Supplementary Materials: FedBCGD: Communication-Efficient Accelerated Block Coordinate Gradient Descent for Federated Learning

Anonymous Authors

1 APPENDIX A: BASIC ASSUMPTIONS AND NOTATIONS

1.1 Basic Assumptions

Before giving our theoretical results, we first present the common assumptions.

Assumption 1 (Convexity). *f_i is μ-strongly-convex for all $i \in [M]$, i.e.,*

$$f_i(\boldsymbol{y}) \geq f_i(\boldsymbol{x}) + \langle \nabla f_i(\boldsymbol{x}), \boldsymbol{y} - \boldsymbol{x} \rangle + \frac{\mu}{2} \|\boldsymbol{y} - \boldsymbol{x}\|^2 \tag{1}$$

for all $\boldsymbol{x}, \boldsymbol{y}$ in its domain and $i \in [M]$. We allow $\mu = 0$, which corresponds to general convex functions.

Assumption 2 (Smoothness). *The gradient of the loss function is Lipschitz continuous with constant β, for all $\boldsymbol{x}_1, \boldsymbol{x}_2 \in \mathbb{R}^d$*

$$\|\nabla f(\boldsymbol{x}_1) - \nabla f(\boldsymbol{x}_2)\| \leq \beta \|\boldsymbol{x}_1 - \boldsymbol{x}_2\|. \tag{2}$$

Assumption 3. *Let ζ be a mini-batch drawn uniformly at random from all samples. We assume that the data is distributed so that, for all $\boldsymbol{x} \in \mathbb{R}^d$*

$$\mathbb{E}_{\zeta|\boldsymbol{x}} [\nabla f_i(\boldsymbol{x}; \zeta)] = \nabla f_i(\boldsymbol{x}). \tag{3}$$

We also can get:

$$\mathbb{E}_{\zeta|\boldsymbol{x}} \left[\|\nabla f_i(\boldsymbol{x}; \zeta_i) - \nabla f_i(\boldsymbol{x})\|^2 \right] \leq \sigma^2. \tag{4}$$

Assumption 4 (Bounded heterogeneity). *The dissimilarity of $f_i(\boldsymbol{x})$ and $f(\boldsymbol{x})$ is bounded as follows:*

$$\frac{1}{M} \sum_{i=1}^{M} \|\nabla f_i(\boldsymbol{x}) - \nabla f(\boldsymbol{x})\|^2 \leq G^2. \tag{5}$$

Assumption 5 (Stochastic gradient smoothness). *The gradient of the loss function is Lipschitz continuous with constant β, for all $\boldsymbol{x}_1, \boldsymbol{x}_2 \in \mathbb{R}^d$*

$$\|\nabla f(\boldsymbol{x}_1; \zeta) - \nabla f(\boldsymbol{x}_2; \zeta)\| \leq \beta \|\boldsymbol{x}_1 - \boldsymbol{x}_2\|. \tag{6}$$

Assumption 2 bounds the variance of stochastic gradients, which is common in stochastic optimization analysis [?]. Assumption 3 bounds the gradient difference between global and local loss functions, which is a widely-used approach to characterize client heterogeneity in federated optimization literature [? ?]. Assumption 5 is a necessary assumption in stochastic gradient noise reduction, an assumption that is used only in the proof of the convergence speed of the FedBCGD+ algorithm.

1.2 Notation

We first define the notations to be used in analyzing the convergence properties of our algorithms.

1. $\boldsymbol{x}^r$ is the r communication rounds global model.

2. $\boldsymbol{x}^r_{(j)}$ is the j-th block of $\boldsymbol{x}$, so that $\boldsymbol{x}^r = \left[\boldsymbol{x}^{r\top}_{(1)}, \dots, \boldsymbol{x}^{r\top}_{(N)}\right]^\top$. Note that $\boldsymbol{x}^r_{(j)}$ is a virtual vector. It is realized at a hub j every r iterations, but we will study the evolution of this virtual vector in every iteration.

3. $\boldsymbol{x}^r_{k,j} \in \mathbb{R}^d$ are the local versions of the coordinates of the weight vector $\boldsymbol{x}^r_{(j)}$ that each client k if hub j updates.

4. $\boldsymbol{x}^\star$ is is the minimum value of the function $f(\boldsymbol{x})$.

5. $\boldsymbol{x}_{k,j,(j)}$ is the j-th block of $\boldsymbol{x}_{k,j}$ at client k in silo j, so that $\boldsymbol{x}_{(j)} = \frac{1}{K} \sum_{k=1}^{K} \boldsymbol{x}_{k,j,(j)}$.

6. $\boldsymbol{y}^{r,t}_{k,j}$ is the local parameter vector that client j in silo k at iteration t.

7. $\nabla_{(j)} f_{k,j}\left(\boldsymbol{y}_{k,j}; \zeta\right)$ is the partial derivative of $f(\boldsymbol{x})$ with respect to coordinate block j, computed at client k in silo j using the coordinates and rows at client k corresponding to minibatch ζ.

8. $\boldsymbol{G}^r = \left[\left(\boldsymbol{G}^r_{(1)}\right)^\top, \dots, \left(\boldsymbol{G}^r_{(N)}\right)^\top\right]^\top$, where $\mathbf{G}^r_{(j)} = \frac{1}{K} \sum_{k=1}^{K} \sum_{t=1}^{T} \nabla_{(j)} f_{k,j}\left(y^{r,t}_{k,j}; \zeta\right)$.

It should be noted that components on $\boldsymbol{x}$, i.e., $\boldsymbol{x}_{(j)}$ are realized every T iterations when the hubs communicate with clients and with other hubs, but we will study the evolution of these virtual vectors at each iteration. Therefore, based on the above definitions, assumptions and our algorithms, we can express the evolution of the virtual global parameter/weight vector in the following forms:

$$\boldsymbol{x}^r = \begin{bmatrix} \boldsymbol{x}^r_{(1)} \\ \boldsymbol{x}^r_{(2)} \\ \vdots \\ \boldsymbol{x}^r_{(N)} \end{bmatrix} = \frac{1}{K} \begin{bmatrix} \sum_{k=1}^{K} x^r_{k,1,(1)} \\ \sum_{k=1}^{K} x^r_{k,2,(2)} \\ \vdots \\ \sum_{k=1}^{K} x^r_{k,N,(N)} \end{bmatrix} \tag{7}$$

$$\boldsymbol{x}^{r+1}=\boldsymbol{x}^{r}-\frac{\eta}{K}\left[\begin{array}{c}\sum_{k=1}^{K}\sum_{t=1}^{T}\nabla_{(1)}f_{k,1}\left(y_{k,1}^{r,t};\zeta\right)\\ \sum_{k=1}^{K}\sum_{t=1}^{T}\nabla_{(2)}f_{k,2}\left(y_{k,2}^{r,t};\zeta\right)\\ \vdots\\ \sum_{k=1}^{K}\sum_{t=1}^{T}\nabla_{(N)}f_{k,N}\left(y_{k,N}^{r,t};\zeta\right).\end{array}\right] \tag{8}$$

In this case, we update all coordinates of the global weight vector $\boldsymbol{x}^r$, virtually at each time step t. We have the virtual gradient at each time instant t as:

$$\mathrm{G}^{\mathrm{r}}=\frac{1}{K}\left[\begin{array}{c}\sum_{k=1}^{K}\sum_{t=1}^{T}\nabla_{(1)}f_{k,1}\left(y_{k,1}^{r,t};\zeta\right)\\ \sum_{k=1}^{K}\sum_{t=1}^{T}\nabla_{(2)}f_{k,2}\left(y_{k,2}^{r,t};\zeta\right)\\ \vdots\\ \sum_{k=1}^{K}\sum_{t=1}^{T}\nabla_{(N)}f_{k,N}\left(y_{k,N}^{r,t};\zeta\right),\end{array}\right] \tag{9}$$

$$\mathbb{E}_{\mathcal{S}}\left[\mathrm{G}^{\mathrm{r}}\right]=\frac{1}{M}\left[\begin{array}{c}\sum_{i=1}^{M}\sum_{t=1}^{T}\nabla_{(1)}f_{i}\left(y_{i}^{r,t}\right)\\ \sum_{i=1}^{M}\sum_{t=1}^{T}\nabla_{(2)}f_{i}\left(y_{i}^{r,t}\right)\\ \vdots\\ \sum_{i=1}^{M}\sum_{t=1}^{T}\nabla_{(N)}f_{i}\left(y_{i}^{r,t}\right).\end{array}\right] \tag{10}$$

We optimize the objective function of the tiered decentralized coordinate descent approach with periodic averaging. The objective is to train a global model $\boldsymbol{x}^r$, which is a d-vector that can be decomposed as follows:

$$\boldsymbol{x}^{r}=\left[\boldsymbol{x}_{(1)}^{r\top},\ldots,\boldsymbol{x}_{(N)}^{r\top}\right]^{\top} \tag{11}$$

where each $\boldsymbol{x}_{(j)}^{r}$ is the block of $\boldsymbol{x}^r$, or coordinates, for block j, r is communication rounds. The goal of the training algorithm is to minimize an objective function with following structures.

2 APPENDIX B: THEORETICAL RESULTS OF FEDBCGD, FEDBCGD+

In this section, we only present the main theoretical results of the proposed FedBCGD, FedBCGD+ algorithms in Theorems 1-2, respectively. The detailed proofs of Theorems 1-2 are given in Appendices respectively.

Moreover, we provide the convergence properties of the proposed FedBCGD algorithm. In addition, we also present the detailed proof for the theoretical results in the next subsection.

Theorem 1 (Convergence rates of FedBCGD). *Suppose that each function $\{f_i\}$ satisfies Assumptions 1, 2, and 3. Then, in each of the following cases, there exist weights $\{w_r\}$ and local step-sizes η, the output of FedBCGD (i.e. $\overline{z}^R$) satisfies the following inequalities.*

***1. Case of strongly convex**: f_i satisfies Assumption 1 for $\mu>0$, $\tilde{\eta}=\frac{\alpha\eta T}{4}$, $\tilde{\eta}\leq\frac{1}{\beta}$ then*

$$\begin{aligned}\mathbb{E}\left[f\left(\bar{z}^{R}\right)\right]-f\left(x^{\star}\right)\leq\left\|x^{0}-x^{\star}\right\|^{2}\mu\exp\left(-\frac{\alpha\mu R}{\beta}\right)+\frac{128\left[\left(1-\frac{K}{M}\right)\frac{1}{K}\right]G^{2}+32\frac{\sigma^{2}}{KT}}{\mu R}\\+\frac{\left(384\beta G^{2}+\frac{192\beta}{T}\sigma^{2}\right)}{\alpha^{2}\mu^{2}R^{2}}+\frac{\left(6144\beta^{2}G^{2}+\frac{3072}{T}\beta^{2}\sigma^{2}\right)}{\alpha^{2}\mu^{3}R^{3}}\end{aligned} \tag{12}$$

***2. Case of general convex**: Each f_i satisfies Assumption 1 for $\mu=0$, $\tilde{\eta}=\frac{\alpha\eta T}{4}$, $\tilde{\eta}\leq\frac{1}{\beta}$ then*

$$\begin{aligned}&\mathbb{E}\left[f\left(\bar{z}^{R}\right)\right]-f\left(x^{\star}\right)\\&\leq\frac{\beta^{\frac{3}{2}}d_{0}}{\alpha R}+\frac{\left(6144\beta^{2}G^{2}+\frac{3072}{T}\beta^{2}\sigma^{2}\right)}{\alpha^{2}R}+\frac{\left[32\left[\left(1-\frac{K}{M}\right)\frac{1}{K}\right]G^{2}+32\frac{\sigma^{2}}{KT}\right]^{\frac{1}{2}}d_{0}^{\frac{1}{2}}}{\sqrt{R}}\\&\quad+\frac{\left(384\beta G^{2}+\frac{192\beta}{T}\sigma^{2}\right)^{\frac{1}{3}}d_{0}^{\frac{2}{3}}}{\alpha^{\frac{2}{3}}R^{\frac{2}{3}}}.\end{aligned} \tag{13}$$

