FedVARP: Tackling the Variance Due to Partial Client Participation in Federated Learning (Supplementary material)

Divyansh Jhunjhunwala[1] **Pranay Sharma**[1] **Aushim Nagarkatti**[1] **Gauri Joshi**[1]

[1]Carnegie Mellon University, Pittsburgh, Pennsylvania, USA

1 NOTATIONS AND BASIC RESULTS

Let $\mathcal{S}^{(t)}$ be the subset of clients sampled in the $t-$th round. Let $\xi^{(t)}$ denotes the randomness due to the stochastic sampling at round t.

$$\begin{aligned}
\text{Normalized Stochastic Gradient: } \Delta_i^{(t)} &= \frac{1}{\tau}\sum_{k=0}^{\tau-1} \nabla f_i(\mathbf{w}_i^{(t,k)}, \xi_i^{(t,k)}) \\
\text{Normalized Gradient: } \mathbf{h}_i^{(t)} &= \frac{1}{\tau}\sum_{k=0}^{\tau-1} \nabla f_i(\mathbf{w}_i^{(t,k)}) \\
\text{Average Normalized Gradient: } \bar{\mathbf{h}}^{(t)} &= \frac{1}{N}\sum_{i=1}^{N} \mathbf{h}_i^{(t)} \\
\text{Server Updates: } \mathbf{w}^{(t+1)} &= \mathbf{w}^{(t)} - \tilde{\eta}_s \frac{1}{M}\sum_{i \in \mathcal{S}^{(t)}} \Delta_i^{(t)}, \qquad \text{where } \tilde{\eta}_s = \eta_s \eta_c \tau
\end{aligned} \tag{1}$$

Lemma 1 (Young's inequality). *Given two same-dimensional vectors* $\mathbf{u}, \mathbf{v} \in \mathbb{R}^d$*, the Euclidean inner product can be bounded as follows:*

$$\langle \mathbf{x}, \mathbf{v} \rangle \leq \frac{\|\mathbf{x}\|^2}{2\gamma} + \frac{\gamma \|\mathbf{v}\|^2}{2}$$

for every constant $\gamma > 0$*.*

Lemma 2 (Jensen's inequality). *Given a convex function* f *and a random variable* X*, the following holds.*

$$f\left(\mathbb{E}[X]\right) \leq \mathbb{E}\left[f(X)\right].$$

Lemma 3 (Sum of squares). *For a positive integer* K*, and a set of vectors* $\mathbf{x}_1, \dots, \mathbf{x}_K$*, the following holds:*

$$\left\| \sum_{k=1}^{K} \mathbf{x}_k \right\|^2 \leq K \sum_{k=1}^{K} \|\mathbf{x}_k\|^2.$$

Lemma 4 (Variance under uniform, without replacement sampling). *Let* $\bar{\mathbf{x}} = \frac{1}{N}\sum_{i=1}^{N} \mathbf{x}_i$*. If* $\bar{\mathbf{x}}$ *is approximated using a mini-batch* $\mathcal{M}$ *of size* M*, sampled uniformly at random, and without replacement, then the following holds.*

$$\mathbb{E}\left[\frac{1}{M}\sum_{i \in \mathcal{M}} \mathbf{x}_i\right] = \bar{\mathbf{x}},$$

Accepted for the 38th Conference on Uncertainty in Artificial Intelligence (UAI 2022).

$$\mathbb{E}\left\|\frac{1}{M}\sum_{i\in\mathcal{M}}\mathbf{x}_i-\bar{\mathbf{x}}\right\|^2=\frac{1}{M}\frac{(N-M)}{(N-1)}\frac{1}{N}\sum_{i=1}^{N}\left\|\mathbf{x}_i-\bar{\mathbf{x}}\right\|^2.$$

Proof.

$$\begin{aligned}
\mathbb{E}\left\|\frac{1}{M}\sum_{i\in\mathcal{M}}\mathbf{x}_i-\bar{\mathbf{x}}\right\|^2 &= \mathbb{E}\left\|\frac{1}{M}\sum_{i=1}^{N}\mathbb{I}(i\in\mathcal{M})\left(\mathbf{x}_i-\bar{\mathbf{x}}\right)\right\|^2 \\
&= \frac{1}{M^2}\mathbb{E}\left[\sum_{i=1}^{N}\left(\mathbb{I}(i\in\mathcal{M})\right)^2\left\|\mathbf{x}_i-\bar{\mathbf{x}}\right\|^2+\sum_{j\in[N]}\sum_{\substack{j\in[N]\\ i\neq j}}\mathbb{I}(i\in\mathcal{M})\mathbb{I}(j\in\mathcal{M})\left\langle\mathbf{x}_i-\bar{\mathbf{x}},\mathbf{x}_j-\bar{\mathbf{x}}\right\rangle\right] \\
&= \frac{1}{M^2}\sum_{i=1}^{N}\frac{M}{N}\left\|\mathbf{x}_i-\bar{\mathbf{x}}\right\|^2+\frac{1}{M^2}\sum_{i\neq j}\frac{M}{N}\frac{(M-1)}{(N-1)}\left\langle\mathbf{x}_i-\bar{\mathbf{x}},\mathbf{x}_j-\bar{\mathbf{x}}\right\rangle \\
&= \frac{1}{M^2}\sum_{i=1}^{N}\left\|\mathbf{x}_i-\bar{\mathbf{x}}\right\|^2\left[\frac{M}{N}-\frac{M}{N}\frac{(M-1)}{(N-1)}\right]+\frac{1}{M^2}\frac{M}{N}\frac{(M-1)}{(N-1)}\underbrace{\left\|\sum_{i=1}^{N}\left(\mathbf{x}_i-\bar{\mathbf{x}}\right)\right\|^2}_{=0} \\
&= \frac{1}{M}\frac{(N-M)}{(N-1)}\frac{1}{N}\sum_{i=1}^{N}\left\|\mathbf{x}_i-\bar{\mathbf{x}}\right\|^2.
\end{aligned}$$

□

2 CONVERGENCE PROOF FOR FEDAVG (THEOREM 1)

In this section we prove the convergence of FedAvg, and provide the complexity and communication guarantees. We organize this section as follows. First, in 2.1 we present some intermediate results, which we use to prove the main theorem. Next, in 2.2, we present the proof of Theorem 1, which is followed by the proofs of the intermediate results in 2.3.

2.1 INTERMEDIATE LEMMAS

Lemma 5. *Suppose the function f satisfies Assumption 1, and the stochastic oracles at the clients satisfy Assumption 2, then the iterates $\{\mathbf{w}^{(t)}\}_t$ generated by* FedAvg *satisfy*

$$\begin{aligned}
\mathbb{E}_{\mathcal{S}^{(t)},\xi^{(t)}}\left[f(\mathbf{w}^{(t+1)})\right] \leq f(\mathbf{w}^{(t)}) &- \frac{\tilde{\eta}_s}{2}\left[\left\|\nabla f(\mathbf{w}^{(t)})\right\|^2+\mathbb{E}_{\xi^{(t)}}\left\|\bar{\mathbf{h}}^{(t)}\right\|^2-\mathbb{E}_{\xi^{(t)}}\left\|\nabla f(\mathbf{w}^{(t)})-\bar{\mathbf{h}}^{(t)}\right\|^2\right] \\
&+ \frac{\tilde{\eta}_s^2 L}{2}\left[\frac{2\sigma^2}{M\tau}+2\mathbb{E}_{\mathcal{S}^{(t)},\xi^{(t)}}\left\|\frac{1}{M}\sum_{i\in\mathcal{S}^{(t)}}\mathbf{h}_i^{(t)}\right\|^2\right],
\end{aligned}$$

where $\tilde{\eta}_s$ is the server learning rate, and $\mathbb{E}_{\mathcal{S}^{(t)},\xi^{(t)}}\left[\cdot\right]$ is expectation over the randomness in the $t-$th round, conditioned on $\mathbf{w}^{(t)}$.

Lemma 6. *Suppose the function f satisfies Assumption 1, and the stochastic oracles at the clients satisfy Assumptions 2, 3. Further, η_c, the client learning rate is chosen such that $\eta_c \leq \frac{1}{2L\tau}$. Then, the iterates $\{\mathbf{w}^{(t)}\}_t$ generated by* FedAvg *satisfy*

$$\begin{aligned}
\mathbb{E}_{\xi^{(t)}}\left\|\nabla f(\mathbf{w}^{(t)})-\bar{\mathbf{h}}^{(t)}\right\|^2 &\leq \frac{1}{N}\sum_{i=1}^{N}\mathbb{E}_{\xi^{(t)}}\left\|\nabla f_i(\mathbf{w}^{(t)})-\mathbf{h}_i^{(t)}\right\|^2 \\
&\leq 2\eta_c^2L^2(\tau-1)\sigma^2+8\eta_c^2L^2\tau(\tau-1)\left[\sigma_g^2+\left\|\nabla f(\mathbf{w}^{(t)})\right\|^2\right].
\end{aligned}$$

Lemma 7. *Suppose the function f satisfies Assumption 1, and the stochastic oracles at the clients satisfy Assumptions 2, 3. Then, the iterates $\{\mathbf{w}^{(t)}\}_t$ generated by* `FedAvg` *satisfy*

$$\mathbb{E}_{\mathcal{S}^{(t)},\xi^{(t)}} \left\| \frac{1}{M}\sum_{i\in\mathcal{S}^{(t)}} \mathbf{h}_i^{(t)} \right\|^2 \le \frac{3}{N}\sum_{i=1}^{N} \mathbb{E}\left\|\mathbf{h}_i^{(t)} - \nabla f_i(\mathbf{w}^{(t)})\right\|^2 + \frac{3(N-M)}{(N-1)M}\sigma_g^2 + 3\mathbb{E}\left\|\nabla f(\mathbf{w}^{(t)})\right\|^2.$$

2.2 PROOF OF THEOREM 1

For the sake of completeness, we first state the complete theorem statement.

Theorem. *Suppose the function f satisfies Assumption 1, and the stochastic oracles at the clients satisfy Assumptions 2,3. Further, the client learning rate η_c, and the server learning rate η_s are chosen such that $\eta_c \le \frac{1}{8L\tau}$, $\eta_s\eta_c \le \frac{1}{24\tau L}$. Then, the iterates $\{\mathbf{w}^{(t)}\}_t$ generated by* `FedAvg` *satisfy*

$$\begin{aligned}\min_{t\in[T]} \mathbb{E}\left\|\nabla f(\mathbf{w}^{(t)})\right\|^2 \le \underbrace{\mathcal{O}\left(\frac{f(\mathbf{w}^{(0)}) - f^*}{\eta_s\eta_c\tau T}\right)}_{\textit{Effect of initialization}} + \underbrace{\mathcal{O}\left(\frac{\eta_s\eta_c L\sigma^2}{M} + \eta_c^2L^2(\tau-1)\sigma^2\right)}_{\textit{Stochastic Gradient Error}} \\ + \underbrace{\mathcal{O}\left(\eta_s\eta_c\tau L\frac{(N-M)}{(N-1)M}\sigma_g^2\right)}_{\textit{Error due to partial participation}} + \underbrace{\mathcal{O}\left(\eta_c^2L^2\tau(\tau-1)\sigma_g^2\right)}_{\textit{Client Drift Error}},\end{aligned}$$

where $f^* = \arg\min_{\mathbf{x}} f(\mathbf{x})$.

Proof. Note that for simplicity, we use the notation $\tilde{\eta}_s = \eta_s\eta_c\tau$. Substituting the bounds in Lemma 6 and Lemma 7 in (2), we get

$$\begin{aligned}
&\mathbb{E}\left[f(\mathbf{w}^{(t+1)}) - f(\mathbf{w}^{(t)})\right] \\
&\le -\frac{\tilde{\eta}_s}{2}\left\|\nabla f(\mathbf{w}^{(t)})\right\|^2 - \frac{\tilde{\eta}_s}{2}\mathbb{E}\left\|\bar{\mathbf{h}}^{(t)}\right\|^2 \\
&\quad + \frac{\tilde{\eta}_s}{2}\left[2\eta_c^2L^2(\tau-1)\sigma^2 + 8\eta_c^2L^2\tau(\tau-1)\left(\sigma_g^2 + \left\|\nabla f(\mathbf{w}^{(t)})\right\|^2\right)\right] \\
&\quad + \frac{\tilde{\eta}_s^2L}{2}\left[\frac{2\sigma^2}{M\tau} + \frac{6(N-M)}{(N-1)M}\sigma_g^2 + 6\mathbb{E}\left\|\nabla f(\mathbf{w}^{(t)})\right\|^2\right] \\
&\quad + 3\tilde{\eta}_s^2L\left[2\eta_c^2L^2(\tau-1)\sigma^2 + 8\eta_c^2L^2\tau(\tau-1)\left(\sigma_g^2 + \left\|\nabla f(\mathbf{w}^{(t)})\right\|^2\right)\right] \\
&\le -\frac{\tilde{\eta}_s}{2}\left(1 - 8\eta_c^2L^2\tau(\tau-1) - 6\tilde{\eta}_sL - 48\tilde{\eta}_sL\eta_c^2L^2\tau(\tau-1)\right)\left\|\nabla f(\mathbf{w}^{(t)})\right\|^2 - \frac{\tilde{\eta}_s}{2}\mathbb{E}\left\|\bar{\mathbf{h}}^{(t)}\right\|^2 \\
&\quad + \frac{\tilde{\eta}_s}{2}\left(1 + 6\tilde{\eta}_sL\right)2\eta_c^2L^2(\tau-1)\left[\sigma^2 + 4\tau\sigma_g^2\right] + \frac{\tilde{\eta}_s^2L}{2}\left[\frac{2\sigma^2}{M\tau} + \frac{6(N-M)}{(N-1)M}\sigma_g^2\right] \\
&\le -\frac{\tilde{\eta}_s}{4}\left\|\nabla f(\mathbf{w}^{(t)})\right\|^2 + 2\tilde{\eta}_s\eta_c^2L^2(\tau-1)\left[\sigma^2 + 4\tau\sigma_g^2\right] + \frac{\tilde{\eta}_s^2L}{2}\left[\frac{2\sigma^2}{M\tau} + \frac{6(N-M)}{(N-1)M}\sigma_g^2\right]. \qquad (2)
\end{aligned}$$

where (2) follows because

$$\begin{aligned}
8\eta_c^2L^2\tau(\tau-1) &\le \frac{1}{8} && (\because \eta_c \le \tfrac{1}{8\tau L}) \\
6\tilde{\eta}_sL &\le \frac{1}{4} && (\because \eta_s\eta_c \le \tfrac{1}{24\tau L}) \\
48\tilde{\eta}_sL\eta_c^2L^2\tau(\tau-1) \le 6\tilde{\eta}_sL &\le \frac{1}{4}. && (\because 8\eta_c^2L^2\tau(\tau-1) \le \tfrac{1}{8})
\end{aligned}$$

