On the Convergence and Stability of Distributed Sub-model Training

Yuyang Deng
Accenture

Fuli Qiao
Penn State

Mehrdad Mahdavi
Penn State

Abstract

As learning models continue to grow in size, enabling on-device local training of these models has emerged as a critical challenge in federated learning. A popular solution is sub-model training, where the server only distributes randomly sampled sub-models to the edge clients, and clients only update these small models. However, those random sampling of sub-models may not give satisfying convergence performance. In this paper, observing the success of SGD with shuffling, we propose a distributed shuffled sub-model training, where the full model is partitioned into several sub-models in advance, and the server shuffles those sub-models, sends each of them to clients at each round, and by the end of local updating period, clients send back the updated sub-models, and server averages them. We establish the convergence rate of this algorithm. We also study the generalization of distributed sub-model training via stability analysis, and find that the sub-model training can improve the generalization via amplifying the stability of training process. The extensive experiments also validate our theoretical findings.

1 INTRODUCTION

We consider optimizing the following objective

$$\min_{\mathbf{w}\in\mathcal{W}\subset\mathbb{R}^d} F(\mathbf{w}) := \frac{1}{N}\sum\nolimits_{i=1}^{N} \left\{ f_i(\mathbf{w}) := \mathbb{E}_{\xi\sim\mathcal{D}_i}[\ell(\mathbf{w};\xi)] \right\}, \tag{1}$$

collaboratively in a distributed setting with N clients, where $f_i(\mathbf{w})$ is the local loss function realized by ith client's data, and $\mathcal{W} \subset \mathbb{R}^d$ is some bounded convex set. To solve the problem with communication efficiency and data privacy, a widely employed method is FedAvg (McMahan et al., 2017) and its variants, where multiple devices (clients) collaborate to train a shared machine learning model without exchanging their data. One major limitation of FedAvg is that each client must maintain a local model with the same complexity as the global model. In modern machine learning (ML), where model complexity can reach millions or even billions of parameters, many clients may lack the memory and computational resources needed to store and optimize the full model. Moreover, devices participating in collaborative learning greatly vary in computational and storage capacity and can only confine to ML models that meet their resources for training. A common approach to address this issue is *partial training* where the server selects sub-models proportional to the computational resources available on each device, either *randomly* or based on predefined rules (e.g., *rolling* or *static partitioning*), and then distribute them to the clients. The clients only update these sub-models and, after a few rounds of local updates, send the updated sub-models back to the server for aggregation. The representative works include PruneFL (Jiang et al., 2022), IST (Yuan et al., 2019), HeteroFL (Diao et al., 2020), FedRolex (Alam et al., 2022), and ReeFL (Lee et al., 2024).

Despite the empirical success of this training paradigm, the convergence of distributed sub-model training has not been well understood. A key research question we seek to rigorously address is: *how does sub-model training, in comparison to full-model training, affect both the convergence and generalization* . Investigating this could provide insights into the trade-offs between computational efficiency and model performance in resource-constrained federated settings. A few recent studies have begun to explore this question, primarily from an optimization perspective, highlighting both the benefits and limitations of sub-model training in federated learning. (Shulgin and Richtárik, 2023) studied distributed fully synchronized sub-model training algorithm on quadratic objective, and showed that the convergence result will suffer from a residual error,

Proceedings of the 29th International Conference on Artificial Intelligence and Statistics (AISTATS) 2026, Tangier, Morocco. PMLR: Volume 300.

unless the objective and masking scheme admit some benign properties. (Demidovich et al., 2023) studied similar algorithm, and proved convergence to the optimal point of the masked objective for general strongly convex losses. For nonconvex loss, (Mohtashami et al., 2022) studied single machine setting, where $N = 1$, and proved that sub-model training converges to first order stationary point of the masked objective. Their convergence bound include factors that quantifies the alignment between gradient on the masked and non-masked models; however, it remains unclear whether these quantities can be effectively controlled. (Zhou et al., 2024) and (Wu et al., 2024) investigate the convergence of sub-model training with local updates on general nonconvex loss functions, making it the most relevant prior work to this study. However, their bound depends on the sum of the norms of the intermediate solutions; if these norms are too large, the convergence bound becomes vacuous.

A desired bound should *explicitly* show how sub-model selection strategy affects the convergence and generalization due to *model drift* caused by partial updating. In this paper, we present the first concrete convergence analysis of distributed sub-model training with different sampling schemes. We first consider FedAvg with Bernoulli random sub-model sampling. Then we consider rolling sub-model training method, where the server partitions the full model into R pieces and at the beginning of each epoch, it shuffles these sub-models and sequentially assigns them to the clients to be updated locally. We establish the convergence rate of both sampling schemes in convex and non-convex settings, highlighting the impact of partial training as captured through sampling. In nonconvex setting, we show that sub-model training will converge to the stationary point of an alternative objective induced by masking. To study the generalization of sub-model training, we further provide the generalization analysis of distributed sub-model training with random and rolling masking, and find that masking can enhance generalization by stabilizing the training process, as long as the residual optimization error from partial training remains controlled.

Contribution. The main results of this work include:

- We establish the rigorous convergence rate of distributed sub-model training with random sub-model selection (Section 2) and shuffled rolling sub-model selection (Section 3). We show that under strongly convex and smooth setting, they both enjoy an $O(\frac{1}{NKR})$ rate plus a residual error due to model masking, where R is the number of communication rounds and K is that of local steps.
- We further establish the convergence results of the two algorithms under nonconvex setting. We show that, under the algorithm dynamic, the model will finally converge to the stationary point of an alternative objective function induced by model masking. To the best of our knowledge, this is the first rigorous analysis of the convergence of permutation-based methods in sub-model training, thoroughly examining the interaction between model drift from partial training on sub-models and the impact of permutation-based sub-model assignments on convergence.
- We analyze the generalization ability of distributed sub-model training with random and rolling masking (Section 4), and find that masking can improve the generalization via stabilizing the training process, as long as the residual optimization error from partial training is controlled.
- We conduct thorough experiments that corroborate our theoretical results (Section 5).

Additional related works, empirical results, and omitted proofs can be found in the appendix.

2 DISTRIBUTED SUB-MODEL TRAINING VIA RANDOM SAMPLING

We consider the FedAvg (a.k.a. Local SGD) to optimize the objective in Eq. 1. The algorithm proceeds for R communication rounds, where at round r, each client, upon receiving the global model from the server, independently performs K local updates using its local data to compute gradients to update its local model parameters accordingly. After completing the K local steps, the clients send their updated parameters to the central server, which averages the models across all clients to produce a global model for the next round. This periodic averaging allows for reduced communication frequency and efficient use of bandwidth while enhancing convergence stability.

Adapting FedAvg for system heterogeneity is key to enabling collaboration among clients with varying computational power. A simple solution is distributed sub-model training using random sub-model sampling, which is first proposed in (Alam et al., 2022). At the beginning of rth communication round, the server generates a d-dimensional Bernoulli random masking $\mathbf{m}_i^r \sim Ber(p_i)$ (each coordinate is 1 with probability p_i, otherwise 0) for ith client and distributes the masked global model $\mathbf{w}_r^i = \mathbf{m}_i^r \odot \mathbf{w}_r$ to the client to perform the following update locally for K steps :

$$\mathbf{w}_{r,k+1}^i = \mathbf{w}_{r,k}^i - \eta \mathbf{m}_i^r \odot \nabla f_i(\mathbf{m}_i^r \odot \mathbf{w}_{r,k}^i; \xi_{r,k}^i),$$

Algorithm 1: Randomly Masked FedAvg

Input: Initial model $\mathbf{w}_0 = \mathbf{0}$, masking probabilities $p_1, \ldots, p_N$ and stepsizes η.
for $i = 1, \ldots, N$ **do**
  for $r = 0, \ldots, R-1$ **do**
    Server generates Bernoulli random masks $\mathbf{m}_1^r, \ldots, \mathbf{m}_N^r$ for each client, according to probability $p_1, \ldots, p_N$.
    Server distributes $\mathbf{w}_r^i = \mathbf{m}_i^r \odot \mathbf{w}_r$ to ith user.
    for $k = 0, \ldots, K-1$ **do**
      $\mathbf{w}_{r,k+1}^i = \mathbf{w}_{r,k}^i - \eta \mathbf{m}_i^r \odot \nabla f_i(\mathbf{m}_i^r \odot \mathbf{w}_{r,k}^i; \xi_{r,k}^i)$
    end
    The ith client sends $\mathbf{w}_{r,K}^i$ back to server.
    Server averages models $\mathbf{w}_{r+1} = \mathcal{P}_{\mathcal{W}}\left(\frac{1}{N}\sum_{i=1}^N (\mathbf{w}_{r,K}^i + (\mathbf{1} - \mathbf{m}_i^r) \odot \mathbf{w}_r)\right)$
  end
end
Output: $\hat{\mathbf{w}} = \mathcal{P}_{\mathcal{W}}\left(\mathbf{w}_R - \frac{1}{L}\frac{1}{N}\sum_{i=1}^N \mathbf{m}_i \odot \nabla f_i(\mathbf{m}_i \odot \mathbf{w}_R)\right)$ where $\mathbf{m}_i \sim Ber(p_i)$.

where $\xi_{r,k}^i$ is the data point uniformly randomly sampled from ith client's dataset. Each client updates only a subset of the model parameters, with the number of updated parameters determined by the masking probability p_i. For clients with limited capacity, the server can assign a smaller p_i to reduce their computational burden. After K local steps, ith client sends model $\mathbf{m}_i^r \odot \mathbf{w}_{r,K}$ to server, and server averages the received models:

$$\mathbf{w}_{r+1} = \mathcal{P}_{\mathcal{W}}\left(\frac{1}{N}\sum_{i=1}^N (\mathbf{w}_{r,K}^i + (\mathbf{1} - \mathbf{m}_i^r) \odot \mathbf{w}_r)\right).$$

where $\mathcal{P}_{\mathcal{W}}(\cdot)$ is the projection operator onto convex set $\mathcal{W}$. In words, server will fill the parameters not selected by $\mathbf{m}_i^r$ with old parameters of the last round model, i.e., $\mathbf{w}_r$, and then average all clients models. After that, server distributes $\mathbf{m}_i^{r+1} \odot \mathbf{w}_{r+1}$ to ith client. The pseudo-code of the algorithm is depicted in Algorithm 1.

2.1 Convergence in Convex Setting

To study the convergence of this simple algorithm, we make the following standard assumptions.

Assumption 1 (Smoothness). *We assume $\forall i \in [N]$, $f_i(\mathbf{x})$ is L-smooth, i.e.,*

$$\forall \mathbf{x}, \mathbf{y} : \|\nabla f_i(\mathbf{x}) - \nabla f_i(\mathbf{y})\| \leq L\|\mathbf{x} - \mathbf{y}\|.$$

Assumption 2 (Strong convexity). *We assume $\forall i \in [N]$, $f_i(\mathbf{x})$ is L-smooth and μ-strongly convex*

$$\forall \mathbf{x}, \mathbf{y} : f_i(\mathbf{y}) \geq f_i(\mathbf{x}) + \langle \nabla f_i(\mathbf{x}), \mathbf{y} - \mathbf{x}\rangle + \frac{\mu}{2}\|\mathbf{y} - \mathbf{x}\|^2.$$

We denote the condition number by $\kappa = L/\mu$.

Assumption 3 (Bounded variance). *The variance of stochastic gradients computed at each local function is bounded, i.e., $\forall i \in [N], \forall \mathbf{w} \in \mathcal{W}, \mathbb{E}_{\mathbf{m}_i \sim Ber(p_i), \xi}[\|\mathbf{m}_i \odot \nabla f_i(\mathbf{m}_i \odot \mathbf{w}; \xi) - \mathbf{m}_i \odot \nabla f_i(\mathbf{m}_i \odot \mathbf{w})\|^2] \leq \delta^2$.*

Assumption 4 (Bounded domain). *The domain $\mathcal{W} \subset \mathbb{R}^d$ is a bounded convex set, with diameter W under ℓ_2 metric, i.e., $\forall \mathbf{w} \in \mathcal{W}, \|\mathbf{w}\| \leq W$.*

Assumption 5 (Bounded gradient). *The gradients computed at each local function are bounded, i.e., $\forall i \in [N], \sup_{\mathbf{w} \in \mathcal{W}} \|\nabla f_i(\mathbf{w})\| \leq G$.*

Definition 1. *We define the masked heterogeneity at optimum as follows:*

$$\frac{1}{N}\sum_{i=1}^N \mathbb{E}_{\mathbf{m}_i \sim Ber(p_i)} \|\mathbf{m}_i \odot \nabla f_i(\mathbf{m}_i \odot \mathbf{w}^*(\mathbf{p}))\|^2 \leq \sigma_*^2.$$

where $\mathbf{w}^*(\mathbf{p}) = \arg\min_{\mathbf{w} \in \mathcal{W}} \frac{1}{N}\sum_{i=1}^N \mathbb{E}_{\mathbf{m}_i \sim Ber(p_i)}[f_i(\mathbf{m}_i \odot \mathbf{w})]$.

Assumptions 1, 2, 4 and 5 are standard in convex optimization. Assumption 3 is the bounded variance of masked gradients which becomes small if the masking probability p_i is high. Definition 1 is also standard in Local SGD analysis (Khaled et al., 2020), but here it is adapted to the masked gradients.

Theorem 1. *Let Assumptions 1- 5 hold. Then Algorithm 1 with $\eta = \frac{\log(KR)^2}{\tilde{\mu} KR}$ and $R \geq \frac{L}{\tilde{\mu}}\log(K^2R^2)$ will output the solution $\hat{\mathbf{w}}$, such that the following statement holds:*

$$F(\hat{\mathbf{w}}) - F(\mathbf{w}^*) \leq L\tilde{O}\left(\frac{\mathbb{E}\|\mathbf{w}_0 - \mathbf{w}^*\|^2}{K^2R^2} + \frac{\tilde{\kappa}\sigma_*^2 + \tilde{\kappa}\delta^2}{\tilde{\mu}^2R^2} + \frac{\delta^2}{\tilde{\mu}^2 NKR}\right) + \underbrace{\left(\frac{5L}{2\tilde{\mu}} + \frac{4}{L}\right)\frac{2G^2 + 2W^2L^2}{N}\sum_{i=1}^N d(1-p_i)}_{\text{Residual error due to masked updates}},$$

where $\mathbf{w}^* := \arg\min_{\mathbf{w} \in \mathcal{W}} F(\mathbf{w}), \bar{\mu} := \frac{1}{N}\sum_{i=1}^N p_i\mu,$

$$\tilde{\mu} := \min_{i \in [N]} p_i\mu,\ \tilde{L} := \max_{i \in [N]} p_iL,\ \tilde{\kappa} := \tilde{L}/\tilde{\mu}.$$

Here we achieve an $O(\frac{1}{R^2} + \frac{1}{NKR})$ rate plus residual error due to masked updates. If each client chooses masking probability to be 1, i.e., enabling full model

training, the residual error vanishes and we recover the convergence of heterogeneous Local SGD (Woodworth et al., 2020b; Khaled et al., 2020).

■ **Comparison to existing works.** (Shulgin and Richtárik, 2023) studied the special scenario where $f(\mathbf{w})$ is quadratic, and also proved that the convergence rate will suffer from a residual error. (Demidovich et al., 2023) studied single machine and distributed fully synchronized versions of Algorithm 1, and proved convergence to the optimal point of the masked objective, i.e., the minimizer of $F_{\mathcal{D}}(\mathbf{w}) := \mathbb{E}_{\mathbf{m}\sim\mathcal{D}}[f(\mathbf{m}\odot\mathbf{w})]$ where $\mathcal{D}$ represents a distribution on masking vector.

2.2 Convergence in Nonconvex Setting

In this section, we will present convergence result of Algorithm 1 in nonconvex setting. We will need the following heterogeneity measure.

Definition 2. *We define the masked gradient dissimilarity as follows:*

$$\max_{\mathbf{w}\in\mathbb{R}^d} \frac{1}{N}\sum_{i=1}^{N} \mathbb{E}_{\mathbf{m}_i} \left\|\mathbf{m}_i\odot\nabla f_i(\mathbf{m}_i\odot\mathbf{w}) - \nabla F_{\mathbf{p}}(\mathbf{w})\right\|^2 = \zeta_{\mathbf{p}}^2,$$

where $\mathbf{m}_i \sim Ber(p_i), i = 1,\dots,N$ *and* $F_{\mathbf{p}}(\mathbf{w}) := \frac{1}{N}\sum_{i=1}^{N}\mathbb{E}_{\mathbf{m}_i\sim Ber(p_i)}[f_i(\mathbf{m}_i\odot\mathbf{w})]$.

Similar definition can be found in the classical Local SGD analysis (Woodworth et al., 2020b), but here dissimilarity is defined over the masked objective. A more aggressive masking scheme will result in a smaller $\zeta_{\mathbf{p}}^2$.

Theorem 2. *Let Assumptions 1 and 3 hold. Then Algorithm 1 with* $\eta = \Theta\left(\frac{1}{L\sqrt{KR}}\right)$ *guarantees that the following statement holds for* $F_{\mathbf{p}}$ *as defined in Eq. (2):*

$$\frac{1}{R}\sum_{r=1}^{R}\mathbb{E}\left\|\nabla F_{\mathbf{p}}(\mathbf{w}_r)\right\|^2 \le O\left(\frac{\tilde{L}\mathbb{E}[F_{\mathbf{p}}(\mathbf{w}_0)]}{\sqrt{RK}} + \frac{K\zeta_{\mathbf{p}}^2}{R} + \frac{K\delta^2}{R} + \frac{\delta^2}{N\sqrt{RK}}\right).$$

We can see that sub-model training will converge to the stationary point of an alternative objective induced by masking, not the raw objective $F(\mathbf{w})$. We achieve $O(\frac{1}{\sqrt{RK}} + \frac{K\zeta_{\mathbf{p}}^2}{R})$ rate, analogous to the rate of FedAvg with full model training (Haddadpour and Mahdavi, 2019).

■ **Comparison to existing works.** (Mohtashami et al., 2022) studied single machine sub-model training setting, i.e., $N = 1$. Given a fixed sequence of masks at each iteration, i.e., $\mathbf{m}_1,\dots,\mathbf{m}_T$, they perform update $\mathbf{w}_{t+1} = \mathbf{w}_t - \eta\mathbf{m}_t\odot\nabla F(\mathbf{m}_t\odot\mathbf{w}_t;\xi_t)$. They proved $\frac{1}{T}\sum_{t=1}^{T}\alpha_t\mathbb{E}\left\|\mathbf{m}_t\odot F(\mathbf{m}_t\odot\mathbf{w}_t)\right\|^2 \le O(\frac{1}{\sqrt{T}})$, where $\alpha_t := \frac{\langle \mathbf{m}_t\odot\nabla F(\mathbf{w}_t), \mathbf{m}_t\odot\nabla F(\mathbf{m}_t\odot\mathbf{w}_t)\rangle}{\left\|\mathbf{m}_t\odot\nabla F(\mathbf{m}_t\odot\mathbf{w}_t)\right\|^2}$. However, it is not shown how small the α_t can be. If α_ts approach zero, the convergence measure loses significance. (Zhou et al., 2024)studies a similar algorithm and examines convergence via the gradient norm; however, their analysis yields a bound that depends on the norm of history models, i.e., $\|\mathbf{w}_t\|$. With the model's norm uncontrolled in the bound, it is possible that the convergence bound becomes vacuous.

■ **On the stationary points of F and $F_{\mathbf{p}}$.** In (Zhou et al., 2024), the authors proved the convergence of the model to the stationary point of F, with residual error depending on norm of iterates, i.e., $\|\mathbf{w}_r\|$. Indeed, our result can also be translated to the stationarity of F, by using norm of iterates. To see this, assume $\tilde{\mathbf{w}}_\epsilon$ is ϵ stationary point of $F_{\mathbf{p}}$. If we evaluate F's gradient at $\tilde{\mathbf{w}}_\epsilon$ we have:

$$\begin{aligned}
&\left\|\nabla F(\tilde{\mathbf{w}}_\epsilon)\right\|^2 \\
&\le 2\left\|\nabla F_{\mathbf{p}}(\tilde{\mathbf{w}}_\epsilon)\right\|^2 + 2\left\|\nabla F_{\mathbf{p}}(\tilde{\mathbf{w}}_\epsilon) - \nabla F(\tilde{\mathbf{w}}_\epsilon)\right\|^2 \\
&\le 2\epsilon^2 \\
&\quad + 2\left\|\frac{1}{N}\sum_{i=1}^{N}\mathbb{E}\left[\mathbf{m}_i\odot\nabla f_i(\mathbf{m}_i\odot\tilde{\mathbf{w}}_\epsilon)\right] - \nabla F(\tilde{\mathbf{w}}_\epsilon)\right\|^2 \\
&\le 2\epsilon^2 + \frac{1}{N}\sum_{i=1}^{N} d(1-p_i)(G^2 + L^2\left\|\tilde{\mathbf{w}}_\epsilon\right\|^2).
\end{aligned}$$

Thus, any ϵ-stationary point of $F_{\mathbf{p}}$ is also $\sqrt{2\epsilon^2 + \frac{1}{N}\sum_{i=1}^{N} d(1-p_i)(G^2 + L^2\left\|\tilde{\mathbf{w}}_\epsilon\right\|^2)}$-stationary point of F.

3 DISTRIBUTED SUB-MODEL TRAINING VIA ROLLING MASKING

Another popular algorithm for distributed sum-model training is via rolling masking (Alam et al., 2022), where each of the whole model parameters is divided into several pre-defined sub-models,

$$\mathbf{m}_i^1\odot\mathbf{w},\dots,\mathbf{m}_i^R\odot\mathbf{w}$$

where $\mathbf{m}_i^j$ is such that j to $(j+s_i) \mod d$ coordinate is 1 and rest are zero, s_i is the sub-model size for ith client. At the beginning of eth epoch, server generates a permutation $\sigma_e : [R] \mapsto [R]$ and shuffles the clients' sub-model according to σ_e:

$$\mathbf{m}_i^{\sigma_e(1)}\odot\mathbf{w},\dots,\mathbf{m}_i^{\sigma_e(R)}\odot\mathbf{w}.$$

Then, ith client will optimize on those sub-models sequentially at each round. At the beginning of rth

Algorithm 2: Rolling Masked FedAvg

Input: Initial model $\mathbf{w}_0 = \mathbf{0}$, pre-defined masking vectors $\left\{\left\{\mathbf{m}_i^j\right\}_{j=1}^R\right\}_{i=1}^N$, stepsizes η.
for $e = 0, \dots, T-1$ **do**
  Server generates random permutation $\sigma_e : [R] \mapsto [R]$.
  $\mathbf{w}_{e,1} = \mathbf{w}_e$
  for $r = 1, \dots, R$ **do**
    Server distributes $\mathbf{w}_{e,r}^i = \mathbf{m}_i^{\sigma_e(r)} \odot \mathbf{w}_{e,r}$ to ith user.
    for $i = 1, ..., N$ **do**
      for $k = 0, \dots, K-1$ **do**
        $\mathbf{w}_{e,r,k+1}^i = \mathbf{w}_{e,r,k}^i - \eta \mathbf{m}_i^{\sigma_e(r)} \odot \nabla f_i(\mathbf{m}_i^{\sigma_e(r)} \odot \mathbf{w}_{e,r,k}^i; \xi_{e,r,k}^i)$
      end
      ith Client sends $\mathbf{w}_{e,r,K}^i$ back to server.
    end
    Server averages models $\mathbf{w}_{e,r+1} = \mathcal{P}_{\mathcal{W}}\left(\frac{\sum_{i=1}^N (\mathbf{w}_{e,r,K}^i + (\mathbf{1} - \mathbf{m}_i^{\sigma_e(r)}) \odot \mathbf{w}_{e,r})}{N}\right)$
    $\mathbf{w}_{e,r+1} = \mathbf{w}_{e,r,K}$
  end
  $\mathbf{w}_{e+1} = \mathbf{w}_{e,R+1}$
end
Output: $\hat{\mathbf{w}} = \mathcal{P}_{\mathcal{W}}\left(\mathbf{w}_T - \frac{1}{L}\frac{1}{N}\sum_{i=1}^N \frac{\sum_{r=1}^R \mathbf{m}_i^r \odot \nabla f_i(\mathbf{m}_i^r \odot \mathbf{w}_T)}{R}\right)$

round, it performs the following local updates:

$$\mathbf{g}_{e,r,k}^i = \mathbf{m}_i^{\sigma_e(r)} \odot \nabla f_i(\mathbf{m}_i^{\sigma_e(r)} \odot \mathbf{w}_{e,r,k}^i; \xi_{e,r,k}^i)$$
$$\mathbf{w}_{e,r,k+1}^i = \mathbf{w}_{e,r,k}^i - \eta \mathbf{g}_{e,r,k}^i.$$

After K local steps, ith client sends model $\mathbf{m}_i^{\sigma_e(r)} \odot \mathbf{w}_{e,r,K}$ to server, and server averages the local models:

$$\mathbf{w}_{e,r+1} = \mathcal{P}_{\mathcal{W}}\left(\frac{1}{N}\sum\nolimits_{i=1}^N (\mathbf{w}_{e,r,K}^i + (\mathbf{1} - \mathbf{m}_i^{\sigma_e(r)}) \odot \mathbf{w}_{e,r})\right).$$

In words, server will fill the parameters not selected by $\mathbf{m}_i^{\sigma_e(r)}$ with old parameters of the model from last round model $\mathbf{w}_{e,r}$, and then average all clients models. Next, server distributes $\mathbf{m}_i^{\sigma_e(r+1)} \odot \mathbf{w}_{e,r+1}$ to ith client to proceed another round of local updates. The pseudo-code is depicted in Algorithm 2.

3.1 Convergence in Convex Setting

Assumption 6 (Bounded variance)**.** *The variance of stochastic gradients computed at each local function is bounded, i.e.,* $\forall i \in [N], \forall r \in [R], \forall \mathbf{w} \in \mathcal{W}, \mathbb{E}_\xi[\|\mathbf{m}_i^r \odot \nabla f_i(\mathbf{m}_i^r \odot \mathbf{w}; \xi) - \mathbf{m}_i^r \odot \nabla f_i(\mathbf{m}_i^r \odot \mathbf{w})\|^2] \le \delta^2$.

Definition 3. *Given a masking configuration* $\mathbf{m} = \left[\left[\mathbf{m}_i^1, \dots, \mathbf{m}_i^R\right]\right]_{i=1}^N \in \{0,1\}^{dNR}$*, we define the masked gradient dissimilarity as follows:*

$$\max_{\mathbf{w}\in\mathcal{W}} \frac{1}{NR}\sum_{i=1}^N\sum_{j=1}^R \left\|\mathbf{m}_i^j \odot \nabla f_i(\mathbf{m}_i^j \odot \mathbf{w}) - \nabla F_{\mathbf{m}}(\mathbf{w})\right\|^2 \le \zeta_{\mathbf{m}}^2,$$

where $F_{\mathbf{m}}(\mathbf{w}) = \frac{1}{N}\sum_{i=1}^N \frac{1}{d}\sum_{j=1}^d f_i(\mathbf{m}_i^j \odot \mathbf{w})$.

Assumption 6 and Definition 3 are analogous to Assumption 3 and Definition 2, but here the masking vectors are deterministic.

Theorem 3. *Let Assumptions 1, 2, 4, 5 and 6 hold. Then Algorithm 2 with* $\eta = \Theta\left(\frac{\log(T^2)}{\mu KRT}\right)$ *and* $T \ge 512\kappa^2 \log T$ *will output the solution* $\hat{\mathbf{w}}$*, such that with probability at least* $1-\nu$*, the following statement holds:*

$$F(\hat{\mathbf{w}}) - F(\mathbf{w}^*) \le O\left(\frac{L\mathbb{E}\left\|\mathbf{w}_0 - \mathbf{w}^*\right\|^2}{T^2}\right) + LO\left(\frac{\kappa^2\log(RK/\nu)}{\mu T^2 R} + \frac{\kappa\zeta_{\mathbf{m}}^2\log^2(T)}{\mu^2T^2R^2} + \frac{\delta^2\log(T)}{\mu^2TN}\right) + \underbrace{\left(\frac{L}{\mu}+\frac{1}{L}\right)\frac{G^2+W^2L^2}{NR}\sum_{i=1}^N\sum_{j=1}^R\left\|\mathbf{m}_i^j - \mathbf{1}\right\|^2}_{\text{Residual error due to masked updates}}.$$

The first part of the rate is contributed from shuffled Local SGD, and the second part is due to masking updates. If each mask $\mathbf{m}_i^j = \mathbf{1}$, i.e., full model training, we can get rid of this residual error. Note that our algorithm is different from Cyclic FedAvg (Cho et al., 2023), where they do the shuffling on the client level, and for each communication round, a subset of clients are picked to do local updates.

The proof of theorem is deferred to Appendix D. We note that in our convergence analysis, we account for the partitioning of the full model into R sub-models. At each epoch, the server shuffles and sequentially assigns these sub-models to clients, introducing analytical challenges due to model drift from partial training and the effects of permutation-based assignments. Our technical contribution lies in jointly addressing these challenges to establish convergence, an aspect that is interesting in its own right.

Remark 1. *(Deng et al., 2024) proposed and studied similar algorithm, but in their work, the clients directly optimize the full models, while in ours, clients optimize local models and server performs averaging periodically. Another relevant work is (Cho et al., 2023), where they studied Local SGD with cyclic client participation. The main difference is that they do client-level shuffling and at each round server only picks a subset of clients to participate training.*

3.2 Convergence in Nonconvex Setting

In this section, we will present convergence result of Algorithm 2 in nonconvex setting.

Theorem 4. *Let Assumptions 1, 3 and 5 hold. Then Algorithm 2 with* $\eta = \Theta\left(\frac{1}{L\sqrt{KRT}}\right)$ *guarantees that with probability at least* $1-\nu$*, for* $F_{\mathbf{m}}$ *defined as in Eq. (15) the following holds true:*

$$\frac{1}{T}\sum_{e=1}^{T}\mathbb{E}\left\|\nabla F_{\mathbf{m}}(\mathbf{w}_e)\right\|^2 \leq O\left(\frac{L\mathbb{E}[F_{\mathbf{m}}(\mathbf{w}_0)]}{\sqrt{RKT}} + \frac{K^2\zeta_{\mathbf{m}}^2}{TL^2} + \frac{\delta^2}{N\sqrt{RKT}L}\right).$$

The convergence rate here matches that of the random masking case, measured by the gradient norm of an alternative objective $F_{\mathbf{m}}$, induced by sub-model selection scheme.

■ **On the stationary points of F and $F_{\mathbf{m}}$.** We can also translate stationarity of $F_{\mathbf{m}}$ to that of F as follows. Similar to random masking setting, we can also translate the stationarity between F and $F_{\mathbf{m}}$. Assume $\tilde{\mathbf{w}}_\epsilon$ is ϵ stationary point of $F_{\mathbf{m}}$, if we evaluate F's gradient at $\tilde{\mathbf{w}}_\epsilon$ we have:

$$\begin{aligned}
&\left\|\nabla F(\tilde{\mathbf{w}}_\epsilon)\right\|^2 \leq 2\left\|\nabla F_{\mathbf{m}}(\tilde{\mathbf{w}}_\epsilon)\right\|^2 + 2\left\|\nabla F_{\mathbf{m}}(\tilde{\mathbf{w}}_\epsilon) - \nabla F(\tilde{\mathbf{w}}_\epsilon)\right\|^2 \\
&\leq 2\epsilon^2 \\
&+ 2\left\|\frac{1}{N}\sum_{i=1}^{N}\left(\frac{1}{R}\sum_{j=1}^{R}\mathbf{m}_i^j \odot \nabla f_i(\mathbf{m}_i^j \odot \tilde{\mathbf{w}}_\epsilon) - \nabla f_i(\tilde{\mathbf{w}}_\epsilon)\right)\right\|^2 \\
&\leq 2\epsilon^2 + \frac{1}{N}\sum_{i=1}^{N}\frac{1}{R}\sum_{j=1}^{R}\left\|\mathbf{1} - \mathbf{m}_i^j\right\|^2 (G^2 + L^2\left\|\tilde{\mathbf{w}}_\epsilon\right\|^2).
\end{aligned}$$

4 ON THE STABILITY OF MASKED TRAINING

In this section, we will study the generalization ability of distributed sub-model training with Bernoulli and rolling based masking. Formally, for a learning algorithm $\mathcal{A}$ and training dataset $\mathcal{S}$ drawn from distribution $\mathcal{D}$, the generalization error is defined as $\epsilon_{gen} := \mathbb{E}_{\mathcal{A},\mathcal{S}}\left|\mathcal{L}_{\mathcal{D}}(\mathcal{A}(\mathcal{S})) - \mathcal{L}_{\mathcal{S}}(\mathcal{A}(\mathcal{S}))\right|$. A classical way to study the above error is algorithmic stability. We adopt the following definition of stable federated learning algorithm from (Sun et al., 2024).

Definition 4. *A federated learning algorithm $\mathcal{A}$ is said to have ϵ-on-average stability if given any two neighboring datasets $\mathcal{S}$ and $\mathcal{S}^{(i)}$, then $\forall i \in [N], j \in [n]$*

$$\mathbb{E}_{\mathcal{A},\mathcal{S},z'_{i,j}}\left|\ell(\mathcal{A}(\mathcal{S}); z'_{i,j}) - \ell(\mathcal{A}(\mathcal{S}^{(i)}); z'_{i,j})\right| \leq \epsilon.$$

where $\mathcal{S}$ and $\mathcal{S}^{(i)}$ are the two dataset only differing at jth point of ith client's dataset, i.e., $\mathcal{S} = \{\ldots, z_{i,j}, \ldots\}$ and $\mathcal{S}^{(i)} = \{\ldots, z'_{i,j}, \ldots\}$.

An immediate implication of ϵ-on-average stability is the following lemma on generalization error.

Lemma 1. *(Sun et al., 2024) If $\mathcal{A}$ is an ϵ-on-average stable algorithm, then*

$$\epsilon_{gen} \leq \mathbb{E}_{\mathcal{A},\mathcal{S}}\left|\mathcal{L}_{\mathcal{D}}(\mathcal{A}(\mathcal{S})) - \mathcal{L}_{\mathcal{S}}(\mathcal{A}(\mathcal{S}))\right| \leq \epsilon.$$

Lemma 1 indicates that any ϵ-on-average stable algorithm will admit the expected generalization error no larger than ϵ as well. Hence, to study the generalization of Algorithm 1, it suffices to bound the difference between the models trained on raw training set and one-point-perturbed dataset as

$$\begin{aligned}
&\mathbb{E}_{\mathcal{A},\mathcal{S},z'_{i,j}}\left|\ell(\mathcal{A}(\mathcal{S}); z'_{i,j}) - \ell(\mathcal{A}(\mathcal{S}^{(i)}); z'_{i,j})\right| \\
&\leq \mathbb{E}_{\mathcal{A},\mathcal{S},z'_{i,j}} G\left\|\mathcal{A}(\mathcal{S}) - \mathcal{A}(\mathcal{S}^{(i)})\right\|.
\end{aligned}$$

Hence, we are aimed at studying $\left\|\mathcal{A}(\mathcal{S}) - \mathcal{A}(\mathcal{S}^{(i)})\right\|$. We will make the following assumption.

Assumption 7 (Convexity)**.** *We assume $f_i(\mathbf{x})$'s are convex, i.e., $\forall \mathbf{x}, \mathbf{y} : f_i(\mathbf{y}) \geq f_i(\mathbf{x}) + \langle \nabla f_i(\mathbf{x}), \mathbf{y} - \mathbf{x}\rangle$.*

Assumption 8 (Point-wise Smoothness)**.** *We assume $\forall \xi \in \Xi$, $\nabla_k \ell(\cdot; \xi)$ is l_ℓ-Lipschitz, i.e.,*

$$\forall \mathbf{w}, \mathbf{w}' \in \mathcal{W} : \left\|\nabla_k \ell(\mathbf{w}; \xi) - \nabla_k \ell(\mathbf{w}'; \xi)\right\| \leq l_\ell \|\mathbf{w} - \mathbf{w}'\|.$$

Assumption 9 (Bounded L_∞ norm)**.** *We assume that $G_\infty := \max_{\mathbf{w}\in\mathcal{W}, i\in[N]}\left\|\nabla f_i(\mathbf{w})\right\|_\infty$ and $W_\infty := \max_{\mathbf{w}\in\mathcal{W}}\left\|\mathbf{w}\right\|_\infty$.*

Theorem 5. *[Stability of Random Masking] Let Assumptions 1, 3, 4 and 7 hold. We assume each client has n training data drawn from its distribution. Let $\hat{\mathbf{w}}$ and $\hat{\mathbf{w}}'$ be the output model of Algorithm 1 on dataset $\mathcal{S}$ and $\mathcal{S}^{(i)}$. Then, if we choose $\eta = \frac{\sqrt{Nn}}{RK}$ and R to be sufficiently large, it holds that $\mathbb{E}\left\|\hat{\mathbf{w}} - \hat{\mathbf{w}}'\right\|$ is bounded by*

$$\mathcal{O}\left(\frac{d\cdot\Psi_{\max}}{\sqrt{Nn}} + \sqrt{\frac{\sigma_*^2 + \delta^2}{Nn}}\right),$$

where

$$\begin{aligned}
\Psi_{\max} := \max_{i,j}\Big[&p_i l_\ell^2 W_1(\mathcal{D}_i, \mathcal{D}_j)^2 \\
&+ (p_i + p_j - p_i p_j)\left(G_\infty^2 + L^2 W_\infty^2\right)\Big]
\end{aligned}$$

and $\bar{p} = \frac{1}{N}\sum_{i=1}^{N} p_i$.

