Supplemental Material of ADFed: Asynchronous Decentralized Federated Learning with Efficient Cluster-Based Aggregation

Jiahuai Mao[1,5], Zehui Xiong[2], Riccardo Spolaor[3], Baosheng Li[4], Yaxi Yang[4], Lei Zhang[1], Man Ho Au[5], Shuo Wang[6]

[1,]*Software Engineering Institute, East China Normal University, Shanghai, China*
[2]*School of Electronics, Electrical Engineering and Computer Science, Queen's University Belfast, UK*
[3]*School of Computer Science and Technology, Shandong University, China*
[4]*Information Systems Technology and Design Pillar, Singapore University of Technology and Design, Singapore*
[5]*Department of Computing, The Hong Kong Polytechnic University, Hong Kong, China*
[6]*Department of Mathematics, University of Padua, Italy*

jiahuai.mao@polyu.edu.hk, z.xiong@qub.ac.uk, rspolaor@sdu.edu.cn, bs.li@stu.xidian.edu.cn, yxyangjnu@gmail.com, leizhang@sei.ecnu.edu.cn, mhaau@polyu.edu.hk, shuo.wang@studenti.unipd.it

Appendix A
Self-Adaptive Global-Best Harmony Search for Optimal Cluster Selection

The SGHS algorithm represents an efficient heuristic global search algorithm, inspired by the tuning behaviors of musicians during improvisational performances. This algorithm continuously adjusts the solution variables within the solution space through iterative processes. Each iteration involves selecting and potentially adjusting solutions from Harmony Memory (HM) based on specific probabilities to gradually approach the global optimal solution and ultimately minimize the value of the function. Its success comes from effectively balancing exploration and exploitation, avoiding local minima, and converging quickly. The key parameters of the SGHS comprise the following:

- Harmony Memory Size (HMS): represents the number of solutions that can be stored in the HM. A larger HMS enhances the diversity of solutions, but can also increase computational complexity. In this study, HMS is set to 10, a value chosen based on previous studies that demonstrated its effectiveness in balancing solution diversity and computational efficiency, as well as our preliminary experiments that validate its suitability for the problem at hand.
- Harmony Memory Considering Rate (HMCR): defines the probability of selecting existing solutions from the stored solutions in HM. A higher HMCR facilitates local search, thus increasing the convergence speed of the algorithm, while a lower HMCR encourages the exploration of new solution spaces. We have set the HMCR to follow a normal distribution with a mean of 0.98 and a standard deviation of 0.01, based on its effectiveness in previous work.
- Pitch Adjusting Rate (PAR): defines the probability of adjusting a solution. PAR is set to a normal distribution with a mean of 0.90 and a standard deviation of 0.05. This value is informed by existing studies suggesting that a moderately high PAR supports fine-tuning of solutions, while maintaining adequate randomness to escape local optima.
- Number of Improvisations (NI): represents the number of iterations the algorithm will perform. We set this value to 2000 based on a trade-off between solution quality and computational time, as observed in our preliminary experiments.
- Distance Bandwidth (DB): defines the maximum distance a solution can move during an adjustment. A larger DB value is advantageous for searching within a broader range, while a smaller DB is suitable for fine-tuning the optimal solution. To effectively balance exploration and exploitation in the proposed SGHS algorithm, the DB value dynamically decreases as the NI increases, as follows:

$$\mathrm{DB} = \begin{cases} \mathrm{DB}^{\max} - 2t\frac{\mathrm{DB}^{\max}-\mathrm{DB}^{\min}}{\mathrm{NI}}, & t < \mathrm{NI}/2 \\ \mathrm{DB}^{\min}, & t \geq \mathrm{NI}/2 \end{cases} \quad (1)$$

where $\mathrm{DB}^{\min} = 2$ and $\mathrm{DB}^{\max} = \frac{N-1}{4}$ are the minimum and maximum distance bandwidths.

As illustrated in Algorithm 4, after setting the fundamental parameters, the coordinator initiates by initializing the HM. HM comprises potential HMS solutions, each calculated by the formula $K_a = 1 + (N-1) \times r$, where r is a uniform random number between 0 and 1, and $1 \leq a \leq$ HMS. Subsequently, the total cost for each solution K_a is assessed using Eq. (14) The coordinator then iteratively computes the optimal number of clusters K^{best} according to lines 3 to 24 of Algorithm 4. To enhance the adaptability of the algorithm, we dynamically adjust the HMCR and PAR. After a specified

learning period (LP, which is 100 in our experiment), we update the mean HMCR and PAR based on the averages of all recorded HMCR and PAE values during this period. This update process is represented as HMCR= $\sum$HMCR*/LP* and PAR= $\sum$PAR*/LP* , where LP* denotes the number of successful replacements of the worst solutions in HM, and HMCR* and PAR* represent the corresponding HMCR and PAR values during LP* valid iterations. Upon completion of all sub-objectives of the algorithm, we obtain the optimal number of clusters K^{best}.

Algorithm 4 SGHS-Driven Optimizer for Minimizing Eq. (14)

Require: Set the Parameters: HMS, HMCR, PAR, NI, DB$^{\max}$, DB$^{\min}$, LP.
Ensure: The best harmony K^{best} minimizing $O(K)$.
Initialize harmony memory HM and evaluate $O(K_a)$ for each harmony $K_a \in$ HM
$t = 1$ ▷ Set iteration counter
while $t \leq$ NI **do** ▷ Main loop for NI iterations
  if $t\%$LP = 0 **then** ▷ Every LP iterations
    Update the mean of HMCR and PAR ▷ Adapt HMCR and PAR parameters
  end if
  Generate r_1, r_2, $r_3 \in (0, 1)$, HMCR, PAR, and compute DB according to Eq. (1)
  Identify K^{best} with the lowest $O(K)$ in HM, and K^{worst} with the highest $O(K)$ ▷ Find best and worst harmonies
  if $r_1 <$ HMCR **then**
    $K^{\text{new}} = \lfloor K_a \pm r_3 \times \text{DB} \rfloor$, $a \in \{1, 2, ..., \text{HMS}\}$
    if $r_2 <$ PAR **then**
      $K^{\text{new}} = K^{\text{best}}$ ▷ Adjust towards best harmony
    end if
  else ▷ Random selection outside memory
    $K^{\text{new}} = \lfloor K^{\min} + r_3 \times (K^{\max} - K^{\min}) \rfloor$ ▷ Generate new random harmony
  end if
  Evaluate $O(K^{\text{new}})$ ▷ Compute cost of new harmony
  if $O(K^{\text{new}}) < O(K^{\text{worst}})$ **then**
    Replace K^{worst} with K^{new} in HM ▷ Update harmony memory
  end if
  if $O(K^{\text{new}}) < O(K^{\text{best}})$ **then**
    Update $K^{\text{best}} = K^{\text{new}}$ ▷ Update best harmony found
  end if
  $t = t + 1$ ▷ Increment iteration counter
end while
return K^{best} ▷ Output the optimal number of clusters

APPENDIX B
COMPLETE CONVERGENCE ANALYSIS

In this section, we theoretically analyze the model convergence rate of our method. Furthermore, we obtain an upper bound on convergence regarding network topology, i.e. number of clusters. First, we make the following assumptions that are widely used in related work [1], [2].