Rearranging the terms, and summing over $t = 0, \dots, T-1$, we get

$$\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}\left\|\nabla f(\mathbf{w}^{(t)})\right\|^2$$

$$
\begin{aligned}
&\leq \frac{4}{\tilde{\eta}_s T}\sum_{t=0}^{T-1}\mathbb{E}\left[f(\mathbf{w}^{(t)}) - f(\mathbf{w}^{(t+1)})\right] + 8\eta_c^2L^2(\tau-1)\left[\sigma^2+4\tau\sigma_g^2\right] + 2\tilde{\eta}_s L\left[\frac{2\sigma^2}{M\tau}+\frac{6(N-M)}{(N-1)M}\sigma_g^2\right] \\
&\leq \frac{4\left[f(\mathbf{w}^{(0)})-f(\mathbf{w}^{(T)})\right]}{\eta_s\eta_c\tau T} + \frac{4\eta_s\eta_c L\sigma^2}{M} + 8\eta_c^2L^2(\tau-1)\sigma^2 + 12\eta_s\eta_c\tau L\frac{(N-M)}{(N-1)M}\sigma_g^2 + 32\eta_c^2L^2\tau(\tau-1)\sigma_g^2.
\end{aligned}
$$

□

2.3 PROOFS OF THE INTERMEDIATE LEMMAS

Proof of Lemma 5. Using L-smoothness (Assumption 1) of f, and only considering the randomness in the t-th round $\{\mathcal{S}^{(t)}, \xi^{(t)}\}$,

$$
\begin{aligned}
\mathbb{E}_{\mathcal{S}^{(t)},\xi^{(t)}} f(\mathbf{w}^{(t+1)}) - f(\mathbf{w}^{(t)}) &\leq \mathbb{E}_{\mathcal{S}^{(t)},\xi^{(t)}}\left\langle \nabla f(\mathbf{w}^{(t)}), \mathbf{w}^{(t+1)} - \mathbf{w}^{(t)}\right\rangle + \frac{L}{2}\mathbb{E}_{\mathcal{S}^{(t)},\xi^{(t)}}\left\|\mathbf{w}^{(t+1)} - \mathbf{w}^{(t)}\right\|^2 \\
&= -\tilde{\eta}_s \underbrace{\mathbb{E}_{\mathcal{S}^{(t)},\xi^{(t)}}\left\langle \nabla f(\mathbf{w}^{(t)}), \frac{1}{M}\sum_{i\in\mathcal{S}^{(t)}}\Delta_i^{(t)}\right\rangle}_{T_1} + \frac{\tilde{\eta}_s^2 L}{2}\underbrace{\mathbb{E}_{\mathcal{S}^{(t)},\xi^{(t)}}\left\|\frac{1}{M}\sum_{i\in\mathcal{S}^{(t)}}\Delta_i^{(t)}\right\|^2}_{T_2}. \qquad (3)
\end{aligned}
$$

Next, we bound the terms T_1 and T_2 separately.

$$
\begin{aligned}
-T_1 &= -\mathbb{E}_{\xi^{(t)}}\left\langle \nabla f(\mathbf{w}^{(t)}), \frac{1}{N}\sum_{i=1}^{N}\mathbf{h}_i^{(t)}\right\rangle && \text{(from Assumption 3 and uniform sampling of clients)} \\
&= \frac{1}{2}\left[\mathbb{E}_{\xi^{(t)}}\left\|\nabla f(\mathbf{w}^{(t)}) - \frac{1}{N}\sum_{i=1}^{N}\mathbf{h}_i^{(t)}\right\|^2 - \left\|\nabla f(\mathbf{w}^{(t)})\right\|^2 - \mathbb{E}_{\xi^{(t)}}\left\|\frac{1}{N}\sum_{i=1}^{N}\mathbf{h}_i^{(t)}\right\|^2\right]. && (4)
\end{aligned}
$$

Next, we bound T_2.

$$
\begin{aligned}
T_2 &\leq 2\mathbb{E}_{\mathcal{S}^{(t)},\xi^{(t)}}\left\|\frac{1}{M}\sum_{i\in\mathcal{S}^{(t)}}\left(\Delta_i^{(t)} - \mathbf{h}_i^{(t)}\right)\right\|^2 + 2\mathbb{E}_{\mathcal{S}^{(t)},\xi^{(t)}}\left\|\frac{1}{M}\sum_{i\in\mathcal{S}^{(t)}}\mathbf{h}_i^{(t)}\right\|^2 && \text{(Young's inequality)} \\
&\leq \frac{2}{M}\frac{1}{N}\sum_{i=1}^{N}\mathbb{E}_{\xi^{(t)}}\left\|\Delta_i^{(t)} - \mathbf{h}_i^{(t)}\right\|^2 + 2\mathbb{E}_{\mathcal{S}^{(t)},\xi^{(t)}}\left\|\frac{1}{M}\sum_{i\in\mathcal{S}^{(t)}}\mathbf{h}_i^{(t)}\right\|^2 && \text{(uniform sampling of clients, } \mathbb{E}[\Delta_i^{(t)}] = \mathbf{h}_i^{(t)}\text{)} \\
&\leq \frac{2}{M}\frac{1}{N}\sum_{i=1}^{N}\frac{\sigma^2}{\tau} + 2\mathbb{E}_{\mathcal{S}^{(t)},\xi^{(t)}}\left\|\frac{1}{M}\sum_{i\in\mathcal{S}^{(t)}}\mathbf{h}_i^{(t)}\right\|^2, && (5)
\end{aligned}
$$

where, (5) follows from the following reasoning.

$$
\begin{aligned}
\mathbb{E}_{\xi^{(t)}}\left\|\Delta_i^{(t)} - \mathbf{h}_i^{(t)}\right\|^2 &= \mathbb{E}_{\xi^{(t)}}\left\|\frac{1}{\tau}\sum_{k=0}^{\tau-1}\left(\nabla f_i(\mathbf{w}_i^{(t,k)}, \xi_i^{(t,k)}) - \nabla f_i(\mathbf{w}_i^{(t,k)})\right)\right\|^2 && \text{(from (1))} \\
&= \frac{1}{\tau^2}\mathbb{E}_{\xi^{(t)}}\left[\sum_{k=0}^{\tau-1}\left\|\nabla f_i(\mathbf{w}_i^{(t,k)}, \xi_i^{(t,k)}) - \nabla f_i(\mathbf{w}_i^{(t,k)})\right\|^2\right. \\
&\quad \left. + \frac{2}{\tau^2}\sum_{j<k}\mathbb{E}_{\xi^{(t)}}\left\langle \underbrace{\mathbb{E}\left[\nabla f_i(\mathbf{w}_i^{(t,k)}, \xi_i^{(t,k)}) - \nabla f_i(\mathbf{w}_i^{(t,k)})|\mathbf{w}_i^{(t,j)}\right]}_{=0}, \nabla f_i(\mathbf{w}_i^{(t,j)}, \xi_i^{(t,j)}) - \nabla f_i(\mathbf{w}_i^{(t,j)})\right\rangle\right] \\
&\leq \frac{\sigma^2}{\tau}. && \text{(Assumption 3)}
\end{aligned}
$$

Substituting the bounds on T_1 (4) and T_2 (5) in (3), we get the result in the lemma. □

Proof of Lemma 6. We borrow some of the proof techniques from [Wang et al., 2020].

$$\mathbb{E}_{\xi^{(t)}}\left\|\nabla f(\mathbf{w}^{(t)})-\bar{\mathbf{h}}^{(t)}\right\|^2 \leq \frac{1}{N}\sum_{i=1}^{N}\mathbb{E}_{\xi^{(t)}}\left\|\nabla f_i(\mathbf{w}^{(t)})-\mathbf{h}_i^{(t)}\right\|^2 \tag{Jensen's inequality}$$

$$= \frac{1}{N}\sum_{i=1}^{N}\mathbb{E}_{\xi^{(t)}}\left\|\frac{1}{\tau}\sum_{k=0}^{\tau-1}\left(\nabla f_i(\mathbf{w}^{(t)})-\nabla f_i(\mathbf{w}_i^{(t,k)})\right)\right\|^2 \tag{from (1)}$$

$$= \frac{L^2}{N}\sum_{i=1}^{N}\frac{1}{\tau}\sum_{k=0}^{\tau-1}\mathbb{E}_{\xi^{(t)}}\left\|\mathbf{w}_i^{(t,k)}-\mathbf{w}^{(t)}\right\|^2 \tag{6}$$

Next, we bound the individual difference $\mathbb{E}_{\xi^{(t)}}\left\|\mathbf{w}_i^{(t,k)}-\mathbf{w}^{(t)}\right\|^2$.

$$\begin{aligned}
\mathbb{E}_{\xi^{(t)}}\left\|\mathbf{w}_i^{(t,k)}-\mathbf{w}^{(t)}\right\|^2 &= \eta_c^2\mathbb{E}_{\xi^{(t)}}\left\|\sum_{j=0}^{k-1}\nabla f_i(\mathbf{w}_i^{(t,j)},\xi_i^{(t,j)})\right\|^2 \\
&= \eta_c^2\left[\mathbb{E}_{\xi^{(t)}}\left\|\sum_{j=0}^{k-1}\left(\nabla f_i(\mathbf{w}_i^{(t,j)},\xi_i^{(t,j)})-\nabla f_i(\mathbf{w}_i^{(t,j)})\right)\right\|^2+\mathbb{E}_{\xi^{(t)}}\left\|\sum_{j=0}^{k-1}\nabla f_i(\mathbf{w}_i^{(t,j)})\right\|^2\right] \\
&\leq \eta_c^2\left[\sum_{j=0}^{k-1}\mathbb{E}_{\xi^{(t)}}\left\|\nabla f_i(\mathbf{w}_i^{(t,j)},\xi_i^{(t,j)})-\nabla f_i(\mathbf{w}_i^{(t,j)})\right\|^2+k\sum_{j=0}^{k-1}\mathbb{E}_{\xi^{(t)}}\left\|\nabla f_i(\mathbf{w}_i^{(t,j)})\right\|^2\right] \\
&\leq \eta_c^2\left[k\sigma^2+k\sum_{j=0}^{k-1}\mathbb{E}_{\xi^{(t)}}\left\|\nabla f_i(\mathbf{w}_i^{(t,j)})\right\|^2\right].
\end{aligned} \tag{7}$$

Summing over $k=0,\ldots,\tau-1$, we get

$$\begin{aligned}
\frac{1}{\tau}\sum_{k=0}^{\tau-1}\mathbb{E}_{\xi^{(t)}}\left\|\mathbf{w}_i^{(t,k)}-\mathbf{w}^{(t)}\right\|^2 &\leq \eta_c^2\frac{1}{\tau}\sum_{k=0}^{\tau-1}\left[k\sigma^2+k\sum_{j=0}^{k-1}\mathbb{E}_{\xi^{(t)}}\left\|\nabla f_i(\mathbf{w}_i^{(t,j)})-\nabla f_i(\mathbf{w}^{(t)})+\nabla f_i(\mathbf{w}^{(t)})\right\|^2\right] \\
&\leq \eta_c^2(\tau-1)\sigma^2+\frac{\eta_c^2L^2}{\tau}\sum_{k=0}^{\tau-1}k\sum_{j=0}^{k-1}\left[\mathbb{E}_{\xi^{(t)}}\left\|\mathbf{w}_i^{(t,j)}-\mathbf{w}^{(t)}\right\|^2+\left\|\nabla f_i(\mathbf{w}^{(t)})\right\|^2\right] \\
&\leq \eta_c^2(\tau-1)\sigma^2+2\eta_c^2L^2\tau(\tau-1)\left[\frac{1}{\tau}\sum_{k=0}^{\tau-1}\mathbb{E}_{\xi^{(t)}}\left\|\mathbf{w}_i^{(t,k)}-\mathbf{w}^{(t)}\right\|^2\right] \\
&\quad+2\eta_c^2\tau(\tau-1)\left\|\nabla f_i(\mathbf{w}^{(t)})\right\|^2.
\end{aligned} \tag{8}$$

Define $D\triangleq 2\eta_c^2L^2\tau(\tau-1)$. We choose η_c small enough such that $D\leq 1/2$. Then, rearranging the terms in

$$\frac{1}{\tau}\sum_{k=0}^{\tau-1}\mathbb{E}_{\xi^{(t)}}\left\|\mathbf{w}_i^{(t,k)}-\mathbf{w}^{(t)}\right\|^2 \leq \frac{\eta_c^2(\tau-1)\sigma^2}{1-D}+\frac{2\eta_c^2\tau(\tau-1)}{1-D}\left\|\nabla f_i(\mathbf{w}^{(t)})\right\|^2. \tag{9}$$

