Supplementary material

A Specifications of Experiments

A.1 Model architecture and dataset

MLP. The 2-layer multilayer perceptron (MLP) has 784 input units and 200 hidden units so that the hidden layer parameters (157,000 parameters) are optimized for solving the upper-level problem and the output layer parameters (2,010 parameters) are optimized for solving the lower-level problem.

CNN. We use the 7-layer CNN [33] model to train CIFAR-10. We optimize the last fully connected layer's parameters for solving the lower-level problem and optimize the rest layers' parameters for solving the upper-level problem.

Dataset. For the hyper-representation experiment, we use full MNIST and CIFAR-10 datasets. In the experiment with heterogeneous local computation, we treat the first 2000 images in MNIST's default training dataset as the training data and the first 1000 images in MNIST's default test dataset as test data.

A.2 Hyperparameter settings

For all comparison methods, we optimize their hyperparameters via grid search guided by the default values in their source codes, to ensure the best performance given the algorithms are convergent.

Comparison to existing methods. First of all, for all methods, 10 clients from 100 clients are chosen randomly and participate in each communication round. For the baseline methods FedNest and LFedNest, we use their published codes in https://github.com/ucr-optml/FedNest. For FBO-AggITD, we use the source codes sent from the authors. For our method, SimFBO, we take the number of local updates,τ_i, for each client i to be 1, $a_i^{(t,k)}$ to be 1, and $\widetilde{p_i}$ to be 0.1. In MNIST-MLP experiment, the stepsizes $[\eta_y, \eta_v, \eta_x]$ and $[\gamma_y, \gamma_v, \gamma_x]$ of our method for updating $[y_i^{t,k}(y^t), v_i^{t,k}(v^t), x_i^{t,k}(x^t)]$ are both [0.2, 0.1, 0.05], respectively. Second, for the CIFAR-10-CNN experiment under the i.i.d. setup, we only draw the result of FBO-AggITD and SimFBO because other algorithms cannot converge under various hyperparameter configurations. We take the best inner stepsize as 0.003 and the best outer stepsize as 0.005 for FBO-AggITD, and the stepsizes $[\eta_y, \eta_v, \eta_x]$ and $[\gamma_y, \gamma_v, \gamma_x]$ of our method for updating $[y_i^{t,k}(y^t), v_i^{t,k}(v^t), x_i^{t,k}(x^t)]$ are both [0.1, 0.05, 0.03], respectively.

Performance under heterogeneous local computation. In this experiment, we compare the results among LFedNest, FedNest, FBO-AggITD, SimFBO, and ShroFBO. For all methods, the numbers of local updates of different clients are randomly chosen in the range from 1 to 10 to simulate the system-level heterogeneity and there are a total of 10 clients participating during the entire procedure. In specific, inner stepsizes for LFedNest, FedNest, FBO-AggITD are [0.002, 0.01, 0.005], respectively while outer stepsizes for LFedNest, FedNest, FBO-AggITD are [0.005, 0.03, 0.04], respectively. Other hyperparameters of the above-mentioned three methods are chosen the same as in the above experiment. Then for ShroFBO, we keep choosing $a_i^{(t,k)}$ for each client i as 1 but the values of $\|a_j^{(t)}\|_1$ for $j \in [1, n]$ would be different since the number of local updates τ_i is randomly chosen between 1 and 10. The value of p_j is chosen as 0.1. Similarly, the stepsizes $[\eta_y, \eta_v, \eta_x]$ and $[\gamma_y, \gamma_v, \gamma_x]$ of our method for updating $[y_i^{t,k}(y^t), v_i^{t,k}(v^t), x_i^{t,k}(x^t)]$ are both [0.03, 0.02, 0.01], respectively. Lastly, SimFBO has the same settings mentioned above.

B Notations

For notational convenience, we define

$$\widetilde{F}(x, y) := \sum_{i=1}^{n} w_i f_i(x, y),\ \widetilde{G}(x, y) := \sum_{i=1}^{n} w_i g_i(x, y),\ \widetilde{R}(w, y, v) := \sum_{i=1}^{n} w_i R_i(x, y, v).$$

For SimBFO, the problem we solve here is

$$\begin{aligned}\min_{x\in\mathbb{R}^p} \widetilde{\Phi}(x) &= \widetilde{F}(x,\widetilde{y}^*(x)) := \sum_{i=1}^n w_i f_i(x,\widetilde{y}^*(x)) = \sum_{i=1}^n w_i \mathbb{E}_\xi\big[f_i(x,\widetilde{y}^*(x);\xi)\big] \\ \text{s.t. } \widetilde{y}^*(x) &= \arg\min_{y\in\mathbb{R}^q} \widetilde{G}(x,y) := \sum_{i=1}^n w_i g_i(x,y) = \sum_{i=1}^n w_i \mathbb{E}_\zeta\big[g_i(x,y;\zeta)\big].\end{aligned}$$

Similarly, we define

$$\widetilde{\Phi}(x) := \widetilde{F}(x,\widetilde{y}^*),\ \nabla\widetilde{\Phi}(x) := \sum_{i=1}^n w_i \bar{\nabla} f(x,\widetilde{y}^*,\widetilde{v}^*) \tag{11}$$

where $\widetilde{y}^* = \arg\min_y \widetilde{G}(x,y)$ and $\widetilde{v}^* = \arg\min_v \widetilde{R}(x,\widetilde{y}^*,v)$.

We can see that $\widetilde{y}^*$ and $\widetilde{v}^*$ are unique due to the strong convexity of $g_i(x,y)$ and $R_i(x,y,v)$. Client updates are aggregated to compute $\{h_{x,i}^{(t)}, h_{y,i}^{(t)}, h_{v,i}^{(t)}\}$ as

$$\begin{aligned}h_{y,i}^{(t)} &= \frac{1}{\|a_i^{(t)}\|_1}\sum_{k=1}^{\tau_i^{(t)}} a_i^{(t,k)} \nabla_y g_i(x_i^{(t,k)}, y_i^{(t,k)};\zeta_i^{(t,k)}),\\ h_{v,i}^{(t)} &= \frac{1}{\|a_i^{(t)}\|_1}\sum_{k=1}^{\tau_i^{(t)}} a_i^{(t,k)} \nabla_v R_i(x_i^{(t,k)}, y_i^{(t,k)}, v_i^{(t,k)};\psi_i^{(t,k)}),\\ h_{x,i}^{(t)} &= \frac{1}{\|a_i^{(t)}\|_1}\sum_{k=1}^{\tau_i^{(t)}} a_i^{(t,k)} \bar{\nabla} f_i(x_i^{(t,k)}, y_i^{(t,k)}, v_i^{(t,k)};\xi_i^{(t,k)}),\end{aligned}$$

and their expectations are

$$\begin{aligned}\widetilde{h}_{y,i}^{(t)} = \mathbb{E}[h_{y,i}^{(t)}] &= \frac{1}{\|a_i^{(t)}\|_1}\sum_{k=1}^{\tau_i^{(t)}} a_i^{(t,k)} \nabla_y g_i(x_i^{(t,k)}, y_i^{(t,k)}),\\ \widetilde{h}_{v,i}^{(t)} = \mathbb{E}[h_{v,i}^{(t)}] &= \frac{1}{\|a_i^{(t)}\|_1}\sum_{k=1}^{\tau_i^{(t)}} a_i^{(t,k)} \nabla_v R_i(x_i^{(t,k)}, y_i^{(t,k)}, v_i^{(t,k)}),\\ \widetilde{h}_{x,i}^{(t)} = \mathbb{E}[h_{x,i}^{(t)}] &= \frac{1}{\|a_i^{(t)}\|_1}\sum_{k=1}^{\tau_i^{(t)}} a_i^{(t,k)} \bar{\nabla} f_i(x_i^{(t,k)}, y_i^{(t,k)}, v_i^{(t,k)})\end{aligned}$$

for all $t \in \{0,1,...,T-1\}$, $k \in \{0,1,...,\tau_i - 1\}$ and $i \in \{1,2,...,n\}$.

To ensure the robustness of server updates, we set $\alpha_{\min} \le a_i^{(t,k)} \le \alpha_{\max}$ for all $t = 0,1,...,T$, $i = 1,2,...n$ and $k = 0,1,...,\tau_i - 1$; we also set $\frac{\beta_{\min}}{n} \le w_i \le \frac{\beta_{\max}}{n}$ and $\frac{\beta'_{\min}}{n} \le p_i \le \frac{\beta'_{\max}}{n}$ for all $i = 1,2,...n$.

At global iteration t, the server samples $|C^{(t)}|$ clients without replacement (**WOR**) uniformly at random. On the server side, the aggregated client i update is weighed by $\widetilde{w}_i = \frac{n}{|C^{(t)}|} w_i$. The aggregates $\{h_y^{(t)}, h_v^{(t)}, h_x^{(t)}\}$ computed at the server are of the form

$$h_y^{(t)} = \sum_{i\in C^{(t)}} \widetilde{w}_i h_{y,i}^{(t)},\quad h_v^{(t)} = \sum_{i\in C^{(t)}} \widetilde{w}_i h_{v,i}^{(t)},\quad h_x^{(t)} = \sum_{i\in C^{(t)}} \widetilde{w}_i h_{x,i}^{(t)},$$

and we also have partial clients expectation as

$$\mathbb{E}_{C_{(t)}}\big[h_y^{(t)}\big] = \mathbb{E}_{C_{(t)}}\Big[\sum_{i=1}^n \mathbb{I}(i\in C^{(t)})\widetilde{w}_i h_{y,i}^{(t)}\Big] = \sum_{i=1}^n w_i h_{y,i}^{(t)},$$

$$\mathbb{E}_{C_{(t)}}\big[h_v^{(t)}\big] = \mathbb{E}_{C_{(t)}}\Bigg[\sum_{i=1}^{n} \mathbb{I}(i \in C^{(t)})\widetilde{w}_i h_{v,i}^{(t)}\Bigg] = \sum_{i=1}^{n} w_i h_{v,i}^{(t)},$$
$$\mathbb{E}_{C_{(t)}}\big[h_x^{(t)}\big] = \mathbb{E}_{C_{(t)}}\Bigg[\sum_{i=1}^{n} \mathbb{I}(i \in C^{(t)})\widetilde{w}_i h_{x,i}^{(t)}\Bigg] = \sum_{i=1}^{n} w_i h_{x,i}^{(t)}.$$

Generally, the expectations we use contain both the expectations of samples and the expectations of clients. And in analysis, we simply define $|C^{(t)}| = P$ for all t. And server updates $y^{(t+1)}$, $v^{(t+1)}$, $x^{(t+1)}$ as

$$\begin{aligned} y^{(t+1)} &= y^{(t)} - \rho^{(t)}\gamma_y h_y^{(t)}, \\ v^{(t+1)} &= \mathcal{P}_r\big(v^{(t)} - \rho^{(t)}\gamma_v h_v^{(t)}\big), \\ x^{(t+1)} &= x^{(t)} - \rho^{(t)}\gamma_x h_x^{(t)}, \end{aligned}$$

where the auxiliary projection function is defined as $\mathcal{P}_r(v) := \min\{1, \frac{r}{\|v\|}\}v$ and $r = \frac{L_f}{\mu_g}$ is the server side auxiliary projection radius. To simplify the problem, we set $a_i^{(t,k)}$ such that $\rho^{(t)} \in [\frac{1}{2}\bar{\rho}, \frac{3}{2}\bar{\rho}]$ and $c_a'\bar{\tau}\alpha_{\min} \le \|a_i^{(t)}\|_1 \le c_a\bar{\tau}\alpha_{\max}$ for some positive constant c_a, c_a' and $i \in C^{(t)}$, where $\bar{\rho} := \frac{1}{T}\sum_{t=0}^{T-1}\rho^{(t)}$ and $\bar{\tau} := \sum_{i=1}^{n}\tau_i$.

C Proofs of Preliminary Lemmas

Lemma 1 (Boundedness of v^*)**.** *Under Assumptions 1 and 2, we have for v^* in eq.* (2)*, $\|v^*\|^2 \le \frac{L_f^2}{\mu_g^2}$.*

Proof. Remind that we define $v^* = \arg\min_v R(x, y^*, v)$, then we have

$$\|v^*\|^2 = \Big\|\big[\nabla_{yy}^2 G(x, y^*)\big]^{-1}\nabla_y F(x, y^*)\Big\|^2 \le \Big\|\big[\nabla_{yy}^2 G(x, y^*)\big]^{-1}\Big\|^2 \|\nabla_y F(x, y^*)\|^2 \overset{(a)}{\le} r^2,$$

where (a) follows Assumptions 1, 2 and defines $r := \frac{L_f}{\mu_g}$. Then, the proof is complete. ☐

Lemma 2 (Boundedness of local v)**.** *Under Assumptions 1 and 2, for each global iteration t, client i, and local iteration $k = 1, 2, ..., \tau_i$, we have*

$$r_i := \|v_i^{(t,k)}\| \le \Big(1 + \frac{\alpha_{\max}}{\alpha_{\min}}\Big)r =: r_{\max},$$

where $r = \frac{L_f}{\mu_g}$ is the server side auxiliary projection radius.

Proof. By the local update rule of $v_i^{(t,k)}$ from step 6 in Algorithm 1, we have

$$\begin{aligned} v_i^{(t,k)} &= v_i^{(t,k-1)} - \eta_v a_i^{(t,k-1)}\nabla_v R_i(x_i^{(t,k-1)}, y_i^{(t,k-1)}, v_i^{(t,k-1)}; \psi_i^{(t,k-1)}) \\ &= \big(I - \eta_v a_i^{(t,k-1)}\nabla_{yy}^2 g_i(x_i^{(t,k-1)}, y_i^{(t,k-1)}; \psi_i^{(t,k-1)})\big)v_i^{(t,k-1)} \\ &\quad + \eta_v a_i^{(t,k-1)}\nabla_y f_i(x_i^{(t,k-1)}, y_i^{(t,k-1)}; \psi_i^{(t,k-1)}). \end{aligned}$$

By taking l_2 norm, we have

$$\|v_i^{(t,k)}\| \le (1 - \eta_v a_i^{(t,k-1)}\mu_g)\|v_i^{(t,k-1)}\| + \eta_v a_i^{(t,k-1)}L_f. \tag{12}$$

Telescope eq. (12) over $j \in [0, ..., k-1]$, and we have

$$\begin{aligned} \|v_i^{(t,k)}\| &\le (1 - \eta_v a_i^{(t,k-1)}\mu_g)\|v_i^{(t,k-1)}\| + \eta_v a_i^{(t,k-1)}L_f \\ &\le (1 - \eta_v\alpha_{\min}\mu_g)^k\|v_i^{(t,0)}\| + \sum_{j=0}^{k-1}(1 - \eta\alpha_{\min}\mu_g)^j\eta_v\alpha_{\max}L_f \\ &\le (1 - \eta_v\alpha_{\min}\mu_g)^k\|v_i^{(t,0)}\| + \frac{\alpha_{\max}L_f}{\alpha_{\min}\mu_g} \\ &\le \Big(1 + \frac{\alpha_{\max}}{\alpha_{\min}}\Big)r. \end{aligned}$$

Then, the proof is complete. ☐

Since both $\alpha_{\max}$ and $\alpha_{\min}$ are preset constants, the bound of local v_i has the same order κ as server-side auxiliary projection radius r.

