Appendix

.1. Pseudocodes

Algorithm 5 FedAvgM (FedProxM)

Input: learning rates (η_l, η_g), control parameters μ and β_1, synchronization interval I and the number of workers N.
Initial state $\boldsymbol{x}_0^{(i)} = \boldsymbol{x}_0 \in \mathbb{R}^d$, $\forall i \in [N]$ and $\boldsymbol{m}_0 = \boldsymbol{0}$.
for $t = 0, 1, \ldots, T-1$ **do**
  On server:
  Server samples a subset $\mathcal{S}_t$ with S active workers from $[N]$ and transmits $\boldsymbol{x}_t$ to $\mathcal{S}_t$.
  On workers:
  for $i \in \mathcal{S}_t$ parallel **do**
    Sets $\boldsymbol{x}_{t,0}^{(i)} = \boldsymbol{x}_t$.
    for $\tau = 0, 1, \ldots, I-1$ **do**
      $\boldsymbol{x}_{t,\tau+1}^{(i)} = \boldsymbol{x}_{t,\tau}^{(i)} - \eta_l \nabla f_i(\boldsymbol{x}_{t,\tau}^{(i)})$. (FedAvgM)
      $\boldsymbol{x}_{t,\tau+1}^{(i)} = \boldsymbol{x}_{t,\tau}^{(i)} - \eta_l (\nabla f_i(\boldsymbol{x}_{t,\tau}^{(i)}) + \mu(\boldsymbol{x}_{t,\tau}^{(i)} - \boldsymbol{x}_t))$. (FedProxM)
    end for
    Sends $\boldsymbol{d}_{t+1}^{(i)} = \boldsymbol{x}_t - \boldsymbol{x}_{t,I}^{(i)}$ to server.
  end for
  On server:
  $\boldsymbol{d}_{t+1} = \frac{1}{S} \sum_{i \in \mathcal{S}_t} \boldsymbol{d}_{t+1}^{(i)}$, $\boldsymbol{m}_{t+1} = \beta_1 \boldsymbol{m}_t + \boldsymbol{d}_{t+1}$.
  $\boldsymbol{x}_{t+1} = \boldsymbol{x}_t - \eta_g \boldsymbol{m}_{t+1}$.
  Sends $\boldsymbol{x}_{t+1}$ to sampled active workers in the next round.
end for
Output: $\boldsymbol{x}_T$

Algorithm 6 MIFAM (MIFA, i.e., MIFAM with $\beta_1 = 0.0$)

Input: learning rates (η_l, η_g), control parameter β_1, synchronization interval I and the number of workers N.
Initial state $\boldsymbol{x}_0^{(i)} = \boldsymbol{x}_0 \in \mathbb{R}^d$, $\boldsymbol{g}_{old}^{(i)} = \boldsymbol{0}$, $\forall i \in [N]$, $\boldsymbol{d}_0 = \frac{1}{N} \sum_{i=1}^N \boldsymbol{g}_{old}^{(i)}$ and $\boldsymbol{m}_0 = \boldsymbol{0}$.
for $t = 0, 1, \ldots, T-1$ **do**
  On server:
  Server samples a subset $\mathcal{S}_t$ with S active workers from $[N]$ and transmits $\boldsymbol{x}_t$ to $\mathcal{S}_t$.
  On workers:
  for $i \in S_t$ parallel **do**
    Sets $\boldsymbol{x}_{t,0}^{(i)} = \boldsymbol{x}_t$.
    for $\tau = 0, 1, \ldots, I-1$ **do**
      $\boldsymbol{x}_{t,\tau+1}^{(i)} = \boldsymbol{x}_{t,\tau}^{(i)} - \eta_l \nabla f_i(\boldsymbol{x}_{t,\tau}^{(i)})$.
    end for
    Computes $\boldsymbol{g}_{t+1}^{(i)} = \boldsymbol{x}_t - \boldsymbol{x}_{t,I}^{(i)}$.
    Sends $\boldsymbol{d}_{t+1}^{(i)} = \boldsymbol{g}_{t+1}^{(i)} - \boldsymbol{g}_{old}^{(i)}$ to Server.
    Sets $\boldsymbol{g}_{old}^{(i)} = \boldsymbol{g}_{t+1}^{(i)}$.
  end for
  On server:
  $\boldsymbol{d}_{t+1} = \boldsymbol{d}_t + \frac{1}{N} \sum_{i \in \mathcal{S}_t} \boldsymbol{d}_{t+1}^{(i)}$, $\boldsymbol{m}_{t+1} = \beta_1 \boldsymbol{m}_t + \boldsymbol{d}_{t+1}$.
  $\boldsymbol{x}_{t+1} = \boldsymbol{x}_t - \eta_g \boldsymbol{m}_{t+1}$.
  Sends $\boldsymbol{x}_{t+1}$ to sampled active workers in the next round.
end for
Output: $\boldsymbol{x}_T$

Algorithm 7 GradMA-W

Input: learning rates (η_l, η_g), the number of all workers N, the number of active workers each round S and synchronization interval I.
Initial state $\boldsymbol{x}_0^{(i)} = \boldsymbol{x}_0 \in \mathbb{R}^d, \forall i \in [N]$.
for $t = 0, 1, \ldots, T-1$ **do**
  On server:
  Server samples a subset $\mathcal{S}_t$ with S active workers and transmits $\boldsymbol{x}_t$ to $\mathcal{S}_t$.
  On workers:
  for $i \in \mathcal{S}_t$ parallel **do**
    $\boldsymbol{x}_{t+1}^{(i)} = \text{Worker\_Update}(\boldsymbol{x}_t^{(i)}, \boldsymbol{x}_t, \eta_l, I)$,
    sends $\boldsymbol{d}_{t+1}^{(i)} = \boldsymbol{x}_t - \boldsymbol{x}_{t+1}^{(i)}$ to server.
  end for
  On server:
  $\boldsymbol{d}_{t+1} = \frac{1}{S}\sum_{i\in\mathcal{S}_t} \boldsymbol{d}_{t+1}^{(i)}$, $\boldsymbol{x}_{t+1} = \boldsymbol{x}_t - \eta_g \boldsymbol{d}_{t+1}$.
  Sends $\boldsymbol{x}_{t+1}$ to sampled active workers in the next round.
end for
Output: $\boldsymbol{x}_T$

Algorithm 8 GradMA-S

Input: learning rates (η_l, η_g), the number of all workers N, the number of sampled active workers per communication round S, control parameters (β_1, β_2), synchronization interval I and memory size m ($S \leq m \leq \min\{d, N\}$).
Initial state $\boldsymbol{x}_0^{(i)} = \boldsymbol{x}_0 \in \mathbb{R}^d, \forall i \in [N]$, $\tilde{\boldsymbol{m}}_0 = \boldsymbol{0}$.
Initial $counter = \{c(i) = 0\}, \forall i \in [N]$.
Initial memory state $\boldsymbol{D} = \{\}$.
$buf = \{\}$, $new\_buf = \{\}$.
for $t = 0, 1, \ldots, T-1$ **do**
  On server:
  Server samples a subset $\mathcal{S}_t$ with S active workers and transmits $\boldsymbol{x}_t$ to $\mathcal{S}_t$.
  $counter, \boldsymbol{D}, buf, new\_buf \leftarrow \text{mem\_red}\ (m, \mathcal{S}_t, counter, \boldsymbol{D}, buf, new\_buf)$.
  On workers:
  for $i \in \mathcal{S}_t$ parallel **do**
    Sets $\boldsymbol{x}_{t,0}^{(i)} = \boldsymbol{x}_t$.
    for $\tau = 0, 1, \ldots, I-1$ **do**
      $\boldsymbol{x}_{t,\tau+1}^{(i)} = \boldsymbol{x}_{t,\tau}^{(i)} - \eta_l \nabla f_i(\boldsymbol{x}_{t,\tau}^{(i)})$.
    end for
    Sends $\boldsymbol{d}_{t+1}^{(i)} = \boldsymbol{x}_t - \boldsymbol{x}_{t,I}^{(i)}$ to server.
  end for
  On server:
  $\boldsymbol{D}, \boldsymbol{x}_{t+1}, \tilde{\boldsymbol{m}}_{t+1} = \text{Server\_Update}([\boldsymbol{d}_{t+1}^{(i)}, i \in \mathcal{S}_t], \tilde{\boldsymbol{m}}_t, \boldsymbol{D}, \eta_g, \beta_1, \beta_2, buf, new\_buf)$.
  Sends $\boldsymbol{x}_{t+1}$ to sampled active workers in the next round.
  $new\_buf = \{\}$.
end for
Output: $\boldsymbol{x}_T$

.2. Complete Empirical Study

.2.1 Experimental Setup

To gauge the effectiveness of Worker_Update() and Server_Update(), we perform ablation study of GradMA. For this purpose, we design Alg. 7 (marked as GradMA-W) and Alg. 8 (marked as GradMA-S), as specified in Appendix .1. Meanwhile, we compare other baselines, including FedAvg [26], FedProx [19], MOON [17], FedMLB [13], Scaffold [12], Fed-

Dyn [1], MimeLite [11], MIFA [5] and slow-momentum variants of FedAvg, FedProx, MIFA, MOON and FedMLB (i.e., FedAvgM [7], FedProxM, MIFAM, MOONM and FedMLBM), in terms of test accuracy and communication efficiency in different FL scenarios. For fairness, we divide the baselines into three groups based on FedAvg's improvements on the worker side, server side, or both. Furthermore, on top of GradMA-S, we empirically study the effect of the control parameters (β_1, β_2) and verify the effectiveness of men_red() by setting varying memory sizes m.

All our experiments are performed on a centralized network with 100 workers. And fix synchronization interval $I = 5$. To explore the performances of the approaches, we set up multiple different scenarios w.r.t. the number of sampled active workers S per communication round and data heterogeneity. Specifically, we set $S \in \{5, 10, 50\}$. Furthermore, we use Dirichlet process $Dp(\omega)$ [1, 46] to strictly partition the training set of each dataset across 100 workers, where the scaling parameter ω controls the degree of data heterogeneity across workers. Notably, a smaller ω corresponds to higher data heterogeneity. We set $\omega \in \{0.01, 0.1, 1.0\}$. A visualization of the data partitions for the four datasets at varying ω values can be found in Fig. 8. Also, the original testing set (without partitioning) of each dataset is used to evaluate the performance of the trained centralized model. For MNIST, a neural network (NN) with three linear hidden layers is implemented for each worker. We fix the total number of iterations to 2500, i.e., $T \times I = 2500$. For CIFAR-10 (CIFAR-100, Tiny-Imagenet), each worker implements a Lenet-5 [16] (VGG-11 [31], Resnet20 [6]) architecture. We fix the total number of iterations to 5000 (10000, 10000), i.e., $T \times I = 5000$ (10000, 10000).

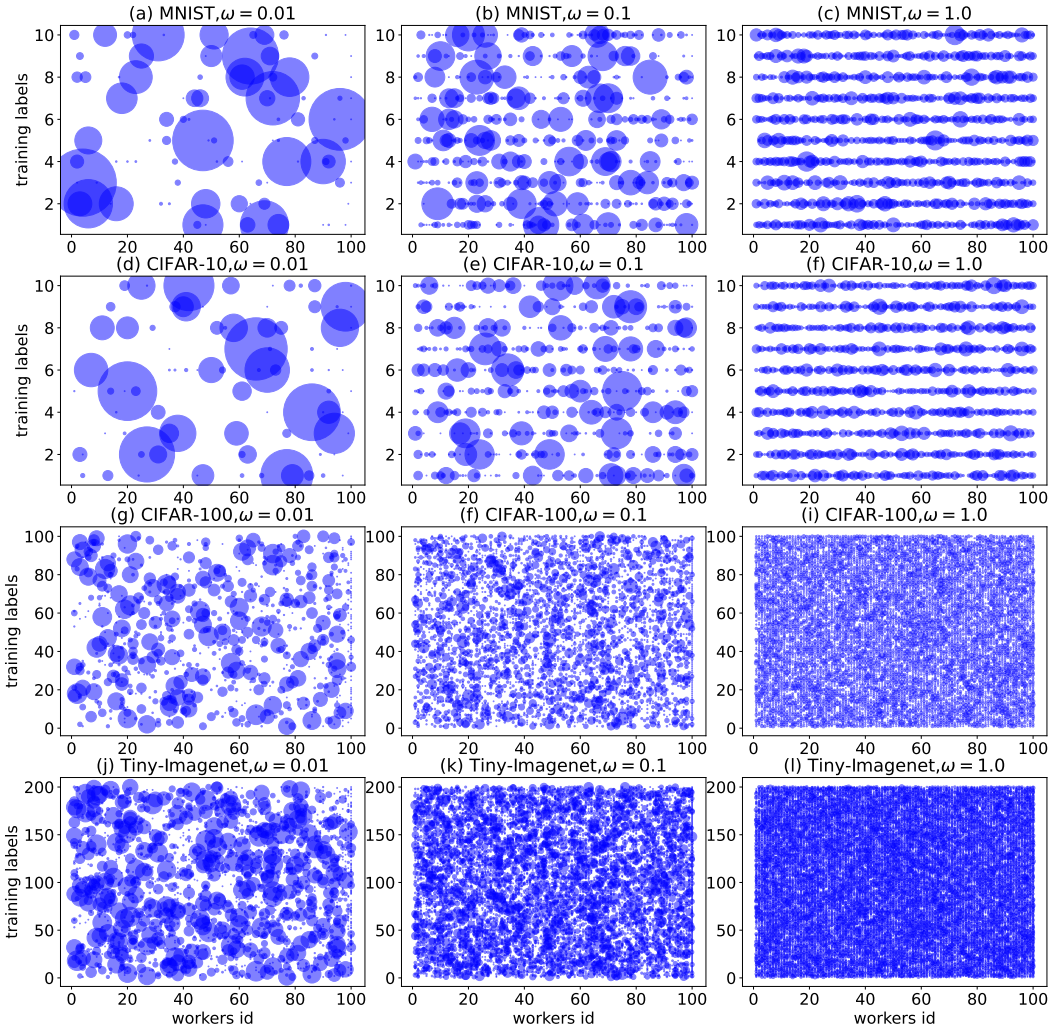

Figure 8. Data heterogeneity among workers is visualized on four datasets (MNIST, CIFAR-10, CIFAR-100 and Tiny-Imagenet), where the x-axis represents the workers id, the y-axis represents the class labels on the training set, and the size of scattered points represents the number of training samples with available labels for that worker.

We perform careful hyper-parameters tuning of all approaches. We set the local learning rate η_l for each worker to $\eta_l \in \{0.001, 0.01, 0.1\}$ and the global learning rate η_g for server to $\eta_g \in \{0.1, 1.0, 10.0\}$. The control parameter μ for FedProx (FedProxM) and α for FedDyn are fine-tuned within $\{0.001, 0.01, 0.1\}$. For control parameters (β_1, β_2), we set $\beta_1, \beta_2 \in \{0.1, 0.5, 0.9\}$ unless otherwise specified. Also, we fix memory size $m = 100$ unless otherwise specified. For the remaining tunable hyper-parameters of MOON (MOONM) and FedMLB (FedMLBM), we follow the settings of [17] and [13], respectively. For fairness, the popular SGD procedure is employed to perform local update steps for each worker. For all experiments, we fix batch size to 64 for all datasets. To ensure reliability, we report the average for each experiment over 3 random seeds.

