Appendix A   PROOF OF THEOREM 1

Theorem 1. *(Non-convex objective, fixed step size, and non-fixed batch size) Suppose that FedAvg algorithm runs with a fixed learning rate $\eta_t = \eta$ satisfying*

$$\sum_{k=1}^{K}\left[\frac{(m_k-2)(m_k+1)}{2} - \frac{1}{L^2\eta^2} + \frac{p_k m_k}{L\eta}\right] \leq 0 \tag{1}$$

where $p_k = {}^{D_k}\!/_D, m_k = {}^{D_k E}\!/_{b_k}$, E is the number of epochs in each client, m_k means the number of local updates in one communication round. Then the expected average squared gradient norms of F satisfy the following bound for all $T \in \mathbb{N}$:

$$\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}\|\nabla F(\mathbf{w}_t)\|_2^2 \leq \frac{F(\mathbf{w}_1) - F^*}{T(G-A-C)} + \frac{L\eta^2}{2K(G-A-C)}\sum_{k=1}^{K}p_k^2\beta_k m_k$$

$$+ \frac{L^2\eta^3}{12K(G-A-C)}\sum_{k=1}^{K}p_k\beta_k(m_k-1)m_k(2m_k-1) \tag{2}$$

where $\beta_k = \sigma_k^2/b_k$, $G = \frac{L^2\eta^3}{4K}\sum_{k=1}^{K}m_k(m_k-1)$, $A = \frac{\eta}{2K}\sum_{k=1}^{K}(m_k+1)$, $C = \frac{L\eta^2}{2K}\sum_{k=1}^{K}m_k$.

Proof. To prove the convergence of *FedAvg* algorithm under non-fixed batch size, we firstly bound the update of one global round $F(\mathbf{w}_{t+1}) - F(\mathbf{w}_t)$ and then summarize all steps from 1 to T to achieve the overall convergence.

We denote $\mathbf{w}_t$ as the t-th global update in *FedAvg* algorithm and $\mathbf{w}_{t+\mu}^k$ as μ-th local update in client k. The $(t+1)$-th global average can be written as

$$\mathbf{w}_{t+1} = \frac{1}{K}\sum_{k=1}^{K}\frac{D_k}{D}\mathbf{w}_{t+m_k}^k = \mathbf{w}_t - \frac{1}{K}\sum_{k=1}^{K}\frac{D_k}{D}\sum_{\mu=0}^{m_k-1}\frac{\eta_\mu}{b_k}\sum_{i=1}^{b_k}\nabla F(\mathbf{w}_{t+\mu}^k; \xi_{i,\mu}^k) \tag{3}$$

The random variables $\xi_{i,\mu}^k$ are Non-IID. Let $p_k = {}^{D_k}\!/_D$. Based on the assumptions, the bound of one step is

$$F(\mathbf{w}_{t+1}) - F(\mathbf{w}_t) \leq \langle\nabla F(\mathbf{w}_t), \mathbf{w}_{t+1} - \mathbf{w}_t\rangle + \frac{L}{2}\|\mathbf{w}_{t+1} - \mathbf{w}_t\|_2^2$$

$$= -\left\langle\nabla F(\mathbf{w}_t), \frac{1}{K}\sum_{k=1}^{K}p_k\sum_{\mu=0}^{m_k-1}\frac{\eta_\mu}{b_k}\sum_{i=1}^{b_k}\nabla F(\mathbf{w}_{t+\mu}^k; \xi_{i,\mu}^k)\right\rangle$$

$$+ \frac{L}{2}\|\frac{1}{K}\sum_{k=1}^{K}p_k\sum_{\mu=0}^{m_k-1}\frac{\eta_\mu}{b_k}\sum_{i=1}^{b_k}\nabla F(\mathbf{w}_{t+\mu}^k; \xi_{i,\mu}^k)\|_2^2 \tag{4}$$