Substituting (9) in (6), we get

$$\mathbb{E}_{\xi^{(t)}}\left\|\nabla f(\mathbf{w}^{(t)})-\bar{\mathbf{h}}^{(t)}\right\|^2 \leq \frac{\eta_c^2L^2(\tau-1)\sigma^2}{1-D}+\frac{D}{1-D}\left\|\nabla f_i(\mathbf{w}^{(t)})-\nabla f(\mathbf{w}^{(t)})+\nabla f(\mathbf{w}^{(t)})\right\|^2$$

$$\leq 2\eta_c^2L^2(\tau-1)\sigma^2+4D\sigma_g^2+4D\left\|\nabla f(\mathbf{w}^{(t)})\right\|^2. \tag{since $D\leq 1/2$}$$

□

Proof of Lemma 7. Also,

$$\mathbb{E}_{\mathcal{S}^{(t)},\xi^{(t)}} \left\| \frac{1}{M} \sum_{i \in \mathcal{S}^{(t)}} \left(\mathbf{h}_i^{(t)} - \nabla f_i(\mathbf{w}^{(t)}) + \nabla f_i(\mathbf{w}^{(t)}) \right) - \nabla f(\mathbf{w}^{(t)}) + \nabla f(\mathbf{w}^{(t)}) \right\|^2$$

$$\leq 3\mathbb{E}_{\mathcal{S}^{(t)},\xi^{(t)}} \left\| \frac{1}{M} \sum_{i \in \mathcal{S}^{(t)}} \left(\mathbf{h}_i^{(t)} - \nabla f_i(\mathbf{w}^{(t)}) \right) \right\|^2 + 3\mathbb{E}_{\mathcal{S}^{(t)}} \left\| \frac{1}{M} \sum_{i \in \mathcal{S}^{(t)}} \nabla f_i(\mathbf{w}^{(t)}) - \nabla f(\mathbf{w}^{(t)}) \right\|^2 + 3\mathbb{E} \left\| \nabla f(\mathbf{w}^{(t)}) \right\|^2$$

$$\leq 3\mathbb{E}_{\mathcal{S}^{(t)},\xi^{(t)}} \left[\frac{1}{M} \sum_{i \in \mathcal{S}^{(t)}} \left\| \mathbf{h}_i^{(t)} - \nabla f_i(\mathbf{w}^{(t)}) \right\|^2 \right] + 3\frac{N-M}{(N-1)M} \frac{1}{N} \sum_{i=1}^{N} \mathbb{E} \left\| \nabla f_i(\mathbf{w}^{(t)}) - \nabla f(\mathbf{w}^{(t)}) \right\|^2 + 3\mathbb{E} \left\| \nabla f(\mathbf{w}^{(t)}) \right\|^2$$

(sampling without replacement, see Lemma 4)

$$\leq \frac{3}{N} \sum_{i=1}^{N} \mathbb{E} \left\| \mathbf{h}_i^{(t)} - \nabla f_i(\mathbf{w}^{(t)}) \right\|^2 + \frac{3(N-M)}{(N-1)M} \sigma_g^2 + 3\mathbb{E} \left\| \nabla f(\mathbf{w}^{(t)}) \right\|^2 .$$

□

3 CONVERGENCE RESULT FOR FEDVARP (THEOREM 2)

In this section we prove the convergence result for FedVARP in Theorem 2, and provide the complexity and communication guarantees.

We organize this section as follows. First, in 3.1 we present some intermediate results, which we use to prove the main theorem. Next, in 3.2, we present the proof of Theorem 2, which is followed by the proofs of the intermediate results in 3.3.

3.1 INTERMEDIATE LEMMAS

Lemma 8. *Suppose the function f satisfies Assumption 1, and the stochastic oracles at the clients satisfy Assumption 2, then the iterates $\{\mathbf{w}^{(t)}\}_t$ generated by FedVARP satisfy*

$$\mathbb{E}_{\mathcal{S}^{(t)},\xi^{(t)}} \left[f(\mathbf{w}^{(t+1)}) \right] \leq f(\mathbf{w}^{(t)}) - \frac{\tilde{\eta}_s}{2} \left[\left\| \nabla f(\mathbf{w}^{(t)}) \right\|^2 + \mathbb{E}_{\xi^{(t)}} \left\| \bar{\mathbf{h}}^{(t)} \right\|^2 - \mathbb{E}_{\xi^{(t)}} \left\| \nabla f(\mathbf{w}^{(t)}) - \bar{\mathbf{h}}^{(t)} \right\|^2 \right] + \frac{\tilde{\eta}_s^2 L}{2} \mathbb{E}_{\mathcal{S}^{(t)},\xi^{(t)}} \left[\left\| \mathbf{v}^{(t)} \right\|^2 \right], \tag{10}$$

where $\tilde{\eta}_s = \eta_s \eta_c \tau$ is the effective server learning rate, and $\mathbb{E}_{\mathcal{S}^{(t)},\xi^{(t)}}[\cdot]$ is expectation over the randomness in the $t-$th round, conditioned on $\mathbf{w}^{(t)}$.

Lemma 9. *Suppose the function f satisfies Assumption 1, and the stochastic oracles at the clients satisfy Assumptions 2. Then, the iterates $\{\mathbf{w}^{(t)}\}_t$ generated by FedVARP satisfy*

$$\mathbb{E}_{\mathcal{S}^{(t)},\xi^{(t)}} \left\| \mathbf{v}^{(t)} - \bar{\mathbf{h}}^{(t)} \right\|^2 \leq \frac{\sigma^2}{M\tau} + \frac{4(N-M)}{M(N-1)} \left[\frac{1}{N} \sum_{i=1}^{N} \mathbb{E}_{\xi^{(t)}} \left[\left\| \mathbf{h}_i^{(t)} - \nabla f_i(\mathbf{w}^{(t)}) \right\|^2 \right] + \tilde{\eta}_s^2 L^2 \left\| \mathbf{v}^{(t-1)} \right\|^2 \right] + \frac{2(N-M)}{M(N-1)} \frac{1}{N} \sum_{i=1}^{N} \left\| \nabla f_i(\mathbf{w}^{(t-1)}) - \mathbf{y}_i^{(t)} \right\|^2 .$$

Lemma 10. *Suppose the function f satisfies Assumption 1, and the stochastic oracles at the clients satisfy Assumptions 2, 3. Then, the iterates $\{\mathbf{w}^{(t)}\}_t$ generated by FedVARP satisfy*

$$\mathbb{E}_{\mathcal{S}^{(t)},\xi^{(t)}} \left[\frac{1}{N} \sum_{j=1}^{N} \left\| \nabla f_j(\mathbf{w}^{(t)}) - \mathbf{y}_j^{(t+1)} \right\|^2 \right] \leq \frac{M}{N} \left[\frac{\sigma^2}{\tau} + \frac{1}{N} \sum_{j=1}^{N} \mathbb{E}_{\xi^{(t)}} \left\| \nabla f_j(\mathbf{w}^{(t)}) - \mathbf{h}_j^{(t)} \right\|^2 \right]$$

$$+\left(1-\frac{M}{N}\right)\left[\left(1+\frac{1}{\beta}\right)\tilde{\eta}_s^2L^2\left\|\mathbf{v}^{(t-1)}\right\|^2+(1+\beta)\frac{1}{N}\sum_{j=1}^{N}\left\|\nabla f_j(\mathbf{w}^{(t-1)})-\mathbf{y}_j^{(t)}\right\|^2\right],$$

for any positive scalar β.

We also use the bound on $\frac{1}{N}\sum_{j=1}^{N}\mathbb{E}_{\xi^{(t)}}\left\|\nabla f_j(\mathbf{w}^{(t)})-\mathbf{h}_j^{(t)}\right\|^2$ from Lemma 6 in the previous section.

3.2 PROOF OF THEOREM 2

For the sake of completeness, first we state the complete theorem statement.

Theorem. *Suppose the function f satisfies Assumption 1, and the stochastic oracles at the clients satisfy Assumptions 2, 3. Further, the client learning rate η_c, and the server learning rate η_s are chosen such that $\eta_c \leq \frac{1}{10L\tau}$, $\eta_s\eta_c \leq \min\left\{\frac{M^{3/2}}{8L\tau N}, \frac{5M}{48\tau L}, \frac{1}{4L\tau}\right\}$. Then, the iterates $\{\mathbf{w}^{(t)}\}_t$ generated by* `FedVARP` *satisfy*

$$\min_{t\in[T]}\mathbb{E}\left\|\nabla f(\mathbf{w}^{(t)})\right\|^2 \leq \underbrace{\mathcal{O}\left(\frac{f(\mathbf{w}^{(0)})-f^*}{\eta_s\eta_c\tau T}\right)}_{\textit{Effect of initialization}}+\underbrace{\mathcal{O}\left(\frac{\eta_s\eta_c L\sigma^2}{M}+\eta_c^2L^2(\tau-1)\sigma^2\right)}_{\textit{Stochastic Gradient Error}}+\underbrace{\mathcal{O}\left(\eta_c^2L^2\tau(\tau-1)\sigma_g^2\right)}_{\textit{Client Drift Error}}$$

where $f^*=\arg\min_{\mathbf{x}} f(\mathbf{x})$.

Corollary 1. *Setting $\eta_c=\frac{1}{\sqrt{T}\tau L}$ and $\eta_s=\sqrt{\tau M}$,* `FedVARP` *converges to a stationary point of the global objective $f(\mathbf{w})$ at a rate given by,*

$$\min_{t\in\{0,\dots,T-1\}}\mathbb{E}\left\|\nabla f(\mathbf{w}^{(t)})\right\|^2 \leq \underbrace{\mathcal{O}\left(\frac{1}{\sqrt{M\tau T}}\right)}_{\textit{stochastic gradient error}}+\underbrace{\mathcal{O}\left(\frac{1}{T}\right)}_{\textit{client drift error}}.$$

Proof. We define the Lyapunov function as below for some α and $\frac{\tilde{\eta}_s^2L}{2}\leq\delta\leq\frac{\tilde{\eta}_s}{2}$. A necessary condition for this to be satisfied is $\tilde{\eta}_s=\eta_s\eta_c\tau\leq 1/L$. The precise choice of α,δ will be discussed later.

$$R^{(t+1)}\triangleq\mathbb{E}\left[f(\mathbf{w}^{(t+1)})+\left(\delta-\frac{\tilde{\eta}_s^2L}{2}\right)\left\|\mathbf{v}^{(t)}\right\|^2+\alpha\frac{1}{N}\sum_{j=1}^{N}\left\|\nabla f_j(\mathbf{w}^{(t)})-\mathbf{y}_j^{(t+1)}\right\|^2\right]. \tag{11}$$

Using Lemma 8,

$$\begin{aligned}R^{(t+1)}&\leq\mathbb{E}\Bigg[f(\mathbf{w}^{(t)})-\frac{\tilde{\eta}_s}{2}\left\|\nabla f(\mathbf{w}^{(t)})\right\|^2+\frac{\tilde{\eta}_s}{2}\frac{1}{N}\sum_{j=1}^{N}\mathbb{E}_{\xi^{(t)}}\left\|\nabla f_j(\mathbf{w}^{(t)})-\mathbf{h}_j^{(t)}\right\|^2-\frac{\tilde{\eta}_s}{2}\mathbb{E}_{\xi^{(t)}}\left\|\bar{\mathbf{h}}^{(t)}\right\|^2+\delta\mathbb{E}_{\mathcal{S}^{(t)},\xi^{(t)}}\left\|\mathbf{v}^{(t)}\right\|^2\\&\qquad+\alpha\frac{1}{N}\sum_{j=1}^{N}\mathbb{E}_{\mathcal{S}^{(t)},\xi^{(t)}}\left[\left\|\nabla f_j(\mathbf{w}^{(t)})-\mathbf{y}_j^{(t+1)}\right\|^2\right]\Bigg] && \text{(Jensen's inequality)}\\&\leq\mathbb{E}\Bigg[f(\mathbf{w}^{(t)})-\frac{\tilde{\eta}_s}{2}\left\|\nabla f(\mathbf{w}^{(t)})\right\|^2+\frac{\tilde{\eta}_s}{2}\frac{1}{N}\sum_{j=1}^{N}\mathbb{E}_{\xi^{(t)}}\left[\left\|\nabla f_j(\mathbf{w}^{(t)})-\mathbf{h}_j^{(t)}\right\|^2\right]+\delta\mathbb{E}_{\mathcal{S}^{(t)},\xi^{(t)}}\left[\left\|\mathbf{v}^{(t)}-\bar{\mathbf{h}}^{(t)}\right\|^2\right]\\&\qquad+\alpha\frac{1}{N}\sum_{j=1}^{N}\mathbb{E}_{\mathcal{S}^{(t)},\xi^{(t)}}\left[\left\|\nabla f_j(\mathbf{w}^{(t)})-\mathbf{y}_j^{(t+1)}\right\|^2\right]\Bigg],\end{aligned} \tag{12}$$

where for the last line we use that $\mathbb{E}_{\mathcal{S}^{(t)},\xi^{(t)}}\left[\left\|\mathbf{v}^{(t)}\right\|^2\right]=\mathbb{E}_{\mathcal{S}^{(t)},\xi^{(t)}}\left[\left\|\mathbf{v}^{(t)}-\bar{\mathbf{h}}^{(t)}\right\|^2\right]+\mathbb{E}_{\xi^{(t)}}\left[\left\|\bar{\mathbf{h}}^{(t)}\right\|^2\right]$ and $\delta\leq\frac{\tilde{\eta}_s}{2}$. Next, define $C^{(t)}\triangleq\frac{1}{N}\sum_{j=1}^{N}\mathbb{E}_{\xi^{(t)}}\left[\left\|\nabla f_j(\mathbf{w}^{(t)})-\mathbf{h}_j^{(t)}\right\|^2\right]$. Substituting the bounds from Lemma 9 and Lemma 10 in (12) we get,

$$\begin{aligned}R^{(t+1)} &\le \mathbb{E}\left[f(\mathbf{w}^{(t)}) - \frac{\tilde{\eta}_s}{2}\left\|\nabla f(\mathbf{w}^{(t)})\right\|^2 + \left(\frac{\tilde{\eta}_s}{2} + \frac{4\delta}{M}\frac{(N-M)}{(N-1)} + \frac{\alpha M}{N}\right)C^{(t)} + \left(\frac{\delta}{M} + \frac{\alpha M}{N}\right)\frac{\sigma^2}{\tau}\right] \\ &\quad + \left(\frac{4\delta\tilde{\eta}_s^2 L^2}{M}\frac{(N-M)}{(N-1)} + \alpha\left(1-\frac{M}{N}\right)\left(1+\frac{1}{\beta}\right)\tilde{\eta}_s^2 L^2\right)\mathbb{E}\left\|\mathbf{v}^{(t-1)}\right\|^2 \\ &\quad + \left(\frac{2\delta}{M}\frac{(N-M)}{(N-1)} + \alpha\left(1-\frac{M}{N}\right)(1+\beta)\right)\frac{1}{N}\sum_{j=1}^{N}\mathbb{E}\left\|\nabla f_j(\mathbf{w}^{(t-1)}) - \mathbf{y}_j^{(t)}\right\|^2. \end{aligned} \tag{13}$$

