results for fine-tuning OPT-125m

on SST-2 and TREC, for comparability to (Li et al., 2024).

Table 6. Mean and Standard Deviation of Maximum Accuracies Across Seeds for RoBERTa-large and i.i.d. data

Dataset	ALIE	FOE	SF	TMA	No Attack	Worst Case
SST-2	93.0 ± 0.4	91.6 ± 0.2	91.9 ± 0.1	92.1 ± 0.6	92.9 ± 0.6	91.6
TREC	95.5 ± 0.3	88.5 ± 0.9	90.5 ± 0.8	91.4 ± 1.9	95.6 ± 0.4	88.5
SNLI	83.5 ± 0.5	77.0 ± 0.8	78.7 ± 1.0	79.6 ± 0.9	84.9 ± 0.2	77.0
RTE	79.4 ± 1.8	73.8 ± 0.7	73.4 ± 0.2	76.4 ± 0.6	79.9 ± 1.1	73.4
MNLI	76.2 ± 0.6	68.1 ± 2.8	68.9 ± 2.4	70.6 ± 1.7	76.4 ± 0.8	68.1

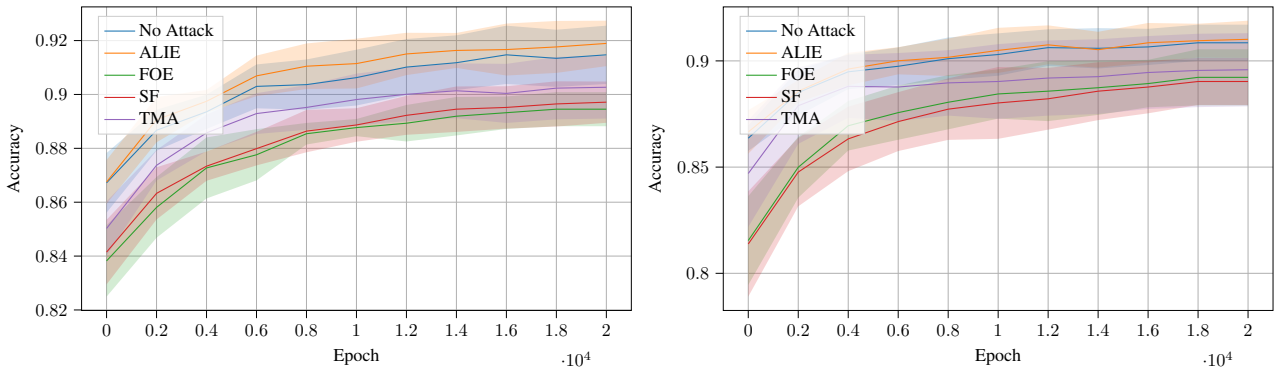
Table 7. Mean and Standard Deviation of Maximum Accuracies Across Seeds for OPT-125m and non-i.i.d. data

Dataset	ALIE	FOE	SF	TMA	No Attack	Worst Case
SST-2	83.8 ± 1.3	82.3 ± 1.0	73.9 ± 4.9	82.6 ± 0.9	83.8 ± 0.5	73.9
TREC	88.3 ± 1.8	75.3 ± 4.3	74.1 ± 5.8	87.5 ± 1.3	92.2 ± 1.3	74.1

Table 8. Mean and Standard Deviation of Maximum Accuracies Across Seeds for OPT-125m and i.i.d. data

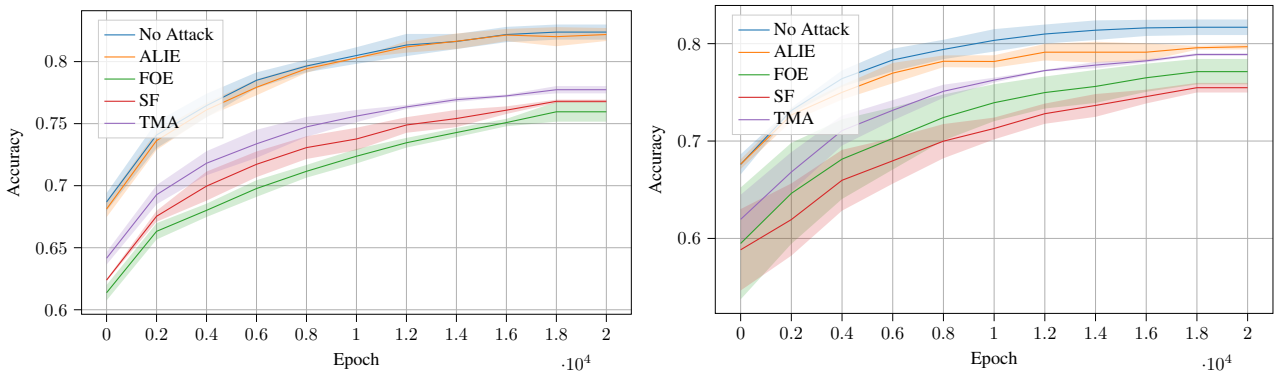
Dataset	ALIE	FOE	SF	TMA	No Attack	Worst Case
SST-2	83.8 ± 0.6	81.0 ± 0.2	81.2 ± 0.2	82.5 ± 0.1	85.5 ± 0.3	81.0
TREC	93.5 ± 0.6	84.9 ± 1.1	88.3 ± 0.8	90.3 ± 1.1	92.4 ± 0.2	84.9

While FEDBYZO consistently achieves satisfying performance even under i.i.d. and non-i.i.d. data distributions and various Byzantine attacks, RoBERTa-large reaches substantially better performance (under Byzantine and non-Byzantine scenarios). Hence, we focus our attention on RoBERTa-large. Further, we provide in the following plots for the accuracies over the epochs for all attacks, datasets, and both i.i.d. and non-i.i.d. data distributions. FEDBYZO exhibits stable performance in all settings.



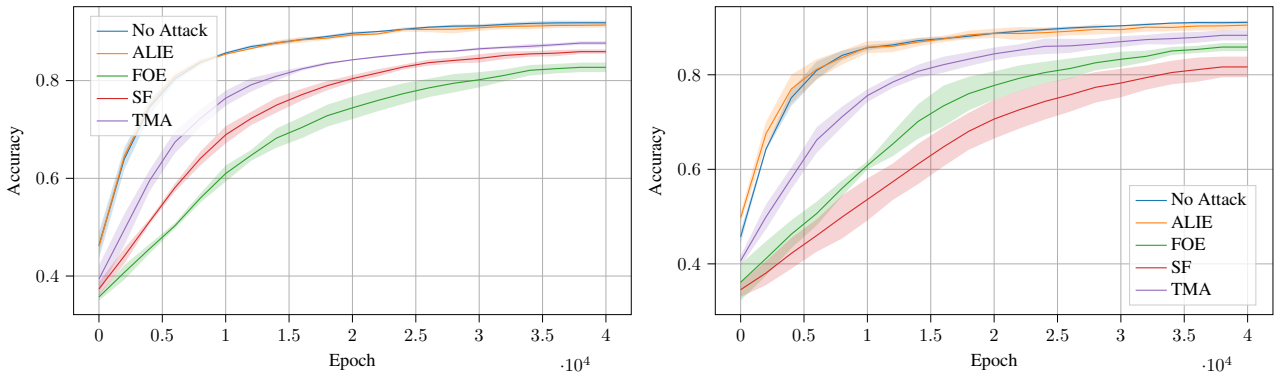
(a) i.i.d. (b) non-i.i.d.

Figure 12. Accuracy comparison of different attacks on fine-tuning RoBERTa-large on SST-2.



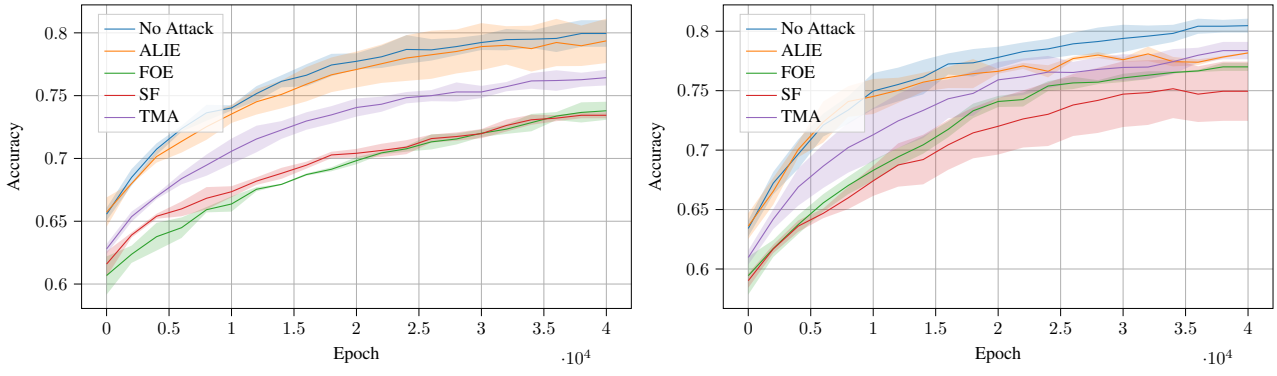
(a) i.i.d. (b) non-i.i.d.

Figure 13. Accuracy comparison of different attacks on fine-tuning RoBERTa-large on SNLI.



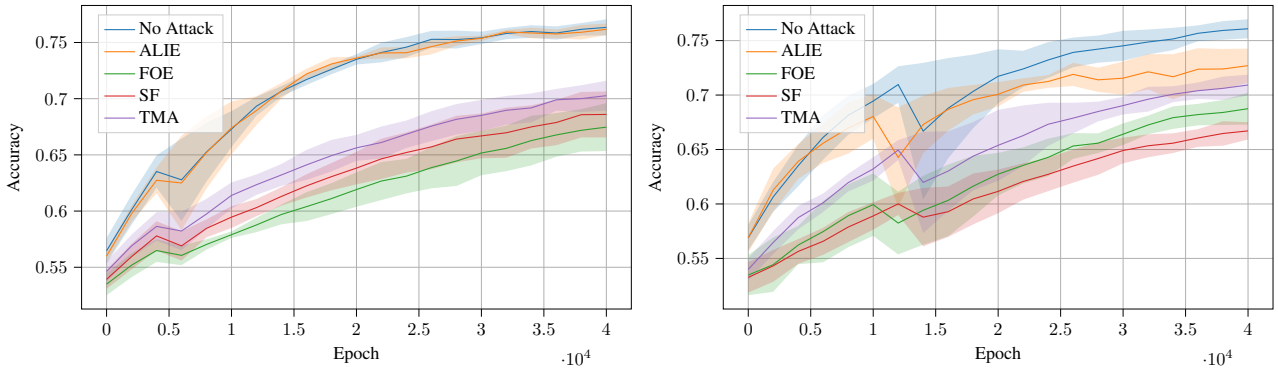
(a) i.i.d. (b) non-i.i.d.

Figure 14. Accuracy comparison of different attacks on fine-tuning RoBERTa-large on TREC.



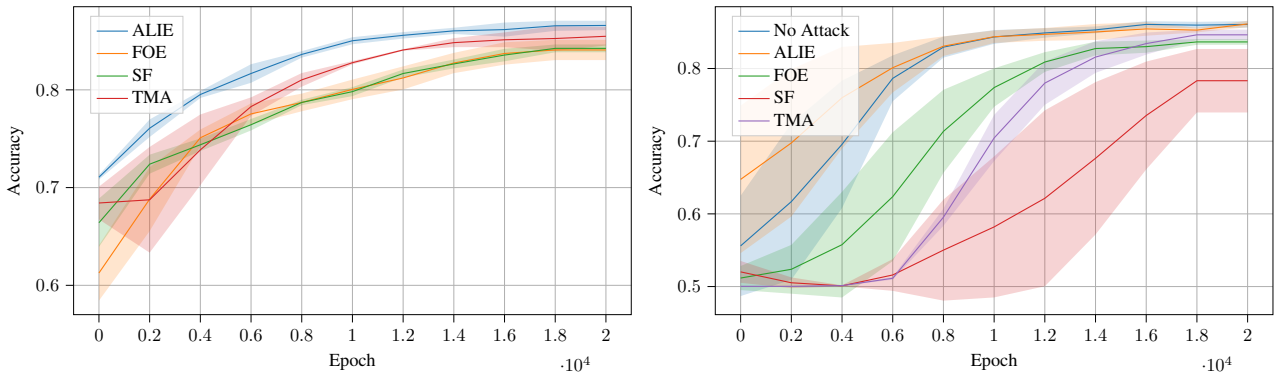
(a) i.i.d. (b) non-i.i.d.

Figure 15. Accuracy comparison of different attacks on fine-tuning RoBERTa-large on RTE.



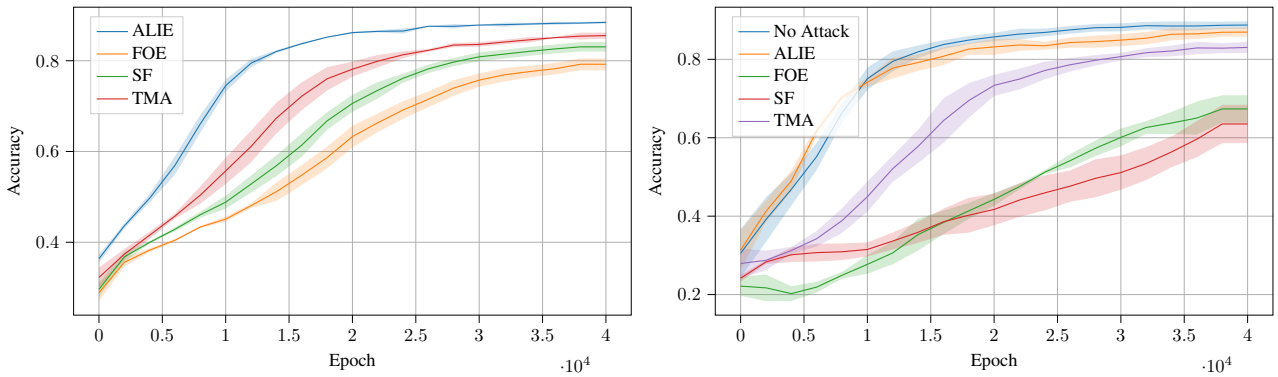
(a) i.i.d. (b) non-i.i.d.

Figure 16. Accuracy comparison of different attacks on fine-tuning RoBERTa-large on MNLI.



(a) i.i.d. (b) non-i.i.d.

Figure 17. Accuracy comparison of different attacks on fine-tuning opt-125m on SST-2.



(a) i.i.d. (b) non-i.i.d.

Figure 18. Accuracy comparison of different attacks on fine-tuning opt-125m on TREC.

A.9 Hyperparameter Study

We provide in the following tables Tables 9 to 12 a sensitivity analysis of FEDBYZO with respect to the number of global epochs T, the number of local epochs K, the number of perturbations ν, the number of clients n, and the number of

Byzantine clients $\mathcal{B}$.

We run our experiments on SST-2 and RoBERTa-large, attacked by FOE. We use non-i.i.d. data with $\alpha = 0.1$. We first show in Table 9 the robustness of FEDBYZO under varying numbers of total and Byzantine clients with $\frac{b}{n} = 0.25$ and observe that the robustness increases with n. We fix $n = 8$, and $b = 2$ as the most challenging setting for the following experiments. We show in Table 10 the stability of FEDBYZO under varying number ν of random perturbations. For comparability, we fix the ratio $\nu T = 20000$ and observe very similar accuracies for $\nu \in \{1, 2, 4, 8\}$. Hence, the number of projections trades almost inversely with the number of global epochs T. A similar behavior can be observed in Table 11 for a varying number of local epochs K, fixing the ratio KT. However, for large K and small T, we can see the negative impacts of local iterations. Lastly, we evaluate FEDBYZO under varying ratios of Byzantine clients, thereby fixing $n = 16$ for better flexibility, and using $\nu = 1$. The results are depicted in Table 12, While the performance reduction from $b = 2$ to $b = 4$ is negligible, we can observe a notable difference for $b = 6$, i.e., when the number of Byzantine clients is close to $n/2$.

Table 9. Accuracy over n, b for $b/n = 0.25$

Clients n	Byzantine b	Acc $\pm$ Std
8	2	0.86 ± 0.02
12	3	0.87 ± 0.06
16	4	0.89 ± 0.02
32	8	0.87 ± 0.03

Table 11. Accuracy over T and K.

Global epochs T	Local epochs K	Acc $\pm$ Std
2000	10	0.81 ± 0.03
4000	5	0.86 ± 0.03
20000	1	0.86 ± 0.02

Table 10. Accuracy over T and ν.

Global epochs T	ν	Acc $\pm$ Std
2500	8	0.87 ± 0.02
5000	4	0.88 ± 0.02
10000	2	0.86 ± 0.02
20000	1	0.86 ± 0.02

Table 12. Accuracy over b/n for $n = 16$.

Clients n	Byzantine b	Acc $\pm$ Std
16	2	0.90 ± 0.01
16	4	0.89 ± 0.02
16	6	0.78 ± 0.05

B Innovations in FEDBYZO

We highlight some challenges addressed by FEDBYZO and key technical innovations.

• **Transformed robust aggregation via embeddings:** Robust aggregation of the full gradients is well understood for achieving Byzantine resilience in FL. However, in FEDBYZO the federator only has access to the model updates along the random perturbations rather than full gradients. A naive approach would be for the federator to first reconstruct approximate gradients from the ZO updates and then apply robust aggregation. However, in general, robust aggregation rules are non-linear, and the aggregate vector may lie outside the subspace spanned by the input vectors. For example, consider vectors $\mathbf{v}_1 = [2, 2, 0]^\top, \mathbf{v}_2 = [0, -1, -1]^\top$, and $\mathbf{v}_3 = [4, 0, -4]^\top$ that lie in the vector space $\mathcal{V}$ spanned by $\{[1, 1, 0]^\top, [0, 1, 1]^\top\}$. Then, $\text{CWTM}_{1/3}(\mathbf{v}_1, \mathbf{v}_2, \mathbf{v}_3) = [2, 0, -1]^\top \notin \mathcal{V}$. In our context, this implies that the global model update $\mathbf{w}^{(t+1)} - \mathbf{w}^{(t)}$ does not necessarily lie in the subspace spanned by the perturbation vectors $\{\mathbf{z}_{t,\ell}^r : r \in [\nu], \ell \in [K]\}$. Thus, communicating the global model update requires either additional communication cost or projecting the global model update back onto the subspace, incurring additional variance and computation.
To address this issue, FEDBYZO introduces *transformed robust aggregation*, *i.e.*, robust aggregation of the clients' model updates when viewed as vectors embedded in the perturbation space $\mathbb{R}^\nu$. By directly aggregating in this lower-dimensional perturbation space, FEDBYZO preserves communication savings (as the aggregated updates continue to belong to the perturbation space). A key challenge is to argue that performing the aggregation in the perturbation space and projecting the result back to the gradient space preserves the robustness guarantees from Definition 2.2. To prove this, we rely on Johnson–Lindenstrauss–style embeddings (Johnson & Lindenstrauss, 1984) to maintain the necessary geometric properties of robust aggregation. Thus, the aggregation remains both efficient and Byzantine-resilient, while limiting the attackers' ability to manipulate the global update outside the chosen subspace.
• **Efficient downlink and uplink communication:** We leverage the structure of ZO updates and transformed robust aggregation to significantly reduce the communication cost on the uplink and the downlink. On the uplink, this is a

consequence of having the clients perform local epochs and transmit only the resulting scalar values for each perturbation direction. On the downlink, as the transformed robust aggregation outcome is guaranteed to lie in the span of the perturbation vectors, it is sufficient to only specify the projections along the perturbation directions. This reduces the communication burden to a handful of scalars, which is a major reduction compared to typical FL settings where the gradient dimension can range from 10^6 to 10^{12}.

• **Shared seed mechanism for synchronizing perturbations:** To ensure correct global model updates, the random perturbation directions must be synchronized between the federator and the clients. A naive, yet inefficient strategy would be to transmit the newly generated perturbation vectors in every round, but this negates any communication gains since their dimension matches that of the model. In FEDBYZO, we address this challenge by adopting a lightweight shared seed protocol, inspired by (Salimans et al., 2017), which enables both the clients and the federator to locally generate identical pseudorandom perturbations. Note that the (one-time) cost of sharing a common seed determined by the federator with all clients is negligible given that standard PRNGs, *e.g.*, in Tensorflow, have a cycle length in the order of 2^{128} (Salmon et al., 2011). Note that the seed is non-collaboratively established by the federator. Thus, Byzantine clients cannot execute coordinated attacks in this phase.

• **Multiple local epochs per client:** We show that FEDBYZO works well with each client performing multiple local epochs. This results in a reduction in the number of global epochs (and hence less frequent communication). Further, FEDBYZO also offers a design choice between the unbiased ZO estimator and the biased ZO estimator. The former is more amenable to theoretical analysis as $\mathbf{Z}_{t,\ell}$'s are independent across the local epochs. However, it has a communication cost of $K\nu$ per global epoch (both on each uplink as well as the downlink). The biased ZO estimator further reduces the communication cost by a factor of K by deliberately using the same perturbation vectors for each local iteration. This, however, introduces additional bias into the training process. We explore this tradeoff under varying numbers of local epochs in Appendix A.4.