Under the assumption that a constant step size is implemented within each inner parallel step, i.e., $\eta_\mu = \eta$, for $\mu = \{0, \cdots, m_k\}$, so the above inequality can be rewritten as

$$F(\mathbf{w}_{t+1}) - F(\mathbf{w}_t) \leq -\frac{\eta}{K}\sum_{k=1}^{K}p_k\sum_{\mu=0}^{m_k-1}\frac{1}{b_k}\sum_{i=1}^{b_k}\langle\nabla F(\mathbf{w}_t), \nabla F(\mathbf{w}_{t+\mu}^k; \xi_{i,\mu}^k)\rangle$$

$$+ \frac{L\eta^2}{2K^2}\|\sum_{k=1}^{K}p_k\sum_{\mu=0}^{m_k-1}\frac{1}{b_k}\sum_{i=1}^{b_k}\nabla F(\mathbf{w}_{t+\mu}^k; \xi_{i,\mu}^k)\|_2^2 \tag{5}$$

The goal here is to investigate the expectation of $F(\mathbf{w}_{t+1}) - F(\mathbf{w}_t)$ over all random variables $\xi_{i,\mu}^k$. Under the unbiased estimator assumption 3, by taking the overall expectation we can immediately get

$$\mathbb{E}\left[\frac{1}{b_k}\sum_{i=1}^{b_k}\nabla F(\mathbf{w}_{t+\mu}^k; \xi_{i,\mu}^k)\right] = \mathbb{E}\left[\frac{1}{b_k}\sum_{i=1}^{b_k}\mathbb{E}_{\xi_{i,\mu}^k}\left[\nabla F(\mathbf{w}_{t+\mu}^k; \xi_{i,\mu}^k | \mathbf{w}_{t+\mu}^k)\right]\right]$$

$$= \mathbb{E}\left[\frac{1}{b_k}\sum_{i=1}^{b_k}\nabla F(\mathbf{w}_{t+\mu}^k)\right]$$

$$= \mathbb{E}\nabla F(\mathbf{w}_{t+\mu}^k) \tag{6}$$

By taking the overall expectation on both side of (5), we have

$$
\mathbb{E}F(\mathbf{w}_{t+1}) - F(\mathbf{w}_t) \leq - \frac{\eta}{K} \sum_{k=1}^{K} p_k \sum_{\mu=0}^{m_k-1} \frac{1}{b_k} \sum_{i=1}^{b_k} \mathbb{E} \left\langle \nabla F(\mathbf{w}_t), \nabla F(\mathbf{w}_{t+\mu}^k; \xi_{i,\mu}^k) \right\rangle
$$

$$
+ \frac{L\eta^2}{2K^2} \mathbb{E} \| \sum_{k=1}^{K} p_k \sum_{\mu=0}^{m_k-1} \frac{1}{b_k} \sum_{i=1}^{b_k} \nabla F(\mathbf{w}_{t+\mu}^k; \xi_{i,\mu}^k) \|_2^2
$$

$$
= \underbrace{- \frac{\eta}{K} \sum_{k=1}^{K} p_k \sum_{\mu=0}^{m_k-1} \mathbb{E} \langle \nabla F(\mathbf{w}_t), \nabla F(\mathbf{w}_{t+\mu}^k) \rangle}_{T_1}
$$

$$
+ \underbrace{\frac{L\eta^2}{2K^2} \mathbb{E} \| \sum_{k=1}^{K} p_k \sum_{\mu=0}^{m_k-1} \frac{1}{b_k} \sum_{i=1}^{b_k} \nabla F(\mathbf{w}_{t+\mu}^k; \xi_{i,\mu}^k) \|_2^2}_{T_2} \tag{7}
$$