Lemma 3 (Basic properties of linear system function R)**.** *Under Assumptions 1 and 2, we have $R_i(x, y, v)$ is μ_g-strongly convex w.r.t v and $\nabla_v R_i(x, y, v)$ is L_R-Lipschitz continuous w.r.t (x, y), where we define $L_R^2 := 2(L_2^2 r_{\max}^2 + L_1^2)$.*

Proof. The strong convexity can be easily observed since $\nabla_{vv}^2 R(x, y, v) = \nabla_{yy}^2 g(x, y) \succeq \mu_g I$. By using the definition of $R_i(x, y, v)$ and assumptions 1 and 2, we have

$$
\begin{aligned}
&\|\nabla_v R_i(x_1, y_1, v) - \nabla_v R_i(x_2, y_2, v)\|^2 \\
&\quad \le 2\|\big(\nabla_{yy}^2 g(x_1, y_1) - \nabla_{yy}^2 g(x_2, y_2)\big)v\|^2 + 2\|\nabla_y f(x_1, y_1) - \nabla_y f(x_2, y_2)\|^2 \\
&\quad \le 2(L_2^2 r_{\max}^2 + L_1^2)(\|x_1 - x_2\|^2 + \|y_1 - y_2\|^2) = L_R^2(\|x_1 - x_2\|^2 + \|y_1 - y_2\|^2).
\end{aligned}
$$

Then, the proof is complete. ☐

Lemma 4 ([15] lemma 2.2, [4] lemma 2 and extensions)**.** *Under Assumptions 1, 2 and 3, we have, for all $x, x_1, x_2 \in \mathbb{R}^{d_x}$ and $y \in \mathbb{R}^{d_y}$,*

$$
\begin{aligned}
&\|\nabla f(x, y) - \nabla \Phi(x)\|^2 \le \widetilde{L}^2 \|y - y^*(x)\|^2, \ \|\nabla \Phi(x_1) - \nabla \Phi(x_2)\|^2 \le L_\Phi^2 \|x_1 - x_2\|^2, \\
&\quad \|y^*(x_1) - y^*(x_2)\|^2 \le L_y^2 \|x_1 - x_2\|^2, \ \|\nabla y^*(x_1) - \nabla y^*(x_2)\|^2 \le L_{yx}^2 \|x_1 - x_2\|^2, \\
&\quad \|v^*(x_1) - v^*(x_2)\|^2 \le L_v^2 \|x_1 - x_2\|^2, \ \|\nabla v^*(x_1) - \nabla v^*(x_2)\|^2 \le L_{vx}^2 \|x_1 - x_2\|^2
\end{aligned}
$$

where the constants are given by

$$
\begin{aligned}
\widetilde{L} = L_1 + \frac{L_1^2}{\mu_g} + L_f\left(\frac{L_2}{\mu_g} + \frac{L_1 L_2}{\mu_g^2}\right), \quad L_\Phi = \widetilde{L} + \frac{\widetilde{L} L_1}{\mu_g}, \\
L_y = \frac{L_1}{\mu_g}, \quad L_v = \left(\frac{2L_1^2}{\mu_g^2} + \frac{2L_f^2 L_2^2}{\mu_g^4}\right)^{\frac{1}{2}} \left(1 + L_y^2\right)^{\frac{1}{2}}, \\
L_{yx} = \frac{L_2 + L_2 L_y}{\mu_g} + \frac{L_1(L_2 + L_2 L_y)}{\mu_g^2}, \quad L_{vx} = \frac{2}{\mu_g}\Big(\big(L_2^2 + r^2 L_3^2 + L_2^2 L_v^2\big)\big(1 + L_y^2\big) + L_2^2 L_v^2\Big)^{\frac{1}{2}}.
\end{aligned}
\tag{13}
$$

Proof. The proof of the first 3 inequalities is provided in [15]. For the fifth inequality, we have

$$
\begin{aligned}
&\big\|v^*(x_1) - v^*(x_2)\big\|^2 \\
&= \Big\|\big[\nabla_{yy}^2 G\big(x_1, y^*(x_1)\big)\big]^{-1} \nabla_y F\big(x_1, y^*(x_1)\big) - \big[\nabla_{yy}^2 G\big(x_2, y^*(x_2)\big)\big]^{-1} \nabla_y F\big(x_2, y^*(x_2)\big)\Big\|^2 \\
&\le 2\Big\|\big[\nabla_{yy}^2 G\big(x_1, y^*(x_1)\big)\big]^{-1}\Big(\nabla_y F\big(x_1, y^*(x_1)\big) - \nabla_y F\big(x_2, y^*(x_2)\big)\Big)\Big\|^2 \\
&\quad + 2\Big\|\Big(\big[\nabla_{yy}^2 G\big(x_1, y^*(x_1)\big)\big]^{-1} - \big[\nabla_{yy}^2 G\big(x_2, y^*(x_2)\big)\big]^{-1}\Big)\nabla_y F\big(x_2, y^*(x_2)\big)\Big\|^2 \\
&\le \frac{2L_1^2}{\mu_g^2}\big(\big\|x_1 - x_2\big\|^2 + \big\|y^*(x_1) - y^*(x_2)\big\|^2\big) \\
&\quad + 2L_f^2\Big\|\big[\nabla_{yy}^2 G\big(x_1, y^*(x_1)\big)\big]^{-1}\big[\nabla_{yy}^2 G\big(x_1, y^*(x_1)\big) - \nabla_{yy}^2 G\big(x_2, y^*(x_2)\big)\big] \\
&\qquad\qquad \cdot \big[\nabla_{yy}^2 G\big(x_2, y^*(x_2)\big)\big]^{-1}\Big\|^2 \\
&\le \left(\frac{2L_1^2}{\mu_g^2} + \frac{2L_f^2 L_2^2}{\mu_g^4}\right)\big(\big\|x_1 - x_2\big\|^2 + \big\|y^*(x_1) - y^*(x_2)\big\|^2\big) \\
&\le \left(\frac{2L_1^2}{\mu_g^2} + \frac{2L_f^2 L_2^2}{\mu_g^4}\right)\big(1 + L_y^2\big)\big\|x_1 - x_2\big\|^2.
\end{aligned}
$$

And for the sixth inequality, we have

$$
\nabla_v R\big(x, y^*(x), v^*(x)\big) = \nabla_{yy}^2 G\big(x, y^*(x)\big)v^*(x) - \nabla_y F\big(x, y^*(x)\big) = 0,
$$

which implies that

$$
\begin{aligned}
\frac{\partial_x \nabla_v R\big(x, y^*(x), v^*(x)\big)}{\partial x} &= \big[v^*(x)\big]^T \Big[\nabla^3_{yyx} G\big(x, y^*(x)\big) + \nabla^3_{yyy} G\big(x, y^*(x)\big)\nabla y^*(x)\Big] \\
&\quad + \nabla^2_{yy} G\big(x, y^*(x)\big)\nabla v^*(x) - \nabla^2_{yx} F\big(x, y^*(x)\big) \\
&\quad - \nabla^2_{yy} F\big(x, y^*(x)\big)\nabla y^*(x) \\
&= \mathbf{0}_{d^y, d^x}
\end{aligned} \tag{14}
$$

and

$$
\frac{\partial_y \nabla_v R(x, y^*(x), v^*(x))}{\partial y} = \big[v^*(x)\big]^T \nabla^3_{yyy} G\big(x, y^*(x)\big) - \nabla^2_{yy} F\big(x, y^*(x)\big) = \mathbf{0}_{d^y, d^y}. \tag{15}
$$

By combining eq. (14) and eq. (15), we have

$$
\nabla^2_{yy} G\big(x, y^*(x)\big)\nabla_x v^*\big(x, y^*(x)\big) = \nabla^2_{yx} F\big(x, y^*(x)\big) - \big[v^*(x)\big]^T \nabla^3_{yyx} G\big(x, y^*(x)\big). \tag{16}
$$

Then we can get that

$$
\begin{aligned}
&\nabla^2_{yy} G\big(x_1, y^*(x_1)\big)\nabla_x v^*\big(x_1, y^*(x_1)\big) - \nabla^2_{yy} G\big(x_2, y^*(x_2)\big)\nabla_x v^*\big(x_2, y^*(x_2)\big) \\
&\quad = \nabla^2_{yy} G\big(x_1, y^*(x_1)\big)\nabla_x v^*\big(x_1, y^*(x_1)\big) - \nabla^2_{yy} G\big(x_2, y^*(x_2)\big)\nabla_x v^*\big(x_1, y^*(x_1)\big) \\
&\qquad + \nabla^2_{yy} G\big(x_2, y^*(x_2)\big)\nabla_x v^*\big(x_1, y^*(x_1)\big) - \nabla^2_{yy} G\big(x_2, y^*(x_2)\big)\nabla_x v^*\big(x_2, y^*(x_2)\big) \\
&\quad = \Big[\nabla_{yx} F\big(x_1, y^*(x_1)\big) - \nabla_{yx} F\big(x_2, y^*(x_2)\big)\Big] \\
&\qquad - \Big(\big[v^*(x_1)\big]^T \nabla^3_{yyx} G\big(x_1, y^*(x_1)\big) - \big[v^*(x_2)\big]^T \nabla^3_{yyx} G\big(x_2, y^*(x_2)\big)\Big)
\end{aligned}
$$

By taking the norm and using Assumption 2, we have

$$
\begin{aligned}
&\Big\|\nabla^2_{yy} G\big(x_2, y^*(x_2)\big)\Big[\nabla_x v^*(x_1) - \nabla_x v^*(x_2)\Big]\Big\|^2 \\
&\quad \leq 4\Big\|\nabla^2_{xy} F\big(x_1, y^*(x_1)\big) - \nabla^2_{xy} F\big(x_2, y^*(x_2)\big)\Big\|^2 \\
&\qquad + 4\Big\|\big[v^*(x_1) - v^*(x_2)\big]^T \nabla^3_{yyx} G\big(x_2, y^*(x_2)\big)\Big\|^2 \\
&\qquad + 4\Big\|\big[v^*(x_1)\big]^T \Big[\nabla^3_{yyx} G\big(x_1, y^*(x_1)\big) - \nabla^3_{yyx} G\big(x_2, y^*(x_2)\big)\Big]\Big\|^2 \\
&\qquad + 4\Big\|\Big[\nabla^2_{yy} G\big(x_1, y^*(x_1)\big) - \nabla^2_{yy} G\big(x_2, y^*(x_2)\big)\Big]\nabla_x v^*(x_1)\Big\|^2 \\
&\quad \leq 4\Big(L_2^2 + r^2 L_3^2 + L_2^2 L_v^2\Big)\Big(\big\|x_1 - x_2\big\|^2 + \big\|y^*(x_1) - y^*(x_2)\big\|^2\Big) + 4\Big(L_2^2 L_v^2\Big)\big\|x_1 - x_2\big\|^2 \\
&\quad \leq 4\Big(\big(L_2^2 + r^2 L_3^2 + L_2^2 L_v^2\big)\big(1 + L_y^2\big) + L_2^2 L_v^2\Big)\big\|x_1 - x_2\big\|^2
\end{aligned} \tag{17}
$$

By using the strong convexity of g_i in Assumption 1, we have

$$
\Big\|\nabla_x v^*\big(x_1, y^*(x_1)\big) - \nabla_x v^*\big(x_2, y^*(x_2)\big)\Big\|^2 \leq L_{vx}^2 \big\|x_1 - x_2\big\|^2, \tag{18}
$$

where $L_{vx} = \frac{2}{\mu_g}\Big(\big(L_2^2 + r^2 L_3^2 + L_2^2 L_v^2\big)\big(1 + L_y^2\big) + L_2^2 L_v^2\Big)^{\frac{1}{2}}$. Then the proof is complete. □

Lemma 5 (Global Heterogeneity Extension)**.** *For any set of non-negative weight* $\{w_i\}_{i=1}^n$ *such that* $\sum_{i=1}^n w_i = 1$*, under Assumption 2 and Lemma 2, we have the bounds of global heterogeneity of*

$\nabla g_i(x,y)$, $\nabla R_i(x,y,v)$ and $\nabla f_i(x,y,v)$ as

$$\sum_{i=1}^{n} w_i \|\nabla_y g_i(x,y)\|^2 \le \beta_{gh}^2 L_1^2 \big\|y - y^*(x)\big\|^2 + \sigma_{gh}^2,$$
$$\sum_{i=1}^{n} w_i \|\nabla_v R_i(x,y,v)\|^2 \le 2r_{\max}^2 L_1^2 + 2L_f^2,$$
$$\sum_{i=1}^{n} w_i \|\bar{\nabla} f_i(x,y,v)\|^2 \le 2r_{\max}^2 L_1^2 + 2L_f^2$$

for all $i \in \{1,...,n\}$.

Proof. Under Assumption 4, we have

$$\begin{aligned}\sum_{i=1}^{n} w_i \|\nabla_y g_i(x,y)\|^2 &\le \beta_{gh}^2 \Big\| \sum_{i=1}^{n} w_i \Big[\nabla_y g_i(x,y) - \nabla_y g_i\big(x, y^*(x)\big)\Big] \Big\|^2 + \sigma_{gh}^2 \\ &\le \beta_{gh}^2 L_1^2 \big\|y - y^*(x)\big\|^2 + \sigma_{gh}^2.\end{aligned}$$

For R_i and f_i, we can easily have

$$\begin{aligned}\|\nabla_v R_i(x,y,v)\|^2 &\le 2\|\nabla_{yy} g(x,y)v\|^2 + 2\|\nabla_y f(x,y)\|^2 \le 2r_{\max}^2 L_1^2 + 2L_f^2, \\ \|\bar{\nabla} f_i(x,y,v)\|^2 &\le 2\|\nabla_{xy} g(x,y)v\|^2 + 2\|\nabla_x f(x,y)\|^2 \le 2r_{\max}^2 L_1^2 + 2L_f^2.\end{aligned}$$

Then the proof is complete. □

Lemma 6. *Under Assumption 2, we have the following bounds*

$$\begin{aligned}\big\|\bar{\nabla} f_i(x^{(t)}, y^{(t)}, v^{(t)}) - \bar{\nabla} f_i(x_i^{(t,k)}, y_i^{(t,k)}, v_i^{(t,k)})\big\|^2 &\le \Delta_{f,i}^{(t,k)}, \\ \big\|\nabla g_i(x^{(t)}, y^{(t)}) - \nabla g_i(x_i^{(t,k)}, y_i^{(t,k)})\big\|^2 &\le \Delta_{g,i}^{(t,k)}, \\ \big\|\nabla_v R_i(x^{(t)}, y^{(t)}, v^{(t)}) - \nabla R_i(x_i^{(t,k)}, y_i^{(t,k)}, v_i^{(t,k)})\big\|^2 &\le \Delta_{R,i}^{(t,k)},\end{aligned}$$

where we define the combinations of client drift as

$$\begin{aligned}\Delta_{f_i}^{(t,k)} = \Delta_{R_i}^{(t,k)} &:= 3\big(L_1^2 + r^2 L_2^2\big)\Big[\big\|x_i^{(t,k)} - x^{(t)}\big\|^2 + \big\|y_i^{(t,k)} - y^{(t)}\big\|^2\Big] + 3L_1^2 \big\|v_i^{(t,k)} - v^{(t)}\big\|^2 \\ \Delta_{g_i}^{(t,k)} &:= L_1^2 \Big[\big\|x_i^{(t,k)} - x^{(t)}\big\|^2 + \big\|y_i^{(t,k)} - y^{(t)}\big\|^2\Big].\end{aligned}$$

Proof. According to the definition, we have

$$\begin{aligned}&\big\|\bar{\nabla} f_i(x^{(t)}, y^{(t)}, v^{(t)}) - \bar{\nabla} f_i(x_i^{(t,k)}, y_i^{(t,k)}, v_i^{(t,k)})\big\|^2 \\ &\quad \le \big\|\nabla_x f_i(x^{(t)}, y^{(t)}, v^{(t)}) - \nabla_x f_i(x_i^{(t,k)}, y_i^{(t,k)}, v_i^{(t,k)}) \\ &\qquad - \nabla_{xy}^2 g_i(x^{(t)}, y^{(t)}, v^{(t)}) v^{(t)} + \nabla_{xy}^2 g_i(x_i^{(t,k)}, y_i^{(t,k)}, v_i^{(t,k)}) v_i^{(t,k)}\big\|^2 \\ &\quad \le 3\big\|\nabla_x f_i(x^{(t)}, y^{(t)}, v^{(t)}) - \nabla_x f_i(x_i^{(t,k)}, y_i^{(t,k)}, v_i^{(t,k)})\big\|^2 \\ &\qquad + 3\big\|\big(\nabla_{xy}^2 g_i(x^{(t)}, y^{(t)}, v^{(t)}) - \nabla_{xy}^2 g_i(x_i^{(t,k)}, y_i^{(t,k)}, v_i^{(t,k)})\big) v^{(t)}\big\|^2 \\ &\qquad + 3\big\|\nabla_{xy}^2 g_i(x_i^{(t,k)}, y_i^{(t,k)}, v_i^{(t,k)})\big(v^{(t)} - v_i^{(t,k)}\big)\big\|^2 \\ &\quad \overset{(a)}{\le} 3\big(L_1^2 + r^2 L_2^2\big)\Big[\big\|x_i^{(t,k)} - x^{(t)}\big\|^2 + \big\|y_i^{(t,k)} - y^{(t)}\big\|^2\Big] + 3L_1^2 \big\|v_i^{(t,k)} - v^{(t)}\big\|^2,\end{aligned}$$

where (a) follows from Assumption2 and Lemma 1. Similarly, we can easily get the results of ∇g_i and $\nabla_v R_i$, then the proof is complete. □

D Proofs of Theorem 1

D.1 Descent in Objective Function

Lemma 7. *Under Asusmption 1, for non-convex and smooth $\widetilde{\Phi}(x)$, the consecutive iterates of Algorithm 1 satisfy:*

$$
\begin{aligned}
&\mathbb{E}[\widetilde{\Phi}(x^{(t+1)})] - \mathbb{E}[\widetilde{\Phi}(x^{(t)})] \\
&\qquad \leq -\frac{\rho^{(t)}\gamma_x}{2}\Big(\mathbb{E}\|\nabla\widetilde{\Phi}(x^{(t)})\|^2 + \mathbb{E}\Big\|\sum_{i=1}^{n} w_i \widetilde{h}_{x,i}^{(t)}\Big\|^2\Big) \\
&\qquad\quad + \frac{\rho^{(t)}\gamma_x}{2}\Big(6(L_1^2 + r^2 L_2^2)\mathbb{E}\|y^{(t)} - \widetilde{y}^*(x^{(t)})\|^2 + 3L_1^2\mathbb{E}\|v^{(x^{(t)})} - \widetilde{v}^*(x^{(t)})\|^2\Big) \\
&\qquad\quad + \frac{3\rho^{(t)}\gamma_x}{2}\sum_{i=1}^{n} w_i \frac{1}{\|a_i^{(t)}\|_1}\sum_{k=0}^{\tau_i - 1} a_i^{(t,k)}\Delta_{f,i}^{(t,k)} + \frac{L_\Phi^2(\rho^{(t)}\gamma_x)^2}{2}\Big\|\sum_{i\in C^{(t)}} \widetilde{w}_i h_{x,i}^{(t)}\Big\|^2
\end{aligned}
$$

for all $t \in \{0, 1, ..., T-1\}$.