• **Memory efficiency:** While it is well established that ZO methods reduce the computation cost when ν is small, the reductions in the memory requirements are especially impressive even when ν is large. (Malladi et al., 2023) showed that ZO methods perform inference utilizing up to a factor of 12 lower memory compared to backpropagation. FEDBYZO inherits this property and can operate on resource-constrained edge devices in fine-tuning tasks where classical approaches are intractable. We note that for large ν, the computation costs for $\mu > 0$ can outweigh that of backpropagation when paired with gradient projection, i.e., $\mu = 0$. The details can be found in Appendix A.1.

C Proofs

C.1 Properties of ZO estimate

For a given model $\mathbf{w}$, let $\nabla F_i^\mu \triangleq \mathbb{E}_{\mathbf{z}\sim\mathcal{U}(\mathbb{S}^d)}[\nabla F_i(\mathbf{w}+\mu\mathbf{z})]$ be a smoothened version of the gradient ∇F_i. Then, for the two-point zero-order estimate from Definition 2.1, we have the following well-known result (Flaxman et al., 2005).

Proposition C.1. *The ZO estimate satisfies on expectation*

$$\mathbb{E}\left[\mathbf{Z}_{t,\ell}\mathbf{g}_i(\mathbf{w}_{t,\ell}^i, \mathbf{Z}_{t,\ell})\right] = \nabla F_i^\mu(\mathbf{w}_{t,\ell}^i).$$

Proposition C.2 (Lemma 2, (Tang et al., 2020))**.** *It holds*

$$\left\|\nabla F_i^\mu(\mathbf{w}_{t,\ell}^i) - \nabla F_i(\mathbf{w}_{t,\ell}^i)\right\| \le L\mu.$$

C.2 Proof of Theorem 5.7

We state the extended form of Theorem 5.7 first. Let $\xi_1, \xi_2, \xi_3, \xi_4$ and ξ_5 be as defined in (7)-(12) These quantities scale as follows: $\xi_1 = \Theta\big(\eta^2K^3\frac{d}{\nu} + \frac{d}{\nu}L^2\eta^2K(K+\frac{d}{|\mathcal{H}|\nu})\big)$, $\xi_2 = \Theta\big(\eta^2K^3(L + (\frac{d}{\nu} + \frac{d^2}{\nu^2|\mathcal{H}|}L^2K^2\eta^2)\left(\zeta^2+\sigma^2+L^2\mu d\right)\big)$, $\xi_3 = \Theta\big(\frac{d}{\nu}(K^2\epsilon'\kappa + \frac{K}{|\mathcal{H}|})(1+L^2\eta^2K\left(K+\frac{d}{|\mathcal{H}|\nu}\right)(1+\frac{\nu}{d}))\big)$, $\xi_4 = \Theta\big(K^2L^2\eta^2(K^2D^2 + \frac{1}{|\mathcal{H}|}\frac{d^2}{\nu^2}L^2) + K\epsilon'\kappa(D^2+\zeta^2+\frac{d}{\nu}L^2)\big)$ and $\xi_5 = \Theta\Big((\zeta^2+\sigma^2+L^2\mu^2d)\big((K^2\epsilon'\kappa)(\frac{d}{\nu} + \frac{d^2}{\nu^2|\mathcal{H}|}L^2K\eta^2)\big) + K^2L(\epsilon'\kappa+\mu) + (K^2\epsilon'\kappa)(\frac{d}{\nu}L^2\eta^2K^2L\mu)\Big)$.

Theorem C.3 (General non-convex landscapes)**.** *Let $0 < \Delta < 1$ and suppose that Assumptions 5.1 to 5.4 hold. Consider* FEDBYZO *with a $(|\mathcal{H}|, \kappa)$-robust aggregation rule. If* (a) $\eta^2 \le \min\{\frac{1}{72K^2L}, \frac{|\mathcal{H}|\nu}{96KdL^2}, 6\frac{D^2}{K^2} + 6\frac{\zeta^2}{K^2} + \frac{32d}{\nu K^2}L^2\}$, (b) $\nu \ge 64\log(\frac{2(|\mathcal{H}|-1)TK}{\Delta})$, *and* (c) $\xi_3 + \xi_4\xi_1 \le \frac{1}{2}$ *are satisfied, then the following convergence guarantee holds with probability $1-\Delta$:*

$$\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}\left[\left\|\nabla F_{\mathcal{H}}(\mathbf{w}_t)\right\|^2\right] \le \frac{4(F_{\mathcal{H}}(\mathbf{w}_1) - F_{\mathcal{H}}^\star)}{\eta KT} + \eta/(2K)\left(\xi_4\xi_2 + \xi_5\right).$$

Note that Theorem C.3 requires $\nu = \Theta(d)$ in order for condition *(c)* to be satisfied.

Sketch of Proof. We provide a brief proof outline in the following for the case when $\mu > 0$. The proof for $\mu = 0$ follows similar steps with some modifications, since $\left\|\left(\nabla F_i(\mathbf{w}_{t,\ell}^i) - \nabla F_i^\mu(\mathbf{w}_{t,\ell}^i)\right)\right\|^2 = 0$ by definition and the gradient estimate is bounded differently. We define the following quantities $\hat{\mathbf{w}}_{t,\ell} \triangleq \frac{1}{|\mathcal{H}|}\sum_{i\in\mathcal{H}} \mathbf{w}_{t,\ell}^i$ and $\bar{\mathbf{g}}_{\mathcal{H}}(\mathbf{w}_{t,\ell}^i, \mathbf{Z}_{t,\ell}) \triangleq \frac{1}{|\mathcal{H}|}\sum_{i\in\mathcal{H}} \mathbf{g}_i(\mathbf{w}_{t,\ell}^i, \mathbf{Z}_{t,\ell})$, and focus on the conceptual strategy and omit all factors. We first decompose the difference of two consecutive models into $\|\nabla F_{\mathcal{H}}(\mathbf{w}_t)\|^2$, $\left\|\sum_{\ell=1}^K R_{t,\ell}\mathbf{Z}_{t,\ell}\right\|^2$ and $\left\|K\nabla F_{\mathcal{H}}(\mathbf{w}_t) - \sum_{\ell=1}^K \mathbf{Z}_{t,\ell}R_{t,\ell}\right\|^2$. The former term is the quantity of interest. The second term can be made negative by an appropriate choice of the learning rate. On expectation and using Assumptions 5.1 and 5.4, the latter can be bounded by i) $\mathbb{E}\left[\|\mathbf{w}_t - \hat{\mathbf{w}}_{t,\ell}\|^2\right]$, ii) $\mathbb{E}\left[\sum_{\ell=1}^K \left\|\mathbf{Z}_{t,\ell}\bar{\mathbf{g}}_{\mathcal{H}}(\mathbf{w}_{t,\ell}^i, \mathbf{Z}_{t,\ell}) - \mathbf{Z}_{t,\ell}R_{t,\ell}\right\|^2\right]$, iii) $\mathbb{E}\left[\left\|\mathbf{w}_{t,\ell}^i - \hat{\mathbf{w}}_{t,\ell}\right\|^2\right]$ and iv) $\mathbb{E}\left[\left\|\sum_{\ell=1}^K \left(\sum_{i\in\mathcal{H}} \frac{\nabla F_i(\mathbf{w}_{t,\ell}^i)}{|\mathcal{H}|} - \mathbf{Z}_{t,\ell}\bar{\mathbf{g}}_{\mathcal{H}}(\mathbf{w}_{t,\ell}^i, \mathbf{Z}_{t,\ell})\right)\right\|^2\right]$. Lemma C.4 (in turn requiring similar derivations as for the proof of Lemma C.5) relates the term ii) to $\mathbb{E}\left[\|\nabla F_{\mathcal{H}}(\mathbf{w}_t)\|^2\right]$ and a term like iii). iv) can be bounded from above by $\left\|\left(\nabla F_i(\mathbf{w}_{t,\ell}^i) - \nabla F_i^\mu(\mathbf{w}_{t,\ell}^i)\right)\right\|^2$ and $\mathbb{E}\left[\left\|\sum_{\ell=1}^K \left(\nabla F_i^\mu(\mathbf{w}_{t,\ell}^i) - \mathbf{Z}_{t,\ell}\mathbf{g}_i(\mathbf{w}_{t,\ell}^i, \mathbf{Z}_{t,\ell})\right)\right\|^2\right]$. While the first is bounded by Proposition C.2, the latter is bounded by Lemma C.5 (in turn requiring Assumption 5.2 and Lemmas C.4 and C.9 in terms of $\mathbb{E}\left[\|\nabla F_{\mathcal{H}}(\mathbf{w}_t)\|^2\right]$ and terms like iii). ii) is bounded by a twice application of a particular Johnson-Lindenstrauss-type Lemma (Lemma C.8) and Lemma C.6 (which relies on (Wang et al., 2024, Lemma B.1) and Lemma C.5. All terms of the kind iii) are bounded using Lemma C.7 (that relies on Lemma C.6) in terms of $\mathbb{E}\left[\|\nabla F_{\mathcal{H}}(\mathbf{w}_t)\|^2\right]$. By appropriate choices of the learning rates so that Lemma C.4 and Lemma C.7 hold and the term that multiplies the quantity $\mathbb{E}\left[\|\nabla F_{\mathcal{H}}(\mathbf{w}_t)\|^2\right]$ of interest is negative and can hence be rearranged and bounded, the proof is completed by telescoping over all global iterations. □

C.2.1 Proof of Theorem 5.7 for $\mu > 0$

We start with proving convergence for $\mu > 0$, and apply similar steps to prove the result for $\mu = 0$ in Appendix C.2.2.

Proof. To prove the convergence of our algorithm for general non-convex functions with local iterations, byzantine resilience and heterogeneity, we rely on the following intermediate lemmas that we state in the following. We assume throughout that Assumptions 5.1 to 5.4 hold, and the robust aggregator satisfies Definition 2.2.

Lemma C.4. *Let*

$$\xi_6 \triangleq 5\cdot 32\eta^2 K\left(K + \frac{d}{|\mathcal{H}|\nu}\right),$$
$$\xi_7 \triangleq 5K\eta^2\frac{d}{|\mathcal{H}|\nu}\left(32\zeta^2 + 8\sigma^2 + L^2\mu^2 d\right) + 5\eta^2 24K^2 L\mu, \text{ and}$$
$$\xi_8 \triangleq 5\cdot 32\eta^2\left(KD^2 + \frac{1}{|\mathcal{H}|}\frac{d}{\nu}L^2\right).$$

Let $\hat{\mathbf{w}}_{t,\ell} \triangleq \frac{1}{|\mathcal{H}|}\sum_{i\in\mathcal{H}} \mathbf{w}_{t,\ell}^i$. For a learning rate satisfying $24K\eta^2L^2 \leq \frac{1}{3K}$ and $2\eta^2\frac{1}{|\mathcal{H}|}\frac{4d}{\nu}4L^2 \leq \frac{1}{3K}$, we have the following upper bound on the averaged local model divergence:

$$\mathbb{E}\left[\|\hat{\mathbf{w}}_{t,\ell} - \mathbf{w}_t\|^2\right] \leq \xi_6\mathbb{E}\left[\|\nabla F_{\mathcal{H}}(\mathbf{w}_t)\|^2\right] + \xi_7 + \xi_8\sum_{\ell'=1}^{\ell-1}\frac{1}{|\mathcal{H}|}\sum_{i\in\mathcal{H}}\mathbb{E}\left[\left\|\mathbf{w}_{t,\ell'}^i - \hat{\mathbf{w}}_{t,\ell'}\right\|^2\right].$$

Lemma C.5. *Let*

$$\xi_9 \triangleq \frac{16d}{\nu}L^2,$$

$$\xi_{10} \triangleq \frac{80 \cdot 32d}{\nu} L^2 \eta^2 \left(KD^2 + \frac{1}{|\mathcal{H}|} \frac{d}{\nu} L^2 \right),$$
$$\xi_{11} \triangleq \frac{16d}{\nu} + \frac{80 \cdot 32d}{\nu} L^2 \eta^2 K \left(K + \frac{d}{|\mathcal{H}|\nu} \right), \text{ and}$$
$$\xi_{12} \triangleq \left(\frac{d}{2\nu} + \frac{80d^2}{\nu^2 |\mathcal{H}|} L^2 K \eta^2 \right) \left(32\zeta^2 + 8\sigma^2 + L^2 \mu^2 d \right) + \frac{80 \cdot 24d}{\nu} L^2 \eta^2 K^2 L \mu.$$

We have the following bound on the gradient estimate variance based on multiple independent perturbations

$$\mathbb{E}\left[\left\| \mathbf{Z}_{t,\ell} \mathbf{g}_i(\mathbf{w}_{t,\ell}^i, \mathbf{Z}_{t,\ell}) - \nabla F_i^\mu(\mathbf{w}_{t,\ell}^i) \right\|^2 \right]$$
$$\leq \xi_9 \mathbb{E}\left[\left\| \mathbf{w}_{t,\ell}^i - \hat{\mathbf{w}}_{t,\ell} \right\|^2 \right] + \xi_{12} + \xi_{11} \mathbb{E}\left[\left\| \nabla F_{\mathcal{H}}(\mathbf{w}_t) \right\|^2 \right] + \xi_{10} \sum_{\ell'=1}^{\ell-1} \frac{1}{|\mathcal{H}|} \sum_{i \in \mathcal{H}} \mathbb{E}\left[\left\| \mathbf{w}_{t,\ell'}^i - \hat{\mathbf{w}}_{t,\ell'} \right\|^2 \right].$$

Lemma C.6. *Let*

$$\xi_{13} \triangleq 6D^2 + 6\zeta^2 + 2\xi_9 + 2\xi_{10} K^2 = 6D^2 + 6\zeta^2 + 2\frac{4d}{\nu} 4L^2 + \frac{8d}{\nu} 4L^2 5 \cdot 32\eta^2 \left(KD^2 + \frac{1}{|\mathcal{H}|} \frac{d}{\nu} L^2 \right) K^2,$$
$$\xi_{14} \triangleq 2\xi_{12} = 2 \left(\frac{d}{2\nu} + \frac{80d^2}{\nu^2 |\mathcal{H}|} L^2 K \eta^2 \right) \left(32\zeta^2 + 8\sigma^2 + L^2 \mu^2 d \right) + \frac{8d}{\nu} 4L^2 5 \eta^2 24 K^2 L \mu,$$
$$\xi_{15} \triangleq 6L, \text{ and}$$
$$\xi_{16} \triangleq 2\xi_{11} = \frac{8d}{\nu} 4 + \frac{8d}{\nu} 4L^2 5 \cdot 32\eta^2 K \left(K + \frac{d}{|\mathcal{H}|\nu} \right).$$

Then the local gradient divergence is bounded from above by

$$\sum_{m=1}^{\ell} \frac{1}{|\mathcal{H}|} \sum_{i \in \mathcal{H}} \mathbb{E}\left[\left\| \mathbf{Z}_{t,m} \mathbf{g}_i(\mathbf{w}_{t,\ell}^i, \mathbf{Z}_{t,\ell}) - \frac{1}{|\mathcal{H}|} \sum_{j \in \mathcal{H}} \mathbf{Z}_{t,m} \mathbf{g}_j(\mathbf{w}_{t,m}^j, \mathbf{Z}_{t,m}) \right\|^2 \right]$$
$$\leq \sum_{m=1}^{\ell} \frac{1}{|\mathcal{H}|} \sum_{i \in \mathcal{H}} \xi_{13} \mathbb{E}\left[\left\| \hat{\mathbf{w}}_{t,m} - \mathbf{w}_{t,m}^i \right\|^2 \right] + \sum_{m=1}^{\ell} (\xi_{14} + \xi_{15}) + \sum_{m=1}^{\ell} \xi_{16} \mathbb{E}\left[\left\| \nabla F_{\mathcal{H}}(\mathbf{w}_t) \right\|^2 \right].$$

Lemma C.7. *Let*

$$\xi_1 \triangleq 4\eta^2 K^3 \frac{16d}{\nu} + \frac{4d}{\nu} 4L^2 5 \cdot 32\eta^2 K \left(K + \frac{d}{|\mathcal{H}|\nu} \right) \text{ and}$$
$$\xi_2 \triangleq 4\eta^2 K^3 \left(\left(\frac{d}{2\nu} + \frac{80d^2}{\nu^2 |\mathcal{H}|} L^2 K \eta^2 \right) \left(32\zeta^2 + 8\sigma^2 + L^2 \mu^2 d \right) + \frac{80 \cdot 24d}{\nu} L^2 \eta^2 K^2 L \mu \right) + 12\eta^2 K^3 L.$$

For a learning rate that satisfies $\eta \leq \sqrt{6\frac{D^2}{K^2} + 6\frac{\zeta^2}{K^2} + \frac{24d}{\nu K^2} L^2}$, we have the following upper bound on the local model divergence

$$\sum_{\ell=1}^{K} \frac{1}{|\mathcal{H}|} \sum_{i \in \mathcal{H}} \mathbb{E}\left[\left\| \mathbf{w}_{t,\ell}^i - \hat{\mathbf{w}}_{t,\ell} \right\|^2 \right] \leq \xi_1 \mathbb{E}\left[\left\| \nabla F_{\mathcal{H}}(\mathbf{w}_t) \right\|^2 \right] + \xi_2.$$

Now we are ready to prove Theorem 5.7. With $R_{t,\ell} \triangleq R\left(\{\mathbf{g}_i(\mathbf{w}_{t,\ell}^i, \mathbf{Z}_{t,\ell})\}_{i=1}^n \right)$, we have by Assumption 5.1 that

$$F_{\mathcal{H}}(\mathbf{w}_{t+1}) - F_{\mathcal{H}}(\mathbf{w}_t) \leq \langle \nabla F_{\mathcal{H}}(\mathbf{w}_t), \mathbf{w}_{t+1} - \mathbf{w}_t \rangle + \frac{L}{2} \left\| \mathbf{w}_{t+1} - \mathbf{w}_t \right\|^2$$
$$= -\eta \left\langle \nabla F_{\mathcal{H}}(\mathbf{w}_t), \sum_{\ell=1}^{K} R_{t,\ell} \mathbf{z}_{t,\ell}^r \right\rangle + \frac{\eta^2 L}{2} \left\| \sum_{\ell=1}^{K} R_{t,\ell} \mathbf{z}_{t,\ell}^r \right\|^2 \tag{1}$$

We first seek an upper bound to the first term:

$$
\begin{aligned}
&-\eta/K\left\langle K\nabla F_{\mathcal{H}}(\mathbf{w}_t), \sum_{\ell=1}^{K} R_{t,\ell}\mathbf{Z}_{t,\ell}\right\rangle \\
&= -\eta/(2K)\left(K^2\left\|\nabla F_{\mathcal{H}}(\mathbf{w}_t)\right\|^2 + \left\|\sum_{\ell=1}^{K} R_{t,\ell}\mathbf{Z}_{t,\ell}\right\|^2 - \left\|K\nabla F_{\mathcal{H}}(\mathbf{w}_t) - \sum_{\ell=1}^{K} R_{t,\ell}\mathbf{Z}_{t,\ell}\right\|^2\right).
\end{aligned}
$$