.2.2 Full Experimental Results

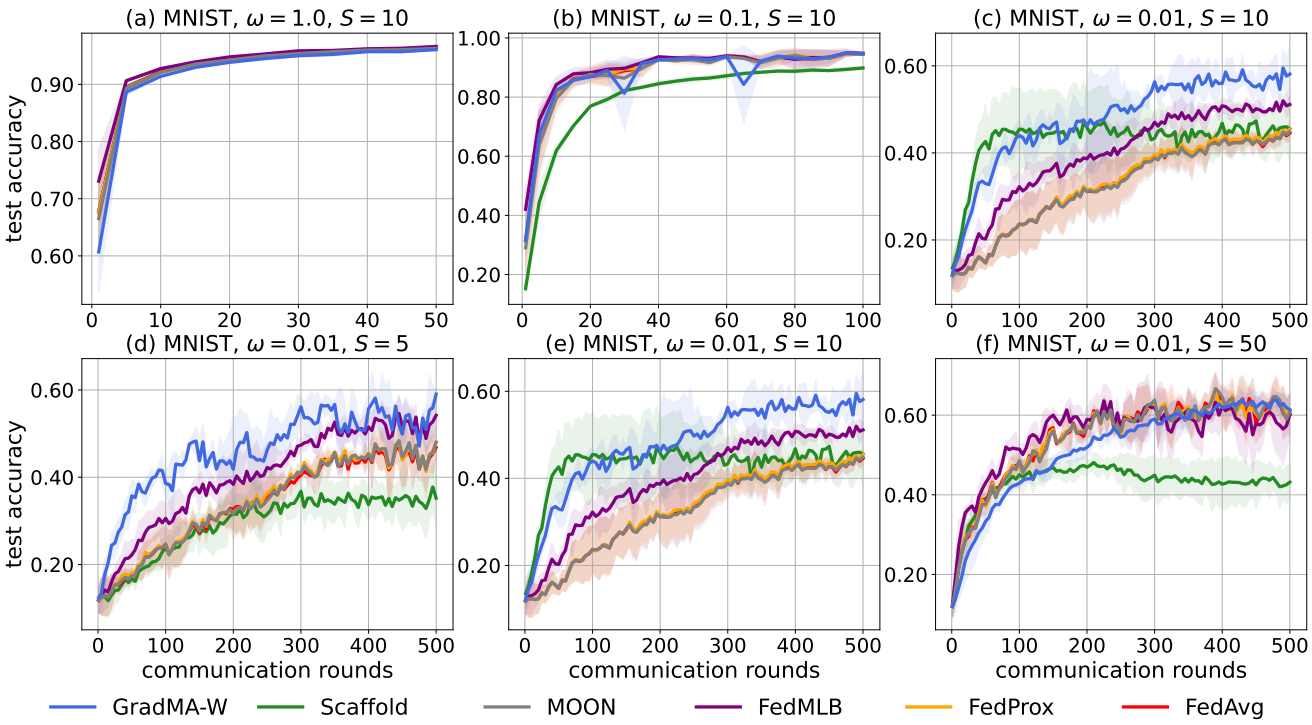

Figure 9. Full test accuracy curves for GradMA-W as well as baselines on MNIST.

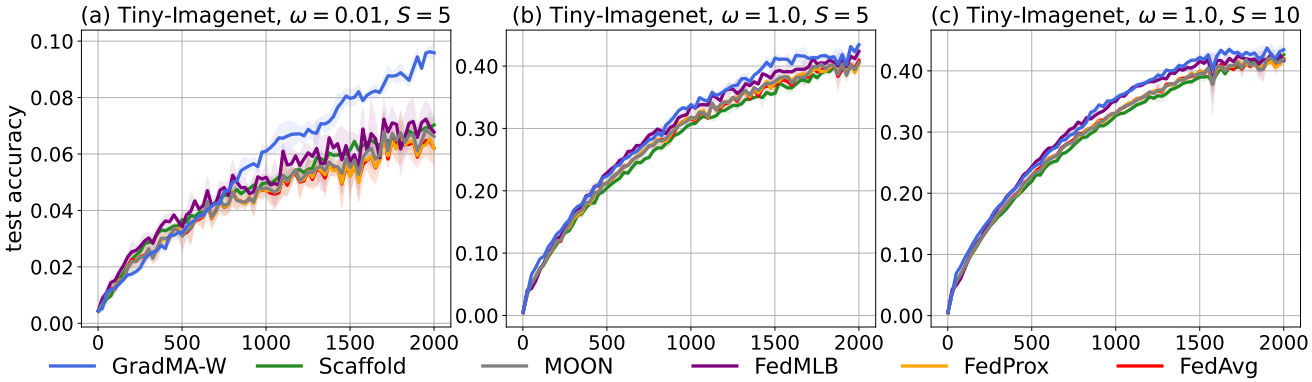

Figure 10. Full test accuracy curves for GradMA-W as well as baselines on Tiny-Imagenet.

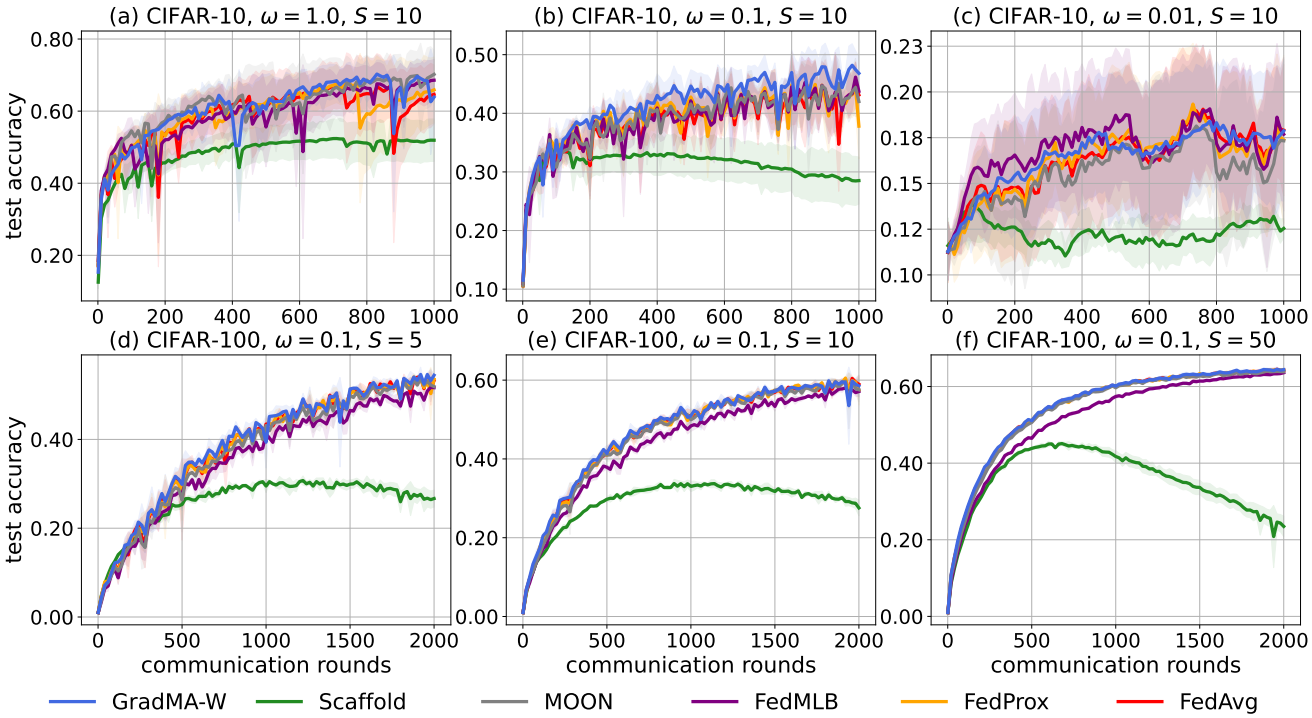

Figure 11. Full test accuracy curves for GradMA-W as well as baselines on CIFAR-10 and CIFAR-100.

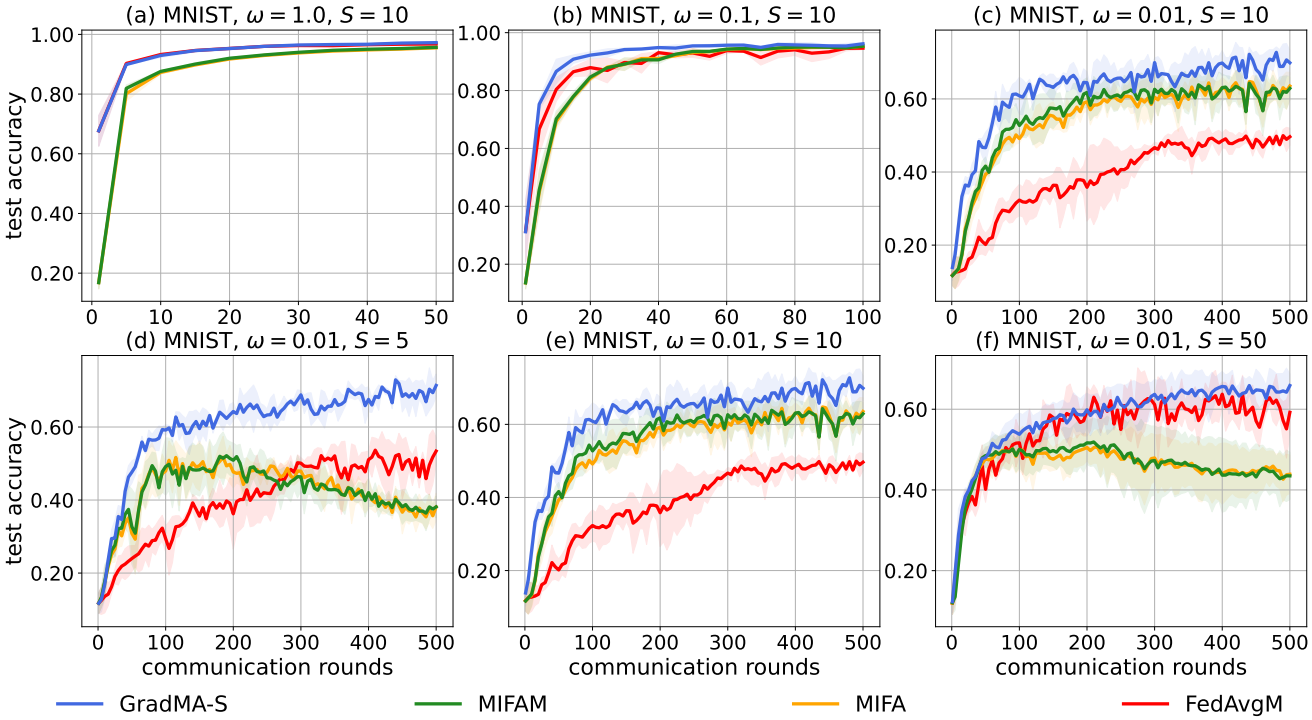

Figure 12. Full test accuracy curves for GradMA-S as well as baselines on MNIST.

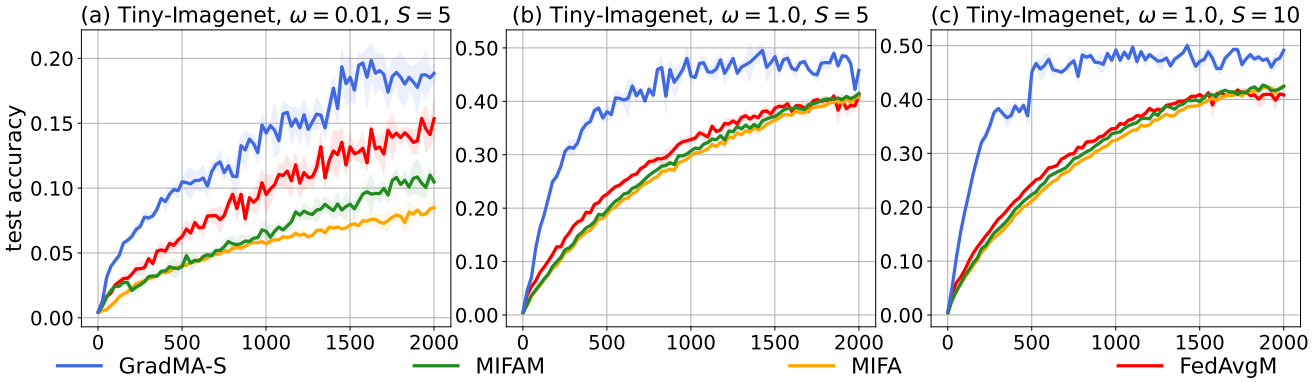

Figure 13. Full test accuracy curves for GradMA-S as well as baselines on Tiny-Imagenet.

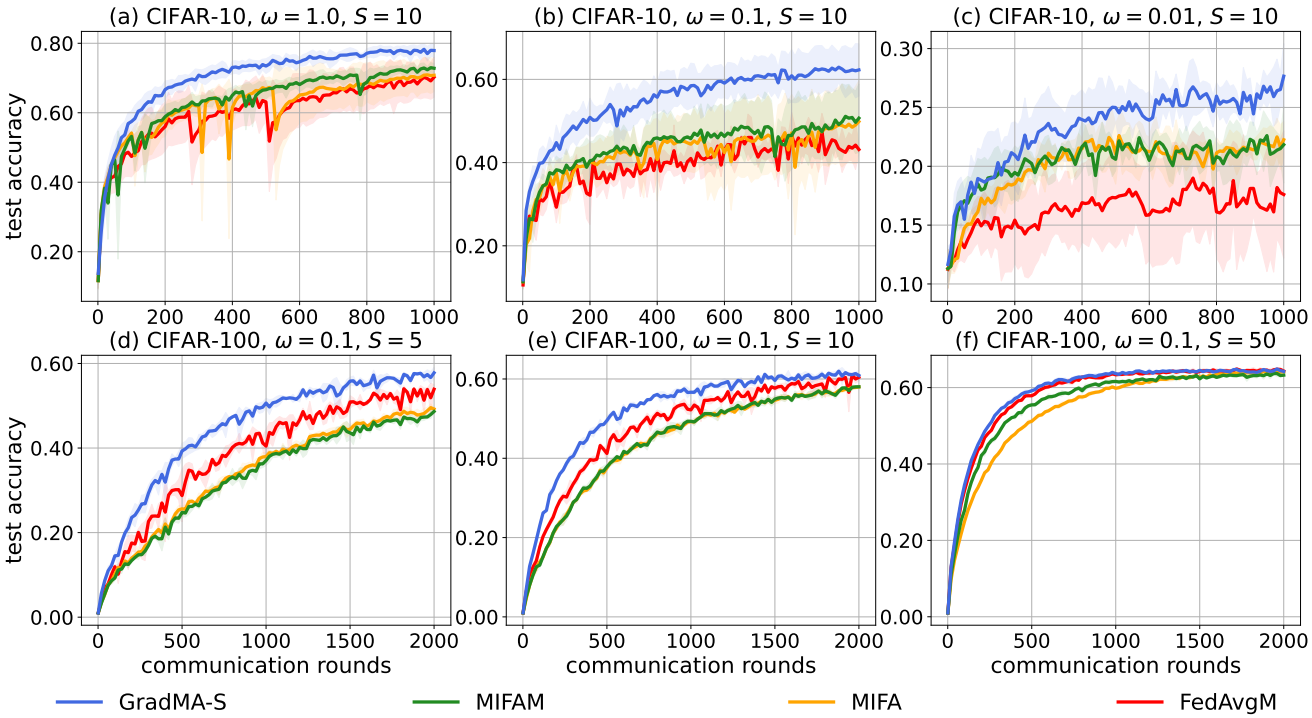

Figure 14. Full test accuracy curves for GradMA-S as well as baselines on CIFAR-10 and CIFAR-100.