Remark 2. *Even though we assume each client has the same amount of data, the analysis can be easily extended to the case where ith client has n_i data samples, but we have to assume our objective is also weighted accordingly, i.e., $F(\mathbf{w}) = \sum_{i=1}^{N}\frac{n_i}{n} f_i(\mathbf{w})$.*

Corollary 1. *Let Assumptions 1, 3, 4 and 7 hold. Then, if we choose $\eta = \frac{\sqrt{Nn}}{RK}$ and R to be sufficiently*

large, Algorithm 1 admits the generalization error ϵ_{gen} bounded by

$$G \cdot \mathcal{O}\left(\frac{d \cdot \Psi_{\max}}{\sqrt{Nn}} + \sqrt{\frac{\sigma_*^2 + \delta^2}{Nn}}\right).$$

Remark 3. *We can see that the main heterogeneity term depends on the masking probability. A smaller masking probability $\{p_i\}_{i=1}^N$ will result in a smaller $\max_{i,j} p_i d l_\ell W_1(\mathcal{D}_i, \mathcal{D}_j)$. Hence masking improves the generalization (empirical risk vs population risk) by stabilizing the training process, as validated by our experiments. However, a lower masking probability p_i will also increase the residual error of the convergence or empirical risk, as stated in Theorem 1. Consequently, the above theorem implies that masking can enhance generalization, as long as the residual optimization error from partial training remains controlled.*

Remark 4. *A related work to ours is (Fu et al., 2023), which also studies the generalization error of sparse training, and has similar conclusion that sparsity can improve the algorithmic stability. The main difference to ours is that they consider regularized ERM algorithm, while our analysis is built for an iterative distributed optimization algorithm.*

Theorem 6. *[Stability of Rolling Masking] Let Assumptions 1, 3, 4 and 7 hold. We assume each client has n training data drawn from its distribution. Let $\hat{\mathbf{w}}$ and $\hat{\mathbf{w}}'$ be the output models of Algorithm 2 on dataset $\mathcal{S}$ and $\mathcal{S}^{(i)}$, respectively. Then, if we choose $\eta = \frac{\sqrt{Nn}}{RK}$ and $T \geq \sqrt{\frac{n}{N}}$, we have $\mathbb{E}\|\hat{\mathbf{w}} - \hat{\mathbf{w}}'\|$ is bounded by:*

$$\mathcal{O}\left(\frac{\Psi_{\max}}{\sqrt{Nn}} + \sqrt{\frac{\sigma_*^2 + \delta^2}{Nn}}\right),$$

where $\Psi_{\max} :=$

$$d_{\mathbf{m}} l_\ell \max_{i,j} W_1(D_i, D_j) + \sqrt{D_{\max}}\,(G_\infty + L W_\infty)$$

$d_{\mathbf{m}} = \max_{i,j} \left\|\mathbf{m}_i^j\right\|_0$, $D_{\max} := \max_{i,j,i',j'} \|\mathbf{m}_i^j - \mathbf{m}_{i'}^{j'}\|_0$.

Here we achieve similar $\frac{1}{\sqrt{Nn}}$ rate as random masking, depending on the size of the largest sub-model ($d_{\mathbf{m}}$), data heterogeneity ($W_1(D_i, D_j)$) and sub-model drift ($D_{\max}$). It indicates that rolling masking can also enjoy a more stable training dynamic, leading to a better generalization rate. The smaller sub-model will also mitigate the impact of data heterogeneity as reflected by term $d_{\mathbf{m}} l_\ell \max_{i,j} W_1(D_i, D_j)$.

Remark 5. *Notice that the full-space Lipschitz constant is related to the coordinate-wise constant by $L_\ell^2 \approx d \cdot l_\ell^2$. Our refined bound $d_m \cdot l_\ell^2 W_{\max}^2$ strictly scales with the subnetwork size d_m. When $d_m \ll d$, this bound is significantly tighter than the full model training with dependence on $L_\ell^2 W_{\max}^2$.*

5 EXPERIMENTS

In this section, we present a comprehensive evaluation of different masking algorithms through a series of experiments designed to assess their performance across various scenarios. Additional results are reported in Appendix B.

5.1 Experiment Setup

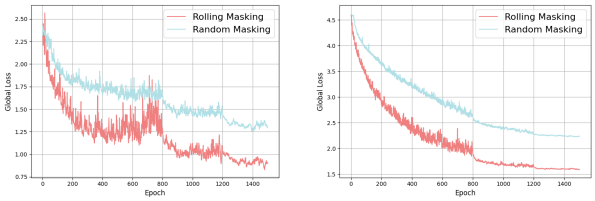
Datasets and Models. We evaluate the performance of different masking in the following scenarios. We train pre-activated ResNet18 models on CIFAR-10 and CIFAR-100. We modify the ResNet18 architecture by replacing batch normalization with static batch normalization and incorporating a scalar module after each convolutional layer.

Data Heterogeneity. To create non-IID data distributions for CIFAR-10 and CIFAR-100 into 100 clients respectively, we follow FedRolex (Alam et al., 2022), restricting each client to have access to only L labels. We evaluate two levels of data heterogeneity. For CIFAR-10, we set $L = 2$ as high data heterogeneity and $L = 5$ as low data heterogeneity, which corresponds to the Dirichlet distribution with $\alpha = 0.1$ and $\alpha = 0.5$, respectively. For CIFAR-100, we set $L = 20$ as high data heterogeneity and $L = 50$ as low data heterogeneity, which also corresponds to the Dirichlet distribution with $\alpha = 0.1$ and $\alpha = 0.5$, respectively. Results for high data heterogeneity and low data heterogeneity are both presented in the following subsections.

Model Heterogeneity. For ResNet18, client model capacities $\beta = \{1, 1/2, 1/4, 1/8, 1/16\}$ are used for evaluation, i.e., 1/16 means the client model capacity is 1/16 of the largest client model capacity (full model). We vary the number of kernels in the convolutional layers while maintaining the same number of nodes in the output layers.

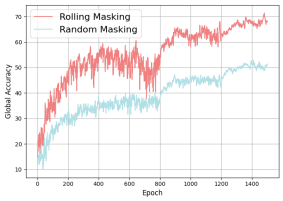
5.2 Convergence of Masked Training

Model-Heterogeneous Setting. We compare the performance of rolling and random masking in model-heterogeneous scenarios, following prior work where client capacities are uniformly distributed and the global server model is the same as the largest client model. Fig. 1(a) and Fig. 1(b) show the global model testing loss under high data heterogeneity, indicating that rolling masking outperforms random in both datasets. Similarly, Fig. 1(c) and Fig. 1(d) illustrate the corresponding global model testing accuracy. Consistent with the loss results, rolling masking outperforms random in both datasets. Under low data heterogeneity, Fig. 2(a) and Fig. 2(b) show the global model testing

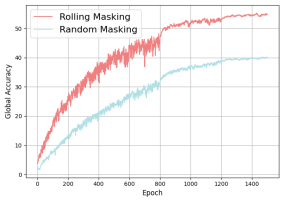


(a) CIFAR-10

(b) CIFAR-100

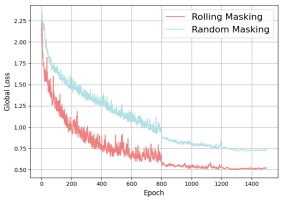


(c) CIFAR-10

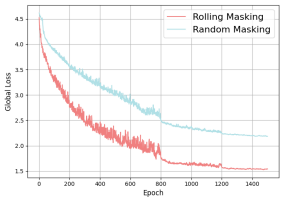


(d) CIFAR-100

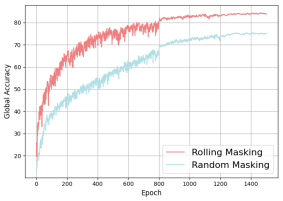
Figure 1: Global testing loss/accuracy of rolling and random masking under high data heterogeneity.



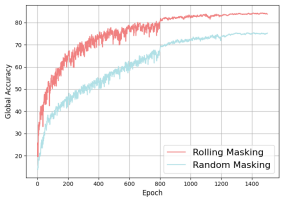
(a) CIFAR-10



(b) CIFAR-100



(c) CIFAR-10



(d) CIFAR-100

Figure 2: Global testing loss/accuracy of rolling and random masking with low data heterogeneity.

Table 1: Generalization of random masking and full model training under high data heterogeneity.

	Global Loss Diff.
Random masking	$\|0.9744 - 0.9618\| = 0.0126$
Full model	$\|0.7592 - 0.7402\| = 0.0190$
	Global Accuracy Diff.
Random masking	$\|66.124 - 65.74\| = 0.384$
Full model	$\|74.322 - 75.81\| = 1.488$

Table 2: Generalization of random masking and full model training under low data heterogeneity.

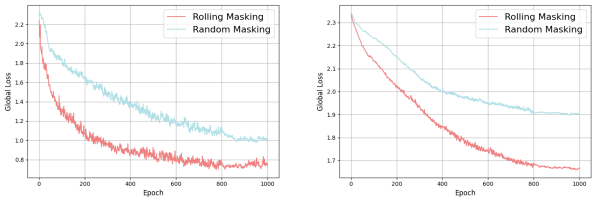
	Global Loss Diff.
Random masking	$\|0.501 - 0.5164\| = 0.0154$
Full model	$\|0.4379 - 0.471\| = 0.0330$
	Global Accuracy Diff.
Random masking	$\|83.014 - 82.92\| = 0.094$
Full model	$\|85.46 - 84.62\| = 0.840$

loss, while Fig. 2(c) and Fig. 2(d) show the global model testing accuracy, respectively. They demonstrate that the results under low data heterogeneity follow the high data heterogeneity.

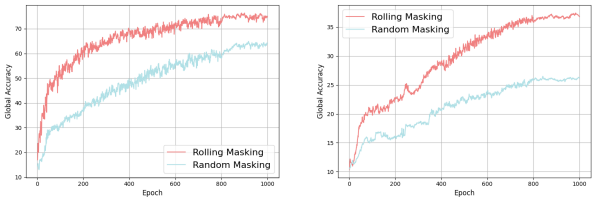
Model-Homogeneous Setting. We compare global model testing loss/accuracy for CIFAR-10 in two model-homogeneous cases: all clients have the largest capacity model (β=1) and all clients have the smallest capacity model (β=1/16), representing the upper and lower performance bounds. For the global testing loss, Fig. 3(a) and Fig. 3(b) show that in both cases, rolling masking outperforms random. For global testing accuracy, Fig. 3(c) and Fig. 3(d) are also consistent with the conclusion. Additionally, the largest model consistently achieves better performance than the smallest model. Similarly, the global testing loss shown in Fig. 4(a) and Fig. 4(b), and the global testing accuracy shown in Fig. 4(c) and Fig. 4(d) show that low data heterogeneity has the same conclusion as the high data heterogeneity in both scenarios.

5.3 Generalization of Masked Training

We evaluate the generalization of random masking and full model training (FedAvg) under both high and low data heterogeneity. To quantify generalization, we measure the gap between the global model's training and testing performance, considering both loss and accuracy. Specifically, we compute the difference between the global model's training loss (evaluated on local training data) and its test loss, as well as the difference between training accuracies (evaluated on local training data) and test accuracies for the global model. A smaller difference indicates better generalization and less overfitting to the local training distribution. Tab. 1 and Tab. 2 show that random masking achieves smaller global loss and accuracy differences compared to full model training, under both cases. It suggests that random masking not only mitigates the adverse impact of non-i.i.d. data distributions but also enhances the stability and robustness of the global model compared to FedAvg training.

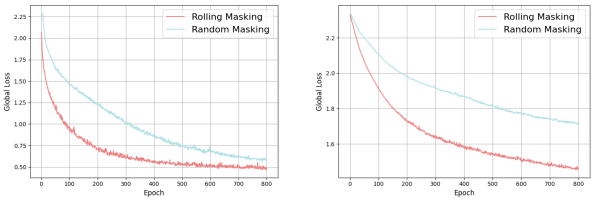


(a) Largest capacity (b) Smallest capacity

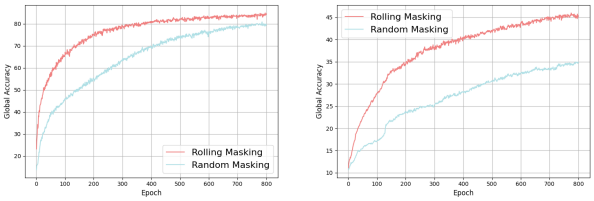


(c) Largest capacity (d) Smallest capacity

Figure 3: Global testing loss/accuracy of rolling and random masking under the largest and smallest client model capacity under high data heterogeneity.



(a) Largest Model Capacity (b) Smallest Model Capacity



(c) Largest Model Capacity (d) Smallest Model Capacity

Figure 4: Global testing loss/accuracy of rolling and random masking under the largest and smallest client model capacity under low data heterogeneity.

6 CONCLUSION

This paper provides a comprehensive analysis of sub-model training in federated learning, addressing a significant gap in the rigorous convergence analysis of this approach. We have established convergence bounds for both randomly selected sub-model training and the partitioned model variant, demonstrating the impact of masking probability on residual error. Additionally, our stability analysis reveals that sub-model training can enhance generalization by stabilizing the training process. The empirical results corroborate our theoretical findings, highlighting the effectiveness of sub-model training in improving on-device local training for large learning models.

Acknowledgment

This work was partially supported by NSF CAREER Award #2239374 NSF CNS Award #1956276.

References

Samiul Alam, Luyang Liu, Ming Yan, and Mi Zhang. Fedrolex: Model-heterogeneous federated learning with rolling sub-model extraction. *arXiv preprint arXiv:2212.01548*, 2022.

Hong-You Chen and Wei-Lun Chao. Fedbe: Making bayesian model ensemble applicable to federated learning. *arXiv preprint arXiv:2009.01974*, 2020.

Yae Jee Cho, Pranay Sharma, Gauri Joshi, Zheng Xu, Satyen Kale, and Tong Zhang. On the convergence of federated averaging with cyclic client participation. In *International Conference on Machine Learning*, pages 5677–5721. PMLR, 2023.

Yae Jee Cho, Luyang Liu, Zheng Xu, Aldi Fahrezi, and Gauri Joshi. Heterogeneous low-rank approximation for federated fine-tuning of on-device foundation models. *arXiv preprint arXiv:2401.06432*, 2024.

Yury Demidovich, Grigory Malinovsky, Egor Shulgin, and Peter Richtárik. Mast: Model-agnostic sparsified training. *arXiv preprint arXiv:2311.16086*, 2023.

Yuyang Deng, Mohammad Mahdi Kamani, Pouria Mahdavinia, and Mehrdad Mahdavi. Distributed personalized empirical risk minimization. *Advances in Neural Information Processing Systems*, 36, 2024.

Enmao Diao, Jie Ding, and Vahid Tarokh. Heterofl: Computation and communication efficient federated learning for heterogeneous clients. *arXiv preprint arXiv:2010.01264*, 2020.

Wenzhi Fang, Dong-Jun Han, and Christopher G. Brinton. Submodel partitioning in hierarchical federated learning: Algorithm design and convergence analysis. In *ICC 2024 - IEEE International Conference on Communications*, pages 268–273, 2024. doi: 10.1109/ICC51166.2024.10622512.

Zihao Fu, Haoran Yang, Anthony Man-Cho So, Wai Lam, Lidong Bing, and Nigel Collier. On the effectiveness of parameter-efficient fine-tuning. In *Proceedings of the AAAI conference on artificial intelligence*, volume 37, pages 12799–12807, 2023.

Saeed Ghadimi, Guanghui Lan, and Hongchao Zhang. Mini-batch stochastic approximation methods for non-convex stochastic composite optimization. *Mathematical Programming*, 155(1):267–305, 2016.

Neel Guha, Ameet Talwalkar, and Virginia Smith. One-shot federated learning. *arXiv preprint arXiv:1902.11175*, 2019.

Farzin Haddadpour and Mehrdad Mahdavi. On the convergence of local descent methods in federated learning. *arXiv preprint arXiv:1910.14425*, 2019.

Farzin Haddadpour, Mohammad Mahdi Kamani, Mehrdad Mahdavi, and Viveck Cadambe. Local sgd with periodic averaging: Tighter analysis and adaptive synchronization. In *Advances in Neural Information Processing Systems*, pages 11080–11092, 2019.

Nam Hyeon-Woo, Moon Ye-Bin, and Tae-Hyun Oh. Fedpara: Low-rank hadamard product for communication-efficient federated learning. *arXiv preprint arXiv:2108.06098*, 2021.

Yuang Jiang, Shiqiang Wang, Victor Valls, Bong Jun Ko, Wei-Han Lee, Kin K Leung, and Leandros Tassiulas. Model pruning enables efficient federated learning on edge devices. *IEEE Transactions on Neural Networks and Learning Systems*, 2022.

Ahmed Khaled, Konstantin Mishchenko, and Peter Richtárik. Tighter theory for local sgd on identical and heterogeneous data. In *AISTAT*, 2020.

Royson Lee, Javier Fernandez-Marques, Shell Xu Hu, Da Li, Stefanos Laskaridis, Łukasz Dudziak, Timothy Hospedales, Ferenc Huszár, and Nicholas Donald Lane. Recurrent early exits for federated learning with heterogeneous clients. In *Forty-first International Conference on Machine Learning*, 2024.

Daliang Li and Junpu Wang. Fedmd: Heterogenous federated learning via model distillation. *arXiv preprint arXiv:1910.03581*, 2019.

Xiang Li, Kaixuan Huang, Wenhao Yang, Shusen Wang, and Zhihua Zhang. On the convergence of fedavg on non-iid data. *arXiv preprint arXiv:1907.02189*, 2019.

Tao Lin, Lingjing Kong, Sebastian U Stich, and Martin Jaggi. Ensemble distillation for robust model fusion in federated learning. *Advances in Neural Information Processing Systems*, 33:2351–2363, 2020.

Brendan McMahan, Eider Moore, Daniel Ramage, Seth Hampson, and Blaise Aguera y Arcas. Communication-efficient learning of deep networks from decentralized data. In *Artificial Intelligence and Statistics*, pages 1273–1282, 2017.

Amirkeivan Mohtashami, Martin Jaggi, and Sebastian Stich. Masked training of neural networks with partial gradients. In *International Conference on Artificial Intelligence and Statistics*, pages 5876–5890. PMLR, 2022.

Zhefeng Qiao, Xianghao Yu, Jun Zhang, and Khaled B Letaief. Communication-efficient federated learning with dual-side low-rank compression. *arXiv preprint arXiv:2104.12416*, 2021.

Felix Sattler, Tim Korjakow, Roman Rischke, and Wojciech Samek. Fedaux: Leveraging unlabeled auxiliary data in federated learning. *IEEE Transactions on Neural Networks and Learning Systems*, 34(9):5531–5543, 2021.

Markus Schneider. Probability inequalities for kernel embeddings in sampling without replacement. In Arthur Gretton and Christian C. Robert, editors, *Proceedings of the 19th International Conference on Artificial Intelligence and Statistics*, volume 51 of *Proceedings of Machine Learning Research*, pages 66–74, Cadiz, Spain, 09–11 May 2016. PMLR. URL `https://proceedings.mlr.press/v51/schneider16.html`.

Egor Shulgin and Peter Richtárik. Towards a better theoretical understanding of independent subnetwork training. *arXiv preprint arXiv:2306.16484*, 2023.

Sebastian U Stich. Local sgd converges fast and communicates little. *arXiv preprint arXiv:1805.09767*, 2018.

Lichao Sun and Lingjuan Lyu. Federated model distillation with noise-free differential privacy. *arXiv preprint arXiv:2009.05537*, 2020.

Zhenyu Sun, Xiaochun Niu, and Ermin Wei. Understanding generalization of federated learning via stability: Heterogeneity matters. In *International Conference on Artificial Intelligence and Statistics*, pages 676–684. PMLR, 2024.

Blake Woodworth, Kumar Kshitij Patel, Sebastian U Stich, Zhen Dai, Brian Bullins, H Brendan McMahan, Ohad Shamir, and Nathan Srebro. Is local sgd better than minibatch sgd? *arXiv preprint arXiv:2002.07839*, 2020a.

Blake E Woodworth, Kumar Kshitij Patel, and Nati Srebro. Minibatch vs local sgd for heterogeneous distributed learning. *Advances in Neural Information Processing Systems*, 33:6281–6292, 2020b.

Feijie Wu, Xingchen Wang, Yaqing Wang, Tianci Liu, Lu Su, and Jing Gao. Fiarse: Model-heterogeneous federated learning via importance-aware submodel extraction. *arXiv preprint arXiv:2407.19389*, 2024.

Dezhong Yao, Wanning Pan, Michael J O'Neill, Yutong Dai, Yao Wan, Hai Jin, and Lichao Sun. Fedhm: Efficient federated learning for heterogeneous models via low-rank factorization. *arXiv preprint arXiv:2111.14655*, 2021.

Binhang Yuan, Cameron R Wolfe, Chen Dun, Yuxin Tang, Anastasios Kyrillidis, and Christopher M Jermaine. Distributed learning of deep neural networks using independent subnet training. *arXiv preprint arXiv:1910.02120*, 2019.

Honglin Yuan and Tengyu Ma. Federated accelerated stochastic gradient descent. *arXiv preprint arXiv:2006.08950*, 2020.

Hanhan Zhou, Tian Lan, Guru Prasadh Venkataramani, and Wenbo Ding. Every parameter matters: Ensuring the convergence of federated learning with dynamic heterogeneous models reduction. *Advances in Neural Information Processing Systems*, 36, 2024.

Zhuangdi Zhu, Junyuan Hong, and Jiayu Zhou. Data-free knowledge distillation for heterogeneous federated learning. In *International conference on machine learning*, pages 12878–12889. PMLR, 2021.

Checklist

1. For all models and algorithms presented, check if you include:
   (a) A clear description of the mathematical setting, assumptions, algorithm, and/or model. [Yes]
   (b) An analysis of the properties and complexity (time, space, sample size) of any algorithm. [Yes]
   (c) (Optional) Anonymized source code, with specification of all dependencies, including external libraries. [Not Applicable]

2. For any theoretical claim, check if you include:
   (a) Statements of the full set of assumptions of all theoretical results. [Yes]
   (b) Complete proofs of all theoretical results. [Yes]
   (c) Clear explanations of any assumptions. [Yes]

3. For all figures and tables that present empirical results, check if you include:
   (a) The code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL). [Not Applicable]
   (b) All the training details (e.g., data splits, hyperparameters, how they were chosen). [Yes]
   (c) A clear definition of the specific measure or statistics and error bars (e.g., with respect to the random seed after running experiments multiple times). [Yes]
   (d) A description of the computing infrastructure used. (e.g., type of GPUs, internal cluster, or cloud provider). [Not Applicable]

4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets, check if you include:
   (a) Citations of the creator If your work uses existing assets. [Yes]
   (b) The license information of the assets, if applicable. [Yes]
   (c) New assets either in the supplemental material or as a URL, if applicable. [Not Applicable]
   (d) Information about consent from data providers/curators. [Yes]
   (e) Discussion of sensible content if applicable, e.g., personally identifiable information or offensive content. [Not Applicable]

5. If you used crowdsourcing or conducted research with human subjects, check if you include:
   (a) The full text of instructions given to participants and screenshots. [Not Applicable]
   (b) Descriptions of potential participant risks, with links to Institutional Review Board (IRB) approvals if applicable. [Not Applicable]
   (c) The estimated hourly wage paid to participants and the total amount spent on participant compensation. [Not Applicable]

Supplementary Materials

APPENDIX

Organization The appendix is organized as follows:

- In Appendix A we discuss additional related works.
- In Appendix B we provide additional experimental results and setup details.
- In Appendix C we provide the proof of randomly masked FedAvg.
- In Appendix D we provide the proof of masking with rolling (FedRolex).
- In Appendix E we provide the proof of the stability of the masked training method.

A ADDITIONAL RELATED WORKS

Convergence analysis of FedAvg. FedAvg (or Local SGD) (McMahan et al., 2017) was used as a solution to reduce communication cost and protect user data privacy in distributed learning. FedAvg is firstly proposed by (McMahan et al., 2017) to alleviate communication bottleneck in the distributed machine learning. (Stich, 2018) was the first to prove that local SGD achieves $O\left(1/T\right)$ convergence rate with only $O(\sqrt{T})$ communication rounds on IID data for smooth strongly-convex loss functions. (Haddadpour et al., 2019) analyzed the convergence of local SGD on nonconvex (PL condition) function, and proposed an adaptive synchronization scheme. (Khaled et al., 2020) gave the tighter bound of local SGD, which directly reduces the $O(\sqrt{T})$ communication rounds in (Stich, 2018) to $O(N)$, under smooth strongly-convex setting. (Yuan and Ma, 2020) proposed the first accelerated local SGD, which further reduced the communication rounds to $O(N^{1/3})$. (Haddadpour and Mahdavi, 2019) gave the analysis of local GD and SGD on smooth nonconvex functions in non-IID setting. Li et al (Li et al., 2019) analyzed the convergence of FedAvg under non-IID data for strongly convex functions. (Woodworth et al., 2020a,b) investigated the difference between local SGD and mini-batch SGD, in both homogeneous and heterogeneous data settings.

Distributed/federated Sub-model training. Distributed/federated sub-model training is proposed to solve clients' insufficient computation and memory issue in federated learning, especially in this large language model era. (Diao et al., 2020) proposed the first federated learning algorithm with heterogeneous client model capacity. (Alam et al., 2022) propose FedRolex, which allows the clients to pick the suitable sub-model to optimize according to their capacity.

From the theoretical perspective, (Mohtashami et al., 2022) is the first to study the sub-model training where they considered the single machine setting, where $N = 1$, and proved that SGD with masked model parameter will converge to first order stationary point of the masked objective. (Shulgin and Richtárik, 2023) studied single machine sub-model training algorithm and considered the special scenario where $f_i(\mathbf{w})$ is quadratic, and proved that the convergence rate will suffer from a residual error unless the objective and sub-model sampling schema have some benign property. (Demidovich et al., 2023) studied similar algorithm, and proved convergence to the optimal point of the masked objective for general strongly convex losses. For nonconvex regime, (Zhou et al., 2024), (Fang et al., 2024) and (Wu et al., 2024) studied convergence of the distributed sub-model training algorithm with local updates on general nonconvex loss function, but their bound depends on the norm of history iterates, i.e., $\|\mathbf{w}_t\|$. When the model's norm is large, the convergence bound becomes vacuous.

Low-rank Federated Learning. Another line of works that reduce client computation/communication burden are low-rank federated learning (Qiao et al., 2021; Hyeon-Woo et al., 2021; Yao et al., 2021; Cho et al., 2024). (Qiao et al., 2021) propose FedDLR, where the clients only send low-rank model to server at communication stage.

Server then performs averaging, decomposes the averaged model into low-rank version again, and sends them back to clients. They also propose an adaptive rank selection which boosts the performance. (Hyeon-Woo et al., 2021) proposed FedPara, where the clients directly optimize on low-rank models instead of full model, to meet the local computation and memory constraints.

Other Model Heterogeneous Federated Learning. Besides sub-model and low-rank federated learning, there is also a body of works for model heterogeneous federated learning, via knowledge distillation (Zhu et al., 2021; Lin et al., 2020; Sun and Lyu, 2020; Guha et al., 2019; Chen and Chao, 2020; Li and Wang, 2019; Sattler et al., 2021). (Li and Wang, 2019) proposed the first knowledge distillation based method for model heterogeneous federated learning, where they leveraged a public dataset to compute the 'consensus', and force each client's model to behave close to this consensus, to share knowledge among clients. (Zhu et al., 2021) proposed the first data-free distillation method for federated learning, where they try to learn a data generator to generate synthetic clients' data for model distillation.

B EXPERIMENT DETAILS

We provide the data settings in Tab. 3 and experimental results under low data heterogeneity in this section.

Table 3: Dataset Description

Dataset	# of Training Clients	# of Training Examples	# of Testing Examples
CIFAR-10	100	50,000	10,000
CiFAR-100	100	50,000	10,000

The experiments are conducted on 2 NVIDIA 6000 GPUs. To compare ours with HeteroFL (Diao et al., 2020), which is referred to as static masking in our experiments. We conducted experiments using CIFAR-100 with high data heterogeneity and the ResNet18 model, and client model capacities are 1/4 and 1/8 of the full model size. Client selection is 10% of clients are randomly selected from a pool of 100 clients per global epoch. The results are shown in Tab. 4, indicating that ours outperforms HeteroFL.

Table 4: Comparisons of global model performance with HeteroFL using CIFAR-100 with ResNet18.

Method	Global Accuracy	Global Loss
Static Masking (HeteroFL)	34.37%	2.5026
Roll Masking (Ours)	35.30%	2.4681

Communication Efficiency and Computational Cost Analysis. We provide measurements of communication cost per epoch and GPU wall-time per epoch across CIFAR-100 and CIFAR-10 under model-heterogeneous and model-homogeneous settings.

We observe that in the model-heterogeneous settings, as shown in Tab. 5 and Tab. 6, rolling Masking consistently communicates fewer parameters than Random Masking while also achieving higher testing accuracy, demonstrating a superior communication-accuracy trade-off. Moreover, in both model-heterogeneous and model-homogeneous settings (Tab. 7 and Tab. 8), rolling Masking yields lower GPU walltime per epoch, indicating that it is computationally more efficient than Random Masking.

C PROOF OF CONVERGENCE OF RANDOMLY MASKED FEDAVG

In this section, we provide detailed proofs for the results and theorems on sub-model training with random masking omitted from the main body. We begin by outlining several general results that serve as helper for the main proofs, and then provide the proofs of Theorem 1 (strongly convex setting) and Theorem 2 (nonconvex setting) in Subsection C.1 and Subsection C.2, respectively.

Table 5: Comparisons of communication cost and GPU walltime on CIFAR-100 under the model-heterogeneous setting with low data heterogeneity.

Method	Comm. Cost/Epoch	GPU Walltime/Epoch
Random Masking (low data heterogeneity)	0.389838 (model rate)	0:00:09.161714
Rolling Masking (low data heterogeneity)	0.388406 (model rate)	0:00:08.922342

Table 6: Comparisons of communication cost and GPU walltime on CIFAR-10 under the model-heterogeneous setting with high data heterogeneity.

Method	Comm. Cost/Epoch	GPU Walltime/Epoch
Random Masking (high data heterogeneity)	0.392070 (model rate)	0:00:09.700524
Rolling Masking (high data heterogeneity)	0.382314 (model rate)	0:00:09.633329

C.1 Proof of Convex Setting

In this subsection, we are going to prove Theorem 1. We begin with a high-level sketch of the proof to briefly illustrate our strategy before presenting the detailed argument. Consider an alternative objective induced by masking:

$$F_{\mathbf{p}}(\mathbf{w}) = \frac{1}{N}\sum\nolimits_{i=1}^{N} \mathbb{E}_{\mathbf{m}_i \sim Ber(p_i)}[f_i(\mathbf{m}_i \odot \mathbf{w})], \tag{2}$$

where $\mathbf{p} = [p_1, \ldots, p_N]$ and define $\mathbf{w}^*(\mathbf{p}) := \arg\min_{\mathbf{w}\in\mathcal{W}} F_{\mathbf{p}}(\mathbf{w})$ as the optimal model given $\mathbf{p}$. Apparently, when $\mathbf{p} = \mathbf{1}$, $F_{\mathbf{p}}(\mathbf{w})$ becomes original objective $F(\mathbf{w})$. Our proof relies on a key Lipschitz property of $\mathbf{w}^*(\mathbf{p})$. That is, if each $f_i(\mathbf{w})$ is strongly convex and with bounded gradient, then

$$\|\mathbf{w}^*(\mathbf{p}) - \mathbf{w}^*(\mathbf{1})\| \leq c \cdot \|\mathbf{p} - \mathbf{1}\|,$$

for some constant c depending on G, μ and $\mathbf{p}$. As a result, we can decompose the objective value into (1) convergence error of the sub-model training to $\mathbf{w}^*(\mathbf{p})$ and (2) residual error due to masking:

$$F(\hat{\mathbf{w}}) - F(\mathbf{w}^*) \leq \frac{5}{2}L\|\tilde{\mathbf{w}} - \mathbf{w}^*(\mathbf{p})\|^2 + \left(\frac{5L}{2\bar{\mu}} + \frac{4}{L}\right)\frac{2G^2 + 2W^2L^2}{N}\sum_{i=1}^{N} d(1-p_i).$$

It remains to prove the convergence of Algorithm 1 to $\mathbf{w}^*(\mathbf{p})$, which can be achieved by standard Local SGD analysis.

We now proceed to the formal proof, making each step of the argument precise. We start by showing the strong convexity of the alternative objective induced in Eq. 2.

Proposition 1. *$F_{\mathbf{p}}(\mathbf{w})$ is $\mu_{\mathbf{p}} := \frac{1}{N}\sum_{i=1}^{N} p_i\mu$ strongly convex, and $L_{\mathbf{p}} := \frac{1}{N}\sum_{i=1}^{N} p_i L$ smooth.*

Proof. The proof mainly follows Lemmas 9 and 11 in (Demidovich et al., 2023). We first examine the smoothness

Table 7: Comparisons of communication cost and GPU walltime on CIFAR-10 under the model-homogeneous setting with low data heterogeneity and the largest model size.

Method	Comm. Cost/Epoch	GPU Walltime/Epoch
Random Masking (low data heterogeneity)	1.0 (model rate)	0:00:10.088458
Rolling Masking (low data heterogeneity)	1.0 (model rate)	0:00:09.967106

Table 8: Comparisons of communication cost and GPU walltime on CIFAR-10 under the model-homogeneous setting with high data heterogeneity and the smallest model size.

Method	Comm. Cost/Epoch	GPU Walltime/Epoch
Random Masking (high data heterogeneity)	0.0625 (model rate)	0:00:05.758695
Rolling Masking (high data heterogeneity)	0.0625 (model rate)	0:00:05.757487

and convexity parameter of $\tilde{F}$. For smoothness:

$$
\begin{aligned}
F_{\mathbf{p}}(\mathbf{w}+\Delta\mathbf{w}) &= \frac{1}{N}\sum_{i=1}^{N}\mathbb{E}_{\mathbf{m}_i\sim Ber(p_i)} f_i(\mathbf{m}_i\odot(\mathbf{w}+\Delta\mathbf{w})) \\
&\leq \frac{1}{N}\sum_{i=1}^{N}\mathbb{E}_{\mathbf{m}_i\sim Ber(p_i)}\left(f_i(\mathbf{m}_i\odot\mathbf{w})+\langle\mathbf{m}_i\odot\Delta\mathbf{w},\,\nabla f_i(\mathbf{m}_i\odot\mathbf{w})\rangle+\frac{L}{2}\left\|\mathbf{m}_i\odot\Delta\mathbf{w}\right\|^2\right) \\
&= F_{\mathbf{p}}(\mathbf{w})+\left\langle\Delta\mathbf{w},\,\frac{1}{N}\sum_{i=1}^{N}\mathbb{E}_{\mathbf{m}_i\sim Ber(p_i)}\mathbf{m}_i\odot\nabla f_i(\mathbf{m}_i\odot\mathbf{w})\right\rangle \\
&\quad+\frac{L}{2}\frac{1}{N}\sum_{i=1}^{N}\mathbb{E}_{\mathbf{m}_i\sim Ber(p_i)}\left\langle(\mathbf{M}_i)^2\Delta\mathbf{w},\,\Delta\mathbf{w}\right\rangle \\
&= F_{\mathbf{p}}(\mathbf{w})+\langle\Delta\mathbf{w},\,\nabla F_{\mathbf{p}}(\mathbf{w})\rangle+\frac{1}{2}\frac{1}{N}\sum_{i=1}^{N}p_iL\left\|\Delta\mathbf{w}\right\|^2,
\end{aligned}
$$

where $\mathbf{M}_i$ is the diagonal matrix with $\mathbf{m}_i$ is its diagonal entries.