The leftmost term is the quantity of interest and will later be brought to the LHS of the equation. The prefactor of the middle term will, by an appropriate choice of the learning rate, be made small enough such that $c \triangleq \frac{\eta^2 L}{2} - \frac{\eta}{2K} \leq 0$, and hence we can bound the term $c\left\|\sum_{\ell=1}^{K} R_{t,\ell}\mathbf{Z}_{t,\ell}\right\|^2$ by 0. It remains to find a bound for the rightmost term $\left\|K\nabla F_{\mathcal{H}}(\mathbf{w}_t) - \sum_{\ell=1}^{K} R_{t,\ell}\mathbf{Z}_{t,\ell}\right\|^2$. Let $\bar{\mathbf{g}}_{\mathcal{H}}(\mathbf{w}_{t,\ell}^i, \mathbf{Z}_{t,\ell}) \triangleq \frac{1}{|\mathcal{H}|}\sum_{i\in\mathcal{H}} \mathbf{g}_i(\mathbf{w}_{t,\ell}^i, \mathbf{Z}_{t,\ell})$. By expansion, we can obtain

$$
\begin{aligned}
&\left\|K\nabla F_{\mathcal{H}}(\mathbf{w}_t) - \sum_{\ell=1}^{K}\mathbf{Z}_{t,\ell}R_{t,\ell}\right\|^2 \\
&= \Bigg\|K\nabla F_{\mathcal{H}}(\mathbf{w}_t) + \sum_{\ell=1}^{K}\Bigg(-\nabla F_{\mathcal{H}}(\hat{\mathbf{w}}_{t,\ell}) + \nabla F_{\mathcal{H}}(\hat{\mathbf{w}}_{t,\ell}) - \sum_{i\in\mathcal{H}}\frac{\nabla F_i(\mathbf{w}_{t,\ell}^i)}{|\mathcal{H}|} \\
&\qquad + \sum_{i\in\mathcal{H}}\frac{\nabla F_i(\mathbf{w}_{t,\ell}^i)}{|\mathcal{H}|} - \mathbf{Z}_{t,\ell}\bar{\mathbf{g}}_{\mathcal{H}}(\mathbf{w}_{t,\ell}^i, \mathbf{Z}_{t,\ell}) + \mathbf{Z}_{t,\ell}\bar{\mathbf{g}}_{\mathcal{H}}(\mathbf{w}_{t,\ell}^i, \mathbf{Z}_{t,\ell}) - \mathbf{Z}_{t,\ell}R_{t,\ell}\Bigg)\Bigg\|^2 \\
&\leq 4K\sum_{\ell=1}^{K}\left\|\nabla F_{\mathcal{H}}(\mathbf{w}_t) - \nabla F_{\mathcal{H}}(\hat{\mathbf{w}}_{t,\ell})\right\|^2 + 4\left\|\sum_{\ell=1}^{K}\left(\mathbf{Z}_{t,\ell}\bar{\mathbf{g}}_{\mathcal{H}}(\mathbf{w}_{t,\ell}^i, \mathbf{Z}_{t,\ell}) - \mathbf{Z}_{t,\ell}R_{t,\ell}\right)\right\|^2 \\
&+ 4K\sum_{\ell=1}^{K}\left\|\nabla F_{\mathcal{H}}(\hat{\mathbf{w}}_{t,\ell}) - \sum_{i\in\mathcal{H}}\frac{\nabla F_i(\mathbf{w}_{t,\ell}^i)}{|\mathcal{H}|}\right\|^2 + 4\left\|\sum_{\ell=1}^{K}\left(\sum_{i\in\mathcal{H}}\frac{\nabla F_i(\mathbf{w}_{t,\ell}^i)}{|\mathcal{H}|} - \mathbf{Z}_{t,\ell}\bar{\mathbf{g}}_{\mathcal{H}}(\mathbf{w}_{t,\ell}^i, \mathbf{Z}_{t,\ell})\right)\right\|^2 \\
&\overset{(a)}{\leq} 4K\sum_{\ell=1}^{K}L^2\left\|\mathbf{w}_t - \hat{\mathbf{w}}_{t,\ell}\right\|^2 + 4\left\|\sum_{\ell=1}^{K}\left(\mathbf{Z}_{t,\ell}\bar{\mathbf{g}}_{\mathcal{H}}(\mathbf{w}_{t,\ell}^i, \mathbf{Z}_{t,\ell}) - \mathbf{Z}_{t,\ell}R_{t,\ell}\right)\right\|^2 + 4K\sum_{\ell=1}^{K}\frac{D^2}{|\mathcal{H}|}\sum_{i\in\mathcal{H}}\left\|\mathbf{w}_{t,\ell}^i - \hat{\mathbf{w}}_{t,\ell}\right\|^2 \\
&+ 4\left\|\sum_{\ell=1}^{K}\left(\sum_{i\in\mathcal{H}}\frac{\nabla F_i(\mathbf{w}_{t,\ell}^i)}{|\mathcal{H}|} - \mathbf{Z}_{t,\ell}\bar{\mathbf{g}}_{\mathcal{H}}(\mathbf{w}_{t,\ell}^i, \mathbf{Z}_{t,\ell})\right)\right\|^2,
\end{aligned}
$$

where (a) holds by Assumption 5.1 and Assumption 5.4. We take the expectation on both sides and obtain

$$
\begin{aligned}
&\mathbb{E}\left[\left\|K\nabla F_{\mathcal{H}}(\mathbf{w}_t) - \sum_{\ell=1}^{K}\mathbf{Z}_{t,\ell}R_{t,\ell}\right\|^2\right] \\
&\leq 4K\sum_{\ell=1}^{K}L^2\mathbb{E}\left[\left\|\mathbf{w}_t - \hat{\mathbf{w}}_{t,\ell}\right\|^2\right] + 4\mathbb{E}\left[\sum_{\ell=1}^{K}\left\|\mathbf{Z}_{t,\ell}\bar{\mathbf{g}}_{\mathcal{H}}(\mathbf{w}_{t,\ell}^i, \mathbf{Z}_{t,\ell}) - \mathbf{Z}_{t,\ell}R_{t,\ell}\right\|^2\right] \\
&+ 4K\sum_{\ell=1}^{K}\frac{D^2}{|\mathcal{H}|}\sum_{i\in\mathcal{H}}\mathbb{E}\left[\left\|\mathbf{w}_{t,\ell}^i - \hat{\mathbf{w}}_{t,\ell}\right\|^2\right] + 4\mathbb{E}\left[\left\|\sum_{\ell=1}^{K}\left(\sum_{i\in\mathcal{H}}\frac{\nabla F_i(\mathbf{w}_{t,\ell}^i)}{|\mathcal{H}|} - \mathbf{Z}_{t,\ell}\bar{\mathbf{g}}_{\mathcal{H}}(\mathbf{w}_{t,\ell}^i, \mathbf{Z}_{t,\ell})\right)\right\|^2\right]. \qquad (2)
\end{aligned}
$$

We continue with bounding the individual terms, and start with the latter term.

$$
\mathbb{E}\left[\left\|\sum_{\ell=1}^{K}\left(\sum_{i\in\mathcal{H}}\frac{1}{|\mathcal{H}|}\nabla F_i(\mathbf{w}_{t,\ell}^i) - \mathbf{Z}_{t,\ell}\bar{\mathbf{g}}_{\mathcal{H}}(\mathbf{w}_{t,\ell}^i, \mathbf{Z}_{t,\ell})\right)\right\|^2\right]
$$

$$
\begin{aligned}
&\leq 2\mathbb{E}\left[\left\|\sum_{\ell=1}^{K}\sum_{i\in\mathcal{H}}\frac{1}{|\mathcal{H}|}\left(\nabla F_i(\mathbf{w}_{t,\ell}^i)-\nabla F_i^{\mu}(\mathbf{w}_{t,\ell}^i)\right)\right\|^2\right]+2\mathbb{E}\left[\left\|\sum_{\ell=1}^{K}\sum_{i\in\mathcal{H}}\frac{1}{|\mathcal{H}|}\left(\nabla F_i^{\mu}(\mathbf{w}_{t,\ell}^i)-\mathbf{Z}_{t,\ell}\mathbf{g}_i(\mathbf{w}_{t,\ell}^i,\mathbf{Z}_{t,\ell})\right)\right\|^2\right]\\
&\overset{(a)}{\leq} 2\frac{K}{|\mathcal{H}|}\sum_{i\in\mathcal{H}}\sum_{\ell=1}^{K}\left\|\left(\nabla F_i(\mathbf{w}_{t,\ell}^i)-\nabla F_i^{\mu}(\mathbf{w}_{t,\ell}^i)\right)\right\|^2+2\frac{1}{|\mathcal{H}|^2}\sum_{i\in\mathcal{H}}\mathbb{E}\left[\left\|\sum_{\ell=1}^{K}\left(\nabla F_i^{\mu}(\mathbf{w}_{t,\ell}^i)-\mathbf{Z}_{t,\ell}\mathbf{g}_i(\mathbf{w}_{t,\ell}^i,\mathbf{Z}_{t,\ell})\right)\right\|^2\right]\\
&\overset{(b)}{\leq} 2K^2L\mu+2\frac{1}{|\mathcal{H}|^2}\sum_{i\in\mathcal{H}}\sum_{\ell=1}^{K}\mathbb{E}\left[\left\|\nabla F_i^{\mu}(\mathbf{w}_{t,\ell}^i)-\mathbf{Z}_{t,\ell}\mathbf{g}_i(\mathbf{w}_{t,\ell}^i,\mathbf{Z}_{t,\ell})\right\|^2\right]\\
&\overset{(c)}{\leq} 2K^2L\mu+2\frac{1}{|\mathcal{H}|^2}\sum_{i\in\mathcal{H}}\sum_{\ell=1}^{K}\Bigg(\xi_9\mathbb{E}\left[\left\|\mathbf{w}_{t,\ell}^i-\hat{\mathbf{w}}_{t,\ell}\right\|^2\right]+\xi_{12}\\
&\qquad\qquad+\xi_{11}\mathbb{E}\left[\left\|\nabla F_{\mathcal{H}}(\mathbf{w}_t)\right\|^2\right]+\xi_{10}\sum_{\ell'=1}^{\ell-1}\frac{1}{|\mathcal{H}|}\sum_{i\in\mathcal{H}}\mathbb{E}\left[\left\|\mathbf{w}_{t,\ell'}^i-\hat{\mathbf{w}}_{t,\ell'}\right\|^2\right]\Bigg)\\
&\overset{(d)}{\leq} 2K^2L\mu+2\frac{1}{|\mathcal{H}|^2}\sum_{i\in\mathcal{H}}\sum_{\ell=1}^{K}\Bigg(\xi_9\mathbb{E}\left[\left\|\mathbf{w}_{t,\ell}^i-\hat{\mathbf{w}}_{t,\ell}\right\|^2\right]+\xi_{12}\\
&\qquad\qquad+\xi_{11}\mathbb{E}\left[\left\|\nabla F_{\mathcal{H}}(\mathbf{w}_t)\right\|^2\right]+\frac{\xi_{10}K(K-1)}{|\mathcal{H}|}\sum_{i\in\mathcal{H}}\mathbb{E}\left[\left\|\mathbf{w}_{t,\ell}^i-\hat{\mathbf{w}}_{t,\ell}\right\|^2\right]\Bigg)\\
&=2K^2L\mu+2\frac{1}{|\mathcal{H}|^2}\sum_{i\in\mathcal{H}}\sum_{\ell=1}^{K}\left((\xi_9+\xi_{10}K(K-1))\mathbb{E}\left[\left\|\mathbf{w}_{t,\ell}^i-\hat{\mathbf{w}}_{t,\ell}\right\|^2\right]+\xi_{12}+\xi_{11}\mathbb{E}\left[\left\|\nabla F_{\mathcal{H}}(\mathbf{w}_t)\right\|^2\right]\right),
\end{aligned}
\tag{3}
$$

where (a) is due to the independence of $\sum_{\ell=1}^{K}\left(\nabla F_i^{\mu}(\mathbf{w}_{t,\ell}^i)-\mathbf{Z}_{t,\ell}\mathbf{g}_i(\mathbf{w}_{t,\ell}^i,\mathbf{Z}_{t,\ell})\right)$ and $\sum_{\ell=1}^{K}\left(\nabla F_j^{\mu}(\mathbf{w}_{t,\ell}^j)-\mathbf{Z}_{t,\ell}\mathbf{g}_j(\mathbf{w}_{t,\ell}^j,\mathbf{Z}_{t,\ell})\right)$ for $i\neq j$. (b) follows from Proposition C.1 and (Wang et al., 2021, Lemma 2). (c) is by the application of Lemma C.5. (d) holds since $\sum_{\ell=1}^{K}\sum_{\ell'=1}^{\ell-1}x_{\ell'}\leq\sum_{\ell=1}^{K}\sum_{\ell'=1}^{\ell}x_{\ell'}\leq\frac{K(K-1)}{2}\sum_{\ell}x_{\ell}$.

We continue with bounding the robustness term using a double-sided application of an extension of the Johnson-Lindenstrauss Lemma as stated in the following Lemma C.8 and the application of Lemma C.6.

Lemma C.8 (Proposition 8, (Li, 2024)). *Let $\mathbf{Z}=(\mathbf{z}_1,\cdots,\mathbf{z}_\nu)^T\in\mathbb{R}^{\nu\times d}$ with $\mathbf{z}_r\sim\mathcal{U}(\mathbb{S}^d),\forall r\in[\nu]$. For a given vector $\mathbf{x}\in\mathbb{R}^d$, we have for $\epsilon>0$, $\delta<1/2$ with probability at least $1-\delta$ that*

$$
(1-\epsilon)\left\|\mathbf{x}\right\|^2\leq\left\|\mathbf{Z}\mathbf{x}\right\|^2\leq(1-\epsilon)\left\|\mathbf{x}\right\|^2
$$

for $\nu\geq 64\epsilon^{-2}\log(2/\delta)$.

For the robustness term, we have

$$
\begin{aligned}
&\mathbb{E}\left[\left\|\sum_{\ell=1}^{K}\mathbf{Z}_{t,\ell}\left(R_{t,\ell}-\bar{\mathbf{g}}_{\mathcal{H}}(\mathbf{w}_{t,\ell}^i,\mathbf{Z}_{t,\ell})\right)\right\|^2\right]\\
&\leq K\sum_{\ell=1}^{K}\left\|\mathbf{Z}_{t,\ell}\left(R_{t,\ell}-\bar{\mathbf{g}}_{\mathcal{H}}(\mathbf{w}_{t,\ell}^i,\mathbf{Z}_{t,\ell})\right)\right\|^2\\
&\overset{(a)}{\leq} K\sum_{\ell=1}^{K}(1+\epsilon)\frac{\kappa}{|\mathcal{H}|}\sum_{i\in\mathcal{H}}\mathbb{E}\left[\left\|\mathbf{g}_i(\mathbf{w}_{t,\ell}^i,\mathbf{Z}_{t,\ell})-\bar{\mathbf{g}}_{\mathcal{H}}(\mathbf{w}_{t,\ell}^i,\mathbf{Z}_{t,\ell})\right\|^2\right]\\
&\overset{(b)}{\leq} K\sum_{\ell=1}^{K}\frac{(1+\epsilon)}{(1-\epsilon)}\frac{\kappa}{|\mathcal{H}|}\sum_{i\in\mathcal{H}}\mathbb{E}\left[\left\|\mathbf{Z}_{t,\ell}(\mathbf{g}_i(\mathbf{w}_{t,\ell}^i,\mathbf{Z}_{t,\ell})-\bar{\mathbf{g}}_{\mathcal{H}}(\mathbf{w}_{t,\ell}^i,\mathbf{Z}_{t,\ell}))\right\|^2\right]\\
&\overset{(c)}{\leq} K\frac{(1+\epsilon)}{(1-\epsilon)}\kappa\left(\sum_{\ell=1}^{K}\frac{1}{|\mathcal{H}|}\sum_{i\in\mathcal{H}}\xi_{13}\mathbb{E}\left[\left\|\hat{\mathbf{w}}_{t,\ell}-\mathbf{w}_{t,\ell}^i\right\|^2\right]+\sum_{\ell=1}^{K}(\xi_{14}+\xi_{15})+\sum_{\ell=1}^{K}\xi_{16}\mathbb{E}\left[\left\|\nabla F_{\mathcal{H}}(\mathbf{w}_t)\right\|^2\right]\right)
\end{aligned}
\tag{4}
$$

where (a) and (b) are by the application of Lemma C.8 for in total $|\mathcal{H}|+1$ projections. By a union bound over Lemma C.8, the distance preservation holds with probability $1-\delta$ for $\epsilon \geq \sqrt{\frac{64}{\nu}\log(\frac{2(|\mathcal{H}|-1)}{\delta})}$. We choose the smallest possible ϵ. This must hold for each iteration, so the distance preservation holds w.p. at least $1-K\delta$ for all local epochs. (c) is by Lemma C.6.

To bound the local model divergence from the global model, by Lemma C.4, we have

$$\begin{aligned}\sum_{\ell=1}^{K}\mathbb{E}\left[\|\hat{\mathbf{w}}_{t,\ell}-\mathbf{w}_t\|^2\right] &\leq \sum_{\ell=1}^{K}\xi_6\mathbb{E}\left[\|\nabla F_{\mathcal{H}}(\mathbf{w}_t)\|^2\right]+\sum_{\ell=1}^{K}\xi_7+\sum_{\ell=1}^{K}\xi_8\sum_{\ell'=1}^{\ell-1}\frac{1}{|\mathcal{H}|}\sum_{i\in\mathcal{H}}\mathbb{E}\left[\left\|\mathbf{w}^i_{t,\ell'}-\hat{\mathbf{w}}_{t,\ell'}\right\|^2\right]\\ &\leq \sum_{\ell=1}^{K}\xi_6\mathbb{E}\left[\|\nabla F_{\mathcal{H}}(\mathbf{w}_t)\|^2\right]+\sum_{\ell=1}^{K}\xi_7+\sum_{\ell=1}^{K}\frac{\xi_8 K(K-1)}{|\mathcal{H}|}\sum_{i\in\mathcal{H}}\mathbb{E}\left[\left\|\mathbf{w}^i_{t,\ell}-\hat{\mathbf{w}}_{t,\ell}\right\|^2\right], \qquad (5)\end{aligned}$$

where the latter follows since $\sum_{\ell=1}^{K}\sum_{\ell'=1}^{\ell-1}x_{\ell'}\leq\sum_{\ell=1}^{K}\sum_{\ell'=1}^{\ell}x_{\ell'}\leq\frac{K(K-1)}{2}\sum_{\ell}x_\ell$.

Plugging (3), (4), and (5) into (2), we obtain

$$\begin{aligned}&\mathbb{E}\left[\left\|K\nabla F_{\mathcal{H}}(\mathbf{w}_t)-\sum_{\ell=1}^{K}\mathbf{Z}_{t,\ell}R_{t,\ell}\right\|^2\right]\\ &\leq 4KL^2\left(\sum_{\ell=1}^{K}\xi_6\mathbb{E}\left[\|\nabla F_{\mathcal{H}}(\mathbf{w}_t)\|^2\right]+\sum_{\ell=1}^{K}\xi_7+\sum_{\ell=1}^{K}\frac{\xi_8K(K-1)}{|\mathcal{H}|}\sum_{i\in\mathcal{H}}\mathbb{E}\left[\left\|\mathbf{w}^i_{t,\ell}-\hat{\mathbf{w}}_{t,\ell}\right\|^2\right]\right)\\ &+4\left(K\frac{(1+\epsilon)}{(1-\epsilon)}\kappa\left(\sum_{\ell=1}^{K}\frac{1}{|\mathcal{H}|}\sum_{i\in\mathcal{H}}\xi_{13}\mathbb{E}\left[\left\|\hat{\mathbf{w}}_{t,\ell}-\mathbf{w}^i_{t,\ell}\right\|^2\right]+\sum_{\ell=1}^{K}(\xi_{14}+\xi_{15})+\sum_{\ell=1}^{K}\xi_{16}\mathbb{E}\left[\|\nabla F_{\mathcal{H}}(\mathbf{w}_t)\|^2\right]\right)\right)\\ &+4K\sum_{\ell=1}^{K}\frac{D^2}{|\mathcal{H}|}\sum_{i\in\mathcal{H}}\mathbb{E}\left[\left\|\mathbf{w}^i_{t,\ell}-\hat{\mathbf{w}}_{t,\ell}\right\|^2\right]\\ &+4\left(2K^2L\mu+2\frac{1}{|\mathcal{H}|^2}\sum_{i\in\mathcal{H}}\sum_{\ell=1}^{K}\left((\xi_9+\xi_{10}K(K-1))\mathbb{E}\left[\left\|\mathbf{w}^i_{t,\ell}-\hat{\mathbf{w}}_{t,\ell}\right\|^2\right]+\xi_{12}+\xi_{11}\mathbb{E}\left[\|\nabla F_{\mathcal{H}}(\mathbf{w}_t)\|^2\right]\right)\right)\\ &\leq \xi_3'\mathbb{E}\left[\|\nabla F_{\mathcal{H}}(\mathbf{w}_t)\|^2\right]+\xi_4'\sum_{\ell=1}^{K}\frac{1}{|\mathcal{H}|}\sum_{i\in\mathcal{H}}\mathbb{E}\left[\left\|\hat{\mathbf{w}}_{t,\ell}-\mathbf{w}^i_{t,\ell}\right\|^2\right]+\xi_5'\\ &\leq \xi_3'\mathbb{E}\left[\|\nabla F_{\mathcal{H}}(\mathbf{w}_t)\|^2\right]+\xi_4'\xi_1\mathbb{E}\left[\|\nabla F_{\mathcal{H}}(\mathbf{w}_t)\|^2\right]+\xi_4'\xi_2+\xi_5',\\ &\leq (\xi_3'+\xi_4'\xi_1)\mathbb{E}\left[\|\nabla F_{\mathcal{H}}(\mathbf{w}_t)\|^2\right]+\xi_4'\xi_2+\xi_5', \qquad (6)\end{aligned}$$

where the penultimate step is by the application of Lemma C.7, and the constants read and can be bounded as