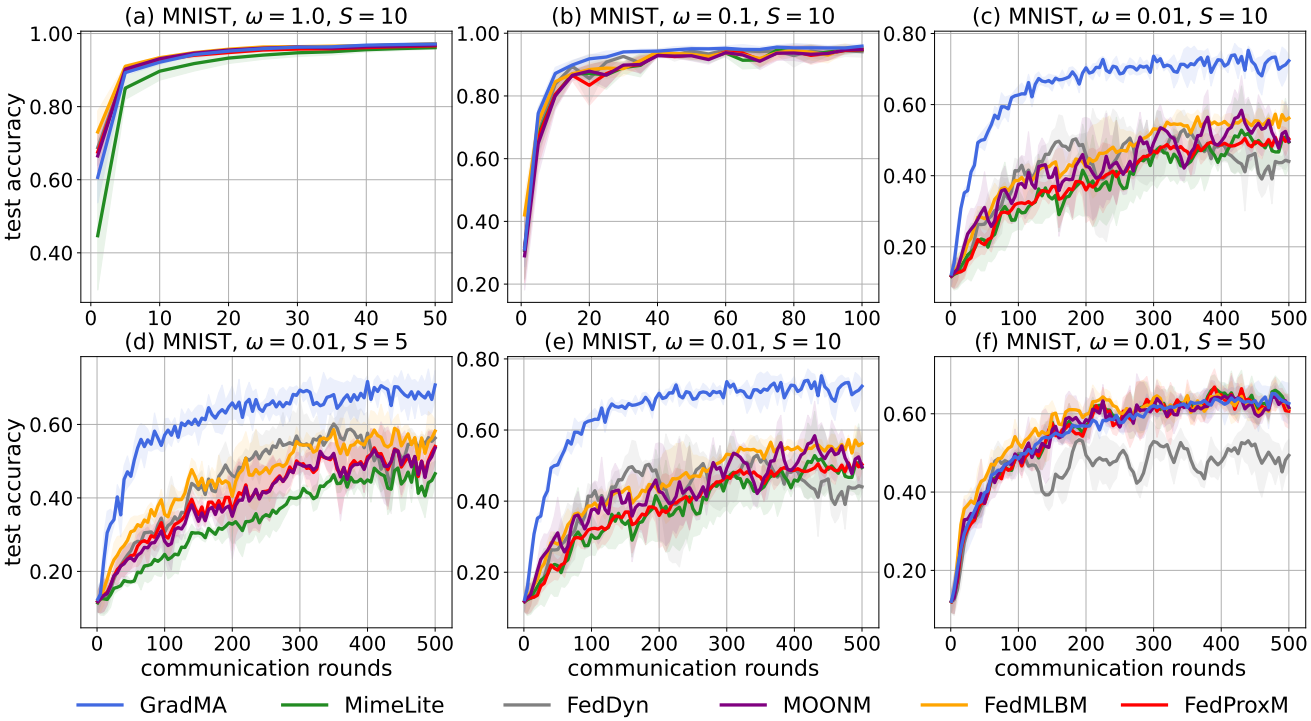

Figure 15. Full test accuracy curves for GradMA as well as baselines on MNIST.

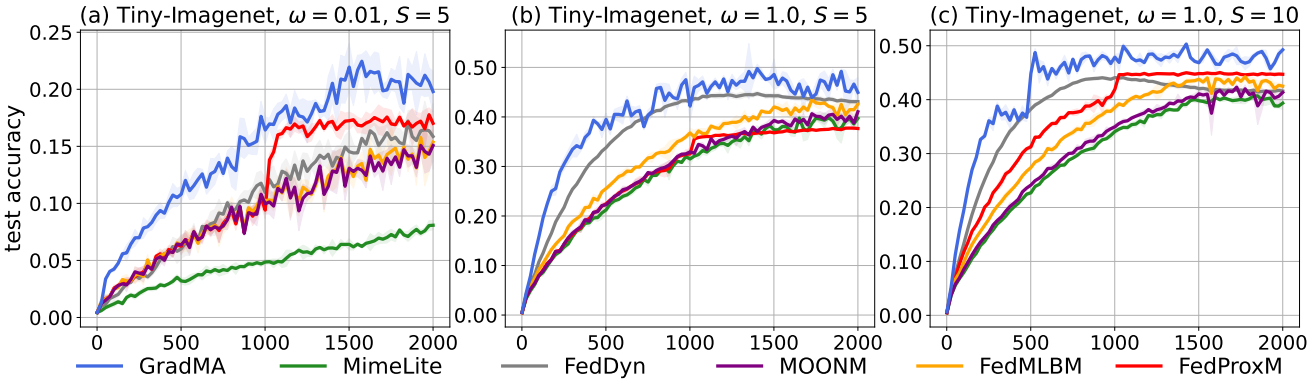

Figure 16. Full test accuracy curves for GradMA as well as baselines on Tiny-Imagenet.

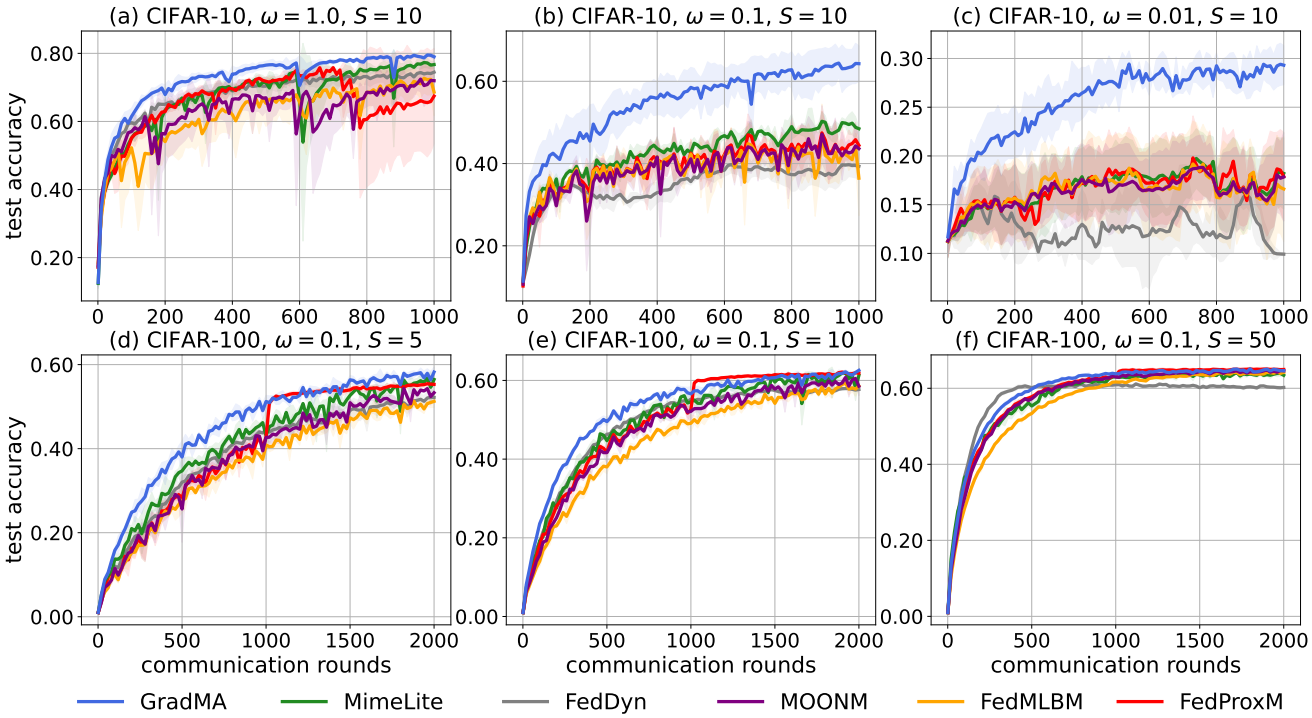

Figure 17. Full test accuracy curves for GradMA as well as baselines on CIFAR-10 and CIFAR-100.

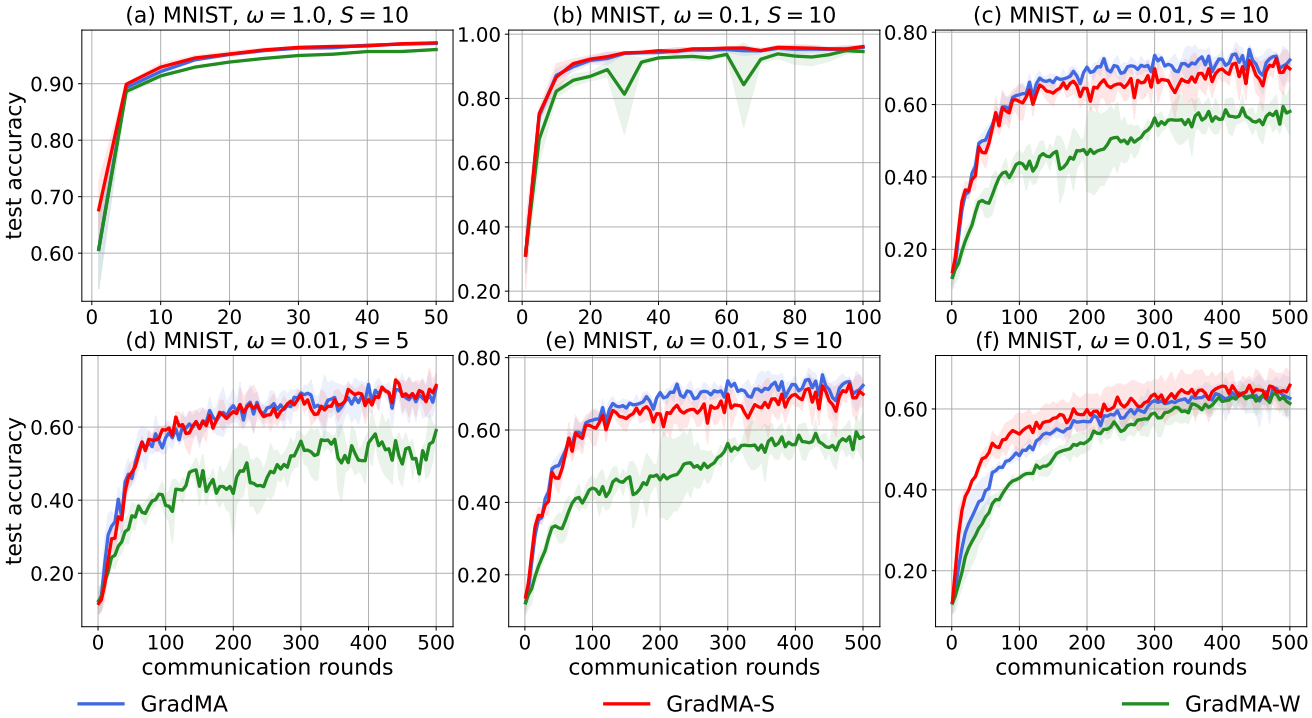

Figure 18. Full test accuracy curves for GradMA, GradMA-S and GradMA-W on MNIST.

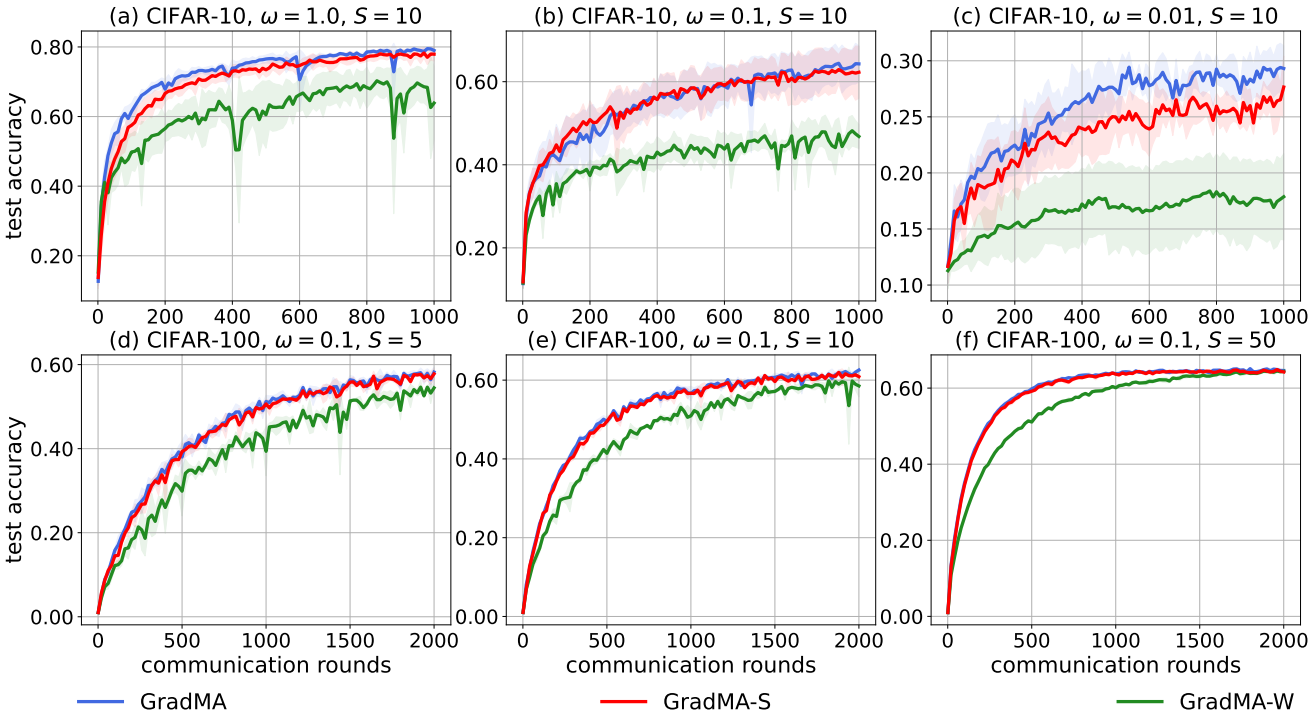

Figure 19. Full test accuracy curves for GradMA, GradMA-S and GradMA-W on CIFAR-10 and CIFAR-100.