For convexity we have:

$$
\begin{aligned}
F_{\mathbf{p}}(\mathbf{w}+\Delta\mathbf{w}) &= \frac{1}{N}\sum_{i=1}^{N}\mathbb{E}_{\mathbf{m}_i\sim Ber(p_i)} f_i(\mathbf{m}_i\odot(\mathbf{w}+\Delta\mathbf{w})) \\
&\geq \frac{1}{N}\sum_{i=1}^{N}\mathbb{E}_{\mathbf{m}_i\sim Ber(p_i)}\left(f_i(\mathbf{m}_i\odot\mathbf{w})+\langle\mathbf{m}_i\odot\Delta\mathbf{w},\,\nabla f_i(\mathbf{m}_i\odot\mathbf{w})\rangle+\frac{\mu}{2}\left\|\mathbf{m}_i\odot\Delta\mathbf{w}\right\|^2\right) \\
&= F_{\mathbf{p}}(\mathbf{w})+\left\langle\Delta\mathbf{w},\,\frac{1}{N}\sum_{i=1}^{N}\mathbb{E}_{\mathbf{m}_i\sim Ber(p_i)}\mathbf{m}_i\odot\nabla f_i(\mathbf{m}_i\odot\mathbf{w})\right\rangle \\
&\quad+\frac{\mu}{2}\frac{1}{N}\sum_{i=1}^{N}\mathbb{E}_{\mathbf{m}_i\sim Ber(p_i)}\left\langle(\mathbf{M}_i)^2\Delta\mathbf{w},\,\Delta\mathbf{w}\right\rangle \\
&= F_{\mathbf{p}}(\mathbf{w})+\langle\Delta\mathbf{w},\,\nabla F_{\mathbf{p}}(\mathbf{w})\rangle+\frac{\mu}{2}\frac{1}{N}\sum_{i=1}^{N}p_i\left\|\Delta\mathbf{w}\right\|^2.
\end{aligned}
$$

So $F_{\mathbf{p}}$ is $\mu_{\mathbf{p}} := \frac{1}{N}\sum_{i=1}^{N}p_i\mu$ strongly convex and $L_{\mathbf{p}} := \frac{1}{N}\sum_{i=1}^{N}p_iL$ smooth. □

Lemma 2. *Given a N-dimensional vector $\mathbf{p} \in [0,1]^N$, we define $\mathbf{w}^*(\mathbf{p}) := \arg\min_{\mathbf{w}\in\mathcal{W}} \Big\{\Phi(\mathbf{p},\mathbf{w}) := \frac{1}{N}\sum_{i=1}^N \mathbb{E}_{\mathbf{m}_i\sim Ber(p_i)} f_i(\mathbf{m}_i \odot \mathbf{w})\Big\}$. Then the following statement holds:*

$$\|\mathbf{w}^*(\mathbf{1}) - \mathbf{w}^*(\mathbf{p})\| \leq \sqrt{\frac{2G^2+2W^2L^2}{\mu_{\mathbf{p}}^2 N}}\sqrt{\sum_{i=1}^N d(1-p_i)}.$$

Proof. First, according to optimality conditions we have:

$$\langle \mathbf{w} - \mathbf{w}^*(\mathbf{p}), \nabla_2\Phi(\mathbf{p}, \mathbf{w}^*(\mathbf{p}))\rangle \geq 0,$$
$$\langle \mathbf{w} - \mathbf{w}^*(\mathbf{1}), \nabla_2\Phi(\mathbf{1}, \mathbf{w}^*(\mathbf{1}))\rangle \geq 0$$

Substituting $\mathbf{w}$ with $\mathbf{w}^*(\mathbf{1})$ and $\mathbf{w}^*(\mathbf{p})$ in the above first and second inequalities respectively yields:

$$\langle \mathbf{w}^*(\mathbf{1}) - \mathbf{w}^*(\mathbf{p}), \nabla_2\Phi(\mathbf{p}, \mathbf{w}^*(\mathbf{p}))\rangle \geq 0,$$
$$\langle \mathbf{w}^*(\mathbf{p}) - \mathbf{w}^*(\mathbf{1}), \nabla_2\Phi(\mathbf{1}, \mathbf{w}^*(\mathbf{1}))\rangle \geq 0.$$

Adding up the above two inequalities yields:

$$\langle \mathbf{w}^*(\mathbf{1}) - \mathbf{w}^*(\mathbf{p}), \nabla_2\Phi(\mathbf{p}, \mathbf{w}^*(\mathbf{p})) - \nabla_2\Phi(\mathbf{1}, \mathbf{w}^*(\mathbf{1}))\rangle \geq 0, \tag{3}$$

Since $\Phi(\mathbf{p},\cdot)$ is $\mu_{\mathbf{p}}$ strongly convex, as shown in Proposition 1, we have:

$$\langle \mathbf{w}^*(\mathbf{1}) - \mathbf{w}^*(\mathbf{p}), \nabla_2\Phi(\mathbf{p}, \mathbf{w}^*(\mathbf{1})) - \nabla_2\Phi(\mathbf{p}, \mathbf{w}^*(\mathbf{p}))\rangle \geq \mu_{\mathbf{p}}\|\mathbf{w}^*(\mathbf{1}) - \mathbf{w}^*(\mathbf{p})\|^2. \tag{4}$$

Adding up (5) and (4) yields:

$$\langle \mathbf{w}^*(\mathbf{1}) - \mathbf{w}^*(\mathbf{p}), \nabla_2\Phi(\mathbf{p}, \mathbf{w}^*(\mathbf{1})) - \nabla_2\Phi(\mathbf{1}, \mathbf{w}^*(\mathbf{1}))\rangle \geq \mu_{\mathbf{p}}\|\mathbf{w}^*(\mathbf{1}) - \mathbf{w}^*(\mathbf{p})\|^2.$$

Now we examine the smoothness of $\nabla_2\Phi(\mathbf{p},\mathbf{w})$ in terms of the first variable.

$$\begin{aligned}
\|\nabla_2\Phi(\mathbf{p},\mathbf{w}) - \nabla_2\Phi(\mathbf{1},\mathbf{w})\|^2 &= \left\|\frac{1}{N}\sum_{i=1}^N \mathbb{E}_{\mathbf{m}_i\sim Ber(p_i)}\left[\mathbf{m}_i \odot \nabla f_i(\mathbf{m}_i\odot\mathbf{w})\right] - \frac{1}{N}\sum_{i=1}^N \mathbf{1}\odot\nabla f_i(\mathbf{1}\odot\mathbf{w})\right\|^2 \\
&\leq \frac{1}{N}\sum_{i=1}^N \left\|\mathbb{E}_{\mathbf{m}_i\sim Ber(p_i)}\left[\mathbf{m}_i \odot \nabla f_i(\mathbf{m}_i\odot\mathbf{w})\right] - \mathbf{1}\odot\nabla f_i(\mathbf{1}\odot\mathbf{w})\right\|^2 \\
&\leq \frac{1}{N}\sum_{i=1}^N \left(2G^2\mathbb{E}_{\mathbf{m}_i\sim Ber(p_i)}\|\mathbf{m}_i - \mathbf{1}\|^2 + 2W^2L^2\mathbb{E}_{\mathbf{m}_i\sim Ber(p_i)}\|\mathbf{m}_i-\mathbf{1}\|^2\right) \\
&= \frac{2G^2+2W^2L^2}{N}\sum_{i=1}^N d(1-p_i)
\end{aligned} \tag{5}$$

Finally, putting pieces together will conclude the proof:

$$\begin{aligned}
&\sqrt{\frac{2G^2+2W^2L^2}{N}}\|\mathbf{w}^*(\mathbf{p}) - \mathbf{w}^*(\mathbf{1})\|\sqrt{\sum_{i=1}^N d(1-p_i)} \geq \mu_{\mathbf{p}}\|\mathbf{w}^*(\mathbf{p}) - \mathbf{w}^*(\mathbf{1})\|^2 \\
\iff &\sqrt{\frac{2G^2+2W^2L^2}{\mu_{\mathbf{p}}^2 N}}\sqrt{\sum_{i=1}^N d(1-p_i)} \geq \|\mathbf{w}^*(\mathbf{p}) - \mathbf{w}^*(\mathbf{1})\|.
\end{aligned}$$

□

Lemma 3 (Optimality Gap)**.** *Let $\Phi(\mathbf{p},\mathbf{w}) := \frac{1}{N}\sum_{i=1}^N \mathbb{E}_{\mathbf{m}_i\sim Ber(p_i)} f_i(\mathbf{m}_i\odot\mathbf{w})$. Let $\hat{\mathbf{w}} = \mathcal{P}_{\mathcal{W}}(\tilde{\mathbf{w}} - \frac{1}{L}\nabla_{\mathbf{w}}\Phi(\mathbf{p},\tilde{\mathbf{w}}))$. If we assume each f_i is L-smooth, μ-strongly convex and with gradient bounded by G, then the following statement holds true:*

$$\mathbb{E}[F(\hat{\mathbf{w}}) - F(\mathbf{w}^*)] \leq \frac{5}{2}L\mathbb{E}\|\tilde{\mathbf{w}} - \mathbf{w}^*(\mathbf{p})\|^2 + \left(\frac{5L}{2\mu_{\mathbf{p}}} + \frac{4}{L}\right)\frac{2G^2+2W^2L^2}{N}\sum_{i=1}^N d(1-p_i),$$

where $\mathbf{w}^* = \arg\min_{\mathbf{w}\in\mathcal{W}} F(\mathbf{w})$.

Proof. Define $\hat{\nabla}_{\mathbf{w}}\Phi(\mathbf{p},\mathbf{w}) := \frac{1}{N}\sum_{i=1}^{N}\mathbf{m}_i \odot \nabla f_i(\mathbf{m}_i \odot \mathbf{w})$ According to property of projection, we have:

$$\begin{aligned}
0 &\leq \left\langle \mathbf{w}-\hat{\mathbf{w}},\, L(\hat{\mathbf{w}}-\tilde{\mathbf{w}}) + \hat{\nabla}_{\mathbf{w}}\Phi(\mathbf{p},\tilde{\mathbf{w}})\right\rangle \\
&= \underbrace{\langle \mathbf{w}-\hat{\mathbf{w}},\, L(\hat{\mathbf{w}}-\tilde{\mathbf{w}}) + \nabla_{\mathbf{w}}\Phi(\mathbf{1},\tilde{\mathbf{w}})\rangle}_{T_1} + \underbrace{\left\langle \mathbf{w}-\hat{\mathbf{w}},\, \hat{\nabla}_{\mathbf{w}}\Phi(\mathbf{p},\tilde{\mathbf{w}}) - \nabla_{\mathbf{w}}\Phi(\mathbf{1},\tilde{\mathbf{w}})\right\rangle}_{T_2}.
\end{aligned}$$

For T_1, we notice:

$$\begin{aligned}
&\langle \mathbf{w}-\hat{\mathbf{w}},\, L(\hat{\mathbf{w}}-\tilde{\mathbf{w}}) + \nabla_{\mathbf{w}}\Phi(\mathbf{1},\tilde{\mathbf{w}})\rangle \\
&= L\langle \mathbf{w}-\tilde{\mathbf{w}},\, \hat{\mathbf{w}}-\tilde{\mathbf{w}}\rangle + L\langle \tilde{\mathbf{w}}-\hat{\mathbf{w}},\, \hat{\mathbf{w}}-\tilde{\mathbf{w}}\rangle + \langle \mathbf{w}-\hat{\mathbf{w}},\, \nabla_{\mathbf{w}}\Phi(\mathbf{1},\tilde{\mathbf{w}})\rangle \\
&= L\langle \mathbf{w}-\tilde{\mathbf{w}},\, \hat{\mathbf{w}}-\tilde{\mathbf{w}}\rangle - L\left\|\tilde{\mathbf{w}}-\hat{\mathbf{w}}\right\|^2 + \langle \mathbf{w}-\hat{\mathbf{w}}_i,\, \nabla_{\mathbf{w}}\Phi(\mathbf{1},\tilde{\mathbf{w}})\rangle \\
&\leq L(\left\|\mathbf{w}-\tilde{\mathbf{w}}\right\|^2 + \frac{1}{4}\left\|\hat{\mathbf{w}}-\tilde{\mathbf{w}}\right\|^2) - L\left\|\tilde{\mathbf{w}}-\hat{\mathbf{w}}\right\|^2 + \underbrace{\langle \mathbf{w}-\hat{\mathbf{w}},\, \nabla_{\mathbf{w}}\Phi(\mathbf{1},\tilde{\mathbf{w}})\rangle}_{\spadesuit}
\end{aligned}$$

where at last step we used Young's inequality. To bound $\spadesuit$, we apply the L smoothness and μ strongly convexity of $\Phi(\mathbf{1},\cdot)$:

$$\begin{aligned}
\langle \mathbf{w}-\hat{\mathbf{w}},\, \nabla_{\mathbf{w}}\Phi(\mathbf{1},\tilde{\mathbf{w}})\rangle &= \langle \mathbf{w}-\tilde{\mathbf{w}},\, \nabla_{\mathbf{w}}\Phi(\mathbf{1},\tilde{\mathbf{w}})\rangle + \langle \tilde{\mathbf{w}}-\hat{\mathbf{w}},\, \nabla_{\mathbf{w}}\Phi(\mathbf{1},\tilde{\mathbf{w}})\rangle \\
&\leq \Phi(\mathbf{1},\mathbf{w}) - \Phi(\mathbf{1},\tilde{\mathbf{w}}) - \frac{\mu}{2}\left\|\tilde{\mathbf{w}}-\mathbf{w}\right\|^2 + \Phi(\mathbf{1},\tilde{\mathbf{w}}) - \Phi(\mathbf{1},\hat{\mathbf{w}}) + \frac{L}{2}\left\|\tilde{\mathbf{w}}-\hat{\mathbf{w}}\right\|^2 \\
&\leq \Phi(\mathbf{1},\mathbf{w}) - \Phi(\mathbf{1},\hat{\mathbf{w}}) - \frac{\mu}{2}\left\|\tilde{\mathbf{w}}-\mathbf{w}\right\|^2 + \frac{L}{2}\left\|\tilde{\mathbf{w}}-\hat{\mathbf{w}}\right\|^2
\end{aligned}$$

Putting above bound back yields:

$$T_1 = \langle \mathbf{w}-\hat{\mathbf{w}},\, L(\hat{\mathbf{w}}-\tilde{\mathbf{w}}) + \nabla_{\mathbf{w}}\Phi(\mathbf{1},\tilde{\mathbf{w}})\rangle \leq \Phi(\mathbf{1},\mathbf{w}) - \Phi(\mathbf{1},\hat{\mathbf{w}}) + L\left\|\tilde{\mathbf{w}}-\mathbf{w}\right\|^2 - \frac{L}{4}\left\|\tilde{\mathbf{w}}-\hat{\mathbf{w}}\right\|^2.$$

Now we switch to bounding T_2. Applying Cauchy-Schwartz yields:

$$\left\langle \mathbf{w}-\hat{\mathbf{w}},\, \hat{\nabla}_{\mathbf{w}}\Phi(\mathbf{p},\tilde{\mathbf{w}}) - \nabla_{\mathbf{w}}\Phi(\mathbf{1},\tilde{\mathbf{w}})\right\rangle \leq \frac{L}{4}\left\|\mathbf{w}-\tilde{\mathbf{w}}\right\|^2 + \frac{L}{4}\left\|\tilde{\mathbf{w}}-\hat{\mathbf{w}}\right\|^2 + \frac{4}{L}\left\|\hat{\nabla}_{\mathbf{w}}\Phi(\mathbf{p},\tilde{\mathbf{w}}) - \nabla_{\mathbf{w}}\Phi(\mathbf{1},\tilde{\mathbf{w}})\right\|^2$$

To bound $\left\|\hat{\nabla}_{\mathbf{w}}\Phi(\mathbf{p},\tilde{\mathbf{w}}) - \nabla_{\mathbf{w}}\Phi(\mathbf{1},\tilde{\mathbf{w}})\right\|^2$, we follow the same steps in (5):

$$\begin{aligned}
\left\|\hat{\nabla}_2\Phi(\mathbf{p},\tilde{\mathbf{w}}) - \nabla_2\Phi(\mathbf{1},\tilde{\mathbf{w}})\right\|^2 &= \left\|\frac{1}{N}\sum_{i=1}^{N}\mathbf{m}_i \odot \nabla f_i(\mathbf{m}_i\odot\tilde{\mathbf{w}}) - \frac{1}{N}\sum_{i=1}^{N}\mathbf{1}\odot\nabla f_i(\mathbf{1}\odot\tilde{\mathbf{w}})\right\|^2 \\
&\leq \frac{1}{N}\sum_{i=1}^{N}\left\|\mathbf{m}_i\odot\nabla f_i(\mathbf{m}_i\odot\tilde{\mathbf{w}}) - \mathbf{1}\odot\nabla f_i(\mathbf{1}\odot\tilde{\mathbf{w}})\right\|^2 \\
&\leq \frac{1}{N}\sum_{i=1}^{N}\left(2G^2\left\|\mathbf{m}_i-\mathbf{1}\right\|^2 + 2W^2L^2\left\|\mathbf{m}_i-\mathbf{1}\right\|^2\right) \\
&= \frac{2G^2+2W^2L^2}{N}\sum_{i=1}^{N}\left\|\mathbf{m}_i-\mathbf{1}\right\|^2. \qquad (6)
\end{aligned}$$

Putting pieces together yields:

$$0 \leq \Phi(\mathbf{1},\mathbf{w}) - \Phi(\mathbf{1},\hat{\mathbf{w}}) + \frac{5L}{4}\left\|\tilde{\mathbf{w}}-\mathbf{w}\right\|^2 + \frac{4}{L}\frac{2G^2+2W^2L^2}{N}\sum_{i=1}^{N}\left\|\mathbf{m}_i-\mathbf{1}\right\|^2.$$

Re-arranging terms and setting $\mathbf{w} = \mathbf{w}^*(\mathbf{1}) = \arg\min_{\mathbf{w}\in\mathcal{W}} \Phi(\mathbf{1}, \mathbf{w})$ yields:

$$\Phi(\mathbf{1}, \hat{\mathbf{w}}) - \Phi(\mathbf{1}, \mathbf{w}^*(\mathbf{1})) \leq \frac{5L}{4} \|\tilde{\mathbf{w}} - \mathbf{w}^*\|^2 + \frac{4}{L} \frac{2G^2 + 2W^2L^2}{N} \sum_{i=1}^{N} \|\mathbf{m}_i - \mathbf{1}\|^2 .$$

Taking expectation over randomness of $\mathbf{m}_i$ yields

$$\mathbb{E}[\Phi(\mathbf{1}, \hat{\mathbf{w}}) - \Phi(\mathbf{1}, \mathbf{w}^*(\mathbf{1}))] \leq \frac{5L}{4} \mathbb{E} \|\tilde{\mathbf{w}} - \mathbf{w}^*\|^2 + \frac{4}{L} \frac{2G^2 + 2W^2L^2}{N} \sum_{i=1}^{N} d(1 - p_i).$$

At last, due to the Lipschitzness property of of $\mathbf{w}^*(\cdot)$ as shown in Lemma 2, it follows that:

$$\begin{aligned}
\frac{5L}{4}\mathbb{E}\|\tilde{\mathbf{w}} - \mathbf{w}^*(\mathbf{1})\|^2 &\leq \frac{5L}{2}\mathbb{E}\|\tilde{\mathbf{w}} - \mathbf{w}^*(\mathbf{p})\|^2 + \frac{5L}{2}\mathbb{E}\|\mathbf{w}^*(\mathbf{p}) - \mathbf{w}^*(\mathbf{1})\|^2 \\
&\leq \frac{5L}{2}\mathbb{E}\|\tilde{\mathbf{w}} - \mathbf{w}^*(\mathbf{p})\|^2 + \frac{5L}{2}\frac{2G^2 + 2W^2L^2}{\mu_{\mathbf{p}}^2 N} \sum_{i=1}^{N} d(1 - p_i),
\end{aligned}$$

as desired.

□

Next we are going to present technical lemmas for proving convergence of Algorithm 1 to $\mathbf{w}^*(\mathbf{p})$. For notational convenience, we drop the subscript and use $\mathbf{w}^*$ to denote $\mathbf{w}^*(\mathbf{p})$, and we define $\tilde{f}_i(\mathbf{w}) := \mathbb{E}_{\mathbf{m}_i \sim Ber(p_i)}[f_i(\mathbf{m}_i \odot \mathbf{w})]$. We define virtual local iterates $\tilde{\mathbf{w}}_{r,k}^i$ be such that,for $j \in \text{supp}(\mathbf{m}_i^r)$, $\tilde{\mathbf{w}}_{r,k}^i[j] = \mathbf{w}_{r,k}^i[j]$; for $j \notin \text{supp}(\mathbf{m}_i^r)$, we set $\tilde{\mathbf{w}}_{r,k}^i[j] = \mathbf{w}_r[j]$. An important property is that, $\mathbf{m}_i^r \odot \nabla \tilde{f}_i(\mathbf{m}_i^r \odot \tilde{\mathbf{w}}_{r,k}^i; \xi) = \mathbf{m}_i^r \odot \nabla \tilde{f}_i(\mathbf{m}_i^r \odot \mathbf{w}_{r,k}^i; \xi)$. Hence, the local updates can be equivalently viewed as conduct on the gradients queried on $\tilde{\mathbf{w}}_{r,k}^i$, i.e.,

$$\mathbf{w}_{r+1} = \mathcal{P}_{\mathcal{W}} \left(\mathbf{w}_r - \eta \sum_{k=0}^{K-1} \frac{1}{N} \sum_{i=1}^{N} \tilde{\mathbf{g}}_{r,k}^i \right)$$

where $\tilde{\mathbf{g}}_{r,k}^i = \mathbf{m}_i \odot \nabla f_i(\mathbf{m}_i \odot \tilde{\mathbf{w}}_{r,k}^i; \xi_{r,k}^i)$.

Lemma 4. *For Algorithm 1, under the condition of Theorem 1, the following statement holds true:*

$$\begin{aligned}
\mathbb{E} \|\mathbf{w}_{r+1} - \mathbf{w}^*\|^2 \leq (1 - \tilde{\mu}\eta)\mathbb{E} \|\mathbf{w}_r - \mathbf{w}^*\|^2 - \frac{1}{2}\eta K \frac{1}{N} \sum_{i=1}^{N} \left(\tilde{f}_i(\mathbf{w}_r) - \tilde{f}_i(\mathbf{w}^*) \right) \\
+ 2\eta \tilde{L} \frac{1}{N} \sum_{i=1}^{N} \sum_{k=0}^{K-1} \mathbb{E} \left\| \tilde{\mathbf{w}}_{r,k}^i - \mathbf{w}_r \right\|^2 + \frac{K\eta^2\delta^2}{N}.
\end{aligned}$$

Proof. According to updating rule we have:

$$
\begin{aligned}
\mathbb{E}\left\|\mathbf{w}_{r+1}-\mathbf{w}^*\right\|^2 &= \mathbb{E}\left\|\mathbf{w}_r-\eta\sum_{k=0}^{K-1}\frac{1}{N}\sum_{i=1}^{N}\tilde{\mathbf{g}}_{r,k}^i-\mathbf{w}^*\right\|^2 \\
&= \mathbb{E}\left\|\mathbf{w}_t-\mathbf{w}^*\right\|^2-\mathbb{E}\left\langle\eta\sum_{k=0}^{K-1}\frac{1}{N}\sum_{i=1}^{N}\tilde{\mathbf{g}}_{r,k}^i,\ \mathbf{w}_t-\mathbf{w}^*\right\rangle+\mathbb{E}\left\|\eta\sum_{k=0}^{K-1}\frac{1}{N}\sum_{i=1}^{N}\tilde{\mathbf{g}}_{r,k}^i\right\|^2 \\
&= \mathbb{E}\left\|\mathbf{w}_r-\mathbf{w}^*\right\|^2+\mathbb{E}\left\|\eta\sum_{k=0}^{K-1}\frac{1}{N}\sum_{i=1}^{N}\tilde{\mathbf{g}}_{r,k}^i\right\|^2 \\
&\quad-\mathbb{E}\left\langle\eta\sum_{k=0}^{K-1}\frac{1}{N}\sum_{i=1}^{N}\mathbf{m}_i^r\odot\nabla f_i(\mathbf{m}_i^r\odot\tilde{\mathbf{w}}_t^i),\ \mathbf{w}_r-\mathbf{w}^*\right\rangle \\
&= \mathbb{E}\left\|\mathbf{w}_r-\mathbf{w}^*\right\|^2+\mathbb{E}\left\|\eta\sum_{k=0}^{K-1}\frac{1}{N}\sum_{i=1}^{N}\tilde{\mathbf{g}}_{r,k}^i\right\|^2 \\
&\quad-\mathbb{E}\left\langle\eta\sum_{k=0}^{K-1}\frac{1}{N}\sum_{i=1}^{N}\mathbf{m}_i^r\odot\nabla f_i(\mathbf{m}_i^r\odot\tilde{\mathbf{w}}_{r,k}^i),\ \mathbf{w}_r-\mathbf{w}^*\right\rangle \\
&= \mathbb{E}\left\|\mathbf{w}_r-\mathbf{w}^*\right\|^2-\left\langle\eta\sum_{k=0}^{K-1}\frac{1}{N}\sum_{i=1}^{N}\nabla\tilde{f}_i(\tilde{\mathbf{w}}_{r,k}^i),\ \mathbf{w}_r-\tilde{\mathbf{w}}_{r,k}^i\right\rangle \\
&\quad-\left\langle\eta\sum_{k=0}^{K-1}\frac{1}{N}\sum_{i=1}^{N}\nabla\tilde{f}_i(\tilde{\mathbf{w}}_{r,k}^i),\ \tilde{\mathbf{w}}_{r,t}^i-\mathbf{w}^*\right\rangle+\mathbb{E}\left\|\eta\sum_{k=0}^{K-1}\frac{1}{N}\sum_{i=1}^{N}\tilde{\mathbf{g}}_{r,k}^i\right\|^2.
\end{aligned}
$$

where at last step we use the fact $\mathbb{E}_{\mathbf{m}_i^r}[\mathbf{m}_i^r\odot\nabla f_i(\mathbf{m}_i^r\odot\tilde{\mathbf{w}}_{r,k}^i)]=\nabla\tilde{f}_i(\tilde{\mathbf{w}}_{r,k}^i)$. Since $\tilde{f}_i$ is $L_i=p_iL$ smooth and μ_i strongly convex, and by definition $\tilde{L}=\max_{i\in[N]}L_i$, $\tilde{\mu}=\min_{i\in[N]}\mu_i$, we have

$$
\begin{aligned}
\mathbb{E}\left\|\mathbf{w}_{r+1}-\mathbf{w}^*\right\|^2 &\le (1-\tilde{\mu}\eta K)\mathbb{E}\left\|\mathbf{w}_r-\mathbf{w}^*\right\|^2-\eta K\frac{1}{N}\sum_{i=1}^{N}\left(\tilde{f}_i(\mathbf{w}_r)-\tilde{f}_i(\mathbf{w}^*)\right)+\frac{K\eta^2\delta^2}{N} \\
&\quad+\eta\tilde{L}\frac{1}{N}\sum_{i=1}^{N}\sum_{k=0}^{K-1}\mathbb{E}\left\|\tilde{\mathbf{w}}_{r,k}^i-\mathbf{w}_r\right\|^2+\eta^2\mathbb{E}\left\|\sum_{k=0}^{K-1}\frac{1}{N}\sum_{i=1}^{N}\mathbf{m}_i^r\odot\nabla f_i(\mathbf{m}_i^r\odot\tilde{\mathbf{w}}_{r,k}^i)\right\|^2 \\
&\le (1-\tilde{\mu}\eta K)\mathbb{E}\left\|\mathbf{w}_r-\mathbf{w}^*\right\|^2-\eta K\frac{1}{N}\sum_{i=1}^{N}\left(\tilde{f}_i(\mathbf{w}_r)-\tilde{f}_i(\mathbf{w}^*)\right) \\
&\quad+\eta\tilde{L}\frac{1}{N}\sum_{i=1}^{N}\sum_{k=0}^{K-1}\mathbb{E}\left\|\tilde{\mathbf{w}}_{r,k}^i-\mathbf{w}_r\right\|^2 \\
&\quad+2\eta^2K\sum_{k=0}^{K-1}\frac{1}{N}\sum_{i=1}^{N}\mathbb{E}\left\|\mathbf{m}_i^r\odot\nabla f_i(\mathbf{m}_i^r\odot\tilde{\mathbf{w}}_{r,k}^i)-\mathbf{m}_i^r\odot\nabla\tilde{f}_i(\mathbf{m}_i^r\odot\mathbf{w}_r)\right\|^2 \\
&\quad+2\eta^2K^2\mathbb{E}\left\|\frac{1}{N}\sum_{i=1}^{N}\mathbf{m}_i^r\odot\nabla f_i(\mathbf{m}_i^r\odot\mathbf{w}_r)\right\|^2+\frac{K\eta^2\delta^2}{N} \\
&\le (1-\tilde{\mu}\eta K)\mathbb{E}\left\|\mathbf{w}_r-\mathbf{w}^*\right\|^2-(\eta K-4\eta^2K^2L)\frac{1}{N}\sum_{i=1}^{N}\left(\tilde{f}_i(\mathbf{w}_r)-\tilde{f}_i(\mathbf{w}^*)\right) \\
&\quad+(\eta\tilde{L}+2\eta^2\tilde{L}^2K)\frac{1}{N}\sum_{i=1}^{N}\sum_{k=0}^{K-1}\mathbb{E}\left\|\tilde{\mathbf{w}}_{r,k}^i-\mathbf{w}_r\right\|^2+\frac{K\eta^2\delta^2}{N},
\end{aligned}
$$

where the last step is due to (Demidovich et al., 2023, Lemma 4) that

$$\mathbb{E}\left\|\frac{1}{N}\sum_{i=1}^{N}\mathbf{m}_i^r\odot\nabla f_i(\mathbf{m}_i^r\odot\mathbf{w}_r)\right\|^2\leq 2L\frac{1}{N}\sum_{i=1}^{N}\mathbb{E}\left(\tilde{f}_i(\mathbf{w}_r)-\tilde{f}_i(\mathbf{w}^*)\right).$$

Since $\eta\leq\frac{1}{4L}$, we can conclude the proof. □

Lemma 5. *For Algorithm 1, under the condition of Theorem 1, the following statement holds true:*

$$\frac{1}{N}\sum_{i=1}^{N}\mathbb{E}\left\|\tilde{\mathbf{w}}_{r,k}^i-\mathbf{w}_r\right\|^2\leq 5K\left(8\eta^2KL\mathbb{E}\left(F_{\mathbf{p}}(\mathbf{w}_r)-F_{\mathbf{p}}(\mathbf{w}^*)\right)+4\eta^2K\sigma_*^2+\eta^2K\delta^2\right)$$

Proof. According to local updating rule we have:

$$\begin{aligned}
&\mathbb{E}\left\|\tilde{\mathbf{w}}_{r,k}^i-\mathbf{w}_r\right\|^2\\
&=(1+\frac{1}{K-1})\mathbb{E}\left\|\tilde{\mathbf{w}}_{r,k-1}^i-\mathbf{w}_r\right\|^2+K\mathbb{E}\left\|\eta\tilde{\mathbf{g}}_{r,k-1}^i\right\|^2\\
&\leq(1+\frac{1}{K-1})\mathbb{E}\left\|\tilde{\mathbf{w}}_{r,k-1}^i-\mathbf{w}_r\right\|^2+K\mathbb{E}\left\|\eta\mathbf{m}_i^r\odot\nabla f_i(\mathbf{m}_i^r\odot\tilde{\mathbf{w}}_{r,k-1}^i)\right\|^2+\eta^2K\delta^2\\
&\leq(1+\frac{1}{K-1})\mathbb{E}\left\|\tilde{\mathbf{w}}_{r,k-1}^i-\mathbf{w}_r\right\|^2+2K\mathbb{E}\left\|\eta\mathbf{m}_i^r\odot\nabla f_i(\mathbf{m}_i^r\odot\tilde{\mathbf{w}}_r)\right\|^2\\
&\quad+2\eta^2\tilde{L}^2K\left\|\tilde{\mathbf{w}}_{r,k-1}^i-\tilde{\mathbf{w}}_r\right\|^2+\eta^2K\delta^2\\
&\leq(1+\frac{2}{K-1})\mathbb{E}\left\|\tilde{\mathbf{w}}_{r,k-1}^i-\mathbf{w}_r\right\|^2+8\eta^2KL\mathbb{E}\left(\tilde{f}_i(\mathbf{w}_r)-\tilde{f}_i(\mathbf{w}^*)\right)\\
&\quad+4\eta^2K\mathbb{E}\left\|\mathbf{m}_i^r\odot\nabla f_i(\mathbf{m}_i^r\odot\mathbf{w}^*)\right\|^2+\eta^2K\delta^2\\
&\leq\sum_{j=1}^{k}(1+\frac{2}{K-1})^{k-j}\left(8\eta^2KL\mathbb{E}\left(\tilde{f}_i(\mathbf{w}_r)-\tilde{f}_i(\mathbf{w}^*)\right)+4\eta^2K\mathbb{E}\left\|\mathbf{m}_i^r\odot\nabla f_i(\mathbf{m}_i^r\odot\mathbf{w}^*)\right\|^2+\eta^2K\delta^2\right)\\
&\leq 5K\left(8\eta^2KL\mathbb{E}\left(\tilde{f}_i(\mathbf{w}_r)-\tilde{f}_i(\mathbf{w}^*)\right)+4\eta^2K\mathbb{E}\left\|\mathbf{m}_i^r\odot\nabla f_i(\mathbf{m}_i^r\odot\mathbf{w}^*)\right\|^2+\eta^2K\delta^2\right)
\end{aligned}$$

where the fourth step is due to $2\eta^2\tilde{L}^2K\leq\frac{1}{K-1}$ and (Demidovich et al., 2023, Lemma 4) that

$$\begin{aligned}
\mathbb{E}\left\|\mathbf{m}_i^r\odot\nabla f_i(\mathbf{m}_i^r\odot\mathbf{w}_r)\right\|^2&\leq 2\mathbb{E}\left\|\mathbf{m}_i^r\odot\nabla f_i(\mathbf{m}_i^r\odot\mathbf{w}_r)-\mathbf{m}_i^r\odot\nabla f_i(\mathbf{m}_i^r\odot\mathbf{w}^*)\right\|^2\\
&\quad+2\mathbb{E}\left\|\mathbf{m}_i^r\odot\nabla f_i(\mathbf{m}_i^r\odot\mathbf{w}^*)\right\|^2\\
&\leq 4L\mathbb{E}\left(\tilde{f}_i(\mathbf{w}_r)-\tilde{f}_i(\mathbf{w}^*)\right)+2\mathbb{E}\left\|\mathbf{m}_i^r\odot\nabla f_i(\mathbf{m}_i^r\odot\mathbf{w}^*)\right\|^2.
\end{aligned}$$

Summing $i=1$ to N yields:

$$\frac{1}{N}\sum_{i=1}^{N}\mathbb{E}\left\|\tilde{\mathbf{w}}_{r,k}^i-\mathbf{w}_r\right\|^2\leq 5K\left(8\eta^2KL\mathbb{E}\left(F_{\mathbf{p}}(\mathbf{w}_r)-F_{\mathbf{p}}(\mathbf{w}^*)\right)+4\eta^2K\sigma_*^2+\eta^2K\delta^2\right),$$

which conclude the proof. □

C.1.1 Proof of Theorem 1

Proof. Evoking Lemma 4 yields:

$$\begin{aligned}
\mathbb{E}\left\|\mathbf{w}_{r+1}-\mathbf{w}^*\right\|^2\leq(1-\tilde{\mu}\eta)\mathbb{E}\left\|\mathbf{w}_r-\mathbf{w}^*\right\|^2-\frac{1}{2}\eta K\frac{1}{N}\sum_{i=1}^{N}\left(\tilde{f}_i(\mathbf{w}_r)-\tilde{f}_i(\mathbf{w}^*)\right)\\
+2\eta\tilde{L}\frac{1}{N}\sum_{i=1}^{N}\sum_{k=0}^{K-1}\mathbb{E}\left\|\tilde{\mathbf{w}}_{r,k}^i-\mathbf{w}_r\right\|^2+\frac{K\eta^2\delta^2}{N}.
\end{aligned}$$

We plug in Lemma 5 and get

$$\begin{aligned}\mathbb{E}\left\|\mathbf{w}_{r+1}-\mathbf{w}^*\right\|^2 &\leq (1-\tilde{\mu}\eta)\mathbb{E}\left\|\mathbf{w}_r-\mathbf{w}^*\right\|^2-\frac{1}{2}\eta K\left(F(\mathbf{w}_r)-F(\mathbf{w}^*)\right)\\ &\quad+2\eta\tilde{L}\cdot 5K^2\left(8\eta^2KL\mathbb{E}\left(F(\mathbf{w}_r)-F(\mathbf{w}^*)\right)+4\eta^2K\sigma_*^2+\eta^2K\delta^2\right)+\frac{K\eta^2\delta^2}{N}\\ &=(1-\tilde{\mu}\eta)\mathbb{E}\left\|\mathbf{w}_r-\mathbf{w}^*\right\|^2-\left(\frac{1}{2}\eta K-(\eta L+2\eta^2L^2K)40\eta^2K^3\right)\left(F_{\mathbf{p}}(\mathbf{w}_r)-F_{\mathbf{p}}(\mathbf{w}^*)\right)\\ &\quad+2\eta\tilde{L}\cdot 5K^2\left(4\eta^2K\sigma_*^2+\eta^2K\delta^2\right)+\frac{K\eta^2\delta^2}{N}.\end{aligned}$$

Since we choose $\eta\leq\frac{1}{16LK}$, we know $\left(\frac{1}{2}\eta K-(\eta L+2\eta^2L^2K)40\eta^2K^3\right)\geq 0$, so we can drop this term and get:

$$\mathbb{E}\left\|\mathbf{w}_{r+1}-\mathbf{w}^*\right\|^2\leq(1-\tilde{\mu}\eta)\mathbb{E}\left\|\mathbf{w}_r-\mathbf{w}^*\right\|^2+2\eta\tilde{L}\cdot 5K^2\left(4\eta^2K\sigma_*^2+\eta^2K\delta^2\right)+\frac{K\eta^2\delta^2}{N}.$$

Unrolling the recursion yields:

$$\mathbb{E}\left\|\mathbf{w}_{r+1}-\mathbf{w}^*\right\|^2\leq(1-\tilde{\mu}\eta K)^r\mathbb{E}\left\|\mathbf{w}_0-\mathbf{w}^*\right\|^2+2\tilde{\kappa}\cdot 5K\left(4\eta^2K\sigma_*^2+\eta^2K\delta^2\right)+\frac{\eta\delta^2}{\tilde{\mu}N}. \tag{7}$$

Plugging in $\eta=\frac{\log(KR)^2}{\tilde{\mu}KR}$ will conclude the proof:

$$\mathbb{E}\left\|\mathbf{w}_R-\mathbf{w}^*\right\|^2\leq O\left(\frac{\mathbb{E}\left\|\mathbf{w}_0-\mathbf{w}^*\right\|^2}{K^2R^2}\right)+\tilde{O}\left(\frac{\tilde{\kappa}\frac{1}{N}\sum_{i=1}^N\left\|\nabla\tilde{f}_i(\mathbf{w}^*)\right\|^2+\tilde{\kappa}\delta^2}{\tilde{\mu}^2R^2}\right)+\tilde{O}\left(\frac{\delta^2}{\tilde{\mu}^2NKR}\right).$$

□

C.2 Proof of Nonconvex Setting

In this subsection, we are going to prove Theorem 2. Since constrained non-convex stochastic optimization suffers from residual noise error unless a large mini-batch is used (Ghadimi et al., 2016), here we assume an unconstrained setting, i.e., $\mathcal{W}=\mathbb{R}^d$. We present the following technical lemma first.

Lemma 6. *For Algorithm 1, under the condition of Theorem 1, the following statement holds true:*

$$\frac{1}{N}\sum_{i=1}^N\mathbb{E}\left\|\tilde{\mathbf{w}}_{r,k}^i-\mathbf{w}_r\right\|^2\leq 5K\left(4\eta^2K\mathbb{E}\left\|\nabla F_{\mathbf{p}}(\mathbf{w}_r)\right\|^2+4\eta^2K\zeta_{\mathbf{p}}^2+\eta^2K\delta^2\right).$$

Proof. According to local updating rule we have:

$$\begin{aligned}&\mathbb{E}\left\|\tilde{\mathbf{w}}_{r,k}^i-\mathbf{w}_r\right\|^2\\ &=(1+\frac{1}{K-1})\mathbb{E}\left\|\tilde{\mathbf{w}}_{r,k-1}^i-\mathbf{w}_r\right\|^2+K\mathbb{E}\left\|\eta\tilde{\mathbf{g}}_{r,k-1}^i\right\|^2\\ &\leq(1+\frac{1}{K-1})\mathbb{E}\left\|\tilde{\mathbf{w}}_{r,k-1}^i-\mathbf{w}_r\right\|^2+K\mathbb{E}\left\|\eta\mathbf{m}_i^r\odot\nabla f_i(\mathbf{m}_i^r\odot\tilde{\mathbf{w}}_{r,k-1}^i)\right\|^2+\eta^2K\delta^2\\ &\leq(1+\frac{1}{K-1})\mathbb{E}\left\|\tilde{\mathbf{w}}_{r,k-1}^i-\mathbf{w}_r\right\|^2+2K\mathbb{E}\left\|\eta\mathbf{m}_i^r\odot\nabla f_i(\mathbf{m}_i^r\odot\mathbf{w}_r)\right\|^2+2\eta^2\tilde{L}^2K\left\|\tilde{\mathbf{w}}_{r,k-1}^i-\mathbf{w}_r\right\|^2\\ &\quad+\eta^2K\delta^2\\ &\leq(1+\frac{2}{K-1})\mathbb{E}\left\|\tilde{\mathbf{w}}_{r,k-1}^i-\mathbf{w}_r\right\|^2+4\eta^2K\mathbb{E}\left\|\nabla F_{\mathbf{p}}(\mathbf{w}_r)\right\|^2\\ &\quad+4\eta^2K\mathbb{E}\left\|\mathbf{m}_i^r\odot\nabla f_i(\mathbf{m}_i^r\odot\mathbf{w}_r)-\nabla F_{\mathbf{p}}(\mathbf{w}_r)\right\|^2+\eta^2K\delta^2\\ &\leq\sum_{j=1}^k(1+\frac{2}{K-1})^{k-j}\left(4\eta^2K\mathbb{E}\left\|\nabla F_{\mathbf{p}}(\mathbf{w}_r)\right\|^2+4\eta^2K\mathbb{E}\left\|\mathbf{m}_i^r\odot\nabla f_i(\mathbf{m}_i^r\odot\mathbf{w}_r)-\nabla F_{\mathbf{p}}(\mathbf{w}_r)\right\|^2+\eta^2K\delta^2\right)\\ &\leq 5K\left(4\eta^2K\mathbb{E}\left\|\nabla F_{\mathbf{p}}(\mathbf{w}_r)\right\|^2+4\eta^2K\mathbb{E}\left\|\mathbf{m}_i^r\odot\nabla f_i(\mathbf{m}_i^r\odot\mathbf{w}_r)-\nabla F_{\mathbf{p}}(\mathbf{w}_r)\right\|^2+\eta^2K\delta^2\right),\end{aligned}$$

where the fourth step is due to $2\eta^2\tilde{L}^2K \leq \frac{1}{K-1}$.