Then, the two bounds T_1 and T_2 are derived respectively.
Bound T_1:

$$
T_1 = - \frac{\eta}{K} \sum_{k=1}^{K} p_k \sum_{\mu=0}^{m_k-1} \mathbb{E} \langle \nabla F(\mathbf{w}_t), \nabla F(\mathbf{w}_{t+\mu}^k) \rangle
$$

$$
= - \frac{\eta}{2K} \sum_{k=1}^{K} p_k \sum_{\mu=0}^{m_k-1} \left(\|\nabla F(\mathbf{w}_t)\|_2^2 + \mathbb{E}\|\nabla F(\mathbf{w}_{t+\mu}^k)\|_2^2 \right) + \frac{\eta}{2K} \sum_{k=1}^{K} p_k \sum_{\mu=0}^{m_k-1} \mathbb{E}\|\nabla F(\mathbf{w}_{t+\mu}^k) - \nabla F(\mathbf{w}_t)\|_2^2
$$

$$
= - \frac{\eta}{2K} \sum_{k=1}^{K} p_k (m_k+1) \|\nabla F(\mathbf{w}_t)\|_2^2 - \frac{\eta}{2K} \sum_{k=1}^{K} p_k \sum_{\mu=1}^{m_k-1} \mathbb{E}\|\nabla F(\mathbf{w}_{t+\mu}^k)\|_2^2
$$

$$
+ \frac{\eta}{2K} \sum_{k=1}^{K} p_k \sum_{\mu=0}^{m_k-1} \mathbb{E}\|\nabla F(\mathbf{w}_{t+\mu}^k) - \nabla F(\mathbf{w}_t)\|_2^2
$$

$$
\leq - \frac{\eta}{2K} \sum_{k=1}^{K} p_k (m_k+1) \|\nabla F(\mathbf{w}_t)\|_2^2 - \frac{\eta}{2K} \sum_{k=1}^{K} p_k \sum_{\mu=1}^{m_k-1} \mathbb{E}\|\nabla F(\mathbf{w}_{t+\mu}^k)\|_2^2
$$

$$
+ \frac{L^2\eta}{2K} \sum_{k=1}^{K} p_k \sum_{\mu=1}^{m_k-1} \underbrace{\mathbb{E}\|\mathbf{w}_{t+\mu}^k - \mathbf{w}_t\|_2^2}_{T_3} \tag{8}
$$

where we use the fact that $\mathbf{w}_{t+\mu}^k = \mathbf{w}_t$, for $k = \{1, \cdots, K\}$ in the second equality. The last inequality is due to the assumption 1. Note that

$$
T_3 = \mathbb{E}\|\mathbf{w}_{t+\mu}^k - \mathbf{w}_t\|_2^2
$$

$$
= \mathbb{E}\| \sum_{s=0}^{\mu-1} \frac{\eta_\mu}{b_k} \sum_{i=1}^{b_k} \nabla F(\mathbf{w}_{t+s}^k; \xi_{i,s}^k) \|_2^2
$$

$$
= \eta^2 \mathbb{E}\| \sum_{s=0}^{\mu-1} \frac{1}{b_k} \sum_{i=1}^{b_k} \nabla F(\mathbf{w}_{t+s}^k; \xi_{i,s}^k) \|_2^2 \leq \mu\eta^2 \mathbb{E} \sum_{s=0}^{\mu-1} \| \frac{1}{b_k} \sum_{i=1}^{b_k} \nabla F(\mathbf{w}_{t+s}^k; \xi_{i,s}^k) \|
$$

$$
= \frac{\mu\eta^2}{b_k^2} \mathbb{E} \sum_{s=0}^{\mu-1} \| \sum_{i=1}^{b_k} \left(\nabla F(\mathbf{w}_{t+s}^k; \xi_{i,s}^k) - \nabla F(\mathbf{w}_{t+s}^k) + \nabla F(\mathbf{w}_{t+s}^k) \right) \|
$$

$$
= \frac{\mu\eta^2}{b_k^2} \mathbb{E} \sum_{s=0}^{\mu-1} \| \sum_{i=1}^{b_k} \left(\nabla F(\mathbf{w}_{t+s}^k; \xi_{i,s}^k) - \nabla F(\mathbf{w}_{t+s}^k) \right) \|_2^2 + \mu\eta^2 \mathbb{E} \sum_{s=0}^{\mu-1} \|\nabla F(\mathbf{w}_{t+s}^k)\|_2^2
$$