Choice of α, δ. Our goal is now to find a suitable δ and α such that,

$$\left(\frac{4\delta\tilde{\eta}_s^2 L^2}{M}\frac{(N-M)}{(N-1)} + \alpha\left(1-\frac{M}{N}\right)\left(1+\frac{1}{\beta}\right)\tilde{\eta}_s^2 L^2\right) \le \delta - \frac{\tilde{\eta}_s^2 L}{2},$$
$$\left(\frac{2\delta}{M}\frac{(N-M)}{(N-1)} + \alpha\left(1-\frac{M}{N}\right)(1+\beta)\right) \le \alpha$$

We define $A = \left(1-\frac{M}{N}\right)\left(1+\frac{1}{\beta}\right)$ and $B = \left(1-\frac{M}{N}\right)(1+\beta)$. In case of full client participation, $M = N$, and $A = B = 0$. The resulting condition on α is

$$\alpha \ge \frac{2\delta}{M(1-B)}\frac{(N-M)}{(N-1)}, \qquad \beta \le \frac{M}{N-M}.$$

We set $\alpha = \frac{2\delta}{M(1-B)}\frac{(N-M)}{(N-1)}$, and our condition on δ then reduces to,

$$\delta \ge \frac{\tilde{\eta}_s^2 L/2}{1 - \frac{4\tilde{\eta}_s^2 L^2}{M}\frac{(N-M)}{(N-1)} - \frac{2A\tilde{\eta}_s^2 L^2}{M(1-B)}\frac{(N-M)}{(N-1)}}$$

We want $\tilde{\eta}_s$ such that,

$$\frac{4\tilde{\eta}_s^2 L^2}{M}\frac{(N-M)}{(N-1)} + \frac{2A\tilde{\eta}_s^2 L^2}{M(1-B)}\frac{(N-M)}{(N-1)} \le \frac{1}{2}$$

A sufficient condition for this is

$$\frac{4\tilde{\eta}_s^2 L^2}{M} \le \frac{1}{4}, \qquad \text{and} \qquad \frac{2A\tilde{\eta}_s^2 L^2}{M(1-B)} \le \frac{1}{4}. \tag{14}$$

For $\beta = \frac{M}{2(N-M)}$, $B = 1 - \frac{M}{2N}$ and $A = \left(1-\frac{M}{N}\right)\left(\frac{2N}{M} - 1\right) \le \frac{2N}{M}$. A sufficient condition for (14) to be satisfied is

$$\tilde{\eta}_s \le \min\left\{\frac{\sqrt{M}}{4L}, \frac{M^{3/2}}{8LN}\right\} \quad \Rightarrow \quad \eta_s\eta_c \le \left\{\frac{\sqrt{M}}{4\tau L}, \frac{M^{3/2}}{8L\tau N}\right\}$$

With (14) we have $\delta \ge \tilde{\eta}_s^2 L$. We set $\delta = 2\tilde{\eta}_s^2 L$ which gives us $\alpha = \frac{8N\tilde{\eta}_s^2 L}{M^2}\frac{(N-M)}{(N-1)}$. Since $\delta \le \frac{\tilde{\eta}_s}{2}$, we also need $\tilde{\eta}_s \le \frac{1}{4L}$. With this choice of α, δ, from (13) we get

$$\begin{aligned}R^{(t+1)} &\le R^{(t)} - \frac{\tilde{\eta}_s}{2}\mathbb{E}\left\|\nabla f(\mathbf{w}^{(t)})\right\|^2 + \left(\frac{\tilde{\eta}_s}{2} + \frac{4\delta}{M} + \frac{\alpha M}{N}\right)C^{(t)} + \left(\frac{\delta}{M} + \frac{\alpha M}{N}\right)\frac{\sigma^2}{\tau} \\ &\le R^{(t)} - \frac{\tilde{\eta}_s}{2}\left\|\nabla f(\mathbf{w}^{(t)})\right\|^2 + 3\tilde{\eta}_s C^{(t)} + \frac{10\tilde{\eta}_s^2 L}{M}\frac{\sigma^2}{\tau} \end{aligned} \tag{15}$$

where we use the condition that $\tilde{\eta}_s \le \frac{5M}{48L}$. Further, we can bound $C^{(t)} = \frac{1}{N}\sum_{j=1}^{N}\mathbb{E}_{\xi^{(t)}}\left[\left\|\nabla f_j(\mathbf{w}^{(t)}) - \mathbf{h}_j^{(t)}\right\|^2\right]$ using Lemma 6, which gives us

$$R^{(t+1)} \le R^{(t)} - \frac{\tilde{\eta}_s}{2}\left(1 - 48\eta_c^2 L^2\tau(\tau-1)\right)\left\|\nabla f(\mathbf{w}^{(t)})\right\|^2 + 6\tilde{\eta}_s\eta_c^2 L^2(\tau-1)\left[\sigma^2 + 4\tau\sigma_g^2\right] + \frac{10\tilde{\eta}_s L\sigma^2}{M\tau}.$$

Using the condition on η_c that $\eta_c \leq \frac{1}{10L\tau}$, and unrolling the recursion we get,

$$R^{(t)} \leq R^{(1)} + \sum_{t=1}^{(t-1)} \left(-\frac{\tilde{\eta}_s}{4} \left\| \nabla f(\mathbf{w}^{(t)}) \right\|^2 + 6\tilde{\eta}_s \eta_c^2 L^2 (\tau - 1) \left[\sigma^2 + 4\tau\sigma_g^2 \right] + \frac{10\tilde{\eta}_s^2 L \sigma^2}{M\tau} \right). \tag{16}$$

Next, we bound $R^{(1)}$. Using (12) and (20) we can bound $R^{(1)}$ as follows,

$$\begin{aligned} R^{(1)} \leq & f(\mathbf{w}^{(0)}) - \frac{\tilde{\eta}_s}{2} \left\| \nabla f(\mathbf{w}^{(0)}) \right\|^2 + \left(\frac{\tilde{\eta}_s}{2} + \frac{\alpha M}{N} \right) C^{(0)} + \frac{\alpha M}{N} \frac{\sigma^2}{\tau} \\ & + \delta \mathbb{E}_{\xi^{(0)}, \mathcal{S}^{(0)}} \left[\left\| \mathbf{v}^{(0)} - \bar{\mathbf{h}}^{(0)} \right\|^2 \right] + \alpha \left(1 - \frac{M}{N} \right) \frac{1}{N} \sum_{i=1}^{N} \left\| \nabla f_i(\mathbf{w}^{(0)}) \right\|^2 \\ \leq & f(\mathbf{w}^{(0)}) - \frac{\tilde{\eta}_s}{4} \left\| \nabla f(\mathbf{w}^{(0)}) \right\|^2 + 4\tilde{\eta}_s \eta_c^2 L^2 (\tau - 1) \left[\sigma^2 + 4\tau\sigma_g^2 \right] + \frac{8\tilde{\eta}_s^2 L \sigma^2}{M\tau} \\ & + \delta \mathbb{E}_{\xi^{(0)}, \mathcal{S}^{(0)}} \left[\left\| \mathbf{v}^{(0)} - \bar{\mathbf{h}}^{(0)} \right\|^2 \right] + \alpha \left(1 - \frac{M}{N} \right) \frac{1}{N} \sum_{i=1}^{N} \left\| \nabla f_i(\mathbf{w}^{(0)}) \right\|^2. \end{aligned} \tag{17}$$

Substituting the bound on $R^{(1)}$ from (17) in (16), and using $t = T$ we get

$$\begin{aligned} R^{(T)} \leq & f(\mathbf{w}^{(0)}) - \sum_{t=0}^{(t-1)} \left(\frac{\tilde{\eta}_s}{4} \left\| \nabla f(\mathbf{w}^{(t)}) \right\|^2 + 6\tilde{\eta}_s \eta_c^2 L^2 (\tau - 1) \left[\sigma^2 + 4\tau\sigma_g^2 \right] + \frac{10\tilde{\eta}_s^2 L \sigma^2}{M\tau} \right) \\ & + 2\tilde{\eta}_s^2 L \mathbb{E}_{\xi^{(0)}, \mathcal{S}^{(0)}} \left[\left\| \mathbf{v}^{(0)} - \bar{\mathbf{h}}^{(0)} \right\|^2 \right] + \frac{8N\tilde{\eta}_s^2 L}{M^2} \left(1 - \frac{M}{N} \right) \frac{1}{N} \sum_{i=1}^{N} \left\| \nabla f(\mathbf{w}^{(0)}) \right\|^2 \end{aligned}$$

Rearranging the terms, we get

$$\begin{aligned} \min_{t \in [0, T-1]} \mathbb{E} \left\| \nabla f(\mathbf{w}^{(t)}) \right\|^2 \leq & \frac{1}{T} \sum_{t=0}^{T-1} \mathbb{E} \left\| \nabla f(\mathbf{w}^{(t)}) \right\|^2 \\ \leq & \frac{4(f(\mathbf{w}^{(0)}) - f^*)}{\tilde{\eta}_s T} + \frac{40\tilde{\eta}_s L \sigma^2}{M\tau} + 24\eta_c^2 L^2 (\tau - 1)\sigma^2 + 96\eta_c^2 L^2 \tau (\tau - 1)\sigma_g^2 \\ & + \frac{1}{T} \left[8\tilde{\eta}_s L \mathbb{E}_{\xi^{(0)}, \mathcal{S}^{(0)}} \left[\left\| \mathbf{v}^{(0)} - \bar{\mathbf{h}}^{(0)} \right\|^2 \right] + \frac{32N\tilde{\eta}_s L}{M^2} \left(1 - \frac{M}{N} \right) \frac{1}{N} \sum_{i=1}^{N} \left\| \nabla f(\mathbf{w}^{(0)}) \right\|^2 \right] \\ = & \mathcal{O}\left(\frac{(f(\mathbf{w}^{(0)}) - f^*)}{\tilde{\eta}_s T} \right) + \mathcal{O}\left(\frac{\tilde{\eta}_s L \sigma^2}{M\tau} \right) + \mathcal{O}(\eta_c^2 L^2 (\tau - 1)\sigma^2 + \eta_c^2 L^2 \tau (\tau - 1)\sigma_g^2) \end{aligned}$$

□

3.3 PROOFS OF THE INTERMEDIATE LEMMAS

Proof of Lemma 8. Using L-smoothness (Assumption 1) of f,

$$f(\mathbf{w}^{(t+1)}) \leq f(\mathbf{w}^{(t)}) - \tilde{\eta}_s \left\langle \nabla f(\mathbf{w}^{(t)}), \mathbf{v}^{(t)} \right\rangle + \frac{\tilde{\eta}_s^2 L}{2} \left\| \mathbf{v}^{(t)} \right\|^2.$$