3. Case of non-convex: *Each f_i satisfies Assumption 2 and $\tilde{\eta} = \frac{\alpha\eta T}{4}$, $\tilde{\eta} \leq \frac{1}{\beta}$, then*

$$
\begin{aligned}
&\frac{1}{R}\sum_{r=1}^{R}\left\|\nabla f\left(x^r\right)\right\|^2 \leq \frac{16\beta d_0}{TK\alpha R} + \frac{2\sqrt{d_0}}{\sqrt{RTM}}\left(\frac{8\beta}{K}\left(1-\frac{K}{M}\right)G^2 + \frac{8\beta\sigma^2}{TK}\left(1-\frac{K}{M}\right) + \frac{8\beta}{TM}\sigma^2\right)^{\frac{1}{2}} \\
&+ 2\left(\frac{d_0}{R}\right)^{\frac{2}{3}}\left[\frac{384\beta^2}{\alpha^2}G^2 + \frac{92\beta^2}{T}\frac{\sigma^2}{\alpha^2} + \frac{\left(16\gamma^2\beta^2\right)}{TM}\sigma^2 + \frac{\left(16\gamma^2\beta^2\right)\sigma^2}{TK}\left(1-\frac{K}{M}\right) + \frac{16\gamma^2\beta^2}{K}\left(1-\frac{K}{M}\right)G^2\right]^{\frac{1}{3}} \\
&+ 2\left(\frac{d_0}{R}\right)^{\frac{3}{4}}\left[\frac{4608}{\alpha^2}\frac{\beta^3}{K}\left(1-\frac{K}{M}\right)G^2 + \frac{1152\beta^3}{KT\alpha^2}\left(1-\frac{K}{M}\right)\sigma^2\right]^{\frac{1}{4}} \\
&+ 2\left(\frac{d_0}{R}\right)^{\frac{4}{5}}\left[\frac{9216}{\alpha^2}\frac{\gamma^2\beta^4}{K}\left(1-\frac{K}{M}\right)G^2 + \frac{2304\beta^4}{K}\frac{\gamma^2}{\alpha^2 T}\left(1-\frac{K}{M}\right)\sigma^2\right]^{\frac{1}{5}}.
\end{aligned} \tag{14}
$$

Theorem 2 (Convergence rates of FedBCGD+). *Suppose that each function $\{f_i\}$ satisfies Assumptions* 1, 2, *and* 3. *Then, in each of the following cases, there exist weights $\{w_r\}$ and local step-sizes η, the output of FedBCGD+ (i.e., $\overline{z}^R$) satisfies the following inequalities.*

1. Case of strongly convex: *Each f_i satisfies Assumption 1 for $\mu > 0$, $\tilde{\eta} = \frac{\alpha\eta T}{4}$, $\tilde{\eta} \leq \min\left(\frac{1}{81\beta}, \frac{S}{15\mu N}\right)$ then*

$$\mathbb{E}\left[f\left(\overline{z}^R\right)\right] - f\left(x^\star\right) \leq \tilde{O}\left(\frac{M\mu}{K}\tilde{D}^2\exp\left(-\min\left\{\frac{M}{30K}, \frac{\mu}{162\beta}\right\}R\right)\right). \tag{15}$$

2. Case of General convex: *Each f_i satisfies Assumption 1 for $\mu = 0$, $\tilde{\eta} \leq \frac{1}{\beta}$ then*

$$\mathbb{E}\left[f\left(\bar{z}^R\right)\right] - f\left(x^\star\right) \leq O\left(\sqrt{\frac{M}{K}}\frac{\beta\tilde{D}^2}{R}\right). \tag{16}$$

3. Case of non-convex: *Each f_i satisfies Assumption 2 and $\tilde{\eta} = \frac{1}{4}\alpha T$, $\tilde{\eta} \leq \frac{1}{24\beta}\left(\frac{K}{M}\right)^{\frac{2}{3}}$ then*

$$\mathbb{E}\left[\left\|\nabla f\left(\bar{z}^R\right)\right\|^2\right] \leq O\left(\frac{\beta F}{R}\left(\frac{M}{K}\right)^{\frac{2}{3}}\right), \tag{17}$$

where $\tilde{D}^2 := \left(\left\|\boldsymbol{x}^0 - \boldsymbol{x}^\star\right\|^2 + \frac{1}{2N\beta^2}\sum_{i=1}^{N}\left\|\boldsymbol{c}_i^0 - \nabla f_i\left(\boldsymbol{x}^\star\right)\right\|^2\right)$ and $F := \left(f\left(\boldsymbol{x}_0\right) - f\left(\boldsymbol{x}^\star\right)\right)$.

3 APPENDIX C: MAIN LEMMAS

In this section, we prove some main lemmas, which play key roles for the proofs of Theorems 1-3.

Lemma 1. *The following holds for any β-smooth and μ-strongly convex function h, and any $\boldsymbol{x}, \boldsymbol{y}, \boldsymbol{z}$ in the domain of h :*

$$\langle\nabla h(\boldsymbol{x}), \boldsymbol{z}-\boldsymbol{y}\rangle \geq h(\boldsymbol{z}) - h(\boldsymbol{y}) + \frac{\mu}{4}\|\boldsymbol{y}-\boldsymbol{z}\|^2 - \beta\|\boldsymbol{z}-\boldsymbol{x}\|^2. \tag{18}$$

Proof. Given any $\boldsymbol{x}, \boldsymbol{y}$, and $\boldsymbol{z}$, we get the following two inequalities using smoothness and strong convexity of h :

$$\langle\nabla h(\boldsymbol{x}), \boldsymbol{z}-\boldsymbol{x}\rangle \geq h(\boldsymbol{z}) - h(\boldsymbol{x}) - \frac{\beta}{2}\|\boldsymbol{z}-\boldsymbol{x}\|^2, \tag{19}$$

$$\langle\nabla h(\boldsymbol{x}), \boldsymbol{x}-\boldsymbol{y}\rangle \geq h(\boldsymbol{x}) - h(\boldsymbol{y}) + \frac{\mu}{2}\|\boldsymbol{y}-\boldsymbol{x}\|^2. \tag{20}$$

Furthermore, applying the relaxed triangle inequality, we can get

$$\frac{\mu}{2}\|\boldsymbol{y}-\boldsymbol{x}\|^2 \geq \frac{\mu}{4}\|\boldsymbol{y}-\boldsymbol{z}\|^2 - \frac{\mu}{2}\|\boldsymbol{x}-\boldsymbol{z}\|^2. \tag{21}$$

Combining all the inequalities together, we have

$$\langle\nabla h(\boldsymbol{x}), \boldsymbol{z}-\boldsymbol{y}\rangle \geq h(\boldsymbol{z}) - h(\boldsymbol{y}) + \frac{\mu}{4}\|\boldsymbol{y}-\boldsymbol{z}\|^2 - \frac{\beta+\mu}{2}\|\boldsymbol{z}-\boldsymbol{x}\|^2. \tag{22}$$

The lemma follows since $\beta \geq \mu$.

Lemma 2 (Bounding heterogeneity). *Recall our bound on the gradient dissimilarity:*

$$\frac{1}{M}\sum_{i=1}^{M}\|\nabla f_i(x) - \nabla f(x)\|^2 \leq G^2. \tag{23}$$

If $\{f_i\}$ are convex, we can relax the assumption to

$$\frac{1}{M}\sum_{i=1}^{M}\|\nabla f_i(\boldsymbol{x})\|^2 \le G^2 + 2\beta\left(f(\boldsymbol{x}) - f^\star\right). \tag{24}$$

Proof. According to the inequality $\frac{1}{n}\sum_{i=1}^{n}\|\mathbf{a}_i - \bar{\mathbf{a}}\|_2^2 = \frac{1}{n}\sum_{i=1}^{n}\|\mathbf{a}_i\|^2 - \|\bar{\mathbf{a}}\|^2$ for $\mathbf{a}_i \in \mathbb{R}^d, \bar{\mathbf{a}} = \frac{1}{n}\sum_{i=1}^{n}\mathbf{a}_i$,

$$\frac{1}{M}\sum_{i=1}^{M}\|\nabla f_i(x) - \nabla f(x)\|^2 \le G^2, \tag{25}$$

$$\begin{aligned}
&\frac{1}{M}\sum_{i=1}^{M}\|\nabla f_i(x)\|^2 \le \|\nabla f(x)\|^2 + G^2 \\
&\le \left\|\nabla f(x) - \nabla f\left(x^\star\right)\right\|^2 + G^2 \\
&\le \frac{1}{M}\sum_{i=1}^{M}\left\|\nabla f_i(x) - \nabla f_i\left(x^\star\right)\right\|^2 + G^2 \\
&\le 2\beta\left(f(x) - f^\star\right) + G^2.
\end{aligned} \tag{26}$$

Lemma 3. ***(Relaxed triangle inequality).*** *Let $\{v_1, \ldots, v_\tau\}$ be τ vectors in $\mathbb{R}^d$. Then the following inequalities are true:*
1. $\left\|v_i + v_j\right\|^2 \le (1+a)\|v_i\|^2 + \left(1 + \frac{1}{a}\right)\left\|v_j\right\|^2$ for any $a > 0$, and
2. $\left\|\sum_{i=1}^{\tau} v_i\right\|^2 \le \tau\sum_{i=1}^{\tau}\|v_i\|^2$.

Lemma 4. *K is the number of selected clients in block j and M is the total number of clients. The following inequalities can be obtained.*

$$\mathbb{E}\left\|\frac{1}{K}\sum_{i=1}^{K}\nabla f_i(x)\right\|^2 \le \mathbb{E}\|\nabla f(x)\|^2 + \mathbb{E}\left(1 - \frac{K}{M}\right)\frac{1}{KM}\sum_{i=1}^{M}\|\nabla f_i(x)\|^2, \tag{27}$$

$$\mathbb{E}\left\|\frac{1}{K}\sum_{i=1}^{K}\nabla f_i(x)\right\|^2 \le \frac{1}{M}\sum_{i=1}^{M}\|\nabla f_i(x)\|^2. \tag{28}$$

Proof. Define $\mathbb{I}_i$ as the random variable which indicates client i is selected in the r-th global epoch.

$$\begin{aligned}
&\mathbb{E}\left\|\frac{1}{K}\sum_{i=1}^{K}\nabla f_i(x)\right\|^2 \\
&= \mathbb{E}\left\|\frac{1}{K}\sum_{i=1}^{M}\nabla f_i(x)\mathbb{I}_i\right\| \\
&= \mathbb{E}\left\langle\frac{1}{K}\sum_{i=1}^{M}\nabla f_i(x)\mathbb{I}_i, \frac{1}{K}\sum_{i=1}^{M}\nabla f_j(x)\mathbb{I}_j\right\rangle \\
&= \mathbb{E}\frac{1}{K^2}\left[\sum_{i,j\in[M], i\neq j}\left\langle\nabla f_i(x), \nabla f_j(x)\right\rangle\mathbb{E}\left[\mathbb{I}_i\mathbb{I}_j\right] + \sum_{i\in[M]}\left\langle\nabla f_i(x), \nabla f_i(x)\right\rangle\mathbb{E}\left[\mathbb{I}_i\right]\right] \\
&= \mathbb{E}\frac{1}{K^2}\left[\sum_{i,j\in[M], i\neq j}\frac{K(K-1)}{M(M-1)}\left\langle\nabla f_i(x), \nabla f_j(x)\right\rangle + \sum_{i\in[M]}\frac{K}{M}\left\langle\nabla f_i(x), \nabla f_i(x)\right\rangle\right] \\
&= \mathbb{E}\frac{1}{K^2}\left[\sum_{i,j\in[M]}\frac{K(K-1)}{M(M-1)}\left\langle\nabla f_i(x), \nabla f_j(x)\right\rangle + \sum_{i\in[M]}\frac{K(M-K)}{M(M-1)}\left\langle\nabla f_i(x), \nabla f_i(x)\right\rangle\right] \\
&\le \mathbb{E}\|\nabla f(x)\|^2 + \mathbb{E}\left(1 - \frac{K}{M}\right)\frac{1}{KM}\sum_{i\in[M]}\|\nabla f_i(x)\|^2 \\
&\le \frac{1}{M}\sum_{i\in[M]}\|\nabla f_i(x)\|^2.
\end{aligned} \tag{29}$$

We will now proceed to the second part of our lemma's exposition.

$$\mathbb{E}\left\|\frac{1}{K}\sum_{i=1}^{K}\nabla f_i(x)\right\|^2 \leq \frac{1}{M}\sum_{i=1}^{M}\|\nabla f_i(x)\|^2 . \tag{30}$$