Assumption 1. *In this paper, we base our analysis on the following commonly accepted assumptions:*

1) ***Lipschitzian Gradient:*** *The loss function f possesses L-Lipschitzian gradients.*
2) ***Spectral Gap:*** *The matrix W^r is doubly stochastic for every training round r. There exists a $\xi \in$ [0, 1) such that for all r, max$\{\lambda_2(\mathbb{E}[{W^r}^T W^r]), \lambda_K(\mathbb{E}[{W^r}^T W^r])\} \leq \xi$. A smaller ξ indicates a more rapid dissemination of information throughout the network, resulting in faster convergence.*
3) ***Unbiased Estimation:*** *Let $\mathcal{B}$ be a mini-batch at communication round r of client C_i in cluster V_k. Then the local gradient estimator is unbiased as follows:*

$$\mathbb{E}_{\mathcal{B}\sim D_i} \nabla F(m_i, \mathcal{B}) = \nabla f(m_i), \tag{2}$$

$$\mathbb{E}_{C_i\sim\mathcal{C}} \mathbb{E}_{\mathcal{B}\sim D_i} \nabla F(m_i, \mathcal{B}) = \nabla f(m). \tag{3}$$

4) ***Bounded Gradient Variance:*** *Suppose the variance of the stochastic gradient is bounded for every m_i, with C_i sampled from distribution $\mathcal{C}$ and $\mathcal{B}$ from distribution D_i. This suggests the existence of constants α and β such that:*

$$\mathbb{E}_{\mathcal{B}\sim D_i} \| \nabla F(m_i, \mathcal{B}) - \nabla f(m_i) \|^2 \leq \alpha, \forall C_i, \forall m_i, \tag{4}$$

$$\mathbb{E}_{C_i\sim\mathcal{C}} \| \nabla f(m_i) - \nabla f(m) \|^2 \leq \beta, \forall m_i. \tag{5}$$

Note that in Eq. (4), α measures the deviation of the estimated gradient over mini-batch $\mathcal{B}$ from the true gradient. β in Eq. (5) denotes the degree of dissimilarity between the local loss functions of the clients and the global loss function $f(m)$, illustrating the statistical diversity present in the non-IID data sets of all clients. Specifically, when the data distributions among clients are independent and identically distributed (IID), all local loss functions are equivalent (i.e., $f(w_i) = f(w_j), \forall C_i, C_j \in \mathcal{C}$), resulting in $\beta = 0$.

5) ***Bounded staleness:*** *$m_{i,k}^r = m_k^{r-\tau_r}$ and there exists a constant Γ such that max$\{\tau_r\} \leq \Gamma$.*

Throughout this paper, we introduce the following definitions for clarity and simplicity.

$$\bar{\kappa} = \frac{K-1}{K}\left(\frac{1}{1-\kappa} + \frac{2\sqrt{\kappa}}{(1-\sqrt{\kappa})^2}\right),$$
$$\lambda = 1 - 24L^2B^2\eta^2\left(\frac{K-1}{K}\Gamma + \bar{\kappa}\right),$$
$$\Psi_1 = \frac{L^2B\eta}{K}(1 + \frac{6LB\eta}{K} + \frac{12L^2B^2\eta^2\Gamma^2}{K^2}),$$
$$\Psi_2 = \frac{LBR\eta^2(\alpha^2 + 6\beta^2 B)}{K^2}(\frac{1}{2} + \frac{LB\eta\Gamma^2}{K}),$$
$$\Psi_3 = \frac{B\eta}{K}(\frac{LB\eta}{K} + \frac{2L^2B^2\eta^2\Gamma^2}{K^2} - \frac{1}{2}).$$

Proposition 1. *According to Algorithm 2 and 4, the difference in data distribution between clusters tends to 0 if K is a constant as N approaches infinity, i.e.,*

$$\lim_{N\to\infty} EMD(L_k, L_{k'}) \to 0, \quad k \in [K], \ K \ll N, \tag{6}$$

where EMD is the earth mover distance and $L_k, L_{k'}$ are the data distribution of client k and k', respectively.

Remark 1. *From Proposition 1, it directly follows that the difference of data distribution between the clusters and the global data across all clients converges to zero. This implies that the data distributions among clusters are IID.*

Therefore, each cluster can be considered as a client, with the data distributions between clusters being IID (independently and identically distributed). In this context, Algorithm 1 and AD-SGD are equivalent. This means that we can give the theorem alike to [2]. However, the most notable distinction is that [2] additionally accounts for non-IID scenarios. The main iteration rule in Algorithm 1 can be rewritten as follows:

$$M^{r+1} \leftarrow M^r W^r - \eta \partial g(\hat{M}^r, k^r), k \in [K], \tag{7}$$

where $\hat{M}^r := M^{r-\tau}, \tau \in \mathbb{Z}^+$ and

$$\begin{aligned} M^r &:= [m_1^r, \ldots, m_K^r] \in \mathbb{R}^{d \times K}, \\ \hat{M}^r &:= [\hat{m}_1^r, \ldots, \hat{m}_K^r] \in \mathbb{R}^{d \times K}, \\ \partial g(\hat{M}^r, k^r) &:= [0, \ldots, \sum_{j=1}^{B} \nabla F(\hat{m}_{k^r}^r, \mathcal{B}_j^r), 0, \ldots, 0] \in \mathbb{R}^{d \times K}. \end{aligned}$$

Theorem 1. *Let Assumption 1.1-1.5 hold and define $f^* = f(m^*)$, the iteration sequence generated by Algorithm 1 has the following property:*

$$\begin{aligned} \frac{1}{R} \sum_{r=0}^{R-1} \mathbb{E}[\|\nabla f(M^r \frac{\mathbf{1}}{K})\|^2] \leq \frac{2K(f(m^0) - f^*)}{RB\eta} \\ + \frac{2L\eta(\alpha^2 + 6\beta^2 B)}{K}. \end{aligned} \tag{8}$$

where m^0 represents the initial model parameters, and m^* denotes the optimal model that minimize the average loss function values. This theorem describes the convergence behavior of local models across all leading clients. To further explore this result, we can appropriately select the learning rate as specified in Theorem 1, leading to the following corollary:

Corollary 1. *Let $\eta = \frac{K}{10LB + \sqrt{\alpha^2 + 6\beta^2 B}\sqrt{RB}}$. If the total number of iteration rounds is sufficiently large, specifically*

$$\begin{aligned} R \geq \frac{L^2 B K^2}{\alpha^2 + 6\beta^2 B} \max\Bigg\{ 192 \left(\frac{K-1}{K}\Gamma + \bar{\kappa} \right), \frac{64\Gamma^4}{K^2}, \\ 1024 K^2 \bar{\kappa}^2, \frac{(8\sqrt{6}\Gamma^{2/3} + 8)^2 (\Gamma + \bar{\kappa}\frac{K}{K-1})^{2/3} (K-1)^{1/2}}{K^{1/6}} \Bigg\}, \end{aligned} \tag{9}$$

we obtain the following convergence rate:

$$\begin{aligned} \frac{\sum_{r=0}^{R-1} \mathbb{E} \left\| \frac{1}{K} \sum_{k=1}^{K} \nabla f(m_k^r) \right\|^2}{R} \leq \mathcal{O}(\frac{L(f(m^0) - f^*)}{R}) \\ + \mathcal{O}(\frac{(f(m^0) - f^* + L)(\alpha^2/B + 6\beta^2)^{1/2}}{R^{1/2}}). \end{aligned} \tag{10}$$

This corollary suggests that for a sufficiently large number of rounds, ADFed exhibits a convergence rate of $O(1/\sqrt{R})$. Additionally, we will demonstrate ADFed's linear speed-up characteristics concerning batch size, number of leaders, and staleness, respectively.

Remark 2. *Linear speed-up w.r.t. batch size.* When R is sufficiently large, the second term on the right-hand side of (10) becomes dominant over the first term. In particular, if $\beta = 0$, the second term converges at a rate of $O(1/\sqrt{BR})$, indicating that increasing the size of the mini-batch improves the convergence efficiency at a linear rate. This finding highlights the linear speed-up relative to batch size. However, when $\beta \neq 0$, ADFed does not exhibit such linear speed-up with respect to batch size. This limitation is inherent, as an increase in the mini-batch size only reduces the stochastic gradient variance within each client, while β reflects the variance between clients, which remains unaffected by the batch size.

Remark 3. *Linear speed-up w.r.t. the number of leaders.* In our analysis, each stochastic gradient update is considered one round, and our convergence rate in Corollary 1 matches that of SGD/mini-batch SGD. This means that the number of updates needed to achieve a certain precision is consistent with SGD/mini-batch SGD, provided that the iteration count is sufficiently large. It further demonstrates a linear speed-up regarding the number of leaders K; K leaders will accelerate the iteration process K-fold in terms of wall-clock time, resulting in K-times faster convergence. Our analysis shows that by eliminating the synchronization of the leaders, ADFed maintains linear speed-up and reduces idle time in heterogeneous environments.

Remark 4. *Linear speed-up w.r.t. staleness.* From (9), we can also observe that linear speed-up is achievable as long as the staleness Γ is bounded by $O(R^{1/4})$, assuming other parameters are constants.

ADFed exhibits notable scalability by achieving linear speed-up in relation to both batch size and the number of leaders. When computational resources allow, increasing the batch size can significantly enhance convergence rates. However, this improvement may be constrained by variability across clusters, highlighting the need to minimize discrepancies in the data distribution among them. In addition to its scalability with batch size, ADFed also demonstrates linear speed-up when increasing the number of leaders. However, it is important to consider that adding more leaders can lead to higher overhead compared to standard clients. Therefore, it is essential to determine an optimal number of leaders that balances improved convergence speed with associated communication costs. Moreover, ADFed exhibits strong robustness in the presence of limited staleness, which refers to delays in updates caused by asynchronous operations. It can sustain linear speed-up as long as the level of staleness remains within a defined range. This characteristic further underscores ADFed's effectiveness in dynamic environments typical of decentralized federated learning.

In the following, we will demonstrate that ADFed with improved privacy and integrity still converges. The iteration sequence generated by Algorithm 3 can be written as:

$$M^{r+1} \leftarrow M^r W^r - \eta \partial \tilde{g}(\hat{M}^r, k^r), \quad k \in [K], \tag{11}$$

where $\tilde{g}(\hat{M}^r, k^r) = \tilde{g}(\hat{M}^r, k^r) + \delta$, δ is the white noise added in the gradient and the variance of δ is σ^2.

Assumption 2. *The variance of noise added to the local gradient is bound, that is, $Var(\delta) = \sigma^2 < \infty$.*

Theorem 2. *Let Assumptions 1.1-1.5 and 2 hold, the iteration*

sequence generated by Algorithm 3 has the following property:

$$\frac{1}{R}\sum_{r=0}^{R-1}\mathbb{E}[\|\nabla f(M^r\frac{\mathbf{1}}{K})\|^2] \leq \underbrace{\frac{2K(f(m^0)-f^*)}{RB\eta}}_{\textit{stochastic error}} + \frac{2L\eta(\alpha^2+6\beta^2B)}{K} + \underbrace{\frac{2L\eta\sigma^2(K^2+1)}{K^3}}_{\textit{noise error}}, \tag{12}$$

Corollary 2. *Let* $\eta = \frac{K}{10LB+\sqrt{\alpha^2+6\beta^2B}\sqrt{RB}}$*, we have*

$$\frac{\sum_{r=0}^{R-1}\mathbb{E}\left\|\frac{1}{K}\sum_{k=1}^{K}\nabla f(m_k^r)\right\|^2}{R} \leq \mathcal{O}(\frac{L(f(m^0)-f^*)}{R}) + \mathcal{O}(\frac{(f(m^0)-f^*+L)(\alpha^2/B+6\beta^2)^{1/2}(\sigma^2/B\frac{K^2+1}{K^2})^{1/2}}{R^{1/2}}). \tag{13}$$

Corollary 2 indicates that the sequence generated by Algorithm 3 has a convergence rate of $O(\sigma/\sqrt{R})$.

Remark 5. *Convergence regarding the variance of LDP noise.* The introduction of LDP increases the variance of gradient estimation (σ^2), directly causing the convergence rate of federated learning to degrade from $O(1/\sqrt{R})$ to $O(\sigma/\sqrt{R})$. This phenomenon reflects the inherent trade-off between privacy protection and model efficiency, necessitating algorithmic design to strike a balance between the two. Both theoretical analysis and experimental validation indicate that by appropriately controlling the noise intensity σ and incorporating adaptive optimization strategies, it is possible to maintain convergence performance while ensuring privacy.