Taking expectation only over the randomness in the t-th round: due to client sampling (inherent in $\mathcal{S}^{(t)}$) and due to stochastic gradients (inherent in $\xi^{(t)} \triangleq \{\xi_i^{(t,k)}\}_{i,k}$), we get

$$\begin{aligned} \mathbb{E}_{\mathcal{S}^{(t)}, \xi^{(t)}} \left[f(\mathbf{w}^{(t+1)}) \right] & \leq f(\mathbf{w}^{(t)}) - \tilde{\eta}_s \mathbb{E}_{\mathcal{S}^{(t)}, \xi^{(t)}} \left\langle \nabla f(\mathbf{w}^{(t)}), \mathbf{v}^{(t)} \right\rangle + \frac{\tilde{\eta}_s^2 L}{2} \mathbb{E}_{\mathcal{S}^{(t)}, \xi^{(t)}} \left\| \mathbf{v}^{(t)} \right\|^2 \\ & \overset{(a)}{=} f(\mathbf{w}^{(t)}) - \tilde{\eta}_s \mathbb{E}_{\xi^{(t)}} \left\langle \nabla f(\mathbf{w}^{(t)}), \bar{\mathbf{h}}^{(t)} \right\rangle + \frac{\tilde{\eta}_s^2 L}{2} \mathbb{E}_{\mathcal{S}^{(t)}, \xi^{(t)}} \left\| \mathbf{v}^{(t)} \right\|^2 \end{aligned}$$

$$= f(\mathbf{w}^{(t)}) - \frac{\tilde{\eta}_s}{2}\left[\left\|\nabla f(\mathbf{w}^{(t)})\right\|^2 + \mathbb{E}_{\xi^{(t)}}\left\|\bar{\mathbf{h}}^{(t)}\right\|^2 - \mathbb{E}_{\xi^{(t)}}\left\|\nabla f(\mathbf{w}^{(t)}) - \bar{\mathbf{h}}^{(t)}\right\|^2\right] + \frac{\tilde{\eta}_s^2 L}{2}\mathbb{E}_{\mathcal{S}^{(t)},\xi^{(t)}}\left\|\mathbf{v}^{(t)}\right\|^2,$$

where (a) follows since

$$\begin{aligned}
\mathbb{E}_{\mathcal{S}^{(t)},\xi^{(t)}}\left[\mathbf{v}^{(t)}\right] &= \mathbb{E}_{\mathcal{S}^{(t)},\xi^{(t)}}\left[\frac{1}{\mathcal{S}^{(t)}}\sum_{i\in\mathcal{S}^{(t)}}\Delta_i^{(t)}\right] - \mathbb{E}_{\mathcal{S}^{(t)}}\left[\frac{1}{\mathcal{S}^{(t)}}\sum_{i\in\mathcal{S}^{(t)}}\mathbf{y}_i^{(t)}\right] + \mathbf{y}^{(t)} \\
&= \frac{1}{N}\sum_{i=1}^{N}\mathbb{E}_{\xi^{(t)}}\left[\Delta_i^{(t)}\right] - \mathbf{y}^{(t)} + \mathbf{y}^{(t)} && \text{(uniform sampling of clients)} \\
&= \frac{1}{N}\sum_{i=1}^{N}\mathbb{E}_{\xi^{(t)}}\left[\frac{1}{\tau}\sum_{k=0}^{\tau-1}\nabla f_i(\mathbf{w}_i^{(t,k)}, \xi_i^{(t,k)})\right] \\
&= \mathbb{E}\left[\bar{\mathbf{h}}^{(t)}\right].
\end{aligned}$$

□

Proof of Lemma 9.

$$\begin{aligned}
&\mathbb{E}_{\mathcal{S}^{(t)},\xi^{(t)}}\left\|\mathbf{v}^{(t)} - \bar{\mathbf{h}}^{(t)}\right\|^2 \\
&= \mathbb{E}_{\mathcal{S}^{(t)},\xi^{(t)}}\left\|\frac{1}{M}\sum_{i\in\mathcal{S}^{(t)}}\left(\Delta_i^{(t)} - \mathbf{y}_i^{(t)} + \frac{1}{N}\sum_{j=1}^{N}\mathbf{y}_j^{(t)} - \bar{\mathbf{h}}^{(t)}\right)\right\|^2 && \text{(Server update direction in } \texttt{FedVARP}\text{)} \\
&= \mathbb{E}_{\mathcal{S}^{(t)},\xi^{(t)}}\left\|\frac{1}{M}\sum_{i\in\mathcal{S}^{(t)}}\left(\Delta_i^{(t)} - \mathbf{h}_i^{(t)} + \mathbf{h}_i^{(t)} - \mathbf{y}_i^{(t)} + \frac{1}{N}\sum_{j=1}^{N}\mathbf{y}_j^{(t)} - \bar{\mathbf{h}}^{(t)}\right)\right\|^2 \\
&\overset{(a)}{=} \mathbb{E}_{\mathcal{S}^{(t)},\xi^{(t)}}\left\|\frac{1}{M}\sum_{i\in\mathcal{S}^{(t)}}\left(\Delta_i^{(t)} - \mathbf{h}_i^{(t)}\right)\right\|^2 + \mathbb{E}_{\mathcal{S}^{(t)},\xi^{(t)}}\left\|\frac{1}{M}\sum_{i\in\mathcal{S}^{(t)}}\left(\mathbf{h}_i^{(t)} - \mathbf{y}_i^{(t)} + \frac{1}{N}\sum_{j=1}^{N}\mathbf{y}_j^{(t)} - \bar{\mathbf{h}}^{(t)}\right)\right\|^2 \\
&\overset{(b)}{=} \frac{1}{M^2}\mathbb{E}_{\mathcal{S}^{(t)},\xi^{(t)}}\left[\sum_{i\in\mathcal{S}^{(t)}}\left\|\Delta_i^{(t)} - \mathbf{h}_i^{(t)}\right\|^2\right] + \mathbb{E}_{\mathcal{S}^{(t)},\xi^{(t)}}\left\|\frac{1}{M}\sum_{i\in\mathcal{S}^{(t)}}\left(\mathbf{h}_i^{(t)} - \mathbf{y}_i^{(t)} + \frac{1}{N}\sum_{j=1}^{N}\mathbf{y}_j^{(t)} - \bar{\mathbf{h}}^{(t)}\right)\right\|^2 \\
&\overset{(c)}{\leq} \frac{\sigma^2}{\tau M} + \underbrace{\mathbb{E}_{\mathcal{S}^{(t)},\xi^{(t)}}\left\|\frac{1}{M}\sum_{i\in\mathcal{S}^{(t)}}\left(\mathbf{h}_i^{(t)} - \mathbf{y}_i^{(t)} + \frac{1}{N}\sum_{j=1}^{N}\mathbf{y}_j^{(t)} - \bar{\mathbf{h}}^{(t)}\right)\right\|^2}_{T_1}.
\end{aligned} \tag{18}$$

where (a) follows from the following reasoning.

$$\begin{aligned}
&\mathbb{E}_{\mathcal{S}^{(t)},\xi^{(t)}}\left\langle\frac{1}{M}\sum_{i\in\mathcal{S}^{(t)}}\left(\Delta_i^{(t)} - \mathbf{h}_i^{(t)}\right), \frac{1}{M}\sum_{i\in\mathcal{S}^{(t)}}\left(\mathbf{h}_i^{(t)} - \mathbf{y}_i^{(t)} + \frac{1}{N}\sum_{j=1}^{N}\mathbf{y}_j^{(t)} - \bar{\mathbf{h}}^{(t)}\right)\right\rangle \\
&= \frac{1}{M^2}\mathbb{E}_{\mathcal{S}^{(t)}}\left[\sum_{i\in\mathcal{S}^{(t)}}\mathbb{E}_{\xi^{(t)}}\left\langle\Delta_i^{(t)} - \mathbf{h}_i^{(t)}, \mathbf{h}_i^{(t)} - \mathbf{y}_i^{(t)} + \mathbf{y}^{(t)} - \bar{\mathbf{h}}^{(t)}\right\rangle + \sum_{i\in\mathcal{S}^{(t)}}\sum_{\substack{j\in\mathcal{S}^{(t)}\\ i\neq j}}\mathbb{E}_{\xi^{(t)}}\left\langle\Delta_i^{(t)} - \mathbf{h}_i^{(t)}, \mathbf{h}_j^{(t)} - \mathbf{y}_j^{(t)} + \mathbf{y}^{(t)} - \bar{\mathbf{h}}^{(t)}\right\rangle\right] \\
&= \frac{1}{M^2}\mathbb{E}_{\mathcal{S}^{(t)}}\left[\sum_{i\in\mathcal{S}^{(t)}}\mathbb{E}_{\xi^{(t)}}\left\langle\Delta_i^{(t)} - \mathbf{h}_i^{(t)}, \mathbf{h}_i^{(t)} - \mathbf{y}_i^{(t)} + \mathbf{y}^{(t)} - \bar{\mathbf{h}}^{(t)}\right\rangle\right] \\
&\qquad\qquad \text{(Assumption 2; independence of stochastic gradients across clients)}
\end{aligned}$$

$$
\begin{aligned}
&= \frac{1}{M^2}\mathbb{E}_{\mathcal{S}^{(t)}}\left[\sum_{i\in\mathcal{S}^{(t)}}\mathbb{E}_{\xi^{(t)}}\left\langle \Delta_i^{(t)} - \mathbf{h}_i^{(t)}, \mathbf{h}_i^{(t)} - \bar{\mathbf{h}}^{(t)}\right\rangle\right] \qquad \text{(since } \{\mathbf{y}_i^{(t)}\} \text{ are independent of } \mathcal{S}^{(t)}, \xi^{(t)}\text{)} \\
&= \frac{1}{(\tau M)^2}\mathbb{E}_{\mathcal{S}^{(t)}}\left[\sum_{i\in\mathcal{S}^{(t)}}\mathbb{E}_{\xi^{(t)}}\left\langle \sum_{k=0}^{\tau-1}\left(\nabla f_i(\mathbf{w}_i^{(t,k)}, \xi_i^{(t,k)}) - \nabla f_i(\mathbf{w}_i^{(t,k)})\right), \sum_{j=0}^{\tau-1}\nabla f_i(\mathbf{w}_i^{(t,j)}) - \frac{1}{N}\sum_{\ell=1}^{N}\sum_{j=0}^{\tau-1}\nabla f_\ell(\mathbf{w}_\ell^{(t,j)})\right\rangle\right] \\
&= \frac{1}{(\tau M)^2}\mathbb{E}_{\mathcal{S}^{(t)}}\left[\sum_{i\in\mathcal{S}^{(t)}}\mathbb{E}_{\xi^{(t)}}\left[\sum_{k=0}^{\tau-1}\left\langle \nabla f_i(\mathbf{w}_i^{(t,k)}, \xi_i^{(t,k)}) - \nabla f_i(\mathbf{w}_i^{(t,k)}), \nabla f_i(\mathbf{w}_i^{(t,k)}) - \frac{1}{N}\sum_{\ell=1}^{N}\nabla f_\ell(\mathbf{w}_\ell^{(t,k)})\right\rangle\right]\right] \\
&\quad + \frac{1}{(\tau M)^2}\mathbb{E}_{\mathcal{S}^{(t)}}\left[\sum_{i\in\mathcal{S}^{(t)}}\mathbb{E}_{\xi^{(t)}}\left[\sum_{k=0}^{\tau-1}\sum_{j\neq k}\left\langle \nabla f_i(\mathbf{w}_i^{(t,k)}, \xi_i^{(t,k)}) - \nabla f_i(\mathbf{w}_i^{(t,k)}), \nabla f_i(\mathbf{w}_i^{(t,j)}) - \frac{1}{N}\sum_{\ell=1}^{N}\nabla f_\ell(\mathbf{w}_\ell^{(t,j)})\right\rangle\right]\right] \\
&= \frac{1}{(\tau M)^2}\mathbb{E}_{\mathcal{S}^{(t)}}\left[\sum_{i\in\mathcal{S}^{(t)}}\mathbb{E}_{\xi^{(t)}}\left[\sum_{k=0}^{\tau-1}\left\langle \mathbb{E}\left[\nabla f_i(\mathbf{w}_i^{(t,k)}, \xi_i^{(t,k)}) - \nabla f_i(\mathbf{w}_i^{(t,k)}) | \mathbf{w}_i^{(t,k)}\right], \nabla f_i(\mathbf{w}_i^{(t,k)}) - \frac{1}{N}\sum_{\ell=1}^{N}\nabla f_\ell(\mathbf{w}_\ell^{(t,k)})\right\rangle\right]\right] \\
&\quad + \frac{2}{(\tau M)^2}\mathbb{E}_{\mathcal{S}^{(t)}}\left[\sum_{i\in\mathcal{S}^{(t)}}\mathbb{E}_{\xi^{(t)}}\left[\sum_{j<k}\left\langle \mathbb{E}\left[\nabla f_i(\mathbf{w}_i^{(t,k)}, \xi_i^{(t,k)}) - \nabla f_i(\mathbf{w}_i^{(t,k)}) | \mathbf{w}_i^{(t,j)}\right], \nabla f_i(\mathbf{w}_i^{(t,j)}) - \frac{1}{N}\sum_{\ell=1}^{N}\nabla f_\ell(\mathbf{w}_\ell^{(t,j)})\right\rangle\right]\right] \\
&= 0.
\end{aligned}
$$

Further, (b) follows since $\mathbb{E}_{\xi^{(t)}}\left[\langle \Delta_i^{(t)} - \mathbf{h}_i^{(t)}, \Delta_j^{(t)} - \mathbf{h}_j^{(t)}\rangle\right] = 0$ for $i \neq j$. Finally, (c) follows from the following reasoning.

$$
\begin{aligned}
\mathbb{E}\left\|\Delta_i^{(t)} - \mathbf{h}_i^{(t)}\right\|^2 &= \frac{1}{\tau^2}\mathbb{E}\left\|\sum_{k=0}^{\tau-1}\left(\nabla f_i(\mathbf{w}_i^{(t,k)}, \xi_i^{(t,k)}) - \nabla f_i(\mathbf{w}_i^{(t,k)})\right)\right\|^2 \\
&= \frac{1}{\tau^2}\mathbb{E}\Bigg[\sum_{k=0}^{\tau-1}\left\|\nabla f_i(\mathbf{w}_i^{(t,k)}, \xi_i^{(t,k)}) - \nabla f_i(\mathbf{w}_i^{(t,k)})\right\|^2 \\
&\qquad + 2\sum_{j<k}\left\langle \mathbb{E}\left[\nabla f_i(\mathbf{w}_i^{(t,k)}, \xi_i^{(t,k)}) - \nabla f_i(\mathbf{w}_i^{(t,k)}) | \mathbf{w}_i^{(t,j)}\right], \nabla f_i(\mathbf{w}_i^{(t,j)}, \xi_i^{(t,j)}) - \nabla f_i(\mathbf{w}_i^{(t,j)})\right\rangle\Bigg] \\
&\leq \frac{\sigma^2}{\tau}. \qquad \text{(Assumption 2)}
\end{aligned}
$$

Next, we bound T_1 in (18).