Proof :

$$\begin{aligned}
&\mathbb{E}\left\|\frac{1}{K}\sum_{i=1}^{K}\nabla f_i(x)\right\|^2 = \mathbb{E}\left\|\frac{1}{K}\sum_{i=1}^{M}\nabla f_i(x)\mathbb{I}_i\right\| \\
&= \mathbb{E}\left\langle\frac{1}{K}\sum_{i=1}^{M}\nabla f_i(x)\mathbb{I}_i, \frac{1}{K}\sum_{i=1}^{M}\nabla f_j(x)\mathbb{I}_j\right\rangle \\
&= \mathbb{E}\frac{1}{K^2}\left[\sum_{i,j\in[M],i\neq j}\left\langle\nabla f_i(x),\nabla f_j(x)\right\rangle\mathbb{E}\left[\mathbb{I}_i\mathbb{I}_j\right] + \sum_{i\in[M]}\left\langle\nabla f_i(x),\nabla f_i(x)\right\rangle\mathbb{E}\left[\mathbb{I}_i\right]\right] \\
&= \mathbb{E}\frac{1}{K^2}\left[\sum_{i,j\in[M],i\neq j}\frac{K(K-1)}{M(M-1)}\left\langle\nabla f_i(x),\nabla f_j(x)\right\rangle + \sum_{i\in[M]}\frac{K}{M}\left\langle\nabla f_i(x),\nabla f_i(x)\right\rangle\right] \\
&= \mathbb{E}\frac{1}{K^2}\left[\sum_{i,j\in[M]}\frac{K(K-1)}{M(M-1)}\left\langle\nabla f_i(x),\nabla f_j(x)\right\rangle + \sum_{i\in[M]}\frac{K(M-K)}{M(M-1)}\left\langle\nabla f_i(x),\nabla f_i(x)\right\rangle\right] \\
&\leq \frac{M^2}{K^2}\frac{K(K-1)}{M(M-1)}\mathbb{E}\|\nabla f(x)\|^2 + \mathbb{E}\frac{1}{K^2}\left[\frac{K}{M} - \frac{K(K-1)}{M(M-1)}\right]\sum_{i\in[M]}\|\nabla f_i(x)\|^2 \\
&\leq \frac{M}{K}\frac{(K-1)}{(M-1)}\mathbb{E}\|\nabla f(x)\|^2 + \frac{1}{K}\left[1-\frac{(K-1)}{(M-1)}\right]\frac{1}{M}\sum_{i\in[M]}\mathbb{E}\|\nabla f_i(x)\|^2 \\
&\leq \frac{M}{K}\frac{(K-1)}{(M-1)}\frac{1}{M}\sum_{i\in[M]}\mathbb{E}\|\nabla f_i(x)\|^2 + \frac{1}{K}\left[1-\frac{(K-1)}{(M-1)}\right]\frac{1}{M}\sum_{i\in[M]}\mathbb{E}\|\nabla f_i(x)\|^2 \\
&= \left(\frac{M}{K}\frac{(K-1)}{(M-1)} + \frac{1}{K}\left[1-\frac{(K-1)}{(M-1)}\right]\right)\frac{1}{M}\sum_{i\in[M]}\mathbb{E}\|\nabla f_i(x)\|^2 \\
&= \frac{1}{M}\sum_{i\in[M]}\mathbb{E}\|\nabla f_i(x)\|^2 .
\end{aligned} \tag{31}$$

Lemma 5 (Bounded drift).

$$\sum_{i=1}^{M}\sum_{t=1}^{T}\mathbb{E}\left\|y_i^{r,t} - x^r\right\|^2 \leq 6T^3\eta^2\sum_{i=1}^{M}\left\|\nabla f_i\left(x^r\right)\right\|^2 + 3MT^2\eta^2\sigma^2. \tag{32}$$

Proof.

$$\begin{aligned}
&\mathbb{E}\left\|y_i^{r,t-1} - x^r - \eta\nabla f_i\left(y_i^{r,t-1};\zeta\right)\right\|^2 \\
&\leq \mathbb{E}\left\|y_i^{r,t-1} - x^r - \eta\nabla f_i\left(y_i^{r,t-1}\right)\right\|^2 + \eta^2\sigma^2 \\
&\overset{a}{\leq} \left(1+\frac{1}{T-1}\right)\mathbb{E}\left\|y_i^{r,t-1} - x^r\right\|^2 + T\eta^2\left\|\nabla f_i\left(y_i^{r,t-1}\right)\right\|^2 + \eta^2\sigma^2 \\
&= \left(1+\frac{1}{T-1}\right)\mathbb{E}\left\|y_i^{r,t-1} - x^r\right\|^2 + T\eta^2\left\|\nabla f_i\left(y_i^{r,t-1}\right) - \nabla f_i\left(x^r\right) + \nabla f_i\left(x^r\right)\right\|^2 + \eta^2\sigma^2 \\
&\leq \left(1+\frac{1}{T-1}\right)\mathbb{E}\left\|y_i^{r,t-1} - x^r\right\|^2 + 2T\eta^2\left\|\nabla f_i\left(y_i^{r,t-1}\right) - \nabla f_i\left(x^r\right)\right\|^2 + 2T\eta^2\left\|\nabla f_i\left(x^r\right)\right\|^2 + \eta^2\sigma^2 \\
&\leq \left(1+\frac{1}{T-1} + 2T\eta^2\beta^2\right)\mathbb{E}^2\left\|y_i^{r,t-1} - x^r\right\|\|^2 + 2T\eta^2\|\nabla f_i\left(x^r\right)\|^2 + \eta^2\sigma^2 \\
&\leq \left(1+\frac{2}{(T-1)}\right)\mathbb{E}^2\left\|y_i^{r,t-1} - x^r\right\|^2 + 2T\eta^2\left\|\nabla f_i\left(x^r\right)\right\|^2 + \eta^2\sigma^2,
\end{aligned} \tag{33}$$

where the inequality $\overset{a}{\leq}$ follows directly from Lemma 3. Let $2T\eta^2\beta^2 \leq \frac{1}{(T-1)}$, and unrolling the above recursion, we have

$$\begin{aligned}
&\mathbb{E}\left\|y_i^{r,t-1} - x^r\right\|^2 \\
&\leq \sum_{\tau=1}^{t-1}\left(2T\eta^2\left\|\nabla f_i\left(x^r\right)\right\|^2 + \eta^2\sigma^2\right)\left(1 + \frac{2}{(T-1)}\right)^\tau \\
&\leq \left(2T\eta^2\left\|\nabla f_i\left(x^r\right)\right\|^2 + \eta^2\sigma^2\right)3T.
\end{aligned} \tag{34}$$

So, we can get

$$\begin{aligned}
&\sum_{i=1}^{M}\sum_{t=1}^{T}\mathbb{E}\left\|y_k^{r,t} - x^r\right\|^2 \leq \sum_{i=1}^{M}\mathbb{E}\left(2T\eta^2\left\|\nabla f_i\left(x^r\right)\right\|^2 + \eta^2\sigma^2\right)3T^2 \\
&\leq 6T^3\eta^2\sum_{i=1}^{M}\mathbb{E}\left\|\nabla f_i\left(x^r\right)\right\|^2 + 3MT^2\eta^2\sigma^2.
\end{aligned} \tag{35}$$

Lemma 6. *The variance of* $\mathbf{G}^r$ *can be bounded by the following inequality*

$$\mathbb{E}\sum_{j=1}^{N}\left\|\frac{1}{K}\sum_{k=1}^{K}\sum_{t=1}^{T}\nabla_{(j)}f_{k,j}\left(y_{k,j}^{r,t};\zeta\right)\right\|^2 \leq 2\frac{T}{K}\sum_{k=1}^{K}\sum_{t=1}^{T}\left\|\nabla f_i\left(y_i^{r,t}\right)\right\|^2 + 2\frac{T}{K}\sigma^2. \tag{36}$$

Proof.

$$\begin{aligned}
&\mathbb{E}\sum_{j=1}^{N}\left\|\frac{1}{K}\sum_{k=1}^{K}\sum_{t=1}^{T}\nabla_{(j)}f_{k,j}\left(y_{k,j}^{r,t};\zeta\right)\right\|^2 \\
&\leq \frac{T}{M}\sum_{k=1}^{K}\sum_{t=1}^{T}\sum_{j=1}^{N}\left\|\nabla_{(j)}f_i\left(y_i^{r,t};\zeta\right)\right\|^2 \\
&\leq \frac{T}{K}\sum_{k=1}^{K}\sum_{t=1}^{T}\left\|\nabla f_i\left(y_i^{r,t};\zeta\right)\right\|^2 \\
&\leq 2\frac{T}{K}\sum_{k=1}^{K}\sum_{t=1}^{T}\left\|\nabla f_i\left(y_i^{r,t}\right)\right\|^2 + 2\frac{T}{K}\sigma^2.
\end{aligned} \tag{37}$$

Lemma 7 (Linear convergence rate)). *For every non-negative sequence* $\{d_{r-1}\}_{r\geq 1}$ *and any parameters* $\mu > 0$, $\eta_{\max} \in (0, 1/\mu]$, $c \geq 0$, $R \geq \frac{1}{2\eta_{\max}\mu}$, *there exists a constant step-size* $\eta \leq \eta_{\max}$ *and weights* $w_r := (1-\mu\eta)^{1-r}$ *such that for* $W_R := \sum_{r=1}^{R+1} w_r$

$$\Psi_R := \frac{1}{W_R}\sum_{r=1}^{R+1}\left(\frac{w_r}{\eta}(1-\mu\eta)d_{r-1} - \frac{w_r}{\eta}d_r + c\eta w_r\right) = \tilde{O}\left(\mu d_0\exp\left(-\mu\eta_{\max}R\right) + \frac{c}{\mu R}\right). \tag{38}$$

Lemma 8 (Sub-linear convergence rate). *For every non-negative sequence* $\{d_{r-1}\}_{r\geq 1}$ *and any parameters* $\eta_{\max} \geq 0, c \geq 0, R \geq 0$, *there exists a constant step-size* $\eta \leq \eta_{\max}$ *and weights* $w_r = \overline{1}$ *such that,*

$$\Psi_R := \frac{1}{R+1}\sum_{r=1}^{R+1}\left(\frac{d_{r-1}}{\eta} - \frac{d_r}{\eta} + c_1\eta + c_2\eta^2\right) \leq \frac{d_0}{\eta_{\max}(R+1)} + \frac{2\sqrt{c_1 d_0}}{\sqrt{R+1}} + 2\left(\frac{d_0}{R+1}\right)^{\frac{2}{3}} c_2^{\frac{1}{3}}. \tag{39}$$

Lemma 9 (Separating mean and variance). *Let* $\{\Xi_1, \ldots, \Xi_\tau\}$ *be* τ *random variables in* $\mathbb{R}^d$ *which are not necessarily independent. First suppose that their mean is* $\mathbb{E}\left[\Xi_i\right] = \xi_i$ *and variance is bounded as* $\mathbb{E}\left[\|\Xi_i - \xi_i\|^2\right] \leq \sigma^2$. *Then, the following holds*

$$\mathbb{E}\left[\left\|\sum_{i=1}^{\tau}\Xi_i\right\|^2\right] \leq \left\|\sum_{i=1}^{\tau}\xi_i\right\|^2 + \tau^2\sigma^2. \tag{40}$$

Now instead suppose that their conditional mean is $\mathbb{E}\left[\Xi_i \mid \Xi_{i-1}, \ldots \Xi_1\right] = \xi_i$ *i.e. the variables* $\{\Xi_i - \xi_i\}$ *form a martingale difference sequence, and the variance is bounded by* $\mathbb{E}\left[\|\Xi_i - \xi_i\|^2\right] \leq \sigma^2$ *as before. Then we can show the tighter bound*

$$\mathbb{E}\left[\left\|\sum_{i=1}^{\tau}\Xi_i\right\|^2\right] \leq 2\left\|\sum_{i=1}^{\tau}\xi_i\right\|^2 + 2\tau\sigma^2. \tag{41}$$

4 APPENDIX D: PROOF OF THEOREM 1

4.1 1. The rate of strongly convex and smooth convergence:

We outline the FEDBCGD algorihtm in Algorithm 1. In round r, we perform the following updates:

$$v^r = \lambda v^{r-1} + \Delta x^{r-1}, \Delta x^{r-1} = \eta \mathbf{G}^r, \tag{42}$$

$$x^r = x^{r-1} + v^{r-1}. \tag{43}$$

Before giving the convergence analysis of Theorem 3, we first present the following lemma.

Lemma 10. *Let* $z^r = x^r + \gamma\left(x^r - x^{r-1}\right), \gamma = \frac{\lambda}{1-\lambda}$, *we can get*

$$z^{r+1} = z^r - \frac{1}{1-\lambda}\eta \mathbf{G}^r. \tag{44}$$

Proof.

$$\begin{aligned}
z^{r+1} &= x^{r+1} + \gamma\left(x^{r+1} - x^r\right) \\
&= x^r + v^{r+1} + \gamma\left(v^{r+1}\right) \\
&= z^r - \gamma\left(v^r\right) + v^{r+1} + \gamma\left(v^{r+1}\right) \\
&= z^r - \gamma v^r + (1+\gamma) v^{r+1} \\
&= z^r - \gamma v^r + (1+\gamma)\left(\lambda v^r - \eta \mathbf{G}^r\right) \\
&\overset{a}{=} z^r + (-\gamma + (1+\gamma)\beta) v^r + (1+\gamma)\left(-\eta \mathbf{G}^r\right) \\
&= z^r - \eta(1+\gamma)\mathbf{G}^r \\
&= z^r - \frac{1}{1-\lambda}\eta \mathbf{G}^r,
\end{aligned} \tag{45}$$

with the equality $\overset{a}{=}$, we let $(-\gamma + (1+\gamma)\lambda) = 0, \gamma = \frac{\lambda}{1-\lambda}$. We complete the proof.