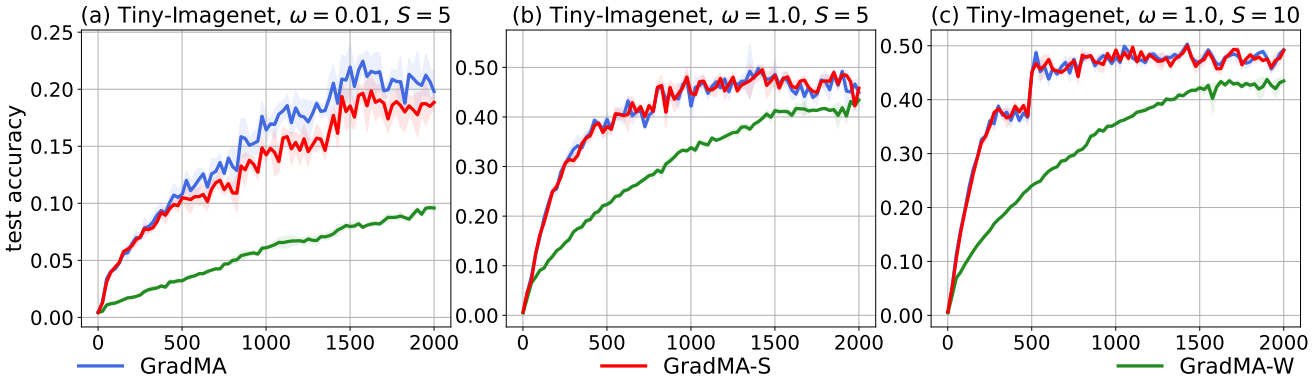

Figure 20. Full test accuracy curves for GradMA, GradMA-S and GradMA-W on Tiny-Imagenet.

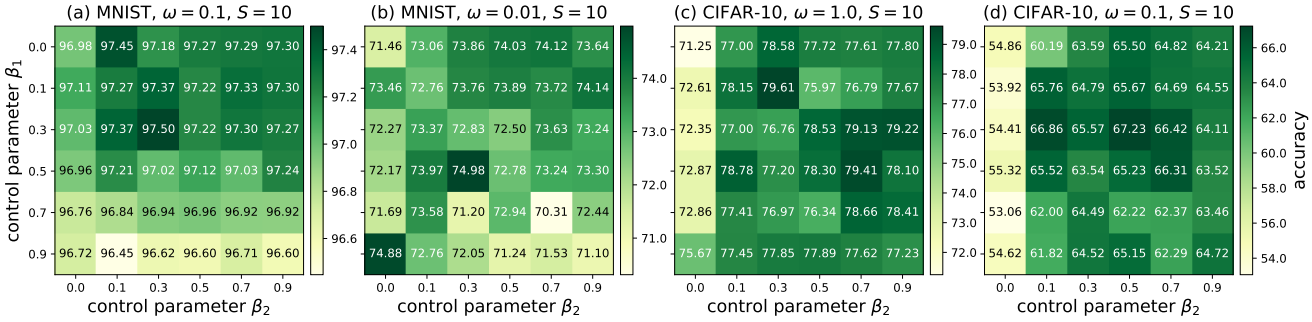

Figure 21. Top test accuracy (%) overview for varying control parameters (β_1,β_2) on MNIST and CIFAR-10.

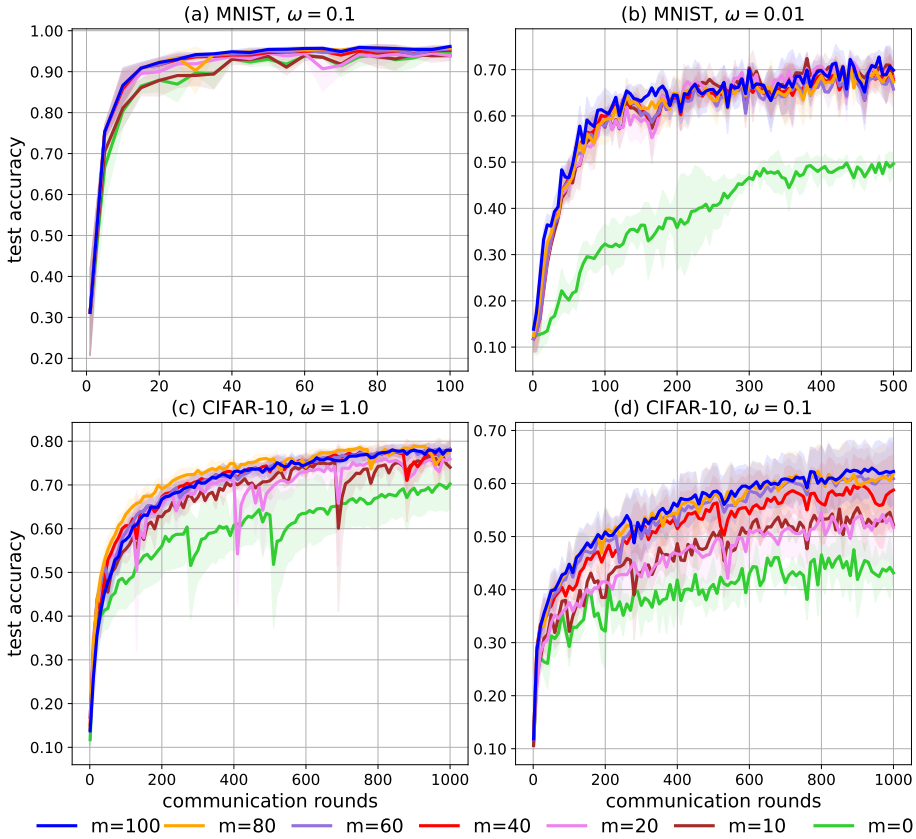

Figure 22. Full test accuracy curves for varying memory sizes m on MNIST and CIFAR-10.

.3. Convergence Proof of GradMA

In this section, we provide the complete theoretical proof for convergence result of GranMA.

We first review the rule for i-th ($i \in [N]$) worker to update local model in Alg. 1 and Alg. 2, as follows:

$$\boldsymbol{g}_{t,\tau}^{(i)} = \nabla f_i(\boldsymbol{x}_{t,\tau}^{(i)}), \tag{9}$$

$$\boldsymbol{G}_{t,\tau}^{(i)} = [\nabla f_i(\boldsymbol{x}_{t,\tau-1}^{(i)}), \nabla f_i(\boldsymbol{x}_t), \boldsymbol{x}_{t,\tau}^{(i)} - \boldsymbol{x}_t], \tag{10}$$

$$\tilde{\boldsymbol{g}}_{t,\tau}^{(i)} = \mathrm{QP}_l(\boldsymbol{g}_{t,\tau}^{(i)}, \boldsymbol{G}_{t,\tau}^{(i)}), \tag{11}$$

$$\boldsymbol{x}_{t,\tau+1}^{(i)} = \boldsymbol{x}_{t,\tau}^{(i)} - \eta_l \tilde{\boldsymbol{g}}_{t,\tau}^{(i)}, \tag{12}$$

where $\tau \in [0, \ldots, I-1]$ and $\boldsymbol{x}_{0,-1}^{(i)} = \boldsymbol{x}_0^{(i)} = \boldsymbol{x}_0$, $\boldsymbol{x}_{t,-1}^{(i)} = \boldsymbol{x}_t^{(i)}, \boldsymbol{x}_{t,0}^{(i)} = \boldsymbol{x}_t$ ($t > 0$).

After receiving update directions sent by active workers, the server updates the centralized model according to the following update rule (see Alg. 1 and Alg. 3):

$$\boldsymbol{d}_{t+1} = \frac{1}{S} \sum_{i \in \mathcal{S}_t} \boldsymbol{d}_{t+1}^{(i)} = \frac{\eta_l}{S} \sum_{i \in \mathcal{S}_t} \sum_{\tau=0}^{I-1} \tilde{\boldsymbol{g}}_{t,\tau}^{(i)}, \tag{13}$$

$$\boldsymbol{m}_{t+1} = \beta_1 \tilde{\boldsymbol{m}}_t + \boldsymbol{d}_{t+1}, \tag{14}$$

$$\tilde{\boldsymbol{m}}_{t+1} = \mathrm{QP}_g(\boldsymbol{m}_{t+1}, \boldsymbol{D}), \tag{15}$$

$$\boldsymbol{x}_{t+1} = \boldsymbol{x}_t - \eta_g \tilde{\boldsymbol{m}}_{t+1}, \tag{16}$$

where $t \in [0, \cdots, T-1]$ and $\tilde{\boldsymbol{m}}_0 = \boldsymbol{0}$. Here, we omit the update rule of $\boldsymbol{D}$ in that Assumption 5 holds as long as the information contained in $\boldsymbol{D}$ is meaningful, without needing to focus on the specific content of $\boldsymbol{D}$.

Furthermore, we set $\tilde{\boldsymbol{d}}_{t+1} = \boldsymbol{d}_{t+1} + \tilde{\boldsymbol{m}}_{t+1} - \boldsymbol{m}_{t+1}$ yields:

$$\tilde{\boldsymbol{m}}_{t+1} = \beta_1 \tilde{\boldsymbol{m}}_t + \tilde{\boldsymbol{d}}_{t+1}, \tag{17}$$

$$\boldsymbol{x}_{t+1} = \boldsymbol{x}_t - \eta_g \tilde{\boldsymbol{m}}_{t+1}. \tag{18}$$

Now, we define an auxiliary sequence such that

$$\boldsymbol{u}_t = \frac{1}{1-\beta_1}\boldsymbol{x}_t - \frac{\beta_1}{1-\beta_1}\boldsymbol{x}_{t-1}, \tag{19}$$

where $t > 0$. If $t = 0$ then $\boldsymbol{u}_t = \boldsymbol{x}_t$.

Lemma .1 *Define the sequence $\{\boldsymbol{u}_t\}_{t\geq 0}$ as in Eq. (19). According to Alg. 1, we have the following relationship*

$$\boldsymbol{u}_{t+1} - \boldsymbol{u}_t = -\frac{\eta_g}{1-\beta_1}\tilde{\boldsymbol{d}}_{t+1}.$$

Proof. Using mathematical induction on Eq. (19), we get:
case $t = 0$,

$$\begin{aligned}\boldsymbol{u}_{t+1} - \boldsymbol{u}_t &= \boldsymbol{u}_1 - \boldsymbol{u}_0 \\ &= \frac{1}{1-\beta_1}\boldsymbol{x}_1 - \frac{\beta_1}{1-\beta_1}\boldsymbol{x}_0 - \boldsymbol{x}_0 = \frac{1}{1-\beta_1}(\boldsymbol{x}_1 - \boldsymbol{x}_0) \\ &= -\frac{\eta_g}{1-\beta_1}\tilde{\boldsymbol{m}}_1 = -\frac{\eta_g}{1-\beta_1}\tilde{\boldsymbol{d}}_1,\end{aligned}$$

and case $t > 0$,

$$\begin{aligned}\boldsymbol{u}_{t+1} - \boldsymbol{u}_t &= \frac{1}{1-\beta_1}\boldsymbol{x}_{t+1} - \frac{\beta_1}{1-\beta_1}\boldsymbol{x}_t - \frac{1}{1-\beta_1}\boldsymbol{x}_t + \frac{\beta_1}{1-\beta_1}\boldsymbol{x}_{t-1} \\ &= \frac{1}{1-\beta_1}\left((\boldsymbol{x}_{t+1} - \boldsymbol{x}_t) - \beta_1(\boldsymbol{x}_t - \boldsymbol{x}_{t-1})\right) \\ &= -\frac{\eta_g}{1-\beta_1}\left(\boldsymbol{m}_{t+1} - \beta_1 \boldsymbol{m}_t\right) = -\frac{\eta_g}{1-\beta_1}\tilde{\boldsymbol{d}}_{t+1}.\end{aligned}$$

Hence, the lemma is proved.

End Proof.

Lemma .2 *Under Assumptions 2-4, then the following relationship generated according to Alg. 1 holds with $\eta_l \leq \frac{1}{4\sqrt{10}LI}$: for any $t \in [0, \cdots, T-1]$ and $\tau \in [0, \cdots, I-1]$,*

$$\frac{1}{N}\sum_{i\in[N]}\mathbb{E}\left[\left\|\boldsymbol{x}_{t,\tau}^{(i)} - \boldsymbol{x}_t\right\|^2\right] \leq 40I^2\eta_l^2\varepsilon_l^2 + 40I^2\eta_l^2\rho^2 + 40I^2\eta_l^2\mathbb{E}\left[\|\nabla f(\boldsymbol{x}_t)\|^2\right],$$

where the expectation $\mathbb{E}$ is w.r.t the sampled active workers per communication round.

Proof. For any worker $i \in [N]$ and $\tau \in [1, \cdots, I-1]$, we have:

$$\begin{aligned}&\mathbb{E}\left[\left\|\boldsymbol{x}_{t,\tau}^{(i)} - \boldsymbol{x}_t\right\|^2\right] \\ &\overset{(a)}{=} \mathbb{E}\left[\left\|\boldsymbol{x}_{t,\tau-1}^{(i)} - \boldsymbol{x}_t - \eta_l\tilde{\boldsymbol{g}}_{t,\tau-1}^{(i)}\right\|^2\right] \\ &= \mathbb{E}\left[\left\|\boldsymbol{x}_{t,\tau-1}^{(i)} - \boldsymbol{x}_t - \eta_l\left(\tilde{\boldsymbol{g}}_{t,\tau-1}^{(i)} - \boldsymbol{g}_{t,\tau-1}^{(i)} + \nabla f_i(\boldsymbol{x}_{t,\tau-1}^{(i)}) - \nabla f_i(\boldsymbol{x}_t) + \nabla f_i(\boldsymbol{x}_t) - \nabla f(\boldsymbol{x}_t) + \nabla f(\boldsymbol{x}_t)\right)\right\|^2\right] \\ &\overset{(b)}{\leq} \left(1 + \frac{1}{2I-1}\right)\mathbb{E}\left[\|\boldsymbol{x}_{t,\tau-1}^{(i)} - \boldsymbol{x}_t\|^2\right] + 8I\eta_l^2\left[\mathbb{E}\left[\|\tilde{\boldsymbol{g}}_{t,\tau-1}^{(i)} - \boldsymbol{g}_{t,\tau-1}^{(i)}\|^2\right] + \mathbb{E}\left[\|\nabla f_i(\boldsymbol{x}_{t,\tau-1}^{(i)}) - \nabla f_i(\boldsymbol{x}_t)\|^2\right]\right.\end{aligned}$$

$$
\begin{aligned}
&+\mathbb{E}\left[\|\nabla f_i(\boldsymbol{x}_t)-\nabla f(\boldsymbol{x}_t)\|^2\right]+\mathbb{E}\left[\|\nabla f(\boldsymbol{x}_t)\|^2\right]\Bigg]\\
&\leq\left(1+\frac{1}{2I-1}+8I\eta_l^2L^2\right)\mathbb{E}\left[\|\boldsymbol{x}_{t,\tau-1}^{(i)}-\boldsymbol{x}_t\|^2\right]+8I\eta_l^2\varepsilon_l^2+8I\eta_l^2\rho^2+8I\eta_l^2\mathbb{E}\left[\|\nabla f(\boldsymbol{x}_t)\|^2\right]\\
&\overset{(c)}{\leq}\left(1+\frac{1}{I-1}\right)\mathbb{E}\left[\|\boldsymbol{x}_{t,\tau-1}^{(i)}-\boldsymbol{x}_t\|^2\right]+8I\eta_l^2\varepsilon_l^2+8I\eta_l^2\rho^2+8I\eta_l^2\mathbb{E}\left[\|\nabla f(\boldsymbol{x}_t)\|^2\right],
\end{aligned}
\tag{20}
$$

where (a) holds by using the Eq. (12), (b) follows from the inequalities $\|\boldsymbol{x}\pm\boldsymbol{y}\|^2\leq(1+\frac{1}{2I-1})\|\boldsymbol{x}\|^2+2I\|\boldsymbol{y}\|^2,\boldsymbol{x},\boldsymbol{y}\in\mathbb{R}^d$ and $\left\|\sum_{i=1}^N\boldsymbol{x}_i\right\|^2\leq N\sum_{i=1}^N\|\boldsymbol{x}_i\|^2,\boldsymbol{x}_i\in\mathbb{R}^d$, and (c) holds by using the fact that $\frac{1}{I-1}\geq\frac{1}{2I-1}+8I\eta_l^2L^2$ holds if $\eta_l\leq\frac{1}{4\sqrt{10}LI}$.