A. Proof of Proposition 1

Proof. The global data distribution is defined as $\bar{\mathcal{L}} := \frac{1}{N}\sum_{i=1}^{N}\mathcal{L}_i$ The data distribution of the cluster k is defined as $\bar{\mathcal{L}}_k := \frac{1}{N_k}\sum_{i=1}^{N_k}\mathcal{L}_i, k \in [K]$. If $N \to \infty$ and K is fixed, We have:

$$\lim_{N\to\infty} N/K \to \infty, K \ll N.$$

This means that the number of clients in each cluster tends to infinity, and according to the law of large numbers,

$$\bar{\mathcal{L}}_k \overset{a.s.}{\to} \bar{\mathcal{L}}, k \in [K], k \ll N.$$

Since we use a greedy strategy to minimize the distribution difference in the clustering process, this directly leads to a reduction in the distribution difference between clusters, that is,

$$\lim_{N\to\infty} \text{EMD}(\bar{\mathcal{L}}_k, \bar{\mathcal{L}}) \to 0, K \ll N,$$

□

B. Proof of Theorem 1

We first define an average error, i.e., $\varepsilon^r = \sum_{k=1}^{K} p_k \|\frac{M^r\mathbf{1}_K}{K} - M^r e_k\|^2$. The specific notations are defined as follows:

$$\partial f(M^r) = K[p_1\nabla f_1(m_1^r), \ldots, p_K\nabla f_K(m_K^r)],$$
$$\partial g(\hat{M}^r, \mathcal{B}^r) = K[p_1\sum_{j=1}^{B}\nabla F(\hat{m}_1^r, B_{1,j}^r), \ldots, p_K\sum_{j=1}^{B}\nabla F(\hat{m}_K^r, B_{K,j}^r)].$$

For simplicity, we use $\mathbf{1}$ to represent $\mathbf{1}_K$. Then, the detailed proofs are as follows:

Proof.

$$\begin{aligned}
&\mathbb{E}[f(\frac{M^{r+1}\mathbf{1}}{K}) - f(\frac{M^r\mathbf{1}}{K})]\\
&\overset{(1)}{\leq} -\eta\mathbb{E}[\langle\nabla f(\frac{M^r\mathbf{1}}{K}), \partial g(\hat{M}^r, k^r)\frac{\mathbf{1}}{K}\rangle] + \frac{L\eta^2}{2}\mathbb{E}[\|\partial g(\hat{M}^r, k^r)\frac{\mathbf{1}_K}{K}\|]\\
&\overset{(2)}{\leq} -\frac{B\eta}{K}\mathbb{E}[\langle\nabla f(\frac{M^r\mathbf{1}}{K}), \partial f(\hat{M}^r)\frac{\mathbf{1}}{K}\rangle] + \frac{LB\eta^2\alpha^2}{2K^2}\\
&\quad + \frac{LB^2\eta^2}{2K^2}\sum_{k=1}^{K}p_k\mathbb{E}[\|\nabla f(\hat{m}_k^r)\|^2]\\
&\overset{(3)}{\leq} \frac{B\eta}{2K}\mathbb{E}[\|\nabla f(\frac{M^r\mathbf{1}}{K}) - \partial f(\hat{M}^r)\frac{\mathbf{1}}{K}\|^2] - \frac{B\eta}{2K}\mathbb{E}[\|\nabla f(M^r\frac{\mathbf{1}}{K})\|^2]\\
&\quad - (\frac{B\eta}{2K} - \frac{LB^2\eta^2}{K^2})\mathbb{E}[\|\partial f(\hat{M}^r)\frac{\mathbf{1}}{K}\|^2] + \frac{L\eta^2(\alpha^2B+6\beta^2B^2)}{2K^2}\\
&\quad + \frac{6L^3B^2\eta^2}{K^2}\mathbb{E}[\hat{\varepsilon}^r]\\
&\overset{(4)}{\leq} \frac{L^2B\eta}{K}\mathbb{E}[\|(M^r-\hat{M}^r)\frac{\mathbf{1}}{K}\|^2] - (\frac{B\eta}{2K} - \frac{LB^2\eta^2}{K^2})\mathbb{E}[\|\partial f(\hat{M}^r)\frac{\mathbf{1}}{K}\|^2]\\
&\quad + (\frac{L^2B\eta}{K} + \frac{6L^3B^2\eta^2}{K^2})\mathbb{E}[\hat{\varepsilon}] - \frac{B\eta}{2K}\mathbb{E}[\|\nabla f(M^r\frac{\mathbf{1}}{K})\|^2]\\
&\quad + \frac{L\eta^2(\alpha^2B+6\beta^2B^2)}{2K^2} + \frac{6L^3B^2\eta^2}{K^2}\mathbb{E}[\hat{\varepsilon}^r]\\
&\overset{(5)}{\leq} (\frac{L^2B\eta}{K} + \frac{6L^3B^2\eta^2}{K^2})\mathbb{E}[\hat{\varepsilon}] + \frac{L\eta^2(\alpha^2B+6\beta^2B^2)}{2K^2}\\
&\quad + \frac{L^2B\eta}{K}(\frac{B(\tau^r)^2\eta^2\alpha^2}{K^2} + \frac{B\tau^r\eta^2}{K^2}\sum_{t=1}^{\tau^r}\sum_{k=1}^{K}p_k\mathbb{E}[\|\nabla f_k(\hat{m}_k^{r-t})\|^2])\\
&\quad - (\frac{B\eta}{2K} - \frac{LB^2\eta^2}{K^2})\mathbb{E}[\|\partial f(\hat{M}^r)\frac{\mathbf{1}}{K}\|^2] - \frac{B\eta}{2K}\mathbb{E}[\|\nabla f(M^r\frac{\mathbf{1}}{K})\|^2],
\end{aligned}$$

where (1), (2) are from Assumption 1.1 and Lemma 4 presented in [2], respectively. We use Lemma 5 in [2] to get (3). By Jensen's inequality, we can derive the result in (4). We then use Lemma 8 in [2] to bound the first term of (4) to directly derive (5).

From Assumption 1.5 and Lemma 5 in [2], the following inequality can be derived:

$$\begin{aligned}
&\mathbb{E}[f(\frac{M^{r+1}\mathbf{1}}{K}) - f(\frac{M^r\mathbf{1}}{K})]\\
&\leq (\frac{L^2B\eta}{K} + \frac{6L^3B^2\eta^2}{K^2})\mathbb{E}[\hat{\varepsilon}] + \frac{L\eta^2(\alpha^2B+6\beta^2B^2)}{2K^2}\\
&\quad + \frac{L^2B^3\eta^3\tau^r}{K^3}\sum_{t=1}^{\tau^r}(12L^2\hat{\varepsilon}^{r-t} + 6\beta^2 + 2\sum_{k=1}^{K}p_k\mathbb{E}[\|\frac{\partial f_k(\hat{M}_k^{r-t})\mathbf{1}}{K}\|^2])\\
&\quad + \frac{L^2\gamma^2B^2\alpha^2\eta^3}{K^3} - (\frac{B\eta}{2K} - \frac{LB^2\eta^2}{K^2})\mathbb{E}[\|\partial f(\hat{M}^r)\frac{\mathbf{1}}{K}\|^2]\\
&\quad - \frac{B\eta}{2K}\mathbb{E}[\|\nabla f(M^r\frac{\mathbf{1}}{K})\|^2].
\end{aligned} \tag{14}$$