Proof. Using the L_Φ in Lemma 4, we have

$$
\begin{aligned}
\mathbb{E}\big[\widetilde{\Phi}(x^{(t+1)})\big] &\leq \mathbb{E}\big[\widetilde{\Phi}(x^{(t)})\big] - \mathbb{E}\Big\langle \nabla\widetilde{\Phi}(x^{(t)}), \rho^{(t)}\gamma_x \sum_{i\in C^{(t)}} \widetilde{w}_i h_{x,i}^{(t)}\Big\rangle \\
&\qquad + \frac{\big(\rho^{(t)}\gamma_x\big)^2 L_\Phi}{2}\mathbb{E}\Big\|\sum_{i\in C^{(t)}} \widetilde{w}_i h_{x,i}^{(t)}\Big\|^2 \\
&\overset{(a)}{=} \mathbb{E}\big[\widetilde{\Phi}(x^{(t)})\big] - \rho^{(t)}\gamma_x \mathbb{E}\Big\langle \nabla\widetilde{\Phi}(x^{(t)}), \sum_{i=1}^{n} w_i \widetilde{h}_{x,i}^{(t)}\Big\rangle \\
&\qquad + \frac{\big(\rho^{(t)}\gamma_x\big)^2 L_\Phi}{2}\mathbb{E}\Big\|\sum_{i\in C^{(t)}} \widetilde{w}_i h_{x,i}^{(t)}\Big\|^2 \\
&= \mathbb{E}\big[\widetilde{\Phi}(x^{(t)})\big] - \frac{\rho^{(t)}\gamma_x}{2}\mathbb{E}\bigg[\Big\|\nabla\widetilde{\Phi}(x^{(t)})\Big\|^2 + \Big\|\sum_{i=1}^{n} w_i \widetilde{h}_{x,i}^{(t)}\Big\|^2\bigg] \\
&\qquad + \frac{\rho^{(t)}\gamma_x}{2}\mathbb{E}\Big\|\nabla\widetilde{\Phi}(x^{(t)}) - \sum_{i=1}^{n} w_i \widetilde{h}_{x,i}^{(t)}\Big\|^2 + \frac{\big(\rho^{(t)}\gamma_x\big)^2 L_\Phi}{2}\mathbb{E}\Big\|\sum_{i\in C^{(t)}} \widetilde{w}_i h_{x,i}^{(t)}\Big\|^2, \qquad (19)
\end{aligned}
$$

where (a) holds because clients are selected without replacement. For the third part of the right-hand side in eq. (19), we have

$$
\begin{aligned}
&\mathbb{E}\Big\|\nabla\widetilde{\Phi}(x^{(t)}) - \sum_{i=1}^{n} w_i \widetilde{h}_{x,i}^{(t,k)}\Big\|^2 \\
&= \mathbb{E}\Big\|\sum_{i=1}^{n} w_i \Big[\bar{\nabla} f_i\big(x^{(t)}, y^*(x^{(t)}), v^*(x^{(t)})\big) - \bar{\nabla} f_i\big(x^{(t)}, y^{(t)}, v^*(x^{(t)})\big) + \bar{\nabla} f_i\big(x^{(t)}, y^{(t)}, v^*(x^{(t)})\big) \\
&\qquad\qquad - \bar{\nabla} f_i\big(x^{(t)}, y^{(t)}, v^{(t)}\big) + \bar{\nabla} f_i(x^{(t)}, y^{(t)}, v^{(t)}) - \widetilde{h}_{x,i}^{(t)}\Big]\Big\|^2 \\
&\leq 3\mathbb{E}\Big\|\sum_{i=1}^{n} w_i \Big[\big(\nabla_x f_i(x^{(t)}, y^*(x^{(t)})) - \nabla_x f_i(x^{(t)}, y^{(t)})\big) \\
&\qquad\qquad - \big(\nabla_{xy}^2 g_i(x^{(t)}, y^*(x^{(t)})) - \nabla_{xy}^2 g_i(x^{(t)}, y^{(t)})\big)v^*(x)\big)\Big]\Big\|^2 \\
&\quad + 3\mathbb{E}\Big\|\sum_{i=1}^{n} w_i \nabla_{xy}^2 g_i(x^{(t)}, y^{(t)})\big(v^*(x^{(t)}) - v^{(t)}\big)\Big\|^2
\end{aligned}
$$

$$
\begin{aligned}
&+ 3\mathbb{E}\Big\|\sum_{i=1}^{n} w_i\big(\bar{\nabla} f_i(x^{(t)}, y^{(t)}, v^{(t)}) - \widetilde{h}_{x,i}^{(t)}\big)\Big\|^2 \\
&\overset{(a)}{\le} 6\big(L_1^2 + r^2 L_2^2\big)\mathbb{E}\big\|y^*(x^{(t)}) - y^{(t)}\big\|^2 + 3L_1^2\mathbb{E}\big\|v^*(x^{(t)}) - v^{(t)}\big\|^2 \\
&\quad + 3\sum_{i=1}^{n} w_i \sum_{k=0}^{\tau_i - 1} \frac{a_i^{(t,k)}}{\|a_i^{(t)}\|_1}\mathbb{E}\big\|\bar{\nabla} f_i(x^{(t)}, y^{(t)}, v^{(t)}) - \bar{\nabla} f_i(x_i^{(t,k)}, y_i^{(t,k)}, v_i^{(t,k)})\big\|^2 \\
&\overset{(b)}{\le} 6\big(L_1^2 + r^2 L_2^2\big)\mathbb{E}\big\|y^*(x^{(t)}) - y^{(t)}\big\|^2 + 3L_1^2\mathbb{E}\big\|v^*(x^{(t)}) - v^{(t)}\big\|^2 + 3\sum_{i=1}^{n} w_i \sum_{k=0}^{\tau_i - 1}\frac{a_i^{(t,k)}}{\|a_i^{(t)}\|_1}\Delta_{f,i}^{(t,k)},
\end{aligned}
$$

where (a) follows from smoothness of $\nabla_x f_i(x, y)$ and L_2-Lipschitz continuity of $\nabla^2_{g_{xy}} g(x, y)$ in Assumption 2 and (b) uses Lemma 6. □

D.2 Bounds of Client Drifts

Lemma 8. *Under Assumption 1, 2 and 4, the local iterates client drifts of* $y_i^{(t,k)}, v_i^{(t,k)}, x_i^{(t,k)}$ *are bounded as*

$$
\begin{aligned}
&\sum_{i=1}^{n} w_i \frac{1}{\|a_i^{(t)}\|_1}\sum_{k=1}^{\tau_i - 1} a_i^{(t,k)}\mathbb{E}\|x_i^{(t,k)} - x^{(t)}\|^2 \le \eta_x^2 \bar{\tau}\sigma_{M1}^2, \\
&\sum_{i=1}^{n} w_i \frac{1}{\|a_i^{(t)}\|_1}\sum_{k=1}^{\tau_i - 1} a_i^{(t,k)}\mathbb{E}\|v_i^{(t,k)} - v^{(t)}\|^2 \le \eta_v^2 \bar{\tau}\sigma_{M1}^2, \\
&\sum_{i=1}^{n} w_i \frac{1}{\|a_i^{(t)}\|_1}\sum_{k=1}^{\tau_i - 1} a_i^{(t,k)}\mathbb{E}\|y_i^{(t,k)} - y^{(t)}\|^2 \le \frac{\eta_y^2\bar{\tau}}{1 - 2\eta_y^2 c_a \bar{\tau}\alpha_{\max}L_1^2}\Big[\alpha_{\max}^2\sigma_g^2 + 2c_a\alpha_{\max}L_1^2\eta_x^2\bar{\tau}\sigma_{M1}^2 \\
&\qquad\qquad + 2c_a\alpha_{\max}L_1^2\mathbb{E}\big\|y^{(t)} - y^*(x^{(t)})\big\|^2 + 2c_a\alpha_{\max}\sigma_{gh}^2\Big]
\end{aligned}
$$

for all $t \in \{0, 1, ..., T-1\}$, $k \in \{0, 1, ..., \tau_i - 1\}$ *and* $i \in \{1, 2, ..., n\}$. *We define* $\sigma_{M1}^2 := \Big(\alpha_{\max}^2(\sigma_f^2 + r_{\max}^2\sigma_{gg}^2) + \alpha_{\max}(L_f^2 + r_{\max}^2 L_1^2)\Big)$. *And* η_y, η_v, η_x *are local stepsizes.*

Proof. For $\big\|y_i^{(t,k)} - y^{(t)}\big\|^2$, we have

$$
\begin{aligned}
&\sum_{i=1}^{n} w_i \sum_{k=0}^{\tau_i - 1}\frac{a_i^{(t,k)}}{\|a_i^{(t)}\|_1}\mathbb{E}\big\|y_i^{(t,k)} - y^{(t)}\big\|^2 \\
&= \eta_y^2 \sum_{i=1}^{n} w_i \sum_{k=0}^{\tau_i - 1}\frac{a_i^{(t,k)}}{\|a_i^{(t)}\|_1}\mathbb{E}\Big\|\sum_{j=0}^{k-1} a_i^{(t,j)}\Big(\nabla_y g_i\big(x_i^{(t,j)}, y_i^{(t,j)}; \zeta_i^{(t,j)}\big) \\
&\qquad\qquad - \nabla_y g_i\big(x_i^{(t,j)}, y_i^{(t,j)}\big) + \nabla_y g_i\big(x_i^{(t,j)}, y_i^{(t,j)}\big)\Big)\Big\|^2 \\
&= \eta_y^2 \sum_{i=1}^{n} w_i \sum_{k=0}^{\tau_i - 1}\frac{a_i^{(t,k)}}{\|a_i^{(t)}\|_1}\sum_{j=0}^{k-1}\Big(a_i^{(t,j)}\Big)^2\mathbb{E}\Big\|\nabla_y g_i\big(x_i^{(t,j)}, y_i^{(t,j)}; \zeta_i^{(t,j)}\big) - \nabla_y g_i\big(x_i^{(t,j)}, y_i^{(t,j)}\big)\Big\|^2 \\
&\quad + \eta_y^2 \sum_{i=1}^{n} w_i \sum_{k=0}^{\tau_i - 1}\frac{a_i^{(t,k)}}{\|a_i^{(t)}\|_1}\mathbb{E}\Big\|\sum_{j=0}^{k-1} a_i^{(t,j)}\nabla_y g_i\big(x_i^{(t,j)}, y_i^{(t,j)}\big)\Big\|^2 \\
&\le \eta_y^2 \sum_{i=1}^{n} w_i \sum_{k=0}^{\tau_i - 1}\frac{a_i^{(t,k)}}{\|a_i^{(t)}\|_1}\sum_{j=0}^{k-1}\Big(a_i^{(t,j)}\Big)^2\sigma_g^2 \\
&\quad + 2\eta_y^2 \sum_{i=1}^{n} w_i \sum_{k=0}^{\tau_i - 1}\frac{a_i^{(t,k)}}{\|a_i^{(t)}\|_1}\sum_{j=0}^{k-1} a_i^{(t,j)}\mathbb{E}\Big\|\nabla_y g_i\big(x_i^{(t,j)}, y_i^{(t,j)}\big) - \nabla_y g_i\big(x^{(t)}, y^{(t)}\big)\Big\|^2
\end{aligned}
$$

$$
\begin{aligned}
&+ 2\eta_y^2 \sum_{i=1}^{n} w_i \sum_{k=0}^{\tau_i-1} \frac{a_i^{(t,k)}}{\|a_i^{(t)}\|_1} \sum_{j=0}^{k-1} a_i^{(t,j)} \mathbb{E}\Big\|\nabla_y g_i\big(x^{(t)}, y^{(t)}\big)\Big\|^2 \\
&\overset{(a)}{\leq} \eta_y^2 \sum_{i=1}^{n} w_i \big\|a_i^{(t)}\big\|_2^2 \sigma_g^2 + 2\eta_y^2 c_a \bar{\tau} \alpha_{\max} \sum_{i=1}^{n} w_i \sum_{k=0}^{\tau_i-1} \frac{a_i^{(t,k)}}{\|a_i^{(t)}\|_1} \mathbb{E}\Big\|\nabla_y g_i\big(x_i^{(t,j)}, y_i^{(t,j)}\big) - \nabla_y g_i\big(x^{(t)}, y^{(t)}\big)\Big\|^2 \\
&+ 2\eta_y^2 \sum_{i=1}^{n} w_i \big\|a_i^{(t)}\big\|_1 \mathbb{E}\Big\|\nabla_y g_i\big(x^{(t)}, y^{(t)}\big)\Big\|^2 \\
&\overset{(b)}{\leq} \eta_y^2 \bar{\tau} \alpha_{\max}^2 \sigma_g^2 + 2\eta_y^2 c_a \bar{\tau} \alpha_{\max} \sum_{i=1}^{n} w_i \sum_{k=0}^{\tau_i-1} \frac{a_i^{(t,k)}}{\|a_i^{(t)}\|_1} \mathbb{E}\Big\|\nabla_y g_i\big(x_i^{(t,j)}, y_i^{(t,j)}\big) - \nabla_y g_i\big(x^{(t)}, y^{(t)}\big)\Big\|^2 \\
&+ 2\eta_y^2 c_a \bar{\tau} \alpha_{\max} \beta_{gh}^2 \mathbb{E}\Bigg\|\sum_{i=1}^{n} w_i \Big[\nabla_y g_i\big(x^{(t)}, y^{(t)}\big) - \nabla_y g_i\big(x^{(t)}, y^*(x^{(t)})\big)\Big]\Bigg\|^2 + 2\eta_y^2 c_a \bar{\tau} \alpha_{\max} \sigma_{gh}^2 \\
&\overset{(c)}{\leq} \eta_y^2 \bar{\tau} \alpha_{\max}^2 \sigma_g^2 + 2\eta_y^2 c_a \bar{\tau} \alpha_{\max} \sum_{i=1}^{n} w_i \sum_{k=0}^{\tau_i-1} \frac{a_i^{(t,k)}}{\|a_i^{(t)}\|_1} L_1^2 \mathbb{E}\Big[\big\|x_i^{(t,j)} - x^{(t)}\big\|^2 + \big\|y_i^{(t,j)} - y^{(t)}\big\|^2\Big] \\
&+ 2\eta_y^2 c_a \bar{\tau} \alpha_{\max} \beta_{gh}^2 L_1^2 \mathbb{E}\big\|y^{(t)} - y^*(x^{(t)})\big\|^2 + 2\eta_y^2 c_a \bar{\tau} \alpha_{\max} \beta_{\max} \sigma_{gh}^2
\end{aligned} \tag{20}
$$

where (a) holds because of Assumption 3 and

$$
\frac{1}{\|a_i^{(t)}\|_1} \sum_{k=0}^{\tau_i-1} a_i^{(t,k)} \sum_{j=0}^{k-1} \Big(a_i^{(t,j)}\Big)^2 \leq \frac{1}{\|a_i^{(t)}\|_1} \sum_{k=0}^{\tau_i-1} a_i^{(t,k)} \sum_{j=0}^{\tau_i-2} \Big(a_i^{(t,j)}\Big)^2 = \sum_{j=0}^{\tau_i-2} \Big(a_i^{(t,j)}\Big)^2 \leq \|a_i\|_2^2
$$

$$
\frac{1}{\|a_i^{(t)}\|_1} \sum_{k=0}^{\tau_i-1} a_i^{(t,k)} \sum_{j=0}^{k-1} a_i^{(t,j)} \leq \frac{1}{\|a_i^{(t)}\|_1} \sum_{k=0}^{\tau_i-1} a_i^{(t,k)} \sum_{j=0}^{\tau_i-2} a_i^{(t,j)} = \sum_{j=0}^{\tau_i-2} a_i^{(t,j)} \leq \|a_i^{(t)}\|_1;
$$