Summing for $i = 1$ to N yields:

$$\frac{1}{N}\sum_{i=1}^{N}\mathbb{E}\left\|\tilde{\mathbf{w}}_{r,k}^{i}-\mathbf{w}_r\right\|^2 \leq 5K\left(4\eta^2K\mathbb{E}\left\|\nabla F_{\mathbf{p}}(\mathbf{w}_r)\right\|^2+4\eta^2K\zeta_{\mathbf{p}}^2+\eta^2K\delta^2\right),$$

which concludes the proof.

□

C.2.1 Proof of Theorem 2

Proof. From $L_{\mathbf{p}}$-smoothness of $F_{\mathbf{p}}$ (Proposition 1), we have

$$\begin{aligned}
\mathbb{E}[F_{\mathbf{p}}(\mathbf{w}_{r+1})] &\leq \mathbb{E}[F_{\mathbf{p}}(\mathbf{w}_r)]+\mathbb{E}\left\langle\nabla F_{\mathbf{p}}(\mathbf{w}_r),\,\mathbf{w}_{r+1}-\mathbf{w}_r\right\rangle+\frac{L_{\mathbf{p}}}{2}\mathbb{E}\left\|\mathbf{w}_{r+1}-\mathbf{w}_r\right\|^2\\
&\leq \mathbb{E}[F_{\mathbf{p}}(\mathbf{w}_r)]-\mathbb{E}\left\langle\nabla F_{\mathbf{p}}(\mathbf{w}_r),\,\eta\sum_{k=0}^{K-1}\frac{1}{N}\sum_{i=1}^{N}\mathbf{m}_i^r\odot\nabla f_i(\mathbf{m}_i^r\odot\tilde{\mathbf{w}}_t^i)\right\rangle\\
&\quad+\frac{L}{2}\mathbb{E}\left\|\eta\sum_{k=0}^{K-1}\frac{1}{N}\sum_{i=1}^{N}\mathbf{m}_i^r\odot\nabla f_i(\mathbf{m}_i^r\odot\tilde{\mathbf{w}}_t^i)\right\|^2+\frac{\eta^2L_{\mathbf{p}}K\delta^2}{2N}\\
&= \mathbb{E}[F_{\mathbf{p}}(\mathbf{w}_r)]-\mathbb{E}\eta K\left\langle\nabla F_{\mathbf{p}}(\mathbf{w}_r),\,\frac{1}{K}\sum_{k=0}^{K-1}\frac{1}{N}\sum_{i=1}^{N}\nabla\tilde{f}_i(\tilde{\mathbf{w}}_t^i)\right\rangle\\
&\quad+\frac{L}{2}\mathbb{E}\left\|\eta\sum_{k=0}^{K-1}\frac{1}{N}\sum_{i=1}^{N}\mathbf{m}_i^r\odot\nabla f_i(\mathbf{m}_i^r\odot\tilde{\mathbf{w}}_t^i)\right\|^2+\frac{\eta^2L_{\mathbf{p}}K\delta^2}{2N}
\end{aligned}$$

Applying the identity $\langle\mathbf{a},\,\mathbf{b}\rangle=\frac{1}{2}\left\|\mathbf{a}\right\|^2+\frac{1}{2}\left\|\mathbf{b}\right\|^2-\frac{1}{2}\left\|\mathbf{a}-\mathbf{b}\right\|^2$ yields:

$$\begin{aligned}
\mathbb{E}[F_{\mathbf{p}}(\mathbf{w}_{r+1})] &\leq \mathbb{E}[F_{\mathbf{p}}(\mathbf{w}_r)]-\frac{1}{2}\eta K\mathbb{E}\left\|\nabla F_{\mathbf{p}}(\mathbf{w}_r)\right\|^2-\frac{1}{2}\eta K\mathbb{E}\left\|\frac{1}{K}\sum_{k=0}^{K-1}\frac{1}{N}\sum_{i=1}^{N}\nabla\tilde{f}_i(\tilde{\mathbf{w}}_t^i)\right\|^2\\
&\quad+\frac{1}{2}\eta K\left\|\nabla F_{\mathbf{p}}(\mathbf{w}_r)-\frac{1}{K}\sum_{k=0}^{K-1}\frac{1}{N}\sum_{i=1}^{N}\nabla\tilde{f}_i(\tilde{\mathbf{w}}_t^i)\right\|^2\\
&\quad+\frac{L_{\mathbf{p}}}{2}\eta^2K^2\mathbb{E}\left\|\frac{1}{K}\sum_{k=0}^{K-1}\frac{1}{N}\sum_{i=1}^{N}\mathbf{m}_i^r\odot\nabla f_i(\mathbf{m}_i^r\odot\tilde{\mathbf{w}}_t^i)\right\|^2+\frac{\eta^2L_{\mathbf{p}}K\delta^2}{2N}\\
&= \mathbb{E}[F_{\mathbf{p}}(\mathbf{w}_r)]-\frac{1}{2}\eta K\mathbb{E}\left\|\nabla F_{\mathbf{p}}(\mathbf{w}_r)\right\|^2\\
&\quad-\left(\frac{1}{2}\eta K-\frac{L_{\mathbf{p}}}{2}\eta^2K^2\right)\mathbb{E}\left\|\frac{1}{K}\sum_{k=0}^{K-1}\frac{1}{N}\sum_{i=1}^{N}\mathbf{m}_i^r\odot\nabla f_i(\mathbf{m}_i^r\odot\tilde{\mathbf{w}}_t^i)\right\|^2\\
&\quad+\frac{1}{2}\eta K\left\|\nabla F_{\mathbf{p}}(\mathbf{w}_r)-\frac{1}{K}\sum_{k=0}^{K-1}\frac{1}{N}\sum_{i=1}^{N}\nabla\tilde{f}_i(\tilde{\mathbf{w}}_t^i)\right\|^2+\frac{\eta^2L_{\mathbf{p}}KV^2}{2N}+\frac{\eta^2L_{\mathbf{p}}K\delta^2}{2N}
\end{aligned}$$

where $V:=\sup_{i\in[N]}\mathbb{E}_{\mathbf{m}}\left\|\nabla\tilde{f}_i(\mathbf{w})-\mathbf{m}\odot\nabla f_i(\mathbf{m}\odot\mathbf{w})\right\|^2$. Since we choose $\eta\leq\frac{1}{KL}$, we know

$\frac{1}{2}\eta K - \frac{L_{\mathbf{p}}}{2}\eta^2 K^2 \leq 0$.

$$\begin{aligned}
\mathbb{E}[F_{\mathbf{p}}(\mathbf{w}_{r+1})] &\leq \mathbb{E}[F_{\mathbf{p}}(\mathbf{w}_r)] - \frac{1}{2}\eta K \mathbb{E}\left\|\nabla F_{\mathbf{p}}(\mathbf{w}_r)\right\|^2 \\
&\quad + \frac{1}{2}\eta K \left\|\nabla F_{\mathbf{p}}(\mathbf{w}_r) - \frac{1}{K}\sum_{k=0}^{K-1}\frac{1}{N}\sum_{i=1}^{N}\nabla \tilde{f}_i(\tilde{\mathbf{w}}_t^i)\right\|^2 \\
&\quad + \frac{\eta^2 L_{\mathbf{p}} K(\delta^2 + V^2)}{2N} \\
&\leq \mathbb{E}[F_{\mathbf{p}}(\mathbf{w}_r)] - \frac{1}{2}\eta K \mathbb{E}\left\|\nabla F_{\mathbf{p}}(\mathbf{w}_r)\right\|^2 + \frac{1}{2}\eta K \frac{1}{K}\sum_{k=0}^{K-1}\frac{1}{N}\sum_{i=1}^{N}\tilde{L}^2\left\|\mathbf{w}_r - \tilde{\mathbf{w}}_t^i\right\|^2 \\
&\quad + \frac{\eta^2 L_{\mathbf{p}} K(\delta^2 + V^2)}{2N}.
\end{aligned}$$

Plugging in Lemma 6 above yields:

$$\begin{aligned}
\mathbb{E}[F_{\mathbf{p}}(\mathbf{w}_{r+1})] &\leq \mathbb{E}[F_{\mathbf{p}}(\mathbf{w}_r)] - \frac{1}{2}\eta K \mathbb{E}\left\|\nabla F_{\mathbf{p}}(\mathbf{w}_r)\right\|^2 \\
&\quad + \frac{1}{2}\eta K \tilde{L}^2 5K\left(4\eta^2 K \mathbb{E}\left\|\nabla F_{\mathbf{p}}(\tilde{\mathbf{w}}_r)\right\|^2 + 4\eta^2 K \zeta_{\mathbf{p}}^2 + \eta^2 K \delta^2\right) + \frac{\eta^2 L_{\mathbf{p}} K(V^2 + \delta^2)}{2N} \\
&= \mathbb{E}[F_{\mathbf{p}}(\mathbf{w}_r)] - \left(\frac{1}{2}\eta K - 10\eta^3 K^3 L^2\right)\mathbb{E}\left\|\nabla F_{\mathbf{p}}(\mathbf{w}_r)\right\|^2 \\
&\quad + 20\eta^3 K^3 L_{\mathbf{p}}^2 \zeta_{\mathbf{p}}^2 + 5\eta^3 K^3 L_{\mathbf{p}}^2 \delta^2 + \frac{\eta^2 L_{\mathbf{p}} K(V^2 + \delta^2)}{2N} \\
&\leq \mathbb{E}[F_{\mathbf{p}}(\mathbf{w}_r)] - \frac{1}{4}\eta K \mathbb{E}\left\|\nabla F_{\mathbf{p}}(\mathbf{w}_r)\right\|^2 + 20\eta^3 K^3 L_{\mathbf{p}}^2 \zeta_{\mathbf{p}}^2 + 5\eta^3 K^3 L_{\mathbf{p}}^2 \delta^2 \\
&\quad + \frac{\eta^2 L_{\mathbf{p}} K(V^2 + \delta^2)}{2N}
\end{aligned}$$

Re-arranging terms yields:

$$\mathbb{E}\left\|\nabla F_{\mathbf{p}}(\mathbf{w}_r)\right\|^2 \leq 4\frac{\mathbb{E}[F_{\mathbf{p}}(\mathbf{w}_r)] - \mathbb{E}[F_{\mathbf{p}}(\mathbf{w}_{r+1})]}{\eta K} + 80\eta^2 K^2 L_{\mathbf{p}}^2 \zeta_{\mathbf{p}}^2 + 20\eta^2 K^2 L_{\mathbf{p}}^2 \delta^2 + \frac{4\eta L_{\mathbf{p}}(V^2 + \delta^2)}{2N}$$

Summing over $r = 1$ to R yields:

$$\frac{1}{R}\sum_{r=1}^{R}\mathbb{E}\left\|\nabla F_{\mathbf{p}}(\mathbf{w}_r)\right\|^2 \leq 4\frac{\mathbb{E}[F_{\mathbf{p}}(\mathbf{w}_0)]}{\eta R K} + 80\eta^2 K^2 L_{\mathbf{p}}^2 \zeta_{\mathbf{p}} + 20\eta^2 K^2 L_{\mathbf{p}}^2 \delta^2 + \frac{4\eta L_{\mathbf{p}}(V^2 + \delta^2)}{2N}.$$

Finally plugging in $\eta = \frac{1}{L\sqrt{RK}}$ will give the desired rate:

$$\frac{1}{R}\sum_{r=1}^{R}\mathbb{E}\left\|\nabla F_{\mathbf{p}}(\mathbf{w}_r)\right\|^2 \leq O\left(\frac{L\mathbb{E}[F_{\mathbf{p}}(\mathbf{w}_0)]}{\sqrt{RK}} + \frac{K\zeta_{\mathbf{p}}^2}{R} + \frac{K\delta^2}{R} + \frac{\delta^2}{N\sqrt{RK}}\right).$$

□

D PROOF OF CONVERGENCE OF ROLLING

In this section, we are going to present convergence proof of Algorithm 2. At the start of each epoch, the server shuffles these sub-models and assigns them sequentially to clients. This introduces complexity in analysis due to the interaction between both the model drift caused by partial training on sub-models and the impact of permutation-based assignments on convergence.

D.1 Proof of Convex Setting

In this section, we will present proof of Algorithm 2 in convex setting (Theorem 3). We present useful lemmas first.

Proposition 2. *Define function* $F_{\mathbf{m}}(\mathbf{w}) = \frac{1}{N}\sum_{i=1}^{N}\frac{1}{d}\sum_{j=1}^{d} f_i(\mathbf{m}_i^j \odot \mathbf{w})$. *If each* f_i *is* L *smooth and* μ *strongly convex, then* $F_{\mathbf{m}}$ *is also* L *smooth and* μ *strongly convex.*

Proof. We first examine the smoothness and convexity parameter of $F_{\mathbf{m}}$. For smoothness:

$$\begin{aligned}
F_{\mathbf{m}}(\mathbf{w}+\Delta\mathbf{w}) &= \frac{1}{N}\sum_{i=1}^{N}\frac{1}{d}\sum_{j=1}^{d} f_i(\mathbf{m}_i^j \odot (\mathbf{w}+\Delta\mathbf{w})) \\
&\le \frac{1}{N}\sum_{i=1}^{N}\frac{1}{d}\sum_{j=1}^{d}\left(f_i(\mathbf{m}_i^j \odot \mathbf{w}) + \left\langle \mathbf{m}_i^j \odot \Delta\mathbf{w},\, \nabla f_i(\mathbf{m}_i^j \odot \mathbf{w})\right\rangle + \frac{L}{2}\left\|\mathbf{m}_i^j \odot \Delta\mathbf{w}\right\|^2\right) \\
&= F_{\mathbf{m}}(\mathbf{w}) + \left\langle \Delta\mathbf{w},\, \frac{1}{N}\sum_{i=1}^{N}\frac{1}{d}\sum_{j=1}^{d}\mathbf{m}_i^j \odot \nabla f_i(\mathbf{m}_i^j \odot \mathbf{w})\right\rangle \\
&\quad + \frac{L}{2}\frac{1}{N}\sum_{i=1}^{N}\frac{1}{d}\sum_{j=1}^{d}\left\langle (\mathbf{M}_i^j)^2\Delta\mathbf{w},\, \Delta\mathbf{w}\right\rangle \\
&\le F_{\mathbf{m}}(\mathbf{w}) + \langle \Delta\mathbf{w},\, \nabla F_{\mathbf{m}}(\mathbf{w})\rangle + \frac{L}{2}\left(\frac{1}{N}\sum_{i=1}^{N}\mu_{\max}(\underbrace{\frac{1}{d}\sum_{j=1}^{d}\mathbf{M}_i^2}_{=\mathbf{I}})\right)\|\Delta\mathbf{w}\|^2 \\
&= F_{\mathbf{m}}(\mathbf{w}) + \langle \Delta\mathbf{w},\, \nabla F_{\mathbf{m}}(\mathbf{w})\rangle + \frac{L}{2}\|\Delta\mathbf{w}\|^2 .
\end{aligned}$$

For convexity:

$$\begin{aligned}
F_{\mathbf{m}}(\mathbf{w}+\Delta\mathbf{w}) &= \frac{1}{N}\sum_{i=1}^{N}\frac{1}{d}\sum_{j=1}^{d} f_i(\mathbf{m}_i^j \odot (\mathbf{w}+\Delta\mathbf{w})) \\
&\ge \frac{1}{N}\sum_{i=1}^{N}\frac{1}{d}\sum_{j=1}^{d}\left(f_i(\mathbf{m}_i^j \odot \mathbf{w}) + \left\langle \mathbf{m}_i^j \odot \Delta\mathbf{w},\, \nabla f_i(\mathbf{m}_i^j \odot \mathbf{w})\right\rangle + \frac{\mu}{2}\left\|\mathbf{m}_i^j \odot \Delta\mathbf{w}\right\|^2\right) \\
&= F_{\mathbf{m}}(\mathbf{w}) + \left\langle \Delta\mathbf{w},\, \frac{1}{N}\sum_{i=1}^{N}\frac{1}{d}\sum_{j=1}^{d}\mathbf{m}_i^j \odot \nabla f_i(\mathbf{m}_i^j \odot \mathbf{w})\right\rangle \\
&\quad + \frac{\mu}{2}\frac{1}{N}\sum_{i=1}^{N}\frac{1}{d}\sum_{j=1}^{d}\left\langle (\mathbf{M}_i^j)^2\Delta\mathbf{w},\, \Delta\mathbf{w}\right\rangle \\
&\ge F_{\mathbf{m}}(\mathbf{w}) + \langle \Delta\mathbf{w},\, \nabla F_{\mathbf{m}}(\mathbf{w})\rangle + \frac{\mu}{2}\left(\frac{1}{N}\sum_{i=1}^{N}\mu_{\min}(\underbrace{\frac{1}{d}\sum_{j=1}^{d}\mathbf{M}_i^2}_{=\mathbf{I}})\right)\|\Delta\mathbf{w}\|^2 \\
&= F_{\mathbf{m}}(\mathbf{w}) + \langle \Delta\mathbf{w},\, \nabla F_{\mathbf{m}}(\mathbf{w})\rangle + \frac{\mu}{2}\|\Delta\mathbf{w}\|^2 .
\end{aligned}$$

So $F_{\mathbf{m}}$ is also L smooth and μ strongly convex. □

Lemma 7. *Given a mask* $\mathbf{m} = \left[\left[\mathbf{m}_i^1, \ldots, \mathbf{m}_i^R\right]\right]_{i=1}^{N} \in \{0,1\}^{dNR}$, *we define* $\mathbf{w}^*(\mathbf{p}) := \arg\min_{\mathbf{w}\in\mathcal{W}}\left\{\Phi(\mathbf{m},\mathbf{w}) := \frac{1}{N}\sum_{i=1}^{N}\frac{1}{R}\sum_{j=1}^{R} f_i(\mathbf{m}_i^j \odot \mathbf{w})\right\}$.

We further define $\bar{\mathbf{m}} = [[\mathbf{1}, \dots, \mathbf{1}]]_{i=1}^{N}$ *and* $\mathbf{w}^*(\bar{\mathbf{m}}) := \arg\min_{\mathbf{w}} F(\mathbf{w})$. *If each* f_i *is* μ *strongly convex, and* $\sup_{\mathbf{w}\in\mathcal{W}} \|\nabla f_i(\mathbf{w})\| \leq G$, *then the following statement holds:*

$$\|\mathbf{w}^*(\bar{\mathbf{m}}) - \mathbf{w}^*(\mathbf{m})\| \leq \sqrt{\frac{2G^2 + 2W^2L^2}{\mu^2 NR}} \|\mathbf{m} - \bar{\mathbf{m}}\|.$$

Proof. We define $\Phi(\mathbf{m}, \mathbf{w}) := \frac{1}{N}\frac{1}{R}\sum_{i=1}^{N}\sum_{j=1}^{R} f_i(\mathbf{m}_i^j \odot \mathbf{w})$. First, according to optimality conditions we have:

$$\langle \mathbf{w} - \mathbf{w}^*(\mathbf{m}), \nabla_2\Phi(\mathbf{m}, \mathbf{w}^*(\mathbf{m}))\rangle \geq 0,$$
$$\langle \mathbf{w} - \mathbf{w}^*(\bar{\mathbf{m}}), \nabla_2\Phi(\bar{\mathbf{m}}, \mathbf{w}^*(\bar{\mathbf{m}}))\rangle \geq 0$$

Substituting $\mathbf{w}$ with $\mathbf{w}^*(\bar{\mathbf{m}})$ and $\mathbf{w}^*(\mathbf{m})$ in the above first and second inequalities respectively yields:

$$\langle \mathbf{w}^*(\bar{\mathbf{m}}) - \mathbf{w}^*(\mathbf{m}), \nabla_2\Phi(\mathbf{m}, \mathbf{w}^*(\mathbf{m}))\rangle \geq 0,$$
$$\langle \mathbf{w}^*(\mathbf{m}) - \mathbf{w}^*(\bar{\mathbf{m}}), \nabla_2\Phi(\bar{\mathbf{m}}, \mathbf{w}^*(\bar{\mathbf{m}}))\rangle \geq 0.$$

Adding up the above two inequalities yields:

$$\langle \mathbf{w}^*(\bar{\mathbf{m}}) - \mathbf{w}^*(\mathbf{m}), \nabla_2\Phi(\mathbf{m}, \mathbf{w}^*(\mathbf{m})) - \nabla_2\Phi(\bar{\mathbf{m}}, \mathbf{w}^*(\bar{\mathbf{m}}))\rangle \geq 0. \tag{8}$$

Since $F(\mathbf{m}, \cdot)$ is μ strongly convex, as shown in Proposition 2, we have:

$$\langle \mathbf{w}^*(\bar{\mathbf{m}}) - \mathbf{w}^*(\mathbf{m}), \nabla_2\Phi(\mathbf{m}, \mathbf{w}^*(\bar{\mathbf{m}})) - \nabla_2\Phi(\mathbf{m}, \mathbf{w}^*(\mathbf{m})) \geq \mu\|\mathbf{w}^*(\bar{\mathbf{m}}) - \mathbf{w}^*(\mathbf{m})\|^2. \tag{9}$$

Adding up (8) and (9) yields:

$$\langle \mathbf{w}^*(\bar{\mathbf{m}}) - \mathbf{w}^*(\mathbf{m}), \nabla_2\Phi(\mathbf{m}, \mathbf{w}^*(\bar{\mathbf{m}})) - \nabla_2\Phi(\bar{\mathbf{m}}, \mathbf{w}^*(\bar{\mathbf{m}})) \geq \mu\|\mathbf{w}^*(\bar{\mathbf{m}}) - \mathbf{w}^*(\mathbf{m})\|^2$$

Now we examine the smoothness of $\nabla_2\Phi(\mathbf{m}, \mathbf{w})$ in terms of the first variable.

$$\begin{aligned}
&\|\nabla_2\Phi(\mathbf{m}, \mathbf{w}) - \nabla_2\Phi(\bar{\mathbf{m}}, \mathbf{w})\|^2 \\
&= \left\| \frac{1}{N}\sum_{i=1}^{N}\frac{1}{R}\sum_{j=1}^{R} \mathbf{m}_i^j \odot \nabla f_i(\mathbf{m}_i^j \odot \mathbf{w}) - \frac{1}{N}\sum_{i=1}^{N}\frac{1}{R}\sum_{j=1}^{R} \mathbf{1} \odot \nabla f_i(\bar{\mathbf{m}}_i^j \odot \mathbf{w}) \right\|^2 \\
&\leq \frac{1}{N}\sum_{i=1}^{N}\frac{1}{R}\sum_{j=1}^{R} \left\| \mathbf{m}_i^j \odot \nabla f_i(\mathbf{m}_i^j \odot \mathbf{w}) - \bar{\mathbf{m}}_i^j \odot \nabla f_i(\mathbf{1} \odot \mathbf{w}) \right\|^2 \\
&\leq \frac{1}{N}\sum_{i=1}^{N}\frac{1}{R}\sum_{j=1}^{R} \left(2G^2 \left\| \mathbf{m}_i^j - \mathbf{1} \right\|^2 + 2W^2L^2 \left\| \mathbf{m}_i^j - \mathbf{1} \right\|^2 \right) \\
&= \frac{2G^2 + 2W^2L^2}{NR} \|\mathbf{m} - \bar{\mathbf{m}}\|^2 .
\end{aligned}$$

Finally, using $\sqrt{\frac{2G^2+2W^2L^2}{NR}}$ smoothness of $\nabla_2\Phi(\cdot, \mathbf{w})$ will conclude the proof:

$$\begin{aligned}
&\sqrt{\frac{2G^2 + 2W^2L^2}{NR}} \|\mathbf{w}^*(\bar{\mathbf{m}}) - \mathbf{w}^*(\mathbf{m})\| \|\mathbf{m} - \bar{\mathbf{m}}\| \geq \mu\|\mathbf{w}^*(\bar{\mathbf{m}}) - \mathbf{w}^*(\mathbf{m})\|^2 \\
&\qquad \iff \sqrt{\frac{2G^2 + 2W^2L^2}{\mu^2 NR}} \|\mathbf{m} - \bar{\mathbf{m}}\| \geq \|\mathbf{w}^*(\bar{\mathbf{m}}) - \mathbf{w}^*(\mathbf{m})\|.
\end{aligned}$$

□

Lemma 8 (Optimality Gap). *Let* $\Phi(\mathbf{m}, \mathbf{w}) := \frac{1}{N}\sum_{i=1}^{N}\frac{1}{R}\sum_{i=1}^{j} f_i(\mathbf{m}_i^j \odot \mathbf{w})$.
Let $\hat{\mathbf{v}} = \mathcal{P}_{\mathcal{W}}(\tilde{\mathbf{w}} - \frac{1}{L}\nabla_{\mathbf{w}}\Phi(\mathbf{m}, \tilde{\mathbf{w}}))$. *If we assume each* f_i *is* L*-smooth,* μ*-strongly convex and with gradient bounded by* G*, then the following statement holds true:*

$$F(\hat{\mathbf{w}}) - F(\mathbf{w}^*) \leq 2L\|\tilde{\mathbf{w}} - \mathbf{w}^*(\mathbf{m})\|^2 + \left(\frac{2L}{\mu} + \frac{4}{L}\right)\frac{2G^2 + 2W^2L^2}{NR}\sum_{j=1}^{R}\|\mathbf{m}_i^j - \mathbf{1}\|^2,$$

where $\mathbf{w}^* = \arg\min_{\mathbf{w}\in\mathcal{W}}\Phi(\mathbf{1}, \mathbf{w})$.

Proof. From Lemma 7, we know $\nabla_{\mathbf{w}}\Phi(\cdot, \mathbf{w})$ is $\sqrt{\frac{2G^2+2W^2L^2}{NR}}$ Lipschitz and we know $\mathbf{w}^*(\boldsymbol{\alpha})$ is $\kappa_\Phi := \frac{\sqrt{N}G}{\mu}$ Lipschitz. According to property of projection, we have:

$$\begin{aligned}
0 &\leq \langle \mathbf{w} - \hat{\mathbf{w}},\, L(\hat{\mathbf{w}} - \tilde{\mathbf{w}}) + \nabla_{\mathbf{w}}\Phi(\mathbf{m}, \tilde{\mathbf{w}})\rangle \\
&= \underbrace{\langle \mathbf{w} - \hat{\mathbf{w}},\, L(\hat{\mathbf{w}} - \tilde{\mathbf{w}}) + \nabla_{\mathbf{w}}\Phi(\mathbf{1}, \tilde{\mathbf{w}})\rangle}_{T_1} + \underbrace{\langle \mathbf{w} - \hat{\mathbf{w}},\, \nabla_{\mathbf{w}}\Phi(\mathbf{m}, \tilde{\mathbf{w}}) - \nabla_{\mathbf{w}}\Phi(\mathbf{1}, \tilde{\mathbf{w}})\rangle}_{T_2}.
\end{aligned}$$

For T_1, we notice:

$$\begin{aligned}
&\langle \mathbf{w} - \hat{\mathbf{w}},\, L(\hat{\mathbf{w}} - \tilde{\mathbf{w}}) + \nabla_{\mathbf{w}}\Phi(\mathbf{1}, \tilde{\mathbf{w}})\rangle \\
&= L\,\langle \mathbf{w} - \tilde{\mathbf{w}},\, \hat{\mathbf{w}} - \tilde{\mathbf{w}}\rangle + L\,\langle \tilde{\mathbf{w}} - \hat{\mathbf{w}},\, \hat{\mathbf{w}} - \tilde{\mathbf{w}}\rangle + \langle \mathbf{w} - \hat{\mathbf{w}},\, \nabla_{\mathbf{w}}\Phi(\mathbf{1}, \tilde{\mathbf{w}})\rangle \\
&= L\,\langle \mathbf{w} - \tilde{\mathbf{w}},\, \hat{\mathbf{w}} - \tilde{\mathbf{w}}\rangle - L\,\|\tilde{\mathbf{w}} - \hat{\mathbf{w}}\|^2 + \langle \mathbf{w} - \hat{\mathbf{w}}_i,\, \nabla_{\mathbf{w}}\Phi(\mathbf{1}, \tilde{\mathbf{w}})\rangle \\
&\leq L(\|\mathbf{w} - \tilde{\mathbf{w}}\|^2 + \frac{1}{4}\|\hat{\mathbf{w}} - \tilde{\mathbf{w}}\|^2) - L\,\|\tilde{\mathbf{w}} - \hat{\mathbf{w}}\|^2 + \underbrace{\langle \mathbf{w} - \hat{\mathbf{w}},\, \nabla_{\mathbf{w}}\Phi(\mathbf{1}, \tilde{\mathbf{w}})\rangle}_{\spadesuit}
\end{aligned}$$

where at last step we used Young's inequality. To bound ♠, we apply the L smoothness and μ strongly convexity of $\Phi(\mathbf{1}, \cdot)$:

$$\begin{aligned}
\langle \mathbf{w} - \hat{\mathbf{w}},\, \nabla_{\mathbf{w}}\Phi(\mathbf{1}, \tilde{\mathbf{w}})\rangle &= \langle \mathbf{w} - \tilde{\mathbf{w}},\, \nabla_{\mathbf{w}}\Phi(\mathbf{1}, \tilde{\mathbf{w}})\rangle + \langle \tilde{\mathbf{w}} - \hat{\mathbf{w}},\, \nabla_{\mathbf{w}}\Phi(\mathbf{1}, \tilde{\mathbf{w}})\rangle \\
&\leq \Phi(\mathbf{1}, \mathbf{w}) - \Phi(\mathbf{1}, \tilde{\mathbf{w}}) - \frac{\mu}{2}\|\tilde{\mathbf{w}} - \mathbf{w}\|^2 + \Phi(\mathbf{1}, \tilde{\mathbf{w}}) - \Phi(\mathbf{1}, \hat{\mathbf{w}}) + \frac{L}{2}\|\tilde{\mathbf{w}} - \hat{\mathbf{w}}\|^2 \\
&\leq \Phi(\mathbf{1}, \mathbf{w}) - \Phi(\mathbf{1}, \hat{\mathbf{w}}) - \frac{\mu}{2}\|\tilde{\mathbf{w}} - \mathbf{w}\|^2 + \frac{L}{2}\|\tilde{\mathbf{w}} - \hat{\mathbf{w}}\|^2
\end{aligned}$$

Putting above bound back yields:

$$\langle \mathbf{w} - \hat{\mathbf{w}},\, L(\hat{\mathbf{w}} - \tilde{\mathbf{w}}) + \nabla_{\mathbf{w}}\Phi(\mathbf{1}, \tilde{\mathbf{w}})\rangle \leq \Phi(\mathbf{1}, \mathbf{w}) - \Phi(\mathbf{1}, \hat{\mathbf{w}}) + \frac{1}{2\eta}\|\tilde{\mathbf{w}} - \mathbf{w}\|^2 - \left(\frac{3L}{4} - \frac{L}{2}\right)\|\tilde{\mathbf{w}} - \hat{\mathbf{w}}\|^2$$

Now we switch to bounding T_2. Applying Cauchy-Schwartz yields:

$$\begin{aligned}
\langle \mathbf{w} - \hat{\mathbf{w}},\, \nabla_{\mathbf{w}}\Phi(\mathbf{m}, \tilde{\mathbf{w}}) - \nabla_{\mathbf{w}}\Phi(\mathbf{1}, \tilde{\mathbf{w}})\rangle &\leq \frac{L}{4}\|\mathbf{w} - \tilde{\mathbf{w}}\|^2 + \frac{L}{4}\|\tilde{\mathbf{w}} - \hat{\mathbf{w}}\|^2 + \frac{4}{L}\|\nabla_{\mathbf{w}}\Phi(\mathbf{m}, \tilde{\mathbf{w}}) - \nabla_{\mathbf{w}}\Phi(\mathbf{1}, \tilde{\mathbf{w}})\|^2 \\
&\leq \frac{L}{4}\|\mathbf{w} - \tilde{\mathbf{w}}\|^2 + \frac{L}{4}\|\tilde{\mathbf{w}} - \hat{\mathbf{w}}\|^2 + \frac{4}{L}\frac{2G^2 + 2W^2L^2}{NR}\|\mathbf{m} - \mathbf{1}\|^2
\end{aligned}$$

where at last step we apply $\sqrt{\frac{2G^2+2W^2L^2}{NR}}$ smoothness of $\nabla_2\Phi(\cdot, \mathbf{w})$. Putting pieces together yields:

$$0 \leq \Phi(\mathbf{1}, \mathbf{w}) - \Phi(\mathbf{1}, \hat{\mathbf{w}}) + \frac{L}{2}\|\tilde{\mathbf{w}} - \mathbf{w}\|^2 + \frac{L}{2}\|\mathbf{w} - \tilde{\mathbf{w}}\|^2 + \frac{4}{L}\frac{2G^2 + 2W^2L^2}{NR}\|\mathbf{m} - \mathbf{1}_i\|^2$$

Re-arranging terms and setting $\mathbf{w} = \mathbf{w}^*(\mathbf{1}) = \arg\min_{\mathbf{w}\in\mathcal{W}}\Phi(\mathbf{1}, \mathbf{w})$ yields:

$$\Phi(\mathbf{1}, \hat{\mathbf{w}}) - \Phi(\mathbf{1}, \mathbf{w}^*(\mathbf{1})) \leq L\|\tilde{\mathbf{w}} - \mathbf{w}^*\|^2 + \frac{4}{L}\frac{2G^2 + 2W^2L^2}{NR}\|\mathbf{m} - \mathbf{1}\|^2.$$

At last, due to the κ_Φ-Lipschitzness property of of $\mathbf{w}^*(\cdot)$ as shown in Lemma 7, it follows that:

$$\begin{aligned} L\|\tilde{\mathbf{w}} - \mathbf{w}^*(\mathbf{1})\|^2 &\leq L\|\tilde{\mathbf{w}} - \mathbf{w}^*(\mathbf{m})\|^2 + L\|\mathbf{w}^*(\mathbf{m}) - \mathbf{w}^*(\mathbf{1})\|^2 \\ &\leq 2L\|\tilde{\mathbf{w}} - \mathbf{w}^*(\mathbf{m})\|^2 + 2\frac{2G^2 + 2W^2L^2}{\mu NR} L\|\mathbf{m} - \mathbf{1}\|^2, \end{aligned}$$

as desired.

□

Lemma 9 (Recursion between each round)**.** *For Algorithm 2, under the assumptions of Theorem 3, the following statement holds:*

$$\mathbf{w}_{e,r+1} = \mathbf{w}_{e,r} - \eta \frac{1}{N} \sum_{i=1}^{N} K \mathbf{m}_i^{\sigma_e(r)} \odot \nabla f_i(\mathbf{m}_i^{\sigma_e(r)} \odot \mathbf{w}_{e,r}) - \eta \frac{1}{N} \sum_{i=1}^{N} \sum_{k=0}^{K-1} \mathbf{r}_{e,r,k}^i - \boldsymbol{\xi}_{e,r}.$$

where $\boldsymbol{\xi}_{e,r} = \eta \frac{1}{N} \sum_{i=1}^{N} \sum_{k=0}^{K-1} \left(\mathbf{m}_i^{\sigma_e(r)} \odot \nabla f_i(\mathbf{m}_i^{\sigma_e(r)} \odot \mathbf{w}_{e,r,k}) - \mathbf{m}_i^{\sigma_e(r)} \odot \nabla f_i(\mathbf{m}_i^{\sigma_e(r)} \odot \mathbf{w}_{e,r,k}; \xi_{e,r,k}^i) \right)$, *and* $\mathbf{r}_{e,r,k}^i = \mathbf{m}_i^{\sigma_e(r)} \odot \mathbf{H}_{e,r,k}^i \left(\mathbf{m}_i^{\sigma_e(r)} \odot \mathbf{w}_{e,r,k}^i - \mathbf{m}_i^{\sigma_e(r)} \odot \mathbf{w}_{e,r} \right)$, $\mathbf{H}_{e,r,k}^i \in \mathbb{R}^{d \times d}$ *are some real matrix such that* $\mathbf{H}_{e,r,k}^i \preceq L\mathbf{I}$.