$$+ \frac{\mu\eta^2}{b_k^2} 2\mathbb{E} \sum_{s=0}^{\mu-1} \mathbb{E} \left\langle \sum_{i=1}^{b_k} \left(\nabla F(\mathbf{w}_{t+s}^k; \xi_{i,s}^k) - \nabla F(\mathbf{w}_{t+s}^k) \right), b_k \nabla F(\mathbf{w}_{t+s}^k) \right\rangle$$

$$= \frac{\mu\eta^2}{b_k^2} \mathbb{E} \sum_{s=0}^{\mu-1} \sum_{i=1}^{b_k} \|\nabla F(\mathbf{w}_{t+s}^k; \xi_{i,s}^k) - \nabla F(\mathbf{w}_{t+s}^k)\|_2^2 + \mu\eta^2 \mathbb{E} \sum_{s=0}^{\mu-1} \|\nabla F(\mathbf{w}_{t+s}^k)\|_2^2$$

$$\leq \frac{\mu^2\eta^2\sigma_k^2}{b_k} + \mu\eta^2 \mathbb{E} \sum_{s=0}^{\mu-1} \|\nabla F(\mathbf{w}_{t+s}^k)\|_2^2 \tag{9}$$

where the first inequality is due to Cauchy-Schwartz inequality. The last equality is due to the fact that if $a_1, a_2, \cdots, a_n$ are i.i.d. and $\mathbb{E}a_i = 0$, for any $i = 1, \cdots, n$, then

$$Var\left(\sum_{i=1}^{n} a_i\right) = \mathbb{E}\|\sum_{i=1}^{n} a_i\|_2^2$$

$$= \sum_{i=1}^{n} Var(a_i)$$

$$= \sum_{i=1}^{n} \mathbb{E}\|a_i\|_2^2 \tag{10}$$

and the last inequality is because of the assumption 4. We plug the above results back into (8) and get

$$T_1 \leq -\frac{\eta}{2K} \sum_{k=1}^{K} p_k(m_k + 1)\|\nabla F(\mathbf{w}_t)\|_2^2 - \frac{\eta}{2K} \sum_{k=1}^{K} p_k \sum_{\mu=1}^{m_k-1} \mathbb{E}\|\nabla F(\mathbf{w}_{t+\mu}^k)\|_2^2$$

$$+ \frac{L^2\eta}{2K} \sum_{k=1}^{K} p_k \sum_{\mu=1}^{m_k-1} \mathbb{E}\|\mathbf{w}_{t+\mu}^k - \mathbf{w}_t\|_2^2$$

$$\leq -\frac{\eta}{2K} \sum_{k=1}^{K} p_k(m_k + 1)\|\nabla F(\mathbf{w}_t)\|_2^2 - \frac{\eta}{2K} \sum_{k=1}^{K} p_k \sum_{\mu=1}^{m_k-1} \mathbb{E}\|\nabla F(\mathbf{w}_{t+\mu}^k)\|_2^2$$

$$+ \frac{L^2\eta}{2K} \sum_{k=1}^{K} p_k \sum_{\mu=1}^{m_k-1} \frac{\mu^2\eta^2\sigma_k^2}{b_k} + \frac{L^2\eta^3}{2K} \sum_{k=1}^{K} p_k \underbrace{\sum_{\mu=1}^{m_k-1} \mu\mathbb{E} \sum_{s=0}^{\mu-1} \|\nabla F(\mathbf{w}_{t+s}^k)\|_2^2}_{T_4} \tag{11}$$

Bound T_4:

$$T_4 = \sum_{\mu=1}^{m_k-1} \mu\mathbb{E} \sum_{s=0}^{\mu-1} \|\nabla F(\mathbf{w}_{t+s}^k)\|_2^2$$

$$= \sum_{\mu=1}^{m_k-1} \mu\mathbb{E} \left[\|\nabla F(\mathbf{w}_t^k)\|_2^2 + \|\nabla F(\mathbf{w}_{t+1}^k)\|_2^2 + \cdots + \|\nabla F(\mathbf{w}_{t+\mu-1}^k)\|_2^2 \right]$$