(b) is obtained from Assumption 4 and (c) follows from Assumption 2. As an easier case, we can easily know that $\bar{\nabla} f_i$ and $\nabla_v R_i$ are bounded from Assumption 2 and Lemma 2. Then we have

$$
\begin{aligned}
&\sum_{i=1}^{n} w_i \frac{1}{\|a_i^{(t)}\|_1} \sum_{k=1}^{\tau_i-1} a_i^{(t,k)} \mathbb{E}\|x_i^{(t,k)} - x^{(t)}\|^2 \leq \eta_x^2 \bar{\tau} \Big(\alpha_{\max}^2 (\sigma_f^2 + r_{\max}^2 \sigma_{gg}^2) + \alpha_{\max}(L_f^2 + r_{\max}^2 L_1^2)\Big), \\
&\sum_{i=1}^{n} w_i \frac{1}{\|a_i^{(t)}\|_1} \sum_{k=1}^{\tau_i-1} a_i^{(t,k)} \mathbb{E}\|v_i^{(t,k)} - v^{(t)}\|^2 \leq \eta_v^2 \bar{\tau} \Big(\alpha_{\max}^2 (\sigma_f^2 + r_{\max}^2 \sigma_{gg}^2) + \alpha_{\max}(L_f^2 + r_{\max}^2 L_1^2)\Big), \\
&\sum_{i=1}^{n} w_i \frac{1}{\|a_i^{(t)}\|_1} \sum_{k=1}^{\tau_i-1} a_i^{(t,k)} \mathbb{E}\|y_i^{(t,k)} - y^{(t)}\|^2 \\
&\leq \frac{\eta_y^2 \bar{\tau}}{1 - 2\eta_y^2 c_a \bar{\tau} \alpha_{\max} L_1^2} \Bigg[\alpha_{\max}^2 \sigma_g^2 + 2c_a \alpha_{\max} L_1^2 \eta_x^2 \bar{\tau} \Big(\alpha_{\max}^2 (\sigma_f^2 + r_{\max}^2 \sigma_{gg}^2) + \alpha_{\max}(L_f^2 + r_{\max}^2 L_1^2)\Big) \\
&\qquad + 2c_a \alpha_{\max} \beta_{gh}^2 L_1^2 \mathbb{E}\big\|y^{(t)} - y^*(x^{(t)})\big\|^2 + 2c_a \alpha_{\max} \sigma_{gh}^2\Bigg]
\end{aligned}
$$

which finished the proof. □

In the later part, we will take $\{2\eta_x^2 \bar{\tau} c_a \alpha_{\max} L_1^2 \leq 1\}$ and $\{4\eta_y^2 c_a \bar{\tau} \alpha_{\max} L_1^2 \leq 1\}$, then we can simplify Lemma 8 as

$$
\begin{aligned}
&\sum_{i=1}^{n} w_i \frac{1}{\|a_i^{(t)}\|_1} \sum_{k=1}^{\tau_i-1} a_i^{(t,k)} \mathbb{E}\|x_i^{(t,k)} - x^{(t)}\|^2 \leq \eta_x^2 \bar{\tau} \sigma_{M1}^2, \\
&\sum_{i=1}^{n} w_i \frac{1}{\|a_i^{(t)}\|_1} \sum_{k=1}^{\tau_i-1} a_i^{(t,k)} \mathbb{E}\|v_i^{(t,k)} - v^{(t)}\|^2 \leq \eta_v^2 \bar{\tau} \sigma_{M1}^2, \\
&\sum_{i=1}^{n} w_i \frac{1}{\|a_i^{(t)}\|_1} \sum_{k=1}^{\tau_i-1} a_i^{(t,k)} \mathbb{E}\|y_i^{(t,k)} - y^{(t)}\|^2 \leq 2\eta_y^2 \bar{\tau} \sigma_{M2}^2 + 4\eta_y^2 \bar{\tau} c_a \alpha_{\max} \beta_{gh}^2 L_1^2 \mathbb{E}\big\|y^{(t)} - y^*(x^{(t)})\big\|^2
\end{aligned} \tag{21}
$$

where we define $\sigma_{M2}^2 := \alpha_{\max}^2\sigma_g^2 + \alpha_{\max}^2(\sigma_f^2 + r_{\max}^2\sigma_{gg}^2) + \alpha_{\max}(L_f^2 + r_{\max}^2 L_1^2) + 2c_a\alpha_{\max}\sigma_{gh}^2$. As the combination of Lemma 6 and Lemma 8, we have

$$\begin{aligned}
\sum_{i=1}^n w_i \sum_{k=1}^{\tau_i-1} \frac{a_i^{(t,k)}}{\|a_i^{(t)}\|_1}\Delta_{f_i}^{(t,k)} &= \sum_{i=1}^n w_i \sum_{k=1}^{\tau_i-1} \frac{a_i^{(t,k)}}{\|a_i^{(t)}\|_1}\Delta_{R_i}^{(t,k)} \\
&\le 3\eta_x^2\bar{\tau}\big(L_1^2 + r^2L_2^2\big)\sigma_{M1}^2 + 3\eta_v^2\bar{\tau}L_1^2\sigma_{M1}^2 + 6\eta_y^2\bar{\tau}\big(L_1^2 + r^2L_2^2\big)\sigma_{M2}^2 \\
&\quad + 12\eta_y^2\bar{\tau}\big(L_1^2 + r^2L_2^2\big)c_a\alpha_{\max}\beta_{gh}^2L_1^2\mathbb{E}\big\|y^{(t)} - y^*(x^{(t)})\big\|^2 \\
\sum_{i=1}^n w_i \sum_{k=1}^{\tau_i-1} \frac{a_i^{(t,k)}}{\|a_i^{(t)}\|_1}\Delta_{g_i}^{(t,k)} &\le \eta_x^2\bar{\tau}L_1^2\sigma_{M1}^2 + 2\eta_y^2\bar{\tau}L_1^2\sigma_{M2}^2 + 4\eta_y^2\bar{\tau}c_a\alpha_{\max}\beta_{gh}^2L_1^4\mathbb{E}\big\|y^{(t)} - y^*(x^{(t)})\big\|^2.
\end{aligned} \tag{22}$$

D.3 Bounds of Aggregated Estimations

Lemma 9. *Suppose the server selects $|C^{(t)}| = P$ clients in each round. Under Assumption 1, 2 and 3, the aggregated estimation of $x^{(t)}$ satisfies*

$$\begin{aligned}
\mathbb{E}\Big\|\sum_{i\in C^{(t)}} \widetilde{w}_i h_{x,i}^{(t)}\Big\|^2 \le & \frac{2n}{P}\sum_{i=1}^n \frac{w_i^2}{\|a_i^{(t)}\|_1^2}\sum_{k=0}^{\tau_i-1}\big(a_i^{(t,k)}\big)^2(\sigma_f^2 + r_i^2\sigma_{gg}^2) + \frac{n(P-1)}{P(n-1)}\mathbb{E}\Big\|\sum_{i=1}^n w_i\widetilde{h}_{x,i}^{(t)}\Big\|^2 \\
& + \frac{2n(n-P)}{P(n-1)}\sum_{i=1}^n \frac{w_i^2}{\|a_i^{(t)}\|_1}\sum_{k=0}^{\tau_i-1} a_i^{(t,k)}\Delta_{f,i}^{(t,k)} + \frac{4(n-P)\beta_{\max}}{P(n-1)}(r_{\max}^2L_1^2 + L_f^2).
\end{aligned}$$

the aggregated estimation of $y^{(t)}$ satisfies

$$\begin{aligned}
\mathbb{E}\Big\|\sum_{i\in C^{(t)}} \widetilde{w}_i h_{y,i}^{(t)}\Big\|^2 \le & \frac{n}{P}\sum_{i=1}^n \frac{w_i^2}{\|a_i^{(t)}\|_1^2}\sum_{k=0}^{\tau_i-1}\big(a_i^{(t,k)}\big)^2\sigma_g^2 + \frac{2(n-P)\beta_{\max}}{P(n-1)}\sigma_{gh}^2 \\
& + \Big(\frac{2n(n-P)}{P(n-1)}\sum_{i=1}^n w_i^2\sum_{k=0}^{\tau_i-1}\frac{a_i^{(t,k)}}{\|a_i^{(t)}\|_1} + 3\sum_{i=1}^n w_i\sum_{k=0}^{\tau_i-1}\frac{a_i^{(t,k)}}{\|a_i^{(t)}\|_1}\Big)\Delta_{g,i}^{(t,k)} \\
& + \Big(\frac{2(n-P)\beta_{\max}\beta_{gh}^2}{P(n-1)} + 3L_1^2\Big)\mathbb{E}\|y^{(t)} - y^*(x^{(t)})\|^2.
\end{aligned}$$

and the aggregated estimation of $v^{(t)}$ satisfies

$$\begin{aligned}
\mathbb{E}\Big\|\sum_{i\in C^{(t)}} \widetilde{w}_i h_{v,i}^{(t)}\Big\|^2 \le & \frac{2n}{P}\sum_{i=1}^n \frac{w_i^2}{\|a_i^{(t)}\|_1^2}\sum_{k=0}^{\tau_i-1}\big(a_i^{(t,k)}\big)^2\big(\sigma_f^2 + r_i^2\sigma_{gg}^2\big) + \frac{4(n-P)\beta_{\max}}{P(n-1)}(r_{\max}^2L_1^2 + L_f^2) \\
& + \Big(\frac{2n(n-P)}{P(n-1)}\sum_{i=1}^n w_i^2\sum_{k=0}^{\tau_i-1}\frac{a_i^{(t,k)}}{\|a_i^{(t)}\|_1} + 3\sum_{i=1}^n w_i\sum_{k=0}^{\tau_i-1}\frac{a_i^{(t,k)}}{\|a_i^{(t)}\|_1}\Big)\Delta_{R,i}^{(t,k)} \\
& + 3L_1^2\mathbb{E}\|v^{(t)} - v^*(x^{(t)})\|^2.
\end{aligned}$$

for all $t \in \{0, 1, ..., T-1\}$, $k \in \{0, 1, ..., \tau_i - 1\}$ and $i \in \{1, 2, ..., n\}$.

Proof. For the aggregated estimation of $x^{(t)}$, we have

$$\begin{aligned}
\mathbb{E}\Big\|\sum_{i\in C^{(t)}} \widetilde{w}_i h_{x,i}^{(t)}\Big\|^2 &= \mathbb{E}\Big\|\sum_{i\in C^{(t)}} \widetilde{w}_i\Big(h_{x,i}^{(t)} - \widetilde{h}_{x,i}^{(t)} + \widetilde{h}_{x,i}^{(t)}\Big)\Big\|^2 \\
&= \mathbb{E}\Big\|\sum_{i\in C^{(t)}} \widetilde{w}_i\Big(h_{x,i}^{(t)} - \widetilde{h}_{x,i}^{(t)}\Big)\Big\|^2 + \mathbb{E}\Big\|\sum_{i\in C^{(t)}} \widetilde{w}_i\widetilde{h}_{x,i}^{(t)}\Big\|^2 \\
&\overset{(a)}{=} \mathbb{E}\Big[\sum_{i\in C^{(t)}} \widetilde{w}_i^2\Big\|h_{x,i}^{(t)} - \widetilde{h}_{x,i}^{(t)}\Big\|^2\Big] + \mathbb{E}\Big\|\sum_{i\in C^{(t)}} \widetilde{w}_i\widetilde{h}_{x,i}^{(t)}\Big\|^2
\end{aligned}$$

$$
\begin{aligned}
&\overset{(b)}{=} \frac{n}{P}\sum_{i=1}^{n} w_i^2 \mathbb{E}\left\| h_{x,i}^{(t)} - \widetilde{h}_{x,i}^{(t)} \right\|^2 + \mathbb{E}\left\| \sum_{i\in C^{(t)}} \widetilde{w}_i \widetilde{h}_{x,i}^{(t)} \right\|^2 \\
&\overset{(c)}{\leq} \frac{2n}{P}\sum_{i=1}^{n} \frac{w_i^2}{\|a_i^{(t)}\|_1^2} \sum_{k=0}^{\tau_i-1} \left(a_i^{(t,k)}\right)^2 (\sigma_f^2 + r_i^2\sigma_{gg}^2) + \mathbb{E}\left\| \sum_{i\in C^{(t)}} \widetilde{w}_i \widetilde{h}_{x,i}^{(t)} \right\|^2, \qquad (23)
\end{aligned}
$$

where (a) holds because clients are selected without replacement; (b) follows from the definition $\widetilde{w}_i = \frac{n}{P} w_i$; (c) uses Assumption 3. For the second term in eq. (23), we have

$$
\begin{aligned}
&\mathbb{E}\left\| \sum_{i\in C^{(t)}} \widetilde{w}_i \widetilde{h}_{x,i}^{(t)} \right\|^2 \\
&\quad= \mathbb{E}\left\| \sum_{i\in C^{(t)}} \widetilde{w}_i \widetilde{h}_{x,i}^{(t)} - \sum_{i=i}^{n} w_i \widetilde{h}_{x,i}^{(t)} + \sum_{i=i}^{n} w_i \widetilde{h}_{x,i}^{(t)} \right\|^2 \\
&\quad\overset{(a)}{=} \mathbb{E}\left\| \sum_{i=1}^{n} \mathbb{I}(i\in C^{(t)}) \widetilde{w}_i \widetilde{h}_{x,i}^{(t)} - \sum_{i=i}^{n} w_i \widetilde{h}_{x,i}^{(t)} \right\|^2 + \mathbb{E}\left\| \sum_{i=i}^{n} w_i \widetilde{h}_{x,i}^{(t)} \right\|^2 \\
&\quad= \sum_{i=1}^{n} \mathbb{E}\left[\left(\mathbb{I}(i\in C^{(t)})^2 \widetilde{w}_i^2 + w_i^2 - 2\mathbb{I}(i\in C^{(t)}) \widetilde{w}_i w_i \right) \left\| h_{x,i}^{(t)} \right\|^2 \right] \\
&\qquad+ \sum_{i\neq j} \mathbb{E}\left\langle (\mathbb{I}(i\in C^{(t)})\widetilde{w}_i - w_i) h_{x,i}^{(t)}, (\mathbb{I}(j\in C^{(t)})\widetilde{w}_j - w_j) h_{x,j}^{(t)} \right\rangle + \mathbb{E}\left\| \sum_{i=1}^{n} w_i \widetilde{h}_{x,i}^{(t)} \right\|^2 \\
&\quad= \mathbb{E}\left\| \sum_{i=1}^{n} w_i \widetilde{h}_{x,i}^{(t)} \right\|^2 + \sum_{i=1}^{n} \mathbb{E}\left[w_i^2 \left(\frac{n}{P} - 1 \right) \left\| h_{x,i}^{(t)} \right\|^2 \right] \\
&\qquad+ \sum_{i\neq j} \mathbb{E}\left[\left(\mathbb{I}(i,j\in C^{(t)}) \widetilde{w}_i \widetilde{w}_j - \mathbb{I}(j\in C^{(t)}) \widetilde{w}_j w_i - \mathbb{I}(i\in C^{(t)}) \widetilde{w}_i w_j + w_i w_j \right) \left\langle h_{x,i}^{(t)}, h_{x,j}^{(t)} \right\rangle \right] \\
&\quad= \mathbb{E}\left\| \sum_{i=1}^{n} w_i \widetilde{h}_{x,i}^{(t)} \right\|^2 + \left(\frac{n}{P} - 1 \right) \sum_{i=1}^{n} \mathbb{E}\left[w_i^2 \left\| \widetilde{h}_{x,i}^{(t)} \right\|^2 \right] \\
&\qquad+ \sum_{i\neq j} \mathbb{E}\left[w_i w_j \left(\frac{n}{P}\left(\frac{P-1}{n-1} \right) - 1 \right) \left\langle h_{x,i}^{(t)}, h_{x,j}^{(t)} \right\rangle \right] \\
&\quad= \frac{n}{P}\left(\frac{P-1}{n-1} \right) \mathbb{E}\left\| \sum_{i=1}^{n} w_i \widetilde{h}_{x,i}^{(t)} \right\|^2 + \frac{n}{P}\left(\frac{n-P}{n-1} \right) \sum_{i=1}^{n} w_i^2 \mathbb{E}\left\| \widetilde{h}_{x,i}^{(t)} \right\|^2, \qquad (24)
\end{aligned}
$$

where (a) holds because clients are selected without replacement. And for the second term in eq. (24), we have