We then rearrange (14) and sum over r from 0 to $R-1$ to

derive the following inequality

$$\begin{aligned}
&\mathbb{E}[f(\frac{M^r\mathbf{1}}{K}) - f(\frac{M^0\mathbf{1}}{K})] \\
&\leq (\frac{L^2B\eta}{K} + \frac{6L^3B^2\eta^2}{K^2}) \sum_{r=0}^{R-1} \mathbb{E}[\hat{\varepsilon}] + \frac{LR\eta^2(\alpha^2 B + 6\beta^2 B^2)}{2K^2} \\
&\quad + \frac{2L^2B^3\eta^3\Gamma^2}{K^3} \sum_{r=0}^{R-1} (6L^2\hat{\varepsilon}^r + \mathbb{E}[\|\frac{\partial f(\hat{M}^r)\mathbf{1}}{K}\|^2]) \\
&\quad + \frac{L^2\Gamma^2\eta^3BR(\alpha^2B + 6\beta^2B^2)}{K^3} - \frac{B\eta}{2K} \sum_{r=0}^{R-1} \mathbb{E}[\|\nabla f(M^r\frac{\mathbf{1}}{K})\|^2] \\
&\quad - (\frac{B\eta}{2K} - \frac{LB^2\eta^2}{K^2}) \sum_{r=0}^{R-1} \mathbb{E}[\|\partial f(\hat{M}^r)\frac{\mathbf{1}}{K}\|^2].
\end{aligned} \tag{15}$$

Rearrange (15), we have

$$\begin{aligned}
&\mathbb{E}[f(\frac{M^r\mathbf{1}}{K}) - f(\frac{M^0\mathbf{1}}{K})] \\
&\leq \Psi_1 \sum_{r=0}^{R-1} \mathbb{E}[\hat{\varepsilon}] + \Psi_2 + \Psi_3 \sum_{r=0}^{R-1} \mathbb{E}[\|\frac{\partial f(\hat{M}^r)\mathbf{1}}{K}\|^2] \\
&\quad - \frac{B\eta}{2K} \sum_{r=0}^{R-1} \mathbb{E}[\|\nabla f(M^r\frac{\mathbf{1}}{K})\|^2],
\end{aligned} \tag{16}$$

where

$$\begin{aligned}
\Psi_1 &= \frac{L^2B\eta}{K}(1 + \frac{6LB\eta}{K} + \frac{12L^2B^2\eta^2\Gamma^2}{K^2}), \\
\Psi_2 &= \frac{LBR\eta^2(\alpha^2 + 6\beta^2B)}{K^2}(\frac{1}{2} + \frac{LB\eta\Gamma^2}{K}), \\
\Psi_3 &= \frac{B\eta}{K}(\frac{LB\eta}{K} + \frac{2L^2B^2\eta^2\Gamma^2}{K^2} - \frac{1}{2}).
\end{aligned}$$

According to the lemma 7 in [2], the first term of (16) is bound, that is,

$$\begin{aligned}
\sum_{r=0}^{R-1} \mathbb{E}[\|\hat{\varepsilon}\|^2] \leq \frac{1}{\lambda}(4B^2\eta^2(\frac{K-1}{K}\Gamma + \bar{\kappa}) \sum_{r=0}^{R-1} \mathbb{E}[\|\sum_{k=1}^{K} p_k \nabla f_k(\hat{m}_k^r)\|^2] \\
+ 2\bar{\kappa}BR\eta^2(\alpha^2 + 6\beta^2B)).
\end{aligned} \tag{17}$$

Substituting it into (16) and define $\lambda = 1 - 24L^2B^2\eta^2(\frac{K-1}{K}\Gamma + \bar{\kappa})$, we have

$$\begin{aligned}
&\mathbb{E}[f(\frac{M^r\mathbf{1}}{K}) - f(\frac{M^0\mathbf{1}}{K})] \\
&\leq \frac{\Psi_1}{\lambda}(4B^2\eta^2(\frac{K-1}{K}\Gamma + \bar{\kappa}) \sum_{r=0}^{R-1} \mathbb{E}[\|\sum_{k=1}^{K} p_k \nabla f_k(\hat{m}_k^r)\|^2] \\
&\quad + 2\bar{\kappa}BR\eta^2(\alpha^2 + 6\beta^2B)) + \Psi_2 + \Psi_3 \sum_{r=0}^{R-1} \mathbb{E}[\|\frac{\partial f(\hat{M}^r)\mathbf{1}}{K}\|^2] \\
&\quad - \frac{B\eta}{2K} \sum_{r=0}^{R-1} \mathbb{E}[\|\nabla f(M^r\frac{\mathbf{1}}{K})\|^2] \\
&\leq \underbrace{(\Psi_3 + \frac{4\Psi_1B^2\eta^2(\frac{K-1}{K}\Gamma + \bar{\kappa})}{\lambda})}_{\mathcal{A}} \sum_{r=0}^{R-1} \mathbb{E}[\|\frac{\partial f(\hat{M}^r)\mathbf{1}}{K}\|^2] \\
&\quad + \frac{2\bar{\kappa}\Psi_1R\eta^2(\alpha^2 + \beta^2B)}{\lambda} + \Psi_2 - \frac{B\eta}{2K} \sum_{r=0}^{R-1} \mathbb{E}[\|\nabla f(M^r\frac{\mathbf{1}}{K})\|^2],
\end{aligned} \tag{18}$$

where $\mathcal{A} \leq 0$ if we set

$$\begin{aligned}
&\frac{L\eta(1 + 2LB\eta\Gamma^2)}{K} + \frac{4B^2\eta^2(6L^3K\eta + 12L^4B^2\eta^2\Gamma^2)(\frac{K-1}{K}\Gamma + \bar{\kappa})}{\lambda K^2} \\
&\quad + L^2 \leq \frac{1}{2}.
\end{aligned}$$

Then, (18) can be rewritten as

$$\begin{aligned}
&\mathbb{E}[f(\frac{M^r\mathbf{1}}{K}) - f(\frac{M^0\mathbf{1}}{K})] \\
&\leq \underbrace{\frac{2\bar{\kappa}\Psi_1R\eta^2(\alpha^2 + \beta^2B)}{\lambda} + \Psi_2}_{\mathcal{B}} - \frac{B\eta}{2K} \sum_{r=0}^{R-1} \mathbb{E}[\|\nabla f(M^r\frac{\mathbf{1}}{K})\|^2],
\end{aligned} \tag{19}$$

where $\mathcal{B} \leq 1$ if we set

$$\frac{KLB\eta(6LB\eta + K) + 12L^3B^3\eta^3\Gamma^2}{K - 24((K-1)\Gamma + K\bar{\kappa})L^2B^2\eta^2}\bar{\kappa} + \frac{LB\eta\Gamma^2}{2K} \leq \frac{1}{4}.$$

Finally, we can rearrange (19) by dividing R to derive the result.