Then, recursively unrolling inequality (20), we get:

$$
\begin{aligned}
\frac{1}{N}\sum_{i\in[N]}\mathbb{E}\left[\left\|\boldsymbol{x}_{t,\tau}^{(i)}-\boldsymbol{x}_t\right\|^2\right]&\leq\sum_{k=0}^{\tau-1}\left(1+\frac{1}{I-1}\right)^k\left[8I\eta_l^2\varepsilon_l^2+8I\eta_l^2\rho^2+8I\eta_l^2\mathbb{E}\left[\|\nabla f(\boldsymbol{x}_t)\|^2\right]\right]\\
&\leq(I-1)\left[\left(1+\frac{1}{I-1}\right)^I-1\right]\left[8I\eta_l^2\varepsilon_l^2+8I\eta_l^2\rho^2+8I\eta_l^2\mathbb{E}\left[\|\nabla f(\boldsymbol{x}_t)\|^2\right]\right]\\
&\overset{(a)}{\leq}5I\left[8I\eta_l^2\varepsilon_l^2+8I\eta_l^2\rho^2+8I\eta_l^2\mathbb{E}\left[\|\nabla f(\boldsymbol{x}_t)\|^2\right]\right]\\
&\leq 40I^2\eta_l^2\varepsilon_l^2+40I^2\eta_l^2\rho^2+40I^2\eta_l^2\mathbb{E}\left[\|\nabla f(\boldsymbol{x}_t)\|^2\right],
\end{aligned}
$$

where (a) holds by using $\sum_{k=0}^{\tau-1}\left(1+\frac{1}{I-1}\right)^k=\frac{1-\left(1+\frac{1}{I-1}\right)^\tau}{1-\left(1+\frac{1}{I-1}\right)}=(I-1)\left(\left(1+\frac{1}{I-1}\right)^\tau-1\right)\leq(I-1)\left(\left(1+\frac{1}{I-1}\right)^I-1\right)\leq 5I$.

So far, we complete the proof.

End Proof.

Lemma .3 *Under Assumptions 2-4, then the following relationship generated according to Alg. 1 holds: for any* $t\in[0,\cdots,T-1]$,

$$
\frac{1}{N}\sum_{i\in[N]}\mathbb{E}\left[\left\|\sum_{\tau=0}^{I-1}\tilde{\boldsymbol{g}}_{t,\tau}^{(i)}\right\|^2\right]\leq 4I^2(\varepsilon_l^2+\rho^2)(1+40I^2\eta_l^2L^2)+4I^2(1+40I^2\eta_l^2L^2)\mathbb{E}\left[\|\nabla f(\boldsymbol{x}_t)\|^2\right],
$$

where the expectation $\mathbb{E}$ *is w.r.t the sampled active workers per communication round.*

Proof.

$$
\begin{aligned}
&\frac{1}{N}\sum_{i\in[N]}\mathbb{E}\left[\left\|\sum_{\tau=0}^{I-1}\tilde{\boldsymbol{g}}_{t,\tau}^{(i)}\right\|^2\right]\\
&\overset{(a)}{\leq}I\frac{1}{N}\sum_{i\in[N]}\sum_{\tau=0}^{I-1}\mathbb{E}\left[\left\|\tilde{\boldsymbol{g}}_{t,\tau}^{(i)}\right\|^2\right]\\
&\leq I\frac{1}{N}\sum_{i\in[N]}\sum_{\tau=0}^{I-1}\mathbb{E}\left[\left\|\tilde{\boldsymbol{g}}_{t,\tau}^{(i)}-\boldsymbol{g}_{t,\tau}^{(i)}+\nabla f_i(\boldsymbol{x}_{t,\tau}^{(i)})-\nabla f_i(\boldsymbol{x}_t)+\nabla f_i(\boldsymbol{x}_t)-\nabla f(\boldsymbol{x}_t)+\nabla f(\boldsymbol{x}_t)\right\|^2\right]\\
&\overset{(b)}{\leq}4I\frac{1}{N}\sum_{i\in[N]}\sum_{\tau=0}^{I-1}\left[\left\|\tilde{\boldsymbol{g}}_{t,\tau}^{(i)}-\boldsymbol{g}_{t,\tau}^{(i)}\right\|^2+\mathbb{E}\left[\left\|\nabla f_i(\boldsymbol{x}_{t,\tau}^{(i)})-\nabla f_i(\boldsymbol{x}_t)\right\|^2\right]+\|\nabla f_i(\boldsymbol{x}_t)-\nabla f(\boldsymbol{x}_t)\|^2+\mathbb{E}\left[\|\nabla f(\boldsymbol{x}_t)\|^2\right]\right]\\
&\leq 4I\frac{1}{N}\sum_{i\in[N]}\sum_{\tau=0}^{I-1}\left[\varepsilon_l^2+L^2\mathbb{E}\left[\left\|\boldsymbol{x}_{t,\tau}^{(i)}-\boldsymbol{x}_t\right\|^2\right]+\rho^2+\mathbb{E}\left[\|\nabla f(\boldsymbol{x}_t)\|^2\right]\right]
\end{aligned}
$$

$$
\begin{aligned}
&\leq 4I^2\left(\varepsilon_l^2+\rho^2+\mathbb{E}\left[\left\|\nabla f(\boldsymbol{x}_t)\right\|^2\right]\right)+4IL^2\frac{1}{N}\sum_{i\in[N]}\sum_{\tau=0}^{I-1}\mathbb{E}\left[\left\|\boldsymbol{x}_{t,\tau}^{(i)}-\boldsymbol{x}_t\right\|^2\right]\\
&\overset{(c)}{\leq} 4I^2\left(\varepsilon_l^2+\rho^2+\mathbb{E}\left[\left\|\nabla f(\boldsymbol{x}_t)\right\|^2\right]\right)+4I^2L^2\left(40I^2\eta_l^2\varepsilon_l^2+40I^2\eta_l^2\rho^2+40I^2\eta_l^2\mathbb{E}\left[\left\|\nabla f(\boldsymbol{x}_t)\right\|^2\right]\right)\\
&\leq 4I^2(\varepsilon_l^2+\rho^2)(1+40I^2\eta_l^2L^2)+4I^2(1+40I^2\eta_l^2L^2)\mathbb{E}\left[\left\|\nabla f(\boldsymbol{x}_t)\right\|^2\right]
\end{aligned}
$$

where (a) and (b) result from the fact that $\left\|\sum_{i=1}^N \boldsymbol{x}_i\right\|^2 \leq N\sum_{i=1}^N\|\boldsymbol{x}_i\|^2, \boldsymbol{x}_i\in\mathbb{R}^d$, and (c) uses the statement from Lemma .2.

So far, the lemma is proved.

End Proof.

Lemma .4 *Under Assumption 5, then the following relationship generated according to Alg. 1 holds: for any* $t\in[0,\cdots,T-1]$,

$$
\mathbb{E}\left[\left\|\tilde{\boldsymbol{d}}_{t+1}\right\|^2\right]\leq\frac{2\eta_l^2}{S^2}\mathbb{E}\left[\left\|\sum_{i\in[N]}\mathbb{I}\{i\in\mathcal{S}_t\}\sum_{\tau=0}^{I-1}\tilde{\boldsymbol{g}}_{t,\tau}^{(i)}\right\|^2\right]+\frac{2\varepsilon_g^2}{1-\beta_2},
$$

where the expectation $\mathbb{E}$ *is w.r.t the sampled active workers per communication round.*

Proof.

$$
\begin{aligned}
\mathbb{E}\left[\left\|\tilde{\boldsymbol{d}}_{t+1}\right\|^2\right]&\overset{(a)}{\leq}\mathbb{E}\left[\left\|\boldsymbol{d}_{t+1}+\tilde{\boldsymbol{m}}_{t+1}-\boldsymbol{m}_{t+1}\right\|^2\right]\\
&\leq 2\mathbb{E}\left[\left\|\boldsymbol{d}_{t+1}\right\|^2\right]+2\mathbb{E}\left[\left\|\tilde{\boldsymbol{m}}_{t+1}-\boldsymbol{m}_{t+1}\right\|^2\right]\\
&\overset{(b)}{\leq} 2\mathbb{E}\left[\left\|\frac{1}{S}\sum_{i\in\mathcal{S}_t}\boldsymbol{d}_{t+1}^{(i)}\right\|^2\right]+\frac{2\varepsilon_g^2}{1-\beta_2}\\
&=\frac{2\eta_l^2}{S^2}\mathbb{E}\left[\left\|\sum_{i\in[N]}\mathbb{I}\{i\in\mathcal{S}_t\}\sum_{\tau=0}^{I-1}\tilde{\boldsymbol{g}}_{t,\tau}^{(i)}\right\|^2\right]+\frac{2\varepsilon_g^2}{1-\beta_2},
\end{aligned}
$$

where (a) uses the fact that $\tilde{\boldsymbol{d}}_{t+1}=\boldsymbol{d}_{t+1}+\tilde{\boldsymbol{m}}_{t+1}-\boldsymbol{m}_{t+1}$, and (b) results from the Eq. (13). Hence, the lemma is proved.

End Proof.

Lemma .5 *Define the sequence* $\{\boldsymbol{u}_t\}_{t\geq0}$ *as in Eq. (19), the following relationship generated according to Alg. 1 holds: for any* $t\in[0,\cdots,T-1]$,

$$
\sum_{t=0}^{T-1}\mathbb{E}\left[\left\|\boldsymbol{u}_t-\boldsymbol{x}_t\right\|^2\right]\leq\frac{2\beta_1^2\eta_g^2\eta_l^2}{(1-\beta_1)^4S^2}\sum_{t=0}^{T-1}\mathbb{E}\left[\left\|\sum_{i\in[N]}\mathbb{I}\{i\in\mathcal{S}_t\}\sum_{\tau=0}^{I-1}\tilde{\boldsymbol{g}}_{t,\tau}^{(i)}\right\|^2\right]+\frac{2\beta_1^2\eta_g^2\varepsilon_g^2T}{(1-\beta_1)^4(1-\beta_2)},
$$

where the expectation $\mathbb{E}$ *is w.r.t the sampled active workers per communication round.*

Proof. Recursively applying Eq. (17) to achieve the update rule for $\tilde{\boldsymbol{m}}_t$ yields:

$$
\tilde{\boldsymbol{m}}_t\overset{(a)}{=}\sum_{k=1}^{t}\beta_1^{t-k}\tilde{\boldsymbol{d}}_k,\forall t\geq1, \tag{21}
$$

where (a) holds by $\tilde{\boldsymbol{m}}_0 = \mathbf{0}$. Furthermore, building on equations (19) and (21), we get:

$$\boldsymbol{u}_t - \boldsymbol{x}_t = \frac{\beta_1}{1-\beta_1}(\boldsymbol{x}_t - \boldsymbol{x}_{t-1}) = -\frac{\beta_1\eta_g}{1-\beta_1}\tilde{\boldsymbol{m}}_t = -\frac{\beta_1\eta_g}{1-\beta_1}\sum_{k=1}^{t}\beta_1^{t-k}\tilde{\boldsymbol{d}}_k. \tag{22}$$

Now, we define $z_t = \sum_{k=1}^{t}\beta_1^{t-k} = \frac{1-\beta_1^t}{1-\beta_1}, \forall t \geq 1$. Using Eq. (22) we obtain:

$$\begin{aligned}
\mathbb{E}\left[\|\boldsymbol{u}_t - \boldsymbol{x}_t\|^2\right] &= \frac{\beta_1^2\eta_g^2}{(1-\beta_1)^2}z_t^2\mathbb{E}\left[\left\|\sum_{k=1}^{t}\frac{\beta_1^{t-k}}{z_t}\tilde{\boldsymbol{d}}_k\right\|^2\right] \\
&\overset{(a)}{\leq} \frac{\beta_1^2\eta_g^2}{(1-\beta_1)^2}z_t\sum_{k=1}^{t}\beta_1^{t-k}\mathbb{E}\left[\left\|\tilde{\boldsymbol{d}}_k\right\|^2\right] \\
&\overset{(b)}{\leq} \frac{\beta_1^2\eta_g^2}{(1-\beta_1)^3}\sum_{k=1}^{t}\beta_1^{t-k}\mathbb{E}\left[\left\|\tilde{\boldsymbol{d}}_k\right\|^2\right],
\end{aligned} \tag{23}$$

where (a) follows from the fact that $\|\sum_{k=1}^{T}\frac{c_k}{c}\boldsymbol{a}_k\|^2 \leq \sum_{k=1}^{T}\frac{c_i}{c}\|\boldsymbol{a}_k\|^2 (\boldsymbol{a}_k \in \mathbb{R}^d)$ holds if $c = \sum_{k=1}^{T}c_k$, and (b) results from $1-\beta^t \leq 1$.

Next, summing Eq. (23) over $t \in \{0, \cdots, T-1\} (T \geq 1)$, we have:

$$\begin{aligned}
\sum_{t=0}^{T-1}\mathbb{E}\left[\|\boldsymbol{u}_t - \boldsymbol{x}_t\|^2\right] &\overset{(a)}{\leq} \frac{\beta_1^2\eta_g^2}{(1-\beta_1)^3}\sum_{t=1}^{T-1}\sum_{k=1}^{t}\beta_1^{t-k}\mathbb{E}\left[\left\|\tilde{\boldsymbol{d}}_k\right\|^2\right] \\
&= \frac{\beta_1^2\eta_g^2}{(1-\beta_1)^3}\left(\sum_{t=k}^{T-1}\beta_1^{t-k}\right)\sum_{k=1}^{T-1}\mathbb{E}\left[\left\|\tilde{\boldsymbol{d}}_k\right\|^2\right] \\
&\overset{(b)}{\leq} \frac{\beta_1^2\eta_g^2}{(1-\beta_1)^4}\sum_{t=1}^{T-1}\mathbb{E}\left[\left\|\tilde{\boldsymbol{d}}_t\right\|^2\right] \\
&\overset{(c)}{\leq} \frac{2\beta_1^2\eta_g^2\eta_l^2}{(1-\beta_1)^4S^2}\sum_{t=0}^{T-1}\mathbb{E}\left[\left\|\sum_{i\in[N]}\mathbb{I}\{i\in\mathcal{S}_t\}\sum_{\tau=0}^{I-1}\tilde{\boldsymbol{g}}_{t,\tau}^{(i)}\right\|^2\right] + \frac{2\beta_1^2\eta_g^2\varepsilon_g^2 T}{(1-\beta_1)^4(1-\beta_2)},
\end{aligned} \tag{24}$$

where (a) uses $\boldsymbol{u}_0 = \boldsymbol{x}_0$, (b) holds by using the inequality $\sum_{t=k}^{T-1}\beta_1^{t-k} = \frac{1-\beta_1^t}{1-\beta_1} \leq \frac{1}{1-\beta_1}$, and (c) follows from the statement of Lemma .4.