4.2 The proof of Theorem 1

Proof. We can then apply $z^{r+1} = z^r - \alpha\eta \mathbf{G}^r, \alpha = \frac{1}{1-\lambda}$ to bound the second moment of the server update as

$$\begin{aligned}
\mathbb{E}\left\|z^{r+1} - x^\star\right\|^2 &= \mathbb{E}\left\|z^{r+1} - x^\star\right\|^2 \\
&\le \mathbb{E}\left\|z^r - x^\star\right\|^2 + \eta\alpha \underbrace{\mathbb{E}\left\langle -\mathbf{G}^r, z^r - x^\star\right\rangle}_{C_1} + \eta^2\alpha^2 \underbrace{\mathbb{E}\left\|\mathbf{G}^r\right\|^2}_{C_2}.
\end{aligned} \tag{46}$$

The term C_1 can be bounded by using perturbed strong-convexity (Lemma 1) with $h = f_k, \boldsymbol{x} = y_k^{r,t}, \boldsymbol{y} = \boldsymbol{x}^\star$, and $\boldsymbol{z} = z^r$ to get

$$\begin{aligned}
C_1 &= -\mathbb{E}\left\langle \mathbf{G}^r, z^r - x^\star\right\rangle \\
&= -\sum_{j=1}^{N}\left\langle \frac{1}{M}\sum_{i=1}^{M}\sum_{t=1}^{T}\nabla_{(j)} f_i\left(y_i^{r,t}\right), z_{(j)}^r - x_{(j)}^\star\right\rangle \\
&= -\left\langle \frac{1}{M}\sum_{k=1}^{M}\sum_{t=1}^{T}\nabla f_i\left(y_i^{r,t}\right), z^r - x^\star\right\rangle \\
&\le -\frac{1}{M}\sum_{k=1}^{M}\sum_{i=1}^{T}\left(f_i\left(z^r\right) - f_i\left(x^\star\right) - \beta\left\|y_i^{r,t} - z^r\right\|^2 + \frac{\mu}{4}\left\|z^r - x^\star\right\|^2\right) \\
&\le T\left(-f\left(z^r\right) + f\left(x^\star\right) - \frac{\mu}{4}\left\|z^r - x^\star\right\|^2\right) + \frac{\beta}{M}\sum_{i=1}^{M}\sum_{t=1}^{T}\left\|y_i^{r,t} - z^r\right\|^2.
\end{aligned} \tag{47}$$

The term C_2 can be bounded by using Lemma 6 in $\overset{a}{\le}$, Lemma 3 in $\overset{b}{\le}$, Lemma 4 and Lemma 2 in $\overset{c}{\le}$.

$$
\begin{aligned}
C_2 &= \mathbb{E}\left\|\mathbf{G}^r\right\|^2 \\
&= \sum_{j=1}^{N} \mathbb{E}\left\|\frac{1}{K}\sum_{k=1}^{K}\sum_{t=1}^{T}\nabla_{(j)} f_{k,j}\left(y_{k,j}^{r,t};\zeta\right)\right\|^2 \\
&= \sum_{j=1}^{N} \mathbb{E}\left\|\frac{1}{K}\sum_{k=1}^{K}\sum_{t=1}^{T}\nabla_{(j)} f_{k,j}\left(y_{k,j}^{r,t};\zeta\right) - \frac{1}{K}\sum_{k=1}^{K}\sum_{t=1}^{T}\nabla_{(j)} f_{k,j}\left(x^r\right) + \frac{1}{K}\sum_{k=1}^{K}\sum_{t=1}^{T}\nabla_{(j)} f_{k,j}\left(x^r\right)\right\|^2 \\
&\leq \sum_{j=1}^{N} 2\mathbb{E}\left\|\frac{1}{K}\sum_{k=1}^{K}\sum_{t=1}^{T}\nabla_{(j)} f_{k,j}\left(y_{k,j}^{r,t};\zeta\right) - \frac{1}{K}\sum_{k=1}^{K}\sum_{t=1}^{T}\nabla_{(j)} f_{k,j}\left(x^r\right)\right\|^2 + \sum_{j=1}^{N} 2\mathbb{E}\left\|\frac{1}{K}\sum_{k=1}^{K}\sum_{t=1}^{T}\nabla_{(j)} f_{k,j}\left(x^r\right)\right\|^2 \\
&\overset{\text{a}}{\leq} 2\frac{T}{M}\sum_{j=1}^{N}\sum_{i=1}^{M}\sum_{t=1}^{T}\left\|\nabla_{(j)} f_i\left(y_i^{r,t};\zeta\right) - \nabla_{(j)} f_i\left(x^r\right)\right\|^2 + \sum_{j=1}^{N} 2T^2\mathbb{E}\left\|\frac{1}{K}\sum_{k=1}^{K}\nabla_{(j)} f_{k,j}\left(x^r\right)\right\|^2 \\
&\leq 4\frac{T}{M}\sum_{i=1}^{M}\sum_{t=1}^{T}\left\|\nabla f_i\left(y_i^{r,t}\right) - \nabla f_i\left(x^r\right)\right\|^2 + \sum_{j=1}^{N} 2T^2\mathbb{E}\left\|\frac{1}{K}\sum_{k=1}^{K}\nabla_{(j)} f_{k,j}\left(x^r\right)\right\|^2 + 4\frac{T}{K}\sigma^2 \\
&\leq 4\frac{T}{M}\sum_{i=1}^{M}\sum_{t=1}^{T}\left\|\nabla_{(j)} f_i\left(y_i^{r,t}\right) - \nabla_{(j)} f_i\left(x^r\right)\right\|^2 + \sum_{j=1}^{N} 2T^2\mathbb{E}\left\|\frac{1}{K}\sum_{k=1}^{K}\nabla_{(j)} f_{k,j}\left(x^r\right)\right\|^2 + 4\frac{T}{K}\sigma^2 \\
&\leq 4\frac{T\beta^2}{M}\sum_{i=1}^{M}\sum_{t=1}^{T}\left\|y_i^{r,t} - x^r\right\|^2 + \sum_{j=1}^{N} 2T^2\mathbb{E}\left\|\frac{1}{K}\sum_{k=1}^{K}\nabla_{(j)} f_{k,j}\left(x^r\right)\right\|^2 + 4\frac{T}{K}\sigma^2 \\
&\leq 4\frac{\beta^2}{M}\sum_{i=1}^{M}\sum_{t=1}^{T}\left\|y_i^{r,t} - x^r\right\|^2 + \sum_{j=1}^{N} 2\mathbb{E}\left\|\frac{T}{K}\sum_{k=1}^{K}\nabla_{(j)} f_{k,j}\left(x^r\right)\right\|^2 + 4\frac{T}{K}\sigma^2 \\
&\leq 4\frac{T\beta^2}{M}\sum_{i=1}^{M}\sum_{t=1}^{T}\left\|y_i^{r,t} - x^r\right\|^2 + 2T^2\sum_{j=1}^{N}\mathbb{E}\left\|\frac{1}{K}\sum_{k=1}^{K}\nabla_{(j)} f_{k,j}\left(x^r\right) - \nabla_{(j)} f\left(x^r\right) + \nabla_{(j)} f\left(x^r\right)\right\|^2 + 4\frac{T}{K}\sigma^2 \\
&\overset{\text{b}}{\leq} 4\frac{T\beta^2}{M}\sum_{k=1}^{M}\sum_{t=1}^{T}\left\|y_i^{r,t} - x^r\right\|^2 + 2T^2\left\|\nabla f\left(x^r\right)\right\|^2 + 2\left(1-\frac{K}{M}\right)T^2\frac{1}{KM}\sum_{i=1}^{M}\left\|\nabla f_i\left(x^r\right)\right\|^2 + 4\frac{T}{K}\sigma^2 \\
&\overset{\text{c}}{\leq} 4\frac{T\beta^2}{M}\sum_{k=1}^{M}\sum_{t=1}^{T}\left\|y_i^{r,t} - x^r\right\|^2 + 4T^2\beta\left(f(x^r) - f\left(x^\star\right)\right) + 2\left(1-\frac{K}{M}\right)\frac{T^2}{K}\left(G^2 + 2\beta\left(f(x^r) - f\left(x^\star\right)\right)\right) + 4\frac{T}{K}\sigma^2,
\end{aligned}
\tag{48}
$$

Combining the bounds on C_1 and C_2 in the original inequality, we can get

$$
\begin{aligned}
&\mathbb{E}\left\|z^{r+1}-x^{\star}\right\|^2=\mathbb{E}\left\|z^{r}-x^{\star}\right\|^2+\eta\alpha\underbrace{\mathbb{E}\left\langle -\mathbf{G}^r, z^r-x^{\star}\right\rangle}_{C_1}+\eta^2\alpha^2\underbrace{\mathbb{E}\left\|\mathbf{G}^r\right\|^2}_{C_2}\\
&\leq \mathbb{E}\left\|z^{r}-x^{\star}\right\|^2+\alpha\eta T\left(-f\left(z^r\right)+f\left(x^{\star}\right)-\frac{\mu}{4}\left\|z^r-x^{\star}\right\|^2\right)+\frac{\alpha\eta\beta}{M}\sum_{k=1}^{M}\sum_{t=1}^{T}\left\|y_k^{r,t}-z^r\right\|^2\\
&+\alpha^2\eta^2\frac{4T\beta^2}{M}\sum_{k=1}^{M}\sum_{t=1}^{T}\left\|y_k^{r,t}-z^r\right\|^2+4T^2\beta\alpha^2\eta^2\left(f\left(z^r\right)-f\left(x^{\star}\right)\right)\\
&+2\left(1-\frac{K}{M}\right)T^2\alpha^2\eta^2\frac{1}{K}\left(G+2\beta\left(f\left(z^r\right)-f\left(x^{\star}\right)\right)\right)+4\alpha^2\eta^2\frac{T}{K}\sigma^2\\
&\overset{\text{a}}{\leq} \mathbb{E}\left\|z^{r}-x^{\star}\right\|^2+\alpha\eta T\left(-f\left(z^r\right)+f\left(x^{\star}\right)-\frac{\mu}{4}\left\|z^r-x^{\star}\right\|^2\right)\\
&+\left(\frac{\alpha\eta\beta}{M}+\frac{4T\beta^2\alpha^2\eta^2}{M}\right)\sum_{k=1}^{M}\sum_{t=1}^{T}\left\|y_k^{r,t}-z^r\right\|^2+2\left(1-\frac{K}{M}\right)T^2\alpha^2\eta^2\frac{1}{K}G\\
&+\left(4T^2\beta\alpha^2\eta^2+\left(1-\frac{K}{M}\right)4T^2\alpha^2\eta^2\frac{1}{K}\beta\right)\left(f\left(z^r\right)-f\left(x^{\star}\right)\right)+4\alpha^2\eta^2\frac{T}{K}\sigma^2\\
&\overset{\text{b}}{\leq} \mathbb{E}\left\|z^{r}-x^{\star}\right\|^2+\alpha\eta T\left(-f\left(z^r\right)+f\left(x^{\star}\right)-\frac{\mu}{4}\left\|z^r-x^{\star}\right\|^2\right)\\
&+\left(\frac{\alpha\eta\beta}{M}+\frac{4T\beta^2\alpha^2\eta^2}{M}\right)\left(6T^3\eta^2\sum_{i=1}^{M}\left\|\nabla f_i\left(z^r\right)\right\|^2+3MT^2\eta^2\sigma^2\right)\\
&+2\left(1-\frac{K}{M}\right)T^2\alpha^2\eta^2\frac{1}{K}G^2+\left(4T^2\beta\alpha^2\eta^2+\left(1-\frac{K}{M}\right)4T^2\alpha^2\eta^2\frac{1}{K}\beta\right)\left(f\left(z^r\right)-f\left(x^{\star}\right)\right)+4\alpha^2\eta^2\frac{T}{K}\sigma^2\\
&\overset{\text{c}}{\leq} \mathbb{E}\left\|z^{r}-x^{\star}\right\|^2+\alpha\eta T\left(-f\left(z^r\right)+f\left(x^{\star}\right)-\frac{\mu}{4}\left\|z^r-x^{\star}\right\|^2\right)+\left(\alpha\eta\beta+4T\beta^2\alpha^2\eta^2\right)6T^3\eta^2 2\beta\left(f\left(z^r\right)-f\left(x^{\star}\right)\right)\\
&+\left(\alpha\eta\beta+4T\beta^2\alpha^2\eta^2\right)6T^3\eta^2G^2+\left(\alpha\eta\beta+4T\beta^2\alpha^2\eta^2\right)3T^2\eta^2\sigma^2\\
&+2\left(1-\frac{K}{M}\right)T^2\alpha^2\eta^2\frac{1}{K}G^2+\left(4T^2\beta\alpha^2\eta^2+\left(1-\frac{K}{M}\right)4T^2\alpha^2\eta^2\frac{1}{K}\beta\right)\left(f\left(z^r\right)-f\left(x^{\star}\right)\right)+4\alpha^2\eta^2\frac{T}{K}\sigma^2\\
&\leq \mathbb{E}\left\|z^{r}-x^{\star}\right\|^2+\alpha\eta T\frac{\mu}{4}\left\|z^r-x^{\star}\right\|^2\\
&+\left[-\alpha\eta T+6T^3\alpha^2\eta^2 2\beta\left(\eta\beta+4T\beta^2\eta^2\right)+\left(4T^2\beta\alpha^2\eta^2+\left(1-\frac{K}{M}\right)4T^2\alpha^2\eta^2\frac{1}{K}\beta\right)\right]\left(f\left(z^r\right)-f\left(x^{\star}\right)\right)\\
&+2\left[\left(1-\frac{K}{M}\right)T^2\frac{1}{K}\right]\alpha^2\eta^2G^2+\left(6\alpha\beta T^3\alpha\eta^3+24\beta^2T^4\alpha^2\eta^4\right)G^2\\
&+4\alpha^2\eta^2\frac{T}{K}\sigma^2+\left(\alpha\eta\beta+4T\beta^2\alpha^2\eta^2\right)3T^2\eta^2\sigma^2,
\end{aligned}
\tag{49}
$$