$$
\begin{aligned}
&\sum_{i=1}^{n} w_i^2 \mathbb{E}\left\| \widetilde{h}_{x,i}^{(t)} \right\|^2 \\
&\quad= \sum_{i=1}^{n} w_i^2 \mathbb{E}\left\| \widetilde{h}_{x,i}^{(t)} - \bar{\nabla} f(x^{(t)}, y^{(t)}, v^{(t)}) + \bar{\nabla} f(x^{(t)}, y^{(t)}, v^{(t)}) \right\|^2 \\
&\quad\overset{(a)}{\leq} 2\sum_{i=1}^{n} w_i^2 \mathbb{E}\left\| \widetilde{h}_{x,i}^{(t)} - \bar{\nabla} f(x^{(t)}, y^{(t)}, v^{(t)}) \right\|^2 + 2\frac{\beta_{\max}}{n} \sum_{i=1}^{n} w_i \mathbb{E}\left\| \bar{\nabla} f(x^{(t)}, y^{(t)}, v^{(t)}) \right\|^2 \\
&\quad\overset{(b)}{\leq} 2\sum_{i=1}^{n} w_i^2 \sum_{k=0}^{\tau_i-1} \frac{a_i^{(t,k)}}{\|a_i^{(t)}\|_1} \mathbb{E}\left\| \bar{\nabla} f(x_i^{(t,k)}, y_i^{(t,k)}, v_i^{(t,k)}) - \bar{\nabla} f(x^{(t)}, y^{(t)}, v^{(t)}) \right\|^2 \\
&\qquad+ \frac{4\beta_{\max}}{n} (r_{\max}^2 L_1^2 + L_f^2) \\
&\quad\overset{(c)}{\leq} 2\sum_{i=1}^{n} w_i^2 \sum_{k=0}^{\tau_i-1} \frac{a_i^{(t,k)}}{\|a_i^{(t)}\|_1} \mathbb{E}\Delta_{f,i}^{(t,k)} + \frac{4\beta_{\max}}{n} (r_{\max}^2 L_1^2 + L_f^2) \qquad (25)
\end{aligned}
$$

where (a) $w_i \le \beta_{\max}/n$ for all $i \in \{1, .., n\}$; the first term of (b) uses Jensen's inequality and the second part of (b) follows from Lemma 5; (c) uses theLemma 6. By incorporating eq. (24) and eq. (25) into eq. (23), we get

$$\begin{aligned}
\mathbb{E}\Big\| \sum_{i \in C^{(t)}} \widetilde{w}_i h_{x,i}^{(t)} \Big\|^2 \le & \frac{2n}{P} \sum_{i=1}^{n} \frac{w_i^2}{\|a_i^{(t)}\|_1^2} \sum_{k=0}^{\tau_i - 1} \big(a_i^{(t,k)}\big)^2 (\sigma_f^2 + r_i^2 \sigma_{gg}^2) + \frac{n(P-1)}{P(n-1)} \mathbb{E}\Big\| \sum_{i=1}^{n} w_i \widetilde{h}_{x,i}^{(t)} \Big\|^2 \\
& + \frac{2n(n-P)}{P(n-1)} \sum_{i=1}^{n} \frac{w_i^2}{\|a_i^{(t)}\|_1} \sum_{k=0}^{\tau_i - 1} a_i^{(t,k)} \Delta_{f,i}^{(t,k)} \\
& + \frac{4(n-P)\beta_{\max}}{P(n-1)} (r_{\max}^2 L_1^2 + L_f^2).
\end{aligned}$$

Similarly, by replacing $h_{x,i}^{(t)}$ and $\bar{\nabla} f$ with $h_{y,i}^{(t)}$ and ∇g, we can easily get

$$\begin{aligned}
\mathbb{E}\Big\| \sum_{i \in C^{(t)}} \widetilde{w}_i h_{y,i}^{(t)} \Big\|^2 \le & \frac{n}{P} \sum_{i=1}^{n} \frac{w_i^2}{\|a_i^{(t)}\|_1^2} \sum_{k=0}^{\tau_i - 1} \big(a_i^{(t,k)}\big)^2 \sigma_g^2 + \frac{n(P-1)}{P(n-1)} \mathbb{E}\Big\| \sum_{i=1}^{n} w_i \widetilde{h}_{y,i}^{(t)} \Big\|^2 \\
& + \frac{2n(n-P)}{P(n-1)} \sum_{i=1}^{n} \frac{w_i^2}{\|a_i^{(t)}\|_1} \sum_{k=0}^{\tau_i - 1} a_i^{(t,k)} \Delta_{g,i}^{(t,k)} + \frac{2(n-P)\beta_{\max}}{P(n-1)} \sigma_{gh}^2 \\
& + \frac{2(n-P)\beta_{\max}\beta_{gh}^2}{P(n-1)} \mathbb{E}\|y^{(t)} - y^*(x^{(t)})\|^2. \qquad (26)
\end{aligned}$$

For the second terms in eq. (26), we have

$$\begin{aligned}
\mathbb{E}\Big\| \sum_{i=1}^{n} w_i \widetilde{h}_{y,i}^{(t)} \Big\|^2 \le & \; 3\mathbb{E}\Big\| \sum_{i=1}^{n} w_i \nabla_y g_i(x^{(t)}, y^*(x^{(t)})) \Big\|^2 \\
& + 3\mathbb{E}\Big\| \sum_{i=1}^{n} w_i \Big(\nabla_y g_i\big(x^{(t)}, y^{(t)}\big) - \nabla_y g_i\big(x^{(t)}, y^*(x^{(t)})\big)\Big) \Big\|^2 \\
& + 3\mathbb{E}\Big\| \sum_{i=1}^{n} w_i \Big(\widetilde{h}_{y,i}^{(t)} - \nabla_y g_i\big(x^{(t)}, y^{(t)}\big)\Big) \Big\|^2 \\
\overset{(a)}{\le} & \; 3\mathbb{E}\Big\| \sum_{i=1}^{n} w_i \Big(\nabla_y g_i\big(x^{(t)}, y^{(t)}\big) - \nabla_y g_i\big(x^{(t)}, y^*(x^{(t)})\big)\Big) \Big\|^2 \\
& + 3\mathbb{E}\Big\| \sum_{i=1}^{n} w_i \Big(\widetilde{h}_{y,i}^{(t)} - \nabla_y g_i\big(x^{(t)}, y^{(t)}\big)\Big) \Big\|^2 \\
\overset{(b)}{\le} & \; 3L_1^2 \mathbb{E}\|y^{(t)} - y^*(x^{(t)})\|^2 \\
& + 3\sum_{i=1}^{n} w_i \sum_{k=0}^{\tau_i - 1} \frac{a_i^{(t,k)}}{\|a_i^{(t)}\|_1} \mathbb{E}\Big\| \nabla_y g_i\big(x_i^{(t,k)}, y_i^{(t,k)}\big) - \nabla_y g_i\big(x^{(t)}, y^{(t)}\big) \Big\|^2 \\
\overset{(b)}{\le} & \; 3L_1^2 \mathbb{E}\|y^{(t)} - y^*(x^{(t)})\|^2 + 3\sum_{i=1}^{n} w_i \sum_{k=0}^{\tau_i - 1} \frac{a_i^{(t,k)}}{\|a_i^{(t)}\|_1} \Delta_{g,i}^{(t,k)}. \qquad (27)
\end{aligned}$$

where (a) follows from $\sum_{i=1}^{n} w_i \nabla_y g_i(x^{(t)}, y^*(x^{(t)})) = 0$; the first term of (b) uses Assumption 2 and the second term of (b) uses Jensen inequality; (c) follows from Lemma 6. By incorporate 27 into 26, we get

$$\begin{aligned}
\mathbb{E}\Big\| \sum_{i \in C^{(t)}} \widetilde{w}_i h_{y,i}^{(t)} \Big\|^2 \le & \frac{n}{P} \sum_{i=1}^{n} \frac{w_i^2}{\|a_i^{(t)}\|_1^2} \sum_{k=0}^{\tau_i - 1} \big(a_i^{(t,k)}\big)^2 \sigma_g^2 + \frac{2(n-P)\beta_{\max}}{P(n-1)} \sigma_{gh}^2 \\
& + \Big(\frac{2n(n-P)}{P(n-1)} \sum_{i=1}^{n} w_i^2 \sum_{k=0}^{\tau_i - 1} \frac{a_i^{(t,k)}}{\|a_i^{(t)}\|_1} + 3\sum_{i=1}^{n} w_i \sum_{k=0}^{\tau_i - 1} \frac{a_i^{(t,k)}}{\|a_i^{(t)}\|_1} \Big) \Delta_{g,i}^{(t,k)}
\end{aligned}$$

$$+\left(\frac{2(n-P)\beta_{\max}\beta_{gh}^2}{P(n-1)}+3L_1^2\right)\mathbb{E}\|y^{(t)}-y^*(x^{(t)})\|^2.$$

Similarly, by replacing by replacing $h_{y,i}^{(t)}$ and ∇g with $h_{R,i}^{(t)}$ and ∇R, we can easily get

$$\begin{aligned}\mathbb{E}\Big\|\sum_{i\in C^{(t)}}\widetilde{w}_i h_{v,i}^{(t)}\Big\|^2 \le& \frac{2n}{P}\sum_{i=1}^{n}\frac{w_i^2}{\|a_i^{(t)}\|_1^2}\sum_{k=0}^{\tau_i-1}\big(a_i^{(t,k)}\big)^2\big(\sigma_f^2+r_i^2\sigma_{gg}^2\big)\\ &+\left(\frac{2n(n-P)}{P(n-1)}\sum_{i=1}^{n}w_i^2\sum_{k=0}^{\tau_i-1}\frac{a_i^{(t,k)}}{\|a_i^{(t)}\|_1}+3\sum_{i=1}^{n}w_i\sum_{k=0}^{\tau_i-1}\frac{a_i^{(t,k)}}{\|a_i^{(t)}\|_1}\right)\Delta_{R,i}^{(t,k)}\\ &+3L_1^2\mathbb{E}\|v^{(t)}-v^*(x^{(t)})\|^2+\frac{4(n-P)\beta_{\max}}{P(n-1)}(r_{\max}^2L_1^2+L_f^2).\end{aligned}\tag{28}$$

Then, the proof is complete. □

D.4 Descent in iterates of the inner- and LS-problem

Lemma 10. *Under the Assumption 1, 2 and 3, the iterates of the inner-problem generated according to Algorithm 1 satisfy*

$$\begin{aligned}&\mathbb{E}\|y^{(t+1)}-\widetilde{y}^*(x^{(t+1)})\|^2-\mathbb{E}\|y^{(t)}-\widetilde{y}^*(x^{(t)})\|^2\\ \le&(\delta_t-\rho^{(t)}\gamma_y\mu_g-\delta_t\rho^{(t)}\gamma_y\mu_g)\mathbb{E}\|y^{(t)}-\widetilde{y}^*(x^{(t)})\|^2+(1+\delta_t)(\rho^{(t)}\gamma_y)^2\mathbb{E}\Big\|\sum_{i\in C^{(t)}}\widetilde{w}_i h_{y,i}^{(t)}\Big\|^2\\ &+(1+\delta_t)\rho^{(t)}\gamma_y\frac{2L_1^2}{\mu_g}\sum_{i=1}^{n}w_i\sum_{k=0}^{\tau_i-1}\frac{a_i^{(t,k)}}{\|a_i^{(t)}\|_1}\mathbb{E}\Big[\big\|x^{(t)}-x_i^{(t,k)}\big\|^2+\big\|y^{(t)}-y_i^{(t,k)}\big\|^2\Big]\\ &+\big(\rho^{(t)}\gamma_x\big)^2\left(L_y^2+\frac{L_{yx}}{2}\right)\mathbb{E}\Big\|\sum_{i\in C^{(t)}}\widetilde{w}_i h_{x,i}^{(t)}\Big\|^2+(\rho^{(t)}\gamma_x)^2\frac{4L_y}{\delta_{t,1}}\mathbb{E}\Big\|\sum_{i=1}^{n}w_i\widetilde{h}_{x,i}^{(t)}\Big\|^2.\end{aligned}$$

and the iterates of the LS problem satisfy

$$\begin{aligned}&\mathbb{E}\|v^{(t+1)}-\widetilde{v}^*(x^{(t+1)})\|^2-\mathbb{E}\|v^{(t)}-\widetilde{v}^*(x^{(t)})\|^2\\ \le&(\delta_t'-\rho^{(t)}\gamma_v\mu_g-\delta_t'\rho^{(t)}\gamma_v\mu_g)\mathbb{E}\|v^{(t)}-\widetilde{v}^*(x^{(t)})\|^2+(1+\delta_t')(\rho^{(t)}\gamma_v)^2\mathbb{E}\Big\|\sum_{i\in C^{(t)}}\widetilde{w}_i h_{v,i}^{(t)}\Big\|^2\\ &+(1+\delta_t')\rho^{(t)}\gamma_v\frac{4L_R^2}{\mu_g}\sum_{i=1}^{n}w_i\sum_{k=0}^{\tau_i-1}\frac{a_i^{(t,k)}}{\|a_i^{(t)}\|_1}\\ &\quad\times\mathbb{E}\Big[\big\|x^{(t)}-x_i^{(t,k)}\big\|^2+\big\|y^{(t)}-y_i^{(t,k)}\big\|^2+\big\|v^{(t)}-v_i^{(t,k)}\big\|^2\Big]\\ &+(1+\delta_t')\rho^{(t)}\gamma_v\frac{4L_R^2}{\mu_g}\mathbb{E}\big\|y^{(t)}-\widetilde{y}^*(x^{(t)})\big\|^2+\big(\rho^{(t)}\gamma_x\big)^2\left(L_v^2+\frac{L_{vx}}{2}\right)\mathbb{E}\Big\|\sum_{i\in C^{(t)}}\widetilde{w}_i h_{x,i}^{(t)}\Big\|^2\\ &+(\rho^{(t)}\gamma_x)^2\frac{4L_v}{\delta_{t,1}'}\mathbb{E}\Big\|\sum_{i=1}^{n}w_i\widetilde{h}_{x,i}^{(t)}\Big\|^2.\end{aligned}$$

for all $t\in\{0,1,...,T-1\}$, $k\in\{0,1,...,\tau_i-1\}$ *and* $i\in\{1,2,...,n\}$.

Proof. For the gap of y and $\widetilde{y}^*$ on server, we have

$$\begin{aligned}\mathbb{E}\big\|y^{(t+1)}-\widetilde{y}^*(x^{(t+1)})\big\|^2=&\mathbb{E}\big\|y^{(t+1)}-\widetilde{y}^*(x^{(t)})\big\|^2+\mathbb{E}\big\|\widetilde{y}^*(x^{(t)})-\widetilde{y}^*(t+1)\big\|^2\\ &+2\mathbb{E}\big\langle y^{(t+1)}-\widetilde{y}^*(x^{(t)}),\widetilde{y}^*(x^{(t)})-\widetilde{y}^*(x^{(t+1)})\big\rangle.\end{aligned}\tag{29}$$

For the last term in eq. (29), we have

$$2\mathbb{E}\big\langle y^{(t+1)}-\widetilde{y}^*(x^{(t)}),\widetilde{y}^*(x^{(t)})-\widetilde{y}^*(x^{(t+1)})\big\rangle$$

$$
\begin{aligned}
&= -\,2\mathbb{E}\Big\langle y^{(t+1)} - \widetilde{y}^*(x^{(t)}), \nabla\widetilde{y}^*(x^{(t)})\big(x^{(t+1)} - x^{(t)}\big)\Big\rangle \\
&\quad -\,2\mathbb{E}\Big\langle y^{(t+1)} - \widetilde{y}^*(x^{(t)}), \widetilde{y}^*(x^{(t+1)}) - \widetilde{y}^*(x^{(t)}) - \nabla\widetilde{y}^*(x^{(t)})\big(x^{(t+1)} - x^{(t)}\big)\Big\rangle \\
&\leq 2\mathbb{E}\big\|y^{(t+1)} - \widetilde{y}^*(x^{(t)})\big\| \cdot \mathbb{E}\Big\|\rho^{(t)}\gamma_x\nabla\widetilde{y}^*(x^{(t)})\sum_{i=1}^{n} w_i\widetilde{h}_{x,i}^{(t)}\Big\| \\
&\quad + 2\mathbb{E}\big\|y^{(t+1)} - \widetilde{y}^*(x^{(t)})\big\| \cdot \mathbb{E}\Big\|\widetilde{y}^*(x^{(t+1)}) - \widetilde{y}^*(x^{(t)}) - \nabla\widetilde{y}^*(x^{(t)})\big(x^{(t+1)} - x^{(t)}\big)\Big\| \\
&\overset{(a)}{\leq} 2\mathbb{E}\big\|y^{(t+1)} - \widetilde{y}^*(x^{(t)})\big\| \cdot \mathbb{E}\Big\|\rho^{(t)}\gamma_x\nabla\widetilde{y}^*(x^{(t)})\sum_{i=1}^{n} w_i\widetilde{h}_{x,i}^{(t)}\Big\| \\
&\quad + L_{yx}\mathbb{E}\big\|y^{(t+1)} - \widetilde{y}^*(x^{(t)})\big\| \cdot \mathbb{E}\big\|x^{(t+1)} - x^{(t)}\big\|^2 \\
&\leq \delta_{t,1}\mathbb{E}\big\|y^{(t+1)} - \widetilde{y}^*(x^{(t)})\big\|^2 + \frac{(\rho^{(t)}\gamma_x)^2 L_y^2}{\delta_{t,1}}\mathbb{E}\Big\|\sum_{i=1}^{n} w_i\widetilde{h}_{x,i}^{(t)}\Big\|^2 \\
&\quad + \frac{L_{yx}}{2}\mathbb{E}\big\|y^{(t+1)} - \widetilde{y}^*(x^{(t)})\big\|^2 \cdot \mathbb{E}\big\|x^{(t+1)} - x^{(t)}\big\|^2 + \frac{L_{yx}}{2}\mathbb{E}\big\|x^{(t+1)} - x^{(t)}\big\|^2 \\
&\leq \delta_{t,1}\mathbb{E}\big\|y^{(t+1)} - \widetilde{y}^*(x^{(t)})\big\|^2 + \frac{(\rho^{(t)}\gamma_x)^2 L_y^2}{\delta_{t,1}}\mathbb{E}\Big\|\sum_{i=1}^{n} w_i\widetilde{h}_{x,i}^{(t)}\Big\|^2 \\
&\quad + L_{yx}(r_{\max}^2 L_1^2 + L_f^2)(\rho^{(t)}\gamma_x)^2\mathbb{E}\big\|y^{(t+1)} - \widetilde{y}^*(x^{(t)})\big\|^2 + \frac{L_{yx}(\rho^{(t)}\gamma_x)^2}{2}\mathbb{E}\Big\|\sum_{i\in C^{(t)}}^{n} \widetilde{w}_i h_{x,i}^{(t)}\Big\|^2 \\
&= \delta_t\mathbb{E}\big\|y^{(t+1)} - \widetilde{y}^*(x^{(t)})\big\|^2 + \frac{(\rho^{(t)}\gamma_x)^2 L_y^2}{\delta_{t,1}}\mathbb{E}\Big\|\sum_{i=1}^{n} w_i\widetilde{h}_{x,i}^{(t)}\Big\|^2 + \frac{L_{yx}(\rho^{(t)}\gamma_x)^2}{2}\mathbb{E}\Big\|\sum_{i\in C^{(t)}}^{n} \widetilde{w}_i h_{x,i}^{(t)}\Big\|^2
\end{aligned}
\tag{30}
$$