$$\leq [1 + 2 + \cdots + (m_k - 1)]\|\nabla F(\mathbf{w}_t^k)\|_2^2 + [1 + 2 + \cdots + (m_k - 1) - 1] \sum_{\mu=1}^{m_k-1} \mathbb{E}\|\nabla F(\mathbf{w}_{t+\mu}^k)\|_2^2$$

$$= \frac{m_k(m_k - 1)}{2}\|\nabla F(\mathbf{w}_t^k)\|_2^2 + \frac{(m_k - 2)(m_k + 1)}{2} \sum_{\mu=1}^{m_k-1} \mathbb{E}\|\nabla F(\mathbf{w}_{t+\mu}^k)\|_2^2 \tag{12}$$

We then plug T_4 back into (11),

$$T_1 \leq -\frac{\eta}{2K} \sum_{k=1}^{K} p_k(m_k + 1)\|\nabla F(\mathbf{w}_t)\|_2^2 - \frac{\eta}{2K} \sum_{k=1}^{K} p_k \sum_{\mu=1}^{m_k-1} \mathbb{E}\|\nabla F(\mathbf{w}_{t+\mu}^k)\|_2^2 + \frac{L^2\eta}{2K} \sum_{k=1}^{K} p_k \sum_{\mu=1}^{m_k-1} \frac{\mu^2\eta^2\sigma_k^2}{b_k}$$

$$+ \frac{L^2\eta^3}{2K} \sum_{k=1}^{K} p_k \frac{m_k(m_k - 1)}{2}\|\nabla F(\mathbf{w}_t^k)\|_2^2 + \frac{L^2\eta^3}{2K} \sum_{k=1}^{K} p_k \frac{(m_k - 2)(m_k + 1)}{2} \sum_{\mu=1}^{m_k-1} \mathbb{E}\|\nabla F(\mathbf{w}_{t+\mu}^k)\|_2^2 \tag{13}$$

3

$$\leq \left[\frac{L^2\eta^3}{4K} \sum_{k=1}^{K} p_k m_k(m_k-1) - \frac{\eta}{2K} \sum_{k=1}^{K} p_k(m_k+1) \right] \|\nabla F(\mathbf{w}_t)\|_2^2$$

$$+ \frac{L^2\eta^3}{2K} \left[\sum_{k=1}^{K} p_k \left[\frac{(m_k-2)(m_k+1)}{2} - \frac{1}{L^2\eta^2} \right] \right] \sum_{\mu=1}^{m_k-1} \mathbb{E}\|\nabla F(\mathbf{w}_{t+\mu}^k)\|_2^2 + \frac{L^2\eta^3}{12K} \sum_{k=1}^{K} \frac{p_k \sigma_k^2 m_k(m_k-1)(2m_k-1)}{b_k}$$

$$\tag{14}$$

On the other hand, bound T_2:

$$T_2 = \frac{L\eta^2}{2K^2} \mathbb{E}\| \sum_{k=1}^{K} p_k \sum_{\mu=0}^{m_k-1} \frac{1}{b_k} \sum_{i=1}^{b_k} \nabla F(\mathbf{w}_{t+\mu}^k; \xi_{i,\mu}^k)\|_2^2$$

$$= \frac{L\eta^2}{2K^2} \mathbb{E}\| \sum_{k=1}^{K} p_k \sum_{\mu=0}^{m_k-1} \frac{1}{b_k} \sum_{i=1}^{b_k} \left(\nabla F(\mathbf{w}_{t+\mu}^k; \xi_{i,\mu}^k) - \nabla F(\mathbf{w}_{t+\mu}^k) + \nabla F(\mathbf{w}_{t+\mu}^k) \right) \|_2^2$$

$$= \frac{L\eta^2}{2K^2} \mathbb{E}\| \sum_{k=1}^{K} p_k \sum_{\mu=0}^{m_k-1} \frac{1}{b_k} \sum_{i=1}^{b_k} \left(\nabla F(\mathbf{w}_{t+\mu}^k; \xi_{i,\mu}^k) - \nabla F(\mathbf{w}_{t+\mu}^k) \right) \|_2^2$$