where the inequality $\overset{\text{b}}{\leq}$ follows Lemma 5, the inequality $\overset{\text{c}}{\leq}$ holds due to Lemma 2. Next, we put $\left(f\left(x^r\right)-f\left(x^*\right)\right)$ term in left.

$$
\begin{aligned}
&\left[\alpha\eta T-6T^3\alpha^2\eta^2\beta\left(\eta\beta+4T\beta^2\eta^2\right)-\left(4T^2\beta\alpha^2\eta^2+\left(1-\frac{K}{M}\right)4T^2\alpha^2\eta^2\frac{1}{K}\beta\right)\right]\left(f\left(z^r\right)-f\left(x^{\star}\right)\right)\\
&\leq \mathbb{E}\left\|z^{r}-x^{\star}\right\|^2-\alpha\eta T\frac{\mu}{4}\left\|z^r-x^{\star}\right\|^2+2\left[\left(1-\frac{K}{M}\right)T^2\frac{1}{K}\right]\alpha^2\eta^2G^2+\left(6\alpha\beta T^3\alpha\eta^3+24\beta^2T^4\alpha^2\eta^4\right)G^2\\
&+4\alpha^2\eta^2\frac{T}{K}\sigma^2+\left(\alpha\eta\beta+4T\beta^2\alpha^2\eta^2\right)3T^2\eta^2\sigma^2.
\end{aligned}
\tag{50}
$$

$$
\begin{aligned}
&\left(f\left(z^{r}\right)-f\left(x^{\star}\right)\right) \\
&\overset{a}{\leq} \frac{\left(1-\mu \frac{\alpha \eta T}{4}\right)}{\frac{\alpha \eta T}{4}} \mathbb{E}\left\|z^{r}-x^{*}\right\|^{2}-\frac{4}{\alpha \eta T} \mathbb{E}\left\|z^{r+1}-x^{*}\right\|^{2} \\
&+2\left[\left(1-\frac{K}{M}\right) T^{2} \frac{1}{K}\right] \alpha^{2} \eta^{2} G^{2}+\left(6 \beta T^{3} \alpha \eta^{3}+24 \beta^{2} T^{4} \alpha^{2} \eta^{4}\right) G^{2}+4 \alpha^{2} \eta^{2} \frac{T}{K} \sigma^{2}+\left(\alpha \eta \beta+4 T \beta^{2} \alpha^{2} \eta^{2}\right) 3 T^{2} \eta^{2} \sigma^{2} \\
&\leq \frac{(1-\mu \tilde{\eta})}{\tilde{\eta}} \mathbb{E}\left\|z^{r}-x^{\star}\right\|^{2}-\frac{1}{\tilde{\eta}} \mathbb{E}\left\|z^{r+1}-x^{\star}\right\|^{2} \\
&+32\left[\left(1-\frac{K}{M}\right) \frac{1}{K}\right] \tilde{\eta} G^{2}+\frac{384 \beta \tilde{\eta}^{2} G^{2}}{\alpha^{2}}+\frac{6144 \beta^{2} \tilde{\eta}^{3} G^{2}}{\alpha^{2}}+64 \tilde{\eta} \frac{\sigma^{2}}{T K}+\frac{192 \beta}{T \alpha^{2}} \tilde{\eta}^{2} \sigma^{2}+\frac{3072}{T \alpha^{2}} \beta^{2} \tilde{\eta}^{3} \sigma^{2} \\
&\leq \frac{(1-\mu \tilde{\eta})}{\tilde{\eta}} \mathbb{E}\left\|z^{r}-x^{\star}\right\|^{2}-\frac{1}{\tilde{\eta}} \mathbb{E}\left\|z^{r+1}-x^{\star}\right\|^{2}+\left[32\left[\left(1-\frac{K}{M}\right) \frac{1}{K}\right] G^{2}+64 \frac{\sigma^{2}}{T K}\right] \tilde{\eta} \\
&+\left(\frac{384}{\alpha^{2}} \beta G^{2}+\frac{192 \beta}{T \alpha^{2}} \sigma^{2}\right) \tilde{\eta}^{2}+\left(\frac{6144}{\alpha^{2}} \beta^{2} G^{2}+\frac{3702}{T \alpha^{2}} \beta^{2} \sigma^{2}\right) \tilde{\eta}^{3},
\end{aligned} \tag{51}
$$

where the inequality $\overset{a}{\leq}$ follows $\left[\alpha\eta T-6T^3\alpha^2\eta^2 2\beta\left(\eta\beta+4T\beta^2\eta^2\right)-\left(4T^2\beta\alpha^2\eta^2+\left(1-\frac{K}{M}\right)4T^2\alpha^2\eta^2\frac{1}{K}\beta\right)\right]\geq\frac{1}{4}\alpha\eta T$. In the last inequalities, we let $\tilde{\eta}=\frac{\alpha\eta T}{4}$,$\tilde{\eta}\leq\frac{1}{8\beta}$. With Lemma 7, we can get

$$
\begin{aligned}
\mathbb{E}\left[f\left(\bar{z}^{R}\right)\right]-f\left(x^{\star}\right) \leq &\left\|x^{0}-x^{\star}\right\|^{2} \mu \exp \left(-\frac{\alpha \mu R}{\beta}\right)+\frac{32\left[\left(1-\frac{K}{M}\right) \frac{1}{K}\right] G^{2}+64 \frac{\sigma^{2}}{K T}}{\mu R} \\
&+\frac{\left(384 \beta G^{2}+\frac{192 \beta}{T} \sigma^{2}\right)}{\alpha^{2} \mu^{2} R^{2}}+\frac{\left(6144 \beta^{2} G^{2}+\frac{3072}{T} \beta^{2} \sigma^{2}\right)}{\alpha^{2} \mu^{3} R^{3}} .
\end{aligned} \tag{52}
$$

4.3 2. The convergence rate of general convex and smooth case:

For general convex case, we have $\mu=0$, then the following inequality holds:

$$
\begin{aligned}
&\left(f\left(z^{r}\right)-f\left(x^{\star}\right)\right) \\
&\leq \frac{1}{\tilde{\eta}} \mathbb{E}\left\|z^{r}-x^{\star}\right\|^{2}-\frac{1}{\tilde{\eta}} \mathbb{E}\left\|z^{r+1}-x^{\star}\right\|^{2}+\left[32\left[\left(1-\frac{K}{M}\right) \frac{1}{K}\right] G^{2}+64 \frac{\sigma^{2}}{K T}\right] \tilde{\eta} \\
&+\left(\frac{384}{\alpha^{2}} \beta G^{2}+\frac{192 \beta}{T \alpha^{2}} \sigma^{2}\right) \tilde{\eta}^{2}+\left(\frac{6144}{\alpha^{2}} \beta^{2} G^{2}+\frac{3072}{T \alpha^{2}} \beta^{2} \sigma^{2}\right) \tilde{\eta}^{3} .
\end{aligned} \tag{53}
$$

With Lemma 8, $\tilde{\eta}\leq\frac{1}{(192T\beta^3+64\beta)}\leq\frac{1}{64\beta}$, we can get,

$$
\begin{aligned}
&\mathbb{E}\left[f\left(\bar{z}^{R}\right)\right]-f\left(x^{\star}\right) \\
&\leq \frac{\beta^{\frac{3}{2}} d_{0}}{\alpha R}+\frac{\left(6144 \beta^{2} G^{2}+\frac{3072}{T} \beta^{2} \sigma^{2}\right)}{\alpha^{2} R}+\frac{\left[32\left[\left(1-\frac{K}{M}\right) \frac{1}{K}\right] G^{2}+64 \frac{\sigma^{2}}{K T}\right]^{\frac{1}{2}} d_{0}^{\frac{1}{2}}}{\sqrt{R}} \\
&+\frac{\left(\frac{384}{\alpha^{2}} \beta G^{2}+\frac{192 \beta}{T \alpha^{2}} \sigma^{2}\right)^{\frac{1}{3}} d_{0}^{\frac{2}{3}}}{\alpha^{\frac{2}{3}} R^{\frac{2}{3}}} .
\end{aligned} \tag{54}
$$

4.4 3. The convergence rate of non-convex and smooth case:

From the smoothness of the function, we can obtain,

$$
\begin{aligned}
\mathbb{E} f\left(z^{r+1}\right) &\leq \mathbb{E} f\left(z^{r}\right)+\mathbb{E}\left\langle\nabla f\left(z^{r}\right), z^{r+1}-z^{r}\right\rangle+\frac{\beta}{2} \mathbb{E}\left\|z^{r+1}-z^{r}\right\|^{2} \\
&\leq \mathbb{E} f\left(z^{r}\right)+\alpha \eta \underbrace{\mathbb{E}\left\langle\nabla f\left(z^{r}\right),-\mathrm{G}^{r}\right\rangle}_{D_{1}}+\frac{\beta}{2} \eta^{2} \alpha^{2} \underbrace{\mathbb{E}\left\|\mathrm{G}^{r}\right\|^{2}}_{D_{2}} .
\end{aligned} \tag{55}
$$

Next we will perform an upper bound analysis on D_1,

$$\begin{aligned}
D_1 &= -\mathbb{E}\left\langle \nabla f\left(z^r\right), \mathbf{G}^r\right\rangle \\
&= -\mathbb{E}\left\langle \nabla f\left(z^r\right) - \nabla f\left(x^r\right), \mathbf{G}^r\right\rangle - \mathbb{E}\left\langle \nabla f\left(x^r\right), \mathbf{G}^r\right\rangle \\
&\leq \frac{1}{2a}\mathbb{E}\left\|\nabla f\left(z^r\right) - \nabla f\left(x^r\right)\right\|^2 + \frac{a}{2}\left\|\mathbb{E}\left[\mathbf{G}^r\right]\right\|^2 - \frac{1}{T}\left\langle T\nabla f\left(x^r\right), \mathbb{E}\left[\mathbf{G}^r\right]\right\rangle \\
&\leq \frac{1}{2a}\mathbb{E}\left\|\nabla f\left(z^r\right) - \nabla f\left(x^r\right)\right\|^2 - \left(\frac{1}{4T} - \frac{a}{2}\right)\left\|\mathbb{E}\left[\mathbf{G}^r\right]\right\|^2 - \frac{1}{2}T\left\|\nabla f\left(x^r\right)\right\|^2 \\
&+ \frac{1}{2T}\left\|T\nabla f\left(x^r\right) - \frac{1}{M}\sum_{i=1}^{M}\sum_{t=1}^{T}\nabla_{(j)} f_i\left(y_i^{r,t}\right)\right\|^2 \\
&\leq \frac{\beta^2}{2a}\mathbb{E}\left\|z^r - x^r\right\|^2 - \left(\frac{1}{4T} - \frac{a}{2}\right)\left\|\mathbb{E}\left[\mathbf{G}^r\right]\right\|^2 - \frac{1}{2}T\left\|\nabla f\left(x^r\right)\right\|^2 + \frac{\beta^2}{2M}\sum_{i=1}^{M}\sum_{t=1}^{T}\left\|x^r - y_i^{r,t}\right\|^2.
\end{aligned} \tag{56}$$