Proof. According to updating rule we have

$$\begin{aligned} \mathbf{w}_{e,r+1} &= \mathbf{w}_{e,r} - \eta \frac{1}{N} \sum_{i=1}^{N} \sum_{k=0}^{K-1} \mathbf{m}_i^{\sigma_e(r)} \odot \nabla f_i(\mathbf{m}_i^{\sigma_e(r)} \odot \mathbf{w}_{e,r,k}; \xi_{e,r,k}^i) \\ &= \mathbf{w}_{e,r} - \eta \frac{1}{N} \sum_{i=1}^{N} \sum_{k=0}^{K-1} \mathbf{m}_i^{\sigma_e(r)} \odot \nabla f_i(\mathbf{m}_i^{\sigma_e(r)} \odot \mathbf{w}_{e,r,k}) - \boldsymbol{\xi}_{e,r} \end{aligned}$$

Notice the following fact:

$$\begin{aligned} \mathbf{m}_i^{\sigma_e(r)} \odot \nabla f_i(\mathbf{m}_i^{\sigma_e(r)} \odot \mathbf{w}_{e,r,k}^i) &= \mathbf{m}_i^{\sigma_e(r)} \odot \nabla f_i(\mathbf{m}_i^{\sigma_e(r)} \odot \mathbf{w}_{e,r}) \\ &\quad + \mathbf{m}_i^{\sigma_e(r)} \odot \nabla f_i(\mathbf{m}_i^{\sigma_e(r)} \odot \mathbf{w}_{e,r,k}^i) - \mathbf{m}_i^{\sigma_e(r)} \odot \nabla f_i(\mathbf{m}_i^{\sigma_e(r)} \odot \mathbf{w}_{e,r}) \\ &= \mathbf{m}_i^{\sigma_e(r)} \odot \nabla f_i(\mathbf{m}_i^{\sigma_e(r)} \odot \mathbf{w}_{e,r}) \\ &\quad + \underbrace{\mathbf{m}_i^{\sigma_e(r)} \odot \mathbf{H}_{e,r,k}^i \left(\mathbf{m}_i^{\sigma_e(r)} \odot \mathbf{w}_{e,r,k}^i - \mathbf{m}_i^{\sigma_e(r)} \odot \mathbf{w}_{e,r} \right)}_{\mathbf{r}_{e,r,k}^i}. \end{aligned}$$

Hence we can conclude that:

$$\mathbf{w}_{e,r+1} = \mathbf{w}_{e,r} - \eta \frac{1}{N} \sum_{i=1}^{N} K \mathbf{m}_i^{\sigma_e(r)} \odot \nabla f_i(\mathbf{m}_i^{\sigma_e(r)} \odot \mathbf{w}_{e,r}) - \eta \frac{1}{N} \sum_{i=1}^{N} \sum_{k=0}^{K-1} \mathbf{r}_{e,r,k}^i - \boldsymbol{\xi}_{e,r}.$$

□

Lemma 10 (Recursion between each epoch)**.** *For Algorithm 2, under the assumptions of Theorem 3, the following statement holds:*

$$\begin{aligned} \mathbf{w}_{e,R} - \mathbf{w}_e = &-\eta RK \nabla F_{\mathbf{m}}(\mathbf{w}_e) - \sum_{r=0}^{R-1} \mathbf{A}_r \left(\eta \frac{1}{N} \sum_{i=1}^{N} \sum_{k=0}^{K-1} \mathbf{r}_{e,r,k}^i + \boldsymbol{\xi}_{e,r} \right) \\ &+ \eta^2 K^2 \sum_{r=0}^{R-2} \mathbf{A}_{r+1} \frac{1}{N} \sum_{i=1}^{N} \mathbf{M}_i^{\sigma_e(r+1)} \mathbf{H}_{e,r+1}^i \mathbf{M}_i^{\sigma_e(r+1)} \sum_{j=0}^{r} \frac{1}{N} \sum_{i=1}^{N} \nabla f_i^{\sigma_e(j)}(\mathbf{w}_e), \end{aligned}$$

where $\mathbf{A}_r := \prod_{r'=R-1}^{r+1} \left(\mathbf{I} - \eta K \frac{1}{N} \sum_{i=1}^{N} \mathbf{M}_i^{\sigma_e(r')} \mathbf{H}_{e,r'}^i \mathbf{M}_i^{\sigma_e(r')} \right)$.

Proof. Define $f_i^{\sigma_e(r)}(\mathbf{w}) := f_i(\mathbf{m}_i^{\sigma_e(r)} \odot \mathbf{w})$. According to Lemma 9 we have

$$\begin{aligned}
\mathbf{w}_{e,r+1} - \mathbf{w}_e &= \mathbf{w}_{e,r} - \mathbf{w}_e - \eta \frac{1}{N} \sum_{i=1}^{N} K \nabla f_i^{\sigma_e(r)}(\mathbf{w}_{e,r}) - \eta \frac{1}{N} \sum_{i=1}^{N} \sum_{k=0}^{K-1} \mathbf{r}_{e,r,k}^i - \boldsymbol{\xi}_{e,r} \\
&= \mathbf{w}_{e,r} - \mathbf{w}_e - \eta K \frac{1}{N} \sum_{i=1}^{N} \nabla f_i^{\sigma_e(r)}(\mathbf{w}_e) \\
&\quad - \eta K \frac{1}{N} \sum_{i=1}^{N} \left(\nabla f_i^{\sigma_e(r)}(\mathbf{w}_{e,r}) - \nabla f_i^{\sigma_e(r)}(\mathbf{w}_e) \right) - \eta \frac{1}{N} \sum_{i=1}^{N} \sum_{k=0}^{K-1} \mathbf{r}_{e,r,k}^i - \boldsymbol{\xi}_{e,r}.
\end{aligned}$$

Applying mean-value theorem on the $\nabla f_i^{\sigma_e(r)}(\cdot)$ yields:

$$\begin{aligned}
\mathbf{w}_{e,r+1} - \mathbf{w}_e &= \mathbf{w}_{e,r} - \mathbf{w}_e - \eta K \frac{1}{N} \sum_{i=1}^{N} \nabla f_i^{\sigma_e(r)}(\mathbf{w}_e) \\
&\quad - \eta K \frac{1}{N} \sum_{i=1}^{N} \mathbf{m}_i^{\sigma_e(r)} \odot \mathbf{H}_{e,r}^i \mathbf{m}_i^{\sigma_e(r)} \odot (\mathbf{w}_{e,r} - \mathbf{w}_e) - \eta \frac{1}{N} \sum_{i=1}^{N} \sum_{k=0}^{K-1} \mathbf{r}_{e,r,k}^i - \boldsymbol{\xi}_{e,r} \\
&= \left(\mathbf{I} - \eta K \frac{1}{N} \sum_{i=1}^{N} \mathbf{M}_i^{\sigma_e(r)} \mathbf{H}_{e,r}^i \mathbf{M}_i^{\sigma_e(r)} \right) (\mathbf{w}_{e,r} - \mathbf{w}_e) - \eta K \frac{1}{N} \sum_{i=1}^{N} \nabla f_i^{\sigma_e(r)}(\mathbf{w}_e) \\
&\quad - \eta \frac{1}{N} \sum_{i=1}^{N} \sum_{k=0}^{K-1} \mathbf{r}_{e,r,k}^i - \boldsymbol{\xi}_{e,r}.
\end{aligned} \tag{10}$$

Unrolling the recursion from $r = R-1$ to 0 yields:

$$\mathbf{w}_{e,R} - \mathbf{w}_e = \sum_{r=0}^{R-1} \mathbf{A}_r \left(-\eta K \frac{1}{N} \sum_{i=1}^{N} \nabla f_i^{\sigma_e(r)}(\mathbf{w}_e) - \eta \frac{1}{N} \sum_{i=1}^{N} \sum_{k=0}^{K-1} \mathbf{r}_{e,r,k}^i - \boldsymbol{\xi}_{e,r} \right).$$

According to summation by part$\sum_{r=0}^{R-1} \mathbf{A}_r \mathbf{b}_r = \mathbf{A}_{R-1} \sum_{r=0}^{R-1} \mathbf{b}_r - \sum_{i=0}^{R-2} (\mathbf{A}_{i+1} - \mathbf{A}_i) \sum_{j=0}^{i} \mathbf{b}_j$ we have

$$\begin{aligned}
\mathbf{w}_{e,R} - \mathbf{w}_e &= - \sum_{r=0}^{R-1} \left(\eta K \frac{1}{N} \sum_{i=1}^{N} \nabla f_i^{\sigma_e(r)}(\mathbf{w}_e) \right) + \eta K \sum_{i=0}^{R-2} (\mathbf{A}_{i+1} - \mathbf{A}_i) \sum_{j=0}^{i} \frac{1}{N} \sum_{i=1}^{N} \nabla f_i^{\sigma_e(j)}(\mathbf{w}_e) \\
&\quad - \sum_{r=0}^{R-1} \mathbf{A}_r \left(\eta \frac{1}{N} \sum_{i=1}^{N} \sum_{k=0}^{K-1} \mathbf{r}_{e,r,k}^i + \boldsymbol{\xi}_{e,r} \right) \\
&= -\eta R K \nabla F_{\mathbf{m}}(\mathbf{w}_e) + \eta K \sum_{r=0}^{R-2} (\mathbf{A}_{r+1} - \mathbf{A}_r) \sum_{j=0}^{r} \frac{1}{N} \sum_{i=1}^{N} \nabla f_i^{\sigma_e(j)}(\mathbf{w}_e) \\
&\quad - \sum_{r=0}^{R-1} \mathbf{A}_r \left(\eta \frac{1}{N} \sum_{i=1}^{N} \sum_{k=0}^{K-1} \mathbf{r}_{e,r,k}^i + \boldsymbol{\xi}_{e,r} \right) \\
&= -\eta R K \nabla F_{\mathbf{m}}(\mathbf{w}_e) \\
&\quad + \eta^2 K^2 \sum_{r=0}^{R-2} \mathbf{A}_{r+1} \frac{1}{N} \sum_{i=1}^{N} \mathbf{M}_i^{\sigma_e(r+1)} \mathbf{H}_{e,r+1}^i \mathbf{M}_i^{\sigma_e(r+1)} \sum_{j=0}^{r} \frac{1}{N} \sum_{i=1}^{N} \nabla f_i^{\sigma_e(j)}(\mathbf{w}_e) \\
&\quad - \sum_{r=0}^{R-1} \mathbf{A}_r \left(\eta \frac{1}{N} \sum_{i=1}^{N} \sum_{k=0}^{K-1} \mathbf{r}_{e,r,k}^i + \boldsymbol{\xi}_{e,r} \right).
\end{aligned}$$

□

Lemma 11 (Local model deviation)**.** *For Algorithm 2, under the assumptions of Theorem 3, the following statement holds with probability at least* $1-\nu$*:*

$$\begin{aligned}
&\sum_{r=0}^{R-1}\sum_{k=0}^{K-1}\frac{1}{N}\sum_{i=1}^{N}\mathbb{E}\left\|\mathbf{w}_{e,r}-\mathbf{w}_{e,r,k+1}^{i}\right\|^2\\
&\leq 48\eta^2RK^3\zeta+\left(12\left(6R^3\eta^4K^5L^2+18R^5\eta^6K^7L^4\right)+48\eta^2RK^3\right)\mathbb{E}\left\|\nabla F_{\mathbf{m}}(\mathbf{w}_e)\right\|^2\\
&\quad+12\left(48R^2\eta^4K^5L^2+432R^4\eta^6K^7L^4\right)G\log(2RK/\nu)\\
&\quad+216R^3\eta^4K^5L^2\frac{\delta^2}{N}+6\eta^2RK^3\delta^2.
\end{aligned}$$

and for any $i\in[N]$ *we have*

$$\begin{aligned}
&\sum_{r=0}^{R-1}\sum_{k=0}^{K-1}\mathbb{E}\left\|\mathbf{w}_{e,r}-\mathbf{w}_{e,r,k+1}^{i}\right\|^2\\
&\leq 48\eta^2RK^3\zeta_i+\left(12\left(6R^3\eta^4K^5L^2+18R^5\eta^6K^7L^4\right)+48\eta^2RK^3\right)\mathbb{E}\left\|\nabla F_{\mathbf{m}}(\mathbf{w}_e)\right\|^2\\
&+12\left(48R^2\eta^4K^5L^2+432R^4\eta^6K^7L^4\right)G\log(2RK/\nu)\\
&+216R^3\eta^4K^5L^2\frac{\delta^2}{N}+6\eta^2RK^3\delta^2.
\end{aligned}$$

Proof. According to updating rule we have:

$$\begin{aligned}
&\mathbb{E}\left\|\mathbf{w}_{e,r}-\mathbf{w}_{e,r,k+1}^{i}\right\|^2\\
&=\left(1+\frac{1}{K-1}\right)\mathbb{E}\left\|\mathbf{w}_{e,r}-\mathbf{w}_{e,r,k}^{i}\right\|^2+\eta^2K\mathbb{E}\left\|\mathbf{m}_i^{\sigma_e(r)}\odot\nabla f_i(\mathbf{m}_i^{\sigma_e(r)}\odot\mathbf{w}_{e,r,k}^{i};\xi_{e,r,k}^{i})\right\|^2\\
&=\left(1+\frac{1}{K-1}\right)\mathbb{E}\left\|\mathbf{w}_{e,r}-\mathbf{w}_{e,r,k}^{i}\right\|^2+\eta^2K\mathbb{E}\left\|\mathbf{m}_i^{\sigma_e(r)}\odot\nabla f_i(\mathbf{m}_i^{\sigma_e(r)}\odot\mathbf{w}_{e,r,k}^{i};\xi_{e,r,k}^{i})\right\|^2+\eta^2K\delta^2\\
&\leq\left(1+\frac{1}{K-1}\right)\mathbb{E}\left\|\mathbf{w}_{e,r}-\mathbf{w}_{e,r,k}^{i}\right\|^2+2\eta^2K\mathbb{E}\left\|\mathbf{m}_i^{\sigma_e(r)}\odot\nabla f_i(\mathbf{m}_i^{\sigma_e(r)}\odot\mathbf{w}_{e,r})\right\|^2+\eta^2K\delta^2\\
&\quad+2\eta^2KL^2\mathbb{E}\left\|\mathbf{w}_{e,r,k}^{i}-\mathbf{w}_{e,r}\right\|^2\\
&\leq\left(1+\frac{1}{K-1}\right)\mathbb{E}\left\|\mathbf{w}_{e,r}-\mathbf{w}_{e,r,k}^{i}\right\|^2+4\eta^2K\mathbb{E}\left\|\mathbf{m}_i^{\sigma_e(r)}\odot\nabla f_i(\mathbf{m}_i^{\sigma_e(r)}\odot\mathbf{w}_{e})\right\|^2\\
&\quad+4\eta^2K\mathbb{E}\left\|\mathbf{m}_i^{\sigma_e(r)}\odot\nabla f_i(\mathbf{m}_i^{\sigma_e(r)}\odot\mathbf{w}_{e,r})-\mathbf{m}_i^{\sigma_e(r)}\odot\nabla f_i(\mathbf{m}_i^{\sigma_e(r)}\odot\mathbf{w}_{e})\right\|^2+\eta^2K\delta^2\\
&\quad+2\eta^2KL^2\mathbb{E}\left\|\mathbf{w}_{e,r,k}^{i}-\mathbf{w}_{e,r}\right\|^2\\
&\leq\left(1+\frac{1}{K-1}\right)\mathbb{E}\left\|\mathbf{w}_{e,r}-\mathbf{w}_{e,r,k}^{i}\right\|^2+4\eta^2K\mathbb{E}\left\|\mathbf{m}_i^{\sigma_e(r)}\odot\nabla f_i(\mathbf{m}_i^{\sigma_e(r)}\odot\mathbf{w}_{e})\right\|^2\\
&\quad+4\eta^2KL^2\mathbb{E}\left\|\mathbf{w}_{e,r}-\mathbf{w}_{e}\right\|^2+\eta^2K\delta^2+2\eta^2KL^2\mathbb{E}\left\|\mathbf{w}_{e,r,k}^{i}-\mathbf{w}_{e,r}\right\|^2.
\end{aligned}$$

Due to $2\eta^2KL^2\leq\frac{1}{K-1}$, we have

$$\begin{aligned}
\mathbb{E}\left\|\mathbf{w}_{e,r}-\mathbf{w}_{e,r,k+1}^{i}\right\|^2&\leq\left(1+\frac{2}{K-1}\right)\mathbb{E}\left\|\mathbf{w}_{e,r}-\mathbf{w}_{e,r,k}^{i}\right\|^2+4\eta^2K\mathbb{E}\left\|\nabla f_i^{\sigma_e(r)}(\mathbf{w}_e)\right\|^2 &(11)\\
&\quad+4\eta^2KL^2\mathbb{E}\left\|\mathbf{w}_{e,r}-\mathbf{w}_{e}\right\|^2+\eta^2K\delta^2. &(12)
\end{aligned}$$

Due to (10) we have

$$\begin{aligned}
\mathbf{w}_{e,r}-\mathbf{w}_e=&\left(\mathbf{I}-\eta K\frac{1}{N}\sum_{i=1}^{N}\mathbf{M}_i^{\sigma_e(r-1)}\mathbf{H}_{e,r-1}^{i}\mathbf{M}_i^{\sigma_e(r-1)}\right)(\mathbf{w}_{e,r-1}-\mathbf{w}_e)\\
&-\eta K\frac{1}{N}\sum_{i=1}^{N}\nabla f_i^{\sigma_e(r-1)}(\mathbf{w}_e)-\eta\frac{1}{N}\sum_{i=1}^{N}\sum_{k=0}^{K-1}\mathbf{r}_{e,r-1,k}^{i}-\boldsymbol{\xi}_{e,r-1},r=1,\ldots,R.
\end{aligned}$$

Unrolling the recursion yields:

$$\begin{aligned}
&\mathbf{w}_{e,r} - \mathbf{w}_e \\
&= \sum_{p=0}^{r-1} \underbrace{\prod_{r'=r-1}^{p+1} \left(\mathbf{I} - \eta K \frac{1}{N} \sum_{i=1}^{N} \mathbf{M}_i^{\sigma_e(r')} \mathbf{H}_{e,r'}^i \mathbf{M}_i^{\sigma_e(r')} \right)}_{\mathbf{A}_p} \\
&\cdot \left(-\eta K \frac{1}{N} \sum_{i=1}^{N} \nabla f_i^{\sigma_e(p)}(\mathbf{w}_e) - \eta \frac{1}{N} \sum_{i=1}^{N} \sum_{k=0}^{K-1} \mathbf{r}_{e,p,k}^i - \boldsymbol{\xi}_{e,p} \right).
\end{aligned}$$

According to summation by part$\sum_{p=0}^{r-1} \mathbf{A}_p \mathbf{b}_p = \mathbf{A}_{r-1} \sum_{j=0}^{r-1} \mathbf{b}_j - \sum_{i=0}^{r-2} (\mathbf{A}_{i+1} - \mathbf{A}_i) \sum_{j=0}^{i} \mathbf{b}_j$ we have

$$\begin{aligned}
&\mathbf{w}_{e,r} - \mathbf{w}_e \\
&= \sum_{p=0}^{r-1} \prod_{r'=r-1}^{p+1} \left(\mathbf{I} - \eta K \frac{1}{N} \sum_{i=1}^{N} \mathbf{M}_i^{\sigma_e(r')} \mathbf{H}_{e,r'}^i \mathbf{M}_i^{\sigma_e(r')} \right) \left(-\eta K \frac{1}{N} \sum_{i=1}^{N} \nabla f_i^{\sigma_e(p)}(\mathbf{w}_e) \right) \\
&\quad + \sum_{p=0}^{r-1} \prod_{r'=r-1}^{p+1} \left(\mathbf{I} - \eta K \frac{1}{N} \sum_{i=1}^{N} \mathbf{M}_i^{\sigma_e(r')} \mathbf{H}_{e,r'}^i \mathbf{M}_i^{\sigma_e(r')} \right) \left(-\eta \frac{1}{N} \sum_{i=1}^{N} \sum_{k=0}^{K-1} \mathbf{r}_{e,p,k}^i - \boldsymbol{\xi}_{e,p} \right) \\
&= \sum_{p=0}^{r-1} \left(-\eta K \frac{1}{N} \sum_{i=1}^{N} \nabla f_i^{\sigma_e(p)}(\mathbf{w}_e) \right) - \sum_{p=0}^{r-2} (\mathbf{A}_{p+1} - \mathbf{A}_p) \sum_{j=0}^{p} \left(-\eta K \frac{1}{N} \sum_{i=1}^{N} \nabla f_i^{\sigma_e(j)}(\mathbf{w}_e) \right) \\
&\quad + \sum_{p=0}^{r-1} \prod_{r'=r-1}^{p+1} \left(\mathbf{I} - \eta K \frac{1}{N} \sum_{i=1}^{N} \mathbf{M}_i^{\sigma_e(r')} \mathbf{H}_{e,r'}^i \mathbf{M}_i^{\sigma_e(r')} \right) \left(-\eta \frac{1}{N} \sum_{i=1}^{N} \sum_{k=0}^{K-1} \mathbf{r}_{e,p,k}^i - \boldsymbol{\xi}_{e,p} \right) \\
&= \sum_{p=0}^{r-1} \left(-\eta K \frac{1}{N} \sum_{i=1}^{N} \nabla f_i^{\sigma_e(p)}(\mathbf{w}_e) \right) \\
&+ \sum_{p=0}^{r-2} \prod_{r'=r-1}^{p+2} \left(\mathbf{I} - \eta K \frac{1}{N} \sum_{i=1}^{N} \mathbf{M}_i^{\sigma_e(r')} \mathbf{H}_{e,r'}^i \mathbf{M}_i^{\sigma_e(r')} \right) \\
&\quad \cdot \eta^2 K^2 \frac{1}{N} \sum_{i=1}^{N} \mathbf{M}_i^{\sigma_e(p+1)} \mathbf{H}_{e,p+1}^i \mathbf{M}_i^{\sigma_e(p+1)} \sum_{j=0}^{p} \left(\frac{1}{N} \sum_{i=1}^{N} \nabla f_i^{\sigma_e(j)}(\mathbf{w}_e) \right) \\
&+ \sum_{p=0}^{r-1} \prod_{r'=r-1}^{p+1} \left(\mathbf{I} - \eta K \frac{1}{N} \sum_{i=1}^{N} \mathbf{M}_i^{\sigma_e(r')} \mathbf{H}_{e,r'}^i \mathbf{M}_i^{\sigma_e(r')} \right) \left(-\eta \frac{1}{N} \sum_{i=1}^{N} \sum_{k=0}^{K-1} \mathbf{r}_{e,p,k}^i - \boldsymbol{\xi}_{e,p} \right).
\end{aligned}$$

Taking expected norm on both side yields:

$$
\begin{aligned}
&\mathbb{E}\left\|\mathbf{w}_{e,r}-\mathbf{w}_e\right\|^2 \\
&=3\eta^2K^2\mathbb{E}\left\|\sum_{p=0}^{r-1}\frac{1}{N}\sum_{i=1}^{N}\nabla f_i^{\sigma_e(p)}(\mathbf{w}_e)\right\|^2 \\
&\quad+3\eta^4K^4r\sum_{p=0}^{r-2}\prod_{r'=r-1}^{p+2}(1+\eta KL)^{2r'}L^2\mathbb{E}\left\|\sum_{j=0}^{p}\frac{1}{N}\sum_{i=1}^{N}\nabla f_i^{\sigma_e(j)}(\mathbf{w}_e)\right\|^2 \\
&\quad+3r\sum_{p=0}^{r-1}\prod_{r'=r-1}^{p+1}(1+\eta KL)^{2r'}\mathbb{E}\left\|\eta\frac{1}{N}\sum_{i=1}^{N}\sum_{k=0}^{K-1}\mathbf{r}_{e,p,k}^i+\boldsymbol{\xi}_{e,p}\right\|^2 \\
&\le 3\eta^2K^2\mathbb{E}\left\|\sum_{p=0}^{r-1}\frac{1}{N}\sum_{i=1}^{N}\nabla f_i^{\sigma_e(p)}(\mathbf{w}_e)\right\|^2+3\eta^4K^4r\sum_{p=0}^{r-2}9L^2\mathbb{E}\left\|\sum_{j=0}^{p}\frac{1}{N}\sum_{i=1}^{N}\nabla f_i^{\sigma_e(j)}(\mathbf{w}_e)\right\|^2 \\
&\quad+18\eta^2r\sum_{p=0}^{r-1}\frac{1}{N}\sum_{i=1}^{N}KL^2\sum_{k=0}^{K-1}\mathbb{E}\left\|\mathbf{w}_{e,p,k}^i-\mathbf{w}_{e,p}\right\|^2+18r\sum_{p=0}^{r-1}\mathbb{E}\left\|\boldsymbol{\xi}_{e,p}\right\|^2.
\end{aligned}
$$

According to Hoeffding-Serfling inequality (Schneider, 2016, Theorem 2) we know

$$
\begin{aligned}
&\left\|\sum_{p=0}^{r-1}\frac{1}{N}\sum_{i=1}^{N}\nabla f_i^{\sigma_e(p)}(\mathbf{w}_e)\right\|^2 \\
&\le 2r^2\left\|\nabla F_{\mathbf{m}}(\mathbf{w}_e)-\frac{1}{r}\sum_{p=0}^{r-1}\frac{1}{N}\sum_{i=1}^{N}\nabla f_i^{\sigma_e(p)}(\mathbf{w}_e)\right\|^2+2r^2\left\|\nabla F_{\mathbf{m}}(\mathbf{w}_e)\right\|^2 \\
&\le 2r^2G\frac{8(1-\frac{r-1}{N})\log(2/\nu)}{r}+2r^2\left\|\nabla F_{\mathbf{m}}(\mathbf{w}_e)\right\|^2.
\end{aligned}
$$

Similarly we know

$$
\left\|\sum_{j=0}^{p}\frac{1}{N}\sum_{i=1}^{N}\nabla f_i^{\sigma_e(j)}(\mathbf{w}_e)\right\|^2\le 2(p+1)^2G\frac{8\log(2RK/\nu)}{p+1}+2(p+1)^2\left\|\nabla F_{\mathbf{m}}(\mathbf{w}_e)\right\|^2.
$$

For $\mathbb{E}\left\|\boldsymbol{\xi}_{e,p}\right\|^2$, we have

$$
\begin{aligned}
&\mathbb{E}\left\|\boldsymbol{\xi}_{e,p}\right\|^2 \\
&=\mathbb{E}\left\|\eta\frac{1}{N}\sum_{i=1}^{N}\sum_{k=0}^{K-1}\left(\mathbf{m}_i^{\sigma_e(p)}\odot\nabla f_i(\mathbf{m}_i^{\sigma_e(p)}\odot\mathbf{w}_{e,p,k})-\mathbf{m}_i^{\sigma_e(p)}\odot\nabla f_i(\mathbf{m}_i^{\sigma_e(p)}\odot\mathbf{w}_{e,p,k};\xi_{e,p,k}^i)\right)\right\|^2 \\
&\le\eta^2K^2\frac{\delta^2}{N}.
\end{aligned}
$$

Putting piece together yields:

$$
\begin{aligned}
&\mathbb{E}\left\|\mathbf{w}_{e,r}-\mathbf{w}_e\right\|^2 \\
&\leq 3\eta^2K^2\left(16rG\log(2/\nu)+2r^2\left\|\nabla F_{\mathbf{m}}(\mathbf{w}_e)\right\|^2\right) \\
&\quad+3\eta^4K^4r\sum_{p=0}^{r-2}9L^2\left(16(p+1)G\log(2RK/\nu)+2(p+1)^2\left\|\nabla F_{\mathbf{m}}(\mathbf{w}_e)\right\|^2\right) \\
&\quad+18\eta^2r\sum_{p=0}^{r-1}KL^2\sum_{k=0}^{K-1}\frac{1}{N}\sum_{i=1}^{N}\mathbb{E}\left\|\mathbf{w}_{e,p,k}^i-\mathbf{w}_{e,p}\right\|^2+18r^2\eta^2K^2\frac{\delta^2}{N} \\
&\leq 3\eta^2K^2\left(16rG\log(2RK/\nu)+2r^2\left\|\nabla F_{\mathbf{m}}(\mathbf{w}_e)\right\|^2\right) \\
&\quad+27\eta^4K^4rL^2\left(16r^2G\log(2RK/\nu)+\frac{2r^3}{3}\left\|\nabla F_{\mathbf{m}}(\mathbf{w}_e)\right\|^2\right) \\
&\quad+18\eta^2r\sum_{p=0}^{r-1}KL^2\sum_{k=0}^{K-1}\frac{1}{N}\sum_{i=1}^{N}\mathbb{E}\left\|\mathbf{w}_{e,p,k}^i-\mathbf{w}_{e,p}\right\|^2+18r^2\eta^2K^2\frac{\delta^2}{N} \\
&=\left(48r\eta^2K^2+432r^3\eta^4K^4L^2\right)G\log(2RK/\nu)+\left(6r^2\eta^2K^2+18r^4\eta^4K^4L^2\right)\left\|\nabla F_{\mathbf{m}}(\mathbf{w}_e)\right\|^2 \\
&\quad+18\eta^2r\sum_{p=0}^{r-1}KL^2\sum_{k=0}^{K-1}\frac{1}{N}\sum_{i=1}^{N}\mathbb{E}\left\|\mathbf{w}_{e,p,k}^i-\mathbf{w}_{e,p}\right\|^2+18r^2\eta^2K^2\frac{\delta^2}{N}. \qquad (14)
\end{aligned}
$$

Plugging above bound back to (12) yields:

$$
\begin{aligned}
&\mathbb{E}\left\|\mathbf{w}_{e,r}-\mathbf{w}_{e,r,k+1}^i\right\|^2 \\
&\leq\left(1+\frac{2}{K-1}\right)\mathbb{E}\left\|\mathbf{w}_{e,r}-\mathbf{w}_{e,r,k}^i\right\|^2+4\eta^2K\mathbb{E}\left\|\nabla f_i^{\sigma_e(r)}(\mathbf{w}_e)\right\|^2+\eta^2K\delta^2 \\
&\quad+4\eta^2KL^2\left(\left(48r\eta^2K^2+432r^3\eta^4K^4L^2\right)G\log(2RK/\nu)+\left(6r^2\eta^2K^2+18r^4\eta^4K^4L^2\right)\left\|\nabla F_{\mathbf{m}}(\mathbf{w}_e)\right\|^2\right) \\
&\quad+4\eta^2KL^2\left(18\eta^2r\sum_{p=0}^{r-1}KL^2\sum_{k=0}^{K-1}\frac{1}{N}\sum_{i=1}^{N}\mathbb{E}\left\|\mathbf{w}_{e,p,k}^i-\mathbf{w}_{e,p}\right\|^2+18r^2\eta^2K^2\frac{\delta^2}{N}\right).
\end{aligned}
$$

Unrolling the recursion from k to 0 yields

$$
\begin{aligned}
&\mathbb{E}\left\|\mathbf{w}_{e,r}-\mathbf{w}_{e,r,k+1}^i\right\|^2\leq\sum_{k'=0}^{k}\left(1+\frac{2}{K-1}\right)^{k'}\left(4\eta^2K\mathbb{E}\left\|\nabla f_i^{\sigma_e(r)}(\mathbf{w}_e)\right\|^2+\eta^2K\delta^2\right) \\
&+\sum_{k'=0}^{k}\left(1+\frac{2}{K-1}\right)^{k'}4\eta^2KL^2 \\
&\quad\cdot\left(\left(48r\eta^2K^2+432r^3\eta^4K^4L^2\right)G\log(2/\nu)+\left(6r^2\eta^2K^2+18r^4\eta^4K^4L^2\right)\left\|\nabla F_{\mathbf{m}}(\mathbf{w}_e)\right\|^2\right) \\
&+\sum_{k'=0}^{k}\left(1+\frac{2}{K-1}\right)^{k'}4\eta^2KL^2\left(18\eta^2r\sum_{p=0}^{r-1}KL^2\sum_{k=0}^{K-1}\frac{1}{N}\sum_{i=1}^{N}\mathbb{E}\left\|\mathbf{w}_{e,p,k}^i-\mathbf{w}_{e,p}\right\|^2+18r^2\eta^2K^2\frac{\delta^2}{N}\right) \\
&\leq 12\eta^2K^2\mathbb{E}\left\|\nabla f_i^{\sigma_e(r)}(\mathbf{w}_e)\right\|^2+3\eta^2K^2\delta^2 \\
&+12\eta^2K^2L^2\left(\left(48R\eta^2K^2+432R^3\eta^4K^4L^2\right)G\log(2/\nu)\right) \\
&+12\eta^2K^2L^2\left(\left(6R^2\eta^2K^2+18R^4\eta^4K^4L^2\right)\left\|\nabla F_{\mathbf{m}}(\mathbf{w}_e)\right\|^2\right) \\
&+12\eta^2K^2L^2\left(18\eta^2R\sum_{p=0}^{R-1}KL^2\sum_{k=0}^{K-1}\frac{1}{N}\sum_{i=1}^{N}\mathbb{E}\left\|\mathbf{w}_{e,p,k}^i-\mathbf{w}_{e,p}\right\|^2+18R^2\eta^2K^2\frac{\delta^2}{N}\right).
\end{aligned}
$$

We further sum above inequality over $k=0$ to $K-1$ and get

$$\sum_{r=0}^{R-1}\sum_{k=0}^{K-1}\frac{1}{N}\sum_{i=1}^{N}\mathbb{E}\left\|\mathbf{w}_{e,r}-\mathbf{w}_{e,r,k+1}^{i}\right\|^2 \leq 12\eta^2K^3\sum_{r=0}^{R-1}\frac{1}{N}\sum_{i=1}^{N}\mathbb{E}\left\|\nabla f_i^{\sigma_e(r)}(\mathbf{w}_e)\right\|^2 + 3\eta^2RK^3\delta^2$$
$$+ 12\eta^2RK^3L^2\left(\left(48R\eta^2K^2+72R^3\eta^4K^4L^2\right)G\log(2/\nu)+\left(6R^2\eta^2K^2+6R^4\eta^4K^4L^2\right)\left\|\nabla F_{\mathbf{m}}(\mathbf{w}_e)\right\|^2\right)$$
$$+ 12\eta^2RK^3L^2\left(18\eta^2RKL^2\sum_{p=0}^{R-1}\sum_{k=0}^{K-1}\frac{1}{N}\sum_{i=1}^{N}\mathbb{E}\left\|\mathbf{w}_{e,p,k}^{i}-\mathbf{w}_{e,p}\right\|^2+18R^2\eta^2K^2\frac{\delta^2}{N}\right).$$

Re-arranging the terms yields:

$$\underbrace{\left(1-216\eta^4R^2K^4L^4\right)}_{\geq\frac{1}{2}}\sum_{r=0}^{R-1}\sum_{k=0}^{K-1}\frac{1}{N}\sum_{i=1}^{N}\mathbb{E}\left\|\mathbf{w}_{e,r}-\mathbf{w}_{e,r,k+1}^{i}\right\|^2 \leq 12\eta^2K^3\sum_{r=0}^{R-1}\frac{1}{N}\sum_{i=1}^{N}\mathbb{E}\left\|\nabla f_i^{\sigma_e(r)}(\mathbf{w}_e)\right\|^2$$
$$+ 3\eta^2RK^3\delta^2$$
$$+ 12\eta^2RK^3L^2\left(\left(48R\eta^2K^2+432R^3\eta^4K^4L^2\right)G\log(2RK/\nu)\right)$$
$$+ 12\eta^2RK^3L^2\left(\left(6R^2\eta^2K^2+18R^4\eta^4K^4L^2\right)\left\|\nabla F_{\mathbf{m}}(\mathbf{w}_e)\right\|^2\right)$$
$$+ 12\eta^2RK^3L^2\left(18R^2\eta^2K^2\frac{\delta^2}{N}\right).$$

Hence we conclude that

$$\sum_{r=0}^{R-1}\sum_{k=0}^{K-1}\frac{1}{N}\sum_{i=1}^{N}\mathbb{E}\left\|\mathbf{w}_{e,r}-\mathbf{w}_{e,r,k+1}^{i}\right\|^2 \leq 24\eta^2K^3\sum_{r=0}^{R-1}\frac{1}{N}\sum_{i=1}^{N}\mathbb{E}\left\|\nabla f_i^{\sigma_e(r)}(\mathbf{w}_e)\right\|^2 + 6\eta^2RK^3\delta^2$$
$$+ 12\left(48R^2\eta^4K^5L^2+432R^4\eta^6K^7L^4\right)G\log(2RK/\nu)$$
$$+ 12\left(6R^3\eta^4K^5L^2+18R^5\eta^6K^7L^4\right)\left\|\nabla F_{\mathbf{m}}(\mathbf{w}_e)\right\|^2$$
$$+ 216R^3\eta^4K^5L^2\frac{\delta^2}{N}.$$

Finally, recall the definition of gradient dissimilarity and get

$$\sum_{r=0}^{R-1}\frac{1}{N}\sum_{i=1}^{N}\mathbb{E}\left\|\nabla f_i^{\sigma_e(r)}(\mathbf{w}_e)\right\|^2 \leq 2\sum_{r=0}^{R-1}\frac{1}{N}\sum_{i=1}^{N}\mathbb{E}\left\|\nabla f_i^{\sigma_e(r)}(\mathbf{w}_e)-\nabla F(\mathbf{w}_e)\right\|^2 + 2R\mathbb{E}\left\|\nabla F(\mathbf{w}_e)\right\|^2,$$

which concludes the proof. The proof of second statement follows the same reasoning. □

D.1.1 Proof of Theorem 3

In the convergence analysis of Theorem 3, we account for the fact that the full model is partitioned into R sub-models. At the start of each epoch, the server shuffles these sub-models and assigns them sequentially to clients. This introduces complexity in analysis due to the interaction between both the model drift caused by partial training on sub-models and the impact of permutation-based assignments on convergence. Our technical contribution is to tackle these challenges jointly to establish the convergence, which could be interesting by its own. Here we briefly illustrate our proof strategy. Consider an alternative objective induced by a mask configuration $\mathbf{m}$:

$$F_{\mathbf{m}}(\mathbf{w}) = \frac{1}{N}\sum\nolimits_{i=1}^{N}\frac{1}{R}\sum\nolimits_{j=1}^{R} f_i(\mathbf{m}_i^j \odot \mathbf{w}) \tag{15}$$

and define $\mathbf{w}^*(\mathbf{m}) := \arg\min_{\mathbf{w}\in\mathcal{W}} F_{\mathbf{m}}(\mathbf{w})$ as the optimal model given a masking configuration $\mathbf{m} = \left[\left[\mathbf{m}_i^1,\ldots,\mathbf{m}_i^R\right]\right]_{i=1}^N \in \{0,1\}^{dNR}$. Apparently, when $\mathbf{m}=\mathbf{1}$, $F_{\mathbf{m}}(\mathbf{w})$ becomes original objective $F(\mathbf{w})$. Our proof relies on a key Lipschitz property of $\mathbf{w}^*(\mathbf{m})$. That is, if $f_i(\mathbf{w})$ is strongly convex and with bounded gradient, then

$$\left\|\mathbf{w}^*(\mathbf{m})-\mathbf{w}^*(\mathbf{1})\right\| \leq c\cdot\left\|\mathbf{m}-\mathbf{1}\right\|,$$

for some constant c. As a result, we can decompose the objective value into (1) convergence error to $\mathbf{w}^*(\mathbf{m})$ and (2) residual error due to masking:

$$F(\hat{\mathbf{w}}) - F(\mathbf{w}^*) \leq 2L\|\tilde{\mathbf{w}} - \mathbf{w}^*(\mathbf{m})\|^2 + \left(\frac{2L}{\mu} + \frac{4}{L}\right)\frac{2G^2 + 2W^2L^2}{NR}\|\mathbf{m} - \mathbf{1}\|^2.$$

Then the heart of the proof is to prove the convergence of Algorithm 2 to $\mathbf{w}^*(\mathbf{m})$. Algorithm 2 performs a variant of Local SGD where each client shuffles its local component function $f_i(\mathbf{m}_i^1 \odot \mathbf{w}), \ldots, f_i(\mathbf{m}_i^R \odot \mathbf{w})$, and conduct local updates on one of them for K steps during one communication round.