$$
\begin{aligned}
&\mathbb{E}_{\mathcal{S}^{(t)},\xi^{(t)}}\left\|\frac{1}{M}\sum_{j\in\mathcal{S}^{(t)}}\left(\mathbf{h}_i^{(t)} - \mathbf{y}_i^{(t)} + \mathbf{y}^{(t)} - \bar{\mathbf{h}}^{(t)}\right)\right\|^2 \\
&= \mathbb{E}_{\mathcal{S}^{(t)},\xi^{(t)}}\left\|\frac{1}{M}\sum_{i\in\mathcal{S}^{(t)}}\left(\mathbf{h}_i^{(t)} - \nabla f_i(\mathbf{w}^{(t-1)}) - (\bar{\mathbf{h}}^{(t)} - \nabla f(\mathbf{w}^{(t-1)})) + \left(\nabla f_i(\mathbf{w}^{(t-1)}) - \mathbf{y}_i^{(t)} + \mathbf{y}^{(t)} - \nabla f(\mathbf{w}^{(t-1)})\right)\right)\right\|^2 \\
&\leq 2\mathbb{E}_{\mathcal{S}^{(t)},\xi^{(t)}}\left\|\frac{1}{M}\sum_{i\in\mathcal{S}^{(t)}}\left(\mathbf{h}_i^{(t)} - \nabla f_i(\mathbf{w}^{(t-1)}) - \left(\bar{\mathbf{h}}^{(t)} - \nabla f(\mathbf{w}^{(t-1)})\right)\right)\right\|^2 \\
&\qquad + 2\mathbb{E}_{\mathcal{S}^{(t)}}\left[\left\|\frac{1}{M}\sum_{i\in\mathcal{S}^{(t)}}\left(\nabla f_i(\mathbf{w}^{(t-1)}) - \mathbf{y}_i^{(t)} + \mathbf{y}^{(t)} - \nabla f(\mathbf{w}^{(t-1)})\right)\right\|^2\right] \\
&= \frac{2(N-M)}{M(N-1)N}\sum_{i=1}^{N}\mathbb{E}_{\xi^{(t)}}\left[\left\|\mathbf{h}_i^{(t)} - \nabla f_i(\mathbf{w}^{(t-1)}) - \bar{\mathbf{h}}^{(t)} + \nabla f(\mathbf{w}^{(t-1)})\right\|^2\right]
\end{aligned}
$$

$$
\begin{aligned}
&\quad + \frac{2(N-M)}{M(N-1)N} \sum_{i=1}^{N} \left\| \nabla f_i(\mathbf{w}^{(t-1)}) - \mathbf{y}_i^{(t)} + \mathbf{y}^{(t)} - \nabla f(\mathbf{w}^{(t-1)}) \right\|^2 && \text{(Lemma 4)} \\
&\overset{(d)}{\leq} \frac{2(N-M)}{M(N-1)N} \left[\sum_{i=1}^{N} \mathbb{E}_{\xi^{(t)}} \left[\left\| \mathbf{h}_i^{(t)} - \nabla f_i(\mathbf{w}^{(t-1)}) \right\|^2 \right] + \sum_{i=1}^{N} \left\| \nabla f_i(\mathbf{w}^{(t-1)}) - \mathbf{y}_i^{(t)} \right\|^2 \right] && (\because \text{Var}(X) \leq \mathbb{E}\left[X^2\right]) \\
&= \frac{2(N-M)}{M(N-1)N} \left[\sum_{i=1}^{N} \mathbb{E}_{\xi^{(t)}} \left[\left\| \mathbf{h}_i^{(t)} - \nabla f_i(\mathbf{w}^{(t)}) + \nabla f_i(\mathbf{w}^{(t)}) - \nabla f_i(\mathbf{w}^{(t-1)}) \right\|^2 \right] + \sum_{i=1}^{N} \left\| \nabla f_i(\mathbf{w}^{(t-1)}) - \mathbf{y}_i^{(t)} \right\|^2 \right] \\
&\leq \frac{2(N-M)}{M(N-1)N} \left[2\sum_{i=1}^{N} \mathbb{E}_{\xi^{(t)}} \left[\left\| \mathbf{h}_i^{(t)} - \nabla f_i(\mathbf{w}^{(t)}) \right\|^2 \right] + 2N\tilde{\eta}_s^2 L^2 \left\| \mathbf{v}^{(t-1)} \right\|^2 + \sum_{i=1}^{N} \left\| \nabla f_i(\mathbf{w}^{(t-1)}) - \mathbf{y}_i^{(t)} \right\|^2 \right]. && (19)
\end{aligned}
$$

Finally, substituting (19) in (18), we get the result in the lemma. □

Proof of Lemma 10.

$$
\begin{aligned}
&\mathbb{E}_{\mathcal{S}^{(t)},\xi^{(t)}} \left[\frac{1}{N} \sum_{j=1}^{N} \left\| \nabla f_j(\mathbf{w}^{(t)}) - \mathbf{y}_j^{(t+1)} \right\|^2 \right] \\
&= \frac{1}{N} \sum_{j=1}^{N} \mathbb{E}_{\mathcal{S}^{(t)},\xi^{(t)}} \left\| \nabla f_j(\mathbf{w}^{(t)}) - \mathbf{y}_j^{(t+1)} \right\|^2 \\
&= \frac{1}{N} \sum_{j=1}^{N} \left[\frac{M}{N} \mathbb{E}_{\xi^{(t)}} \left[\left\| \nabla f_j(\mathbf{w}^{(t)}) - \Delta_j^{(t)} \right\|^2 \right] + \left(1 - \frac{M}{N}\right) \left\| \nabla f_j(\mathbf{w}^{(t)}) - \mathbf{y}_j^{(t)} \right\|^2 \right] && (20) \\
&= \frac{1}{N} \sum_{j=1}^{N} \left[\frac{M}{N} \mathbb{E}_{\xi^{(t)}} \left[\left\| \nabla f_j(\mathbf{w}^{(t)}) - \Delta_j^{(t)} \right\|^2 \right] + \left(1 - \frac{M}{N}\right) \left\| \nabla f_j(\mathbf{w}^{(t)}) - \nabla f_j(\mathbf{w}^{(t-1)}) + \nabla f_j(\mathbf{w}^{(t-1)}) - \mathbf{y}_j^{(t)} \right\|^2 \right] \\
&\leq \frac{1}{N} \sum_{j=1}^{N} \left[\frac{M\sigma^2}{N\tau} + \frac{M}{N} \mathbb{E}_{\xi^{(t)}} \left\| \nabla f_j(\mathbf{w}^{(t)}) - \mathbf{h}_j^{(t)} \right\|^2 + \left(1 - \frac{M}{N}\right) \left\| \nabla f_j(\mathbf{w}^{(t)}) - \nabla f_j(\mathbf{w}^{(t-1)}) + \nabla f_j(\mathbf{w}^{(t-1)}) - \mathbf{y}_j^{(t)} \right\|^2 \right] \\
&\leq \frac{M}{N} \left[\frac{\sigma^2}{\tau} + \frac{1}{N} \sum_{j=1}^{N} \mathbb{E}_{\xi^{(t)}} \left\| \nabla f_j(\mathbf{w}^{(t)}) - \mathbf{h}_j^{(t)} \right\|^2 \right] \\
&+ \left(1 - \frac{M}{N}\right) \left[\left(1 + \frac{1}{\beta}\right) \tilde{\eta}_s^2 L^2 \left\| \mathbf{v}^{(t-1)} \right\|^2 + (1+\beta) \frac{1}{N} \sum_{j=1}^{N} \left\| \nabla f_j(\mathbf{w}^{(t-1)}) - \mathbf{y}_j^{(t)} \right\|^2 \right]. && (\beta \text{ is a positive constant})
\end{aligned}
$$

□

4 CONVERGENCE RESULT FOR CLUSTERFEDVARP (THEOREM 3)

In this section we prove the convergence result for ClusterFedVARP in Theorem 3, and provide the complexity and communication guarantees.

We organize this section as follows. First, in 4.1 we present some intermediate results, which we use to prove the main theorem. Next, in 4.2, we present the proof of Theorem 3, which is followed by the proofs of the intermediate results in 4.3.

4.1 INTERMEDIATE LEMMAS

The proof of ClusterFedVARP follows closely the proof of FedVARP. We borrow Lemma 8 from Section 3, and the next lemma is analogous to Lemma 9.

Algorithm 1 `ClusterFedVARP`

```
Input: initial model $\mathbf{w}^{(0)}$, server learning rate $\eta_s$, client learning rate $\eta$, local SGD steps $\tau$, $\tilde{\eta}_s = \eta_s \eta_c \tau$, number of rounds $T$, number of clusters $K$, initial cluster states $\mathbf{y}_k^{(0)} = \mathbf{0}$ for all $k \in [K]$, cluster identities $c_i \in [K]$ for all $i \in [N]$, cluster sets $\mathcal{C}_k = \{i : c_i = k\}$ for all $k \in [K]$
for $t = 1, 2, \dots, T$ do
    Sample $\mathcal{S}^{(t)} \subseteq [N]$ uniformly without replacement
    for $i \in \mathcal{S}^{(t)}$ do
        $\Delta_i^{(t)} \leftarrow \texttt{LocalSGD}(i, \mathbf{w}^{(t)}, \tau, \eta)$
    end for
    // At Server:
    $\mathbf{v}^{(t)} = \frac{1}{|\mathcal{S}^{(t)}|} \sum_{i \in \mathcal{S}^{(t)}} \left( \Delta_i^{(t)} - \mathbf{y}_{c_i}^{(t)} \right) + \frac{1}{N} \sum_{j=1}^{N} \mathbf{y}_{c_j}^{(t)}$
    $\mathbf{w}^{(t+1)} = \mathbf{w}^{(t)} - \tilde{\eta}_s \mathbf{v}^{(t)}$
    //State update
    for $k \in [K]$ do
        $\mathbf{y}_k^{(t+1)} = \begin{cases} \dfrac{\sum_{i \in \mathcal{S}^{(t)} \cap \mathcal{C}_k} \Delta_i^{(t)}}{|\mathcal{S}^{(t)} \cap \mathcal{C}_k|} & \text{if } |\mathcal{S}^{(t)} \cap \mathcal{C}_k| \neq 0 \\ \mathbf{y}_k^{(t)} & \text{otherwise} \end{cases}$
    end for
end for
procedure $\textsc{LocalSGD}(i, \mathbf{w}^{(t)}, \tau, \eta)$
    Set $\mathbf{w}_i^{(t,0)} = \mathbf{w}^{(t)}$
    for $k = 0, 1 \dots, \tau - 1$ do
        Compute stochastic gradient $\nabla f_i(\mathbf{w}_i^{(t,k)}, \xi_i^{(t,k)})$
        $\mathbf{w}_i^{(t,k+1)} = \mathbf{w}_i^{(t,k)} - \eta_c \nabla f_i(\mathbf{w}_i^{(t,k)}, \xi_i^{(t,k)})$
    end for
    Return $(\mathbf{w}^{(t)} - \mathbf{w}_i^{(t,\tau)}) / \eta_c \tau$
end procedure
```

Lemma 11. *Suppose the function f satisfies Assumption 1, and the stochastic oracles at the clients satisfy Assumptions 2. Then, the iterates $\{\mathbf{w}^{(t)}\}_t$ generated by* `FedVARP` *satisfy*

$$\mathbb{E}_{\mathcal{S}^{(t)},\xi^{(t)}}\left\|\mathbf{v}^{(t)}-\bar{\mathbf{h}}^{(t)}\right\|^2 \leq \frac{\sigma^2}{M\tau}+\frac{4(N-M)}{M(N-1)}\left[\frac{1}{N}\sum_{i=1}^{N}\mathbb{E}_{\xi^{(t)}}\left[\left\|\mathbf{h}_i^{(t)}-\nabla f_i(\mathbf{w}^{(t)})\right\|^2\right]+\tilde{\eta}_s^2L^2\left\|\mathbf{v}^{(t-1)}\right\|^2\right]$$
$$+\frac{2(N-M)}{M(N-1)}\frac{1}{N}\sum_{i=1}^{N}\left\|\nabla f_i(\mathbf{w}^{(t-1)})-\mathbf{y}_{c_i}^{(t)}\right\|^2.$$

Lemma 12. *Suppose the function f satisfies Assumption 1, and the stochastic oracles at the clients satisfy Assumptions 2, 3. Then, the iterates $\{\mathbf{w}^{(t)}\}_t$ generated by* `FedVARP` *satisfy*

$$\mathbb{E}_{\xi^{(t)},\mathcal{S}^{(t)}}\left[\frac{1}{N}\sum_{j=1}^{N}\left\|\nabla f_j(\mathbf{w}^{(t)})-\mathbf{y}_{c_j}^{(t+1)}\right\|^2\right] \leq 4(1-p)\left[\sigma_K^2+\frac{\sigma^2}{\tau}+\frac{1}{N}\sum_{j=1}^{N}\mathbb{E}_t\left\|\nabla f_j(\mathbf{w}^{(t)})-\mathbf{h}_j^{(t)}\right\|^2\right]$$
$$+p\left[\left(1+\frac{1}{\beta}\right)\tilde{\eta}_s^2L^2\left\|\mathbf{v}^{(t-1)}\right\|^2+(1+\beta)\frac{1}{N}\sum_{j=1}^{N}\left\|\nabla f_j(\mathbf{w}^{(t-1)})-\mathbf{y}_{c_j}^{(t)}\right\|^2\right].$$

for any positive scalar β. Note that keeping $r=1$ (which implies $\sigma_K^2=0$) we recover our earlier result in Lemma 10 (upto multiplicative constants).

We also use the bound on $\frac{1}{N}\sum_{j=1}^{N}\mathbb{E}_{\xi^{(t)}}\left\|\nabla f_j(\mathbf{w}^{(t)})-\mathbf{h}_j^{(t)}\right\|^2$ from Lemma 6 in the previous section.