Next we will find the upper bound constraint on $\mathbb{E}\left\|z^r - x^r\right\|^2$

$$\begin{aligned}
& z^r = x^r + \gamma\left(x^r - x^{r-1}\right) \\
& \left\|z^r - x^r\right\|^2 \leq \gamma^2\left\|x^r - x^{r-1}\right\|^2 \\
& \left\|x^r - x^{r-1}\right\|^2 = \gamma^2\left\|\eta^2\mathbf{G}^r + \beta v^{r-1}\right\|^2 \\
& \leq \eta^2\gamma^2\underbrace{\left\|\sum_{s=0}^{r}\beta^{r-s}\mathbf{G}^s\right\|^2}_{T_1}.
\end{aligned} \tag{57}$$

For the first term T_1, taking the total expectation, we get

$$\begin{aligned}
\mathbb{E}\left[T_1\right] &\leq \left(\sum_{s=0}^{r}\beta^{r-s}\right)\sum_{s=0}^{r}\beta^{r-s}\mathbb{E}\left[\left\|\mathbf{G}^s\right\|^2\right] \\
&\leq \left(\sum_{s=0}^{r}\beta^{r-s}\right)\sum_{s=0}^{r}\beta^{r-s}\mathbb{E}\left[\left\|\mathbf{G}^s\right\|^2\right] \\
&\leq \frac{1}{1-\beta}\sum_{s=0}^{r}\beta^{r-s}\mathbb{E}\left[\left\|\mathbf{G}^s\right\|^2\right].
\end{aligned} \tag{58}$$

Finally, we can get,

$$\mathbb{E}\left\|z^r - x^r\right\|^2 \leq \frac{\eta^2\gamma^2}{1-\beta}\sum_{s=0}^{r}\beta^{r-s}\mathbb{E}\left[\left\|\mathbf{G}^s\right\|^2\right], \tag{59}$$

and

$$\begin{aligned}
& \sum_{r=1}^{R}\left\|\sum_{s=0}^{r}\beta^{r-s}\mathbf{G}^s\right\|^2 \leq \frac{1}{1-\beta}\sum_{r=1}^{R}\mathbb{E}\left[\left\|\mathbf{G}^r\right\|^2\right]\sum_{s=0}^{r}\beta^{R-s} \\
& \leq \frac{1}{(1-\beta)^2}\sum_{r=1}^{R}\mathbb{E}\left[\left\|\mathbf{G}^r\right\|^2\right].
\end{aligned} \tag{60}$$

Next we will perform an upper bound analysis on D_2,

$$
\begin{aligned}
D_2 &= \mathbb{E}\left\|\mathbf{G}^r\right\|^2 = \sum_{j=1}^{N} \mathbb{E}\left\|\frac{1}{K}\sum_{k=1}^{K}\sum_{t=1}^{T} \nabla_{(j)} f_{k,j}\left(y_{k,j}^{r,t};\zeta\right)\right\|^2 \\
&\leq \frac{1}{M^2}\mathbb{E}\left\|\sum_{i=1}^{M}\sum_{t=1}^{T}\nabla f_i\left(y_i^{r,t};\zeta\right)\right\|^2 + \mathbb{E}\frac{1}{KM}\left(1-\frac{K}{M}\right)\sum_{i=1}^{M}\left\|\sum_{t=1}^{T}\nabla f_i\left(y_i^{r,t};\zeta\right)\right\|^2 \\
&\leq \frac{1}{M^2}\mathbb{E}\left\|\sum_{i=1}^{M}\sum_{t=1}^{T}\nabla f_i\left(y_i^{r,t}\right)\right\|^2 + \mathbb{E}\frac{1}{KM}\left(1-\frac{K}{M}\right)\left[3T\beta^2\sum_{i=1}^{M}\sum_{t=1}^{T}\left\|y_i^{r,t}-x\right\|^2 + MT^2G^2 + MT^2\|\nabla f(x)\|^2\right] \\
&+ \frac{T}{M}\sigma^2 + \frac{T}{K}\left(1-\frac{K}{M}\right)\sigma^2.
\end{aligned} \tag{61}
$$

Combining the bounds on D_1, D_2 in the original inequality, we can get

$$
\begin{aligned}
\mathbb{E}f\left(z^{r+1}\right) &\leq \mathbb{E}f\left(z^r\right) + \frac{\alpha\eta\beta^2}{2a}\mathbb{E}\left\|z^r - x^r\right\|^2 - \alpha\eta\left(\frac{1}{4T}-\frac{a}{2}\right)\left\|\mathbb{E}\left[\mathbf{G}^r\right]\right\|^2 - \frac{\alpha\eta T}{2}\left\|\nabla f\left(x^r\right)\right\|^2 \\
&+ \frac{\alpha\eta\beta^2}{2M}\sum_{i=1}^{M}\sum_{t=1}^{T}\left\|y_i^{r,t}-x\right\|^2 + \frac{\beta}{2}\eta^2\alpha^2\mathbb{E}\left\|\mathbf{G}^r\right\|^2.
\end{aligned} \tag{62}
$$

Summing the left and right sides of the above inequality from 1 to R simultaneously, we have

$$
\begin{aligned}
\mathbb{E}f\left(z^{R+1}\right) &\leq \mathbb{E}f\left(z^0\right) + \frac{\alpha\eta\beta^2}{2a}\sum_{r=1}^{R}\mathbb{E}\left\|z^r-x^r\right\|^2 - \alpha\eta\left(\frac{1}{4T}-\frac{a}{2}\right)\sum_{r=1}^{R}\left\|\mathbb{E}\left[\mathbf{G}^r\right]\right\|^2 - \frac{\alpha\eta T}{2}\sum_{r=1}^{R}\left\|\nabla f\left(x^r\right)\right\|^2 \\
&+ \frac{\alpha\eta\beta^2}{2M}\sum_{r=1}^{R}\sum_{i=1}^{M}\sum_{t=1}^{T}\left\|y_i^{r,t}-x\right\|^2 + \frac{\beta}{2}\eta^2\alpha^2\sum_{r=1}^{R}\mathbb{E}\left\|\mathbf{G}^r\right\|^2 \\
&\leq \mathbb{E}f\left(z^0\right) + \frac{\alpha\eta\beta^2}{2a}\frac{\eta^2\gamma^2}{(1-\beta)^2}\sum_{r=1}^{R}\mathbb{E}\left[\left\|\mathbf{G}^r\right\|^2\right] - \alpha\eta\left(\frac{1}{4T}-\frac{a}{2}\right)\sum_{r=1}^{R}\left\|\mathbb{E}\left[\mathbf{G}^r\right]\right\|^2 - \frac{\alpha\eta T}{2}\sum_{r=1}^{R}\left\|\nabla f\left(x^r\right)\right\|^2 \\
&+ \frac{\alpha\eta\beta^2}{2M}\sum_{r=1}^{R}\sum_{i=1}^{M}\sum_{t=1}^{T}\left\|y_i^{r,t}-x\right\|^2 + \frac{\beta}{2}\eta^2\alpha^2\sum_{r=1}^{R}\mathbb{E}\left\|\mathbf{G}^r\right\|^2 \\
&\leq \mathbb{E}f\left(z^0\right) + \left(\frac{\alpha\eta\beta^2}{2a}\frac{\eta^2\gamma^2}{(1-\beta)^2} + \frac{\beta}{2}\eta^2\alpha^2\right)\sum_{r=1}^{R}\left[\frac{1}{M^2}\mathbb{E}\left\|\sum_{i=1}^{M}\sum_{t=1}^{T}\nabla f_i\left(y_i^{r,t}\right)\right\|\right. \\
&\left.+\mathbb{E}\frac{1}{KM}\left(1-\frac{K}{M}\right)\left[3T\beta^2\sum_{i=1}^{M}\sum_{t=1}^{T}\left\|y_i^{r,t}-x\right\|^2 + MT^2G^2 + MT^2\|\nabla f(x)\|^2\right] + \frac{T}{M}\sigma^2 + \left(1-\frac{K}{M}\right)\frac{T}{K}\sigma^2\right] \\
&- \alpha\eta\left(\frac{1}{4T}-\frac{a}{2}\right)\sum_{r=1}^{R}\left\|\mathbb{E}\left[\mathbf{G}^r\right]\right\|^2 - \frac{\alpha\eta T}{2}\sum_{r=1}^{R}\left\|\nabla f\left(x^r\right)\right\|^2 + \frac{\alpha\eta\beta^2}{2M}\sum_{r=1}^{R}\sum_{i=1}^{M}\sum_{t=1}^{T}\left\|y_i^{r,t}-x\right\|^2 \\
&\leq \mathbb{E}f\left(z^0\right) + \sum_{r=1}^{R}\left[\frac{C_1}{M^2}\mathbb{E}\left\|\sum_{i=1}^{M}\sum_{t=1}^{T}\nabla f_i\left(y_i^{r,t}\right)\right\|^2\right. \\
&\left.+\mathbb{E}\frac{C_1}{KM}\left(1-\frac{K}{M}\right)\left[3T\beta^2\sum_{i=1}^{M}\sum_{t=1}^{T}\left\|y_i^{r,r}-x\right\|^2 + MT^2G^2 + MT^2\|\nabla f(x)\|^2\right] + C_1\frac{T}{M}\sigma^2 + C_1\left(1-\frac{K}{M}\right)\frac{T\sigma^2}{K}\right] \\
&- \alpha\eta\left(\frac{1}{4T}-\frac{a}{2}\right)\sum_{r=1}^{R}\left\|\mathbb{E}\left[\mathbf{G}^r\right]\right\|^2 - \frac{\alpha\eta T}{2}\sum_{r=1}^{R}\left\|\nabla f\left(x^r\right)\right\|^2 + \frac{\alpha\eta\beta^2}{2M}\sum_{r=1}^{R}\sum_{i=1}^{M}\sum_{t=1}^{T}\left\|y_i^{r,t}-x\right\|^2,
\end{aligned} \tag{63}
$$

let

$$
\left(\frac{\alpha\eta\beta^2}{2a}\frac{\eta^2\gamma^2}{(1-\beta)^2} + \frac{\beta}{2}\eta^2\alpha^2\right) = C_1, \tag{64}
$$

$$
\begin{aligned}
&\leq \mathbb{E}f\left(z^{0}\right)+\left(\frac{\alpha\eta\beta^{2}}{2M}+\frac{3T\beta^{2}C_{1}}{KM}\left(1-\frac{K}{M}\right)\right)\sum_{r=1}^{R}\sum_{i=1}^{M}\sum_{t=1}^{T}\mathbb{E}\left\|y_{i}^{r,t}-x\right\|^{2}\\
&+\frac{RC_{1}T^{2}G^{2}}{K}\left(1-\frac{K}{M}\right)+\left(\left(1-\frac{K}{M}\right)\frac{C_{1}T^{2}}{K}-\frac{\alpha\eta T}{2}\right)\sum_{r=1}^{R}\left\|\nabla f\left(x^{r}\right)\right\|^{2}\\
&+\left(\frac{C_{1}}{M^{2}}-\alpha\eta\left(\frac{1}{4T}-\frac{a}{2}\right)\right)\sum_{r=1}^{R}\left\|\mathbb{E}\left[\mathbf{G}^{r}\right]\right\|^{2}+C_{1}\frac{TR}{M}\sigma^{2}+C_{1}\frac{TR\sigma^{2}}{K}\left(1-\frac{K}{M}\right)\\
&\leq \mathbb{E}f\left(z^{0}\right)+\left(\frac{\alpha\eta\beta^{2}}{2}+\frac{3T\beta^{2}C_{1}}{K}\left(1-\frac{K}{M}\right)\right)\sum_{r=1}^{R}\left[6T^{3}\eta^{2}\frac{1}{M}\sum_{i=1}^{M}\left\|\nabla f_{i}\left(x^{r}\right)\right\|^{2}+3T^{2}\eta^{2}\sigma^{2}\right]\\
&+\frac{RC_{1}T^{2}G^{2}}{K}\left(1-\frac{K}{M}\right)+\left(\frac{C_{1}T^{2}}{K}\left(1-\frac{K}{M}\right)-\frac{\alpha\eta T}{2}\right)\sum_{r=1}^{R}\left\|\nabla f\left(x^{r}\right)\right\|^{2}\\
&+\left(\frac{C_{1}}{M^{2}}-\alpha\eta\left(\frac{1}{4T}-\frac{a}{2}\right)\right)\sum_{r=1}^{R}\left\|\mathbb{E}\left[\mathbf{G}^{r}\right]\right\|^{2}+C_{1}\frac{TR}{M}\sigma^{2}+C_{1}\frac{TR\sigma^{2}}{K}\\
&\leq \mathbb{E}f\left(z^{0}\right)+\left(\frac{\alpha\eta\beta^{2}}{2}+\frac{3T\beta^{2}C_{1}}{K}\left(1-\frac{K}{M}\right)\right)\sum_{r=1}^{R}\left[12T^{3}\eta^{2}G^{2}+12T^{3}\eta^{2}\left\|\nabla f\left(x^{r}\right)\right\|^{2}+3T^{2}\eta^{2}\sigma^{2}\right]\\
&+\frac{RC_{1}T^{2}G^{2}}{K}\left(1-\frac{K}{M}\right)+\left(\frac{C_{1}T^{2}}{K}\left(1-\frac{K}{M}\right)-\frac{\alpha\eta T}{2}\right)\sum_{r=1}^{R}\left\|\nabla f\left(x^{r}\right)\right\|^{2}\\
&+\left(\frac{C_{1}}{M^{2}}-\alpha\eta\left(\frac{1}{4T}-\frac{a}{2}\right)\right)\sum_{r=1}^{R}\left\|\mathbb{E}\left[\mathbf{G}^{r}\right]\right\|^{2}+C_{1}\frac{TR}{M}\sigma^{2}+C_{1}\frac{TR\sigma^{2}}{K}\left(1-\frac{K}{M}\right).
\end{aligned}
\tag{65}
$$