Having outlined the key ideas, we turn to rigorously establish the convergence rate.

Proof. Due to updating rule we know

$$\begin{aligned}
\mathbb{E}\left\|\mathbf{w}_{e+1} - \mathbf{w}^*\right\|^2 &= \mathbb{E}\left\|\mathcal{P}_{\mathcal{W}}(\mathbf{w}_{e,R}) - \mathbf{w}^*\right\|^2 \\
&\leq \mathbb{E}\left\|\mathbf{w}_{e,R} - \mathbf{w}_e + \mathbf{w}_e - \mathbf{w}^*\right\|^2 \\
&\leq \mathbb{E}\left\|\mathbf{w}_e - \mathbf{w}^*\right\|^2 + 2\left\langle \mathbf{w}_e - \mathbf{w}^*,\, -\eta K \nabla F_{\mathbf{m}}(\mathbf{w}_e)\right\rangle \\
&\quad + 2\left\langle \mathbf{w}_e - \mathbf{w}^*,\, \eta^2 K^2 \sum_{r=0}^{R-2} \mathbf{A}_{r+2} \frac{1}{N}\sum_{i=1}^{N} \mathbf{M}_i^{\sigma_e(r)} \mathbf{H}_{e,r}^i \mathbf{M}_i^{\sigma_e(r)} \sum_{j=0}^{r} \frac{1}{N}\sum_{i=1}^{N} \nabla f_i^{\sigma_e(j)}(\mathbf{w}_e)\right\rangle \\
&\quad + 2\left\langle \mathbf{w}_e - \mathbf{w}^*,\, \sum_{r=0}^{R-1} \mathbf{A}_r \left(\eta \frac{1}{N}\sum_{i=1}^{N}\sum_{k=0}^{K-1} \mathbf{r}_{e,r,k}^i + \boldsymbol{\xi}_{e,r}\right)\right\rangle + \mathbb{E}\left\|\mathbf{w}_e - \mathbf{w}_{e,R}\right\|^2,
\end{aligned}$$

where we plug in Lemma 10.

Applying Cauchy-Schwartz inequality and taking the expectation over randomness of ξ yields:

$$\begin{aligned}
&\mathbb{E}\left\|\mathbf{w}_{e+1} - \mathbf{w}^*\right\|^2 \\
&\leq \mathbb{E}\left\|\mathbf{w}_e - \mathbf{w}^*\right\|^2 + 2\left\langle \mathbf{w}_e - \mathbf{w}^*,\, -\eta RK \nabla F_{\mathbf{m}}(\mathbf{w}_e)\right\rangle \\
&\quad + 2\left\|\mathbf{w}_e - \mathbf{w}^*\right\| \left\|\eta^2 K^2 \sum_{r=0}^{R-2} \mathbf{A}_{r+2} \frac{1}{N}\sum_{i=1}^{N} \mathbf{M}_i^{\sigma_e(r)} \mathbf{H}_{e,r}^i \mathbf{M}_i^{\sigma_e(r)} \sum_{j=0}^{r} \frac{1}{N}\sum_{i=1}^{N} \nabla f_i^{\sigma_e(j)}(\mathbf{w}_e)\right\| \\
&\quad + 2\left\|\mathbf{w}_e - \mathbf{w}^*\right\| \left\|\sum_{r=0}^{R-1} \mathbf{A}_r \left(\eta \frac{1}{N}\sum_{i=1}^{N}\sum_{k=0}^{K-1} \mathbf{r}_{e,r,k}^i + \boldsymbol{\xi}_{e,r}\right)\right\| + \mathbb{E}\left\|\mathbf{w}_e - \mathbf{w}_{e,R}\right\|^2 \\
&\leq (1 - \mu\eta RK)\mathbb{E}\left\|\mathbf{w}_e - \mathbf{w}^*\right\|^2 - \underbrace{\eta RK\left(F_{\mathbf{m}}(\mathbf{w}_e) - F_{\mathbf{m}}(\tilde{\mathbf{w}}^*)\right)}_{\geq 0} \\
&\quad \underbrace{+ 2\eta^2 K^2 \left\|\mathbf{w}_e - \mathbf{w}^*\right\| \left(\sum_{r=0}^{R-2} (1 + \eta KL)^R L \left\|\sum_{j=0}^{r} \frac{1}{N}\sum_{i=1}^{N} \nabla f_i^{\sigma_e(j)}(\mathbf{w}_e)\right\|\right)}_{T_1} \\
&\quad \underbrace{+ 2\left\|\mathbf{w}_e - \mathbf{w}^*\right\| \left(\sum_{r=0}^{R-1} (1 + \eta KL)^R \left\|\eta \frac{1}{N}\sum_{i=1}^{N}\sum_{k=0}^{K-1} \mathbf{r}_{e,r,k}^i\right\|\right)}_{T_2} + \mathbb{E}\left\|\mathbf{w}_e - \mathbf{w}_{e,R}\right\|^2.
\end{aligned}$$

Notice that we choose η such that $\eta KL \leq \frac{1}{R}$, so we know $(1 + \eta KL)^R \leq e \leq 3$.

To bound T_1, we again use Hoeffding-Serfling inequality:

$$\begin{aligned}
&\sum_{r=0}^{R-2} 3L \left\| \sum_{j=0}^{r} \frac{1}{N} \sum_{i=1}^{N} \nabla f_i^{\sigma_e(j)}(\mathbf{w}_e) \right\| \\
&\leq \sum_{r=0}^{R-2} 3L \left(\left\| \nabla F_{\mathbf{m}}(\mathbf{w}_e) - \frac{1}{r} \sum_{j=0}^{r} \frac{1}{N} \sum_{i=1}^{N} \nabla f_i^{\sigma_e(j)}(\mathbf{w}_e) \right\| + r \left\| \nabla F_{\mathbf{m}}(\mathbf{w}_e) \right\| \right) \\
&\leq \sum_{r=0}^{R-2} 3L \left(\sqrt{G \frac{8 \log(2RK/\nu)}{r}} + r \left\| \nabla F_{\mathbf{m}}(\mathbf{w}_e) \right\| \right) \\
&\leq 3L \left(\frac{2R^{3/2}}{3} \sqrt{G 8 \log(2RK/\nu)} + R^2 \left\| \nabla F_{\mathbf{m}}(\mathbf{w}_e) \right\| \right).
\end{aligned}$$

Hence we can bound T_1 as:

$$\begin{aligned}
T_1 &\leq 2\eta^2 K^2 \left\| \mathbf{w}_e - \tilde{\mathbf{w}}^* \right\| \left(3L \left(\frac{2R^{3/2}}{3} \sqrt{G 8 \log(2RK/\nu)} + R^2 \left\| \nabla F_{\mathbf{m}}(\mathbf{w}_e) \right\| \right) \right) \\
&= 4L\eta^2 K^2 \left\| \mathbf{w}_e - \tilde{\mathbf{w}}^* \right\| R^{3/2} \sqrt{8G \log(2RK/\nu)} + 3L\eta^2 R^2 K^2 \left\| \nabla F_{\mathbf{m}}(\mathbf{w}_e) \right\|^2 \\
&\leq 4L\eta^2 K^2 \left(\frac{1}{4\sqrt{\eta LK}} R^{1/2} \left\| \mathbf{w}_e - \tilde{\mathbf{w}}^* \right\| \cdot 4\sqrt{\eta LK} R \sqrt{G 8 \log(2RK/\nu)} \right) \\
&\quad + 6L\eta^2 R^2 K^2 \left\| \mathbf{w}_e - \tilde{\mathbf{w}}^* \right\| \left\| \nabla F_{\mathbf{m}}(\mathbf{w}_e) \right\| \\
&\leq 4L\eta^2 K^2 \left(\frac{1}{16\eta LK} R \left\| \mathbf{w}_e - \mathbf{w}^* \right\|^2 + 4\eta L R^2 K G \log(2RK/\nu) \right) \\
&\quad + 6L\eta^2 R^2 K^2 \left\| \mathbf{w}_e - \tilde{\mathbf{w}}^* \right\| \left\| \nabla F_{\mathbf{m}}(\mathbf{w}_e) \right\| \\
&= \frac{1}{4} \eta RK \left\| \mathbf{w}_e - \tilde{\mathbf{w}}^* \right\|^2 + 16\eta^3 K^3 L^2 R^2 G \log(2RK/\nu) + 3L\eta^2 R^2 K^2 \left\| \nabla F_{\mathbf{m}}(\mathbf{w}_e) \right\|^2 \\
&\quad + 3L\eta^2 R^2 K^2 \left\| \mathbf{w}_e - \tilde{\mathbf{w}}^* \right\|^2.
\end{aligned}$$

To bound T_2 we notice

$$\begin{aligned}
T_2 &\leq 6\eta \left(\frac{1}{4} \sqrt{RK} \left\| \mathbf{w}_e - \tilde{\mathbf{w}}^* \right\| \cdot 4 \frac{1}{\sqrt{RK}} \left(\sum_{r=0}^{R-1} \sum_{k=0}^{K-1} L \frac{1}{N} \sum_{i=1}^{N} \left\| \mathbf{w}_{e,r,k}^i - \mathbf{w}_{e,r} \right\| \right) \right) \\
&\leq \frac{6}{32} \eta RK \left\| \mathbf{w}_e - \tilde{\mathbf{w}}^* \right\|^2 + \frac{48\eta}{RK} \left(\sum_{r=0}^{R-1} \sum_{k=0}^{K-1} L \frac{1}{N} \sum_{i=1}^{N} \left\| \mathbf{w}_{e,r,k}^i - \mathbf{w}_{e,r} \right\| \right)^2 \\
&\leq \frac{3}{8} \eta RK \left\| \mathbf{w}_e - \tilde{\mathbf{w}}^* \right\|^2 + 48\eta \sum_{r=0}^{R-1} \sum_{k=0}^{K-1} L \frac{1}{N} \sum_{i=1}^{N} \left\| \mathbf{w}_{e,r,k}^i - \mathbf{w}_{e,r} \right\|^2.
\end{aligned}$$

We plug in Lemma 11 and get

$$\begin{aligned}
&48\eta L \sum_{r=0}^{R-1} \sum_{k=0}^{K-1} \frac{1}{N} \sum_{i=1}^{N} \left\| \mathbf{w}_{e,r,k}^i - \mathbf{w}_{e,r} \right\|^2 \\
&\leq 48^2 \eta^3 LRK^3 \zeta + 48\eta L \left(12 \left(6R^3 \eta^4 K^5 L^2 + 18 R^5 \eta^6 K^7 L^4 \right) + 48 \eta^2 RK^3 \right) \mathbb{E} \left\| \nabla F_{\mathbf{m}}(\mathbf{w}_e) \right\|^2 \\
&\quad + 576\eta L \left(48 R^2 \eta^4 K^5 L^2 + 432 R^4 \eta^6 K^7 L^4 \right) G \log(2RK/\nu) \\
&\quad + 10368 R^3 \eta^5 K^5 L^3 \frac{\delta^2}{N} + 288 L \eta^3 RK^3 \delta^2.
\end{aligned}$$

Putting pieces together yields:

$$\begin{aligned}
&\mathbb{E}\left\|\mathbf{w}_{e+1}-\mathbf{w}^*\right\|^2 \\
&\leq \left(1-\frac{3}{8}\mu\eta RK-3L\eta^2R^2K^2\right)\mathbb{E}\left\|\mathbf{w}_e-\mathbf{w}^*\right\|^2-\eta RK\left(F_{\mathbf{m}}(\mathbf{w}_e)-F_{\mathbf{m}}(\tilde{\mathbf{w}}^*)\right) \\
&\quad+\left(16\eta^3K^3L^2R^2+576\eta L\left(48R^2\eta^4K^5L^2+432R^4\eta^6K^7L^4\right)\right)G\log(2RK/\nu) \\
&\quad+\mathbb{E}\left\|\mathbf{w}_e-\mathbf{w}_{e,R}\right\|^2+48^2\eta^3LRK^3\zeta \\
&\quad+\left(3456R^3\eta^5K^5L^3+10368R^5\eta^7K^7L^5+2304L\eta^3RK^3+3L\eta^2R^2K^2\right)\mathbb{E}\left\|\nabla F_{\mathbf{m}}(\mathbf{w}_e)\right\|^2 \\
&\quad+10368R^3\eta^5K^5L^3\frac{\delta^2}{N}+288L\eta^3RK^3\delta^2.
\end{aligned}$$

Due to our choice of $\eta=\frac{2\log(T^2)}{\mu RKT}$ and $T\geq 128\kappa L\log T$,
we know $\eta RK\leq\frac{2\log(T^2)}{\mu RK512\kappa\log T^2}RK=\frac{1}{256L^2}$

$$\begin{aligned}
3\eta^2R^2K^2L&\leq\frac{1}{256L}\eta RK\\
2304L\eta^3RK^3&\leq\frac{2304}{256^2L}\eta RK\\
3456R^3\eta^5K^5L^3&\leq\frac{3456}{256^4L}\eta RK\\
10368R^5\eta^7K^7L^5&\leq\frac{10368}{256^6L}\eta RK.
\end{aligned}$$

Hence we have:

$$\begin{aligned}
&\mathbb{E}\left\|\mathbf{w}_{e+1}-\mathbf{w}^*\right\|^2 \\
&\leq \left(1-\frac{3}{8}\mu\eta RK-3L\eta^2R^2K^2\right)\mathbb{E}\left\|\mathbf{w}_e-\mathbf{w}^*\right\|^2-\eta RK\left(F_{\mathbf{m}}(\mathbf{w}_e)-F_{\mathbf{m}}(\tilde{\mathbf{w}}^*)\right) \\
&\quad+\left(16\eta^3K^3L^2R^2+576\eta L\left(48R^2\eta^4K^5L^2+432R^4\eta^6K^7L^4\right)\right)G\log(2RK/\nu) \\
&\quad+\mathbb{E}\left\|\mathbf{w}_e-\mathbf{w}_{e,R}\right\|^2+48^2\eta^3LRK^3\zeta \\
&\quad+\frac{\eta RK}{16L}\mathbb{E}\left\|\nabla F_{\mathbf{m}}(\mathbf{w}_e)\right\|^2 \\
&\quad+10368R^3\eta^5K^5L^3\frac{\delta^2}{N}+288L\eta^3RK^3\delta^2 \\
&\leq \left(1-\frac{3}{8}\mu\eta RK-3L\eta^2R^2K^2\right)\mathbb{E}\left\|\mathbf{w}_e-\mathbf{w}^*\right\|^2 \\
&\quad-\frac{7}{8}\eta RK\left(F_{\mathbf{m}}(\mathbf{w}_e)-F_{\mathbf{m}}(\tilde{\mathbf{w}}^*)\right) \\
&\quad+\left(16\eta^3K^3L^2R^2+576\eta L\left(48R^2\eta^4K^5L^2+432R^4\eta^6K^7L^4\right)\right)G\log(2RK/\nu) \\
&\quad+\mathbb{E}\left\|\mathbf{w}_e-\mathbf{w}_{e,R}\right\|^2+48^2\eta^3LRK^3\zeta+10368R^3\eta^5K^5L^3\frac{\delta^2}{N}+288L\eta^3RK^3\delta^2,
\end{aligned}$$

where at last step we use the L smoothness of $F_{\mathbf{m}}$ such that
$\left\|\nabla F_{\mathbf{m}}(\mathbf{w}_e)\right\|^2\leq 2L\left(F_{\mathbf{m}}(\mathbf{w}_e)-F_{\mathbf{m}}(\tilde{\mathbf{w}}^*)\right)$. It remains to bound $\left\|\mathbf{w}_{e+1}-\mathbf{w}_e\right\|^2$.

We again evoke Lemma 10:

$$
\begin{aligned}
&\mathbb{E}\left\|\mathbf{w}_{e,R}-\mathbf{w}_e\right\|^2\\
&\leq 3\left\|\eta RK\nabla F_{\mathbf{m}}(\mathbf{w}_e)\right\|^2\\
&+3\eta^4K^4R\sum_{r=0}^{R-2}\prod_{r'=R-1}^{r+2}(1+\eta KL)^{2r'}L^2\mathbb{E}\left\|\sum_{j=0}^{r}\frac{1}{N}\sum_{i=1}^{N}\nabla f_i^{\sigma_e(j)}(\mathbf{w}_e)\right\|^2\\
&\quad+3\mathbb{E}\left\|\sum_{r=0}^{R-1}\mathbf{A}_r\left(\eta\frac{1}{N}\sum_{i=1}^{N}\sum_{k=0}^{K-1}\mathbf{r}_{e,r,k}^i+\boldsymbol{\xi}_{e,r}\right)\right\|^2\\
&\leq 3\mathbb{E}\left\|\eta RK\nabla F_{\mathbf{m}}(\mathbf{w}_e)\right\|^2+27L^2\eta^4K^4R\sum_{r=0}^{R-2}\mathbb{E}\left\|\sum_{j=0}^{r}\frac{1}{N}\sum_{i=1}^{N}\nabla f_i^{\sigma_e(j)}(\mathbf{w}_e)\right\|^2\\
&\quad+27R\sum_{r=0}^{R-1}\mathbb{E}\left\|\eta\frac{1}{N}\sum_{i=1}^{N}\sum_{k=0}^{K-1}\mathbf{r}_{e,r,k}^i\right\|^2+27LR\sum_{r=0}^{R-1}\mathbb{E}\left\|\boldsymbol{\xi}_{e,r}\right\|^2\\
&\leq 3\mathbb{E}\left\|\eta RK\nabla F_{\mathbf{m}}(\mathbf{w}_e)\right\|^2+27L^2\eta^4K^4R\underbrace{\sum_{r=0}^{R-2}\mathbb{E}\left\|\sum_{j=0}^{r}\frac{1}{N}\sum_{i=1}^{N}\nabla f_i^{\sigma_e(j)}(\mathbf{w}_e)\right\|^2}_{T_1}\\
&\quad+27\eta^2RKL^2\underbrace{\sum_{r=0}^{R-1}\frac{1}{N}\sum_{i=1}^{N}\sum_{k=0}^{K-1}\mathbb{E}\left\|\mathbf{w}_{e,r,k}^i-\mathbf{w}_{e,r}^i\right\|^2}_{T_2}+27\eta^2R^2K^2\frac{\delta^2}{N}.
\end{aligned}
$$

For T_1, we know that

$$
\begin{aligned}
T_1&\leq 2\sum_{r=0}^{R-2}r^2\mathbb{E}\left\|\nabla F_{\mathbf{m}}(\mathbf{w}_e)\right\|^2\\
&\quad+2\sum_{r=0}^{R-2}r^2\mathbb{E}\left\|\frac{1}{r}\sum_{j=0}^{r}\frac{1}{N}\sum_{i=1}^{N}\nabla f_i^{\sigma_e(j)}(\mathbf{w}_e)-\nabla F_{\mathbf{m}}(\mathbf{w}_e)\right\|^2\\
&\leq 2\frac{R^3}{3}\mathbb{E}\left\|\nabla F_{\mathbf{m}}(\mathbf{w}_e)\right\|^2+2\sum_{r=0}^{R-2}r^2 8G\frac{\log(2RK/\nu)}{r}\\
&\leq 2R^3\mathbb{E}\left\|\nabla F_{\mathbf{m}}(\mathbf{w}_e)\right\|^2+16R^2G\log(2RK/\nu).
\end{aligned}
$$

with probability at least $1-\nu$.

For T_2, we again evoke Lemma 11:

$$
\begin{aligned}
T_2\leq&48\eta^2RK^3\zeta+\left(12\left(6R^3\eta^4K^5L^2+18R^5\eta^6K^7L^4\right)+48\eta^2RK^3\right)\mathbb{E}\left\|\nabla F_{\mathbf{m}}(\mathbf{w}_e)\right\|^2\\
&+12\left(48R^2\eta^4K^5L^2+432R^4\eta^6K^7L^4\right)G\log(2RK/\nu)\\
&+216R^3\eta^4K^5L^2\frac{\delta^2}{N}+6\eta^2RK^3\delta^2.
\end{aligned}
$$

Putting pieces together yields:

$$
\begin{aligned}
&\mathbb{E}\left\|\mathbf{w}_{e+1}-\mathbf{w}_e\right\|^2 \\
&\leq 3\mathbb{E}\left\|\eta RK\nabla F_{\mathbf{m}}(\mathbf{w}_e)\right\|^2+27L^2\eta^4K^4R\left(2R^3\mathbb{E}\left\|\nabla F_{\mathbf{m}}(\mathbf{w}_e)\right\|^2+16R^2G\log(2RK/\nu)\right) \\
&\quad+27\eta^2RKL^2\left(48\eta^2RK^3\zeta+\left(12\left(6R^3\eta^4K^5L^2+18R^5\eta^6K^7L^4\right)\right)\mathbb{E}\left\|\nabla F_{\mathbf{m}}(\mathbf{w}_e)\right\|^2\right) \\
&\quad+27\eta^2RKL^2\left(48\eta^2RK^3\right)\mathbb{E}\left\|\nabla F_{\mathbf{m}}(\mathbf{w}_e)\right\|^2 \\
&\quad+27\eta^2RKL^2\left(12\left(48R^2\eta^4K^5L^2+432R^4\eta^6K^7L^4\right)G\log(2RK/\nu)\right) \\
&\quad+27\eta^2RKL^2\left(216R^3\eta^4K^5L^2\frac{\delta^2}{N}+6\eta^2RK^3\delta^2\right)+27\eta^2R^2K^2\frac{\delta^2}{N} \\
&=\left(3\eta^2R^2K^2+54L^2\eta^4R^4K^4+1944R^4\eta^6K^6L^4+5832R^6\eta^8K^8L^6\right)\mathbb{E}\left\|\nabla F_{\mathbf{m}}(\mathbf{w}_e)\right\| \\
&+1296\eta^4R^2K^4L^2\mathbb{E}\left\|\nabla F_{\mathbf{m}}(\mathbf{w}_e)\right\|^2 \\
&\quad+\left(27\eta^2RKL^212\left(48R^2\eta^4K^5L^2+432R^4\eta^6K^7L^4\right)+432L^2\eta^4R^3K^4\right)G\log(2RK/\nu) \\
&\quad+27\eta^2RKL^2\left(216R^3\eta^4K^5L^2\frac{\delta^2}{N}+6\eta^2RK^3\delta^2\right)+27\eta^2R^2K^2\frac{\delta^2}{N}+1296\eta^4R^2K^4L^2\zeta.
\end{aligned}
$$

Due to our choice of $\eta=\frac{4\log(T^2)}{\mu RKT}$ and $T\geq 512\kappa^2\log T$,
we know $\eta RK\leq\frac{2\log(T^2)}{\mu RK512\kappa^2\log T^2}RK=\frac{1}{256\kappa L}$

$$
\begin{aligned}
3\eta^2R^2K^2&\leq\frac{3}{256L}\eta RK\leq\frac{1}{16L}\eta RK\\
54L^2\eta^4R^4K^4&\leq\frac{54}{256^3L}\eta RK\leq\frac{1}{16L}\eta RK\\
1944R^4\eta^6K^6L^4&\leq\frac{1944}{256^5L}\eta RK\leq\frac{1}{16L}\eta RK\\
5832R^6\eta^8K^8L^6&\leq\frac{5832}{256^7L}\eta RK\leq\frac{1}{16L}\eta RK\\
1296\eta^4R^2K^4L^2&\leq\frac{1296}{256^3L}\eta RK\leq\frac{1}{16L}\eta RK.
\end{aligned}
$$

Hence we know

$$
\begin{aligned}
\mathbb{E}\left\|\mathbf{w}_{e+1}-\mathbf{w}_e\right\|^2\leq&\frac{5}{16L}\eta RK\mathbb{E}\left\|\nabla F_{\mathbf{m}}(\mathbf{w}_e)\right\|^2\\
&+\left(324\eta^2RKL^2\left(48R^2\eta^4K^5L^2+432R^4\eta^6K^7L^4\right)+432L^2\eta^4R^3K^4\right)G\log(2RK/\nu)\\
&+27\eta^2RKL^2\left(216R^3\eta^4K^5L^2\frac{\delta^2}{N}+6\eta^2RK^3\delta^2\right)+27\eta^2R^2K^2\frac{\delta^2}{N}\\
&+1296\eta^4R^2K^4L^2\zeta^2.
\end{aligned}
\tag{16}
$$

Putting pieces together yields:

$$
\begin{aligned}
\mathbb{E}\left\|\mathbf{w}_{e+1}-\mathbf{w}^*\right\|^2\leq&\left(1-\frac{3}{8}\mu\eta RK+3L\eta^2R^2K^2\right)\mathbb{E}\left\|\mathbf{w}_e-\mathbf{w}^*\right\|^2-\frac{1}{4}\eta RK\underbrace{\left(F_{\mathbf{m}}(\mathbf{w}_e)-F_{\mathbf{m}}(\tilde{\mathbf{w}}^*)\right)}_{\geq 0}\\
&+\left(16\eta^3K^3L^2R^2+576\eta L\left(48R^2\eta^4K^5L^2+432R^4\eta^6K^7L^4\right)\right)G\log(2RK/\nu)\\
&+48^2\eta^3LRK^3\zeta+10368R^3\eta^5K^5L^3\frac{\delta^2}{N}+288L\eta^3RK^3\delta^2+27\eta^2R^2K^2\frac{\delta^2}{N}\\
&+\left(324\eta^2RKL^2\left(48R^2\eta^4K^5L^2+432R^4\eta^6K^7L^4\right)+432L^2\eta^4R^3K^4\right)G\log(2RK/\nu)\\
&+27\eta^2RKL^2\left(216R^3\eta^4K^5L^2\frac{\delta^2}{N}+6\eta^2RK^3\delta^2\right)\\
&+1296\eta^4R^2K^4L^2\zeta^2.
\end{aligned}
$$

Due to our choice of η, we know $3L\eta^2R^2K^2 \leq 3L\eta RK\frac{\mu}{256L^2} \leq \frac{1}{8}\mu\eta RK$. We unroll the recursion from T to 0 to get:

$$\begin{aligned}
\mathbb{E}\left\|\mathbf{w}_{e+1}-\mathbf{w}^*\right\|^2 \leq & \left(1-\frac{1}{4}\mu\eta RK\right)^T \mathbb{E}\left\|\mathbf{w}_0-\mathbf{w}^*\right\|^2 \\
& + 4\left(16\eta^2K^2\kappa LR + 576\kappa\left(48R\eta^4K^4L^2 + 432R^3\eta^6K^6L^4\right)\right)G\log(2RK/\nu) \\
& + 4\cdot 48^2\eta^2\kappa K^2\zeta + 4\cdot 10368R^2\eta^4K^4\kappa L^2\frac{\delta^2}{N} + 4\cdot 288\kappa\eta^2K^2\delta^2 + 4\cdot 27\eta RK\frac{\delta^2}{\mu N} \\
& + 4\left(324\eta\kappa L\left(48R^2\eta^4K^5L^2 + 432R^4\eta^6K^7L^4\right) + 432\kappa L\eta^3R^2K^3\right)G\log(2RK/\nu) \\
& + 4\cdot 27\eta\kappa L\left(216R^3\eta^4K^5L^2\frac{\delta^2}{N} + 6\eta^2RK^3\delta^2\right) \\
& + 4\cdot 1296\eta^3RK^3\kappa L\zeta.
\end{aligned}$$

Plugging $\eta = \frac{4\log T^2}{\mu TRK}$ yields:

$$\mathbb{E}\left\|\mathbf{w}_T-\mathbf{w}^*\right\|^2 \leq O\left(\frac{\mathbb{E}\left\|\mathbf{w}_0-\mathbf{w}^*\right\|^2}{T^2}\right) + O\left(\frac{\kappa^2\log(2RK/\nu)}{\mu T^2R}\right) + O\left(\frac{\kappa\zeta\log^2(T)}{\mu^2T^2R^2}\right) + O\left(\frac{\delta^2\log(T)}{\mu^2TN}\right).$$

□

D.2 Proof of Convergence in Nonconvex Setting (Theorem 4)

In this section we will present the proof of convergence of Algorithm 2 in nonconvex setting. Since constrained non-convex stochastic optimization suffers from residual noise error unless a large mini-batch is used (Ghadimi et al., 2016), here we assume an unconstrained setting, i.e., $\mathcal{W} = \mathbb{R}^d$.

Proof. From the smoothness of $F_{\mathbf{m}}$ and Lemma 10, we have

$$\begin{aligned}
F_{\mathbf{m}}(\mathbf{w}_{e+1}) &\leq F_{\mathbf{m}}(\mathbf{w}_e) + \langle\nabla F_{\mathbf{m}}(\mathbf{w}_e),\, \mathbf{w}_{e+1}-\mathbf{w}_e\rangle + \frac{L}{2}\left\|\mathbf{w}_{e+1}-\mathbf{w}_e\right\|^2 \\
&\leq F_{\mathbf{m}}(\mathbf{w}_e) - \eta RK\left\|\nabla F_{\mathbf{m}}(\mathbf{w}_e)\right\|^2 + \langle\nabla F_{\mathbf{m}}(\mathbf{w}_e),\, \mathbf{g}_e\rangle + \frac{L}{2}\left\|\mathbf{w}_{e+1}-\mathbf{w}_e\right\|^2
\end{aligned}$$

where

$$\begin{aligned}
\mathbf{g}_e = & \eta^2K^2\sum_{r=0}^{R-2}\mathbf{A}_{r+1}\frac{1}{N}\sum_{i=1}^{N}\mathbf{M}_i^{\sigma_e(r+1)}\mathbf{H}_{e,r+1}^i\mathbf{M}_i^{\sigma_e(r+1)}\sum_{j=0}^{r}\frac{1}{N}\sum_{i=1}^{N}\nabla f_i^{\sigma_e(j)}(\mathbf{w}_e) \\
& - \sum_{r=0}^{R-1}\mathbf{A}_r\left(\eta\frac{1}{N}\sum_{i=1}^{N}\sum_{k=0}^{K-1}\mathbf{r}_{e,r,k}^i + \boldsymbol{\xi}_{e,r}\right).
\end{aligned}$$

Taking expectation over randomness of samples yields:

$$\begin{aligned}
&\mathbb{E}[F_{\mathbf{m}}(\mathbf{w}_{e+1})] \\
&\leq \mathbb{E}[F_{\mathbf{m}}(\mathbf{w}_e)] - \eta RK\mathbb{E}\left\|\nabla F_{\mathbf{m}}(\mathbf{w}_e)\right\|^2 + \mathbb{E}\left\langle \nabla F_{\mathbf{m}}(\mathbf{w}_e),\, \mathbf{h}_e\right\rangle + \frac{L}{2}\mathbb{E}\left\|\mathbf{w}_{e+1} - \mathbf{w}_e\right\|^2 \\
&\leq \mathbb{E}[F_{\mathbf{m}}(\mathbf{w}_e)] - \eta RK\mathbb{E}\left\|\nabla F_{\mathbf{m}}(\mathbf{w}_e)\right\|^2 \\
&+ \underbrace{\mathbb{E}\left\langle \nabla F_{\mathbf{m}}(\mathbf{w}_e),\, \eta^2K^2\mathbf{h}_e\right\rangle}_{T_1} + \underbrace{\mathbb{E}\left\langle \nabla F_{\mathbf{m}}(\mathbf{w}_e),\, -\sum_{r=0}^{R-1}\mathbf{A}_r\left(\eta\frac{1}{N}\sum_{i=1}^{N}\sum_{k=0}^{K-1}\mathbf{r}_{e,r,k}^i\right)\right\rangle}_{T_2} \\
&\quad + \frac{L}{2}\mathbb{E}\left\|\mathbf{w}_{e+1} - \mathbf{w}_e\right\|^2
\end{aligned}$$

where

$$\begin{aligned}
\mathbf{h}_e = \sum_{r=0}^{R-2}\prod_{r'=R-1}^{r+2}\left(\mathbf{I} - \eta K\frac{1}{N}\sum_{i=1}^{N}\mathbf{M}_i^{\sigma_e(r')}\mathbf{H}_{e,r'}^i\mathbf{M}_i^{\sigma_e(r')}\right) \\
\cdot\frac{1}{N}\sum_{i=1}^{N}\mathbf{M}_i^{\sigma_e(r+1)}\mathbf{H}_{e,r+1}^i\mathbf{M}_i^{\sigma_e(r+1)}\sum_{j=0}^{r}\frac{1}{N}\sum_{i=1}^{N}\nabla f_i^{\sigma_e(j)}(\mathbf{w}_e).
\end{aligned}$$

For T_1:

$$T_1 = \mathbb{E}\left\langle \nabla F_{\mathbf{m}}(\mathbf{w}_e),\, \eta^2K^2\mathbf{h}_e\right\rangle \leq \eta^2K^2\mathbb{E}\left[\frac{1}{\sqrt{\eta K}}\left\|\nabla F_{\mathbf{m}}(\mathbf{w}_e)\right\|\sqrt{\eta K}\left\|\mathbf{h}_e\right\|\right].$$

Then we bound $\|\mathbf{h}_e\|$ as

$$\left\|\mathbf{h}_e\right\| \leq \sum_{r=0}^{R-2}(1+\eta KL)^R L\left\|\sum_{j=0}^{r}\frac{1}{N}\sum_{i=1}^{N}\nabla f_i^{\sigma_e(j)}(\mathbf{w}_e)\right\|.$$

and applying Hoeffding-Serfling inequality (Schneider, 2016, Theorem 2) gives:

$$\begin{aligned}
&\sum_{r=0}^{R-2}L\left\|\sum_{j=0}^{r}\frac{1}{N}\sum_{i=1}^{N}\nabla f_i^{\sigma_e(j)}(\mathbf{w}_e)\right\| \\
&\leq \sum_{r=0}^{R-2}L\left(\left\|\nabla F_{\mathbf{m}}(\mathbf{w}_e) - \frac{1}{r}\sum_{j=0}^{r}\frac{1}{N}\sum_{i=1}^{N}\nabla f_i^{\sigma_e(j)}(\mathbf{w}_e)\right\| + r\left\|\nabla F_{\mathbf{m}}(\mathbf{w}_e)\right\|\right) \\
&\leq \sum_{r=0}^{R-2}L\left(\sqrt{G\frac{8\log(2RK/\nu)}{r}} + r\left\|\nabla F_{\mathbf{m}}(\mathbf{w}_e)\right\|\right) \\
&\leq L\left(\frac{2R^{3/2}}{3}\sqrt{G8\log(2RK/\nu)} + R^2\left\|\nabla F_{\mathbf{m}}(\mathbf{w}_e)\right\|\right).
\end{aligned}$$

Putting pieces together yields:

$$
\begin{aligned}
T_1 &\leq \eta^2 K^2 \mathbb{E}\left[\|\nabla F_{\mathbf{m}}(\mathbf{w}_e)\| \, \|\mathbf{h}_e\|\right] \\
&\leq \eta^2 K^2 \mathbb{E}\left[\|\nabla F_{\mathbf{m}}(\mathbf{w}_e)\| \, 3L \left(\frac{2R^{3/2}}{3}\sqrt{G8\log(2RK/\nu)} + R^2 \|\nabla F_{\mathbf{m}}(\mathbf{w}_e)\|\right)\right] \\
&\leq 2L\mathbb{E}\left[\sqrt{\frac{\eta KR}{4L}} \|\nabla F_{\mathbf{m}}(\mathbf{w}_e)\| \left(4\sqrt{L}\eta^{3/2}K^{3/2}R\sqrt{G8\log(2RK/\nu)}\right)\right] \\
&\quad + 3L\eta^2 R^2 K^2 \mathbb{E}\left[\|\nabla F_{\mathbf{m}}(\mathbf{w}_e)\|^2\right] \\
&\leq L\mathbb{E}\left[\frac{\eta KR}{4L}\|\nabla F_{\mathbf{m}}(\mathbf{w}_e)\|^2 + \left(128L\eta^3K^3R^2G\log(2RK/\nu)\right)\right] \\
&\quad + 3L\eta^2 R^2 K^2 \mathbb{E}\left[\|\nabla F_{\mathbf{m}}(\mathbf{w}_e)\|^2\right] \\
&= \left(\frac{\eta KR}{4} + 3L\eta^2R^2K^2\right)\mathbb{E}\left[\|\nabla F_{\mathbf{m}}(\mathbf{w}_e)\|^2\right] + 128L^2\eta^3K^3R^2G\log(2RK/\nu).
\end{aligned}
$$

For T_2 we have:

$$
\begin{aligned}
T_2 &= \mathbb{E}\left\langle \nabla F_{\mathbf{m}}(\mathbf{w}_e), -\sum_{r=0}^{R-1}\mathbf{A}_r\left(\eta\frac{1}{N}\sum_{i=1}^{N}\sum_{k=0}^{K-1}\mathbf{r}_{e,r,k}^i\right)\right\rangle \\
&\leq \eta\mathbb{E}\left[\sqrt{RK}\|\nabla F_{\mathbf{m}}(\mathbf{w}_e)\|\frac{1}{\sqrt{RK}}\left\|\sum_{r=0}^{R-1}\mathbf{A}_r\left(\frac{1}{N}\sum_{i=1}^{N}\sum_{k=0}^{K-1}\mathbf{r}_{e,r,k}^i\right)\right\|\right] \\
&\leq \frac{1}{4}\eta RK\mathbb{E}\|\nabla F_{\mathbf{m}}(\mathbf{w}_e)\|^2 + \eta\frac{1}{RK}\mathbb{E}\left\|\sum_{r=0}^{R-1}\mathbf{A}_r\left(\frac{1}{N}\sum_{i=1}^{N}\sum_{k=0}^{K-1}\mathbf{r}_{e,r,k}^i\right)\right\|^2.
\end{aligned}
$$