□

C. Proof of Theorem 2

We first provide key lemmas that are crucial for the proof of Theorem 2.

Lemma 1. *Let Assumptions 1.3 and 2 hold, we have*

$$\mathbb{E}[\|\frac{\partial\tilde{g}(\hat{M}^r, k_r)\mathbf{1}}{K}\|^2] \leq \frac{\alpha^2B + \sigma^2}{K^2} + \frac{B^2}{K^2} \sum_{k=1}^{K} p_k \mathbb{E}[\|\nabla f_k(\hat{m}_k^r)\|^2], \forall r \geq 0.$$

Proof.

$$\begin{aligned}
\mathbb{E}[\|\frac{\partial\tilde{g}(\hat{M}^r, k_r)\mathbf{1}}{K}\|^2] &= \mathbb{E}[\|\frac{(\partial g(\hat{M}^r, k_r) + \delta)\mathbf{1}}{K}\|^2] \\
&= \sum_{k=1}^{K} p_k \mathbb{E}[\|\frac{\sum_{j=1}^{B} \nabla F(\hat{m}_j^r, \mathcal{B}_{k,j}^r) + \delta}{K}\|^2] \\
&= \sum_{k=1}^{K} p_k \mathbb{E}[\|\frac{\sum_{j=1}^{B} \nabla F(\hat{m}_j^r, \mathcal{B}_{k,j}^r)}{K}\|^2] + \frac{\mathbb{E}[\|\delta\|^2]}{K^2} \\
&\leq \frac{\alpha^2B}{K^2} + \frac{B^2}{K^2} \sum_{k=1}^{K} p_k \mathbb{E}[\|\nabla f_k(\hat{m}_k^r)\|^2] + \frac{\sigma^2}{K^2}
\end{aligned}$$

□

Lemma 2. *Let Assumption 2 hold, for any $r \geq -1$ we have*

$$\begin{aligned}
&\mathbb{E}[\|\frac{M^{r+1}\mathbf{1}}{K} - M^{r+1}e_k\|^2] \leq 2\eta^2(\alpha^2B + 6\beta^2B^2)\bar{\kappa} \\
&+ \frac{K-1}{K}B^2\eta^2 \sum_{j=0}^{r} \mathbb{E}[(24L^2\hat{\varepsilon}^j + 4\mathbb{E}[\|\sum_{k=1}^{K} p_k \nabla f_k(\hat{m}_k^r)\|^2])(\kappa^{r-j} \\
&+ 2(r-j)\kappa^{\frac{r-j}{2}})] + 2\eta^2\sigma^2(\frac{1}{K^2} + 1).
\end{aligned}$$

Proof. We first have

$$\mathbb{E}[\|\frac{M^{r+1}\mathbf{1}}{K} - M^{r+1}e_k\|^2] = 0, \quad r = -1.$$

Note that $M^0W^r = M^0$ for all r and $M^0\mathbf{1}/n - M^0e_k = 0$.

When $r \geq 0$, we have

$$\begin{aligned}
&\mathbb{E}[\|\frac{M^{r+1}\mathbf{1}}{K} - M^{r+1}e_k\|^2] \\
&= \mathbb{E}[\|\frac{M^r\mathbf{1} - \eta\partial\tilde{g}(\hat{M}^r,k^r)\mathbf{1}}{K} - (M^rW^r - \eta\partial\tilde{g}(\hat{M}^r,k^r))e_k\|^2] \\
&= \mathbb{E}[\|\frac{M^r\mathbf{1} - \eta\partial g(\hat{M}^r,k^r)\mathbf{1}}{K} - (M^rW^r - \eta\partial g(\hat{M}^r,k^r))e_k\|^2] \\
&\quad + \mathbb{E}[\|\frac{\eta\delta\mathbf{1}}{K} + \eta\delta e_k\|^2] \\
&\leq 2\eta^2(\alpha^2B + 6\beta^2B^2)\bar{\kappa} \\
&+ \frac{K-1}{K}B^2\eta^2\sum_{j=0}^{r}\mathbb{E}[(24L^2\hat{\varepsilon}^j + 2\mathbb{E}[\|\sum_{k=1}^{K}p_k\nabla f_k(\hat{m}_k^r)\|^2])(\kappa^{r-j} \\
&+ 2(r-j)\kappa^{\frac{r-j}{2}})] + 2\eta^2\sigma^2(\frac{1}{K^2}+1).
\end{aligned}$$

□

Lemma 3. *Let Assumptions 1.1-1.5 and 2 hold, from Lemma 2, we have*

$$\begin{aligned}
\frac{1}{R}\sum_{r=0}^{R-1}\mathbb{E}[\hat{\varepsilon}^r] \leq \frac{1}{\lambda}\big(&2\bar{\kappa}\eta^2(\alpha^2B + 6\beta^2B^2) + \frac{4B^2\eta^2}{R}(\frac{K-1}{K}\Gamma \\
&+ \bar{\kappa})\sum_{r=0}^{R-1}\mathbb{E}[\|\sum_{k=1}^{K}p_k\nabla f_k(\hat{m}_k^r)\|^2] + 2\eta^2\sigma^2(\frac{1}{K^2}+1)\big)
\end{aligned}$$

Proof. The result is derived according to the proof in [2]. The difference is that we need to reconsider the effect of noise added in the local gradient. □

Lemma 4. *Let Assumptions 1.5 and 2 hold, we have*

$$\begin{aligned}
\mathbb{E}[\|\frac{M^r\mathbf{1} - \hat{M}^r\mathbf{1}}{K}\|^2] \leq \frac{B^2\eta^2\tau^r}{K^2}&\sum_{t=1}^{\tau^r}\sum_{k=1}^{K}p_k\mathbb{E}[\|\nabla f_k(\hat{m_k^{r-t}})\|^2] \\
&+ \frac{\alpha^2B\eta^2(\tau^r)^2}{K^2} + \frac{\eta^2\sigma^2\Gamma^2}{K^2}.
\end{aligned}$$

Proof.

$$\begin{aligned}
&\mathbb{E}[\|\frac{M^r\mathbf{1} - \hat{M}^r\mathbf{1}}{K}\|^2] = \mathbb{E}[\|\frac{\eta\sum_{t=1}^{\tau^r}\partial\tilde{g}(\hat{M}^{r-t},k^{r-t})\mathbf{1}}{K}\|^2] \\
&= \mathbb{E}[\|\frac{\eta\sum_{t=1}^{\tau^r}(\partial g(\hat{M}^{r-t},k^{r-t}) + \delta)\mathbf{1}}{K}\|^2] \\
&= \mathbb{E}[\|\frac{\eta\sum_{t=1}^{\tau^r}\partial g(\hat{M}^{r-t},k^{r-t})\mathbf{1}}{K}\|^2] + \mathbb{E}[\|\frac{\eta\sum_{t=1}^{\tau^r}\delta\mathbf{1}}{K}\|^2] \\
&\leq \frac{B^2\eta^2\tau^r}{K^2}\sum_{t=1}^{\tau^r}\sum_{k=1}^{K}p_k\mathbb{E}[\|\nabla f_k(\hat{m_k^{r-t}})\|^2] \\
&\quad + \frac{\alpha^2B\eta^2(\tau^r)^2}{K^2} + \frac{\eta^2\sigma^2\Gamma^2}{K^2},
\end{aligned}$$

where the last inequality is from Lemma 1, Assumption 1.5 and 2. □

Based on these lemmas, we now proceed to formally present the detailed proof of Theorem 2.