where (a) follows from lemma 4; (b) define $\delta_t := \delta_{t,1} + L_{yx}(r_{\max}^2 L_1^2 + L_f^2)(\rho^{(t)}\gamma_x)^2/2$. By incorporating eq. (30) into eq. (29), we have

$$
\begin{aligned}
\mathbb{E}\big\|y^{(t+1)} - \widetilde{y}^*(x^{(t+1)})\big\|^2 \leq{}& (1+\delta_t)\mathbb{E}\big\|y^{(t+1)} - \widetilde{y}^*(x^{(t)})\big\|^2 + \mathbb{E}\big\|\widetilde{y}^*(x^{(t)}) - \widetilde{y}^*(x^{(t+1)})\big\|^2 \\
&+ (\rho^{(t)}\gamma_x)^2\frac{L_y^2}{\delta_{t,1}}\mathbb{E}\Big\|\sum_{i=1}^{n} w_i\widetilde{h}_{x,i}^{(t)}\Big\|^2 + (\rho^{(t)}\gamma_x)^2\frac{L_{yx}}{2}\mathbb{E}\Big\|\sum_{i\in C^{(t)}}^{n} \widetilde{w}_i h_{x,i}^{(t)}\Big\|^2.
\end{aligned}
\tag{31}
$$

Similarly, we have

$$
\begin{aligned}
\mathbb{E}\big\|v^{(t+1)} - \widetilde{v}^*(x^{(t+1)})\big\|^2 \leq{}& (1+\delta_t')\mathbb{E}\big\|v^{(t+1)} - \widetilde{v}^*(x^{(t)})\big\|^2 + \mathbb{E}\big\|\widetilde{v}^*(x^{(t)}) - \widetilde{v}^*(x^{(t+1)})\big\|^2 \\
&+ (\rho^{(t)}\gamma_x)^2\frac{L_v^2}{\delta_{t,1}'}\mathbb{E}\Big\|\sum_{i=1}^{n} w_i\widetilde{h}_{x,i}^{(t)}\Big\|^2 + (\rho^{(t)}\gamma_x)^2\frac{L_{vx}}{2}\mathbb{E}\Big\|\sum_{i\in C^{(t)}}^{n} \widetilde{w}_i h_{x,i}^{(t)}\Big\|^2,
\end{aligned}
\tag{32}
$$

where $\delta_t' := \delta_{t,1}' + L_{vx}(r_{\max}^2 L_1^2 + L_f^2)(\rho^{(t)}\gamma_x)^2/2$. For the first part in eq. (29), we have

$$
\begin{aligned}
&\mathbb{E}\big\|y^{(t+1)} - \widetilde{y}^*(x^{(t)})\big\|^2 \\
&\quad = \mathbb{E}\Big\|y^{(t)} - \widetilde{y}^*(x^{(t)}) - \rho^{(t)}\gamma_y\sum_{i\in C(t)} \widetilde{w}_i h_{y,i}^{(t)}\Big\|^2 \\
&\quad = \mathbb{E}\Big\|y^{(t)} - \widetilde{y}^*(x^{(t)})\Big\|^2 + \big(\rho^{(t)}\gamma_y\big)^2\mathbb{E}\Big\|\sum_{i\in C(t)} \widetilde{w}_i h_{y,i}^{(t)}\Big\|^2 \\
&\qquad - 2\rho^{(t)}\gamma_y\mathbb{E}\Big\langle y^{(t)} - \widetilde{y}^*(x^{(t)}), \sum_{i\in C(t)} \widetilde{w}_i h_{y,i}^{(t)}\Big\rangle.
\end{aligned}
\tag{33}
$$

For the last term in eq. (33), we have

$$\begin{aligned}
&-\mathbb{E}\Big\langle y^{(t)}-\widetilde{y}^*(x^{(t)}),\sum_{i\in C(t)}\widetilde{w}_i h_{y,i}^{(t)}\Big\rangle\\
&=-\mathbb{E}\Big\langle y^{(t)}-\widetilde{y}^*(x^{(t)}),\sum_{i=1}^{n} w_i\widetilde{h}_{y,i}^{(t)}-\nabla_y\widetilde{G}(x^{(t)},y^{(t)})\\
&\qquad\qquad+\nabla_y\widetilde{G}(x^{(t)},y^{(t)})-\nabla_y\widetilde{G}\big(x^{(t)},\widetilde{y}^*(x^{(t)})\big)\Big\rangle\\
&=-\sum_{i=1}^{n}w_i\sum_{k=0}^{\tau_i-1}\frac{a_i^{(t,k)}}{\|a_i^{(t)}\|_1}\mathbb{E}\Big\langle y^{(t)}-\widetilde{y}^*(x^{(t)}),\nabla_y g_i\big(x_i^{(t,k)},y_i^{(t,k)}\big)-\nabla_y g_i\big(x^{(t)},y^{(t)}\big)\Big\rangle\\
&\quad-\sum_{i=1}^{n}w_i\sum_{k=0}^{\tau_i-1}\frac{a_i^{(t,k)}}{\|a_i^{(t)}\|_1}\mathbb{E}\Big\langle y^{(t)}-\widetilde{y}^*(x^{(t)}),\nabla_y g_i\big(x^{(t)},y^{(t)}\big)-\nabla_y g_i\big(x^{(t)},y^*(x^{(t)})\big)\Big\rangle\\
&\overset{(a)}{\le}\sum_{i=1}^{n}w_i\sum_{k=0}^{\tau_i-1}\frac{a_i^{(t,k)}}{\|a_i^{(t)}\|_1}\mathbb{E}\Big[\frac{1}{\mu_g}\Big\|\nabla_y g_i\big(x_i^{(t,k)},y_i^{(t,k)}\big)-\nabla_y g_i\big(x^{(t)},y^{(t)}\big)\Big\|^2\\
&\qquad\qquad+\frac{\mu_g}{2}\big\|y^{(t)}-\widetilde{y}^*(x^{(t)})\big\|^2\Big]-\sum_{i=1}^{n}w_i\sum_{k=0}^{\tau_i-1}\frac{a_i^{(t,k)}}{\|a_i^{(t)}\|_1}\mu_g\mathbb{E}\big\|y^{(t)}-\widetilde{y}^*(x^{(t)})\big\|^2\\
&\overset{(b)}{\le}\frac{L_1^2}{\mu_g}\sum_{i=1}^{n}w_i\sum_{k=0}^{\tau_i-1}\frac{a_i^{(t,k)}}{\|a_i^{(t)}\|_1}\mathbb{E}\Big[\big\|x^{(t)}-x_i^{(t,k)}\big\|^2+\big\|y^{(t)}-y_i^{(t,k)}\big\|^2\Big]-\frac{\mu_g}{2}\mathbb{E}\big\|y^{(t)}-\widetilde{y}^*(x^{(t)})\big\|^2,
\end{aligned}\tag{34}$$

where (a) follows from the strong convexity of g_i; (b) uses Assumption 2. Incorporate eq. (34) into eq. (33) and we have

$$\begin{aligned}
\mathbb{E}\big\|y^{(t+1)}-\widetilde{y}^*(x^{(t)})\big\|^2=&\Big(1-\rho^{(t)}\gamma_y\mu_g\Big)\mathbb{E}\Big\|y^{(t)}-\widetilde{y}^*(x^{(t)})\Big\|^2+\big(\rho^{(t)}\gamma_y\big)^2\mathbb{E}\Big\|\sum_{i\in C(t)}\widetilde{w}_i h_{y,i}^{(t)}\Big\|^2\\
&+2\rho^{(t)}\gamma_y\frac{L_1^2}{\mu_g}\sum_{i=1}^{n}w_i\sum_{k=0}^{\tau_i-1}\frac{a_i^{(t,k)}}{\|a_i^{(t)}\|_1}\mathbb{E}\Big[\big\|x^{(t)}-x_i^{(t,k)}\big\|^2+\big\|y^{(t)}-y_i^{(t,k)}\big\|^2\Big].
\end{aligned}\tag{35}$$

For the first part in 29, using Lemma 4, we have

$$\mathbb{E}\big\|\widetilde{y}^*(x^{(t)})-\widetilde{y}^*(x^{(t+1)})\big\|^2\le L_y^2\big\|x^{(t)}-x^{(t+1)}\big\|^2=L_y^2\big(\rho^{(t)}\gamma_x\big)^2\mathbb{E}\Big\|\sum_{i\in C^{(t)}}\widetilde{w}_i h_{x,i}^{(t)}\Big\|^2.\tag{36}$$

By incorporating eq. (35) and eq. (36) into eq. (29), we have

$$\begin{aligned}
&\mathbb{E}\|y^{(t+1)}-\widetilde{y}^*(x^{(t+1)})\|^2-\mathbb{E}\|y^{(t)}-\widetilde{y}^*(x^{(t)})\|^2\\
&\le(\delta_t-\rho^{(t)}\gamma_y\mu_g-\delta_t\rho^{(t)}\gamma_y\mu_g)\mathbb{E}\|y^{(t)}-\widetilde{y}^*(x^{(t)})\|^2+(1+\delta_t)(\rho^{(t)}\gamma_y)^2\mathbb{E}\Big\|\sum_{i\in C^{(t)}}\widetilde{w}_i h_{y,i}^{(t)}\Big\|^2\\
&\quad+(1+\delta_t)\rho^{(t)}\gamma_y\frac{2L_1^2}{\mu_g}\sum_{i=1}^{n}w_i\sum_{k=0}^{\tau_i-1}\frac{a_i^{(t,k)}}{\|a_i^{(t)}\|_1}\mathbb{E}\Big[\big\|x^{(t)}-x_i^{(t,k)}\big\|^2+\big\|y^{(t)}-y_i^{(t,k)}\big\|^2\Big]\\
&\quad+\big(\rho^{(t)}\gamma_x\big)^2\Big(L_y^2+\frac{L_{yx}}{2}\Big)\mathbb{E}\Big\|\sum_{i\in C^{(t)}}\widetilde{w}_i h_{x,i}^{(t)}\Big\|^2+(\rho^{(t)}\gamma_x)^2\frac{L_y^2}{\delta_{t,1}}\mathbb{E}\Big\|\sum_{i=1}^{n}w_i\widetilde{h}_{x,i}^{(t)}\Big\|^2.
\end{aligned}$$

Following similar steps of eq. (34) by replacing $h_{y,i}^{(t)}$ and $\nabla_y g_i$ with $h_{v,i}^{(t)}$ and $\nabla_v R_i$, we can easily have

$$-\mathbb{E}\Big\langle v^{(t)}-\widetilde{v}^*(x^{(t)}),\sum_{i\in C(t)}\widetilde{w}_i h_{v,i}^{(t)}\Big\rangle$$

$$
\begin{aligned}
&= -\sum_{i=1}^{n} w_i \sum_{k=0}^{\tau_i-1} \frac{a_i^{(t,k)}}{\|a_i^{(t)}\|_1} \mathbb{E}\Big\langle v^{(t)} - \widetilde{v}^*(x^{(t)}), \nabla_v R_i\big(x_i^{(t,k)}, y_i^{(t,k)}, v_i^{(t,k)}\big) - \nabla_v R_i\big(x^{(t)}, y^{(t)}, v^{(t)}\big)\Big\rangle \\
&\quad -\sum_{i=1}^{n} w_i \sum_{k=0}^{\tau_i-1} \frac{a_i^{(t,k)}}{\|a_i^{(t)}\|_1} \mathbb{E}\Big\langle v^{(t)} - \widetilde{v}^*(x^{(t)}), \nabla_v R_i\big(x^{(t)}, y^{(t)}, v^{(t)}\big) - \nabla_v R_i\big(x^{(t)}, y^*(x^{(t)}), v^{(t)}\big)\Big\rangle \\
&\quad -\sum_{i=1}^{n} w_i \sum_{k=0}^{\tau_i-1} \frac{a_i^{(t,k)}}{\|a_i^{(t)}\|_1} \mathbb{E}\Big\langle v^{(t)} - \widetilde{v}^*(x^{(t)}), \nabla_v R_i\big(x^{(t)}, y^*(x^{(t)}), v^{(t)}\big) \\
&\qquad\qquad\qquad\qquad\qquad - \nabla_v R_i\big(x^{(t)}, y^*(x^{(t)}), v^*(x^{(t)})\big)\Big\rangle \\
&\overset{(a)}{\leq} \frac{2}{\mu_g} \sum_{i=1}^{n} w_i \sum_{k=0}^{\tau_i-1} \frac{a_i^{(t,k)}}{\|a_i^{(t)}\|_1} \mathbb{E}\Big\|\nabla_v R_i\big(x_i^{(t,k)}, y_i^{(t,k)}, v_i^{(t,k)}\big) - \nabla_v R_i\big(x^{(t)}, y^{(t)}, v^{(t)}\big)\Big\|^2 \\
&\quad + \frac{2}{\mu_g} \sum_{i=1}^{n} w_i \sum_{k=0}^{\tau_i-1} \frac{a_i^{(t,k)}}{\|a_i^{(t)}\|_1} \mathbb{E}\Big\|\nabla_v R_i\big(x^{(t)}, y^{(t)}, v^{(t)}\big) - \nabla_v R_i\big(x^{(t)}, y^*(x^{(t)}), v^{(t)}\big)\Big\|^2 \\
&\quad - \frac{\mu_g}{2} \sum_{i=1}^{n} w_i \sum_{k=0}^{\tau_i-1} \frac{a_i^{(t,k)}}{\|a_i^{(t)}\|_1} \mathbb{E}\big\|v^{(t)} - \widetilde{v}^*(x^{(t)})\big\|^2 \\
&\overset{(b)}{\leq} \frac{2L_R^2}{\mu_g} \sum_{i=1}^{n} w_i \sum_{k=0}^{\tau_i-1} \frac{a_i^{(t,k)}}{\|a_i^{(t)}\|_1} \mathbb{E}\Big[\big\|x^{(t)} - x_i^{(t,k)}\big\|^2 + \big\|y^{(t)} - y_i^{(t,k)}\big\|^2 + \big\|v^{(t)} - v_i^{(t,k)}\big\|^2\Big] \\
&\quad + \frac{2L_R^2}{\mu_g} \mathbb{E}\big\|y^{(t)} - \widetilde{y}^*(x^{(t)})\big\|^2 - \frac{\mu_g}{2} \mathbb{E}\big\|v^{(t)} - \widetilde{v}^*(x^{(t)})\big\|^2,
\end{aligned}
$$

where (a) uses the strong convexity of R_i; (b) follows from Lemma 3. Since the projection is non-expansive, we can easily have that

$$
\begin{aligned}
\mathbb{E}\|v^{(t+1)} - v^*(x^{(t)})\|^2 &= \mathbb{E}\|\mathcal{P}_r(v^{(t)} - \rho^{(t)}\gamma_v \sum_{i \in C^{(t)}} \widetilde{w}_i h_{v,i}^{(t)}) - v^*(x^{(t)})\|^2 \\
&\leq \mathbb{E}||v^{(t)} - v^*(x^{(t)}) - \rho^{(t)}\gamma_v \sum_{i \in C^{(t)}} \widetilde{w}_i h_{v,i}^{(t)}||^2
\end{aligned}
$$