For $\mathbb{E}\left\|\sum_{r=0}^{R-1}\mathbf{A}_r\left(\frac{1}{N}\sum_{i=1}^{N}\sum_{k=0}^{K-1}\mathbf{r}_{e,r,k}^i\right)\right\|^2$ we have

$$
\begin{aligned}
\mathbb{E}\left\|\sum_{r=0}^{R-1}\mathbf{A}_r\left(\frac{1}{N}\sum_{i=1}^{N}\sum_{k=0}^{K-1}\mathbf{r}_{e,r,k}^i\right)\right\|^2 &\leq R\sum_{r=0}^{R-1}(1+\eta KL)^{2R}\left\|\frac{1}{N}\sum_{i=1}^{N}\sum_{k=0}^{K-1}\mathbf{r}_{e,r,k}^i\right\|^2 \\
&\leq RK\sum_{r=0}^{R-1}\sum_{k=0}^{K-1}(1+\eta KL)^{2R}\frac{1}{N}\sum_{i=1}^{N}\left\|\mathbf{w}_{e,r,k}^i - \mathbf{w}_{e,r}\right\|^2.
\end{aligned}
$$

We plug in Lemma 11 to get

$$
\begin{aligned}
&\mathbb{E}\left\|\sum_{r=0}^{R-1}\mathbf{A}_r\left(\frac{1}{N}\sum_{i=1}^{N}\sum_{k=0}^{K-1}\mathbf{r}_{e,r,k}^{i}\right)\right\|^2 \\
&\leq RK\sum_{r=0}^{R-1}\sum_{k=0}^{K-1}(1+\eta KL)^{2R} \\
&\cdot\left(48\eta^2RK^3\zeta+\left(12\left(6R^3\eta^4K^5L^2+18R^5\eta^6K^7L^4\right)\mathbb{E}\left\|\nabla F_{\mathbf{m}}(\mathbf{w}_e)\right\|^2\right)\right) \\
&\quad+(1+\eta KL)^{2R}\left(48\eta^2RK^3\right)\mathbb{E}\left\|\nabla F_{\mathbf{m}}(\mathbf{w}_e)\right\|^2 \\
&\quad+RK\sum_{r=0}^{R-1}\sum_{k=0}^{K-1}(1+\eta KL)^{2R}12\left(48R^2\eta^4K^5L^2+432R^4\eta^6K^7L^4\right)G\log(2RK/\nu) \\
&\quad+RK\sum_{r=0}^{R-1}\sum_{k=0}^{K-1}(1+\eta KL)^{2R}\left(216R^3\eta^4K^5L^2\frac{\delta^2}{N}+6\eta^2RK^3\delta^2\right) \\
&\leq 9R^2K^2\left(48\eta^2RK^3\zeta+\left(12\left(6R^3\eta^4K^5L^2+18R^5\eta^6K^7L^4\right)+48\eta^2RK^3\right)\mathbb{E}\left\|\nabla F_{\mathbf{m}}(\mathbf{w}_e)\right\|^2\right) \\
&\quad+9R^2K^212\left(48R^2\eta^4K^5L^2+432R^4\eta^6K^7L^4\right)G\log(2RK/\nu) \\
&\quad+9R^2K^2\left(216R^3\eta^4K^5L^2\frac{\delta^2}{N}+6\eta^2RK^3\delta^2\right).
\end{aligned}
$$

Putting pieces together yields:

$$
\begin{aligned}
T_2&\leq\frac{1}{4}\eta RK\mathbb{E}\left\|\nabla F_{\mathbf{m}}(\mathbf{w}_e)\right\|^2+\eta\frac{1}{RK}\mathbb{E}\left\|\sum_{r=0}^{R-1}\mathbf{A}_r\left(\frac{1}{N}\sum_{i=1}^{N}\sum_{k=0}^{K-1}\mathbf{r}_{e,r,k}^{i}\right)\right\|^2 \\
&\leq\frac{1}{4}\eta RK\mathbb{E}\left\|\nabla F_{\mathbf{m}}(\mathbf{w}_e)\right\|^2 \\
&\quad+9\eta RK\left(48\eta^2RK^3\zeta\right)\mathbb{E}\left\|\nabla F_{\mathbf{m}}(\mathbf{w}_e)\right\|^2 \\
&\quad+\left(12\left(6R^3\eta^4K^5L^2+18R^5\eta^6K^7L^4\right)+48\eta^2RK^3\right)\mathbb{E}\left\|\nabla F_{\mathbf{m}}(\mathbf{w}_e)\right\|^2 \\
&\quad+9\eta RK12\left(48R^2\eta^4K^5L^2+432R^4\eta^6K^7L^4\right)G\log(2RK/\nu) \\
&\quad+9\eta RK\left(216R^3\eta^4K^5L^2\frac{\delta^2}{N}+6\eta^2RK^3\delta^2\right).
\end{aligned}
$$

Plugging T_1 and T_2 back yields:

$$
\begin{aligned}
&\mathbb{E}[F_{\mathbf{m}}(\mathbf{w}_{e+1})] \\
&\leq\mathbb{E}[F_{\mathbf{m}}(\mathbf{w}_e)] \\
&\quad-\left(\frac{1}{2}\eta RK-3L\eta^2R^2K^2\right)\mathbb{E}\left\|\nabla F_{\mathbf{m}}(\mathbf{w}_e)\right\|^2 \\
&\quad-9\eta RK\left(12\left(6R^3\eta^4K^5L^2+18R^5\eta^6K^7L^4\right)\mathbb{E}\left\|\nabla F_{\mathbf{m}}(\mathbf{w}_e)\right\|^2\right)\mathbb{E}\left\|\nabla F_{\mathbf{m}}(\mathbf{w}_e)\right\|^2 \\
&\quad+9\eta RK\left(48\eta^2RK^3\right)\mathbb{E}\left\|\nabla F_{\mathbf{m}}(\mathbf{w}_e)\right\|^2 \\
&\quad+128L^2\eta^3K^3R^2G\log(2RK/\nu) \\
&\quad+9\eta RK\cdot48\eta^2RK^3\zeta+9\eta RK12\left(48R^2\eta^4K^5L^2+432R^4\eta^6K^7L^4\right)G\log(2RK/\nu) \\
&\quad+9\eta RK\left(216R^3\eta^4K^5L^2\frac{\delta^2}{N}+6\eta^2RK^3\delta^2\right)+\frac{L}{2}\mathbb{E}\left\|\mathbf{w}_{e+1}-\mathbf{w}_e\right\|^2.
\end{aligned}
$$

From (16) we know

$$\begin{aligned}
&\mathbb{E}\left\|\mathbf{w}_{e+1}-\mathbf{w}_e\right\|^2 \\
&\leq \frac{5}{16L}\eta RK\mathbb{E}\left\|\nabla F_{\mathbf{m}}(\mathbf{w}_e)\right\|^2 \\
&\quad + \left(324\eta^2 RKL^2\left(48R^2\eta^4K^5L^2+432R^4\eta^6K^7L^4\right)+432L^2\eta^4R^3K^4\right)G\log(2RK/\nu) \\
&\quad + 27\eta^2RKL^2\left(216R^3\eta^4K^5L^2\frac{\delta^2}{N}+6\eta^2RK^3\delta^2\right)+27\eta^2R^2K^2\frac{\delta^2}{N} \\
&\quad + 1296\eta^4R^2K^4L^2\zeta^2.
\end{aligned}$$

Putting pieces together yields:

$$\begin{aligned}
&\mathbb{E}[F_{\mathbf{m}}(\mathbf{w}_{e+1})] \\
&\leq \mathbb{E}[F_{\mathbf{m}}(\mathbf{w}_e)] \\
&\quad -\left(\frac{3}{16}\eta RK-3L\eta^2R^2K^2\right)\mathbb{E}\left\|\nabla F_{\mathbf{m}}(\mathbf{w}_e)\right\|^2 \\
&\quad -9\left(12\left(6R^4\eta^5K^6L^2+18R^6\eta^7K^8L^4\right)+48\eta^3R^2K^4\right)\mathbb{E}\left\|\nabla F_{\mathbf{m}}(\mathbf{w}_e)\right\|^2 \\
&\quad + 128L^2\eta^3K^3R^2G\log(2RK/\nu) \\
&\quad + 9\cdot 48\eta^3R^2K^4\zeta+108\left(48R^3\eta^5K^6L^2+432R^5\eta^7K^8L^4\right)G\log(2RK/\nu) \\
&\quad + 9\left(216R^4\eta^5K^6L^2\frac{\delta^2}{N}+6\eta^3R^2K^4\delta^2\right) \\
&\quad + \frac{L}{2}\left(324\eta^2RKL^2\left(48R^2\eta^4K^5L^2+432R^4\eta^6K^7L^4\right)\right)G\log(2RK/\nu) \\
&\quad + \frac{L}{2}432L^2\eta^4R^3K^4G\log(2RK/\nu) \\
&\quad + \frac{L}{2}27\eta^2RKL^2\left(216R^3\eta^4K^5L^2\frac{\delta^2}{N}+6\eta^2RK^3\delta^2\right)+27\eta^2R^2K^2\frac{\delta^2}{N} \\
&\quad + 648\eta^4R^2K^4L^3\zeta^2.
\end{aligned}$$

Since $\eta \leq \frac{1}{c\cdot L\sqrt{RKT}}$ for some sufficiently large c, we know

$$\begin{aligned}
3\eta^2R^2K^2L &\leq \frac{1}{64}\eta RK, \\
648\eta^5R^4K^6L^2 &\leq \frac{1}{64}\eta RK, \\
162\eta^7R^6K^8L^4 &\leq \frac{1}{64}\eta RK, \\
432\eta^3R^2K^4 &\leq \frac{1}{64}\eta RK.
\end{aligned}$$

Hence we have

$$\begin{aligned}
&\frac{1}{T}\sum_{e=1}^{T}\mathbb{E}\left\|\nabla F_{\mathbf{m}}(\mathbf{w}_e)\right\|^2 \\
&\leq O\left(\frac{\mathbb{E}[F_{\mathbf{m}}(\mathbf{w}_0)]}{\eta RKT}\right) \\
&\quad + O\left(L^2\eta^2K^2RG\log(2RK/\nu)+\eta^2RK^3\zeta^2+\eta RK\frac{\delta^2}{N}\right)
\end{aligned}$$

Since we choose $\eta = \Theta\left(\frac{1}{L\sqrt{RKT}}\right)$ we have

$$\begin{aligned}&\frac{1}{T}\sum_{e=1}^{T}\mathbb{E}\left\|\nabla F_{\mathbf{m}}(\mathbf{w}_e)\right\|^2\\&\leq O\left(\frac{L\mathbb{E}[F_{\mathbf{m}}(\mathbf{w}_0)]}{\sqrt{RKT}}+\frac{KG\log(2RK/\nu)}{T}+\frac{K^2\zeta^2}{TL^2}+\frac{\delta^2}{N\sqrt{RKT}L}\right).\end{aligned}$$

□

E STABILITY OF MASKED TRAINING METHOD

In this section, we will present the proof of Theorem 5. We define two helper quantities:

$$\zeta^2_{\mathbf{p},\max}=\max_{\mathbf{w}\in\mathbb{R}^d}\max_{i,j}\mathbb{E}_{\mathbf{m}_i}\mathbb{E}_{\mathbf{m}_j}\left\|\mathbf{m}_i\odot\nabla f_i(\mathbf{m}_i\odot\mathbf{w})-\mathbf{m}_j\odot\nabla f_j(\mathbf{m}_j\odot\mathbf{w})\right\|^2$$

and

$$\zeta^2_{\mathbf{m},\max}=\max_{\mathbf{w}\in\mathcal{W}}\max_{i,j,i',j'}\left\|\mathbf{m}_i^j\odot\nabla f_i(\mathbf{m}_i^j\odot\mathbf{w})-\mathbf{m}_{i'}^{j'}\odot\nabla f_{i'}(\mathbf{m}_{i'}^{j'}\odot\mathbf{w})\right\|^2.$$

We first introduce the following technical lemma.

Lemma 12. *Assuming that the partial derivative $\nabla_k\ell(x;\xi)$ is l_ℓ-Lipschitz continuous with respect to the data input ξ for any coordinate k. Then for Algorithm 1, the following statement holds true:*

$$\zeta^2_{p,\max}\leq\max_{i,j}\left[2p_id\cdot l_\ell^2W_1(\mathcal{D}_i,\mathcal{D}_j)^2+4d(p_i+p_j-2p_ip_j)\left(G_\infty^2+L^2W_\infty^2\right)\right]. \tag{17}$$

Proof. By definition, the target quantity can be decoupled using the triangle inequality: $\|a-b\|^2\leq 2\|a-c\|^2+2\|c-b\|^2$. By introducing the cross-term $\mathbf{m}_i\odot\nabla f_j(\mathbf{m}_i\odot\mathbf{w})$, we have:

$$\begin{aligned}&\mathbb{E}_{\mathbf{m}_i,\mathbf{m}_j}\left\|\mathbf{m}_i\odot\nabla f_i(\mathbf{m}_i\odot\mathbf{w})-\mathbf{m}_j\odot\nabla f_j(\mathbf{m}_j\odot\mathbf{w})\right\|^2\\&\leq 2\underbrace{\mathbb{E}_{\mathbf{m}_i}\left\|\mathbf{m}_i\odot\left(\nabla f_i(\mathbf{m}_i\odot\mathbf{w})-\nabla f_j(\mathbf{m}_i\odot\mathbf{w})\right)\right\|^2}_{\text{Term A: Subspace Statistical Heterogeneity}}\\&\quad+2\underbrace{\mathbb{E}_{\mathbf{m}_i,\mathbf{m}_j}\left\|\mathbf{m}_i\odot\nabla f_j(\mathbf{m}_i\odot\mathbf{w})-\mathbf{m}_j\odot\nabla f_j(\mathbf{m}_j\odot\mathbf{w})\right\|^2}_{\text{Term B: System Heterogeneity Variance}}\end{aligned} \tag{18}$$

Bounding Term A (Statistical Heterogeneity):
To obtain a tight bound for highly sparse subnetworks, we evaluate the squared norm coordinate-wise rather than trivially relaxing the mask projection. Let $\mathbf{m}_{i,k}\in\{0,1\}$ denote the k-th coordinate of the mask $\mathbf{m}_i$. Term A can be written as:

$$\begin{aligned}\text{Term A}&=\mathbb{E}_{\mathbf{m}_i}\left[\sum_{k=1}^{d}\mathbf{m}_{i,k}^2\left(\nabla_k f_i(\mathbf{m}_i\odot\mathbf{w})-\nabla_k f_j(\mathbf{m}_i\odot\mathbf{w})\right)^2\right]\\&=\mathbb{E}_{\mathbf{m}_i}\left[\sum_{k=1}^{d}\mathbf{m}_{i,k}\left(\int_\xi\nabla_k\ell(\mathbf{m}_i\odot\mathbf{w};\xi)d\mathcal{D}_i(\xi)-\int_\xi\nabla_k\ell(\mathbf{m}_i\odot\mathbf{w};\xi)d\mathcal{D}_j(\xi)\right)^2\right],\end{aligned} \tag{19}$$

where we used the fact that $\mathbf{m}_{i,k}^2=\mathbf{m}_{i,k}$ for binary variables. Under the coordinate-wise l_ℓ-Lipschitz assumption, we apply the Kantorovich-Rubinstein duality to bound the discrepancy by the 1-Wasserstein distance:

$$\left(\int_\xi\nabla_k\ell(x;\xi)d\mathcal{D}_i(\xi)-\int_\xi\nabla_k\ell(x;\xi)d\mathcal{D}_j(\xi)\right)^2\leq l_\ell^2W_1(\mathcal{D}_i,\mathcal{D}_j)^2. \tag{20}$$

Substituting this upper bound back into (19), the dependency on the non-linear gradient function is fully decoupled from the mask. Since $\mathbb{E}[\mathbf{m}_{i,k}] = p_i$, we obtain:

$$\text{Term A} \leq \sum_{k=1}^{d} \mathbb{E}[\mathbf{m}_{i,k}] \cdot l_\ell^2 W_1(\mathcal{D}_i, \mathcal{D}_j)^2 = p_i d \cdot l_\ell^2 W_1(\mathcal{D}_i, \mathcal{D}_j)^2. \tag{21}$$

Bounding Term B (System Heterogeneity):
To tightly bound the system heterogeneity, we further decouple the effect of the mask on the gradient output and input by introducing the cross-term $\mathbf{m}_i \odot \nabla f_j(\mathbf{m}_j \odot \mathbf{w})$:

$$\begin{aligned}\text{Term B} &\leq 2\mathbb{E}_{\mathbf{m}_i,\mathbf{m}_j} \left\| \mathbf{m}_i \odot \left(\nabla f_j(\mathbf{m}_i \odot \mathbf{w}) - \nabla f_j(\mathbf{m}_j \odot \mathbf{w}) \right) \right\|^2 \\ &\quad + 2\mathbb{E}_{\mathbf{m}_i,\mathbf{m}_j} \left\| (\mathbf{m}_i - \mathbf{m}_j) \odot \nabla f_j(\mathbf{m}_j \odot \mathbf{w}) \right\|^2 \\ &\leq 2\mathbb{E}_{\mathbf{m}_i,\mathbf{m}_j} \left\| \nabla f_j(\mathbf{m}_i \odot \mathbf{w}) - \nabla f_j(\mathbf{m}_j \odot \mathbf{w}) \right\|^2 + 2\mathbb{E}_{\mathbf{m}_i,\mathbf{m}_j} \left\| (\mathbf{m}_i - \mathbf{m}_j) \odot \nabla f_j(\mathbf{m}_j \odot \mathbf{w}) \right\|^2. \end{aligned} \tag{22}$$

Under Assumptions 2 and 3, the gradients and optimization weights are bounded coordinate-wise by G_∞ and W_∞, and f_j is L-smooth. For any fixed realization of $\mathbf{m}_i$ and $\mathbf{m}_j$, their distance can be measured by the Hamming distance $\|\mathbf{m}_i - \mathbf{m}_j\|_0$. The two sub-terms can be bounded as follows:

1. $\|\nabla f_j(\mathbf{m}_i \odot \mathbf{w}) - \nabla f_j(\mathbf{m}_j \odot \mathbf{w})\|^2 \leq L^2 \|(\mathbf{m}_i - \mathbf{m}_j) \odot \mathbf{w}\|^2 \leq L^2 \|\mathbf{m}_i - \mathbf{m}_j\|_0 W_\infty^2$.
2. $\|(\mathbf{m}_i - \mathbf{m}_j) \odot \nabla f_j(\mathbf{m}_j \odot \mathbf{w})\|^2 \leq \|\mathbf{m}_i - \mathbf{m}_j\|_0 G_\infty^2$.

Therefore, Term B $\leq 2(L^2 W_\infty^2 + G_\infty^2)\mathbb{E}_{\mathbf{m}_i,\mathbf{m}_j}[\|\mathbf{m}_i - \mathbf{m}_j\|_0]$.

Since $\mathbf{m}_i \sim \text{Ber}(p_i)$ and $\mathbf{m}_j \sim \text{Ber}(p_j)$ are independent, the probability that their k-th coordinates differ is given by $\mathbb{P}(\mathbf{m}_{i,k} \neq \mathbf{m}_{j,k}) = p_i(1-p_j) + p_j(1-p_i) = p_i + p_j - 2p_i p_j$. Thus, the expected Hamming distance is exactly:

$$\mathbb{E}_{\mathbf{m}_i,\mathbf{m}_j}[\|\mathbf{m}_i - \mathbf{m}_j\|_0] = d(p_i + p_j - 2p_i p_j). \tag{23}$$

Substituting this expectation yields the final bound for Term B:

$$\text{Term B} \leq 2d(p_i + p_j - 2p_i p_j)\left(G_\infty^2 + L^2 W_\infty^2\right). \tag{24}$$

Final Synthesis:
Substituting the bounds for Term A (21) and Term B (24) back into Equation (18), and taking the supremum over all weights $\mathbf{w}$ and client pairs, we obtain:

$$\zeta_{p,\max}^2 \leq \max_{i,j} \left[2p_i d \cdot l_\ell^2 W_1(\mathcal{D}_i, \mathcal{D}_j)^2 + 4d(p_i + p_j - 2p_i p_j)\left(G_\infty^2 + L^2 W_\infty^2\right) \right]. \tag{25}$$

This completes the proof. □

Lemma 13. *For Algorithm 2, the following statement holds true:*

$$\zeta_{\mathbf{m},\max}^2 \leq d_m \cdot l_\ell^2 \max_{i,i'} W_1(D_i, D_{i'})^2 + 4D_{\max}\left(G_\infty^2 + L^2 W_\infty^2\right).$$

Proof. **Definitions.** Let $W_{\max} := \max_{i,i'} W_1(D_i, D_{i'})$ denote the maximum 1-Wasserstein distance between any two clients' data distributions. Let $D_{\max} := \max_{i,j,i',j'} \|\mathbf{m}_i^j - \mathbf{m}_{i'}^{j'}\|_0$ be the maximum Hamming distance between any two predefined masks.

Proof of the Upper Bound for $\zeta_{\mathbf{m},\max}^2$:

To upper bound the worst-case divergence $\zeta_{\mathbf{m},\max}^2$, we decouple the error into statistical heterogeneity and system (mask) heterogeneity. For simplicity in notation, let $\mathbf{m}_1 := \mathbf{m}_i^j$ and $\mathbf{m}_2 := \mathbf{m}_{i'}^{j'}$.

By applying the triangle inequality $\|a-b\|^2 \leq 2\|a-c\|^2 + 2\|c-b\|^2$, with the cross-term $\mathbf{m}_1 \odot \nabla f_{i'}(\mathbf{m}_1 \odot \mathbf{w})$, we obtain:

$$\begin{aligned}
&\|\mathbf{m}_1 \odot \nabla f_i(\mathbf{m}_1 \odot \mathbf{w}) - \mathbf{m}_2 \odot \nabla f_{i'}(\mathbf{m}_2 \odot \mathbf{w})\|^2 \\
&\leq 2 \underbrace{\left\|\mathbf{m}_1 \odot \left(\nabla f_i(\mathbf{m}_1 \odot \mathbf{w}) - \nabla f_{i'}(\mathbf{m}_1 \odot \mathbf{w})\right)\right\|^2}_{\text{Term A: Statistical Heterogeneity}} \\
&\quad + 2 \underbrace{\|\mathbf{m}_1 \odot \nabla f_{i'}(\mathbf{m}_1 \odot \mathbf{w}) - \mathbf{m}_2 \odot \nabla f_{i'}(\mathbf{m}_2 \odot \mathbf{w})\|^2}_{\text{Term B: System Heterogeneity}}
\end{aligned} \tag{26}$$

Bounding Term A (Statistical Heterogeneity):

Instead of trivially relaxing the mask projection (i.e., using $\|\mathbf{m} \odot \mathbf{v}\|^2 \leq \|\mathbf{v}\|^2$), which yields an overly loose bound for highly sparse subnetworks, we treat the mask $\mathbf{m}_1$ as a projection operator onto a lower-dimensional subspace.

Let $\text{supp}(\mathbf{m}_1)$ denote the support set of the mask, and let $d_m = \|\mathbf{m}_1\|_0$ be the exact number of active parameters in this subnetwork. Term A can be written exactly as the squared norm restricted to this subspace:

$$\begin{aligned}
\text{Term A} &= \left\|\mathbf{m}_1 \odot \left(\nabla f_i(\mathbf{m}_1 \odot \mathbf{w}) - \nabla f_{i'}(\mathbf{m}_1 \odot \mathbf{w})\right)\right\|^2 \\
&= \sum_{k \in \text{supp}(\mathbf{m}_1)} \left(\int_\xi \nabla_k \ell(\mathbf{m}_1 \odot \mathbf{w}; \xi) dD_i(\xi) - \int_\xi \nabla_k \ell(\mathbf{m}_1 \odot \mathbf{w}; \xi) dD_{i'}(\xi) \right)^2 .
\end{aligned} \tag{27}$$

where $\nabla_k \ell$ is the partial derivative with respect to the k-th coordinate.

To bound this tightly, we assume the gradient discrepancy caused by data heterogeneity is uniformly bounded across coordinates. Let the partial derivative $\nabla_k \ell(\mathbf{x}; \xi)$ be l_ξ-Lipschitz continuous with respect to the data input ξ for any coordinate k. Applying the Kantorovich-Rubinstein duality coordinate-wise yields:

$$\left| \int_\xi \nabla_k \ell(\mathbf{x}; \xi) dD_i(\xi) - \int_\xi \nabla_k \ell(\mathbf{x}; \xi) dD_{i'}(\xi) \right| \leq l_\ell W_1(D_i, D_{i'}). \tag{28}$$

Substituting this into (27), we sum the bounds strictly over the d_m active coordinates rather than the entire dimension d:

$$\begin{aligned}
\text{Term A} &\leq \sum_{k \in \text{supp}(\mathbf{m}_1)} \left(l_\ell W_1(D_i, D_{i'})\right)^2 \\
&= d_m \cdot l_\ell^2 W_1(D_i, D_{i'})^2 \\
&\leq d_m \cdot l_\ell^2 W_{\max}^2.
\end{aligned} \tag{29}$$

Bounding Term B (System Heterogeneity):
We further decouple the effect of the mask on the gradient output and the gradient input by introducing the cross-term $\mathbf{m}_2 \odot \nabla f_{i'}(\mathbf{m}_1 \odot \mathbf{w})$:

$$\begin{aligned}
\text{Term B} &\leq 2\left\|(\mathbf{m}_1 - \mathbf{m}_2) \odot \nabla f_{i'}(\mathbf{m}_1 \odot \mathbf{w})\right\|^2 + 2\left\|\mathbf{m}_2 \odot \left(\nabla f_{i'}(\mathbf{m}_1 \odot \mathbf{w}) - \nabla f_{i'}(\mathbf{m}_2 \odot \mathbf{w})\right)\right\|^2 \\
&\leq 2\left\|(\mathbf{m}_1 - \mathbf{m}_2) \odot \nabla f_{i'}(\mathbf{m}_1 \odot \mathbf{w})\right\|^2 + 2\left\|\nabla f_{i'}(\mathbf{m}_1 \odot \mathbf{w}) - \nabla f_{i'}(\mathbf{m}_2 \odot \mathbf{w})\right\|^2 .
\end{aligned} \tag{30}$$

Under Assumptions 2 and 3, we bound the two sub-terms using the Hamming distance $\|\mathbf{m}_1 - \mathbf{m}_2\|_0$:

1. The first sub-term is bounded by the non-zero elements of the mask difference:

$$\|(\mathbf{m}_1 - \mathbf{m}_2) \odot \nabla f_{i'}\|^2 \leq \|\mathbf{m}_1 - \mathbf{m}_2\|_0 \cdot \|\nabla f_{i'}\|_\infty^2 \leq D_{\max} G_\infty^2.$$

2. The second sub-term is bounded via L-smoothness and the bounded weight space:

$$\begin{aligned}
\|\nabla f_{i'}(\mathbf{m}_1 \odot \mathbf{w}) - \nabla f_{i'}(\mathbf{m}_2 \odot \mathbf{w})\|^2 &\leq L^2 \|(\mathbf{m}_1 - \mathbf{m}_2) \odot \mathbf{w}\|^2 \\
&\leq L^2 \|\mathbf{m}_1 - \mathbf{m}_2\|_0 \|\mathbf{w}\|_\infty^2 \leq D_{\max} L^2 W_\infty^2.
\end{aligned}$$

Substituting these back yields the bound for Term B:

$$\text{Term B} \leq 2D_{\max}\left(G_\infty^2 + L^2 W_\infty^2\right). \tag{31}$$

Final Synthesized Bound:
By substituting the statistical bound (29) and the system heterogeneity bound (31) back into the original decoupled inequality (26), and taking the supremum over all weights $\mathbf{w} \in \mathcal{W}$ and all client/mask combinations, we arrive at the final upper bound:

$$\begin{aligned}\zeta_{\mathbf{m},\max}^2 &= \max_{\mathbf{w}\in\mathcal{W}} \max_{i,j,i',j'} \left\| \mathbf{m}_i^j \odot \nabla f_i(\mathbf{m}_i^j \odot \mathbf{w}) - \mathbf{m}_{i'}^{j'} \odot \nabla f_{i'}(\mathbf{m}_{i'}^{j'} \odot \mathbf{w}) \right\|^2 \\ &\leq d_m \cdot l_\ell^2 W_{\max}^2 + 4D_{\max}\left(G_\infty^2 + L^2 W_\infty^2\right). \end{aligned} \tag{32}$$

This completes the proof. □

Lemma 14. *For Algorithm 1, under the condition of Theorem 1, the following statement holds true:*

$$\mathbb{E}\left\|\tilde{\mathbf{w}}_{r,k}^i - \mathbf{w}_r\right\|^2 \leq 5K\left(4\eta^2 K\mathbb{E}\left\|\nabla F_{\mathbf{p}}(\mathbf{w}_r)\right\|^2 + 4\eta^2 K\zeta_{\mathbf{p},\max}^2 + \eta^2 K\delta^2\right)$$

Proof. According to local updating rule we have:

$$\begin{aligned}
&\mathbb{E}\left\|\tilde{\mathbf{w}}_{r,k}^i - \mathbf{w}_r\right\|^2 \\
&= (1+\frac{1}{K-1})\mathbb{E}\left\|\tilde{\mathbf{w}}_{r,k-1}^i - \mathbf{w}_r\right\|^2 + K\mathbb{E}\left\|\eta\tilde{\mathbf{g}}_{r,k-1}^i\right\|^2 \\
&\leq (1+\frac{1}{K-1})\mathbb{E}\left\|\tilde{\mathbf{w}}_{r,k-1}^i - \mathbf{w}_r\right\|^2 + K\mathbb{E}\left\|\eta\mathbf{m}_i^r \odot \nabla f_i(\mathbf{m}_i^r \odot \tilde{\mathbf{w}}_{r,k-1}^i)\right\|^2 + \eta^2 K\delta^2 \\
&\leq (1+\frac{1}{K-1})\mathbb{E}\left\|\tilde{\mathbf{w}}_{r,k-1}^i - \mathbf{w}_r\right\|^2 + 2K\mathbb{E}\left\|\eta\mathbf{m}_i^r \odot \nabla f_i(\mathbf{m}_i^r \odot \mathbf{w}_r)\right\|^2 \\
&\quad + 2\eta^2\tilde{L}^2 K\left\|\tilde{\mathbf{w}}_{r,k-1}^i - \mathbf{w}_r\right\|^2 + \eta^2 K\delta^2 \\
&\leq (1+\frac{2}{K-1})\mathbb{E}\left\|\tilde{\mathbf{w}}_{r,k-1}^i - \mathbf{w}_r\right\|^2 + 4\eta^2 K\mathbb{E}\left\|\nabla F_{\mathbf{p}}(\mathbf{w}_r)\right\|^2 + 4\eta^2 K\zeta_{\mathbf{p},\max}^2 + \eta^2 K\delta^2 \\
&\leq \sum_{j=1}^{k}(1+\frac{2}{K-1})^{k-j}\left(4\eta^2 K\mathbb{E}\left\|\nabla F_{\mathbf{p}}(\mathbf{w}_r)\right\|^2 + 4\eta^2 K\zeta_{\mathbf{p},\max}^2 + \eta^2 K\delta^2\right) \\
&\leq 5K\left(4\eta^2 K\mathbb{E}\left\|\nabla F_{\mathbf{p}}(\mathbf{w}_r)\right\|^2 + 4\eta^2 K\zeta_{\mathbf{p},\max}^2 + \eta^2 K\delta^2\right)
\end{aligned}$$

where the fourth step is due to $2\eta^2\tilde{L}^2 K \leq \frac{1}{K-1}$, which conclude the proof. □

E.1 Proof of Theorem 5

Recall the output of Algorithm 1:

$$\begin{aligned}
&\mathbb{E}\left\|\hat{\mathbf{w}} - \hat{\mathbf{w}}'\right\| \\
&= \mathbb{E}\left\| \begin{array}{l} \mathcal{P}_{\mathcal{W}}\left(\mathbf{w}_R - (1/L)\frac{1}{N}\sum_{i=1}^{N}\mathbf{m}_i \odot \nabla f_i(\mathbf{m}_i \odot \mathbf{w}_R)\right) \\ \quad - \mathcal{P}_{\mathcal{W}}\left(\mathbf{w}'_R - (1/L)\frac{1}{N}\sum_{i=1}^{N}\mathbf{m}_i \odot \nabla f_i(\mathbf{m}_i \odot \mathbf{w}'_R)\right) \end{array} \right\| \\
&\leq \mathbb{E}\left\| \begin{array}{l} \mathbf{w}_R - \mathbf{w}'_R - (1/L)\frac{1}{N}\sum_{i=1}^{N}\mathbf{m}_i \odot \nabla f_i(\mathbf{m}_i \odot \mathbf{w}_R) \\ + (1/L)\frac{1}{N}\sum_{i=1}^{N}\mathbf{m}_i \odot \nabla f_i(\mathbf{m}_i \odot \mathbf{w}'_R) \end{array} \right\| \\
&\leq 2\mathbb{E}\left\|\mathbf{w}_R - \mathbf{w}'_R\right\|.
\end{aligned}$$

Hence it suffices to bound $\mathbb{E}\left\|\mathbf{w}_R - \mathbf{w'}_R\right\|$.