4.2 PROOF OF THEOREM 3

For the sake of completeness, first we state the complete theorem statement.

Theorem. *Suppose the function f satisfies Assumption 1, and the individual client functions satisfy Assumptions 2, 4. Further, the client learning rate η_c, and the server learning rate η_s are chosen such that $\eta_c \leq \frac{1}{10L\tau}$, $\eta_s\eta_c \leq \min\left\{\frac{\sqrt{M}(1-p)}{8L\tau}, \frac{M}{16\tau L}, \frac{1}{4L\tau}\right\}$. Then, the iterates $\{\mathbf{w}^{(t)}\}_t$ generated by* `ClusterFedVARP` *satisfy*

$$\min_{t\in[T]}\mathbb{E}\left\|\nabla f(\mathbf{w}^{(t)})\right\|^2 \leq \underbrace{\mathcal{O}\left(\frac{f(\mathbf{w}^{(0)})-f^*}{\eta_s\eta_c\tau T}\right)}_{\textit{Effect of initialization}}+\underbrace{\mathcal{O}\left(\frac{\eta_s\eta_c L\sigma^2}{M}+\eta_c^2L^2(\tau-1)\sigma^2\right)}_{\textit{Stochastic Gradient Error}}$$
$$+\underbrace{\mathcal{O}\left(\frac{\eta_s\eta_c\tau L\sigma_K^2}{M}\frac{(N-M)}{(N-1)}\right)}_{\textit{Error due clustering}}+\underbrace{\mathcal{O}\left(\eta_c^2L^2\tau(\tau-1)\sigma_g^2\right)}_{\textit{Client Drift Error}}$$

where $f^*=\arg\min_{\mathbf{x}} f(\mathbf{x})$.

Corollary 2. *Setting $\eta_c=\frac{1}{\sqrt{T}\tau L}$ and $\eta_s=\sqrt{\tau M}$,* `ClusterFedVARP` *converges to a stationary point of the global objective $f(\mathbf{w})$ at a rate given by,*

$$\min_{t\in\{0,\ldots,T-1\}}\mathbb{E}\left\|\nabla f(\mathbf{w}^{(t)})\right\|^2 \leq \underbrace{\mathcal{O}\left(\frac{1}{\sqrt{M\tau T}}\right)}_{\textit{stochastic gradient error}}+\underbrace{\mathcal{O}\left(\frac{(N-M)}{(N-1)}\sqrt{\frac{\tau}{MT}}\right)}_{\textit{partial participation error}}+\underbrace{\mathcal{O}\left(\frac{1}{T}\right)}_{\textit{client drift error}}$$

Proof. The proof is analogous to the proof of Theorem 2 in Section 3, with $1-\frac{M}{N}$ replaced by $p=\frac{\binom{N-r}{M}}{\binom{N}{M}}$. We use the same Lyapunov function defined in (11), with $\frac{\tilde{\eta}_s^2L}{2}\leq\delta\leq\frac{\tilde{\eta}_s}{2}$. Using Lemma 8, we get

$$R^{(t+1)} \leq \mathbb{E}\left[f(\mathbf{w}^{(t)})-\frac{\tilde{\eta}_s}{2}\left\|\nabla f(\mathbf{w}^{(t)})\right\|^2+\frac{\tilde{\eta}_s}{2}A^{(t)}+\delta\mathbb{E}_{\mathcal{S}^{(t)},\xi^{(t)}}\left\|\mathbf{v}^{(t)}-\bar{\mathbf{h}}^{(t)}\right\|^2\right.$$

$$+\alpha \frac{1}{N}\sum_{j=1}^{N} \mathbb{E}_{\mathcal{S}^{(t)},\xi^{(t)}} \left\|\nabla f_j(\mathbf{w}^{(t)}) - \mathbf{y}_{c_j}^{(t+1)}\right\|^2 \Bigg], \tag{21}$$

where $A^{(t)} \triangleq \frac{1}{N}\sum_{j=1}^{N} \mathbb{E}_{\xi^{(t)}}\left[\left\|\nabla f_j(\mathbf{w}^{(t)}) - \mathbf{h}_j^{(t)}\right\|^2\right]$. Substituting the bounds from Lemma 11 and Lemma 12 in (21) we get,

$$\begin{aligned} R^{(t+1)} \leq\ & \mathbb{E}\left[f(\mathbf{w}^{(t)}) - \frac{\tilde{\eta}_s}{2}\left\|\nabla f(\mathbf{w}^{(t)})\right\|^2 + \left(\frac{\tilde{\eta}_s}{2} + \frac{4\delta}{M}\frac{(N-M)}{(N-1)} + 4\alpha(1-p)\right)A^{(t)} + \frac{\delta\sigma^2}{M\tau} + 4\alpha(1-p)\left(\frac{\sigma^2}{\tau} + \sigma_K^2\right)\right] \\ & + \left(\frac{4\delta\tilde{\eta}_s^2 L^2}{M}\frac{(N-M)}{(N-1)} + \alpha p\left(1+\frac{1}{\beta}\right)\tilde{\eta}_s^2 L^2\right)\mathbb{E}\left\|\mathbf{v}^{(t-1)}\right\|^2 \\ & + \left(\frac{2\delta}{M}\frac{(N-M)}{(N-1)} + \alpha p(1+\beta)\right)\frac{1}{N}\sum_{j=1}^{N}\mathbb{E}\left\|\nabla f_j(\mathbf{w}^{(t-1)}) - \mathbf{y}_{c_j}^{(t)}\right\|^2. \end{aligned} \tag{22}$$

Choice of α, δ. Our goal is now to find a suitable δ and α such that,

$$\left(\frac{4\delta\tilde{\eta}_s^2 L^2}{M}\frac{(N-M)}{(N-1)} + \alpha p\left(1+\frac{1}{\beta}\right)\tilde{\eta}_s^2 L^2\right) \leq \delta - \frac{\tilde{\eta}_s^2 L}{2},$$
$$\left(\frac{2\delta}{M}\frac{(N-M)}{(N-1)} + \alpha p(1+\beta)\right) \leq \alpha$$

We define $A = p\left(1+\frac{1}{\beta}\right)$ and $B = p(1+\beta)$. The resulting condition on α is

$$\alpha \geq \frac{2\delta}{M(1-B)}\frac{(N-M)}{(N-1)}, \qquad \beta \leq \frac{1}{p} - 1.$$

We set $\alpha = \frac{2\delta}{M(1-B)}\frac{(N-M)}{(N-1)}$, and our condition on δ then reduces to,

$$\delta \geq \frac{\tilde{\eta}_s^2 L/2}{1 - \frac{4\tilde{\eta}_s^2 L^2}{M}\frac{(N-M)}{(N-1)} - \frac{2A\tilde{\eta}_s^2 L^2}{M(1-B)}\frac{(N-M)}{(N-1)}}$$

We want $\tilde{\eta}_s$ such that,

$$\frac{4\tilde{\eta}_s^2 L^2}{M}\frac{(N-M)}{(N-1)} + \frac{2A\tilde{\eta}_s^2 L^2}{M(1-B)}\frac{(N-M)}{(N-1)} \leq \frac{1}{2}$$

A sufficient condition for this is

$$\frac{4\tilde{\eta}_s^2 L^2}{M} \leq \frac{1}{4}, \qquad \frac{2A\tilde{\eta}_s^2 L^2}{M(1-B)} \leq \frac{1}{4}. \tag{23}$$

For $\beta = \frac{1}{2p} - \frac{1}{2}$, $B = \frac{p}{2} + \frac{1}{2}$ and $A \leq \frac{2}{1-p}$. Hence, a sufficient condition for (23) to be satisfied is

$$\tilde{\eta}_s \leq \frac{\sqrt{M}(1-p)}{8L} \quad \Rightarrow \quad \eta_s\eta_c \leq \frac{\sqrt{M}(1-p)}{8L\tau}$$

With (23) we have $\delta \geq \tilde{\eta}_s^2 L$. We set $\delta = 2\tilde{\eta}_s^2 L$ which gives us $\alpha = \frac{8\tilde{\eta}_s^2 L}{M(1-p)}\frac{(N-M)}{(N-1)}$. Since $\delta \leq \frac{\tilde{\eta}_s}{2}$, we also need $\tilde{\eta}_s \leq \frac{1}{4L}$. With this choice of α, δ, from (22) we get

$$\begin{aligned} R^{(t+1)} &\leq R^{(t)} - \frac{\tilde{\eta}_s}{2}\mathbb{E}\left\|\nabla f(\mathbf{w}^{(t)})\right\|^2 + \left(\frac{\tilde{\eta}_s}{2} + \frac{4\delta}{M} + 4\alpha(1-p)\right)A^{(t)} + \frac{\delta\sigma^2}{M\tau} + 4\alpha(1-p)\left(\frac{\sigma^2}{\tau} + \sigma_K^2\right) \\ &\leq R^{(t)} - \frac{\tilde{\eta}_s}{2}\left\|\nabla f(\mathbf{w}^{(t)})\right\|^2 + 3\tilde{\eta}_s A^{(t)} + \frac{40\tilde{\eta}_s^2 L}{M}\frac{\sigma^2}{\tau} + \frac{32\tilde{\eta}_s^2 L}{M}\frac{(N-M)}{(N-1)}\sigma_K^2, \end{aligned} \tag{24}$$

where we use the condition that $\tilde{\eta}_s \leq \frac{M}{16L}$. Further, we can bound $A^{(t)} = \frac{1}{N}\sum_{j=1}^{N} \mathbb{E}_{\xi^{(t)}}\left[\left\|\nabla f_j(\mathbf{w}^{(t)}) - \mathbf{h}_j^{(t)}\right\|^2\right]$ using Lemma 6, which gives us

$$\begin{aligned}R^{(t+1)} &\leq R^{(t)} - \frac{\tilde{\eta}_s}{2}\left(1 - 48\eta_c^2 L^2 \tau(\tau-1)\right)\left\|\nabla f(\mathbf{w}^{(t)})\right\|^2 + 6\tilde{\eta}_s\eta_c^2 L^2(\tau-1)\left[\sigma^2 + 4\tau\sigma_g^2\right] \\ &\quad + \frac{40\tilde{\eta}_s^2 L}{M}\frac{\sigma^2}{\tau} + \frac{32\tilde{\eta}_s^2 L}{M}\frac{(N-M)}{(N-1)}\sigma_K^2.\end{aligned}$$

Using the condition on η_c that $\eta_c \leq \frac{1}{10L\tau}$, and unrolling the recursion we get,

$$R^{(t)} \leq R^{(1)} + \sum_{t=1}^{(t-1)}\left(-\frac{\tilde{\eta}_s}{4}\left\|\nabla f(\mathbf{w}^{(t)})\right\|^2 + 6\tilde{\eta}_s\eta_c^2 L^2(\tau-1)\left[\sigma^2 + 4\tau\sigma_g^2\right] + \frac{40\tilde{\eta}_s^2 L}{M}\frac{\sigma^2}{\tau} + \frac{32\tilde{\eta}_s^2 L}{M}\frac{(N-M)}{(N-1)}\sigma_K^2\right). \tag{25}$$

Next, we bound $R^{(1)}$. Using (21) and (31) we can bound $R^{(1)}$ as follows,

$$\begin{aligned}R^{(1)} \leq & f(\mathbf{w}^{(0)}) - \frac{\tilde{\eta}_s}{2}\left\|\nabla f(\mathbf{w}^{(0)})\right\|^2 + \left(\frac{\tilde{\eta}_s}{2} + 4\alpha(1-p)\right)A^{(0)} + 4\alpha(1-p)\left(\frac{\sigma^2}{\tau} + \sigma_K^2\right) \\ & + \delta\mathbb{E}_{\xi^{(0)},\mathcal{S}^{(0)}}\left[\left\|\mathbf{v}^{(0)} - \bar{\mathbf{h}}^{(0)}\right\|^2\right] + \alpha p\frac{1}{N}\sum_{i=1}^{N}\left\|\nabla f_i(\mathbf{w}^{(0)})\right\|^2 \\ \leq & f(\mathbf{w}^{(0)}) - \frac{\tilde{\eta}_s}{4}\left\|\nabla f(\mathbf{w}^{(0)})\right\|^2 + 4\tilde{\eta}_s\eta_c^2 L^2(\tau-1)\left[\sigma^2 + 4\tau\sigma_g^2\right] + \frac{32\tilde{\eta}_s^2 L}{M}\frac{(N-M)}{(N-1)}\left(\frac{\sigma^2}{\tau} + \sigma_K^2\right) \\ & + \delta\mathbb{E}_{\xi^{(0)},\mathcal{S}^{(0)}}\left[\left\|\mathbf{v}^{(0)} - \bar{\mathbf{h}}^{(0)}\right\|^2\right] + \alpha p\frac{1}{N}\sum_{i=1}^{N}\left\|\nabla f_i(\mathbf{w}^{(0)})\right\|^2.\end{aligned} \tag{26}$$

Substituting the bound on $R^{(1)}$ from (26) in (25), and using $t = T$ we get

$$\begin{aligned}R^{(T)} \leq & f(\mathbf{w}^{(0)}) - \sum_{t=0}^{(T-1)}\left(\frac{\tilde{\eta}_s}{4}\left\|\nabla f(\mathbf{w}^{(t)})\right\|^2 + 6\tilde{\eta}_s\eta_c^2 L^2(\tau-1)\left[\sigma^2 + 4\tau\sigma_g^2\right] + \frac{40\tilde{\eta}_s^2 L}{M}\frac{\sigma^2}{\tau} + \frac{32\tilde{\eta}_s^2 L}{M}\frac{(N-M)}{(N-1)}\sigma_K^2\right) \\ & + 2\tilde{\eta}_s^2 L\mathbb{E}_{\xi^{(0)},\mathcal{S}^{(0)}}\left[\left\|\mathbf{v}^{(0)} - \bar{\mathbf{h}}^{(0)}\right\|^2\right] + \frac{8p\tilde{\eta}_s^2 L}{M(1-p)}\frac{(N-M)}{(N-1)}\frac{1}{N}\sum_{i=1}^{N}\left\|\nabla f(\mathbf{w}^{(0)})\right\|^2\end{aligned}$$