$$\begin{aligned}\xi_3' &\triangleq 4KL^2\sum_{\ell=1}^{K}\xi_6+4K\frac{(1+\epsilon)}{(1-\epsilon)}\kappa\sum_{\ell=1}^{K}\xi_{16}+4\cdot 2\frac{1}{|\mathcal{H}|^2}\sum_{i\in\mathcal{H}}\sum_{\ell=1}^{K}\xi_{11}\\ &\leq 4K^2L^2\xi_6+\left(8K^2\frac{(1+\epsilon)}{(1-\epsilon)}\kappa+8\frac{1}{|\mathcal{H}|}K\right)\xi_{11}\\ &\leq \xi_3 \triangleq 4K^2L^2 5\cdot 32\eta^2K\left(K+\frac{d}{|\mathcal{H}|\nu}\right)+\left(8K^2\frac{(1+\epsilon)}{(1-\epsilon)}\kappa+8\frac{1}{|\mathcal{H}|}K\right)\left(\frac{16d}{\nu}+\frac{80\cdot 32d}{\nu}L^2\eta^2K\left(K+\frac{d}{|\mathcal{H}|\nu}\right)\right)\\ \xi_4' &\triangleq 4KL^2\xi_8K(K-1)+4K\frac{(1+\epsilon)}{(1-\epsilon)}\kappa\xi_{13}+4KD^2+4\cdot 2\frac{1}{|\mathcal{H}|}(\xi_9+\xi_{10}K(K-1))\\ &\leq 4K^3L^2\xi_8+4K\frac{(1+\epsilon)}{(1-\epsilon)}\kappa\xi_{13}+4KD^2+8\frac{1}{|\mathcal{H}|}(\xi_9+\xi_{10}K^2)\end{aligned}$$

$$
\begin{aligned}
&\leq 4K^3L^2\xi_8 + 8K\frac{(1+\epsilon)}{(1-\epsilon)}\kappa\left(3D^2+3\zeta^2+\xi_9+\xi_{10}K^2\right) + 4KD^2 + 8\frac{1}{|\mathcal{H}|}(\xi_9+\xi_{10}K^2)\\
&\leq 4K^3L^2\xi_8 + 8K\frac{(1+\epsilon)}{(1-\epsilon)}\kappa\left(3D^2+3\zeta^2\right) + 4KD^2 + 8\left(\frac{1}{|\mathcal{H}|}+K\frac{(1+\epsilon)}{(1-\epsilon)}\kappa\right)(\xi_9+\xi_{10}K^2)\\
&\leq 20\cdot 32K^3L^2\eta^2\left(KD^2+\frac{1}{|\mathcal{H}|}\frac{d}{\nu}L^2\right) + 8K\frac{(1+\epsilon)}{(1-\epsilon)}\kappa\left(3D^2+3\zeta^2\right) + 4KD^2\\
&+ 8\left(\frac{1}{|\mathcal{H}|}+K\frac{(1+\epsilon)}{(1-\epsilon)}\kappa\right)\left(\frac{16d}{\nu}L^2+\frac{80\cdot 32d}{\nu}L^2\eta^2\left(KD^2+\frac{1}{|\mathcal{H}|}\frac{d}{\nu}L^2\right)K^2\right)\\
&\leq \xi_4 \triangleq 20\cdot 32K^2L^2\eta^2\left(KD^2+\frac{1}{|\mathcal{H}|}\frac{d}{\nu}L^2\right)\left(K+\frac{4d}{\nu}\right)\\
&\qquad + 8K\frac{(1+\epsilon)}{(1-\epsilon)}\kappa\left(3D^2+3\zeta^2+\frac{16d}{\nu}L^2\right) + 4KD^2 + 8\frac{1}{|\mathcal{H}|}\frac{16d}{\nu}L^2\\
\xi_5' &\triangleq 4KL^2\sum_{\ell=1}^{K}\xi_7 + 4K\frac{(1+\epsilon)}{(1-\epsilon)}\kappa\sum_{\ell=1}^{K}(\xi_{14}+\xi_{15}) + 4\cdot 2K^2L\mu + 4\cdot 2\frac{1}{|\mathcal{H}|^2}\sum_{i\in\mathcal{H}}\sum_{\ell=1}^{K}\xi_{12}\\
&\leq 4K^2L^2\xi_7 + 4K^2\frac{(1+\epsilon)}{(1-\epsilon)}\kappa(\xi_{14}+\xi_{15}) + 8K^2L\mu + 8\frac{1}{|\mathcal{H}|}K\xi_{12}\\
&\leq 4K^2L^2\xi_7 + 4K^2\frac{(1+\epsilon)}{(1-\epsilon)}\kappa(2\xi_{12}+6L) + 8K^2L\mu + 8\frac{1}{|\mathcal{H}|}K\xi_{12}\\
&\leq 4K^2L^2\xi_7 + 4K^2\frac{(1+\epsilon)}{(1-\epsilon)}\kappa 6L + 8K^2L\mu + \left(8\frac{1}{|\mathcal{H}|}K+8K^2\frac{(1+\epsilon)}{(1-\epsilon)}\kappa\right)\xi_{12}\\
&\leq 4K^2L^2\left(5K\eta^2\frac{d}{|\mathcal{H}|\nu}\left(32\zeta^2+8\sigma^2+L^2\mu^2d\right)+5\eta^2 24K^2L\mu\right) + 4K^2\frac{(1+\epsilon)}{(1-\epsilon)}\kappa 6L + 8K^2L\mu\\
&+ \left(8\frac{1}{|\mathcal{H}|}K+8K^2\frac{(1+\epsilon)}{(1-\epsilon)}\kappa\right)\left(\left(\frac{d}{2\nu}+\frac{80d^2}{\nu^2|\mathcal{H}|}L^2K\eta^2\right)\left(32\zeta^2+8\sigma^2+L^2\mu^2d\right)+\frac{80\cdot 24d}{\nu}L^2\eta^2K^2L\mu\right)\\
&\leq \xi_5 \triangleq \left(32\zeta^2+8\sigma^2+L^2\mu^2d\right)\left(20K^3L^2\eta^2\frac{d}{|\mathcal{H}|\nu}+\left(8\frac{1}{|\mathcal{H}|}K+8K^2\frac{(1+\epsilon)}{(1-\epsilon)}\kappa\right)\left(\frac{d}{2\nu}+\frac{80d^2}{\nu^2|\mathcal{H}|}L^2K\eta^2\right)\right)\\
&\qquad + 4K^2L^2 5\eta^2 24K^2L\mu + 4K^2\frac{(1+\epsilon)}{(1-\epsilon)}\kappa 6L + 8K^2L\mu + \left(8\frac{1}{|\mathcal{H}|}K+8K^2\frac{(1+\epsilon)}{(1-\epsilon)}\kappa\right)\left(\frac{80\cdot 24d}{\nu}L^2\eta^2K^2L\mu\right)
\end{aligned}
$$

By taking the expectation over (1) and replacing $\mathbb{E}\left[\left\|K\nabla F_{\mathcal{H}}(\mathbf{w}_t)-\sum_{\ell=1}^{K}\mathbf{Z}_{t,\ell}R_{t,\ell}\right\|^2\right]$ by (6), we can write

$$
\begin{aligned}
&\mathbb{E}\left[F_{\mathcal{H}}(\mathbf{w}_{t+1})\right]-\mathbb{E}\left[F_{\mathcal{H}}(\mathbf{w}_t)\right]\\
&\leq -\eta/(2)K\mathbb{E}\left[\left\|\nabla F_{\mathcal{H}}(\mathbf{w}_t)\right\|^2\right]+\left(\frac{\eta^2L}{2}-\frac{\eta}{2K}\right)\mathbb{E}\left[\left\|\sum_{\ell=1}^{K}R_{t,\ell}\mathbf{Z}_{t,\ell}\right\|^2\right]\\
&+\eta/(2K)\mathbb{E}\left[\left\|K\nabla F_{\mathcal{H}}(\mathbf{w}_t)-\sum_{\ell=1}^{K}R_{t,\ell}\mathbf{Z}_{t,\ell}\right\|^2\right]\\
&\overset{(a)}{\leq} -\eta/(2)K\mathbb{E}\left[\left\|\nabla F_{\mathcal{H}}(\mathbf{w}_t)\right\|^2\right]+\eta/(2K)\left((\xi_3+\xi_4\xi_1)\mathbb{E}\left[\left\|\nabla F_{\mathcal{H}}(\mathbf{w}_t)\right\|^2\right]+\xi_4\xi_2+\xi_5\right)\\
&\overset{(b)}{\leq} -\frac{\eta K}{4}\mathbb{E}\left[\left\|\nabla F_{\mathcal{H}}(\mathbf{w}_t)\right\|^2\right]+\eta/(2K)\left(\xi_4\xi_2+\xi_5\right),
\end{aligned}
$$

where (a) holds when $\eta\leq\frac{1}{KL}$ and (b) assumes that $\frac{\eta}{2K}(\xi_3+\xi_4\xi_1)\leq\frac{\eta K}{4}$.

Reordering and telescoping over t, we obtain

$$
\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}\left[\left\|\nabla F_{\mathcal{H}}(\mathbf{w}_t)\right\|^2\right]\leq\frac{4(\mathbb{E}\left[F_{\mathcal{H}}(\mathbf{w}_1)\right]-\mathbb{E}\left[F_{\mathcal{H}}(\mathbf{w}_{T+1})\right])}{T\eta K}+\eta/(2K)\left(\xi_4\xi_2+\xi_5\right)
$$

with probability $1-\delta KT$ by a union bound argument over all global iterations T. We let $\Delta \triangleq \delta KT$, and obtain $\epsilon \geq \sqrt{\frac{64}{\nu}\log(\frac{2(|\mathcal{H}|-1)}{\delta})} = \sqrt{\frac{64}{\nu}\log(\frac{2(|\mathcal{H}|-1)TK}{\Delta})}$. Since it is required to satisfy $\epsilon < 1$, the proof holds for $\nu \geq 64\log(\frac{2(|\mathcal{H}|-1)TK}{\Delta})$. Noting that $F_{\mathcal{H}}(\mathbf{w}_{T+1}) \geq F_{\mathcal{H}}^{\star}$ by definition concludes the proof. The requirements on the learning rate are summarized as follows:

- $24K\eta^2L^2 \leq \frac{1}{3K} \rightarrow \eta \leq \frac{1}{\sqrt{72K^2L}}$
- $2\eta^2\frac{1}{|\mathcal{H}|}\frac{4d}{\nu}4L^2 \leq \frac{1}{3K} \rightarrow \eta \leq \sqrt{\frac{|\mathcal{H}|\nu}{96KdL^2}}$
- $\eta \leq \sqrt{6\frac{D^2}{K^2}+6\frac{\zeta^2}{K^2}+\frac{32d}{\nu K^2}L^2}$

The constants are summarized as

$$\xi_1 \triangleq 4\eta^2K^3\frac{16}{\nu}+\frac{4d}{\nu}4L^2 5\cdot 32\eta^2K\left(K+\frac{d}{|\mathcal{H}|\nu}\right)$$

$$\xi_2 \triangleq 4\eta^2K^3\left(\left(\frac{d}{2\nu}+\frac{80d^2}{\nu^2|\mathcal{H}|}L^2K\eta^2\right)\left(32\zeta^2+8\sigma^2+L^2\mu^2d\right)+\frac{80\cdot 24d}{\nu}L^2\eta^2K^2L\mu\right)+12\eta^2K^3L$$

$$\xi_3 \triangleq 4K^2L^2 5\cdot 32\eta^2K\left(K+\frac{d}{|\mathcal{H}|\nu}\right)+\left(8K^2\frac{(1+\epsilon)}{(1-\epsilon)}\kappa+8\frac{1}{|\mathcal{H}|}K\right)\left(\frac{16d}{\nu}+\frac{80\cdot 32d}{\nu}L^2\eta^2K\left(K+\frac{d}{|\mathcal{H}|\nu}\right)\right)$$

$$\xi_4 \triangleq 20\cdot 32K^2L^2\eta^2\left(KD^2+\frac{1}{|\mathcal{H}|}\frac{d}{\nu}L^2\right)\left(K+\frac{4d}{\nu}\right)+8K\frac{(1+\epsilon)}{(1-\epsilon)}\kappa\left(3D^2+3\zeta^2+\frac{16d}{\nu}L^2\right)+4KD^2+8\frac{1}{|\mathcal{H}|}\frac{16d}{\nu}L^2$$

$$\xi_5 \triangleq \left(32\zeta^2+8\sigma^2+L^2\mu^2d\right)\left(20K^3L^2\eta^2\frac{d}{|\mathcal{H}|\nu}+\left(8\frac{1}{|\mathcal{H}|}K+8K^2\frac{(1+\epsilon)}{(1-\epsilon)}\kappa\right)\left(\frac{d}{2\nu}+\frac{80d^2}{\nu^2|\mathcal{H}|}L^2K\eta^2\right)\right)$$
$$+4K^2L^2 5\eta^2 24K^2L\mu+4K^2\frac{(1+\epsilon)}{(1-\epsilon)}\kappa 6L+8K^2L\mu+\left(8\frac{1}{|\mathcal{H}|}K+8K^2\frac{(1+\epsilon)}{(1-\epsilon)}\kappa\right)\left(\frac{80\cdot 24d}{\nu}L^2\eta^2K^2L\mu\right),$$

and, with $\epsilon' \triangleq \frac{(1+\epsilon)}{(1-\epsilon)}$, can be approximated by

$$\xi_1 = \Theta\left(\eta^2K^3\frac{d}{\nu}+\frac{d}{\nu}L^2\eta^2K\left(K+\frac{d}{|\mathcal{H}|\nu}\right)\right) \tag{7}$$

$$\xi_2 = \Theta\left(\eta^2K^3\left(L+\left(\frac{d}{\nu}+\frac{d^2}{\nu^2|\mathcal{H}|}L^2K^2\eta^2\right)\left(\zeta^2+\sigma^2+L^2\mu d\right)\right)\right) \tag{8}$$

$$\xi_3 = \Theta\left(\frac{d}{\nu}\left(K^2\epsilon'\kappa+\frac{K}{|\mathcal{H}|}\right)\left(1+L^2\eta^2K\left(K+\frac{d}{|\mathcal{H}|\nu}\right)(1+\frac{\nu}{d})\right)\right) \tag{9}$$

$$\xi_4 = \Theta\left(K^2L^2\eta^2\left(K^2D^2+\frac{1}{|\mathcal{H}|}\frac{d^2}{\nu^2}L^2\right)+K\epsilon'\kappa\left(D^2+\zeta^2+\frac{d}{\nu}L^2\right)\right) \tag{10}$$

$$\xi_5 = \Theta\Bigg(\left(\zeta^2+\sigma^2+L^2\mu^2d\right)\left(\left(K^2\epsilon'\kappa\right)\left(\frac{d}{\nu}+\frac{d^2}{\nu^2|\mathcal{H}|}L^2K\eta^2\right)\right) \tag{11}$$

$$+K^2L(\epsilon'\kappa+\mu)+\left(K^2\epsilon'\kappa\right)\left(\frac{d}{\nu}L^2\eta^2K^2L\mu\right)\Bigg). \tag{12}$$

□

We now continue to prove all intermediate Lemmas C.4 to C.7.

Proof of Lemma C.4. By definition, $\mathbb{E}\left[\left\|\hat{\mathbf{w}}_{t,1}-\mathbf{w}_t\right\|^2\right]=0$. For $\ell \in \{2,\cdots,K\}$, we have

$$\mathbb{E}\left[\left\|\hat{\mathbf{w}}_{t,\ell}-\mathbf{w}_t\right\|^2\right]$$

$$
\begin{aligned}
&= \mathbb{E}\left[\left\|\hat{\mathbf{w}}_{t,\ell-1} - \frac{\eta}{|\mathcal{H}|}\sum_{i\in\mathcal{H}}\mathbf{Z}_{t,\ell-1}\mathbf{g}_i(\mathbf{w}^i_{t,\ell-1},\mathbf{Z}_{t,\ell-1}) - \mathbf{w}_t\right\|^2\right]\\
&= \mathbb{E}\Bigg[\Bigg\|\hat{\mathbf{w}}_{t,\ell-1} - \mathbf{w}_t - \eta\Bigg(\frac{1}{|\mathcal{H}|}\sum_{i\in\mathcal{H}}\mathbf{Z}_{t,\ell-1}\mathbf{g}_i(\mathbf{w}^i_{t,\ell-1},\mathbf{Z}_{t,\ell-1}) - \frac{1}{|\mathcal{H}|}\sum_{i\in\mathcal{H}}\nabla F^\mu_i(\mathbf{w}^i_{t,\ell-1}) + \frac{1}{|\mathcal{H}|}\sum_{i\in\mathcal{H}}\nabla F^\mu_i(\mathbf{w}^i_{t,\ell-1})\\
&\qquad - \frac{1}{|\mathcal{H}|}\sum_{i\in\mathcal{H}}\nabla F_i(\mathbf{w}^i_{t,\ell-1}) + \frac{1}{|\mathcal{H}|}\sum_{i\in\mathcal{H}}\nabla F_i(\mathbf{w}^i_{t,\ell-1}) - \nabla F_{\mathcal{H}}(\hat{\mathbf{w}}_{t,\ell-1}) + \nabla F_{\mathcal{H}}(\hat{\mathbf{w}}_{t,\ell-1}) - \nabla F_{\mathcal{H}}(\mathbf{w}_t) + \nabla F_{\mathcal{H}}(\mathbf{w}_t)\Bigg)\Bigg\|^2\Bigg]\\
&\le (1+\frac{1}{\tau})\mathbb{E}\left[\|\hat{\mathbf{w}}_{t,\ell-1} - \mathbf{w}_t\|^2\right] +\\
&(1+\tau)\mathbb{E}\Bigg[\Bigg\| - \eta\Bigg(\frac{1}{|\mathcal{H}|}\sum_{i\in\mathcal{H}}\mathbf{Z}_{t,\ell-1}\mathbf{g}_i(\mathbf{w}^i_{t,\ell-1},\mathbf{Z}_{t,\ell-1}) - \frac{1}{|\mathcal{H}|}\sum_{i\in\mathcal{H}}\nabla F^\mu_i(\mathbf{w}^i_{t,\ell-1}) + \frac{1}{|\mathcal{H}|}\sum_{i\in\mathcal{H}}\nabla F^\mu_i(\mathbf{w}^i_{t,\ell-1})\\
&\qquad - \frac{1}{|\mathcal{H}|}\sum_{i\in\mathcal{H}}\nabla F_i(\mathbf{w}^i_{t,\ell-1}) + \frac{1}{|\mathcal{H}|}\sum_{i\in\mathcal{H}}\nabla F_i(\mathbf{w}^i_{t,\ell-1}) - \nabla F_{\mathcal{H}}(\hat{\mathbf{w}}_{t,\ell-1}) + \nabla F_{\mathcal{H}}(\hat{\mathbf{w}}_{t,\ell-1}) - \nabla F_{\mathcal{H}}(\mathbf{w}_t) + \nabla F_{\mathcal{H}}(\mathbf{w}_t)\Bigg)\Bigg\|^2\Bigg]\\
&\overset{(b)}{\le} (1+\frac{1}{\tau})\mathbb{E}\left[\|\hat{\mathbf{w}}_{t,\ell-1} - \mathbf{w}_t\|^2\right]\\
&\quad + 2(1+\tau)\eta^2\mathbb{E}\Bigg[\Bigg\|\frac{1}{|\mathcal{H}|}\sum_{i\in\mathcal{H}}\nabla F^\mu_i(\mathbf{w}^i_{t,\ell-1}) - \frac{1}{|\mathcal{H}|}\sum_{i\in\mathcal{H}}\nabla F_i(\mathbf{w}^i_{t,\ell-1}) + \frac{1}{|\mathcal{H}|}\sum_{i\in\mathcal{H}}\nabla F_i(\mathbf{w}^i_{t,\ell-1}) - \nabla F_{\mathcal{H}}(\hat{\mathbf{w}}_{t,\ell-1})\\
&\qquad\qquad + \nabla F_{\mathcal{H}}(\hat{\mathbf{w}}_{t,\ell-1}) - \nabla F_{\mathcal{H}}(\mathbf{w}_t) + \nabla F_{\mathcal{H}}(\mathbf{w}_t)\Bigg\|^2\Bigg]\\
&\quad + 2\eta^2\frac{1}{|\mathcal{H}|^2}\sum_{i\in\mathcal{H}}\mathbb{E}\left[\left\|\mathbf{Z}_{t,\ell-1}\mathbf{g}_i(\mathbf{w}^i_{t,\ell-1},\mathbf{Z}_{t,\ell-1}) - \nabla F^\mu_i(\mathbf{w}^i_{t,\ell-1})\right\|^2\right]\\
&\le (1+\frac{1}{\tau})\mathbb{E}\left[\|\hat{\mathbf{w}}_{t,\ell-1} - \mathbf{w}_t\|^2\right] + 8(1+\tau)\eta^2\mathbb{E}\left[\left\|\frac{1}{|\mathcal{H}|}\sum_{i\in\mathcal{H}}\nabla F^\mu_i(\mathbf{w}^i_{t,\ell-1}) - \frac{1}{|\mathcal{H}|}\sum_{i\in\mathcal{H}}\nabla F_i(\mathbf{w}^i_{t,\ell-1})\right\|^2\right]\\
&\quad + 8(1+\tau)\eta^2\mathbb{E}\left[\|\nabla F_{\mathcal{H}}(\mathbf{w}_t)\|^2\right]\\
&\quad + 8(1+\tau)\eta^2\mathbb{E}\left[\left\|\frac{1}{|\mathcal{H}|}\sum_{i\in\mathcal{H}}\nabla F_i(\mathbf{w}^i_{t,\ell-1}) - \nabla F_{\mathcal{H}}(\hat{\mathbf{w}}_{t,\ell-1})\right\|^2\right] + 8(1+\tau)\eta^2\mathbb{E}\left[\|\nabla F_{\mathcal{H}}(\hat{\mathbf{w}}_{t,\ell-1}) - \nabla F_{\mathcal{H}}(\mathbf{w}_t)\|^2\right]\\
&\quad + 2\eta^2\frac{1}{|\mathcal{H}|^2}\sum_{i\in\mathcal{H}}\mathbb{E}\left[\left\|\mathbf{Z}_{t,\ell-1}\mathbf{g}_i(\mathbf{w}^i_{t,\ell-1},\mathbf{Z}_{t,\ell-1}) - \nabla F^\mu_i(\mathbf{w}^i_{t,\ell-1})\right\|^2\right]
\end{aligned}
\tag{13}
$$