Moving $\|\nabla f(x^r)\|^2$ to the left, we can obtain

$$
\begin{aligned}
&\left(\frac{\alpha\eta T}{2}-\frac{C_{1}T^{2}}{K}-12T^{3}\eta^{2}\left(\frac{\alpha\eta\beta^{2}}{2}+\frac{3T\beta^{2}C_{1}}{K}\left(1-\frac{K}{M}\right)\right)\right)\sum_{r=1}^{R}\left\|\nabla f\left(x^{r}\right)\right\|^{2}\\
&\leq \mathbb{E}f\left(z^{0}\right)+12T^{3}\eta^{2}G^{2}\left(\frac{\alpha\eta\beta^{2}}{2}+\frac{3T\beta^{2}C_{1}}{K}\left(1-\frac{K}{M}\right)\right)R+\frac{RC_{1}T^{2}G^{2}}{K}\left(1-\frac{K}{M}\right)\\
&+3T^{2}\eta^{2}\sigma^{2}\left(\frac{\alpha\eta\beta^{2}}{2}+\frac{3T\beta^{2}C_{1}}{K}\left(1-\frac{K}{M}\right)\right)R+C_{1}\frac{TR}{M}\sigma^{2}+C_{1}\frac{TR\sigma^{2}}{K}\left(1-\frac{K}{M}\right).
\end{aligned}
\tag{66}
$$

Let $\left(\frac{\alpha\eta T}{2}-\frac{C_1T^2}{K}-12T^3\eta^2\left(\frac{\alpha\eta\beta^2}{2}+\frac{3T\beta^2C_1}{K}\left(1-\frac{K}{M}\right)\right)\right)\leq\frac{\alpha\eta T}{4}$,$\tilde{\eta}=\frac{1}{4}\alpha\eta T$,$\tilde{\eta}\leq\frac{1}{16\beta}$, we can get,

$$
\begin{aligned}
&\tilde{\eta}\frac{1}{R}\sum_{r=1}^{R}\left\|\nabla f\left(x^{r}\right)\right\|^{2}\leq\frac{\mathbb{E}f\left(z^{0}\right)}{R}+12T^{3}\eta^{2}G^{2}\left(\frac{\alpha\eta\beta^{2}}{2}+\frac{3T\beta^{2}C_{1}}{K}\left(1-\frac{K}{M}\right)\right)+\frac{C_{1}T^{2}G^{2}}{K}\left(1-\frac{K}{M}\right)\\
&+3T^{2}\eta^{2}\sigma^{2}\left(\frac{\alpha\eta\beta^{2}}{2}+\frac{3T\beta^{2}C_{1}}{K}\left(1-\frac{K}{M}\right)\right)+C_{1}\frac{T}{M}\sigma^{2}+C_{1}\frac{T\sigma^{2}}{K}\left(1-\frac{K}{M}\right).
\end{aligned}
\tag{67}
$$

Moving $\tilde{\eta}$ to the left, we can obtain

$$\begin{aligned}
&\frac{1}{R}\sum_{r=1}^{R}\left\|\nabla f\left(x^{r}\right)\right\|^{2}\\
&\leq \frac{\mathbb{E}f\left(z^{0}\right)}{\tilde{\eta}R}+\frac{192T\tilde{\eta}}{\alpha^{2}}G^{2}\left(\frac{2\tilde{\eta}\beta^{2}}{T}+\frac{3T\beta^{2}}{K}\left(16\frac{\tilde{\eta}^{3}\gamma^{2}\beta^{2}}{T^{2}}+8\frac{\beta\tilde{\eta}^{2}}{T^{2}}\right)\left(1-\frac{K}{M}\right)\right)\\
&+\frac{\left(16\tilde{\eta}^{2}\gamma^{2}\beta^{2}+8\beta\tilde{\eta}\right)G^{2}}{K}\left(1-\frac{K}{M}\right)\\
&+\frac{48\tilde{\eta}\sigma^{2}}{\alpha^{2}}\left(\frac{2\tilde{\eta}\beta^{2}}{T}+\frac{3T\beta^{2}}{K}\left(16\frac{\tilde{\eta}^{3}\gamma^{2}\beta^{2}}{T^{2}}+8\frac{\beta\tilde{\eta}^{2}}{T^{2}}\right)\left(1-\frac{K}{M}\right)\right)\\
&+\left(16\frac{\tilde{\eta}^{2}\gamma^{2}\beta^{2}}{T^{2}}+8\frac{\beta\tilde{\eta}}{T^{2}}\right)\frac{T}{M}\sigma^{2}+\left(16\frac{\tilde{\eta}^{2}\gamma^{2}\beta^{2}}{T^{2}}+8\frac{\beta\tilde{\eta}}{T^{2}}\right)\left(1-\frac{K}{M}\right)\frac{T\sigma^{2}}{K}\\
&\leq \frac{\mathbb{E}f\left(z^{0}\right)}{\tilde{\eta}R}+G^{2}\left(\frac{384\beta^{2}\tilde{\eta}^{2}}{\alpha^{2}}+\frac{9216}{\alpha^{2}}\frac{\gamma^{2}\beta^{4}\tilde{\eta}^{4}}{K}\left(1-\frac{K}{M}\right)\frac{4608}{\alpha^{2}}\frac{\beta^{3}\tilde{\eta}^{3}}{K}\left(1-\frac{K}{M}\right)\right)\\
&+\frac{\left(16\tilde{\eta}^{2}\gamma^{2}\beta^{2}+8\beta\tilde{\eta}\right)G^{2}}{K}\left(1-\frac{K}{M}\right)\\
&+\left(\frac{92\beta^{2}}{T}\frac{\tilde{\eta}^{2}\sigma^{2}}{\alpha^{2}}+\frac{2304\beta^{4}}{K}\frac{\tilde{\eta}^{4}\gamma^{2}\sigma^{2}}{\alpha^{2}T}\left(1-\frac{K}{M}\right)+\frac{1152\beta^{3}}{K}\frac{\tilde{\eta}^{3}\sigma^{2}}{T\alpha^{2}}\left(1-\frac{K}{M}\right)\right)\\
&+\frac{\left(16\tilde{\eta}^{2}\gamma^{2}\beta^{2}+8\beta\tilde{\eta}\right)}{TM}\sigma^{2}+\frac{\left(16\tilde{\eta}^{2}\gamma^{2}\beta^{2}+8\beta\tilde{\eta}\right)\sigma^{2}}{TK}\left(1-\frac{K}{M}\right)\\
&\leq \frac{\mathbb{E}f\left(z^{0}\right)}{\tilde{\eta}R}+\frac{8\beta}{TM}\sigma^{2}\tilde{\eta}+\frac{8\beta}{TK}\left(1-\frac{K}{M}\right)\sigma^{2}\tilde{\eta}+\frac{8\beta}{K}\left(1-\frac{K}{M}\right)G^{2}\tilde{\eta}\\
&+\frac{384\beta^{2}}{\alpha^{2}}G^{2}\tilde{\eta}^{2}+\frac{92\beta^{2}}{\alpha^{2}T}\sigma^{2}\tilde{\eta}^{2}+\frac{\left(16\gamma^{2}\beta^{2}\right)}{TM}\sigma^{2}\tilde{\eta}^{2}+\frac{16\gamma^{2}\beta^{2}}{TK}\left(1-\frac{K}{M}\right)\sigma^{2}\tilde{\eta}^{2}+\frac{16\gamma^{2}\beta^{2}}{K}\left(1-\frac{K}{M}\right)G^{2}\tilde{\eta}^{2}\\
&+\frac{4608}{\alpha^{2}}\frac{\beta^{3}}{K}\left(1-\frac{K}{M}\right)G^{2}\tilde{\eta}^{3}+\frac{1152\beta^{3}}{KT\alpha^{2}}\left(1-\frac{K}{M}\right)\sigma^{2}\tilde{\eta}^{3}\\
&+\frac{9216}{\alpha^{2}}\frac{\gamma^{2}\beta^{4}}{K}\left(1-\frac{K}{M}\right)G^{2}\tilde{\eta}^{4}+\frac{2304\beta^{4}}{K}\frac{\gamma^{2}}{\alpha^{2}T}\left(1-\frac{K}{M}\right)\sigma^{2}\tilde{\eta}^{4}.
\end{aligned}\tag{68}$$

With Lemma 8, $\tilde{\eta}\leq\frac{1}{16\beta}$, we can get

$$\begin{aligned}
&\frac{1}{R}\sum_{r=1}^{R}\left\|\nabla f\left(x^{r}\right)\right\|^{2}\leq\frac{16\beta d_{0}}{TK\alpha R}+\frac{2\sqrt{d_{0}}}{\sqrt{RTM}}\left(\frac{8\beta}{K}\left(1-\frac{K}{M}\right)G^{2}+\frac{8\beta\sigma^{2}}{TK}\left(1-\frac{K}{M}\right)+\frac{8\beta}{TM}\sigma^{2}\right)^{\frac{1}{2}}\\
&+2\left(\frac{d_{0}}{R}\right)^{\frac{2}{3}}\left[\frac{384\beta^{2}}{\alpha^{2}}G^{2}+\frac{92\beta^{2}}{T}\frac{\sigma^{2}}{\alpha^{2}}+\frac{\left(16\gamma^{2}\beta^{2}\right)}{TM}\sigma^{2}+\frac{\left(16\gamma^{2}\beta^{2}\right)\sigma^{2}}{TK}\left(1-\frac{K}{M}\right)+\frac{16\gamma^{2}\beta^{2}}{K}\left(1-\frac{K}{M}\right)G^{2}\right]^{\frac{1}{3}}\\
&+2\left(\frac{d_{0}}{R}\right)^{\frac{3}{4}}\left[\frac{4608}{\alpha^{2}}\frac{\beta^{3}}{K}\left(1-\frac{K}{M}\right)G^{2}+\frac{1152\beta^{3}}{KT\alpha^{2}}\left(1-\frac{K}{M}\right)\sigma^{2}\right]^{\frac{1}{4}}\\
&+2\left(\frac{d_{0}}{R}\right)^{\frac{4}{5}}\left[\frac{9216}{\alpha^{2}}\frac{\gamma^{2}\beta^{4}}{K}\left(1-\frac{K}{M}\right)G^{2}+\frac{2304\beta^{4}}{K}\frac{\gamma^{2}}{\alpha^{2}T}\left(1-\frac{K}{M}\right)\sigma^{2}\right]^{\frac{1}{5}}.
\end{aligned}\tag{69}$$

5 APPENDIX E: PROOF OF THEOREM 2.

5.1 1. The rate of strongly convex and smooth convergence:

We update the local control variates only for clients $i\in\mathcal{S}^{r}$

$$\boldsymbol{c}_{i}^{r}=\begin{cases}\tilde{\boldsymbol{c}}_{i}^{r} & \text{if } i\in\mathcal{S}^{r}\\ \boldsymbol{c}_{i}^{r-1} & \text{otherwise}\end{cases}.\tag{70}$$

Compute the new global parameters and global control variate using only updates from the clients $i\in\mathcal{K}_{j}^{r}$:

$$\boldsymbol{c}_{(j)}^{r}=\boldsymbol{c}_{(j)}^{r}+\frac{K}{M}\sum_{k=1}^{K}\Delta\boldsymbol{c}_{k,j,(j)}^{r},\tag{71}$$

$$\boldsymbol{c}_{(j)}^{r}=\frac{1}{M}\sum_{i=1}^{M}\boldsymbol{c}_{i,(j)}^{r}=\frac{1}{M}\left(\sum_{i\in\mathcal{K}_{j}^{r}}\boldsymbol{c}_{i,(j)}^{r}+\sum_{i\notin\mathcal{K}_{j}^{r}}\boldsymbol{c}_{i,(j)}^{r-1}\right),\tag{72}$$

$$\boldsymbol{c}^{r}=\left[\boldsymbol{c}_{(1)}^{r\top},\ldots,\boldsymbol{c}_{(N)}^{r\top}\right]^{\top}.\tag{73}$$

We define client-drift to be how much the clients move from their starting point:

$$\mathcal{E}_{r}=\frac{1}{MT}\sum_{i=1}^{M}\sum_{t=1}^{T}\left(\left\|y_{i}^{r,t}-z^{r}\right\|^{2}\right).\tag{74}$$

Because we are sampling the clients, not all the client control-variates get updated every round. This leads to some 'lag' which we call control-lag:

$$C_{r}=\frac{1}{M}\sum_{i=1}^{M}\left\|\mathbb{E}[c_{i}^{r}]-\nabla f_{i}\left(z^{r}\right)\right\|^{2}.\tag{75}$$

With Lemma 10, we have

$$\mathbb{E}\left\|z^{r+1}-x^{\star}\right\|^{2}\leq\mathbb{E}\left\|z^{r}-x^{\star}\right\|^{2}+2\eta\alpha\underbrace{\mathbb{E}\left\langle-\mathrm{G}^{r},z^{r}-x^{\star}\right\rangle}_{E_{1}}+\eta^{2}\alpha^{2}\underbrace{\mathbb{E}\left\|\mathrm{G}^{r}\right\|^{2}}_{E_{2}}.\tag{76}$$

Before giving the convergence analysis of Theorem 1, we first present the following lemma.