$$+ \frac{L\eta^2}{2K^2} \mathbb{E}\| \sum_{k=1}^{K} p_k \sum_{\mu=0}^{m_k-1} \nabla F(\mathbf{w}_{t+\mu}^k)\|_2^2$$

$$\leq \frac{L\eta^2}{2K} \mathbb{E} \sum_{k=1}^{K} p_k^2 \| \sum_{\mu=0}^{m_k-1} \frac{1}{b_k} \sum_{i=1}^{b_k} \left(\nabla F(\mathbf{w}_{t+\mu}^k; \xi_{i,\mu}^k) - \nabla F(\mathbf{w}_{t+\mu}^k) \right) \|_2^2$$

$$+ \frac{L\eta^2}{2K^2} \mathbb{E}\| \sum_{k=1}^{K} p_k \sum_{\mu=0}^{m_k-1} \nabla F(\mathbf{w}_{t+\mu}^k)\|_2^2$$

$$\leq \frac{L\eta^2}{2K} \sum_{k=1}^{K} \frac{p_k^2 m_k \sigma_k^2}{b_k} + \frac{L\eta^2}{2K} \sum_{k=1}^{K} p_k^2 m_k \sum_{\mu=0}^{m_k-1} \mathbb{E}\|\nabla F(\mathbf{w}_{t+\mu}^k)\|_2^2 \tag{15}$$

where the third equality is due to the assumption 3, and the last two inequality is because of the assumption 4 and Cauchy-Schwartz inequality.

Combine the results in T_1 and T_2, we have

$$\mathbb{E}F(\mathbf{w}_{t+1}) - F(\mathbf{w}_t) \leq T_1 + T_2$$

$$\leq \left[\frac{L^2\eta^3}{4K} \sum_{k=1}^{K} p_k m_k(m_k-1) - \frac{\eta}{2K} \sum_{k=1}^{K} p_k(m_k+1) \right] \|\nabla F(\mathbf{w}_t)\|_2^2$$

$$+ \frac{L^2\eta^3}{2K} \left[\sum_{k=1}^{K} p_k \left[\frac{(m_k-2)(m_k+1)}{2} - \frac{1}{L^2\eta^2} \right] \right] \sum_{\mu=1}^{m_k-1} \mathbb{E}\|\nabla F(\mathbf{w}_{t+\mu}^k)\|_2^2$$

$$+ \frac{L^2\eta^3}{12K} \sum_{k=1}^{K} \frac{p_k \sigma_k^2 m_k(m_k-1)(2m_k-1)}{b_k} + \frac{L\eta^2}{2K} \sum_{k=1}^{K} \frac{p_k^2 m_k \sigma_k^2}{b_k}$$

$$+ \frac{L\eta^2}{2K} \sum_{k=1}^{K} p_k^2 m_k \left[\|\nabla F(\mathbf{w}_t)\|_2^2 + \sum_{\mu=1}^{m_k-1} \mathbb{E}\|\nabla F(\mathbf{w}_{t+\mu}^k)\|_2^2 \right]$$

$$= \left[\frac{L^2\eta^3}{4K} \sum_{k=1}^{K} p_k m_k(m_k-1) - \frac{\eta}{2K} \sum_{k=1}^{K} p_k(m_k+1) + \frac{L\eta^2}{2K} \sum_{k=1}^{K} p_k^2 m_k \right] \|\nabla F(\mathbf{w}_t)\|_2^2$$

$$+ \frac{L^2\eta^3}{2K} \left[\sum_{k=1}^{K} p_k \left[\frac{(m_k-2)(m_k+1)}{2} - \frac{1}{L^2\eta^2} + \frac{p_k m_k}{L\eta} \right] \right] \sum_{\mu=1}^{m_k-1} \mathbb{E}\|\nabla F(\mathbf{w}_{t+\mu}^k)\|_2^2$$

$$+ \frac{L^2\eta^3}{12K}\sum_{k=1}^{K}\frac{p_k\sigma_k^2 m_k(m_k-1)(2m_k-1)}{b_k} + \frac{L\eta^2}{2K}\sum_{k=1}^{K}\frac{p_k^2 m_k\sigma_k^2}{b_k} \tag{16}$$