$$
\begin{aligned}
&\overset{(c)}{\le} (1+\frac{1}{\tau})\mathbb{E}\left[\|\hat{\mathbf{w}}_{t,\ell-1} - \mathbf{w}_t\|^2\right] + 8(1+\tau)\eta^2 L\mu + 8(1+\tau)\eta^2\mathbb{E}\left[\|\nabla F_{\mathcal{H}}(\mathbf{w}_t)\|^2\right]\\
&\quad + 8(1+\tau)\eta^2\frac{D^2}{|\mathcal{H}|}\sum_{i\in\mathcal{H}}\mathbb{E}\left[\left\|\mathbf{w}^i_{t,\ell-1} - \hat{\mathbf{w}}_{t,\ell-1}\right\|^2\right] + 8(1+\tau)\eta^2L^2\mathbb{E}\left[\|\hat{\mathbf{w}}_{t,\ell-1} - \mathbf{w}_t\|^2\right]\\
&\quad + 2\eta^2\frac{1}{|\mathcal{H}|^2}\sum_{i\in\mathcal{H}}\frac{4d}{\nu}\left(4L^2\mathbb{E}\left[\left\|\mathbf{w}^i_{t,\ell-1} - \hat{\mathbf{w}}_{t,\ell-1}\right\|^2\right] + 4\zeta^2 + 4L^2\mathbb{E}\left[\|\hat{\mathbf{w}}_{t,\ell-1} - \mathbf{w}_t\|^2\right] + 4\mathbb{E}\left[\|\nabla F_{\mathcal{H}}(\mathbf{w}_t)\|^2\right]\right)\\
&\quad + 2\eta^2\frac{1}{|\mathcal{H}|}\frac{4d}{\nu}\sigma^2 + 2\eta^2\frac{1}{|\mathcal{H}|}\frac{L^2\mu^2d^2}{2\nu}
\end{aligned}
$$

$$
\begin{aligned}
&\overset{(d)}{\leq} \left((1+\frac{1}{\tau})+8(1+\tau)\eta^2L^2+2\eta^2\frac{1}{|\mathcal{H}|}\frac{4d}{\nu}4L^2\right)\mathbb{E}\left[\left\|\hat{\mathbf{w}}_{t,\ell-1}-\mathbf{w}_t\right\|^2\right]+8(1+\tau)\eta^2L\mu\\
&+\left(8(1+\tau)\eta^2+2\eta^2\frac{1}{|\mathcal{H}|}\frac{4d}{\nu}4\right)\mathbb{E}\left[\left\|\nabla F_{\mathcal{H}}(\mathbf{w}_t)\right\|^2\right]\\
&+\left(8(1+\tau)\eta^2\frac{D^2}{|\mathcal{H}|}+2\eta^2\frac{1}{|\mathcal{H}|^2}\frac{4d}{\nu}4L^2\right)\sum_{i\in\mathcal{H}}\mathbb{E}\left[\left\|\mathbf{w}^i_{t,\ell-1}-\hat{\mathbf{w}}_{t,\ell-1}\right\|^2\right]+2\eta^2\frac{1}{|\mathcal{H}|}\frac{4d}{\nu}4\zeta^2\\
&+2\eta^2\frac{1}{|\mathcal{H}|}\frac{4d}{\nu}\sigma^2+2\eta^2\frac{1}{|\mathcal{H}|}\frac{L^2\mu^2d^2}{2\nu}
\end{aligned}
$$

where (a) is by independence for $i\neq j$, and (b) is because $\|\mathbf{x}+\mathbf{y}\|^2=(1+1/\tau)\|\mathbf{x}\|^2+(1+\tau)\|\mathbf{y}\|^2$, $\tau>0$. (c) follows from Assumption 5.1, Assumption 5.4 and an intermediate step in the proof of Lemma C.5.

To ensure the bound holds uniformly for all $\ell\in[K]$, we now choose $\tau=3K-1$ and the learning rate small enough so that $\left((1+\frac{1}{\tau})+8(1+\tau)\eta^2L^2+2\eta^2\frac{1}{|\mathcal{H}|}\frac{4d}{\nu}4L^2\right)\leq 1+\frac{1}{K-1}$, i.e., that $8(1+\tau)\eta^2L^2\leq\frac{1}{3K}$ and $2\eta^2\frac{1}{|\mathcal{H}|}\frac{4d}{\nu}4L^2\leq\frac{1}{3K}$. With this choice of the learning rate, we have

$$
\begin{aligned}
&\mathbb{E}\left[\left\|\hat{\mathbf{w}}_{t,\ell}-\mathbf{w}_t\right\|^2\right]\\
&\leq(1+\frac{1}{K-1})\mathbb{E}\left[\left\|\hat{\mathbf{w}}_{t,\ell-1}-\mathbf{w}_t\right\|^2\right]+8(1+\tau)\eta^2L\mu+\left(8(1+\tau)\eta^2+2\eta^2\frac{1}{|\mathcal{H}|}\frac{4d}{\nu}4\right)\mathbb{E}\left[\left\|\nabla F_{\mathcal{H}}(\mathbf{w}_t)\right\|^2\right]\\
&+\left(8(1+\tau)\eta^2\frac{D^2}{|\mathcal{H}|}+2\eta^2\frac{1}{|\mathcal{H}|^2}\frac{4d}{\nu}4L^2\right)\sum_{i\in\mathcal{H}}\mathbb{E}\left[\left\|\mathbf{w}^i_{t,\ell-1}-\hat{\mathbf{w}}_{t,\ell-1}\right\|^2\right]+2\eta^2\frac{1}{|\mathcal{H}|}\frac{4d}{\nu}4\zeta^2\\
&+2\eta^2\frac{1}{|\mathcal{H}|}\frac{4d}{\nu}\sigma^2+2\eta^2\frac{1}{|\mathcal{H}|}\frac{L^2\mu^2d^2}{2\nu}\\
&\overset{(d)}{\leq}\xi_6'(\ell)\mathbb{E}\left[\left\|\nabla F_{\mathcal{H}}(\mathbf{w}_t)\right\|^2\right]+\xi_7'(\ell)+\xi_8'\sum_{\ell'=1}^{\ell-1}\sum_{i\in\mathcal{H}}\mathbb{E}\left[\left\|\mathbf{w}^i_{t,\ell'}-\hat{\mathbf{w}}_{t,\ell'}\right\|^2\right],
\end{aligned}
$$

where (e) is by the recursive application of (d) and the fact that $(1+\frac{1}{K})^\ell\leq(1+\frac{1}{\ell})^\ell\leq e$ for all $\ell\in[K]$. This concludes the proof. The constants are given as

$$
\begin{aligned}
\xi_6'(\ell)&\triangleq 5(\ell-1)\left(8(1+\tau)\eta^2+2\eta^2\frac{1}{|\mathcal{H}|}\frac{4d}{\nu}4\right)\leq\xi_6\triangleq 5\cdot 32\eta^2K\left(K+\frac{d}{|\mathcal{H}|\nu}\right)\\
\xi_7'(\ell)&\triangleq 5(\ell-1)\left(2\eta^2\frac{1}{|\mathcal{H}|}\frac{4d}{\nu}4\zeta^2+2\eta^2\frac{1}{|\mathcal{H}|}\frac{4d}{\nu}\sigma^2+2\eta^2\frac{1}{|\mathcal{H}|}\frac{L^2\mu^2d^2}{2\nu}+8(1+\tau)\eta^2L\mu\right)\\
&\leq\xi_7\triangleq 5K\eta^2\frac{d}{|\mathcal{H}|\nu}\left(32\zeta^2+8\sigma^2+L^2\mu^2d\right)+5\eta^224K^2L\mu\\
\xi_8'&\triangleq 5\left(8(1+\tau)\eta^2D^2+2\eta^2\frac{1}{|\mathcal{H}|}\frac{4d}{\nu}4L^2\right)\leq\xi_8\triangleq 5\cdot 32\eta^2\left(KD^2+\frac{1}{|\mathcal{H}|}\frac{d}{\nu}L^2\right)
\end{aligned}
$$

□

Proof of Lemma C.5. For the proof of the zero-order approximated gradient variance, we rely on the following intermediate lemma.

Lemma C.9. *The second moment of the gradient estimate can be bounded from above as*

$$
\mathbb{E}\left[\left\|\mathbf{z}g(\mathbf{w},\mathbf{z},\mu,\mathcal{D})\right\|^2\right]\leq 2d\left\|\nabla F_{\mathcal{H}}(\mathbf{w},\mathcal{D})\right\|^2+\frac{L^2\mu^2d^2}{2}.
$$

Proof.

$$
\mathbb{E}\left[\left\|\mathbf{z}g(\mathbf{w},\mathbf{z},\mu,\mathcal{D})\right\|^2\right]=\mathbb{E}\left[\left\|d\frac{F(\mathbf{w}+\mu\mathbf{z})-F(\mathbf{w}-\mu\mathbf{z})}{2\mu}\right\|^2\right]
$$

$$
\begin{aligned}
&= \mathbb{E}\left[\left\| d\frac{F(\mathbf{w}+\mu\mathbf{z}) - F(\mathbf{w}) + F(\mathbf{w}) - F(\mathbf{w}-\mu\mathbf{z})}{2\mu}\right\|^2\right] \\
&\leq \frac{1}{2}\mathbb{E}\left[\left\| d\frac{F(\mathbf{w}+\mu\mathbf{z}) - F(\mathbf{w})}{\mu}\right\|^2\right] + \frac{1}{2}\mathbb{E}\left[\left\| d\frac{F(\mathbf{w}-\mu\mathbf{z}) - F(\mathbf{w})}{\mu}\right\|^2\right] \\
&= \mathbb{E}\left[\left\| d\frac{F(\mathbf{w}+\mu\mathbf{z}) - F(\mathbf{w})}{\mu}\right\|^2\right] \\
&\leq 2d\left\|\nabla F_{\mathcal{H}}(\mathbf{w})\right\|^2 + \frac{L^2\mu^2 d^2}{2},
\end{aligned}
$$

where the penultimate step holds by symmetry and the last step is from (Gao et al., 2018, Lemma 4.1). □

We bound the zero-order approximated gradient variance as follows: Since $\mathbb{E}\left[\|Z - \mathbb{E}[Z]\|^2\right] \leq \mathbb{E}\left[\|Z\|^2\right]$, and $\mathbb{E}\left[\mathbf{z}_{t,\ell}^r g_i(\mathbf{w}_{t,\ell}^i, \mathbf{z}_{t,\ell}^r)\right] = \nabla F_i^\mu(\mathbf{w}_{t,\ell}^i)$, we have

$$
\begin{aligned}
&\mathbb{E}\left[\left\|\mathbf{Z}_{t,\ell}\mathbf{g}_i(\mathbf{w}_{t,\ell}^i, \mathbf{Z}_{t,\ell}) - \nabla F_i^\mu(\mathbf{w}_{t,\ell}^i)\right\|^2\right] = \mathbb{E}\left[\left\|\frac{1}{\nu}\sum_{r=1}^{\nu}\mathbf{z}_{t,\ell}^r g_i(\mathbf{w}_{t,\ell}^i, \mathbf{z}_{t,\ell}^r) - \nabla F_i^\mu(\mathbf{w}_{t,\ell}^i)\right\|^2\right] \\
&\overset{(a)}{=} \frac{1}{\nu^2}\sum_{r=1}^{\nu}\mathbb{E}\left[\left\|\mathbf{z}_{t,\ell}^r g_i(\mathbf{w}_{t,\ell}^i, \mathbf{z}_{t,\ell}^r) - \nabla F_i^\mu(\mathbf{w}_{t,\ell}^i)\right\|^2\right] \\
&\leq \frac{1}{\nu^2}\sum_{r=1}^{\nu}\mathbb{E}\left[\left\|\mathbf{z}_{t,\ell}^r g_i(\mathbf{w}_{t,\ell}^i, \mathbf{z}_{t,\ell}^r)\right\|^2\right] \\
&\overset{(b)}{\leq} \frac{2d}{\nu^2}\sum_{r=1}^{\nu}\mathbb{E}\left[\mathbb{E}\left[\left\|\mathbf{g}_i(\mathbf{w}_{t,\ell}^i)\right\|^2\right]\right] + \frac{L^2\mu^2 d^2}{2\nu} \\
&\leq \frac{2d}{\nu^2}\sum_{r=1}^{\nu}\mathbb{E}\left[\mathbb{E}\left[\left\|\mathbf{g}_i(\mathbf{w}_{t,\ell}^i) - \nabla F_i(\mathbf{w}_{t,\ell}^i) + \nabla F_i(\mathbf{w}_{t,\ell}^i)\right\|^2\right]\right] + \frac{L^2\mu^2 d^2}{2\nu} \\
&\leq \frac{2d}{\nu^2}\sum_{r=1}^{\nu}\mathbb{E}\left[2\mathbb{E}\left[\left\|\mathbf{g}_i(\mathbf{w}_{t,\ell}^i) - \nabla F_i(\mathbf{w}_{t,\ell}^i)\right\|^2\right] + 2\left\|\nabla F_i(\mathbf{w}_{t,\ell}^i)\right\|^2\right] + \frac{L^2\mu^2 d^2}{2\nu} \\
&\overset{(c)}{\leq} \frac{4d}{\nu^2}\sum_{r=1}^{\nu}\mathbb{E}\left[\left\|\nabla F_i(\mathbf{w}_{t,\ell}^i)\right\|^2\right] + \frac{4d}{\nu}\sigma^2 + \frac{L^2\mu^2 d^2}{2\nu}
\end{aligned} \tag{14}
$$

where (a) is due to the independence of $\mathbf{z}_{t,\ell}^r g_i(\mathbf{w}_{t,\ell}^i, \mathbf{z}_{t,\ell}^r)$ and $\mathbf{z}_{t,\ell}^{r'} g_i(\mathbf{w}_{t,\ell}^i, \mathbf{z}_{t,\ell}^{r'})$ for $r \neq r'$. (b) is by Lemma C.9. (c) follows from Assumption 5.2.

$$
\begin{aligned}
&\mathbb{E}\left[\left\|\nabla F_i(\mathbf{w}_{t,\ell}^i)\right\|^2\right] \\
&\quad= \mathbb{E}\left[\left\|\nabla F_i(\mathbf{w}_{t,\ell}^i) - \nabla F_i(\hat{\mathbf{w}}_{t,\ell}) + \nabla F_i(\hat{\mathbf{w}}_{t,\ell}) - \nabla F_{\mathcal{H}}(\hat{\mathbf{w}}_{t,\ell}) + \nabla F_{\mathcal{H}}(\hat{\mathbf{w}}_{t,\ell}) - \nabla F_{\mathcal{H}}(\mathbf{w}_t) + \nabla F_{\mathcal{H}}(\mathbf{w}_t)\right\|^2\right] \\
&\quad\leq 4\mathbb{E}\left[\left\|\nabla F_i(\mathbf{w}_{t,\ell}^i) - \nabla F_i(\hat{\mathbf{w}}_{t,\ell})\right\|^2\right] + 4\mathbb{E}\left[\left\|\nabla F_i(\hat{\mathbf{w}}_{t,\ell}) - \nabla F_{\mathcal{H}}(\hat{\mathbf{w}}_{t,\ell})\right\|^2\right] \\
&\quad+ 4\mathbb{E}\left[\left\|\nabla F_{\mathcal{H}}(\hat{\mathbf{w}}_{t,\ell}) - \nabla F_{\mathcal{H}}(\mathbf{w}_t)\right\|^2\right] + 4\mathbb{E}\left[\left\|\nabla F_{\mathcal{H}}(\mathbf{w}_t)\right\|^2\right] \\
&\quad\leq 4L^2\mathbb{E}\left[\left\|\mathbf{w}_{t,\ell}^i - \hat{\mathbf{w}}_{t,\ell}\right\|^2\right] + 4\zeta^2 + 4L^2\mathbb{E}\left[\left\|\hat{\mathbf{w}}_{t,\ell} - \mathbf{w}_t\right\|^2\right] + 4\mathbb{E}\left[\left\|\nabla F_{\mathcal{H}}(\mathbf{w}_t)\right\|^2\right].
\end{aligned}
$$

Substituting the result in (14), we obtain

$$
\mathbb{E}\left[\left\|\mathbf{Z}_{t,\ell}\mathbf{g}_i(\mathbf{w}_{t,\ell}^i, \mathbf{Z}_{t,\ell}) - \nabla F_i^\mu(\mathbf{w}_{t,\ell}^i)\right\|^2\right] = \mathbb{E}\left[\left\|\frac{1}{\nu}\sum_{r=1}^{\nu}\mathbf{z}_{t,\ell}^r g_i(\mathbf{w}_{t,\ell}^i, \mathbf{z}_{t,\ell}^r) - \nabla F_i^\mu(\mathbf{w}_{t,\ell}^i)\right\|^2\right]
$$

$$
\begin{aligned}
&\leq \frac{4d}{\nu^2}\sum_{r=1}^{\nu}\left(4L^2\mathbb{E}\left[\left\|\mathbf{w}_{t,\ell}^i-\hat{\mathbf{w}}_{t,\ell}\right\|^2\right]+4\zeta^2+4L^2\mathbb{E}\left[\left\|\hat{\mathbf{w}}_{t,\ell}-\mathbf{w}_t\right\|^2\right]+4\mathbb{E}\left[\left\|\nabla F_{\mathcal{H}}(\mathbf{w}_t)\right\|^2\right]\right)+\frac{4d}{\nu}\sigma^2+\frac{L^2\mu^2d^2}{2\nu}\\
&=\frac{4d}{\nu}\left(4L^2\mathbb{E}\left[\left\|\mathbf{w}_{t,\ell}^i-\hat{\mathbf{w}}_{t,\ell}\right\|^2\right]+4\zeta^2+4L^2\mathbb{E}\left[\left\|\hat{\mathbf{w}}_{t,\ell}-\mathbf{w}_t\right\|^2\right]+4\mathbb{E}\left[\left\|\nabla F_{\mathcal{H}}(\mathbf{w}_t)\right\|^2\right]\right)+\frac{4d}{\nu}\sigma^2+\frac{L^2\mu^2d^2}{2\nu}\\
&\leq\frac{4d}{\nu}\left(4L^2\mathbb{E}\left[\left\|\mathbf{w}_{t,\ell}^i-\hat{\mathbf{w}}_{t,\ell}\right\|^2\right]+4\zeta^2+4\mathbb{E}\left[\left\|\nabla F_{\mathcal{H}}(\mathbf{w}_t)\right\|^2\right]\right)+\frac{4d}{\nu}\sigma^2+\frac{L^2\mu^2d^2}{2\nu}\\
&+\frac{4d}{\nu}4L^2\left(\xi_6\mathbb{E}\left[\left\|\nabla F_{\mathcal{H}}(\mathbf{w}_t)\right\|^2\right]+\xi_7+\xi_8\sum_{\ell'=1}^{\ell-1}\frac{1}{|\mathcal{H}|}\sum_{i\in\mathcal{H}}\mathbb{E}\left[\left\|\mathbf{w}_{t,\ell'}^i-\hat{\mathbf{w}}_{t,\ell'}\right\|^2\right]\right)\\
&\leq\xi_9'\mathbb{E}\left[\left\|\mathbf{w}_{t,\ell}^i-\hat{\mathbf{w}}_{t,\ell}\right\|^2\right]+\xi_{12}'+\xi_{11}'\mathbb{E}\left[\left\|\nabla F_{\mathcal{H}}(\mathbf{w}_t)\right\|^2\right]+\xi_{10}'\sum_{\ell'=1}^{\ell-1}\frac{1}{|\mathcal{H}|}\sum_{i\in\mathcal{H}}\mathbb{E}\left[\left\|\mathbf{w}_{t,\ell'}^i-\hat{\mathbf{w}}_{t,\ell'}\right\|^2\right],
\end{aligned}
$$

where the latter is by Lemma C.4, and we have

$$
\begin{aligned}
\xi_9'&\triangleq\xi_9\triangleq\frac{4d}{\nu}4L^2\\
\xi_{10}'&\triangleq\frac{4d}{\nu}4L^2\xi_8\leq\xi_{10}\triangleq\frac{4d}{\nu}4L^25\cdot32\eta^2\left(KD^2+\frac{1}{|\mathcal{H}|}\frac{d}{\nu}L^2\right)\\
\xi_{11}'&\triangleq\frac{4d}{\nu}4+\frac{4d}{\nu}4L^2\xi_6\leq\xi_{11}\triangleq\frac{4d}{\nu}4+\frac{4d}{\nu}4L^25\cdot32\eta^2K\left(K+\frac{d}{|\mathcal{H}|\nu}\right)\\
\xi_{12}'&\triangleq\frac{4d}{\nu}4\zeta^2+\frac{4d}{\nu}\sigma^2+\frac{L^2\mu^2d^2}{2\nu}+\frac{4d}{\nu}4L^2\xi_7\\
&\leq\frac{d}{2\nu}\left(32\zeta^2+8\sigma^2+L^2\mu^2d\right)+\frac{4d}{\nu}4L^2\left(5K\eta^2\frac{d}{|\mathcal{H}|\nu}\left(32\zeta^2+8\sigma^2+L^2\mu^2d\right)+5\eta^224K^2L\mu\right)\\
&\leq\xi_{12}\triangleq\left(\frac{d}{2\nu}+\frac{80d^2}{\nu^2|\mathcal{H}|}L^2K\eta^2\right)\left(32\zeta^2+8\sigma^2+L^2\mu^2d\right)+\frac{4d}{\nu}4L^25\eta^224K^2L\mu.
\end{aligned}
$$

This concludes the proof. □

Proof of Lemma C.6. We start with stating an intermediate lemma proven by (Wang et al., 2024).