Hence, the lemma is proved.

End Proof.

Lemma .6 *According to Assumptions 1-5 and setting* $\eta_l \leq \frac{1}{4\sqrt{10}LI}$, $\eta_g\eta_l \leq \frac{(1-\beta_1)^2S(N-1)}{IL(\beta_1 S(N-1)+4N(S-1))}$ *and* $320I^2\eta_l^2L^2 + \frac{64I\eta_g\eta_l L(1+40I^2\eta_l^2L^2)}{(1-\beta_1)^2}\frac{N-S}{S(N-1)} \leq 1$, *then the iterates generated by Alg. 1 satisfy: for all* $t \in [0, \cdots, T-1]$,

$$\begin{aligned}
\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}\left[\|\nabla f(\boldsymbol{x}_t)\|^2\right] \leq{}& \frac{8(1-\beta_1)(f(\boldsymbol{x}_0)-f^\star)}{I\eta_g\eta_l T} + 8\varepsilon_l^2 + 320I^2\eta_l^2L^2\varepsilon_l^2 + \frac{64I\eta_g\eta_l L\varepsilon_l^2(1+40I^2\eta_l^2L^2)}{(1-\beta_1)^2}\frac{N-S}{S(N-1)} \\
&+ \frac{20\eta_g L\varepsilon_g^2}{(1-\beta_1)^2(1-\beta_2)I\eta_l} + \frac{8\varepsilon_g^2}{(1-\beta_2)I^2\eta_l^2} + 320I^2\eta_l^2L^2\rho^2 + \frac{64I\eta_g\eta_l L\rho^2(1+40I^2\eta_l^2L^2)}{(1-\beta_1)^2}\frac{N-S}{S(N-1)},
\end{aligned}$$

where the expectation $\mathbb{E}$ *is w.r.t the sampled active workers per communication round.*

Proof. Based on L-smooth of f and expectation w.r.t. the sampled active workers per communication round, we have:

$$\mathbb{E}[f(\boldsymbol{u}_{t+1})] \leq \mathbb{E}[f(\boldsymbol{u}_t)] + \mathbb{E}\left[\langle\nabla f(\boldsymbol{u}_t), \boldsymbol{u}_{t+1} - \boldsymbol{u}_t\rangle\right] + \frac{L}{2}\mathbb{E}\left[\|\boldsymbol{u}_{t+1} - \boldsymbol{u}_t\|^2\right]$$

$$
\begin{aligned}
&\overset{(a)}{=} \mathbb{E}[f(\boldsymbol{u}_t)] - \frac{\eta_g}{1-\beta_1}\mathbb{E}\left[\left\langle \nabla f(\boldsymbol{u}_t), \tilde{\boldsymbol{d}}_{t+1}\right\rangle\right] + \frac{L\eta_g^2}{2(1-\beta_1)^2}\mathbb{E}\left[\left\|\tilde{\boldsymbol{d}}_{t+1}\right\|^2\right] \\
&= \mathbb{E}[f(\boldsymbol{u}_t)] \underbrace{-\frac{\eta_g}{1-\beta_1}\mathbb{E}\left[\left\langle \nabla f(\boldsymbol{u}_t) - \nabla f(\boldsymbol{x}_t), \tilde{\boldsymbol{d}}_{t+1}\right\rangle\right]}_{T_1} \underbrace{-\frac{\eta_g}{1-\beta_1}\mathbb{E}\left[\left\langle \nabla f(\boldsymbol{x}_t), \tilde{\boldsymbol{d}}_{t+1}\right\rangle\right]}_{T_2} + \underbrace{\frac{L\eta_g^2}{2(1-\beta_1)^2}\mathbb{E}\left[\left\|\tilde{\boldsymbol{d}}_{t+1}\right\|^2\right]}_{T_3},
\end{aligned} \tag{25}
$$

where (a) holds because of the statement of Lemma .1.

Firstly, we note that

$$
\begin{aligned}
T_1 &= -\frac{\eta_g}{1-\beta_1}\mathbb{E}\left[\left\langle \nabla f(\boldsymbol{u}_t) - \nabla f(\boldsymbol{x}_t), \tilde{\boldsymbol{d}}_{t+1}\right\rangle\right] \\
&= -\frac{\eta_g}{1-\beta_1}\mathbb{E}\left[\left\langle \nabla f(\boldsymbol{u}_t) - \nabla f(\boldsymbol{x}_t), \frac{1}{S}\sum_{i\in\mathcal{S}_t}\boldsymbol{d}_{t+1}^{(i)}\right\rangle\right] - \frac{\eta_g}{1-\beta_1}\mathbb{E}\left[\langle \nabla f(\boldsymbol{u}_t) - \nabla f(\boldsymbol{x}_t), \tilde{\boldsymbol{m}}_{t+1} - \boldsymbol{m}_{t+1}\rangle\right] \\
&= \underbrace{-\frac{\eta_g}{1-\beta_1}\mathbb{E}\left[\left\langle \nabla f(\boldsymbol{u}_t) - \nabla f(\boldsymbol{x}_t), \frac{1}{N}\sum_{i\in[N]}\boldsymbol{d}_{t+1}^{(i)}\right\rangle\right]}_{T_{1,1}} \underbrace{-\frac{\eta_g}{1-\beta_1}\mathbb{E}\left[\langle \nabla f(\boldsymbol{u}_t) - \nabla f(\boldsymbol{x}_t), \tilde{\boldsymbol{m}}_{t+1} - \boldsymbol{m}_{t+1}\rangle\right]}_{T_{1,2}}.
\end{aligned}
$$

We proceed by analysis $T_{1,1}$,

$$
\begin{aligned}
T_{1,1} &= -\frac{\eta_g}{1-\beta_1}\mathbb{E}\left[\left\langle \nabla f(\boldsymbol{u}_t) - \nabla f(\boldsymbol{x}_t), \frac{1}{N}\sum_{i\in[N]}\boldsymbol{d}_{t+1}^{(i)}\right\rangle\right] \\
&\overset{(a)}{=} -\frac{\eta_g}{1-\beta_1}\mathbb{E}\left[\left\langle \nabla f(\boldsymbol{u}_t) - \nabla f(\boldsymbol{x}_t), \frac{\eta_l}{N}\sum_{i\in[N]}\sum_{\tau=0}^{I-1}\tilde{\boldsymbol{g}}_{t,\tau}^{(i)}\right\rangle\right] \\
&\overset{(b)}{\leq} \frac{1-\beta_1}{2\beta_1 L}\mathbb{E}\left[\|\nabla f(\boldsymbol{u}_t) - \nabla f(\boldsymbol{x}_t)\|^2\right] + \frac{\beta_1 L\eta_g^2\eta_l^2}{2(1-\beta_1)^3}\mathbb{E}\left[\left\|\frac{1}{N}\sum_{i\in[N]}\sum_{\tau=0}^{I-1}\tilde{\boldsymbol{g}}_{t,\tau}^{(i)}\right\|^2\right] \\
&\leq \frac{(1-\beta_1)L}{2\beta_1}\mathbb{E}\left[\|\boldsymbol{u}_t - \boldsymbol{x}_t\|^2\right] + \frac{\beta_1 L\eta_g^2\eta_l^2}{2(1-\beta_1)^3N^2}\mathbb{E}\left[\left\|\sum_{i\in[N]}\sum_{\tau=0}^{I-1}\tilde{\boldsymbol{g}}_{t,\tau}^{(i)}\right\|^2\right],
\end{aligned}
$$

where (a) follows by using Eq. (13), and (b) holds because of the fact that $\pm\langle \boldsymbol{a}, \boldsymbol{b}\rangle \leq \frac{1}{2}\|\boldsymbol{a}\|^2 + \frac{1}{2}\|\boldsymbol{a}\|^2$ $(\boldsymbol{a}, \boldsymbol{b} \in \mathbb{R}^d)$ where $\boldsymbol{a} = -\frac{\sqrt{1-\beta_1}}{\sqrt{\beta_1 L}}(\nabla f(\boldsymbol{u}_t) - \nabla f(\boldsymbol{x}_t))$ and $\boldsymbol{b} = \frac{\sqrt{\beta_1 L}\eta_g\eta_l}{(1-\beta_1)^{3/2}}\frac{1}{N}\sum_{i\in[N]}\sum_{\tau=0}^{I-1}\tilde{\boldsymbol{g}}_{t,\tau}^{(i)}$. And we proceed by analysis $T_{1,2}$,

$$
\begin{aligned}
T_{1,2} &= -\frac{\eta_g}{1-\beta_1}\mathbb{E}\left[\langle \nabla f(\boldsymbol{u}_t) - \nabla f(\boldsymbol{x}_t), \tilde{\boldsymbol{m}}_{t+1} - \boldsymbol{m}_{t+1}\rangle\right] \\
&\overset{(a)}{\leq} \frac{1-\beta_1}{2\beta_1 L}\mathbb{E}\left[\|\nabla f(\boldsymbol{u}_t) - \nabla f(\boldsymbol{x}_t)\|^2\right] + \frac{\beta_1 L\eta_g^2}{2(1-\beta_1)^3}\mathbb{E}\left[\|\tilde{\boldsymbol{m}}_{t+1} - \boldsymbol{m}_{t+1}\|^2\right] \\
&\leq \frac{(1-\beta_1)L}{2\beta_1}\mathbb{E}\left[\|\boldsymbol{u}_t - \boldsymbol{x}_t\|^2\right] + \frac{\beta_1 L\eta_g^2\varepsilon_g^2}{2(1-\beta_1)^3(1-\beta_2)},
\end{aligned}
$$

where (a) results from the fact that $\pm\langle \boldsymbol{a}, \boldsymbol{b}\rangle \leq \frac{1}{2}\|\boldsymbol{a}\|^2 + \frac{1}{2}\|\boldsymbol{b}\|^2$ $(\boldsymbol{a}, \boldsymbol{b} \in \mathbb{R}^d)$ where $\boldsymbol{a} = -\frac{\sqrt{1-\beta_1}}{\sqrt{\beta_1 L}}(\nabla f(\boldsymbol{u}_t) - \nabla f(\boldsymbol{x}_t))$ and $\boldsymbol{b} = \frac{\sqrt{\beta_1 L}\eta_g}{(1-\beta_1)^{3/2}}(\tilde{\boldsymbol{m}}_{t+1} - \boldsymbol{m}_{t+1})$.

Secondly, we observe that

$$
T_2 = -\frac{\eta_g}{1-\beta_1}\mathbb{E}\left[\left\langle \nabla f(\boldsymbol{x}_t), \tilde{\boldsymbol{d}}_{t+1}\right\rangle\right]
$$

$$= \underbrace{-\frac{\eta_g}{1-\beta_1}\mathbb{E}\left[\langle \nabla f(\boldsymbol{x}_t), \boldsymbol{d}_{t+1}\rangle\right]}_{T_{2,1}} \underbrace{-\frac{\eta_g}{1-\beta_1}\mathbb{E}\left[\langle \nabla f(\boldsymbol{x}_t), \tilde{\boldsymbol{m}}_{t+1} - \boldsymbol{m}_{t+1}\rangle\right]}_{T_{2,2}}.$$

We proceed by analysis $T_{2,1}$,

$$\begin{aligned}
T_{2,1} &= -\frac{\eta_g}{1-\beta_1}\mathbb{E}\left[\langle \nabla f(\boldsymbol{x}_t), \boldsymbol{d}_{t+1}\rangle\right] \\
&= -\frac{\eta_g}{1-\beta_1}\mathbb{E}\left[\left\langle \nabla f(\boldsymbol{x}_t), \frac{\eta_l}{N}\sum_{i\in[N]}\sum_{\tau=0}^{I-1}\tilde{\boldsymbol{g}}_{t,\tau}^{(i)}\right\rangle\right] \\
&= -\frac{\eta_g}{1-\beta_1}\mathbb{E}\left[\left\langle \nabla f(\boldsymbol{x}_t), \frac{\eta_l}{N}\sum_{i\in[N]}\sum_{\tau=0}^{I-1}\tilde{\boldsymbol{g}}_{t,\tau}^{(i)} + \eta_l I\nabla f(\boldsymbol{x}_t) - \eta_l I\nabla f(\boldsymbol{x}_t)\right\rangle\right] \\
&= -\frac{I\eta_g\eta_l}{1-\beta_1}\mathbb{E}\left[\left\|\nabla f(\boldsymbol{x}_t)\right\|^2\right] + \frac{\eta_g}{1-\beta_1}\mathbb{E}\left[\left\langle -\sqrt{\eta_l I}\nabla f(\boldsymbol{x}_t), \frac{\sqrt{\eta_l}}{\sqrt{I}N}\sum_{i\in[N]}\sum_{\tau=0}^{I-1}\left(\tilde{\boldsymbol{g}}_{t,\tau}^{(i)} - \nabla f_i(\boldsymbol{x}_t)\right)\right\rangle\right] \\
&\overset{(a)}{=} -\frac{I\eta_g\eta_l}{1-\beta_1}\mathbb{E}\left[\left\|\nabla f(\boldsymbol{x}_t)\right\|^2\right] + \frac{\eta_g}{1-\beta_1}\mathbb{E}\Bigg[\frac{\eta_l I}{2}\left\|\nabla f(\boldsymbol{x}_t)\right\|^2 + \frac{\eta_l}{2IN^2}\left\|\sum_{i\in[N]}\sum_{\tau=0}^{I-1}\left(\tilde{\boldsymbol{g}}_{t,\tau}^{(i)} - \nabla f_i(\boldsymbol{x}_t)\right)\right\|^2 \\
&\qquad\qquad - \frac{\eta_l}{2IN^2}\left\|\sum_{i\in[N]}\sum_{\tau=0}^{I-1}\tilde{\boldsymbol{g}}_{t,\tau}^{(i)}\right\|^2\Bigg] \\
&= -\frac{I\eta_g\eta_l}{2(1-\beta_1)}\mathbb{E}\left[\left\|\nabla f(\boldsymbol{x}_t)\right\|^2\right] + \frac{\eta_g\eta_l}{2I(1-\beta_1)N^2}\mathbb{E}\left\|\sum_{i\in[N]}\sum_{\tau=0}^{I-1}\left(\tilde{\boldsymbol{g}}_{t,\tau}^{(i)} - \boldsymbol{g}_{t,\tau}^{(i)} + \nabla f_i(\boldsymbol{x}_{t,\tau}^{(i)}) - \nabla f_i(\boldsymbol{x}_t)\right)\right\|^2 \\
&\qquad\qquad - \frac{\eta_g\eta_l}{2I(1-\beta_1)N^2}\mathbb{E}\left\|\sum_{i\in[N]}\sum_{\tau=0}^{I-1}\tilde{\boldsymbol{g}}_{t,\tau}^{(i)}\right\|^2 \\
&\overset{(b)}{\leq} -\frac{I\eta_g\eta_l}{2(1-\beta_1)}\mathbb{E}\left[\left\|\nabla f(\boldsymbol{x}_t)\right\|^2\right] + \frac{\eta_g\eta_l}{(1-\beta_1)N}\sum_{i\in[N]}\sum_{\tau=0}^{I-1}\left[\mathbb{E}\left\|\tilde{\boldsymbol{g}}_{t,\tau}^{(i)} - \boldsymbol{g}_{t,\tau}^{(i)}\right\|^2 + \mathbb{E}\left\|\nabla f_i(\boldsymbol{x}_{t,\tau}^{(i)}) - \nabla f_i(\boldsymbol{x}_t)\right\|^2\right] \\
&\qquad\qquad - \frac{\eta_g\eta_l}{2I(1-\beta_1)N^2}\mathbb{E}\left\|\sum_{i\in[N]}\sum_{\tau=0}^{I-1}\tilde{\boldsymbol{g}}_{t,\tau}^{(i)}\right\|^2 \\
&\leq -\frac{I\eta_g\eta_l}{2(1-\beta_1)}\mathbb{E}\left[\left\|\nabla f(\boldsymbol{x}_t)\right\|^2\right] + \frac{I\eta_g\eta_l\varepsilon_l^2}{(1-\beta_1)} + \frac{\eta_g\eta_l L^2}{(1-\beta_1)N}\sum_{i\in[N]}\sum_{\tau=0}^{I-1}\mathbb{E}\left[\left\|\boldsymbol{x}_{t,\tau}^{(i)} - \boldsymbol{x}_t\right\|^2\right] \\
&\qquad\qquad - \frac{\eta_g\eta_l}{2I(1-\beta_1)N^2}\mathbb{E}\left\|\sum_{i\in[N]}\sum_{\tau=0}^{I-1}\tilde{\boldsymbol{g}}_{t,\tau}^{(i)}\right\|^2,
\end{aligned}$$