Thus, we can have the similar result

$$
\begin{aligned}
&\mathbb{E}\|v^{(t+1)} - \widetilde{v}^*(x^{(t+1)})\|^2 - \mathbb{E}\|v^{(t)} - \widetilde{v}^*(x^{(t)})\|^2 \\
&\quad \leq (\delta_t' - \rho^{(t)}\gamma_v\mu_g - \delta_t'\rho^{(t)}\gamma_v\mu_g)\mathbb{E}\|v^{(t)} - \widetilde{v}^*(x^{(t)})\|^2 + (1+\delta_t')(\rho^{(t)}\gamma_v)^2 \mathbb{E}\Big\|\sum_{i \in C^{(t)}} \widetilde{w}_i h_{v,i}^{(t)}\Big\|^2 \\
&\qquad + (1+\delta_t')\rho^{(t)}\gamma_v \frac{4L_R^2}{\mu_g} \sum_{i=1}^{n} w_i \sum_{k=0}^{\tau_i-1} \frac{a_i^{(t,k)}}{\|a_i^{(t)}\|_1} \\
&\qquad\quad \cdot \mathbb{E}\Big[\big\|x^{(t)} - x_i^{(t,k)}\big\|^2 + \big\|y^{(t)} - y_i^{(t,k)}\big\|^2 + \big\|v^{(t)} - v_i^{(t,k)}\big\|^2\Big] \\
&\qquad + (1+\delta_t')\rho^{(t)}\gamma_v \frac{4L_R^2}{\mu_g} \mathbb{E}\big\|y^{(t)} - \widetilde{y}^*(x^{(t)})\big\|^2 + \big(\rho^{(t)}\gamma_x\big)^2 \Big(L_v^2 + \frac{L_{vx}}{2}\Big) \mathbb{E}\Big\|\sum_{i \in C^{(t)}} \widetilde{w}_i h_{x,i}^{(t)}\Big\|^2 \\
&\qquad + (\rho^{(t)}\gamma_x)^2 \frac{L_v^2}{\delta_{t,1}'} \mathbb{E}\Big\|\sum_{i=1}^{n} w_i \widetilde{h}_{x,i}^{(t)}\Big\|^2.
\end{aligned}
$$

Then, the proof is complete. □

D.5 Descent in the Lyapunov Function

We define the Lyapunov function as

$$
\Psi(x^{(t)}) := \mathbb{E}\Big[\widetilde{\Phi}(x^{(t)})\Big] + K_1\mathbb{E}\|y^{(t)} - \widetilde{y}^*(x^{(t)})\|^2 + K_2\mathbb{E}\|v^{(t)} - \widetilde{v}^*(x^{(t)})\|^2, \tag{37}
$$

where the coefficients are given by

$$K_1 = \left[\frac{40(L_1^2 + r^2 L_2^2)}{\mu_g} + \frac{384 L_R^2 L_1^2}{\mu_g^3}\right]\frac{1+\beta_{gh}^2}{c_{\gamma_y}}, \;\; K_2 = \frac{6L_1^2}{\mu_g c_{\gamma_v}}; \;\; \delta_t = \frac{\rho^{(t)}\gamma_y \mu_g}{4}, \;\; \delta_t' = \frac{\rho^{(t)}\gamma_v \mu_g}{4}.$$

For server and local stepsizes, we choose

$$\begin{gathered}\gamma_x = \mathcal{O}\Big(\sqrt{\frac{P}{\bar{\tau}T}}\Big), \quad \gamma_y = c_{\gamma_y}\gamma_x, \quad \gamma_v = c_{\gamma_v}\gamma_x, \\ \eta_x = \mathcal{O}\Big(\frac{1}{\bar{\tau}\sqrt{T}}\Big),\; \eta_y = \mathcal{O}\Big(\frac{1}{\bar{\tau}\sqrt{T}}\Big),\; \eta_v = \mathcal{O}\Big(\frac{1}{\bar{\tau}\sqrt{T}}\Big). \end{gathered} \tag{38}$$

where $c_{\gamma_y} \geq \frac{256 K_1 L_y}{\mu_g}$ and $c_{\gamma_v} \geq \frac{256 K_2 L_v}{\mu_g}$. We also set

$$\begin{aligned}\rho^{(t)}\gamma_x \leq \min\Bigg\{ & \frac{\mu_g}{12 L_1^2 c_{\gamma_v}}, \frac{\mu_g}{36 L_1^2 c_{\gamma_v}}\Big(\frac{2\beta_{\max}}{P} + 3\Big)^{-1}, \frac{P\mu_g c_{\gamma_v}}{36 L_1^2 \beta_{\max}}\Big(L_v^2 + \frac{L_{vx}}{2}\Big)^{-1}, \frac{P}{6 L_\Phi \beta_{\max}}, \frac{4}{L_\Phi}, \\ & \frac{4}{c_{\gamma_y}}, \frac{4}{c_{\gamma_v}}, \frac{\mu_g}{12 c_{\gamma_y}}\Big[\big(\frac{2\beta_{\max}}{P} + 3\big)L_1^2 + \frac{2\beta_{\max}\beta_{gh}^2}{P} + 3L_1^2\Big]^{-1}, \frac{c_{\gamma_y}\mu_g}{4 L_{yx}(L_f^2 + r_{\max}^2 L_1^2)}, \\ & \frac{c_{\gamma_v}\mu_g}{4 L_{vx}(L_f^2 + r_{\max}^2 L_1^2)}, \frac{1}{8}\Big(K_1(L_y^2 + \frac{L_{yx}}{2}) + K_2(L_v^2 + \frac{L_{vx}}{2})\Big)^{-1}\Bigg\}; \\ & \eta_x^2\bar{\tau} \leq \frac{1}{2 c_a \alpha_{\max} L_1^2}, \quad \eta_y^2 \bar{\tau} \leq \min\Big\{\frac{1}{4 c_a \alpha_{\max} L_1^2}, \frac{\mu_g^2}{96 c_a \alpha_{\max} L_1^4}\Big\}. \end{aligned} \tag{39}$$

To simplify our proof, we define constants:

$$\begin{aligned} \bar{\tau} &:= \frac{1}{n}\sum_{i=1}^{n}\tau_i, \quad \bar{\rho} := \frac{1}{T}\sum_{t=0}^{T-1}\rho^{(t)}, \\ M_1 &:= 2\big[\frac{L_\Phi}{2} + K_1\big(L_y^2 + \frac{L_{yx}}{2}\big) + K_2\big(L_v^2 + \frac{L_{vx}}{2}\big) + 2K_2 c_{\gamma_v}^2\big]\beta_{\max}(L_f^2 + r_{\max}^2 L_1^2) \\ &\quad + K_1 c_{\gamma_y}^2 \beta_{\max}\sigma_{gh}^2 \\ M_2 &:= 2\big[\frac{L_\Phi}{2} + K_1\big(L_y^2 + \frac{L_{yx}}{2}\big) + K_2\big(L_v^2 + \frac{L_{vx}}{2}\big) + 2K_2 c_{\gamma_v}^2\big](\sigma_f^2 + r_{\max}^2 \sigma_{gg}^2) + K_1 c_{\gamma_y}^2 \sigma_g^2, \\ M_3 &:= 3\left[\frac{3}{2} + 24 K_2 c_{\gamma_v} + \frac{\beta_{\max}}{P}\Big(\frac{17}{4} + 16 K_2 c_{\gamma_v}\Big)\right](L_1^2 + r^2 L_2^2) \\ &\quad + K_1\left[4 c_{\gamma_y}\frac{L_1^2}{\mu_g} + 8 c_{\gamma_v}\Big(\frac{2\beta_{\max}}{P} + 3\Big)L_1^2\right] + K_2 c_{\gamma_v}\frac{4L_R^2}{\mu_g}. \end{aligned} \tag{40}$$

We apply Lemma 7 and Lemma 10 to eq. (37), and incorporate Lemma 9, then we have

$$\begin{aligned} &\Psi(x^{(t+1)}) - \Psi(x^{(t)}) \\ &= \mathbb{E}\Big[\widetilde{\Phi}(x^{(t+1)}) - \widetilde{\Phi}(x^{(t)})\Big] + K_1\mathbb{E}\Big[\|y^{(t+1)} - \widetilde{y}^*(x^{(t+1)})\|^2 - \|y^{(t)} - \widetilde{y}^*(x^{(t)})\|^2\Big] \\ &\quad + K_2\mathbb{E}\Big[\|v^{(t+1)} - \widetilde{v}^*(x^{(t+1)})\|^2 - \|v^{(t)} - \widetilde{v}^*(x^{(t)})\|^2\Big] \\ &\overset{(a)}{=} -\frac{\rho^{(t)}\gamma_x}{2}\mathbb{E}\Big\|\nabla\widetilde{\Phi}(x^{(t)})\Big\|^2 \\ &\quad + 3(\rho^{(t)}\gamma_x)\left[\frac{3}{2} + 24K_2 c_{\gamma_v} + \frac{\beta_{\max}}{P}\Big(\frac{17}{4} + 16 K_2 c_{\gamma_v}\Big)\right](L_1^2 + r^2 L_2^2) \\ &\qquad \cdot \Big((\eta_x^2 + \eta_v^2)\bar{\tau}\sigma_{M1}^2 + 2\eta_y^2\bar{\tau}\sigma_{M2}^2\Big) \\ &\quad + \rho^{(t)}\gamma_x K_1\left[4c_{\gamma_y}\frac{L_1^2}{\mu_g} + 8c_{\gamma_v}\Big(\frac{2\beta_{\max}}{P} + 3\Big)L_1^2\right]\Big(\eta_x^2\bar{\tau}\sigma_{M1}^2 + 2\eta_y^2\bar{\tau}\sigma_{M2}^2\Big) \\ &\quad + \rho^{(t)}\gamma_x K_2 c_{\gamma_v}\frac{4L_R^2}{\mu_g}\Big((\eta_x^2 + \eta_v^2)\bar{\tau}\sigma_{M1}^2 + 2\eta_y^2\bar{\tau}\sigma_{M2}^2\Big) \end{aligned}$$

$$
\begin{aligned}
&+ (\rho^{(t)}\gamma_x)^2\Big[\frac{L_\Phi}{2} + K_1\big(L_y^2 + \frac{L_{yx}}{2}\big) + K_2\big(L_v^2 + \frac{L_{vx}}{2}\big) + 2K_2 c_{\gamma_v}^2\Big]\frac{2n}{P} \\
&\quad\cdot \sum_{i=1}^{n}\frac{w_i^2\|a_i^{(t)}\|_2^2}{\|a_i^{(t)}\|_1^2}(\sigma_f^2 + r_i^2\sigma_{gg}^2) \\
&+ (\rho^{(t)}\gamma_x)^2 K_1 c_{\gamma_y}^2\frac{n}{P}\sum_{i=1}^{n}\frac{w_i^2\|a_i^{(t)}\|_2^2}{\|a_i^{(t)}\|_1^2}\sigma_g^2 \\
&+ (\rho^{(t)}\gamma_x)^2\Big[\frac{L_\Phi}{2} + K_1\big(L_y^2 + \frac{L_{yx}}{2}\big) + K_2\big(L_v^2 + \frac{L_{vx}}{2}\big) + 2K_2 c_{\gamma_v}^2\Big]\frac{4(n-P)\beta_{\max}}{P(n-1)}(L_f^2 + r_{\max}^2 L_1^2) \\
&+ (\rho^{(t)}\gamma_x)^2 K_1 c_{\gamma_y}^2\frac{2(n-P)\beta_{\max}}{P(n-1)}\sigma_{gh}^2
\end{aligned} \tag{41}
$$

where (a) holds when we set K_1, K_2, δ_t and δ_t' as eq. (40). We rearrange eq. (41) and separate it into three error terms. Here, we define the error from full synchronization as

$$
\begin{aligned}
\epsilon_{sync}^{(t)} := (\rho^{(t)}\gamma_x)^2&\Big[\frac{L_\Phi}{2} + K_1\big(L_y^2 + \frac{L_{yx}}{2}\big) + K_2\big(L_v^2 + \frac{L_{vx}}{2}\big) + 2K_2 c_{\gamma_v}^2\Big]\frac{2n}{P} \\
&\cdot \sum_{i=1}^{n}\frac{w_i^2\|a_i^{(t)}\|_2^2}{\|a_i^{(t)}\|_1^2}(\sigma_f^2 + r_i^2\sigma_{gg}^2) + (\rho^{(t)}\gamma_x)^2 K_1 c_{\gamma_y}^2\frac{n}{P}\sum_{i=1}^{n}\frac{w_i^2\|a_i^{(t)}\|_2^2}{\|a_i^{(t)}\|_1^2}\sigma_g^2
\end{aligned} \tag{42}
$$

and the error from partial participation as

$$
\begin{aligned}
\epsilon_{part}^{(t)} := (\rho^{(t)}\gamma_x)^2&\Big[\frac{L_\Phi}{2} + K_1\big(L_y^2 + \frac{L_{yx}}{2}\big) + K_2\big(L_v^2 + \frac{L_{vx}}{2}\big) + 2K_2 c_{\gamma_v}^2\Big] \\
&\cdot \frac{4(n-P)\beta_{\max}}{P(n-1)}(L_f^2 + r_{\max}^2 L_1^2) + (\rho^{(t)}\gamma_x)^2 K_1 c_{\gamma_y}^2\frac{2(n-P)\beta_{\max}}{P(n-1)}\sigma_{gh}^2
\end{aligned} \tag{43}
$$

Next, we define the error due to client drifts as

$$
\begin{aligned}
\epsilon_{cd}^{(t)} :=\ & 3(\rho^{(t)}\gamma_x)\Big[\frac{3}{2} + 24K_2 c_{\gamma_v} + \frac{\beta_{\max}}{P}\Big(\frac{17}{4} + 16K_2 c_{\gamma_v}\Big)\Big](L_1^2 + r^2 L_2^2) \\
&\cdot \Big((\eta_x^2 + \eta_v^2)\bar{\tau}\sigma_{M1}^2 + 2\eta_y^2\bar{\tau}\sigma_{M2}^2\Big) + \rho^{(t)}\gamma_x K_2 c_{\gamma_v}\frac{4L_R^2}{\mu_g}\Big((\eta_x^2 + \eta_v^2)\bar{\tau}\sigma_{M1}^2 + 2\eta_y^2\bar{\tau}\sigma_{M2}^2\Big) \\
&+ \rho^{(t)}\gamma_x K_1\Big[4c_{\gamma_y}\frac{L_1^2}{\mu_g} + 8c_{\gamma_v}\Big(\frac{2\beta_{\max}}{P} + 3\Big)L_1^2\Big]\Big(\eta_x^2\bar{\tau}\sigma_{M1}^2 + 2\eta_y^2\bar{\tau}\sigma_{M2}^2\Big).
\end{aligned} \tag{44}
$$

Then we have the Descent in the Lyapunov function as

$$
\Psi(x^{(t+1)}) - \Psi(x^{(t)}) \leq -\frac{\rho^{(t)}\gamma_x}{2}\mathbb{E}\Big\|\nabla\widetilde{\Phi}(x^{(t)})\Big\|^2 + \epsilon_{part}^{(t)} + \epsilon_{sync}^{(t)} + \epsilon_{cd}^{(t)}. \tag{45}
$$