Lemma C.10 (Extracted from Lemma B.1, (Wang et al., 2024))**.** *The following holds for the divergence of the local gradients:*

$$
\left\|\frac{1}{|\mathcal{H}|}\sum_{j\in\mathcal{H}}\nabla F_j(\mathbf{w}_{t,\ell}^j)-\nabla F_i(\mathbf{w}_{t,\ell}^i)\right\|^2\leq3\frac{D^2}{|\mathcal{H}|}\sum_{j\in\mathcal{H}}\left\|\hat{\mathbf{w}}_{t,\ell}-\mathbf{w}_{t,\ell}^j\right\|^2+3L+3\zeta^2\left\|\hat{\mathbf{w}}_{t,\ell}-\mathbf{w}_{t,\ell}^i\right\|^2.
$$

Following similar lines as in the proof of (Wang et al., 2024, Lemma B.2), for some local iteration $m\in[\ell]$, we have

$$
\begin{aligned}
&\mathbb{E}\left[\left\|\mathbf{Z}_{t,m}\mathbf{g}_i(\mathbf{w}_{t,\ell}^i,\mathbf{Z}_{t,\ell})-\frac{1}{|\mathcal{H}|}\sum_{j\in\mathcal{H}}\mathbf{Z}_{t,m}\mathbf{g}_j(\mathbf{w}_{t,m}^j,\mathbf{Z}_{t,m})\right\|^2\right]\\
&=\mathbb{E}\Bigg[\Bigg\|\mathbf{Z}_{t,m}\mathbf{g}_i(\mathbf{w}_{t,m}^i,\mathbf{Z}_{t,m})-\nabla F_i(\mathbf{w}_{t,m}^i)+\nabla F_i(\mathbf{w}_{t,m}^i)\\
&-\frac{1}{|\mathcal{H}|}\sum_{j\in\mathcal{H}}\nabla F_j(\mathbf{w}_{t,m}^j)+\frac{1}{|\mathcal{H}|}\sum_{j\in\mathcal{H}}\nabla F_j(\mathbf{w}_{t,m}^j)-\frac{1}{|\mathcal{H}|}\sum_{j\in\mathcal{H}}\mathbf{Z}_{t,m}\mathbf{g}_j(\mathbf{w}_{t,m}^j,\mathbf{Z}_{t,m})\Bigg\|\Bigg]\\
&\leq2\mathbb{E}\left[\left\|\nabla F_i(\mathbf{w}_{t,m}^i)-\frac{1}{|\mathcal{H}|}\sum_{j\in\mathcal{H}}\nabla F_j(\mathbf{w}_{t,m}^j)\right\|^2\right]
\end{aligned}
$$

$$
\begin{aligned}
&+ 2\mathbb{E}\left[\left\|\mathbf{Z}_{t,m}\mathbf{g}_i(\mathbf{w}_{t,m}^i,\mathbf{Z}_{t,m}) - \nabla F_i(\mathbf{w}_{t,m}^i) + \frac{1}{|\mathcal{H}|}\sum_{j\in\mathcal{H}}\nabla F_j(\mathbf{w}_{t,m}^j) - \frac{1}{|\mathcal{H}|}\sum_{j\in\mathcal{H}}\mathbf{Z}_{t,m}\mathbf{g}_j(\mathbf{w}_{t,m}^j,\mathbf{Z}_{t,m})\right\|^2\right]\\
&\overset{(a)}{\leq} 2\mathbb{E}\left[\left\|\nabla F_i(\mathbf{w}_{t,m}^i) - \frac{1}{|\mathcal{H}|}\sum_{j\in\mathcal{H}}\nabla F_j(\mathbf{w}_{t,m}^j)\right\|^2\right]\\
&+ 2\mathbb{E}\left[\left\|\mathbf{Z}_{t,m}\mathbf{g}_i(\mathbf{w}_{t,m}^i,\mathbf{Z}_{t,m}) - \nabla F_i(\mathbf{w}_{t,m}^i)\right\|^2\right]\\
&\leq 2\mathbb{E}\left[3\frac{D^2}{|\mathcal{H}|}\sum_{j\in\mathcal{H}}\left\|\hat{\mathbf{w}}_{t,m} - \mathbf{w}_{t,m}^j\right\|^2 + 3L + 3\zeta^2\left\|\hat{\mathbf{w}}_{t,m} - \mathbf{w}_{t,m}^i\right\|^2\right]\\
&+ 2\mathbb{E}\left[\left\|\left(\mathbf{Z}_{t,m}\mathbf{g}_i(\mathbf{w}_{t,m}^i,\mathbf{Z}_{t,m}) - \nabla F_i(\mathbf{w}_{t,m}^i)\right)\right\|^2\right],
\end{aligned}
\tag{15}
$$

where (a) is since $\frac{1}{|\mathcal{H}|}\sum_{i\in\mathcal{H}}\left\|\mathbf{x}_i - \frac{1}{|\mathcal{H}|}\sum_{j\in\mathcal{H}}\mathbf{x}_j\right\|^2 = \frac{1}{|\mathcal{H}|}\sum_{j\in\mathcal{H}}\|\mathbf{x}_j\|^2 - \left\|\frac{1}{|\mathcal{H}|}\sum_{j\in\mathcal{H}}\mathbf{x}_j\right\|^2 \leq \frac{1}{|\mathcal{H}|}\sum_{i\in\mathcal{H}}\|\mathbf{x}_i\|^2$ for $\mathbf{x}_i\in\mathbb{R}^d, \forall i$, where we set $\mathbf{x}_i = \mathbf{Z}_{t,m}\mathbf{g}_i(\mathbf{w}_{t,m}^i,\mathbf{Z}_{t,m}) - \nabla F_i(\mathbf{w}_{t,m}^i)$.

Summing over all benign clients, we obtain

$$
\begin{aligned}
&\sum_{m=1}^{\ell}\frac{1}{|\mathcal{H}|}\sum_{i\in\mathcal{H}}\mathbb{E}\left[\left\|\mathbf{Z}_{t,m}\mathbf{g}_i(\mathbf{w}_{t,\ell}^i,\mathbf{Z}_{t,\ell}) - \frac{1}{|\mathcal{H}|}\sum_{j\in\mathcal{H}}\mathbf{Z}_{t,m}\mathbf{g}_j(\mathbf{w}_{t,m}^j,\mathbf{Z}_{t,m})\right\|^2\right]\\
&\overset{(a)}{\leq} 2\sum_{m=1}^{\ell}\mathbb{E}\left[\left(3\frac{D^2}{|\mathcal{H}|} + 3\frac{\zeta^2}{|\mathcal{H}|}\right)\sum_{i\in\mathcal{H}}\left\|\hat{\mathbf{w}}_{t,m} - \mathbf{w}_{t,m}^i\right\|^2 + 3L\right]\\
&+ 2\sum_{m=1}^{\ell}\frac{1}{|\mathcal{H}|}\sum_{i\in\mathcal{H}}\left(\xi_9\mathbb{E}\left[\left\|\mathbf{w}_{t,m}^i - \hat{\mathbf{w}}_{t,m}\right\|^2\right] + \xi_{12} + \xi_{11}\mathbb{E}\left[\left\|\nabla F_{\mathcal{H}}(\mathbf{w}_t)\right\|^2\right] + \xi_{10}\sum_{\ell'=1}^{m-1}\sum_{i\in\mathcal{H}}\mathbb{E}\left[\left\|\mathbf{w}_{t,\ell'}^i - \hat{\mathbf{w}}_{t,\ell'}\right\|^2\right]\right)\\
&\leq 2\sum_{m=1}^{\ell}\frac{1}{|\mathcal{H}|}\sum_{i\in\mathcal{H}}\Bigg(\left(3D^2 + 3\zeta^2 + \xi_9\right)\mathbb{E}\left[\left\|\mathbf{w}_{t,m}^i - \hat{\mathbf{w}}_{t,m}\right\|^2\right] + \xi_{12} + 3L\\
&+ \xi_{11}\mathbb{E}\left[\left\|\nabla F_{\mathcal{H}}(\mathbf{w}_t)\right\|^2\right] + \xi_{10}\sum_{\ell'=1}^{m-1}\sum_{i\in\mathcal{H}}\mathbb{E}\left[\left\|\mathbf{w}_{t,\ell'}^i - \hat{\mathbf{w}}_{t,\ell'}\right\|^2\right]\Bigg)\\
&\leq 2\sum_{m=1}^{\ell}\frac{1}{|\mathcal{H}|}\sum_{i\in\mathcal{H}}\Bigg(\left(3D^2 + 3\zeta^2 + \xi_9\right)\mathbb{E}\left[\left\|\mathbf{w}_{t,m}^i - \hat{\mathbf{w}}_{t,m}\right\|^2\right] + \xi_{12} + 3L\\
&+ \xi_{11}\mathbb{E}\left[\left\|\nabla F_{\mathcal{H}}(\mathbf{w}_t)\right\|^2\right] + \xi_{10}\ell(\ell-1)\frac{1}{|\mathcal{H}|}\sum_{i\in\mathcal{H}}\mathbb{E}\left[\left\|\mathbf{w}_{t,m}^i - \hat{\mathbf{w}}_{t,m}\right\|^2\right]\Bigg)\\
&\leq \sum_{m=1}^{\ell}\frac{1}{|\mathcal{H}|}\sum_{i\in\mathcal{H}}\xi_{13}'(\ell)\mathbb{E}\left[\left\|\hat{\mathbf{w}}_{t,m} - \mathbf{w}_{t,m}^i\right\|^2\right] + \sum_{m=1}^{\ell}(\xi_{14}+\xi_{15}) + \sum_{m=1}^{\ell}\xi_{16}\mathbb{E}\left[\left\|\nabla F_{\mathcal{H}}(\mathbf{w}_t)\right\|^2\right],
\end{aligned}
$$

where (a) is due to Lemma C.5 and (b) is since $\sum_{m=1}^{\ell}\sum_{\ell'=1}^{m}x_m \leq \frac{\ell(\ell-1)}{2}\sum_m x_m$. Thereby,

$$
\begin{aligned}
\xi_{13}'(\ell) &\triangleq 2(3D^2 + 3\zeta^2 + \xi_9 + \xi_{10}\ell(\ell-1)) \leq \xi_{13} \triangleq 6D^2 + 6\zeta^2 + 2\xi_9 + 2\xi_{10}K^2\\
\xi_{14} &\triangleq 2\xi_{12}\\
\xi_{15} &\triangleq 6L\\
\xi_{16} &\triangleq 2\xi_{11}.
\end{aligned}
$$

This concludes the proof. □

Proof of Lemma C.7. From Lemma C.6, we have

$$\sum_{\ell=1}^{K}\frac{1}{|\mathcal{H}|}\sum_{i\in\mathcal{H}}\mathbb{E}\left[\left\|\mathbf{w}_{t,\ell}^{i}-\hat{\mathbf{w}}_{t,\ell}\right\|^{2}\right]=\sum_{\ell=1}^{K}\frac{1}{|\mathcal{H}|}\sum_{i\in\mathcal{H}}\mathbb{E}\left[\left\|\eta\sum_{m=1}^{\ell}\left(\mathbf{Z}_{t,m}\mathbf{g}_{i}(\mathbf{w}_{t,\ell}^{i},\mathbf{Z}_{t,\ell})-\frac{1}{|\mathcal{H}|}\sum_{j\in\mathcal{H}}\mathbf{Z}_{t,m}\mathbf{g}_{j}(\mathbf{w}_{t,m}^{j},\mathbf{Z}_{t,m})\right)\right\|^{2}\right]$$

$$\leq\eta^{2}\sum_{\ell=1}^{K}\ell\sum_{m=1}^{\ell}\frac{1}{|\mathcal{H}|}\sum_{i\in\mathcal{H}}\mathbb{E}\left[\left\|\mathbf{Z}_{t,m}\mathbf{g}_{i}(\mathbf{w}_{t,\ell}^{i},\mathbf{Z}_{t,\ell})-\frac{1}{|\mathcal{H}|}\sum_{j\in\mathcal{H}}\mathbf{Z}_{t,m}\mathbf{g}_{j}(\mathbf{w}_{t,m}^{j},\mathbf{Z}_{t,m})\right\|^{2}\right]$$

$$=\eta^{2}\sum_{\ell=1}^{K}\ell\sum_{m=1}^{\ell}\sum_{i\in\mathcal{H}}\frac{1}{|\mathcal{H}|}\xi_{13}\mathbb{E}\left[\left\|\hat{\mathbf{w}}_{t,m}-\mathbf{w}_{t,m}^{i}\right\|^{2}\right]+\eta^{2}\sum_{\ell=1}^{K}\ell\sum_{m=1}^{\ell}\xi_{14}+\eta^{2}\sum_{\ell=1}^{K}\ell\sum_{m=1}^{\ell}\xi_{15}+\eta^{2}\sum_{\ell=1}^{K}\ell\sum_{m=1}^{\ell}\xi_{16}\mathbb{E}\left[\left\|\nabla F_{\mathcal{H}}(\mathbf{w}_{t})\right\|^{2}\right]$$

$$\leq\eta^{2}\sum_{\ell=1}^{K}\frac{1}{|\mathcal{H}|}\sum_{i\in\mathcal{H}}K^{2}\xi_{13}\mathbb{E}\left[\left\|\hat{\mathbf{w}}_{t,\ell}-\mathbf{w}_{t,\ell}^{i}\right\|^{2}\right]+\eta^{2}\sum_{\ell=1}^{K}\ell\sum_{m=1}^{\ell}\xi_{14}+\eta^{2}\sum_{\ell=1}^{K}\ell\sum_{m=1}^{\ell}\xi_{15}+\eta^{2}\sum_{\ell=1}^{K}\ell\sum_{m=1}^{\ell}\xi_{16}\mathbb{E}\left[\left\|\nabla F_{\mathcal{H}}(\mathbf{w}_{t})\right\|^{2}\right].$$

We rewrite the equation as

$$\left(1-\eta^{2}K^{2}\xi_{13}\right)\sum_{\ell=1}^{K}\frac{1}{|\mathcal{H}|}\sum_{i\in\mathcal{H}}\mathbb{E}\left[\left\|\mathbf{w}_{t,\ell}^{i}-\hat{\mathbf{w}}_{t,\ell}\right\|^{2}\right]\leq\eta^{2}\sum_{\ell=1}^{K}\ell\sum_{m=1}^{\ell}\xi_{14}+\eta^{2}\sum_{\ell=1}^{K}\ell\sum_{m=1}^{\ell}\xi_{15}+\eta^{2}\sum_{\ell=1}^{K}\ell\sum_{m=1}^{\ell}\xi_{16}\mathbb{E}\left[\left\|\nabla F_{\mathcal{H}}(\mathbf{w}_{t})\right\|^{2}\right]$$

and choose the learning rate small enough so that $\left(1-\eta^{2}K^{2}\xi_{13}\right)\geq\frac{1}{2}$. Letting $\xi_{13}=6D^{2}+6\zeta^{2}+\frac{32d}{\nu}L^{2}+\frac{160\cdot32d}{\nu}L^{2}\eta^{2}\left(KD^{2}+\frac{1}{|\mathcal{H}|}\frac{d}{\nu}L^{2}\right)K^{2}$, we require that $\eta\leq\sqrt{6\frac{D^{2}}{K^{2}}+6\frac{\zeta^{2}}{K^{2}}+\frac{32d}{\nu K^{2}}L^{2}+\frac{160\cdot32d}{\nu}L^{2}\eta^{2}\left(KD^{2}+\frac{1}{|\mathcal{H}|}\frac{d}{\nu}L^{2}\right)}$. Since $\frac{160\cdot32d}{\nu}L^{2}\eta^{2}\left(KD^{2}+\frac{1}{|\mathcal{H}|}\frac{d}{\nu}L^{2}\right)\geq0$, it suffices to let $\eta\leq\sqrt{6\frac{D^{2}}{K^{2}}+6\frac{\zeta^{2}}{K^{2}}+\frac{32d}{\nu K^{2}}L^{2}}$. Hence, we obtain

$$\sum_{\ell=1}^{K}\frac{1}{|\mathcal{H}|}\sum_{i\in\mathcal{H}}\mathbb{E}\left[\left\|\mathbf{w}_{t,\ell}^{i}-\hat{\mathbf{w}}_{t,\ell}\right\|^{2}\right]\leq\xi_{2}'+\xi_{1}'\mathbb{E}\left[\left\|\nabla F_{\mathcal{H}}(\mathbf{w}_{t})\right\|^{2}\right],$$

where

$$\xi_{1}'\triangleq2\eta^{2}\sum_{\ell=1}^{K}\ell\sum_{m=1}^{\ell}\xi_{16}\leq2\eta^{2}K^{3}\xi_{16}\leq\xi_{1}\triangleq4\eta^{2}K^{3}\xi_{11},$$

$$\xi_{2}'\triangleq2\eta^{2}\sum_{\ell=1}^{K}\ell\sum_{m=1}^{\ell}\xi_{14}+2\eta^{2}\sum_{\ell=1}^{K}\ell\sum_{m=1}^{\ell}\xi_{15}\leq\xi_{2}\triangleq4\eta^{2}K^{3}\xi_{12}+12\eta^{2}K^{3}L.$$

□

C.2.2 Proof of Theorem 5.7 for $\mu=0$

Proof. For the proof of the projected gradient variance, we rely on the following intermediate lemma.

Lemma C.11. *The second moment of the projected gradient according to Definition 2.1 for $\mu=0$ is bounded as*

$$\mathbb{E}\left[\left\|\mathbf{z}g(\mathbf{w},\mathbf{z},\mu,\mathcal{D})\right\|^{2}\right]\leq d\left\|\nabla F(\mathbf{w},\mathcal{D})\right\|^{2}.$$

Proof of Lemma C.11.

$$\begin{aligned}\mathbb{E}\left[\left\|\mathbf{z}g(\mathbf{w},\mathbf{z},\mu,\mathcal{D})\right\|^{2}\right]&=\mathbb{E}\left[\left\|d\mathbf{z}\left\langle\nabla F(\mathbf{w},\mathcal{D}),\mathbf{z}\right\rangle\right\|^{2}\right]\\&=\mathbb{E}\left[\left\|d\mathbf{z}\mathbf{z}^{T}\nabla F(\mathbf{w},\mathcal{D})\right\|^{2}\right]\\&=\mathbb{E}\left[d^{2}\nabla F(\mathbf{w},\mathcal{D})^{T}\mathbf{z}\mathbf{z}^{T}\mathbf{z}\mathbf{z}^{T}\nabla F(\mathbf{w},\mathcal{D})\right]\end{aligned}$$

$$
\begin{aligned}
&= d^2 \nabla F(\mathbf{w}, \mathcal{D})^T \mathbb{E}\left[\mathbf{z}\mathbf{z}^T\right] \nabla F(\mathbf{w}, \mathcal{D}) \\
&= d \nabla F(\mathbf{w}, \mathcal{D})^T \nabla F(\mathbf{w}, \mathcal{D}) \\
&= d \left\|\nabla F(\mathbf{w}, \mathcal{D})\right\|^2,
\end{aligned}
$$

where the penultimate step is by (Gao et al., 2018, Lemma 7.3), which states that $\mathbb{E}\left[\mathbf{z}\mathbf{z}^T\right] = \frac{1}{d}\mathbf{I}$, for $\mathbf{I}$ being the identity matrix. □

Accordingly, the bound in Lemma C.9 for the zero-order estimate is an upper bound to the result of Lemma C.11 when choosing $\mu = 0$ in Lemma C.9. Further, we observe that Proposition C.2 still holds for $\mu = 0$, due to a non-zero bias in the gradient projection case. Since those are the only two intermediate results where the case of gradient projection differs from the zero-order estimate, the result established in Theorem 5.7 holds for the gradient projection case when choosing $\mu = 0$. □

C.3 Proof of Theorem 5.6

Proof. We assume for all that follows that the objective F exhibits a G-Lipschitz behavior. Similar to the proof of Theorem 5.7, we first state necessary intermediate lemmas, which we prove in the sequel. All lemmas hold under Assumptions 5.1 to 5.5 and a robust aggregator according to Definition 2.2.