For the client i with perturbed data, by the updating rule we know with probability $1 - \frac{1}{n}$:

$$\begin{aligned}
&\mathbb{E}\left\|\mathbf{w}_{r,k+1}^i - \mathbf{w'}_{r,k+1}^i\right\| \\
&\leq \mathbb{E}\left\|\mathbf{w}_{r,k}^i - \mathbf{w'}_{r,k}^i - \eta\left(\mathbf{m}_i^r \odot \nabla f_i(\mathbf{m}_i^r \odot \mathbf{w}_{r,k}^i; \xi_{r,k}^i) - \mathbf{m}_i^r \odot \nabla f_i(\mathbf{m}_i^r \odot \mathbf{w'}_{r,k}^i; \xi_{r,k}^i)\right)\right\| \\
&\leq \mathbb{E}\left\|\mathbf{w}_{r,k} - \mathbf{w'}_{r,k}\right\|.
\end{aligned}$$

With probability $\frac{1}{n}$, we have

$$\begin{aligned}
&\mathbb{E}\left\|\mathbf{w}_{r,k+1}^i - \mathbf{w'}_{r,k+1}^i\right\| \\
&\leq \mathbb{E}\left\|\mathbf{w}_{r,k}^i - \mathbf{w'}_{r,k}^i - \eta\left(\mathbf{m}_i^r \odot \nabla f_i(\mathbf{m}_i^r \odot \mathbf{w}_{r,k}^i; \xi_{r,k}^i) - \mathbf{m}_i^r \odot \nabla f_i(\mathbf{m}_i^r \odot \mathbf{w'}_{r,k}^i; \xi'^i_{r,k})\right)\right\| \\
&\leq \mathbb{E}\left\|\mathbf{w}_{r,k}^i - \mathbf{w'}_{r,k}^i - \eta\left(\mathbf{m}_i^r \odot \nabla f_i(\mathbf{m}_i^r \odot \mathbf{w}_{r,k}^i; \xi_{r,k}^i) - \mathbf{m}_i^r \odot \nabla f_i(\mathbf{m}_i^r \odot \mathbf{w'}_{r,k}^i; \xi_{r,k}^i)\right)\right\| \\
&\quad + \left\|\eta\left(\mathbf{m}_i^r \odot \nabla f_i(\mathbf{m}_i^r \odot \mathbf{w'}_{r,k}^i; \xi_{r,k}^i) - \mathbf{m}_i^r \odot \nabla f_i(\mathbf{m}_i^r \odot \mathbf{w'}_{r,k}^i; \xi'^i_{r,k})\right)\right\| \\
&\leq \mathbb{E}\left\|\mathbf{w}_{r,k}^i - \mathbf{w'}_{r,k}^i\right\| + \eta\mathbb{E}\left\|\mathbf{m}_i^r \odot \nabla f_i(\mathbf{m}_i^r \odot \mathbf{w'}_{r,k}^i; \xi_{r,k}^i)\right\| \\
&\quad + \eta\mathbb{E}\left\|\mathbf{m}_i^r \odot \nabla f_i(\mathbf{m}_i^r \odot \mathbf{w'}_{r,k}^i; \xi'^i_{r,k})\right\| \\
&\leq \mathbb{E}\left\|\mathbf{w}_{r,k}^i - \mathbf{w'}_{r,k}^i\right\| + \eta\mathbb{E}\left\|\nabla F_{\mathbf{p}}(\mathbf{w}_r)\right\| + \eta\mathbb{E}\left\|\nabla F_{\mathbf{p}}(\mathbf{w'}_r)\right\| \\
&\quad + 2\eta\left\|\mathbf{w'}_r - \mathbf{w'}_{r,k}^i\right\| + 2\eta\delta + 2\eta\zeta_{\mathbf{p},\max}
\end{aligned}$$

For $j \neq i$, we have $\mathbb{E}\left\|\mathbf{w}_{r+1}^j - \mathbf{w'}_{r+1}^j\right\| \leq \mathbb{E}\left\|\mathbf{w}_{r,K-1}^j - \mathbf{w'}_{r,K-1}^j\right\| \leq \mathbb{E}\left\|\mathbf{w}_r^j - \mathbf{w'}_r^j\right\|$. Combining two cases we have

$$\begin{aligned}
\frac{1}{N}\sum_{j=1}^N \mathbb{E}\left\|\mathbf{w}_{r,k+1}^j - \mathbf{w'}_{r,k+1}^j\right\| \leq \frac{1}{N}\sum_{j=1}^N \mathbb{E}\left\|\mathbf{w}_{r,k}^j - \mathbf{w'}_{r,k}^j\right\| + \frac{2\eta\delta}{Nn} + \frac{2\eta\zeta_{\mathbf{p},\max}}{Nn} + \frac{1}{Nn}2\eta\left\|\mathbf{w'}_r - \mathbf{w'}_{r,k}^i\right\| \\
+ \frac{1}{Nn}\eta\mathbb{E}\left\|\nabla F_{\mathbf{p}}(\mathbf{w}_r)\right\| + \frac{1}{Nn}\eta\mathbb{E}\left\|\nabla F_{\mathbf{p}}(\mathbf{w'}_r)\right\|.
\end{aligned}$$

Performing telescoping sum from $k = K-1$ to 0 yields:

$$\begin{aligned}
\mathbb{E}\left\|\mathbf{w}_{r+1} - \mathbf{w'}_{r+1}\right\| \leq \mathbb{E}\left\|\mathbf{w}_r - \mathbf{w'}_r\right\| + \frac{2\eta K\delta}{Nn} + \frac{2\eta K\zeta_{\mathbf{p},\max}}{Nn} + \frac{2\eta}{Nn}\sum_{k=1}^K \mathbb{E}\left\|\mathbf{w'}_r - \mathbf{w'}_{r,k}^i\right\| \\
+ \frac{1}{Nn}\eta K\mathbb{E}\left\|\nabla F_{\mathbf{p}}(\mathbf{w}_r)\right\| + \frac{1}{Nn}\eta K\mathbb{E}\left\|\nabla F_{\mathbf{p}}(\mathbf{w'}_r)\right\|.
\end{aligned}$$

Performing telescoping sum from $r = R-1$ to 0, and using the fact $\mathbf{w}_0 = \mathbf{w'}_0$ yields:

$$\begin{aligned}
\mathbb{E}\left\|\mathbf{w}_R - \mathbf{w'}_R\right\| &\leq \frac{2\eta RK\delta}{Nn} + \frac{2\eta RK\zeta_{\mathbf{p},\max}}{Nn} + \frac{2\eta}{Nn}\sum_{r=1}^R\sum_{k=1}^K \mathbb{E}\left\|\mathbf{w'}_r - \mathbf{w'}_{r,k}^i\right\| \\
&\quad + \frac{1}{Nn}\eta K\sum_{r=1}^R \mathbb{E}\left\|\nabla F_{\mathbf{p}}(\mathbf{w}_r)\right\| + \frac{1}{Nn}\eta K\sum_{r=1}^R \mathbb{E}\left\|\nabla F_{\mathbf{p}}(\mathbf{w'}_r)\right\| \\
&\leq \frac{2\eta RK\delta}{Nn} + \frac{2\eta RK\zeta_{\mathbf{p},\max}}{Nn} + \frac{2\eta}{Nn}\sum_{r=1}^R\sum_{k=1}^K \mathbb{E}\left\|\mathbf{w'}_r - \mathbf{w'}_{r,k}^i\right\| \\
&\quad + \frac{1}{Nn}\eta RK\sqrt{\frac{1}{R}\sum_{r=1}^R \mathbb{E}\left\|\nabla F_{\mathbf{p}}(\mathbf{w}_r)\right\|^2} + \frac{1}{Nn}\eta RK\sqrt{\frac{1}{R}\sum_{r=1}^R \mathbb{E}\left\|\nabla F_{\mathbf{p}}(\mathbf{w'}_r)\right\|^2}.
\end{aligned}$$

To bound $\mathbb{E}\left\|\mathbf{w}_{r,k}^i - \mathbf{w}_{r,k}\right\|$, evoking Lemma 14 gives:

$$\mathbb{E}\left\|\mathbf{w}_{r,k}^i - \mathbf{w}_{r,k}\right\| \leq \sqrt{\mathbb{E}\left\|\mathbf{w}_{r,k}^i - \mathbf{w}_{r,k}\right\|^2} \leq \sqrt{5K\left(4\eta^2 K\mathbb{E}\left\|\nabla F_{\mathbf{p}}(\mathbf{w}_r)\right\|^2 + 4\eta^2 K\zeta_{\mathbf{p},\max}^2 + \eta^2 K\delta^2\right)}.$$

Hence we have

$$\begin{aligned}\mathbb{E}\left\|\mathbf{w}_R - \mathbf{w'}_R\right\| \leq{}& \frac{2\eta RK\delta}{Nn} + \frac{2\eta RK\zeta_{\mathbf{p},\max}}{Nn} \qquad (33)\\ &+ \frac{2\eta}{Nn}\sum_{r=1}^{R}\sum_{k=1}^{K}\sqrt{5K\left(4\eta^2 K\mathbb{E}\left\|\nabla F_{\mathbf{p}}(\mathbf{w}_r)\right\|^2 + 4\eta^2 K\zeta_{\mathbf{p},\max}^2 + \eta^2 K\delta^2\right)}\\ &+ \frac{1}{Nn}\eta RK\sqrt{\frac{1}{R}\sum_{r=1}^{R}\mathbb{E}\left\|\nabla F_{\mathbf{p}}(\mathbf{w}_r)\right\|^2} + \frac{1}{Nn}\eta RK\sqrt{\frac{1}{R}\sum_{r=1}^{R}\mathbb{E}\left\|\nabla F_{\mathbf{p}}(\mathbf{w'}_r)\right\|^2}. \qquad (34)\end{aligned}$$

Now we will bound the gradient norm $\mathbb{E}\left\|\nabla F_{\mathbf{p}}(\mathbf{w}_r)\right\|^2$. Due to Eq.(7) we have

$$\begin{aligned}\mathbb{E}\left\|\mathbf{w}_{r+1} - \mathbf{w}^*\right\|^2 \leq{}& \mathbb{E}\left\|\mathbf{w}_r - \mathbf{w}^*\right\|^2 - \frac{1}{8}\eta K\left(F_{\mathbf{p}}(\mathbf{w}_r) - F_{\mathbf{p}}(\mathbf{w}^*)\right)\\ &+ 2\eta\tilde{L}\cdot 5K^2\left(4\eta^2 K\sigma_*^2 + \eta^2 K\delta^2\right) + \frac{K\eta^2\delta^2}{N}.\end{aligned}$$

Due to the fact $\left\|\nabla F_{\mathbf{p}}(\mathbf{w}_r)\right\|^2 \leq 2L\left(F_{\mathbf{p}}(\mathbf{w}_r) - F_{\mathbf{p}}(\mathbf{w}^*)\right)$ we have

$$\begin{aligned}\mathbb{E}\left\|\mathbf{w}_{r+1} - \mathbf{w}^*\right\|^2 \leq{}& \mathbb{E}\left\|\mathbf{w}_r - \mathbf{w}^*\right\|^2 - \frac{1}{16L}\eta K\mathbb{E}\left\|\nabla F_{\mathbf{p}}(\mathbf{w}_r)\right\|^2\\ &+ 2\eta\tilde{L}\cdot 5K^2\left(4\eta^2 K\sigma_*^2 + \eta^2 K\delta^2\right) + \frac{K\eta^2\delta^2}{N}.\end{aligned}$$

Summing over $r = 1$ to R and dividing both sides by R yields:

$$\begin{aligned}&\frac{1}{16R\tilde{L}}\eta K\sum_{r=1}^{R}\mathbb{E}\left\|\nabla F_{\mathbf{p}}(\mathbf{w}_r)\right\|^2 \leq \frac{\mathbb{E}\left\|\mathbf{w}_1 - \mathbf{w}^*\right\|^2}{R} + 2\eta\tilde{L}\cdot 5K^2\left(4\eta^2 K\sigma_*^2 + \eta^2 K\delta^2\right) + \frac{K\eta^2\delta^2}{N}\\ \iff &\frac{1}{R}\sum_{r=1}^{R}\mathbb{E}\left\|\nabla F_{\mathbf{p}}(\mathbf{w}_r)\right\|^2 \leq \frac{16\tilde{L}\mathbb{E}\left\|\mathbf{w}_1 - \mathbf{w}^*\right\|^2}{\eta KR} + 32\tilde{L}^2\cdot 5K\left(4\eta^2 K\sigma_*^2 + \eta^2 K\delta^2\right) + \frac{32L\eta\delta^2}{N}.\end{aligned}$$

Plugging above bound back to (34) yields:

$$\begin{aligned}&\mathbb{E}\left\|\mathbf{w}_R - \mathbf{w'}_R\right\|\\ &\leq \frac{2\eta RK\delta}{Nn} + \frac{2\eta RK\zeta_{\max}}{Nn} + \frac{2\eta}{Nn}\sum_{r=1}^{R}K\sqrt{5K\left(4\eta^2 K\mathbb{E}\left\|\nabla F_{\mathbf{p}}(\mathbf{w}_r)\right\|^2 + 4\eta^2 K\zeta_{\mathbf{p},\max}^2 + \eta^2 K\delta^2\right)}\\ &\quad + 2\frac{1}{Nn}\eta RK\sqrt{\frac{16\tilde{L}\mathbb{E}\left\|\mathbf{w}_1 - \mathbf{w}^*\right\|^2}{\eta KR} + 32\tilde{L}^2\cdot 5K\left(4\eta^2 K\sigma_*^2 + \eta^2 K\delta^2\right) + \frac{32L\eta\delta^2}{N}}\\ &\leq O\left(\begin{array}{l}\dfrac{\eta RK(\delta + \zeta_{\mathbf{p},\max})}{Nn}\\ \quad + \dfrac{RK\eta}{Nn}\sqrt{\eta^2 K^2\left(\dfrac{\tilde{L}}{\eta KR} + \tilde{L}^2 K\left(\eta^2 K\sigma_*^2 + \eta^2 K\delta^2\right) + \dfrac{L\eta\delta^2}{N}\right) + \eta^2 K^2(\zeta_{\mathbf{p},\max}^2 + \delta^2)}\end{array}\right)\\ &\quad + \frac{\eta RK}{Nn}O\left(\sqrt{\frac{\tilde{L}}{\eta KR} + \tilde{L}^2 K\left(\eta^2 K\sigma_*^2 + \eta^2 K\delta^2\right) + \frac{L\eta\delta^2}{N}}\right).\end{aligned}$$

Choosing $\eta = \frac{\sqrt{Nn}}{RK}$ and $\frac{R}{\sqrt{Nn}} \geq \tilde{L}$ will conclude the proof:

$$\begin{aligned}\mathbb{E}\left\|\mathbf{w}_R - \mathbf{w'}_R\right\| &\leq O\left(\frac{\delta + \zeta_{\mathbf{p},\max}}{\sqrt{Nn}} + \frac{1}{\sqrt{Nn}}\left(\sqrt{\frac{\tilde{L}}{\sqrt{Nn}} + \tilde{L}^2\frac{Nn}{R^2}\sigma_*^2 + \tilde{L}^2\frac{Nn}{R^2}\delta^2 + \frac{\sqrt{Nn}}{RK}\frac{L\delta^2}{N}}\right)\right) \\ &\leq O\left(\frac{\delta + \zeta_{\mathbf{p},\max}}{\sqrt{Nn}} + \frac{1}{\sqrt{Nn}}\sqrt{\frac{\tilde{L}}{\sqrt{Nn}} + \sigma_*^2 + \delta^2 + \frac{\delta^2}{KN}}\right).\end{aligned}$$

Finally, applying Lemma 12 will conclude the proof.

E.2 Proof of Theorem 6

Recall the output of Algorithm 2:

$$\begin{aligned}\mathbb{E}\left\|\hat{\mathbf{w}} - \hat{\mathbf{w}}'\right\| = \mathbb{E}\Bigg\|&\mathcal{P}_{\mathcal{W}}\left(\mathbf{w}_T - \frac{1}{L}\frac{1}{N}\sum_{i=1}^{N}\frac{1}{R}\sum_{r=1}^{R}\mathbf{m}_i^{\sigma_e(r)} \odot \nabla f_i(\mathbf{m}_i^{\sigma_e(r)} \odot \mathbf{w}_T)\right) \\ &-\mathcal{P}_{\mathcal{W}}\left(\mathbf{w}'_T - \frac{1}{L}\frac{1}{N}\sum_{i=1}^{N}\frac{1}{R}\sum_{r=1}^{R}\mathbf{m}_i^{\sigma_e(r)} \odot \nabla f_i(\mathbf{m}_i^{\sigma_e(r)} \odot \mathbf{w}'_T)\right)\Bigg\| \\ &\leq 2\mathbb{E}\left\|\mathbf{w}_T - \mathbf{w}'_T\right\|.\end{aligned}$$

Hence it suffices to bound $\mathbb{E}\left\|\mathbf{w}_T - \mathbf{w}'_T\right\|$.

For the client i with perturbed data, by the updating rule and convexity of f_i, we know with probability $1 - \frac{1}{n}$:

$$\begin{aligned}&\mathbb{E}\left\|\mathbf{w}^i_{e,r,k+1} - \mathbf{w'}^i_{e,r,k+1}\right\| \\ &\leq \mathbb{E}\left\|\begin{array}{l}\mathbf{w}^i_{e,r,k} - \mathbf{w'}^i_{e,r,k} \\ \quad - \eta\mathbf{m}_i^{\sigma_e(r)} \odot \nabla f_i(\mathbf{m}_i^{\sigma_e(r)} \odot \mathbf{w}^i_{e,r,k}; \xi^i_{e,r,k}) \\ \quad\quad - \eta\mathbf{m}_i^{\sigma_e(r)} \odot \nabla f_i(\mathbf{m}_i^{\sigma_e(r)} \odot \mathbf{w'}^i_{e,r,k}; \xi'^i_{e,r,k})\end{array}\right\| \\ &\leq \mathbb{E}\left\|\mathbf{w}_{e,r,k} - \mathbf{w}'_{e,r,k}\right\|.\end{aligned}$$

With probability $\frac{1}{n}$, we have

$$
\begin{aligned}
&\mathbb{E}\left\|\mathbf{w}^i_{e,r,k+1} - \mathbf{w'}^i_{e,r,k+1}\right\| \\
&\leq \mathbb{E}\left\|\begin{aligned}&\mathbf{w}^i_{e,r,k} - \mathbf{w'}^i_{e,r,k} \\ &\quad - \eta \mathbf{m}_i^{\sigma_e(r)} \odot \nabla f_i(\mathbf{m}_i^{\sigma_e(r)} \odot \mathbf{w}^i_{e,r,k}; \xi^i_{e,r,k}) \\ &\qquad - \eta \mathbf{m}_i^{\sigma_e(r)} \odot \nabla f_i(\mathbf{m}_i^{\sigma_e(r)} \odot \mathbf{w'}^i_{e,r,k}; \xi'^i_{e,r,k})\end{aligned}\right\| \\
&\leq \mathbb{E}\left\|\mathbf{w}^i_{e,r,k} - \mathbf{w'}^i_{e,r,k}\right\| + \eta\left\|\mathbf{m}_i^{\sigma_e(r)} \odot \nabla f_i(\mathbf{m}_i^{\sigma_e(r)} \odot \mathbf{w}^i_{e,r,k}; \xi^i_{e,r,k})\right\| \\
&\quad + \eta\left\|\mathbf{m}_i^{\sigma_e(r)} \odot \nabla f_i(\mathbf{m}_i^{\sigma_e(r)} \odot \mathbf{w'}^i_{e,r,k}; \xi'^i_{e,r,k})\right\| \\
&\leq \mathbb{E}\left\|\mathbf{w}^i_{e,r,k} - \mathbf{w'}^i_{e,r,k}\right\| + 2\eta\delta + \eta\left\|\mathbf{m}_i^{\sigma_e(r)} \odot \nabla f_i(\mathbf{m}_i^{\sigma_e(r)} \odot \mathbf{w}^i_{e,r,k})\right\| \\
&\quad + \eta\left\|\mathbf{m}_i^{\sigma_e(r)} \odot \nabla f'_i(\mathbf{m}_i^{\sigma_e(r)} \odot \mathbf{w'}^i_{e,r,k})\right\| \\
&\leq \mathbb{E}\left\|\mathbf{w}^i_{e,r,k} - \mathbf{w'}^i_{e,r,k}\right\| + 2\eta\delta + \eta\left\|\nabla F_{\mathbf{m}}(\mathbf{w}^i_{e,r,k})\right\| + \eta\left\|F'_{\mathbf{m}}(\mathbf{w'}^i_{e,r,k})\right\| + 2\eta\zeta_{\mathbf{m},\max} \\
&\leq \mathbb{E}\left\|\mathbf{w}^i_{e,r,k} - \mathbf{w'}^i_{e,r,k}\right\| + \eta\mathbb{E}\left\|\nabla F_{\mathbf{m}}(\mathbf{w}_e)\right\| + \eta\mathbb{E}\left\|\nabla F'_{\mathbf{m}}(\mathbf{w'}_e)\right\| + \eta L\left\|\mathbf{w}_e - \mathbf{w}^i_{e,r,k}\right\| \\
&\quad + \eta L\left\|\mathbf{w'}_e - \mathbf{w'}^i_{e,r,k}\right\| + 2\eta\delta + 2\eta\zeta_{\mathbf{m},\max} \\
&\leq \mathbb{E}\left\|\mathbf{w}^i_{e,r,k} - \mathbf{w'}^i_{e,r,k}\right\| + \eta\mathbb{E}\left\|\nabla F_{\mathbf{m}}(\mathbf{w}_e)\right\| + \eta\mathbb{E}\left\|\nabla F'_{\mathbf{m}}(\mathbf{w'}_e)\right\| + \eta L\left\|\mathbf{w}_{e,r} - \mathbf{w}^i_{e,r,k}\right\| \\
&\quad + \eta L\left\|\mathbf{w'}_{e,r} - \mathbf{w'}^i_{e,r,k}\right\| + 2\eta\delta + 2\eta\zeta_{\mathbf{m},\max} \\
&\quad + \eta L\left\|\mathbf{w}_e - \mathbf{w}_{e,r}\right\| + \eta L\left\|\mathbf{w'}_e - \mathbf{w'}_{e,r}\right\|.
\end{aligned}
$$

To bound $\mathbb{E}\left\|\mathbf{w}_e - \mathbf{w}_{e,r}\right\|$ and $\mathbb{E}\left\|\mathbf{w'}_e - \mathbf{w'}_{e,r}\right\|$, we evoke (14):

$$
\begin{aligned}
\mathbb{E}\left\|\mathbf{w}_{e,r} - \mathbf{w}_e\right\|^2 \leq{}& \left(48r\eta^2K^2 + 432r^3\eta^4K^4L^2\right)G\mathrm{log}(2RK/\nu) \qquad (35)\\
&+ \left(6r^2\eta^2K^2 + 18r^4\eta^4K^4L^2\right)\left\|\nabla F_{\mathbf{m}}(\mathbf{w}_e)\right\|^2 \\
&+ 18\eta^2 r\sum_{p=0}^{r-1}KL^2\sum_{k=0}^{K-1}\frac{1}{N}\sum_{i=1}^{N}\mathbb{E}\left\|\mathbf{w}^i_{e,p,k} - \mathbf{w}_{e,p}\right\|^2 + 18r^2\eta^2K^2\frac{\delta^2}{N}. \qquad (36)
\end{aligned}
$$

Hence we have

$$
\begin{aligned}
&\mathbb{E}\left\|\mathbf{w}^i_{e,r,k+1} - \mathbf{w'}^i_{e,r,k+1}\right\| \\
&\leq \mathbb{E}\left\|\mathbf{w}^i_{e,r,k} - \mathbf{w'}^i_{e,r,k}\right\| + \left(\eta + \eta L\sqrt{6r^2\eta^2K^2 + 18r^4\eta^4K^4L^2}\right)\mathbb{E}\left\|\nabla F_{\mathbf{m}}(\mathbf{w}_e)\right\| \\
&\quad + \left(\eta + \eta L\sqrt{6r^2\eta^2K^2 + 18r^4\eta^4K^4L^2}\right)\mathbb{E}\left\|\nabla F'_{\mathbf{m}}(\mathbf{w'}_e)\right\| \\
&\quad + \eta L\sqrt{18\eta^2 r\sum_{p=0}^{r-1}KL^2\sum_{k=0}^{K-1}\frac{1}{N}\sum_{i=1}^{N}\mathbb{E}\left\|\mathbf{w}^i_{e,p,k} - \mathbf{w}_{e,p}\right\|^2 + 18r^2\eta^2K^2\frac{\delta^2}{N}} \\
&\quad + \eta L\sqrt{18\eta^2 r\sum_{p=0}^{r-1}KL^2\sum_{k=0}^{K-1}\frac{1}{N}\sum_{i=1}^{N}\mathbb{E}\left\|\mathbf{w'}^i_{e,p,k} - \mathbf{w'}_{e,p}\right\|^2 + 18r^2\eta^2K^2\frac{\delta^2}{N}} \\
&\quad + 4\eta L\sqrt{\left(48r\eta^2K^2 + 432r^3\eta^4K^4L^2\right)G\mathrm{log}(2RK/\nu)} \\
&\quad + \eta L\mathbb{E}\left\|\mathbf{w}_{e,r} - \mathbf{w}^i_{e,r,k}\right\| + \eta L\mathbb{E}\left\|\mathbf{w'}_{e,r} - \mathbf{w'}^i_{e,r,k}\right\| + 2\eta\delta + 2\eta\zeta_{\mathbf{m},\max}.
\end{aligned}
$$

For $j \neq i$, we have $\mathbb{E}\left\|\mathbf{w}^j_{e,r,k+1} - \mathbf{w'}^j_{e,r,k+1}\right\| \leq \mathbb{E}\left\|\mathbf{w}^j_{e,r,k} - \mathbf{w'}^j_{e,r,k}\right\|$. Combining two cases we have

$$
\begin{aligned}
&\frac{1}{N}\sum_{j=1}^{N}\mathbb{E}\left\|\mathbf{w}^j_{e,r,k+1} - \mathbf{w'}^j_{e,r,k+1}\right\| \\
&\leq \frac{1}{N}\sum_{j=1}^{N}\mathbb{E}\left\|\mathbf{w}^j_{e,r,k} - \mathbf{w'}^j_{e,r,k}\right\| \\
&+ \frac{1}{Nn}\left(\eta + \eta L\sqrt{6r^2\eta^2K^2 + 18r^4\eta^4K^4L^2}\right)\left(\mathbb{E}\left\|\nabla F_{\mathbf{m}}(\mathbf{w}_e)\right\| + \mathbb{E}\left\|\nabla F'_{\mathbf{m}}(\mathbf{w'}_e)\right\|\right) \\
&+ \frac{1}{Nn}\eta L\sqrt{18\eta^2 r\sum_{p=0}^{r-1}KL^2\sum_{k=0}^{K-1}\frac{1}{N}\sum_{i=1}^{N}\mathbb{E}\left\|\mathbf{w}^i_{e,p,k} - \mathbf{w}_{e,p}\right\|^2 + 18r^2\eta^2K^2\frac{\delta^2}{N}} \\
&+ \frac{1}{Nn}\eta L\sqrt{18\eta^2 r\sum_{p=0}^{r-1}KL^2\sum_{k=0}^{K-1}\frac{1}{N}\sum_{i=1}^{N}\mathbb{E}\left\|\mathbf{w'}^i_{e,p,k} - \mathbf{w}'_{e,p}\right\|^2 + 18r^2\eta^2K^2\frac{\delta^2}{N}} \\
&+ 4\frac{1}{Nn}\eta L\sqrt{\left(48r\eta^2K^2 + 432r^3\eta^4K^4L^2\right)G\log(2RK/\nu)} \\
&+ \frac{1}{Nn}\eta L\mathbb{E}\left\|\mathbf{w}_{e,r} - \mathbf{w}^i_{e,r,k}\right\| + \frac{1}{Nn}\eta L\mathbb{E}\left\|\mathbf{w'}_{e,r} - \mathbf{w'}^i_{e,r,k}\right\| \\
&+ 2\frac{1}{Nn}\eta\delta + 2\frac{1}{Nn}\eta\zeta_{\mathbf{m},\max}.
\end{aligned}
$$

Performing telescoping sum from $k = K-1$ to 0 yields:

$$
\begin{aligned}
&\mathbb{E}\left\|\mathbf{w}_{e,r+1} - \mathbf{w'}_{e,r+1}\right\| \\
&\leq \mathbb{E}\left\|\mathbf{w}_{e,r} - \mathbf{w'}_{e,r}\right\| \\
&+ \frac{1}{Nn}K\left(\eta + \eta L\sqrt{6r^2\eta^2K^2 + 18r^4\eta^4K^4L^2}\right) \\
&\cdot \left(\mathbb{E}\left\|\nabla F_{\mathbf{m}}(\mathbf{w}_e)\right\| + \mathbb{E}\left\|\nabla F'_{\mathbf{m}}(\mathbf{w'}_e)\right\|\right) \\
&+ \frac{1}{Nn}K\eta L\sqrt{18\eta^2 r\sum_{p=0}^{r-1}KL^2\sum_{k=0}^{K-1}\frac{1}{N}\sum_{i=1}^{N}\mathbb{E}\left\|\mathbf{w}^i_{e,p,k} - \mathbf{w}_{e,p}\right\|^2 + 18r^2\eta^2K^2\frac{\delta^2}{N}} \\
&+ \frac{1}{Nn}K\eta L\sqrt{18\eta^2 r\sum_{p=0}^{r-1}KL^2\sum_{k=0}^{K-1}\frac{1}{N}\sum_{i=1}^{N}\mathbb{E}\left\|\mathbf{w'}^i_{e,p,k} - \mathbf{w}'_{e,p}\right\|^2 + 18r^2\eta^2K^2\frac{\delta^2}{N}} \\
&+ 4\frac{1}{Nn}K\eta L\sqrt{\left(48r\eta^2K^2 + 432r^3\eta^4K^4L^2\right)G\log(2RK/\nu)} \\
&+ \frac{1}{Nn}\eta L\sum_{k=0}^{K-1}\mathbb{E}\left\|\mathbf{w}_{e,r} - \mathbf{w}^i_{e,r,k}\right\| + \frac{1}{Nn}\eta L\sum_{k=0}^{K-1}\mathbb{E}\left\|\mathbf{w'}_{e,r} - \mathbf{w'}^i_{e,r,k}\right\| \\
&+ 2\frac{K}{Nn}\eta\delta + 2\frac{K}{Nn}\eta\zeta_{\mathbf{m},\max}.
\end{aligned}
$$

Performing telescoping sum from $r = R-1$ to 0, and using the fact $\mathbf{w}_0 = \mathbf{w}'_0$ yields:

$$
\begin{aligned}
&\mathbb{E}\left\|\mathbf{w}_{e+1}-\mathbf{w'}_{e+1}\right\| \\
&\leq \mathbb{E}\left\|\mathbf{w}_{e}-\mathbf{w'}_{e}\right\| \\
&+\frac{1}{Nn}RK\left(\eta+\eta L\sqrt{6R^2\eta^2K^2+18R^4\eta^4K^4L^2}\right) \\
&\quad\cdot\left(\mathbb{E}\left\|\nabla F_{\mathbf{m}}(\mathbf{w}_e)\right\|+\mathbb{E}\left\|\nabla F'_{\mathbf{m}}(\mathbf{w'}_e)\right\|\right) \\
&+\frac{RK\eta L}{Nn}\sqrt{18\eta^2RKL^2\sum_{p=0}^{R-1}\sum_{k=0}^{K-1}\frac{1}{N}\sum_{i=1}^{N}\mathbb{E}\left\|\mathbf{w}^i_{e,p,k}-\mathbf{w}_{e,p}\right\|^2+18R^2\eta^2K^2\frac{\delta^2}{N}} \\
&+\frac{RK\eta L}{Nn}\sqrt{18\eta^2RKL^2\sum_{p=0}^{R-1}\sum_{k=0}^{K-1}\frac{1}{N}\sum_{i=1}^{N}\mathbb{E}\left\|\mathbf{w'}^i_{e,p,k}-\mathbf{w}'_{e,p}\right\|^2+18R^2\eta^2K^2\frac{\delta^2}{N}} \\
&+4\frac{RK\eta L}{Nn}\sqrt{\left(48R\eta^2K^2+432R^3\eta^4K^4L^2\right)G\log(2RK/\nu)} \\
&+\frac{1}{Nn}\eta L\sum_{r=0}^{R-1}\sum_{k=0}^{K-1}\mathbb{E}\left\|\mathbf{w}_{e,r}-\mathbf{w}^i_{e,r,k}\right\|+\frac{1}{Nn}\eta L\sum_{r=0}^{R-1}\sum_{k=0}^{K-1}\mathbb{E}\left\|\mathbf{w'}_{e,r}-\mathbf{w'}^i_{e,r,k}\right\| \\
&+2\frac{RK}{Nn}\eta\delta+2\frac{RK}{Nn}\eta\zeta_{\mathbf{m},\max}.
\end{aligned}
$$

Note the fact that

$$
\begin{aligned}
\sum_{r=0}^{R-1}\sum_{k=0}^{K-1}\mathbb{E}\left\|\mathbf{w}_{e,r}-\mathbf{w}^i_{e,r,k}\right\| &= RK\cdot\frac{1}{RK}\sum_{r=0}^{R-1}\sum_{k=0}^{K-1}\mathbb{E}\sqrt{\left\|\mathbf{w}_{e,r}-\mathbf{w}^i_{e,r,k}\right\|^2} \\
&\leq RK\cdot\sqrt{\frac{1}{RK}\sum_{r=0}^{R-1}\sum_{k=0}^{K-1}\mathbb{E}\left\|\mathbf{w}_{e,r}-\mathbf{w}^i_{e,r,k}\right\|^2},
\end{aligned}
$$

where we apply the Jensen's inequality and concavity of square root function. Hence we have

$$
\begin{aligned}
&\mathbb{E}\left\|\mathbf{w}_{e+1}-\mathbf{w'}_{e+1}\right\| \\
&\leq \mathbb{E}\left\|\mathbf{w}_{e}-\mathbf{w'}_{e}\right\| \\
&+\frac{1}{Nn}RK\left(\eta+\eta L\sqrt{6R^2\eta^2K^2+18R^4\eta^4K^4L^2}\right)\left(\mathbb{E}\left\|\nabla F_{\mathbf{m}}(\mathbf{w}_e)\right\|+\mathbb{E}\left\|\nabla F'_{\mathbf{m}}(\mathbf{w'}_e)\right\|\right) \\
&+\frac{RK\eta L}{Nn}\left(\sqrt{18\eta^2RKL^2\sum_{p=0}^{R-1}\sum_{k=0}^{K-1}\frac{1}{N}\sum_{i=1}^{N}\mathbb{E}\left\|\mathbf{w}^i_{e,p,k}-\mathbf{w}_{e,p}\right\|^2}+\sqrt{18R^2\eta^2K^2\frac{\delta^2}{N}}\right) \\
&+\frac{RK\eta L}{Nn}\left(\sqrt{18\eta^2RKL^2\sum_{p=0}^{R-1}\sum_{k=0}^{K-1}\frac{1}{N}\sum_{i=1}^{N}\mathbb{E}\left\|\mathbf{w'}^i_{e,p,k}-\mathbf{w'}_{e,p}\right\|^2}+\sqrt{18R^2\eta^2K^2\frac{\delta^2}{N}}\right) \\
&+4\frac{1}{Nn}RK\eta L\sqrt{\left(48R\eta^2K^2+432R^3\eta^4K^4L^2\right)G\log(2RK/\nu)} \\
&+\frac{1}{Nn}RK\eta L\sqrt{\frac{1}{RK}\sum_{r=0}^{R-1}\sum_{k=0}^{K-1}\mathbb{E}\left\|\mathbf{w}_{e,r}-\mathbf{w}^i_{e,r,k}\right\|^2} \\
&+\frac{1}{Nn}RK\eta L\sqrt{\frac{1}{RK}\sum_{r=0}^{R-1}\sum_{k=0}^{K-1}\mathbb{E}\left\|\mathbf{w'}_{e,r}-\mathbf{w'}^i_{e,r,k}\right\|^2} \\
&+2\frac{RK}{Nn}\eta\delta+2\frac{RK}{Nn}\eta\zeta_{\mathbf{m},\max} \\
&=\mathbb{E}\left\|\mathbf{w}_{e}-\mathbf{w'}_{e}\right\|+\frac{1}{Nn}RK\left(\eta+\eta L\sqrt{6R^2\eta^2K^2+18R^4\eta^4K^4L^2}\right) \\
&\cdot\left(\mathbb{E}\left\|\nabla F_{\mathbf{m}}(\mathbf{w}_e)\right\|+\mathbb{E}\left\|\nabla F'_{\mathbf{m}}(\mathbf{w'}_e)\right\|\right) \\
&+\frac{RK\eta L}{Nn}\left(1+\sqrt{18\eta^2RKL^2}\right) \\
&\cdot\left(\sqrt{\frac{1}{RK}\sum_{r=0}^{R-1}\sum_{k=0}^{K-1}\mathbb{E}\left\|\mathbf{w}_{e,r}-\mathbf{w}^i_{e,r,k}\right\|^2}+\sqrt{\frac{1}{RK}\sum_{r=0}^{R-1}\sum_{k=0}^{K-1}\mathbb{E}\left\|\mathbf{w'}_{e,r}-\mathbf{w'}^i_{e,r,k}\right\|^2}\right) \\
&+\frac{2RK\eta L}{Nn}\sqrt{18R^2\eta^2K^2\frac{\delta^2}{N}}+2\frac{RK}{Nn}\eta(\delta+\zeta_{\mathbf{m},\max}) \\
&+\frac{4RK\eta L}{Nn}\sqrt{\left(48R\eta^2K^2+432R^3\eta^4K^4L^2\right)G\log(2RK/\nu)}.
\end{aligned}
$$

Now we plug in Lemma 5:

$$
\begin{aligned}
&\mathbb{E}\left\|\mathbf{w}_{e+1}-\mathbf{w'}_{e+1}\right\| \leq \mathbb{E}\left\|\mathbf{w}_{e}-\mathbf{w'}_{e}\right\| \\
&+\frac{\eta RK}{Nn}\left(1+L\sqrt{R^2\eta^2K^2+R^4\eta^4K^4L^2}\right) \\
&\times\left(\sqrt{\mathbb{E}\left\|\nabla F_{\mathbf{m}}(\mathbf{w}_e)\right\|^2}+\sqrt{\mathbb{E}\left\|\nabla F'_{\mathbf{m}}(\mathbf{w'}_e)\right\|^2}\right) \\
&+\frac{\eta RK}{Nn}\left(L\left(1+\sqrt{\eta^2RKL^2}\right)\sqrt{R^3\eta^4K^5L^2+R^5\eta^6K^7L^4+\eta^2RK^3}\right) \\
&\times\left(\sqrt{\mathbb{E}\left\|\nabla F_{\mathbf{m}}(\mathbf{w}_e)\right\|^2}+\sqrt{\mathbb{E}\left\|\nabla F'_{\mathbf{m}}(\mathbf{w'}_e)\right\|^2}\right) \\
&+\frac{RK\eta L}{Nn}\times\left(1+\sqrt{\eta^2RKL^2}\right) \\
&\cdot\left(\sqrt{\eta^2RK^3\zeta+\left(R^2\eta^4K^5L^2+R^4\eta^6K^7L^4\right)G\log(2RK/\nu)+R^3\eta^4K^5L^2\frac{\delta^2}{N}+\eta^2RK^3\delta^2}\right) \\
&+\frac{RK\eta L}{Nn}\sqrt{R^2\eta^2K^2\frac{\delta^2}{N}}+\frac{RK}{Nn}\eta(\delta+\zeta_{\mathbf{m},\max}) \\
&+\frac{RK\eta L}{Nn}\sqrt{\left(R\eta^2K^2+R^3\eta^4K^4L^2\right)G\log(2RK/\nu)}.
\end{aligned}
$$

Performing telescoping sum yields:

$$
\begin{aligned}
&\mathbb{E}\left\|\mathbf{w}_{T}-\mathbf{w'}_{T}\right\| \\
&\leq\frac{\eta TRK}{Nn}\left(1+L\sqrt{R^2\eta^2K^2+R^4\eta^4K^4L^2}\right) \\
&\times\frac{1}{T}\sum_{t=1}^{T}\left(\sqrt{\mathbb{E}\left\|\nabla F_{\mathbf{m}}(\mathbf{w}_e)\right\|^2}+\sqrt{\mathbb{E}\left\|\nabla F'_{\mathbf{m}}(\mathbf{w'}_e)\right\|^2}\right) \\
&+\frac{\eta TRK}{Nn}L\left(1+\sqrt{\eta^2RKL^2}\right)\sqrt{R^3\eta^4K^5L^2+R^5\eta^6K^7L^4+\eta^2RK^3} \\
&\times\frac{1}{T}\sum_{t=1}^{T}\left(\sqrt{\mathbb{E}\left\|\nabla F_{\mathbf{m}}(\mathbf{w}_e)\right\|^2}+\sqrt{\mathbb{E}\left\|\nabla F'_{\mathbf{m}}(\mathbf{w'}_e)\right\|^2}\right) \\
&+\frac{TRK\eta L}{Nn}\left(1+\sqrt{\eta^2RKL^2}\right) \\
&\cdot\left(\sqrt{\eta^2RK^3\zeta+\left(R^2\eta^4K^5L^2+R^4\eta^6K^7L^4\right)G\log(2RK/\nu)+R^3\eta^4K^5L^2\frac{\delta^2}{N}+\eta^2RK^3\delta^2}\right) \\
&+\frac{TRK\eta L}{Nn}\sqrt{R^2\eta^2K^2\frac{\delta^2}{N}}+\frac{TRK}{Nn}\eta(\delta+\zeta_{\mathbf{m},\max}) \\
&+\frac{TRK\eta L}{Nn}\sqrt{\left(R\eta^2K^2+R^3\eta^4K^4L^2\right)G\log(2RK/\nu)}
\end{aligned}
$$

Now we will bound the gradient norm $\mathbb{E}\left\|\nabla F_{\mathbf{m}}(\mathbf{w}_r)\right\|^2$. Due to Eq.(7), we have

$$
\begin{aligned}
\frac{1}{T}\sum_{e=1}^{T}\mathbb{E}\left\|\nabla F_{\mathbf{m}}(\mathbf{w}_e)\right\|^2 \leq& O\left(\frac{\mathbb{E}[F_{\mathbf{m}}(\mathbf{w}_0)]}{\eta RKT}\right) \\
&+O\left(L^2\eta^2K^2RG\log(2RK/\nu)+\eta^2RK^3\zeta^2+\eta RK\frac{\delta^2}{N}\right)
\end{aligned}
$$

Choosing $\eta = \frac{\sqrt{Nn}}{TRK}$ and $T \geq \sqrt{\frac{n}{N}}$ yields:

$$\begin{aligned}
&\mathbb{E}\left\|\mathbf{w}_T - \mathbf{w'}_T\right\| \\
&\leq O\left(\frac{\delta + \zeta_{\mathbf{m},\max}}{\sqrt{Nn}} + \frac{1}{\sqrt{Nn}}\left(\sqrt{\frac{\mathbb{E}[F_{\mathbf{m}}(\mathbf{w}_0)]}{\sqrt{Nn}}} + \frac{L^2 G \log(2RK/\nu)}{T^2 R} + \frac{K\zeta^2}{T^2 RK} + \frac{\sqrt{Nn}\delta^2}{TN}\right)\right) \\
&\leq O\left(\frac{\delta + \zeta_{\mathbf{m},\max}}{\sqrt{Nn}} + \frac{1}{\sqrt{Nn}}\left(\sqrt{\frac{\mathbb{E}[F_{\mathbf{m}}(\mathbf{w}_0)]}{\sqrt{Nn}}} + \frac{L^2 G \log(2RK/\nu)}{T^2 R} + \frac{K\zeta^2}{T^2 R} + \frac{\sqrt{Nn}\delta^2}{TN}\right)\right).
\end{aligned}$$

Plugging Lemma 13 will conclude the proof.