where (a) follows from the fact that $\pm\langle \boldsymbol{a}, \boldsymbol{b}\rangle = \frac{1}{2}(\|\boldsymbol{a}\|^2 + \|\boldsymbol{b}\|^2 - \|\boldsymbol{a}-\boldsymbol{b}\|^2)$, and (b) uses the inequality $\left\|\sum_{i=1}^N \boldsymbol{x}_i\right\|^2 \leq N\sum_{i=1}^N \|\boldsymbol{x}_i\|^2, \boldsymbol{x}_i \in \mathbb{R}^d$. And we proceed by analysis $T_{2,2}$,

$$\begin{aligned}
T_{2,2} &= -\frac{\eta_g}{1-\beta_1}\mathbb{E}\left[\langle \nabla f(\boldsymbol{x}_t), \tilde{\boldsymbol{m}}_{t+1} - \boldsymbol{m}_{t+1}\rangle\right] \\
&\overset{(a)}{\leq} \frac{I\eta_g\eta_l}{4(1-\beta_1)}\mathbb{E}\left[\left\|\nabla f(\boldsymbol{x}_t)\right\|^2\right] + \frac{\eta_g}{(1-\beta_1)I\eta_l}\mathbb{E}\left[\left\|\tilde{\boldsymbol{m}}_{t+1} - \boldsymbol{m}_{t+1}\right\|^2\right]
\end{aligned}$$

$$\leq \frac{I\eta_g\eta_l}{4(1-\beta_1)}\mathbb{E}\left[\|\nabla f(\boldsymbol{x}_t)\|^2\right] + \frac{\eta_g\varepsilon_g^2}{(1-\beta_1)(1-\beta_2)I\eta_l},$$

where (a) using the fact that $\pm\langle \boldsymbol{a}, \boldsymbol{b}\rangle \leq \frac{1}{2}\|\boldsymbol{a}\|^2 + \frac{1}{2}\|\boldsymbol{b}\|^2$ $(\boldsymbol{a}, \boldsymbol{b} \in \mathbb{R}^d)$ where $\boldsymbol{a} = -\frac{\sqrt{I\eta_l}}{\sqrt{2}}\nabla f(\boldsymbol{x}_t)$ and $\boldsymbol{b} = \frac{\sqrt{2}}{\sqrt{I\eta_l}}(\tilde{\boldsymbol{m}}_{t+1} - \boldsymbol{m}_{t+1})$.

Next, we utilize the statement of Lemma .4 to derive the following upper bound of T_3,

$$T_3 = \frac{L\eta_g^2}{2(1-\beta_1)^2}\mathbb{E}\left[\left\|\tilde{\boldsymbol{d}}_{t+1}\right\|^2\right] \leq \frac{L\eta_g^2\eta_l^2}{(1-\beta_1)^2S^2}\mathbb{E}\left[\left\|\sum_{i\in[N]}\mathbb{I}\{i\in\mathcal{S}_t\}\sum_{\tau=0}^{I-1}\tilde{\boldsymbol{g}}_{t,\tau}^{(i)}\right\|^2\right] + \frac{L\eta_g^2\varepsilon_g^2}{(1-\beta_1)^2(1-\beta_2)}.$$

Substituting the upper bounds of $T_{1,1}$, $T_{1,2}$ into T_1 and $T_{2,1}$, $T_{2,2}$ into T_2 and T_1, T_2, T_3 into (25) yields:

$$\begin{aligned}
&\mathbb{E}[f(\boldsymbol{u}_{t+1})] \\
&\leq \mathbb{E}[f(\boldsymbol{u}_t)] \underbrace{-\frac{\eta_g}{1-\beta_1}\mathbb{E}\left[\left\langle \nabla f(\boldsymbol{u}_t) - \nabla f(\boldsymbol{x}_t), \tilde{\boldsymbol{d}}_{t+1}\right\rangle\right]}_{T_1} \underbrace{-\frac{\eta_g}{1-\beta_1}\mathbb{E}\left[\left\langle \nabla f(\boldsymbol{x}_t), \tilde{\boldsymbol{d}}_{t+1}\right\rangle\right]}_{T_2} + \underbrace{\frac{L\eta_g^2}{2(1-\beta_1)^2}\mathbb{E}\left[\left\|\tilde{\boldsymbol{d}}_{t+1}\right\|^2\right]}_{T_3} \\
&\leq \mathbb{E}[f(\boldsymbol{u}_t)] - \frac{I\eta_g\eta_l}{(1-\beta_1)}\left(\frac{1}{4} - 40I^2\eta_l^2L^2\right)\mathbb{E}\left[\|\nabla f(\boldsymbol{x}_t)\|^2\right] + \frac{(1-\beta_1)L}{\beta_1}\mathbb{E}\left[\|\boldsymbol{u}_t - \boldsymbol{x}_t\|^2\right] + \frac{\beta_1 L\eta_g^2\varepsilon_g^2}{2(1-\beta_1)^3(1-\beta_2)} \\
&+ \frac{I\eta_g\eta_l\varepsilon_l^2}{(1-\beta_1)} + \frac{\eta_g\varepsilon_g^2}{(1-\beta_1)(1-\beta_2)I\eta_l} + \frac{40I^3\eta_g\eta_l^3L^2\varepsilon_l^2}{(1-\beta_1)} + \frac{40I^3\eta_g\eta_l^3L^2\rho^2}{(1-\beta_1)} + \frac{L\eta_g^2\varepsilon_g^2}{(1-\beta_1)^2(1-\beta_2)} \\
&- \frac{\eta_g\eta_l}{2I(1-\beta_1)N^2}\left(1 - \frac{\beta_1 I\eta_g\eta_l L}{(1-\beta_1)^2}\right)\mathbb{E}\left[\left\|\sum_{i\in[N]}\sum_{\tau=0}^{I-1}\tilde{\boldsymbol{g}}_{t,\tau}^{(i)}\right\|^2\right] + \frac{L\eta_g^2\eta_l^2}{(1-\beta_1)^2S^2}\mathbb{E}\left[\left\|\sum_{i\in[N]}\mathbb{I}\{i\in\mathcal{S}_t\}\sum_{\tau=0}^{I-1}\tilde{\boldsymbol{g}}_{t,\tau}^{(i)}\right\|^2\right].
\end{aligned} \tag{26}$$

Using the inequality (26), making a simple arrangement and doing the summation operation from $t=0$ to $T-1$, we get:

$$\begin{aligned}
&\mathbb{E}[f(\boldsymbol{u}_T)] - \mathbb{E}[f(\boldsymbol{u}_0)] \\
&\leq -\frac{I\eta_g\eta_l}{(1-\beta_1)}\left(\frac{1}{4} - 40I^2\eta_l^2L^2\right)\sum_{t=0}^{T-1}\mathbb{E}\left[\|\nabla f(\boldsymbol{x}_t)\|^2\right] + \frac{(1-\beta_1)L}{\beta_1}\sum_{t=0}^{T-1}\mathbb{E}\left[\|\boldsymbol{u}_t - \boldsymbol{x}_t\|^2\right] + \frac{\beta_1 L\eta_g^2\varepsilon_g^2 T}{2(1-\beta_1)^3(1-\beta_2)} \\
&\quad + \frac{I\eta_g\eta_l\varepsilon_l^2 T}{(1-\beta_1)} + \frac{\eta_g\varepsilon_g^2 T}{(1-\beta_1)(1-\beta_2)I\eta_l} + \frac{40I^3\eta_g\eta_l^3L^2\varepsilon_l^2 T}{(1-\beta_1)} + \frac{40I^3\eta_g\eta_l^3L^2\rho^2 T}{(1-\beta_1)} + \frac{L\eta_g^2\varepsilon_g^2 T}{(1-\beta_1)^2(1-\beta_2)} \\
&\quad - \frac{\eta_g\eta_l}{2I(1-\beta_1)N^2}\left(1 - \frac{\beta_1 I\eta_g\eta_l L}{(1-\beta_1)^2}\right)\sum_{t=0}^{T-1}\mathbb{E}\left[\left\|\sum_{i\in[N]}\sum_{\tau=0}^{I-1}\tilde{\boldsymbol{g}}_{t,\tau}^{(i)}\right\|^2\right] + \frac{L\eta_g^2\eta_l^2}{(1-\beta_1)^2S^2}\sum_{t=0}^{T-1}\mathbb{E}\left[\left\|\sum_{i\in[N]}\mathbb{I}\{i\in\mathcal{S}_t\}\sum_{\tau=0}^{I-1}\tilde{\boldsymbol{g}}_{t,\tau}^{(i)}\right\|^2\right] \\
&\overset{(a)}{\leq} -\frac{I\eta_g\eta_l}{(1-\beta_1)}\left(\frac{1}{4} - 40I^2\eta_l^2L^2\right)\sum_{t=0}^{T-1}\mathbb{E}\left[\|\nabla f(\boldsymbol{x}_t)\|^2\right] + \frac{5\beta_1\eta_g^2 L\varepsilon_g^2 T}{2(1-\beta_1)^3(1-\beta_2)} \\
&\quad + \frac{I\eta_g\eta_l\varepsilon_l^2 T}{(1-\beta_1)} + \frac{\eta_g\varepsilon_g^2 T}{(1-\beta_1)(1-\beta_2)I\eta_l} + \frac{40I^3\eta_g\eta_l^3L^2\varepsilon_l^2 T}{(1-\beta_1)} + \frac{40I^3\eta_g\eta_l^3L^2\rho^2 T}{(1-\beta_1)} + \frac{\eta_g^2 L\varepsilon_g^2 T}{(1-\beta_1)^2(1-\beta_2)} \\
&\quad \underbrace{-\frac{\eta_g\eta_l}{2I(1-\beta_1)N^2}\left(1 - \frac{\beta_1 I\eta_g\eta_l L}{(1-\beta_1)^2}\right)\sum_{t=0}^{T-1}\mathbb{E}\left[\left\|\sum_{i\in[N]}\sum_{\tau=0}^{I-1}\tilde{\boldsymbol{g}}_{t,\tau}^{(i)}\right\|^2\right] + \frac{2\eta_g^2\eta_l^2 L}{(1-\beta_1)^3S^2}\sum_{t=0}^{T-1}\mathbb{E}\left[\left\|\sum_{i\in[N]}\mathbb{I}\{i\in\mathcal{S}_t\}\sum_{\tau=0}^{I-1}\tilde{\boldsymbol{g}}_{t,\tau}^{(i)}\right\|^2\right]}_{T_4},
\end{aligned} \tag{27}$$

where (a) holds by using the statement of Lemma .5. Next, we derive the upper bound for T_4. To simplify the proof process, we set $\boldsymbol{q}_{t,i} = \sum_{\tau=0}^{I-1} \tilde{\boldsymbol{g}}_{t,\tau}^{(i)}$ yields:

$$\begin{aligned}
\mathbb{E}\left[\left\|\sum_{i\in[N]}\sum_{\tau=0}^{I-1}\tilde{\boldsymbol{g}}_{t,\tau}^{(i)}\right\|^2\right] &= \mathbb{E}\left[\left\|\sum_{i\in[N]}\boldsymbol{q}_{t,i}\right\|^2\right] \\
&= \sum_{i\in[N]}\mathbb{E}\left[\left\|\boldsymbol{q}_{t,i}\right\|^2\right] + \sum_{i\neq j}\mathbb{E}\left[\langle \boldsymbol{q}_{t,i}, \boldsymbol{q}_{t,j}\rangle\right] \\
&= \sum_{i\in[N]} N\mathbb{E}\left[\left\|\boldsymbol{q}_{t,i}\right\|^2\right] - \frac{1}{2}\sum_{i\neq j}\mathbb{E}\left[\left\|\boldsymbol{q}_{t,i} - \boldsymbol{q}_{t,j}\right\|^2\right], \qquad (28)
\end{aligned}$$

and

$$\begin{aligned}
\mathbb{E}\left[\left\|\sum_{i\in[N]}\mathbb{I}\{i\in\mathcal{S}_t\}\sum_{\tau=0}^{I-1}\tilde{\boldsymbol{g}}_{t,\tau}^{(i)}\right\|^2\right] &= \mathbb{E}\left[\left\|\sum_{i\in[N]}\mathbb{P}\{i\in\mathcal{S}_t\}\boldsymbol{q}_{t,i}\right\|^2\right] \\
&= \sum_{i\in[N]}\mathbb{P}\{i\in\mathcal{S}_t\}\mathbb{E}\left[\left\|\boldsymbol{q}_{t,i}\right\|^2\right] + \sum_{i\neq j}\mathbb{P}\{i,j\in\mathcal{S}_t\}\mathbb{E}\left[\langle \boldsymbol{q}_{t,i}, \boldsymbol{q}_{t,j}\rangle\right] \\
&= \frac{S}{N}\sum_{i\in[N]}\mathbb{E}\left[\left\|\boldsymbol{q}_{t,i}\right\|^2\right] + \frac{S(S-1)}{N(N-1)}\sum_{i\neq j}\mathbb{E}\left[\langle \boldsymbol{q}_{t,i}, \boldsymbol{q}_{t,j}\rangle\right] \\
&= \frac{S^2}{N}\sum_{i\in[N]}\mathbb{E}\left[\left\|\boldsymbol{q}_{t,i}\right\|^2\right] - \frac{S(S-1)}{2N(N-1)}\sum_{i\neq j}\mathbb{E}\left[\left\|\boldsymbol{q}_{t,i} - \boldsymbol{q}_{t,j}\right\|^2\right], \qquad (29)
\end{aligned}$$

where $\mathbb{P}\{i\in\mathcal{S}_t\} = \frac{S}{N}$.