Proof.

$$\begin{aligned}
&\mathbb{E}[f(\frac{M^{r+1}\mathbf{1}}{K}) - f(\frac{M^r\mathbf{1}}{K})] \\
&\overset{(1)}{\leq} -\eta\mathbb{E}[\langle\nabla f(\frac{M^r\mathbf{1}}{K}), \partial\tilde{g}(\hat{M}^r,k^r)\frac{\mathbf{1}}{K}\rangle] + \frac{L\eta^2}{2}\mathbb{E}[\|\partial\tilde{g}(\hat{M}^r,k^r)\frac{\mathbf{1}}{K}\|] \\
&\overset{(2)}{\leq} -\frac{B\eta}{K}\mathbb{E}[\langle\nabla f(\frac{M^r\mathbf{1}}{K}), (\partial f(\hat{M}^r) + \delta)\frac{\mathbf{1}}{K}\rangle] + \frac{L\eta^2(\alpha^2B + \sigma^2)}{2K^2} \\
&\quad + \frac{LB^2\eta^2}{2K^2}\sum_{k=1}^{K}p_k\mathbb{E}[\|\nabla f_k(\hat{m}_k^r)\|^2] \\
&\overset{(3)}{=} \frac{B\eta}{2K}\mathbb{E}[\|\nabla f(\frac{M^r\mathbf{1}}{K}) - \partial f(\hat{M}^r))\frac{\mathbf{1}}{K}\|^2] - \frac{B\eta}{2K}\mathbb{E}[\|\nabla f(\frac{M^r\mathbf{1}}{K})\|^2] \\
&\quad - \frac{B\eta}{2K}\mathbb{E}[\|\partial f(\hat{M}^r)\frac{\mathbf{1}}{K}\|^2] + \frac{LB^2\eta^2}{2K^2}\sum_{k=1}^{K}p_k\mathbb{E}[\|\nabla f_k(\hat{m}_k^r)\|^2] \\
&\quad + \frac{L\eta^2(\alpha^2B + \sigma^2)}{2K^2} \\
&\overset{(4)}{\leq} \frac{B\eta}{2K}\mathbb{E}[\|\nabla f(\frac{M^r\mathbf{1}}{K}) - \partial f(\hat{M}^r))\frac{\mathbf{1}}{K}\|^2] - \frac{B\eta}{2K}\mathbb{E}[\|\nabla f(\frac{M^r\mathbf{1}}{K})\|^2] \\
&\quad - (\frac{B\eta}{2K} - \frac{LB^2\eta^2}{K^2})\mathbb{E}[\|\partial f(\hat{M}^r)\frac{\mathbf{1}}{K}\|^2] + \frac{6L^3B^2\eta^2}{K^2}\mathbb{E}[\hat{\varepsilon}^r] \\
&\quad + \frac{L\eta^2(\alpha^2B + \sigma^2 + 6\beta^2B^2)}{2K^2} \\
&\overset{(5)}{\leq} \frac{L^2B\eta}{K}\mathbb{E}[\|(M^r - \hat{M}^r)\frac{\mathbf{1}}{K}\|^2] + (\frac{L^2B\eta}{K} + \frac{6L^3B^2\eta^2}{K^2})\mathbb{E}[\hat{\varepsilon}] \\
&\quad - (\frac{B\eta}{2K} - \frac{LB^2\eta^2}{K^2})\mathbb{E}[\|\partial f(\hat{M}^r)\frac{\mathbf{1}}{K}\|^2] - \frac{B\eta}{2K}\mathbb{E}[\|\nabla f(M^r\frac{\mathbf{1}}{K})\|^2] \\
&\quad + \frac{L\eta^2(\alpha^2B + 6\beta^2B^2 + \sigma^2)}{2K^2} + \frac{6L^3B^2\eta^2}{K^2}\mathbb{E}[\hat{\varepsilon}^r] \\
&\overset{(6)}{\leq} (\frac{L^2B\eta}{K} + \frac{6L^3B^2\eta^2}{K^2})\mathbb{E}[\hat{\varepsilon}] + \frac{L\eta^2(\alpha^2B + 6\beta^2B^2 + \sigma^2)}{2K^2} \\
&\quad + \frac{L^2B\eta}{K}(\frac{B(\tau^r)^2\eta^2\alpha^2}{K^2} + \frac{B\tau^r\eta^2}{K^2}\sum_{t=1}^{\tau^r}\sum_{k=1}^{K}p_k\mathbb{E}[\|\nabla f_k(\hat{m}_k^{r-t})\|^2]) \\
&\quad - (\frac{B\eta}{2K} - \frac{LB^2\eta^2}{K^2})\mathbb{E}[\|\partial f(\hat{M}^r)\frac{\mathbf{1}}{K}\|^2] - \frac{B\eta}{2K}\mathbb{E}[\|\nabla f(M^r\frac{\mathbf{1}}{K})\|^2] \\
&\quad + \frac{L^2B\eta^3\sigma^2\Gamma^2}{K^3},
\end{aligned}$$

where (1) is from Assumption 1.1, (2) is from Assumption 2 and Lemma 1, and (3) is due to $-<a,b>=\frac{\|a+b\|^2}{2} - \frac{\|a\|^2}{2} - \frac{\|b\|^2}{2}$. We use Lemma 5 in [2] to get (4). By Jensen's inequality, we can derive the result in (5). We then use Lemma 4 to bound the first term of (5) to directly derive (6).