Lemma C.12. *Let*

$$
\begin{aligned}
\xi_6 &\triangleq 5 \cdot 16\eta^2 K^2 \\
\xi_7 &\triangleq 5K\eta^2 \frac{\varphi^2 G^2 d}{|\mathcal{H}|\nu} + 5\eta^2 16 K^2 L\mu \\
\xi_8 &\triangleq 5 \cdot 16\eta^2 \left(KD^2\right)
\end{aligned}
$$

For a learning rate satisfying $\eta \leq \sqrt{\frac{1}{32L^2K^2}}$, we have the following upper bound on the averaged local model divergence:

$$
\mathbb{E}\left[\left\|\hat{\mathbf{w}}_{t,\ell} - \mathbf{w}_t\right\|^2\right] \leq \xi_6 \mathbb{E}\left[\left\|\nabla F_{\mathcal{H}}(\mathbf{w}_t)\right\|^2\right] + \xi_7 + \xi_8 \sum_{\ell'=1}^{\ell-1} \frac{1}{|\mathcal{H}|} \sum_{i \in \mathcal{H}} \mathbb{E}\left[\left\|\mathbf{w}_{t,\ell'}^i - \hat{\mathbf{w}}_{t,\ell'}\right\|^2\right]
$$

Lemma C.13. *The following holds for the gradient estimate variance*

$$
\mathbb{E}\left[\left\|\mathbf{Z}_{t,\ell}\mathbf{g}_i(\mathbf{w}_{t,\ell}^i, \mathbf{Z}_{t,\ell}) - \nabla F_i^{\mu}(\mathbf{w}_{t,\ell}^i)\right\|^2\right] \leq \frac{\varphi^2 G^2 d}{\nu}
$$

Lemma C.14. *Let*

$$
\begin{aligned}
\xi_{13} &\triangleq 6D^2 + 6\zeta^2 \\
\xi_{14} &\triangleq 2\frac{\varphi^2 G^2 d}{\nu} \\
\xi_{15} &\triangleq 6L
\end{aligned}
$$

Then

$$
\begin{aligned}
&\sum_{m=1}^{\ell} \frac{1}{|\mathcal{H}|} \sum_{i \in \mathcal{H}} \mathbb{E}\left[\left\|\mathbf{Z}_{t,m}\mathbf{g}_i(\mathbf{w}_{t,\ell}^i, \mathbf{Z}_{t,\ell}) - \frac{1}{|\mathcal{H}|} \sum_{j \in \mathcal{H}} \mathbf{Z}_{t,m}\mathbf{g}_j(\mathbf{w}_{t,m}^j, \mathbf{Z}_{t,m})\right\|^2\right] \\
&\leq \sum_{m=1}^{\ell} \frac{1}{|\mathcal{H}|} \sum_{i \in \mathcal{H}} \xi_{13} \mathbb{E}\left[\left\|\hat{\mathbf{w}}_{t,m} - \mathbf{w}_{t,m}^i\right\|^2\right] + \sum_{m=1}^{\ell} (\xi_{14} + \xi_{15})
\end{aligned}
$$

Lemma C.15. *Let*

$$\xi_2 \triangleq 4\eta^2 K^3 \frac{\varphi^2 G^2 d}{\nu} + 12\eta^2 K^3 L.$$

For a learning rate that satisfies $\eta \leq \sqrt{\frac{1}{12K^2(D^2+\zeta^2)}}$*, we have*

$$\sum_{\ell=1}^{K} \frac{1}{|\mathcal{H}|} \sum_{i\in\mathcal{H}} \mathbb{E}\left[\left\|\mathbf{w}_{t,\ell}^i - \hat{\mathbf{w}}_{t,\ell}\right\|^2\right] \leq \xi_2$$

To proof Theorem 5.6, we first follow the same lines as in the proof of Theorem 5.7, arriving at Equation (2). We continue with bounding the individual terms.

$$\begin{aligned}
&\mathbb{E}\left[\left\|\sum_{\ell=1}^{K}\left(\sum_{i\in\mathcal{H}}\frac{1}{|\mathcal{H}|}\nabla F_i(\mathbf{w}_{t,\ell}^i) - \mathbf{Z}_{t,\ell}\bar{\mathbf{g}}_{\mathcal{H}}(\mathbf{w}_{t,\ell}^i, \mathbf{Z}_{t,\ell})\right)\right\|^2\right] \\
&\leq 2\mathbb{E}\left[\left\|\sum_{\ell=1}^{K}\sum_{i\in\mathcal{H}}\frac{1}{|\mathcal{H}|}\left(\nabla F_i(\mathbf{w}_{t,\ell}^i) - \nabla F_i^{\mu}(\mathbf{w}_{t,\ell}^i)\right)\right\|^2\right] + 2\mathbb{E}\left[\left\|\sum_{\ell=1}^{K}\sum_{i\in\mathcal{H}}\frac{1}{|\mathcal{H}|}\left(\nabla F_i^{\mu}(\mathbf{w}_{t,\ell}^i) - \mathbf{Z}_{t,\ell}\mathbf{g}_i(\mathbf{w}_{t,\ell}^i, \mathbf{Z}_{t,\ell})\right)\right\|^2\right] \\
&\overset{(a)}{\leq} 2\frac{K}{|\mathcal{H}|}\sum_{i\in\mathcal{H}}\sum_{\ell=1}^{K}\left\|\left(\nabla F_i(\mathbf{w}_{t,\ell}^i) - \nabla F_i^{\mu}(\mathbf{w}_{t,\ell}^i)\right)\right\|^2 + 2\frac{1}{|\mathcal{H}|^2}\sum_{i\in\mathcal{H}}\mathbb{E}\left[\left\|\sum_{\ell=1}^{K}\left(\nabla F_i^{\mu}(\mathbf{w}_{t,\ell}^i) - \mathbf{Z}_{t,\ell}\mathbf{g}_i(\mathbf{w}_{t,\ell}^i, \mathbf{Z}_{t,\ell})\right)\right\|^2\right] \\
&\overset{(b)}{\leq} 2K^2L\mu + 2\frac{1}{|\mathcal{H}|^2}\sum_{i\in\mathcal{H}}\sum_{\ell=1}^{K}\mathbb{E}\left[\left\|\nabla F_i^{\mu}(\mathbf{w}_{t,\ell}^i) - \mathbf{Z}_{t,\ell}\mathbf{g}_i(\mathbf{w}_{t,\ell}^i, \mathbf{Z}_{t,\ell})\right\|^2\right] \\
&\overset{(c)}{\leq} 2K^2L\mu + 2\frac{1}{|\mathcal{H}|^2}\sum_{i\in\mathcal{H}}\sum_{\ell=1}^{K}\frac{\varphi^2 G^2 d}{\nu} \\
&\overset{(c)}{\leq} 2K^2L\mu + 2\frac{1}{|\mathcal{H}|}K\frac{\varphi^2 G^2 d}{\nu}
\end{aligned} \tag{16}$$

where (a) is due to the independence of $\sum_{\ell=1}^{K}\left(\nabla F_i^{\mu}(\mathbf{w}_{t,\ell}^i) - \mathbf{Z}_{t,\ell}\mathbf{g}_i(\mathbf{w}_{t,\ell}^i, \mathbf{Z}_{t,\ell})\right)$ and $\sum_{\ell=1}^{K}\left(\nabla F_j^{\mu}(\mathbf{w}_{t,\ell}^j) - \mathbf{Z}_{t,\ell}\mathbf{g}_j(\mathbf{w}_{t,\ell}^j, \mathbf{Z}_{t,\ell})\right)$ for $i \neq j$. (b) follows from Proposition C.1 and (Wang et al., 2021, Lemma 2). (c) is by the application of Lemma C.13.

We continue with bounding the robustness term using as the main ingredient a double-sided application of an extended Johnson-Lindenstrauss Lemma as stated in Lemma C.8 and the application of Lemma C.14. We have

$$\begin{aligned}
&\mathbb{E}\left[\left\|\sum_{\ell=1}^{K}\mathbf{Z}_{t,\ell}\left(R_{t,\ell} - \bar{\mathbf{g}}_{\mathcal{H}}(\mathbf{w}_{t,\ell}^i, \mathbf{Z}_{t,\ell})\right)\right\|^2\right] \\
&\leq K\sum_{\ell=1}^{K}\left\|\mathbf{Z}_{t,\ell}\left(R_{t,\ell} - \bar{\mathbf{g}}_{\mathcal{H}}(\mathbf{w}_{t,\ell}^i, \mathbf{Z}_{t,\ell})\right)\right\|^2 \\
&\overset{(a)}{\leq} K\sum_{\ell=1}^{K}(1+\epsilon)\frac{\kappa}{|\mathcal{H}|}\sum_{i\in\mathcal{H}}\mathbb{E}\left[\left\|\mathbf{g}_i(\mathbf{w}_{t,\ell}^i, \mathbf{Z}_{t,\ell}) - \bar{\mathbf{g}}_{\mathcal{H}}(\mathbf{w}_{t,\ell}^i, \mathbf{Z}_{t,\ell})\right\|^2\right] \\
&\overset{(b)}{\leq} K\sum_{\ell=1}^{K}\frac{(1+\epsilon)}{(1-\epsilon)}\frac{\kappa}{|\mathcal{H}|}\sum_{i\in\mathcal{H}}\mathbb{E}\left[\left\|\mathbf{Z}_{t,\ell}(\mathbf{g}_i(\mathbf{w}_{t,\ell}^i, \mathbf{Z}_{t,\ell}) - \bar{\mathbf{g}}_{\mathcal{H}}(\mathbf{w}_{t,\ell}^i, \mathbf{Z}_{t,\ell}))\right\|^2\right] \\
&\overset{(c)}{\leq} K\frac{(1+\epsilon)}{(1-\epsilon)}\kappa\left(\sum_{\ell=1}^{K}\frac{1}{|\mathcal{H}|}\sum_{i\in\mathcal{H}}\xi_{13}\mathbb{E}\left[\left\|\hat{\mathbf{w}}_{t,\ell} - \mathbf{w}_{t,\ell}^i\right\|^2\right] + \sum_{\ell=1}^{K}(\xi_{14}+\xi_{15})\right)
\end{aligned} \tag{17}$$

where (a) and (b) are by the application of Lemma C.8 for in total $|\mathcal{H}|+1$ projections. By a union bound over Lemma C.8, the distance preservation holds with probability $1-\delta$ for $\epsilon \geq \sqrt{\frac{64}{\nu}\log(\frac{2(|\mathcal{H}|-1)}{\delta})}$. We choose the smallest possible ϵ. This must hold for each iteration, so the distance preservation holds w.p. at least $1-K\delta$ for all local epochs. (c) is by Lemma C.6.

To bound the local model divergence from the global model, by Lemma C.4, we have

$$\begin{aligned}\sum_{\ell=1}^{K}\mathbb{E}\left[\left\|\hat{\mathbf{w}}_{t,\ell}-\mathbf{w}_t\right\|^2\right] &\leq \sum_{\ell=1}^{K}\xi_6\mathbb{E}\left[\left\|\nabla F_{\mathcal{H}}(\mathbf{w}_t)\right\|^2\right]+\sum_{\ell=1}^{K}\xi_7+\sum_{\ell=1}^{K}\xi_8\sum_{\ell'=1}^{\ell-1}\frac{1}{|\mathcal{H}|}\sum_{i\in\mathcal{H}}\mathbb{E}\left[\left\|\mathbf{w}_{t,\ell'}^i-\hat{\mathbf{w}}_{t,\ell'}\right\|^2\right]\\ &\leq \sum_{\ell=1}^{K}\xi_6\mathbb{E}\left[\left\|\nabla F_{\mathcal{H}}(\mathbf{w}_t)\right\|^2\right]+\sum_{\ell=1}^{K}\xi_7+\sum_{\ell=1}^{K}\frac{\xi_8K(K-1)}{|\mathcal{H}|}\sum_{i\in\mathcal{H}}\mathbb{E}\left[\left\|\mathbf{w}_{t,\ell}^i-\hat{\mathbf{w}}_{t,\ell}\right\|^2\right], \qquad (18)\end{aligned}$$

where the latter follows since $\sum_{\ell=1}^{K}\sum_{\ell'=1}^{\ell-1}x_{\ell'}\leq\sum_{\ell=1}^{K}\sum_{\ell'=1}^{\ell}x_{\ell'}\leq\frac{K(K-1)}{2}\sum_{\ell}x_{\ell}$.

Plugging (3), (17), and (18) into (2), we obtain

$$\begin{aligned}&\mathbb{E}\left[\left\|K\nabla F_{\mathcal{H}}(\mathbf{w}_t)-\sum_{\ell=1}^{K}\mathbf{Z}_{t,\ell}R_{t,\ell}\right\|^2\right]\\ &\leq 4KL^2\left(\sum_{\ell=1}^{K}\xi_6\mathbb{E}\left[\left\|\nabla F_{\mathcal{H}}(\mathbf{w}_t)\right\|^2\right]+\sum_{\ell=1}^{K}\xi_7+\sum_{\ell=1}^{K}\frac{\xi_8K(K-1)}{|\mathcal{H}|}\sum_{i\in\mathcal{H}}\mathbb{E}\left[\left\|\mathbf{w}_{t,\ell}^i-\hat{\mathbf{w}}_{t,\ell}\right\|^2\right]\right)\\ &+4\left(K\frac{(1+\epsilon)}{(1-\epsilon)}\kappa\left(\sum_{\ell=1}^{K}\frac{1}{|\mathcal{H}|}\sum_{i\in\mathcal{H}}\xi_{13}\mathbb{E}\left[\left\|\hat{\mathbf{w}}_{t,\ell}-\mathbf{w}_{t,\ell}^i\right\|^2\right]+\sum_{\ell=1}^{K}(\xi_{14}+\xi_{15})\right)\right)\\ &+4K\sum_{\ell=1}^{K}\frac{D^2}{|\mathcal{H}|}\sum_{i\in\mathcal{H}}\mathbb{E}\left[\left\|\mathbf{w}_{t,\ell}^i-\hat{\mathbf{w}}_{t,\ell}\right\|^2\right]+4\left(2K^2L\mu+2\frac{1}{|\mathcal{H}|}K\frac{\varphi^2G^2d}{\nu}\right)\\ &\leq \xi_3'\mathbb{E}\left[\left\|\nabla F_{\mathcal{H}}(\mathbf{w}_t)\right\|^2\right]+\xi_4'\sum_{\ell=1}^{K}\frac{1}{|\mathcal{H}|}\sum_{i\in\mathcal{H}}\mathbb{E}\left[\left\|\hat{\mathbf{w}}_{t,\ell}-\mathbf{w}_{t,\ell}^i\right\|^2\right]+\xi_5'\\ &\leq \xi_3'\mathbb{E}\left[\left\|\nabla F_{\mathcal{H}}(\mathbf{w}_t)\right\|^2\right]+\xi_4'\xi_2+\xi_5', \qquad (19)\end{aligned}$$

where

$$\begin{aligned}\xi_3' &\triangleq 4KL^2\sum_{\ell=1}^{K}\xi_6\leq\xi_3\triangleq 4K^4L^2 5\cdot 16\eta^2\\ \xi_4' &\triangleq 4KL^2\xi_8K(K-1)+4K\frac{(1+\epsilon)}{(1-\epsilon)}\kappa\xi_{13}+4KD^2\\ &\leq 4KL^2 5\cdot 16\eta^2\left(KD^2\right)K(K-1)+4K\kappa\frac{(1+\epsilon)}{(1-\epsilon)}6(D^2+\zeta^2)+4KD^2\\ &\leq \xi_4\triangleq 20\cdot 16K^4L^2\eta^2D^2+4K\frac{(1+\epsilon)}{(1-\epsilon)}\kappa 6(D^2+\zeta^2)+4KD^2\\ \xi_5' &\triangleq 4KL^2\sum_{\ell=1}^{K}\xi_7+4K\frac{(1+\epsilon)}{(1-\epsilon)}\kappa\sum_{\ell=1}^{K}(\xi_{14}+\xi_{15})+4\cdot 2K^2L\mu+4\cdot 2\frac{1}{|\mathcal{H}|^2}K\frac{\varphi^2G^2d}{\nu}\\ &\leq 4KL^2\sum_{\ell=1}^{K}\left(5K\eta^2\frac{\varphi^2G^2d}{|\mathcal{H}|\nu}+5\eta^2 16K^2L\mu\right)+4K\frac{(1+\epsilon)}{(1-\epsilon)}\kappa\sum_{\ell=1}^{K}(2\frac{\varphi^2G^2d}{\nu}+6L)\\ &+4\cdot 2K^2L\mu+4\cdot 2\frac{1}{|\mathcal{H}|^2}K\frac{\varphi^2G^2d}{\nu}\\ &\leq 4K^2L^2\left(5K\eta^2\frac{\varphi^2G^2d}{|\mathcal{H}|\nu}+5\eta^2 16K^2L\mu\right)+4K^2\frac{(1+\epsilon)}{(1-\epsilon)}\kappa(2\frac{\varphi^2G^2d}{\nu}+6L)\end{aligned}$$

$$
\begin{aligned}
&+ 4 \cdot 2K^2 L\mu + 4 \cdot 2 \frac{1}{|\mathcal{H}|^2} K \frac{\varphi^2 G^2 d}{\nu} \\
&\leq \xi_5 \triangleq 4K \frac{\varphi^2 G^2 d}{\nu} \left(5K^2 L^2 \eta^2 \frac{1}{|\mathcal{H}|} + 2K \frac{(1+\epsilon)}{(1-\epsilon)} \kappa + \frac{8}{|\mathcal{H}|^2} \right) + 4K^2 L \left(5\eta^2 16 K^2 L^2 \mu + 6\kappa \frac{(1+\epsilon)}{(1-\epsilon)} + 2\mu \right)
\end{aligned}
$$

By taking the expectation over (1) and replacing $\mathbb{E}\left[\left\| K \nabla F_{\mathcal{H}}(\mathbf{w}_t) - \sum_{\ell=1}^{K} \mathbf{Z}_{t,\ell} R_{t,\ell} \right\|^2 \right]$ by (6), we can write

$$
\begin{aligned}
&\mathbb{E}\left[F_{\mathcal{H}}(\mathbf{w}_{t+1}) \right] - \mathbb{E}\left[F_{\mathcal{H}}(\mathbf{w}_t) \right] \\
&\leq -\eta/(2) K \mathbb{E}\left[\left\| \nabla F_{\mathcal{H}}(\mathbf{w}_t) \right\|^2 \right] + \left(\frac{\eta^2 L}{2} - \frac{\eta}{2K} \right) \mathbb{E}\left[\left\| \sum_{\ell=1}^{K} R_{t,\ell} \mathbf{Z}_{t,\ell} \right\|^2 \right] \\
&+ \eta/(2K) \mathbb{E}\left[\left\| K \nabla F_{\mathcal{H}}(\mathbf{w}_t) - \sum_{\ell=1}^{K} R_{t,\ell} \mathbf{Z}_{t,\ell} \right\|^2 \right] \\
&\overset{(a)}{\leq} -\eta/(2) K \mathbb{E}\left[\left\| \nabla F_{\mathcal{H}}(\mathbf{w}_t) \right\|^2 \right] + \eta/(2K) \left(\xi_3 \mathbb{E}\left[\left\| \nabla F_{\mathcal{H}}(\mathbf{w}_t) \right\|^2 \right] + \xi_4 \xi_2 + \xi_5 \right) \\
&\overset{(b)}{\leq} -\frac{\eta K}{4} \mathbb{E}\left[\left\| \nabla F_{\mathcal{H}}(\mathbf{w}_t) \right\|^2 \right] + \eta/(2K) \left(\xi_4 \xi_2 + \xi_5 \right),
\end{aligned}
$$

where (a) holds when $\eta \leq \frac{1}{KL}$ and (b) assumes that $\eta/(2K)\xi_3 = \eta/(2K) 4K^2 L^2 5 \cdot 16 \eta^2 K^2 \leq \frac{\eta K}{4}$. The learning rate must hence satisfy $\eta^2 \leq \frac{1}{8 \cdot K^2 L^2 5 \cdot 16}$, and consequently $\eta \leq \frac{1}{26KL}$.