Rearranging the terms, we get

$$\begin{aligned}&\min_{t\in[0,T-1]}\mathbb{E}\left\|\nabla f(\mathbf{w}^{(t)})\right\|^2 \leq \frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}\left\|\nabla f(\mathbf{w}^{(t)})\right\|^2 \\ &\leq \frac{4(f(\mathbf{w}^{(0)}) - f^*)}{\tilde{\eta}_s T} + \frac{160\tilde{\eta}_s L}{M}\frac{\sigma^2}{\tau} + \frac{128\tilde{\eta}_s L}{M}\frac{(N-M)}{(N-1)}\sigma_K^2 + 24\eta_c^2 L^2(\tau-1)\sigma^2 + 96\eta_c^2 L^2\tau(\tau-1)\sigma_g^2 \\ &\quad + \frac{1}{T}\left[8\tilde{\eta}_s L\mathbb{E}_{\xi^{(0)},\mathcal{S}^{(0)}}\left[\left\|\mathbf{v}^{(0)} - \bar{\mathbf{h}}^{(0)}\right\|^2\right] + \frac{32p\tilde{\eta}_s L}{M(1-p)}\frac{(N-M)}{(N-1)}\frac{1}{N}\sum_{i=1}^{N}\left\|\nabla f(\mathbf{w}^{(0)})\right\|^2\right] \\ &= \mathcal{O}\left(\frac{f(\mathbf{w}^{(0)}) - f^*}{\tilde{\eta}_s T}\right) + \mathcal{O}\left(\frac{\tilde{\eta}_s L\sigma^2}{M\tau}\right) + \mathcal{O}\left(\frac{\tilde{\eta}_s L\sigma_K^2}{M}\frac{(N-M)}{(N-1)}\right) + \mathcal{O}(\eta_c^2 L^2(\tau-1)\sigma^2 + \eta_c^2 L^2\tau(\tau-1)\sigma_g^2),\end{aligned}$$

which concludes the proof. □

4.3 PROOFS OF THE INTERMEDIATE LEMMAS

Proof of Lemma 11. The proof is analogous to proof of Lemma 11. The only difference being that in (19), we do not bound the term $\frac{N-M}{N-1}$ with 1. □

Proof of Lemma 12.

$$\mathbb{E}_{\xi^{(t)},\mathcal{S}^{(t)}}\left[\frac{1}{N}\sum_{j=1}^{N}\left\|\nabla f_j(\mathbf{w}^{(t)})-\mathbf{y}_{c_j}^{(t+1)}\right\|^2\right]=\frac{1}{N}\sum_{j=1}^{N}\mathbb{E}_{\xi^{(t)},\mathcal{S}^{(t)}}\left[\left\|\nabla f_j(\mathbf{w}^{(t)})-\mathbf{y}_{c_j}^{(t+1)}\right\|^2\right]. \tag{27}$$

Let $\mathcal{C}_k^{(t)}=\{i:c_i=k \text{ and } i\in\mathcal{S}^{(t)}\}$, i.e., the set of sampled clients which belong to the k-th cluster. For a specific cluster $c_j\in[K]$

$$\begin{aligned}
&\mathbb{E}_{\mathcal{S}^{(t)}}\left[\left\|\nabla f_j(\mathbf{w}^{(t)})-\mathbf{y}_{c_j}^{(t+1)}\right\|^2\right]\\
&=\mathbb{E}_{\mathcal{C}_{c_j}^{(t)}}\left[\mathbb{I}\left(|\mathcal{C}_{c_j}^{(t)}|\neq 0\right)\left\|\nabla f_j(\mathbf{w}^{(t)})-\frac{\sum_{l\in\mathcal{C}_{c_j}^{(t)}}\Delta_l^{(t)}}{|\mathcal{C}_{c_j}^{(t)}|}\right\|^2+\mathbb{I}\left(|\mathcal{C}_{c_j}^{(t)}|=0\right)\left\|\nabla f_j(\mathbf{w}^{(t)})-\mathbf{y}_{c_j}^{(t)}\right\|^2\right]\\
&\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\text{(Cluster center update in }\texttt{ClusterFedVARP}\text{)}\\
&=\mathbb{E}_{\mathcal{C}_{c_j}^{(t)}}\left[\mathbb{I}\left(|\mathcal{C}_{c_j}^{(t)}|\neq 0\right)\left\|\nabla f_j(\mathbf{w}^{(t)})-\frac{1}{|\mathcal{C}_{c_j}^{(t)}|}\sum_{l\in\mathcal{C}_{c_j}^{(t)}}\left(\Delta_l^{(t)}-\nabla f_l(\mathbf{w}^{(t)})+\nabla f_l(\mathbf{w}^{(t)})\right)\right\|^2\right]\\
&\quad+\mathbb{E}_{\mathcal{C}_{c_j}^{(t)}}\left[\mathbb{I}\left(|\mathcal{C}_{c_j}^{(t)}|=0\right)\left\|\nabla f_j(\mathbf{w}^{(t)})-\mathbf{y}_{c_j}^{(t)}\right\|^2\right]\\
&\leq\mathbb{E}_{\mathcal{C}_{c_j}^{(t)}}\left[\mathbb{I}\left(|\mathcal{C}_{c_j}^{(t)}|\neq 0\right)\left(\frac{2}{|\mathcal{C}_{c_j}^{(t)}|}\sum_{l\in C_{c_j}^{(t)}}\left\|\nabla f_j(\mathbf{w}^{(t)})-\nabla f_l(\mathbf{w}^{(t)})\right\|^2+\frac{2}{|\mathcal{C}_{c_j}^{(t)}|}\sum_{l\in C_{c_j}^{(t)}}\left\|\nabla f_l(\mathbf{w}^{(t)})-\Delta_l^{(t)}\right\|^2\right)\right]\\
&\quad+\mathbb{E}_{\mathcal{C}_{c_j}^{(t)}}\left[\mathbb{I}\left(|\mathcal{C}_{c_j}^{(t)}|=0\right)\left\|\nabla f_j(\mathbf{w}^{(t)})-\mathbf{y}_{c_j}^{(t)}\right\|^2\right]\qquad\text{(Jensen's inequality; Young's inequality)}\\
&\leq\mathbb{E}_{\mathcal{C}_{c_j}^{(t)}}\left[\mathbb{I}\left(|\mathcal{C}_{c_j}^{(t)}|\neq 0\right)\left(4\sigma_K^2+\frac{2}{|\mathcal{C}_{c_j}^{(t)}|}\sum_{l\in C_{c_j}^{(t)}}\left\|\nabla f_l(\mathbf{w}^{(t)})-\Delta_l^{(t)}\right\|^2\right)\right]\\
&\quad+\mathbb{E}_{\mathcal{C}_{c_j}^{(t)}}\left[\mathbb{I}\left(|\mathcal{C}_{c_j}^{(t)}|=0\right)\left\|\nabla f_j(\mathbf{w}^{(t)})-\mathbf{y}_{c_j}^{(t)}\right\|^2\right]\qquad\qquad\qquad\qquad\text{(Assumption [])}
\end{aligned}$$

Substituting in (27) we get

$$\begin{aligned}
&\sum_{i=1}^{N}\mathbb{E}_{\mathcal{S}^{(t)}}\left[\left\|\nabla f_j(\mathbf{w}^{(t)})-\mathbf{y}_{c_j}^{(t+1)}\right\|^2\right]\\
&\leq\sum_{k=1}^{K}\left(4r\mathbb{E}_{\mathcal{C}_k^{(t)}}\left[\mathbb{I}(|\mathcal{C}_k^{(t)}|\neq 0)\right]\sigma_K^2+2\mathbb{E}_{\mathcal{C}_k^{(t)}}\left[\mathbb{I}(|\mathcal{C}_k^{(t)}|\neq 0)\frac{r}{|\mathcal{C}_k^{(t)}|}\sum_{l\in\mathcal{C}_k^{(t)}}\left\|\nabla f_l(\mathbf{w}^{(t)})-\Delta_l^{(t)}\right\|^2\right]\right)\\
&\quad+\sum_{j=1}^{N}\mathbb{E}_{\mathcal{C}_{c_j}^{(t)}}\left[\mathbb{I}(\mathcal{C}_{c_j}^{(t)}=0)\right]\left\|\nabla f_j(\mathbf{w}^{(t)})-\mathbf{y}_{c_j}^{(t)}\right\|^2\qquad\qquad\qquad(|\mathcal{C}_k^{(t)}|\leq r)\\
&=\sum_{k=1}^{K}\left(4r(1-p)\sigma_K^2+2\mathbb{E}_{\mathcal{C}_k^{(t)}}\left[\mathbb{I}(|\mathcal{C}_k^{(t)}|\neq 0)\frac{r}{|\mathcal{C}_k^{(t)}|}\sum_{l\in\mathcal{C}_k^{(t)}}\left\|\nabla f_l(\mathbf{w}^{(t)})-\Delta_l^{(t)}\right\|^2\right]\right)\\
&\quad+p\sum_{j=1}^{N}\left\|\nabla f_j(\mathbf{w}^{(t)})-\mathbf{y}_{c_j}^{(t)}\right\|^2,
\end{aligned}\tag{28}$$

where $p=\mathbb{E}_{\mathcal{C}_{c_j}^{(t)}}\left[\mathbb{I}(\mathcal{C}_{c_j}^{(t)}=0)\right]=\frac{\binom{N-r}{M}}{\binom{N}{r}}$ is the probability that no client from a particular cluster is sampled in $\mathcal{S}^{(t)}$ (same for all j since we assumed equal number of devices in each cluster). Note that,

$$\mathbb{E}_{\mathcal{C}_k^{(t)}}\left[\mathbb{I}(|\mathcal{C}_k^{(t)}| \neq 0)\frac{r}{|\mathcal{C}_k^{(t)}|}\sum_{l\in\mathcal{C}_k^{(t)}}\left\|\nabla f_l(\mathbf{w}^{(t)})-\Delta_l^{(t)}\right\|^2\right] = \mathbb{E}_{\mathcal{C}_k^{(t)}}\left[\frac{r}{|\mathcal{C}_k^{(t)}|}\sum_{l\in\mathcal{C}_k}\mathbb{I}(|\mathcal{C}_k^{(t)}| \neq 0, l\in\mathcal{C}_k^{(t)})\left\|\nabla f_l(\mathbf{w}^{(t)})-\Delta_l^{(t)}\right\|^2\right] \tag{29}$$

$$\begin{aligned}
&= \sum_{l\in\mathcal{C}_k}\left\|\nabla f_l(\mathbf{w}^{(t)})-\Delta_l^{(t)}\right\|^2\sum_{z=1}^{r}\frac{r}{z}\mathbb{P}(|\mathcal{C}_k^{(t)}|=z, l\in\mathcal{C}_k^{(t)})\\
&= \sum_{l\in\mathcal{C}_k}\left\|\nabla f_l(\mathbf{w}^{(t)})-\Delta_l^{(t)}\right\|^2\sum_{z=1}^{r}\frac{r}{z}\frac{\binom{r-1}{z-1}\binom{N-r}{M-z}}{\binom{N}{M}}\\
&= \sum_{l\in\mathcal{C}_k}\left\|\nabla f_l(\mathbf{w}^{(t)})-\Delta_l^{(t)}\right\|^2\sum_{z=1}^{r}\frac{\binom{r}{z}\binom{N-r}{M-z}}{\binom{N}{M}}\\
&= (1-p)\sum_{l\in\mathcal{C}_k}\left\|\nabla f_l(\mathbf{w}^{(t)})-\Delta_l^{(t)}\right\|^2. \qquad (30)
\end{aligned}$$

Substituting the bounds from (28), (30) in (27), we get

$$\begin{aligned}
&\frac{1}{N}\sum_{j=1}^{N}\mathbb{E}_{\xi^{(t)},\mathcal{S}^{(t)}}\left[\left\|\nabla f_j(\mathbf{w}^{(t)})-\mathbf{y}_{c_j}^{(t+1)}\right\|^2\right]\\
&\leq 4(1-p)\sigma_K^2+2(1-p)\frac{1}{N}\sum_{j=1}^{N}\mathbb{E}_{\xi^{(t)}}\left[\left\|\nabla f_j(\mathbf{w}^{(t)})-\Delta_j^{(t)}\right\|^2\right]+p\frac{1}{N}\sum_{j=1}^{N}\left\|\nabla f_j(\mathbf{w}^{(t)})-\mathbf{y}_{c_j}^{(t)}\right\|^2 \qquad (31)\\
&\leq (1-p)\left[4\sigma_K^2+\frac{4\sigma^2}{\tau}+\frac{4}{N}\sum_{j=1}^{N}\mathbb{E}_t\left[\left\|\nabla f_j(\mathbf{w}^{(t)})-\mathbf{h}_j^{(t)}\right\|^2\right]\right]\\
&\quad +p\left(1+\frac{1}{\beta}\right)\tilde{\eta}_s^2L^2\left\|\mathbf{v}^{(t-1)}\right\|^2+p(1+\beta)\frac{1}{N}\sum_{j=1}^{N}\left\|\nabla f_j(\mathbf{w}^{(t-1)})-\mathbf{y}_{c_j}^{(t)}\right\|^2,
\end{aligned}$$

for any positive constant β. □