If $\sum_{k=1}^{K}p_k\left[\frac{(m_k-2)(m_k+1)}{2} - \frac{1}{L^2\eta^2} + \frac{p_k m_k}{L\eta}\right] \leq 0$, the second term of (16) on the right hand side can be discarded. Then, we have

$$\mathbb{E}F(\mathbf{w}_{t+1}) - F(\mathbf{w}_t) \leq \left[\frac{L^2\eta^3}{4K}\sum_{k=1}^{K}p_k m_k(m_k-1) - \frac{\eta}{2K}\sum_{k=1}^{K}p_k(m_k+1) + \frac{L\eta^2}{2K}\sum_{k=1}^{K}p_k^2 m_k\right]\|\nabla F(\mathbf{w}_t)\|_2^2$$
$$+ \frac{L^2\eta^3}{12K}\sum_{k=1}^{K}\frac{p_k\sigma_k^2 m_k(m_k-1)(2m_k-1)}{b_k} + \frac{L\eta^2}{2K}\sum_{k=1}^{K}\frac{p_k^2 m_k\sigma_k^2}{b_k} \tag{17}$$

By taking the summation we have

$$\mathbb{E}F(\mathbf{w}_T) - F(\mathbf{w}_1) \leq \sum_{t=1}^{T}\left[\frac{L^2\eta^3}{4K}\sum_{k=1}^{K}p_k m_k(m_k-1) - \frac{\eta}{2K}\sum_{k=1}^{K}p_k(m_k+1) + \frac{L\eta^2}{2K}\sum_{k=1}^{K}p_k^2 m_k\right]\|\nabla F(\mathbf{w}_t)\|_2^2$$
$$+ \frac{L^2\eta^3}{12K}\sum_{k=1}^{K}\frac{p_k\sigma_k^2 m_k(m_k-1)(2m_k-1)}{b_k} + \frac{L\eta^2}{2K}\sum_{k=1}^{K}\frac{p_k^2 m_k\sigma_k^2}{b_k} \tag{18}$$

Under assumption 2, we have
$$F^* - F(\mathbf{w}_1) \leq F(\mathbf{w}_t) - F(\mathbf{w}_1) \tag{19}$$

Combining both, we can immediately get the following bound

$$\mathbb{E}\sum_{t=1}^{T}\eta_t\|\nabla F(\mathbf{w}_t)\|_2^2 \leq \frac{(F(\mathbf{w}_1) - F^*)\eta}{G - A - C} + \sum_{t=1}^{T}\frac{L\eta^3}{2K(G-A-C)}\sum_{k=1}^{K}\frac{p_k^2 m_k\sigma_k^2}{b_k}$$
$$+ \sum_{t=1}^{T}\frac{L^2\eta^4}{12K(G-A-C)}\sum_{k=1}^{K}\frac{p_k\sigma_k^2 m_k(m_k-1)(2m_k-1)}{b_k} \tag{20}$$

where $A = \frac{\eta}{2K}\sum_{k=1}^{K}p_k(m_k+1), G = \frac{L^2\eta^3}{4K}\sum_{k=1}^{K}p_k m_k(m_k-1), C = \frac{L\eta^2}{2K}\sum_{k=1}^{K}p_k^2 m_k$, and let $\beta_k = \frac{\sigma_k^2}{b_k}$, we obtain the bound on the expected average squared gradient norms of F under fixed learning rate as following:

$$\frac{1}{T}\mathbb{E}\sum_{t=1}^{T}\|\nabla F(\mathbf{w}_t)\|_2^2 \leq \frac{F(\mathbf{w}_1) - F^*}{T(G-A-C)} + \frac{L\eta^2}{2K(G-A-C)}\sum_{k=1}^{K}p_k^2\beta_k m_k$$
$$+ \frac{L^2\eta^3}{12K(G-A-C)}\sum_{k=1}^{K}p_k\beta_k m_k(m_k-1)(2m_k-1) \tag{21}$$

which completes the proof. $\qquad\square$