Substituting equalities (28) and (29) into T_4 yields:

$$\begin{aligned}
T_4 &= -\frac{\eta_g\eta_l}{2I(1-\beta_1)N^2}\left(1 - \frac{\beta_1 I\eta_g\eta_l L}{(1-\beta_1)^2}\right)\sum_{t=0}^{T-1}\mathbb{E}\left[\left\|\sum_{i\in[N]}\sum_{\tau=0}^{I-1}\tilde{\boldsymbol{g}}_{t,\tau}^{(i)}\right\|^2\right] + \frac{2\eta_g^2\eta_l^2 L}{(1-\beta_1)^3S^2}\sum_{t=0}^{T-1}\mathbb{E}\left[\left\|\sum_{i\in[N]}\mathbb{I}\{i\in\mathcal{S}_t\}\sum_{\tau=0}^{I-1}\tilde{\boldsymbol{g}}_{t,\tau}^{(i)}\right\|^2\right] \\
&= -\frac{\eta_g\eta_l}{2I(1-\beta_1)N^2}\left(1 - \frac{\beta_1 I\eta_g\eta_l L}{(1-\beta_1)^2}\right)\sum_{t=0}^{T-1}\left[\sum_{i\in[N]} N\mathbb{E}\left[\left\|\boldsymbol{q}_{t,i}\right\|^2\right] - \frac{1}{2}\sum_{i\neq j}\mathbb{E}\left[\left\|\boldsymbol{q}_{t,i} - \boldsymbol{q}_{t,j}\right\|^2\right]\right] \\
&\qquad + \frac{2\eta_g^2\eta_l^2 L}{(1-\beta_1)^3S^2}\sum_{t=0}^{T-1}\left[\frac{S^2}{N}\sum_{i\in[N]}\mathbb{E}\left[\left\|\boldsymbol{q}_{t,i}\right\|^2\right] - \frac{S(S-1)}{2N(N-1)}\sum_{i\neq j}\mathbb{E}\left[\left\|\boldsymbol{q}_{t,i} - \boldsymbol{q}_{t,j}\right\|^2\right]\right] \\
&= -\frac{\eta_g\eta_l}{2I(1-\beta_1)N}\left(1 - \frac{\beta_1 I\eta_g\eta_l L}{(1-\beta_1)^2} - \frac{4I\eta_g\eta_l L}{(1-\beta_1)^2}\right)\sum_{t=0}^{T-1}\sum_{i\in[N]}\mathbb{E}\left[\left\|\boldsymbol{q}_{t,i}\right\|^2\right] \\
&\qquad + \frac{\eta_g\eta_l}{4I(1-\beta_1)N^2}\left(1 - \frac{\beta_1 I\eta_g\eta_l L}{(1-\beta_1)^2} - \frac{4I\eta_g\eta_l L}{(1-\beta_1)^2}\frac{N(S-1)}{S(N-1)}\right)\sum_{t=0}^{T-1}\sum_{i\neq j}\mathbb{E}\left[\left\|\boldsymbol{q}_{t,i} - \boldsymbol{q}_{t,j}\right\|^2\right] \\
&\overset{(a)}{=} -\frac{\eta_g\eta_l}{2I(1-\beta_1)N}\left(1 - \frac{\beta_1 I\eta_g\eta_l L}{(1-\beta_1)^2} - \frac{4I\eta_g\eta_l L}{(1-\beta_1)^2}\right)\sum_{t=0}^{T-1}\sum_{i\in[N]}\mathbb{E}\left[\left\|\boldsymbol{q}_{t,i}\right\|^2\right] \\
&\quad + \frac{\eta_g\eta_l}{4I(1-\beta_1)N^2}\left(1 - \frac{\beta_1 I\eta_g\eta_l L}{(1-\beta_1)^2} - \frac{4I\eta_g\eta_l L}{(1-\beta_1)^2}\frac{N(S-1)}{S(N-1)}\right)\sum_{t=0}^{T-1}\left[2N\sum_{i\in[N]}\mathbb{E}\left[\left\|\boldsymbol{q}_{t,i}\right\|^2\right] - 2\mathbb{E}\left[\left\|\sum_{i\in[N]}\boldsymbol{q}_{t,i}\right\|^2\right]\right]
\end{aligned}$$

$$
\begin{aligned}
&= \frac{2\eta_g^2\eta_l^2 L}{N(1-\beta_1)^3}\left(1-\frac{N(S-1)}{S(N-1)}\right)\sum_{t=0}^{T-1}\sum_{i\in[N]}\mathbb{E}\left[\|\boldsymbol{q}_{t,i}\|^2\right] \\
&\qquad - \frac{\eta_g\eta_l}{2I(1-\beta_1)N^2}\left(1-\frac{\beta_1 I\eta_g\eta_l L}{(1-\beta_1)^2}-\frac{4I\eta_g\eta_l L}{(1-\beta_1)^2}\frac{N(S-1)}{S(N-1)}\right)\sum_{t=0}^{T-1}\mathbb{E}\left[\|\sum_{i\in[N]}\boldsymbol{q}_{t,i}\|^2\right] \\
&= \frac{2\eta_g^2\eta_l^2 L}{(1-\beta_1)^3}\frac{N-S}{NS(N-1)}\sum_{t=0}^{T-1}\sum_{i\in[N]}\mathbb{E}\left[\|\boldsymbol{q}_{t,i}\|^2\right] \\
&\qquad - \frac{\eta_g\eta_l}{2I(1-\beta_1)N^2}\left(1-\frac{\beta_1 I\eta_g\eta_l L}{(1-\beta_1)^2}-\frac{4I\eta_g\eta_l L}{(1-\beta_1)^2}\frac{N(S-1)}{S(N-1)}\right)\sum_{t=0}^{T-1}\mathbb{E}\left[\|\sum_{i\in[N]}\boldsymbol{q}_{t,i}\|^2\right] \\
&\overset{(b)}{\leq} \frac{2\eta_g^2\eta_l^2 L}{(1-\beta_1)^3}\frac{N-S}{NS(N-1)}\sum_{t=0}^{T-1}\sum_{i\in[N]}\mathbb{E}\left[\|\boldsymbol{q}_{t,i}\|^2\right] \\
&\overset{(c)}{\leq} \frac{2\eta_g^2\eta_l^2 L}{(1-\beta_1)^3}\frac{N-S}{S(N-1)}\left[4I^2(\varepsilon_l^2+\rho^2)(1+40I^2\eta_l^2L^2)T+4I^2(1+40I^2\eta_l^2L^2)\sum_{t=0}^{T-1}\mathbb{E}\left[\|\nabla f(\boldsymbol{x}_t)\|^2\right]\right] \\
&= \frac{8I^2\eta_g^2\eta_l^2 L(\varepsilon_l^2+\rho^2)(1+40I^2\eta_l^2L^2)T}{(1-\beta_1)^3}\frac{N-S}{S(N-1)}+\frac{8I^2\eta_g^2\eta_l^2 L(1+40I^2\eta_l^2L^2)}{(1-\beta_1)^3}\frac{N-S}{S(N-1)}\sum_{t=0}^{T-1}\mathbb{E}\left[\|\nabla f(\boldsymbol{x}_t)\|^2\right],
\end{aligned}
$$

where (a) holds by using $\sum_{i\neq j}\mathbb{E}\left[\|\boldsymbol{q}_{t,i}-\boldsymbol{q}_{t,j}\|^2\right] = 2N\sum_{i\in[N]}\mathbb{E}\left[\|\boldsymbol{q}_{t,i}\|^2\right] - 2\mathbb{E}\left[\|\sum_{i\in[N]}\boldsymbol{q}_{t,i}\|^2\right]$, (b) results from the fact that $1-\frac{\beta_1 I\eta_g\eta_l L}{(1-\beta_1)^2}-\frac{4I\eta_g\eta_l L}{(1-\beta_1)^2}\frac{N(S-1)}{S(N-1)}\geq 0$ holds if $\eta_g\eta_l \leq \frac{(1-\beta_1)^2S(N-1)}{IL(\beta_1 S(N-1)+4N(S-1))}$, and (c) follows from the statement of Lemma .3.

Furthermore, substituting the upper bound of T_4 into (27), we get:

$$
\begin{aligned}
&\mathbb{E}[f(\boldsymbol{u}_T)]-\mathbb{E}[f(\boldsymbol{u}_0)] \\
&\leq -\frac{I\eta_g\eta_l}{(1-\beta_1)}\left(\frac{1}{4}-40I^2\eta_l^2L^2\right)\sum_{t=0}^{T-1}\mathbb{E}\left[\|\nabla f(\boldsymbol{x}_t)\|^2\right]+\frac{5\beta_1\eta_g^2L\varepsilon_g^2T}{2(1-\beta_1)^3(1-\beta_2)} \\
&\qquad +\frac{I\eta_g\eta_l\varepsilon_l^2T}{(1-\beta_1)}+\frac{\eta_g\varepsilon_g^2T}{(1-\beta_1)(1-\beta_2)I\eta_l}+\frac{40I^3\eta_g\eta_l^3L^2\varepsilon_l^2T}{(1-\beta_1)}+\frac{40I^3\eta_g\eta_l^3L^2\rho^2T}{(1-\beta_1)}+\frac{\eta_g^2L\varepsilon_g^2T}{(1-\beta_1)^2(1-\beta_2)} \\
&\underbrace{\quad -\frac{\eta_g\eta_l}{2I(1-\beta_1)N^2}\left(1-\frac{\beta_1 I\eta_g\eta_l L}{(1-\beta_1)^2}\right)\sum_{t=0}^{T-1}\mathbb{E}\left[\left\|\sum_{i\in[N]}\sum_{\tau=0}^{I-1}\tilde{\boldsymbol{g}}_{t,\tau}^{(i)}\right\|^2\right]+\frac{2\eta_g^2\eta_l^2L}{(1-\beta_1)^3S^2}\sum_{t=0}^{T-1}\mathbb{E}\left[\left\|\sum_{i\in[N]}\mathbb{I}\{i\in\mathcal{S}_t\}\sum_{\tau=0}^{I-1}\tilde{\boldsymbol{g}}_{t,\tau}^{(i)}\right\|^2\right]}_{T_4} \\
&\leq -\frac{I\eta_g\eta_l}{(1-\beta_1)}\left(\frac{1}{4}-40I^2\eta_l^2L^2\right)\sum_{t=0}^{T-1}\mathbb{E}\left[\|\nabla f(\boldsymbol{x}_t)\|^2\right]+\frac{5\beta_1\eta_g^2L\varepsilon_g^2T}{2(1-\beta_1)^3(1-\beta_2)} \\
&\qquad +\frac{I\eta_g\eta_l\varepsilon_l^2T}{(1-\beta_1)}+\frac{\eta_g\varepsilon_g^2T}{(1-\beta_1)(1-\beta_2)I\eta_l}+\frac{40I^3\eta_g\eta_l^3L^2\varepsilon_l^2T}{(1-\beta_1)}+\frac{40I^3\eta_g\eta_l^3L^2\rho^2T}{(1-\beta_1)}+\frac{\eta_g^2L\varepsilon_g^2T}{(1-\beta_1)^2(1-\beta_2)} \\
&\quad +\frac{8I^2\eta_g^2\eta_l^2L(\varepsilon_l^2+\rho^2)(1+40I^2\eta_l^2L^2)T}{(1-\beta_1)^3}\frac{N-S}{S(N-1)}+\frac{8I^2\eta_g^2\eta_l^2L(1+40I^2\eta_l^2L^2)}{(1-\beta_1)^3}\frac{N-S}{S(N-1)}\sum_{t=0}^{T-1}\mathbb{E}\left[\|\nabla f(\boldsymbol{x}_t)\|^2\right] \\
&= -\frac{I\eta_g\eta_l}{(1-\beta_1)}\left(\frac{1}{4}-40I^2\eta_l^2L^2-\frac{8I\eta_g\eta_l L(1+40I^2\eta_l^2L^2)}{(1-\beta_1)^2}\frac{N-S}{S(N-1)}\right)\sum_{t=0}^{T-1}\mathbb{E}\left[\|\nabla f(\boldsymbol{x}_t)\|^2\right]+\frac{5\beta_1\eta_g^2L\varepsilon_g^2T}{2(1-\beta_1)^3(1-\beta_2)} \\
&\qquad +\frac{I\eta_g\eta_l\varepsilon_l^2T}{(1-\beta_1)}+\frac{\eta_g\varepsilon_g^2T}{(1-\beta_1)(1-\beta_2)I\eta_l}+\frac{40I^3\eta_g\eta_l^3L^2\varepsilon_l^2T}{(1-\beta_1)}+\frac{40I^3\eta_g\eta_l^3L^2\rho^2T}{(1-\beta_1)}+\frac{\eta_g^2L\varepsilon_g^2T}{(1-\beta_1)^2(1-\beta_2)}
\end{aligned}
$$

$$+\frac{8I^2\eta_g^2\eta_l^2L(\varepsilon_l^2+\rho^2)(1+40I^2\eta_l^2L^2)T}{(1-\beta_1)^3}\frac{N-S}{S(N-1)}$$

$$\overset{(a)}{\leq}-\frac{I\eta_g\eta_l}{8(1-\beta_1)}\sum_{t=0}^{T-1}\mathbb{E}\left[\|\nabla f(\boldsymbol{x}_t)\|^2\right]+\frac{5\beta_1\eta_g^2L\varepsilon_g^2T}{2(1-\beta_1)^3(1-\beta_2)}+\frac{I\eta_g\eta_l\varepsilon_l^2T}{(1-\beta_1)}+\frac{\eta_g\varepsilon_g^2T}{(1-\beta_1)(1-\beta_2)I\eta_l}+\frac{40I^3\eta_g\eta_l^3L^2\varepsilon_l^2T}{(1-\beta_1)}$$

$$+\frac{40I^3\eta_g\eta_l^3L^2\rho^2T}{(1-\beta_1)}+\frac{\eta_g^2L\varepsilon_g^2T}{(1-\beta_1)^2(1-\beta_2)}+\frac{8I^2\eta_g^2\eta_l^2L(\varepsilon_l^2+\rho^2)(1+40I^2\eta_l^2L^2)T}{(1-\beta_1)^3}\frac{N-S}{S(N-1)},\tag{30}$$

where (a) results from the fact that $\frac{1}{4}-40I^2\eta_l^2L^2-\frac{8I\eta_g\eta_lL(1+40I^2\eta_l^2L^2)}{(1-\beta_1)^2}\frac{N-S}{S(N-1)}\geq\frac{1}{8}$ holds if $320I^2\eta_l^2L^2+\frac{64I\eta_g\eta_lL(1+40I^2\eta_l^2L^2)}{(1-\beta_1)^2}\frac{N-S}{S(N-1)}\leq 1$. Now, we use the statement of Assumption 1 yields:

$$f^\star-\mathbb{E}[f(\boldsymbol{x}_0)]\leq\mathbb{E}[f(\boldsymbol{u}_T)]-\mathbb{E}[f(\boldsymbol{u}_0)].\tag{31}$$

This holds as $\boldsymbol{u}_0=\boldsymbol{x}_0$. Finally, the proof of the lemma is completed by substituting inequality (31) into inequality (30) and making a simple arrangement.

End Proof.