Reordering and telescoping over t, we obtain

$$
\frac{1}{T} \sum_{t=1}^{T} \mathbb{E}\left[\left\| \nabla F_{\mathcal{H}}(\mathbf{w}_t) \right\|^2 \right] \leq \frac{4(\mathbb{E}\left[F_{\mathcal{H}}(\mathbf{w}_1) \right] - \mathbb{E}\left[F_{\mathcal{H}}(\mathbf{w}_{T+1}) \right])}{T\eta K} + \eta/(2K) \left(\xi_4 \xi_2 + \xi_5 \right)
$$

with probability $1 - \delta KT$ by a union bound argument over all global iterations T. Noting that $F_{\mathcal{H}}(\mathbf{w}_{T+1}) \geq F_{\mathcal{H}}^{\star}$ by definition concludes the proof.

Since the learning rates must satisfy $\eta \leq \frac{1}{6KL} \leq \sqrt{\frac{1}{32 L^2 K^2}}$, $\eta \leq \frac{1}{4K\sqrt{D^2+\zeta^2}} \leq \sqrt{\frac{1}{12K^2(D^2+\zeta^2)}}$ and $\eta \leq \frac{1}{26KL}$ for all lemmas to hold, and hence $\eta \leq \min\left\{ \frac{1}{26KL}, \frac{1}{4K\sqrt{D^2+\zeta^2}} \right\}$, we have

$$
\begin{aligned}
&\eta/(2K) \left(\xi_4 \xi_2 + \xi_5 \right) \\
&= \frac{\eta}{2K} \left(4\eta^2 K^3 \frac{\varphi^2 G^2 d}{\nu} + 12\eta^2 K^3 L \right) \left(20 \cdot 16 K^4 L^2 \eta^2 D^2 + 4K \frac{(1+\epsilon)}{(1-\epsilon)} \kappa 6 (D^2 + \zeta^2) + 4KD^2 \right) \\
&+ 4K \frac{\varphi^2 G^2 d}{\nu} \frac{\eta}{2K} \left(5K^2 L^2 \eta^2 \frac{1}{|\mathcal{H}|} + 2K \frac{(1+\epsilon)}{(1-\epsilon)} \kappa + \frac{8}{|\mathcal{H}|^2} \right) + 4K^2 L \frac{\eta}{2K} \left(5\eta^2 16 K^2 L^2 \mu + 6\kappa \frac{(1+\epsilon)}{(1-\epsilon)} + 2\mu \right) \\
&= 2\eta^3 K^2 \left(\frac{\varphi^2 G^2 d}{\nu} + 3L \right) \left(20 \cdot 16 K^4 L^2 \eta^2 D^2 + 4K \frac{(1+\epsilon)}{(1-\epsilon)} \kappa 6 (D^2 + \zeta^2) + 4KD^2 \right) \\
&+ 2 \frac{\varphi^2 G^2 d}{\nu} \eta \left(5K^2 L^2 \eta^2 \frac{1}{|\mathcal{H}|} + 2K \frac{(1+\epsilon)}{(1-\epsilon)} \kappa + \frac{8}{|\mathcal{H}|^2} \right) + 4K^2 L \frac{\eta}{2K} \left(5\eta^2 16 K^2 L^2 \mu + 6\kappa \frac{(1+\epsilon)}{(1-\epsilon)} + 2\mu \right) \\
&= 2\eta K \left(\frac{\varphi^2 G^2 d}{L^2 \nu} + 3 \frac{1}{L} \right) \left(13 K D^2 + 4 \frac{(1+\epsilon)}{(1-\epsilon)} \kappa 6 (D^2 + \zeta^2) + 4D^2 \right) \\
&+ 2\eta \frac{\varphi^2 G^2 d}{\nu} \left(\frac{1}{5|\mathcal{H}|} + 2K \frac{(1+\epsilon)}{(1-\epsilon)} \kappa + \frac{8}{|\mathcal{H}|^2} \right) + 2KL\eta \left(6\kappa \frac{(1+\epsilon)}{(1-\epsilon)} + 6\mu \right).
\end{aligned}
$$

We let $\Delta \triangleq \delta KT$, and obtain $\epsilon \geq \sqrt{\frac{64}{\nu}\log(\frac{2(|\mathcal{H}|-1)}{\delta})} = \sqrt{\frac{64}{\nu}\log(\frac{2(|\mathcal{H}|-1)TK}{\Delta})}$. Since it is required to satisfy $\epsilon < 1$, the proof holds for $\nu \geq 64\log(\frac{2(|\mathcal{H}|-1)TK}{\Delta})$. This concludes the proof.

□

Proof of Lemma C.13. For the proof of the zero-order approximated gradient variance, we rely on the following intermediate lemma.

Lemma C.16 (Lemma 5.3, (Tang et al., 2020)). *Let F be G-Lipschitz. Then for any $\mathbf{w} \in \mathbb{R}^d$, $\mathbf{z}$ and $\mu > 0$, we have for a numerical constant $\varphi > 0$ according to (Tang et al., 2020) that*

$$\mathbb{E}\left[\left\|\mathbf{z}^T g(\mathbf{w}, \mathbf{z}, \mu, \mathcal{D})\right\|^2\right] \leq \varphi^2 G^2 d.$$

We bound the zero-order approximated gradient variance as follows: Since $\mathbb{E}\left[\|Z - \mathbb{E}[Z]\|^2\right] \leq \mathbb{E}\left[\|Z\|^2\right]$, and $\mathbb{E}\left[\mathbf{z}_{t,\ell}^r g_i(\mathbf{w}_{t,\ell}^i, \mathbf{z}_{t,\ell}^r)\right] = \nabla F_i^\mu(\mathbf{w}_{t,\ell}^i)$, we have

$$\begin{aligned}
&\mathbb{E}\left[\left\|\mathbf{Z}_{t,\ell}\mathbf{g}_i(\mathbf{w}_{t,\ell}^i, \mathbf{Z}_{t,\ell}) - \nabla F_i^\mu(\mathbf{w}_{t,\ell}^i)\right\|^2\right] = \mathbb{E}\left[\left\|\frac{1}{\nu}\sum_{r=1}^{\nu}\mathbf{z}_{t,\ell}^r g_i(\mathbf{w}_{t,\ell}^i, \mathbf{z}_{t,\ell}^r) - \nabla F_i^\mu(\mathbf{w}_{t,\ell}^i)\right\|^2\right] \\
&\overset{(a)}{=} \frac{1}{\nu^2}\sum_{r=1}^{\nu}\mathbb{E}\left[\left\|\mathbf{z}_{t,\ell}^r g_i(\mathbf{w}_{t,\ell}^i, \mathbf{z}_{t,\ell}^r) - \nabla F_i^\mu(\mathbf{w}_{t,\ell}^i)\right\|^2\right] \\
&\leq \frac{1}{\nu^2}\sum_{r=1}^{\nu}\mathbb{E}\left[\left\|\mathbf{z}_{t,\ell}^r g_i(\mathbf{w}_{t,\ell}^i, \mathbf{z}_{t,\ell}^r)\right\|^2\right] \overset{(b)}{\leq} \frac{\varphi^2 G^2 d}{\nu}
\end{aligned}$$

where (a) is due to the independence of $\mathbf{z}_{t,\ell}^r g_i(\mathbf{w}_{t,\ell}^i, \mathbf{z}_{t,\ell}^r)$ and $\mathbf{z}_{t,\ell}^{r'} g_i(\mathbf{w}_{t,\ell}^i, \mathbf{z}_{t,\ell}^{r'})$ for $r \neq r'$. (b) is by Lemma C.16. □

Proof of Lemma C.12. By definition, $\mathbb{E}\left[\|\hat{\mathbf{w}}_{t,1} - \mathbf{w}_t\|^2\right] = 0$. From (13), we have for $\ell \in \{2, \cdots, K\}$,

$$\begin{aligned}
&\mathbb{E}\left[\|\hat{\mathbf{w}}_{t,\ell} - \mathbf{w}_t\|^2\right] \\
&\leq (1 + \frac{1}{\tau})\mathbb{E}\left[\|\hat{\mathbf{w}}_{t,\ell-1} - \mathbf{w}_t\|^2\right] + 8(1+\tau)\eta^2\mathbb{E}\left[\left\|\frac{1}{|\mathcal{H}|}\sum_{i\in\mathcal{H}}\nabla F_i^\mu(\mathbf{w}_{t,\ell-1}^i) - \frac{1}{|\mathcal{H}|}\sum_{i\in\mathcal{H}}\nabla F_i(\mathbf{w}_{t,\ell-1}^i)\right\|^2\right] \\
&+ 8(1+\tau)\eta^2\mathbb{E}\left[\|\nabla F_{\mathcal{H}}(\mathbf{w}_t)\|^2\right] \\
&+ 8(1+\tau)\eta^2\mathbb{E}\left[\left\|\frac{1}{|\mathcal{H}|}\sum_{i\in\mathcal{H}}\nabla F_i(\mathbf{w}_{t,\ell-1}^i) - \nabla F_{\mathcal{H}}(\hat{\mathbf{w}}_{t,\ell-1})\right\|^2\right] + 8(1+\tau)\eta^2\mathbb{E}\left[\|\nabla F_{\mathcal{H}}(\hat{\mathbf{w}}_{t,\ell-1}) - \nabla F_{\mathcal{H}}(\mathbf{w}_t)\|^2\right] \\
&+ 2\eta^2\frac{1}{|\mathcal{H}|^2}\sum_{i\in\mathcal{H}}\mathbb{E}\left[\left\|\mathbf{Z}_{t,\ell-1}\mathbf{g}_i(\mathbf{w}_{t,\ell-1}^i, \mathbf{Z}_{t,\ell-1}) - \nabla F_i^\mu(\mathbf{w}_{t,\ell-1}^i)\right\|^2\right] \\
&\overset{(a)}{\leq} (1 + \frac{1}{\tau})\mathbb{E}\left[\|\hat{\mathbf{w}}_{t,\ell-1} - \mathbf{w}_t\|^2\right] + 8(1+\tau)\eta^2 L\mu + 8(1+\tau)\eta^2\mathbb{E}\left[\|\nabla F_{\mathcal{H}}(\mathbf{w}_t)\|^2\right] \\
&+ 8(1+\tau)\eta^2\frac{D^2}{|\mathcal{H}|}\sum_{i\in\mathcal{H}}\mathbb{E}\left[\left\|\mathbf{w}_{t,\ell-1}^i - \hat{\mathbf{w}}_{t,\ell-1}\right\|^2\right] + 8(1+\tau)\eta^2 L^2\mathbb{E}\left[\|\hat{\mathbf{w}}_{t,\ell-1} - \mathbf{w}_t\|^2\right] \\
&+ 2\eta^2\frac{1}{|\mathcal{H}|^2}\sum_{i\in\mathcal{H}}\frac{\varphi^2 G^2 d}{\nu}
\end{aligned}$$

$$
\begin{aligned}
&= \left((1+\frac{1}{\tau}) + 8(1+\tau)\eta^2 L^2\right) \mathbb{E}\left[\|\hat{\mathbf{w}}_{t,\ell-1} - \mathbf{w}_t\|^2\right] + 8(1+\tau)\eta^2 L\mu \\
&+ \left(8(1+\tau)\eta^2\right) \mathbb{E}\left[\|\nabla F_{\mathcal{H}}(\mathbf{w}_t)\|^2\right] \\
&+ \left(8(1+\tau)\eta^2 \frac{D^2}{|\mathcal{H}|}\right) \sum_{i\in\mathcal{H}} \mathbb{E}\left[\left\|\mathbf{w}^i_{t,\ell-1} - \hat{\mathbf{w}}_{t,\ell-1}\right\|^2\right] + 2\eta^2 \frac{1}{|\mathcal{H}|}\frac{\varphi^2 G^2 d}{\nu},
\end{aligned}
$$

where (a) follows from Assumption 5.1, Assumption 5.4 and an intermediate step in the proof of Lemma C.5.

To ensure the bound holds uniformly for all $\ell \in [K]$, we now choose $\tau = 2K-1$ and the learning rate small enough so that $\left((1+\frac{1}{\tau}) + 8(1+\tau)\eta^2 L^2\right) \leq 1 + \frac{1}{K-1}$, i.e., that $8(1+\tau)\eta^2 L^2 \leq \frac{1}{2K}$. Hence, the learning rate is required to satisfy $\eta \leq \sqrt{\frac{1}{32L^2K^2}}$ With this choice of the learning rate, we have

$$
\begin{aligned}
&\mathbb{E}\left[\|\hat{\mathbf{w}}_{t,\ell} - \mathbf{w}_t\|^2\right] \\
&\leq (1+\frac{1}{K-1})\mathbb{E}\left[\|\hat{\mathbf{w}}_{t,\ell-1} - \mathbf{w}_t\|^2\right] + 8(1+\tau)\eta^2 L\mu + \left(8(1+\tau)\eta^2\right) \mathbb{E}\left[\|\nabla F_{\mathcal{H}}(\mathbf{w}_t)\|^2\right] \\
&+ \left(8(1+\tau)\eta^2 \frac{D^2}{|\mathcal{H}|}\right) \sum_{i\in\mathcal{H}} \mathbb{E}\left[\left\|\mathbf{w}^i_{t,\ell-1} - \hat{\mathbf{w}}_{t,\ell-1}\right\|^2\right] + 2\eta^2 \frac{1}{|\mathcal{H}|}\frac{\varphi^2 G^2 d}{\nu} \\
&\overset{(d)}{\leq} \xi'_6(\ell)\mathbb{E}\left[\|\nabla F_{\mathcal{H}}(\mathbf{w}_t)\|^2\right] + \xi'_7(\ell) + \xi'_8 \sum_{\ell'=1}^{\ell-1}\sum_{i\in\mathcal{H}} \mathbb{E}\left[\left\|\mathbf{w}^i_{t,\ell'} - \hat{\mathbf{w}}_{t,\ell'}\right\|^2\right],
\end{aligned}
$$

where (e) is by the recursive application of (d) and the fact that $(1+\frac{1}{K})^\ell \leq (1+\frac{1}{\ell})^\ell \leq e \leq 5$ for all $\ell \in [K]$. This concludes the proof. The constants are given as

$$
\begin{aligned}
\xi'_6(\ell) &\triangleq 5(\ell-1)\left(8(1+\tau)\eta^2\right) \leq \xi_6 \triangleq 5\cdot 16\eta^2 K^2 \\
\xi'_7(\ell) &\triangleq 5(\ell-1)\left(2\eta^2\frac{1}{|\mathcal{H}|}\frac{\varphi^2 G^2 d}{\nu} + 8(1+\tau)\eta^2 L\mu\right) \\
&\leq \xi_7 \triangleq 5K\eta^2 \frac{\varphi^2 G^2 d}{|\mathcal{H}|\nu} + 5\eta^2 16K^2 L\mu \\
\xi'_8 &\triangleq 5\left(8(1+\tau)\eta^2 D^2\right) \leq \xi_8 \triangleq 5\cdot 16\eta^2\left(KD^2\right)
\end{aligned}
$$

□

Proof of Lemma C.14. Using (15) , we obtain

$$
\begin{aligned}
&\sum_{m=1}^{\ell} \frac{1}{|\mathcal{H}|}\sum_{i\in\mathcal{H}} \mathbb{E}\left[\left\|\mathbf{Z}_{t,m}\mathbf{g}_i(\mathbf{w}^i_{t,\ell}, \mathbf{Z}_{t,\ell}) - \frac{1}{|\mathcal{H}|}\sum_{j\in\mathcal{H}} \mathbf{Z}_{t,m}\mathbf{g}_j(\mathbf{w}^j_{t,m}, \mathbf{Z}_{t,m})\right\|^2\right] \\
&\overset{(a)}{\leq} 2\sum_{m=1}^{\ell} \mathbb{E}\left[\left(3\frac{D^2}{|\mathcal{H}|} + 3\frac{\zeta^2}{|\mathcal{H}|}\right)\sum_{i\in\mathcal{H}}\left\|\hat{\mathbf{w}}_{t,m} - \mathbf{w}^i_{t,m}\right\|^2 + 3L\right] \\
&+ 2\sum_{m=1}^{\ell}\frac{1}{|\mathcal{H}|}\sum_{i\in\mathcal{H}}\left(\frac{\varphi^2 G^2 d}{\nu}\right) \\
&\leq \sum_{m=1}^{\ell}\frac{1}{|\mathcal{H}|}\sum_{i\in\mathcal{H}} \xi_{13}\mathbb{E}\left[\left\|\hat{\mathbf{w}}_{t,m} - \mathbf{w}^i_{t,m}\right\|^2\right] + \sum_{m=1}^{\ell}(\xi_{14}+\xi_{15})
\end{aligned}
$$

where (a) is due to Lemma C.13. Thereby,

$$
\xi_{13} \triangleq 6D^2 + 6\zeta^2
$$

$$\xi_{14} \triangleq 2\frac{\varphi^2 G^2 d}{\nu}$$
$$\xi_{15} \triangleq 6L$$

□

Proof of Lemma C.15. From Lemma C.14, we have

$$\begin{aligned}
&\sum_{\ell=1}^{K}\frac{1}{|\mathcal{H}|}\sum_{i\in\mathcal{H}}\mathbb{E}\left[\left\|\mathbf{w}_{t,\ell}^{i}-\hat{\mathbf{w}}_{t,\ell}\right\|^2\right]\\
&=\sum_{\ell=1}^{K}\frac{1}{|\mathcal{H}|}\sum_{i\in\mathcal{H}}\mathbb{E}\left[\left\|\eta\sum_{m=1}^{\ell}\left(\mathbf{Z}_{t,m}\mathbf{g}_i(\mathbf{w}_{t,\ell}^{i},\mathbf{Z}_{t,\ell})-\frac{1}{|\mathcal{H}|}\sum_{j\in\mathcal{H}}\mathbf{Z}_{t,m}\mathbf{g}_j(\mathbf{w}_{t,m}^{j},\mathbf{Z}_{t,m})\right)\right\|^2\right]\\
&\leq\eta^2\sum_{\ell=1}^{K}\ell\sum_{m=1}^{\ell}\frac{1}{|\mathcal{H}|}\sum_{i\in\mathcal{H}}\mathbb{E}\left[\left\|\mathbf{Z}_{t,m}\mathbf{g}_i(\mathbf{w}_{t,\ell}^{i},\mathbf{Z}_{t,\ell})-\frac{1}{|\mathcal{H}|}\sum_{j\in\mathcal{H}}\mathbf{Z}_{t,m}\mathbf{g}_j(\mathbf{w}_{t,m}^{j},\mathbf{Z}_{t,m})\right\|^2\right]\\
&=\eta^2\sum_{\ell=1}^{K}\ell\sum_{m=1}^{\ell}\sum_{i\in\mathcal{H}}\frac{1}{|\mathcal{H}|}\xi_{13}\mathbb{E}\left[\left\|\hat{\mathbf{w}}_{t,m}-\mathbf{w}_{t,m}^{i}\right\|^2\right]+\eta^2\sum_{\ell=1}^{K}\ell\sum_{m=1}^{\ell}\xi_{14}+\eta^2\sum_{\ell=1}^{K}\ell\sum_{m=1}^{\ell}\xi_{15}\\
&\leq\eta^2\sum_{\ell=1}^{K}\frac{1}{|\mathcal{H}|}\sum_{i\in\mathcal{H}}K^2\xi_{13}\mathbb{E}\left[\left\|\hat{\mathbf{w}}_{t,\ell}-\mathbf{w}_{t,\ell}^{i}\right\|^2\right]+\eta^2 put\sum_{\ell=1}^{K}\ell\sum_{m=1}^{\ell}\xi_{14}+\eta^2\sum_{\ell=1}^{K}\ell\sum_{m=1}^{\ell}\xi_{15}
\end{aligned}$$

We rewrite the expression as

$$\left(1-\eta^2K^2\xi_{13}\right)\sum_{\ell=1}^{K}\frac{1}{|\mathcal{H}|}\sum_{i\in\mathcal{H}}\mathbb{E}\left[\left\|\mathbf{w}_{t,\ell}^{i}-\hat{\mathbf{w}}_{t,\ell}\right\|^2\right]\leq\eta^2\sum_{\ell=1}^{K}\ell\sum_{m=1}^{\ell}\xi_{14}+\eta^2\sum_{\ell=1}^{K}\ell\sum_{m=1}^{\ell}\xi_{15}$$

and choose the learning rate small enough so that $\left(1-\eta^2K^2\xi_{13}\right)=\left(1-6\eta^2K^2(D^2+\zeta^2)\right)\geq\frac{1}{2}$. Hence, we obtain

$$\sum_{\ell=1}^{K}\frac{1}{|\mathcal{H}|}\sum_{i\in\mathcal{H}}\mathbb{E}\left[\left\|\mathbf{w}_{t,\ell}^{i}-\hat{\mathbf{w}}_{t,\ell}\right\|^2\right]\leq\xi_2',$$

where

$$\xi_2'\triangleq 2\eta^2\sum_{\ell=1}^{K}\ell\sum_{m=1}^{\ell}\xi_{14}+2\eta^2\sum_{\ell=1}^{K}\ell\sum_{m=1}^{\ell}\xi_{15}\leq\xi_2\triangleq 4\eta^2K^3\frac{\varphi^2G^2d}{\nu}+12\eta^2K^3L.$$

The learning rate must satisfy $\eta\leq\sqrt{\frac{1}{12K^2(D^2+\zeta^2)}}$. This concludes the proof. □