From Assumption 1.1 and Lemma 5 in [2], we have

$$\begin{aligned}
&\mathbb{E}[f(\frac{M^{r+1}\mathbf{1}}{K}) - f(\frac{M^r\mathbf{1}}{K})] \\
&\leq (\frac{L^2B\eta}{K} + \frac{6L^3B^2\eta^2}{K^2})\mathbb{E}[\hat{\varepsilon}] + \frac{L\eta^2(\alpha^2B + 6\beta^2B^2 + \sigma^2)}{2K^2} \\
&\quad + \frac{L^2B^3\eta^3\tau^r}{K^3}\sum_{t=1}^{\tau^r}(12L^2\hat{\varepsilon}^{r-t} + 6\beta^2 + 2\sum_{k=1}^{K}p_k\mathbb{E}[\|\frac{\partial f_k(\hat{M}_k^{r-t})\mathbf{1}}{K}\|^2]) \\
&\quad + \frac{L^2\gamma^2B^2\alpha^2\eta^3}{K^3} - (\frac{B\eta}{2K} - \frac{LB^2\eta^2}{K^2})\mathbb{E}[\|\partial f(\hat{M}^r)\frac{\mathbf{1}}{K}\|^2] \\
&\quad - \frac{B\eta}{2K}\mathbb{E}[\|\nabla f(M^r\frac{\mathbf{1}}{K})\|^2] + \frac{L^2B\eta^3\sigma^2\Gamma^2}{K^3}.
\end{aligned} \tag{20}$$

We then rearrange (20) and sum over r from 0 to $R-1$ to

derive the following inequality

$$\begin{aligned}
&\mathbb{E}[f(\frac{M^r\mathbf{1}}{K}) - f(\frac{M^0\mathbf{1}}{K})] \\
&\leq (\frac{L^2B\eta}{K} + \frac{6L^3B^2\eta^2}{K^2}) \sum_{r=0}^{R-1} \mathbb{E}[\hat{\varepsilon}] + \frac{LR\eta^2(\alpha^2B + 6\beta^2B^2 + \sigma^2)}{2K^2} \\
&\quad + \frac{2L^2B^3\eta^3\Gamma^2}{K^3} \sum_{r=0}^{R-1} (6L^2\hat{\varepsilon}^r + \mathbb{E}[\| \frac{\partial f(\hat{M}^r)\mathbf{1}}{K} \|^2]) \\
&\quad + \frac{L^2\Gamma^2\eta^3BR(\alpha^2B + 6\beta^2B^2 + \sigma^2)}{K^3} - \frac{B\eta}{2K} \sum_{r=0}^{R-1} \mathbb{E}[\|\nabla f(M^r \frac{\mathbf{1}}{K})\|^2] \\
&\quad - (\frac{B\eta}{2K} - \frac{LB^2\eta^2}{K^2}) \sum_{r=0}^{R-1} \mathbb{E}[\|\partial f(\hat{M}^r) \frac{\mathbf{1}}{K} \|^2].
\end{aligned} \tag{21}$$

Rearrange (21), we have

$$\begin{aligned}
&\mathbb{E}[f(\frac{M^r\mathbf{1}}{K}) - f(\frac{M^0\mathbf{1}}{K})] \\
&\leq \Psi_1 \sum_{r=0}^{R-1} \mathbb{E}[\hat{\varepsilon}] + \Psi_2^* + \Psi_3 \sum_{r=0}^{R-1} \mathbb{E}[\| \frac{\partial f(\hat{M}^r)\mathbf{1}}{K} \|^2] \\
&\quad - \frac{B\eta}{2K} \sum_{r=0}^{R-1} \mathbb{E}[\|\nabla f(M^r \frac{\mathbf{1}}{K})\|^2],
\end{aligned} \tag{22}$$

where Ψ_1, Ψ_3 are defined in Theorem 1, and

$$\Psi_2^* = \frac{LBR\eta^2(\alpha^2 + 6\beta^2B + \sigma^2)}{K^2} (\frac{1}{2} + \frac{LB\eta\Gamma^2}{K})$$

From Lemma 3, we can bound the first term in (22), that is,

$$\begin{aligned}
&\mathbb{E}[f(\frac{M^r\mathbf{1}}{K}) - f(\frac{M^0\mathbf{1}}{K})] \\
&\quad \leq \frac{\Psi_1}{\lambda} \big(2R\bar{\kappa}\eta^2(\alpha^2B + 6\beta^2B^2) + 4B^2\eta^2(\frac{K-1}{K}\Gamma \\
&\quad + \bar{\kappa}) \sum_{r=0}^{R-1} \mathbb{E}[\| \sum_{k=1}^{K} p_k \nabla f_k(\hat{m}_k^r) \|^2] + 2R\eta^2\sigma^2(\frac{1}{K^2} + 1)\big) \\
&\quad + \Psi_2^* + \Psi_3 \sum_{r=0}^{R-1} \mathbb{E}[\| \frac{\partial f(\hat{M}^r)\mathbf{1}}{K} \|^2] \\
&\quad - \frac{B\eta}{2K} \sum_{r=0}^{R-1} \mathbb{E}[\|\nabla f(M^r \frac{\mathbf{1}}{K})\|^2] \\
&\leq \underbrace{(\Psi_3 + \frac{4\Psi_1B^2\eta^2(\frac{K-1}{K}\Gamma + \bar{\kappa})}{\lambda})}_{\mathcal{A}} \sum_{r=0}^{R-1} \mathbb{E}[\| \frac{\partial f(\hat{M}^r)\mathbf{1}}{K} \|^2] \\
&\quad + \frac{2\Psi_1R\eta^2(\bar{\kappa}(\alpha^2B + 6\beta^2B^2) + \sigma^2(1/K^2 + 1))}{\lambda} \\
&\quad + \Psi_2^* - \frac{B\eta}{2K} \sum_{r=0}^{R-1} \mathbb{E}[\|\nabla f(M^r \frac{\mathbf{1}}{K})\|^2].
\end{aligned} \tag{23}$$

Let $\mathcal{A} \leq 0$, (18) can be rewritten as

$$\begin{aligned}
&\mathbb{E}[f(\frac{M^r\mathbf{1}}{K}) - f(\frac{M^0\mathbf{1}}{K})] \\
&\leq \underbrace{\frac{2\Psi_1R\eta^2(\bar{\kappa}(\alpha^2B + 6\beta^2B^2) + \sigma^2(\frac{1}{K^2} + 1))}{\lambda} + \Psi_2^*}_{\mathcal{I}} \\
&\quad - \frac{B\eta}{2K} \sum_{r=0}^{R-1} \mathbb{E}[\|\nabla f(M^r \frac{\mathbf{1}}{K})\|^2],
\end{aligned} \tag{24}$$

Let $\mathcal{I} \leq 1$, we finally have the result. □

REFERENCES

[1] A. Koloskova, N. Loizou, S. Boreiri, M. Jaggi, and S. Stich, "A unified theory of decentralized sgd with changing topology and local updates," in *International Conference on Machine Learning*. PMLR, 2020, pp. 5381–5393.

[2] X. Lian, W. Zhang, C. Zhang, and J. Liu, "Asynchronous decentralized parallel stochastic gradient descent," in *International Conference on Machine Learning*. PMLR, 2018, pp. 3043–3052.