D.6 Proof of Theorem 1

Proof. Summing eq. (45), we have

$$
\begin{aligned}
\min_t \mathbb{E}\Big\|\nabla\widetilde{\Phi}(x^{(t)})\Big\|^2 &\leq \frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}\Big\|\nabla\widetilde{\Phi}(x^{(t)})\Big\|^2 \\
&\overset{(a)}{\leq} 2 \times \frac{1}{T}\sum_{t=0}^{T-1}\frac{\rho^{(t)}}{\bar{\rho}}\mathbb{E}\Big\|\nabla\widetilde{\Phi}(x^{(t)})\Big\|^2 \\
&\leq 2 \times \frac{2}{T}\Big(\frac{\Psi(x^{(0)})}{\bar{\rho}\gamma_x} - \frac{\Psi(x^{(T)})}{\bar{\rho}\gamma_x}\Big) + 2 \times \frac{1}{T}\sum_{t=0}^{T-1}\frac{2}{\bar{\rho}\gamma_x}\Big(\epsilon_{part}^{(t)} + \epsilon_{sync}^{(t)} + \epsilon_{cd}^{(t)}\Big),
\end{aligned} \tag{46}
$$

where (a) uses $\rho^{(t)} \in [\frac{1}{2}\bar{\rho}, \frac{3}{2}\bar{\rho}]$. For the error with partial participation in eq. (42), we have

$$\frac{1}{T}\sum_{t=0}^{T-1}\frac{2}{\bar{\rho}\gamma_x}\epsilon_{part}^{(t)} \overset{(a)}{\leq} \frac{1}{T}\sum_{t=0}^{T-1}\frac{4}{\bar{\rho}\gamma_x}(\rho^{(t)}\gamma_x)^2\frac{(n-P)}{P(n-1)}M_1 \overset{(b)}{\leq} 6\bar{\rho}\gamma_x\frac{(n-P)}{P(n-1)}M_1 \tag{47}$$

where (a) simplifies the problem by defining

$$\begin{aligned}M_1 :=& 2\big[\frac{L_\Phi}{2} + K_1\big(L_y^2 + \frac{L_{yx}}{2}\big) + K_2\big(L_v^2 + \frac{L_{vx}}{2}\big) + 2K_2c_{\gamma_v}^2\big]\beta_{\max}(L_f^2 + r_{\max}^2L_1^2)\\ &+ K_1c_{\gamma_y}^2\beta_{\max}\sigma_{gh}^2\end{aligned}$$

and (b) holds due to $\rho^{(t)} \in [\frac{1}{2}\bar{\rho}, \frac{3}{2}\bar{\rho}]$. By the definition of $\rho^{(t)}$ in eq. (6), we can easily see that $\bar{\rho} = \mathcal{O}(\bar{\tau})$. Then we have

$$\frac{1}{T}\sum_{t=0}^{T-1}\frac{2}{\bar{\rho}\gamma_x}\epsilon_{part}^{(t)} \leq 6M_1\frac{n-P}{P(n-1)}\bar{\rho}\gamma_x = \mathcal{O}\Big(\frac{M_1(n-P)}{n}\sqrt{\frac{\bar{\tau}}{PT}}\Big) \tag{48}$$

by taking $\gamma_x = \mathcal{O}\Big(\sqrt{\frac{P}{\bar{\tau}T}}\Big)$. Similarly, for the error with full synchronization in eq. (43), we have

$$\begin{aligned}\frac{1}{T}\sum_{t=0}^{T-1}\frac{2}{\bar{\rho}\gamma_x}\epsilon_{sync}^{(t)} &\overset{(a)}{\leq} \frac{1}{T}\sum_{t=0}^{T-1}\frac{2}{\bar{\rho}\gamma_x}(\rho^{(t)}\gamma_x)^2\frac{n}{P}\sum_{i=1}^{n}\frac{w_i^2\|a_i^{(t)}\|_2^2}{\|a_i^{(t)}\|_1^2}M_2\\ &\overset{(b)}{\leq} \frac{3}{T}\sum_{t=0}^{T-1}\rho^{(t)}\gamma_x\frac{\beta_{\max}}{P}\sum_{i=1}^{n}w_i\frac{\alpha_{\max}}{c'_a\bar{\tau}\alpha_{\min}}M_2\\ &\leq \frac{3\alpha_{\max}\beta_{\max}}{c'_a\alpha_{\min}}\frac{\bar{\rho}}{P\bar{\tau}}\gamma_xM_2\end{aligned} \tag{49}$$

where (a) simplifies the problem by defining

$$M_2 := 2\big[\frac{L_\Phi}{2} + K_1\big(L_y^2 + \frac{L_{yx}}{2}\big) + K_2\big(L_v^2 + \frac{L_{vx}}{2F}\big) + 2K_2c_{\gamma_v}^2\big](\sigma_f^2 + r_{\max}^2\sigma_{gg}^2) + K_1c_{\gamma_y}^2\sigma_g^2,$$

and (b) holds since $c'_a\bar{\tau}\alpha_{\min} \leq \|a_i^{(t)}\|_1 \leq c_a\bar{\tau}\alpha_{\max}$. Then we have

$$\frac{1}{T}\sum_{t=0}^{T-1}\frac{2}{\bar{\rho}\gamma_x}\epsilon_{sync}^{(t)} \leq \frac{3\alpha_{\max}\beta_{\max}}{c'_a\alpha_{\min}}\frac{\bar{\rho}}{P\bar{\tau}}\gamma_xM_2 = \mathcal{O}\Big(\frac{M_2}{\sqrt{P\bar{\tau}T}}\Big) \tag{50}$$

by taking $\gamma_x = \mathcal{O}\Big(\sqrt{\frac{P}{\bar{\tau}T}}\Big)$. Similarly, for the error due to client drifts in eq. (44), we have

$$\begin{aligned}\frac{1}{T}\sum_{t=0}^{T-1}\frac{2}{\bar{\rho}\gamma_x}\epsilon_{cd}^{(t)} &\leq \frac{1}{T}\sum_{t=0}^{T-1}\frac{2}{\bar{\rho}\gamma_x}\rho^{(t)}\gamma_x\Big((\eta_x^2 + \eta_v^2)\bar{\tau}\sigma_{M1}^2 + 2\eta_y^2\bar{\tau}\sigma_{M2}^2\Big)M_3\\ &\leq 3\Big((\eta_x^2 + \eta_v^2)\bar{\tau}\sigma_{M1}^2 + 2\eta_y^2\bar{\tau}\sigma_{M2}^2\Big)M_3\end{aligned} \tag{51}$$

We define constant M_3 as

$$\begin{aligned}M_3 :=& 3\left[\frac{3}{2} + 24K_2c_{\gamma_v} + \frac{\beta_{\max}}{P}\Big(\frac{17}{4} + 16K_2c_{\gamma_v}\Big)\right](L_1^2 + r^2L_2^2)\\ &+ K_1\left[4c_{\gamma_y}\frac{L_1^2}{\mu_g} + 8c_{\gamma_v}\Big(\frac{2\beta_{\max}}{P} + 3\Big)L_1^2\right] + K_2c_{\gamma_v}\frac{4L_R^2}{\mu_g}\end{aligned}$$

Then we have

$$\frac{1}{T}\sum_{t=0}^{T-1}\frac{2}{\bar{\rho}\gamma_x}\epsilon_{cd}^{(t)} \leq 3\Big((\eta_x^2 + \eta_v^2)\bar{\tau}\sigma_{M1}^2 + 2\eta_y^2\bar{\tau}\sigma_{M2}^2\Big)M_3 = \mathcal{O}\Big(\frac{M_3}{\bar{\tau}T}\Big), \tag{52}$$

by setting $\eta_x = \mathcal{O}\Big(\frac{1}{\bar{\tau}\sqrt{T}}\Big)$, $\eta_y = \mathcal{O}\Big(\frac{1}{\bar{\tau}\sqrt{T}}\Big)$ and $\eta_v = \mathcal{O}\Big(\frac{1}{\bar{\tau}\sqrt{T}}\Big)$. Last but not least, for the first tern on the right-hand side of eq. (46), we have

$$\frac{2}{T}\Big(\frac{\Psi(x^{(0)})}{\bar{\rho}\gamma_x} - \frac{\Psi(x^{(T)})}{\bar{\rho}\gamma_x}\Big) = \mathcal{O}\Big(\sqrt{\frac{1}{P\bar{\tau}T}}\Big) \tag{53}$$

when we take $\gamma_x = \mathcal{O}\Big(\sqrt{\frac{P}{\bar{\tau}T}}\Big)$. Finally, by combining eq. (48), eq. (50), eq. (52) and eq. (53), we have

$$\min_t \mathbb{E}\Big\|\nabla\widetilde{\Phi}(x^{(t)})\Big\|^2 = \mathcal{O}\Big(\frac{M_1(n-P)}{n}\sqrt{\frac{\bar{\tau}}{PT}}\Big) + \mathcal{O}\Big(M_2\sqrt{\frac{1}{P\bar{\tau}T}}\Big) + \mathcal{O}\Big(\frac{M_3}{\bar{\tau}T}\Big). \tag{54}$$

Then, the first part of the proof of Theorem 1 is complete. Next, we provide the detail of complexity analysis. First, for nearly full client participation, which means that $\frac{n-P}{n-1} \approx 0$, we can easily have

$$\min_t \mathbb{E}\Big\|\nabla\widetilde{\Phi}(x^{(t)})\Big\|^2 = \mathcal{O}\Big(M_2\sqrt{\frac{1}{n\bar{\tau}T}}\Big) + \mathcal{O}\Big(\frac{M_3}{\bar{\tau}T}\Big) \leq \epsilon. \tag{55}$$

As a result, we can see that the per-client sample complexity $\bar{\tau}T = \mathcal{O}(n^{-1}\epsilon^{-2})$. Since the local update rounds contribute to saving communication rounds, we take $\bar{\tau} = \mathcal{O}(\frac{T}{n})$, then we have $T = \mathcal{O}(\epsilon^{-1})$. Second, for partial client participation, we have

$$\min_t \mathbb{E}\Big\|\nabla\widetilde{\Phi}(x^{(t)})\Big\|^2 = \mathcal{O}\Big(\frac{M_1(n-P)}{n}\sqrt{\frac{\bar{\tau}}{PT}}\Big) + \mathcal{O}\Big(M_2\sqrt{\frac{1}{P\bar{\tau}T}}\Big) + \mathcal{O}\Big(\frac{M_3}{\bar{\tau}T}\Big) \leq \epsilon. \tag{56}$$

when participating client number P is not close to full client number n, we can find that the local update date rounds will increase the partial participation error, which may affect the whole convergence rate. As a consequence, taking $\bar{\tau} = \mathcal{O}(\frac{n}{n-P})$ will result in the best performance. We can see that

$$\min_t \mathbb{E}\Big\|\nabla\widetilde{\Phi}(x^{(t)})\Big\|^2 = \mathcal{O}\Big((M_1 + M_2)\sqrt{\frac{n-P}{nPT}}\Big) + \mathcal{O}\Big(\frac{M_3}{T}\Big) \leq \epsilon. \tag{57}$$

Since $T \gg P$, we have the per-client sample complexity $\bar{\tau}T = \mathcal{O}(P^{-1}\epsilon^{-2})$ and communication rounds $T = \mathcal{O}(P^{-1}\epsilon^{-2})$. Then, we finish the proof of Theorem 1. □

E Proof of Theorem 2

Proof. Recall the definitions:

$$\Phi(x^{(t)}) = F(x^{(t)}, y^*(x^{(t)})) = \sum_{i=1}^{n} p_i f_i(x^{(t)}, y^*(x^{(t)})),$$
$$\widetilde{\Phi}(x^{(t)}) = \widetilde{F}(x^{(t)}, \widetilde{y}^*(x^{(t)})) = \sum_{i=1}^{n} w_i f_i(x^{(t)}, \widetilde{y}^*(x^{(t)})).$$

Then we have

$$\begin{aligned}
&\nabla\Phi(x^{(t)}) - \nabla\widetilde{\Phi}(x^{(t)}) \\
&= \sum_{i=1}^{n}\Big[p_i\bar{\nabla}f_i\big(x^{(t)}, y^*(x^{(t)}), v^*(x^{(t)})\big) - w_i\bar{\nabla}f_i\big(x^{(t)}, \widetilde{y}^*(x^{(t)}), \widetilde{v}^*(x^{(t)})\big)\Big] \\
&= \sum_{i=1}^{n} p_i\Big[\bar{\nabla}f_i\big(x^{(t)}, y^*(x^{(t)}), v^*(x^{(t)})\big) - \bar{\nabla}f_i\big(x^{(t)}, \widetilde{y}^*(x^{(t)}), \widetilde{v}^*(x^{(t)})\big)\Big] \\
&\quad + \sum_{i=1}^{n}(p_i - w_i)\bar{\nabla}f_i\big(x^{(t)}, \widetilde{y}^*(x^{(t)}), \widetilde{v}^*(x^{(t)})\big) \\
&= \bar{\nabla}F\big(x^{(t)}, y^*(x^{(t)}), v^*(x^{(t)})\big) - \bar{\nabla}F\big(x^{(t)}, \widetilde{y}^*(x^{(t)}), \widetilde{v}^*(x^{(t)})\big)
\end{aligned}$$

$$+\sum_{i=1}^{n}\frac{p_i-w_i}{w_i}w_i\bar{\nabla}f_i\big(x^{(t)},\widetilde{y}^*(x^{(t)}),\widetilde{v}^*(x^{(t)})\big). \tag{58}$$

By taking the norm of eq. (58) and using Assumption 2, we have

$$\begin{aligned}
&\Big\|\nabla\Phi(x^{(t)})-\nabla\widetilde{\Phi}(x^{(t)})\Big\|^2\\
&\overset{(a)}{\leq} 6\big(L_1^2+r^2L_2^2\big)\big\|y^*(x^{(t)})-\widetilde{y}^*(x^{(t)})\big\|^2+6L_1^2\big\|v^*(x^{(t)})-\widetilde{v}^*(x^{(t)})\big\|^2\\
&\quad+2\Big\|\sum_{i=1}^{n}\frac{p_i-w_i}{w_i}w_i\bar{\nabla}f_i\big(x^{(t)},\widetilde{y}^*(x^{(t)}),\widetilde{v}^*(x^{(t)})\big)\Big\|^2\\
&\overset{(b)}{\leq} 6\big(L_1^2+r^2L_2^2\big)\big\|y^*(x^{(t)})-\widetilde{y}^*(x^{(t)})\big\|^2+6L_1^2\big\|v^*(x^{(t)})-\widetilde{v}^*(x^{(t)})\big\|^2\\
&\quad+2\Big\|\frac{\beta'_{\max}-\beta_{\min}}{\beta_{\min}}\sum_{i=1}^{n}w_i\bar{\nabla}f_i\big(x^{(t)},\widetilde{y}^*(x^{(t)}),\widetilde{v}^*(x^{(t)})\big)\Big\|^2\\
&=6\Big(L_1^2+r^2L_2^2\Big)\big\|y^*(x^{(t)})-\widetilde{y}^*(x^{(t)})\big\|^2+6L_1^2\big\|v^*(x^{(t)})-\widetilde{v}^*(x^{(t)})\big\|^2\\
&\quad+2\Big(\frac{\beta'_{\max}-\beta_{\min}}{\beta_{\min}}\Big)^2\Big\|\nabla\widetilde{\Phi}(x)\Big\|^2,
\end{aligned} \tag{59}$$

where (a) uses Assumption 2; (b) uses the setting $\frac{\beta_{\min}}{n}\leq w_i\leq\frac{\beta_{\max}}{n}$ and $\frac{\beta'_{\min}}{n}\leq p_i\leq\frac{\beta'_{\max}}{n}$ for all $i=1,2,...n$. Then we can see that

$$\begin{aligned}
\min_t\Big\|\nabla\Phi(x^{(t)})\Big\|^2&\leq\frac{1}{T}\sum_{t=0}^{T-1}\Big\|\nabla\Phi(x^{(t)})\Big\|^2\\
&\leq\frac{2}{T}\sum_{t=0}^{T-1}\Big(\Big\|\nabla\Phi(x^{(t)})-\nabla\widetilde{\Phi}(x^{(t)})\Big\|^2+\Big\|\nabla\widetilde{\Phi}(x^{(t)})\Big\|^2\Big)\\
&\leq\frac{2}{T}\sum_{t=0}^{T-1}\Big[1+2\Big(\frac{\beta'_{\max}-\beta_{\min}}{\beta_{\min}}\Big)^2\Big]\Big\|\nabla\widetilde{\Phi}(x^{(t)})\Big\|^2\\
&\quad+12\Big(L_1^2+r^2L_2^2\Big)\frac{1}{T}\sum_{t=0}^{T-1}\big\|y^*(x^{(t)})-\widetilde{y}^*(x^{(t)})\big\|^2\\
&\quad+12L_1^2\frac{1}{T}\sum_{t=0}^{T-1}\big\|v^*(x^{(t)})-\widetilde{v}^*(x^{(t)})\big\|^2.
\end{aligned} \tag{60}$$

By taking $w_i=p_i$ in eq. (60) for all i, we have $y^*(x^{(t)})=\widetilde{y}^*(x^{(t)})$ and $v^*(x^{(t)})=\widetilde{v}^*(x^{(t)})$, which results in

$$\begin{aligned}
\min_t\Big\|\nabla\Phi(x^{(t)})\Big\|^2&\leq\frac{2}{T}\Big[1+2\Big(\frac{\beta'_{\max}-\beta_{\min}}{\beta_{\min}}\Big)^2\Big]\sum_{t=0}^{T-1}\Big\|\nabla\widetilde{\Phi}(x^{(t)})\Big\|^2\\
&=\mathcal{O}\Big(\frac{M_1(n-P)}{(n-1)}\sqrt{\frac{\bar{\tau}}{PT}}\Big)+\mathcal{O}\Big(M_2\sqrt{\frac{1}{P\bar{\tau}T}}\Big)+\mathcal{O}\Big(\frac{M_4}{\bar{\tau}T}\Big).
\end{aligned}$$

Since we have the convergence rate of SimFBO and ShroFBO, the complexity analysis is the same as the complexity analysis Theorem 1. Thus, we finish the proof. □