Lemma 11. *We can get the bound of E_2*

$$\mathbb{E}\left\|\mathrm{G}^{r}\right\|^{2}\leq\left(\frac{4T^{2}}{MT}\right)\sum_{i=1}^{M}\sum_{t=1}^{T}\mathbb{E}\left\|y_{i}^{r,t}-z^{r}\right\|^{2}+\left(\frac{8T^{2}}{M}\right)\sum_{i=1}^{M}\left\|\mathbb{E}c_{i}^{r}-\nabla f_{i}\left(x^{r}\right)\right\|^{2}+\left(\frac{4T^{2}}{M}\right)2\beta\left(f\left(z^{r}\right)-f^{\star}\right).\tag{77}$$

Proof.

$$\begin{aligned}
E_{2}&=\mathbb{E}\left\|\mathrm{G}^{r}\right\|^{2}\\
&=\sum_{j=1}^{N}\mathbb{E}\left\|\frac{1}{K}\sum_{k=1}^{K}\sum_{t=1}^{T}\nabla_{(j)}f_{k,j}\left(y_{k,j}^{r,t};\zeta\right)+c_{(j)}^{r}-c_{k,j,(j)}^{r}+\nabla_{(j)}f_{k,j}\left(x^{r}\right)-\nabla_{(j)}f_{k,j}\left(x^{r};\zeta\right)\right\|^{2}\\
&=\sum_{j=1}^{N}\mathbb{E}\left\|\frac{1}{K}\sum_{k=1}^{K}\sum_{t=1}^{T}\nabla_{(j)}f_{k,j}\left(y_{k,j}^{r,t};\zeta\right)+c_{(j)}^{r}-c_{k,j,(j)}^{r}+\nabla_{(j)}f_{k,j}\left(x^{r}\right)-\nabla_{(j)}f_{k,j}\left(x^{r};\zeta\right)\right\|^{2}\\
&\leq\frac{T}{M}\sum_{i=1}^{M}\sum_{t=1}^{T}\mathbb{E}\left\|\nabla f_{i}\left(y_{i}^{r,t};\zeta\right)-\nabla f_{i}\left(z^{r};\zeta\right)+c^{r}-c_{i}^{r}+\nabla f_{i}\left(z^{r}\right)\right\|^{2}\\
&\leq\frac{T}{M}\sum_{k=1}^{M}\sum_{t=1}^{T}\mathbb{E}\left\|\nabla f_{i}\left(y_{i}^{r,t};\zeta\right)-\nabla f_{i}\left(z^{r};\zeta\right)+c^{r}-c_{i}^{r}+\nabla f_{i}\left(z^{r}\right)+\nabla f_{i}\left(x^{\star}\right)-\nabla f_{i}\left(x^{\star}\right)\right\|^{2}\\
&\leq\left(\frac{4T}{M}\right)\sum_{i=1}^{M}\sum_{t=1}^{T}\mathbb{E}\left\|y_{i}^{r,t}-z^{r}\right\|^{2}+\left(\frac{4T^{2}}{M}\right)\sum_{i=1}^{M}\mathbb{E}\left\|c_{i}^{r}-\nabla f_{i}\left(x^{r}\right)\right\|^{2}\\
&+\left(4T^{2}\right)\mathbb{E}\left\|c^{r}\right\|^{2}+\left(\frac{4T^{2}}{M}\right)\sum_{i=1}^{M}\mathbb{E}\left\|f_{i}\left(z^{r}\right)-\nabla f_{i}\left(x^{\star}\right)\right\|^{2}\\
&\leq\left(\frac{4T}{M}\right)\sum_{i=1}^{M}\sum_{i=1}^{T}\mathbb{E}\left\|y_{i}^{r,t}-z^{r}\right\|^{2}+\left(\frac{4T^{2}}{M}\right)\sum_{i=1}^{M}\left\|\mathbb{E}c_{i}^{r}-\nabla f_{i}\left(z^{r}\right)\right\|^{2}\\
&+\left(4T^{2}\right)\left\|\mathbb{E}c^{r}\right\|^{2}+\left(\frac{4T^{2}}{M}\right)2\beta\left(f\left(z^{r}\right)-f^{\star}\right).
\end{aligned}\tag{78}$$

Lemma 12. *We can get the bound of E_1*

$$E_1 \leq -f\left(z^r\right)+f\left(x^{\star}\right)+\frac{\beta}{M} \sum_{i=1}^{M} \sum_{i=1}^{T}\left(\left\|y_i^{r, t}-z^r\right\|^2\right)-\frac{\mu}{4} T\left\|z^r-x^{\star}\right\|^2 . \tag{79}$$

Proof. The term E_1 can be bounded by using perturbed strong-convexity (Lemma 1) with $h=f_i, \boldsymbol{x}=y_i^{r,t}, \boldsymbol{y}=\boldsymbol{x}^{\star}$, and $\boldsymbol{z}=z^r$. Next we will calculate the upper bound for E_1.

$$\begin{aligned}
& E_1=-\mathbb{E}\left\langle\mathrm{G}^r, z^r-x^{\star}\right\rangle=-\sum_{j=1}^{N}\left\langle\frac{1}{M} \sum_{i=1}^{M} \sum_{t=1}^{T} \nabla_{(j)} f_i\left(y_i^{r, t}\right), z_{(j)}^r-x_{(j)}^{\star}\right\rangle \\
& =-\left\langle\frac{1}{M} \sum_{i=1}^{M,} \sum_{t=1}^{T} \nabla f_i\left(y_i^{r, t}\right), z^r-x^{\star}\right\rangle \\
& \leq-\frac{1}{M} \sum_{i=1}^{M} \sum_{i=1}^{T}\left(f_i\left(z^r\right)-f_i\left(x^{\star}\right)-\beta\left\|y_i^{r, t}-z^r\right\|^2+\frac{\mu}{4}\left\|z^r-x^{\star}\right\|^2\right) \\
& \leq-f\left(z^r\right)+f\left(x^{\star}\right)+\frac{\beta}{M} \sum_{i=1}^{M} \sum_{i=1}^{T}\left(\left\|y_i^{r, t}-z^r\right\|^2\right)-\frac{\mu}{4} T\left\|z^r-x^{\star}\right\|^2 .
\end{aligned} \tag{80}$$

We will now bound the final source of error which is the client-drift.

Lemma 13. *f_i satisfies Assumptions 1-4. Then, we can bound the drift as*

$$\frac{1}{TM} \sum_{t=1}^{T} \sum_{i=1}^{M} \mathbb{E}\left\|y_i^{r, t}-z^r\right\|^2 \leq 18 T^2 \beta \eta^2\left(f\left(z^r\right)-f\left(x^{\star}\right)\right)+18 T^2 \eta^2 C_{r-1} . \tag{81}$$

Proof. First, we observe that if $T=1, \mathcal{E}_r=0$ since $\boldsymbol{y}_i^{r,0}=z^r$ for all $i \in[M]$ and that Ξ_{r-1} and the right hand side are both positive. Thus the Lemma is trivially true if $T=1$ and we will henceforth assume $T \geq 2$. Starting from the update rule for $i \in[M]$ and $t \in[T]$

$$\begin{aligned}
& \frac{1}{M} \sum_{i \in M} \mathbb{E}\left\|y_i^{r, t}-z^r\right\|^2 \\
& =\frac{1}{M} \sum_{i \in M} \mathbb{E}\left\|y_i^{r, t-1}+\eta \nabla f_i\left(y_i^{r, t} ; \zeta\right)-\eta \nabla f_i\left(z^r ; \zeta\right)+\eta c^r-\eta c_i^r+\eta \nabla f_i\left(z^r\right)-z^r\right\|^2 \\
& \leq(1+a) \frac{1}{M} \sum_{i \in M} \mathbb{E}\left\|y_i^{r, t-1}-z^r+\eta \nabla f_i\left(y_i^{r, t-1} ; \zeta\right)-\eta \nabla f_i\left(z^r ; \zeta\right)\right\|^2 \\
& +\left(1+\frac{1}{a}\right) \eta^2 \frac{1}{M} \sum_{i \in M} \mathbb{E}\left\|c^r-c_i^r+\nabla f_k\left(z^r\right)\right\|^2 \\
& \leq(1+a) \frac{1}{M} \sum_{i \in M} \mathbb{E}\left\|y_i^{r, t-1}-z^r\right\|^2+\left(1+\frac{1}{a}\right) \eta^2 \frac{1}{M} \sum_{i \in M} \mathbb{E}\left\|c^r-c_i^r+\nabla f_i\left(z^r\right)\right\|^2 .
\end{aligned} \tag{82}$$

Once again using our relaxed triangle inequality to expand the other term $\frac{1}{M} \sum_{i \in M} \mathbb{E}\left\|c^r-c_i^r+\nabla f_i\left(x^r\right)\right\|^2$, we get

$$\begin{aligned}
& \frac{1}{M} \sum_{i \in M} \mathbb{E}\left\|c^r-c_i^r+\nabla f_i\left(x^r\right)\right\|^2 \\
& =\frac{1}{M} \sum_{i=1}^{M} \mathbb{E}\left\|c^r-c_i^r+\nabla f_i\left(x^r\right)-\nabla f_i\left(x^{\star}\right)+\nabla f_i\left(x^{\star}\right)\right\|^2 \\
& \leq 3\|\mathbb{E} c^r\|^2+\frac{3}{M} \sum_{i=1}^{M}\left\|\mathbb{E} c_i-\nabla f_i\left(x^{\star}\right)\right\|^2+\frac{3}{M} \sum_{i=1}^{M}\left\|\nabla f_i\left(x^r\right)-\nabla f_i\left(x^{\star}\right)\right\|^2 \\
& \leq \frac{6}{M} \sum_{i=1}^{M}\left\|\mathbb{E} c_i^r-\nabla f_i\left(x^{\star}\right)\right\|^2+6 \beta\left(f\left(x^r\right)-f\left(x^{\star}\right)\right) .
\end{aligned} \tag{83}$$

The last step used the smoothness of f_i. Combining the bounds on in the original inequality and using $a = \frac{1}{T-1}$, we have

$$\begin{aligned}\frac{1}{M}\sum_{i=1}^{M}\mathbb{E}\left\|y_i^{r,t-1} - z^r\right\|^2 &\leq \frac{\left(1+\frac{1}{T-1}\right)}{M}\sum_{i=1}^{M}\mathbb{E}\left\|y_i^{r,t-1} - z^r\right\|^2 + 6\eta^2 T\beta\left(f\left(z^r\right) - f\left(x^\star\right)\right)\\ &+ \frac{6T\eta^2}{M}\sum_{i=1}^{M}\left\|\mathbb{E}c_i^r - \nabla f_i\left(x^\star\right)\right\|^2.\end{aligned} \tag{84}$$

Unrolling the recursion, we get the following for any $t \in \{1, \ldots, T\}$,

$$\begin{aligned}\frac{1}{M}\sum_{i=1}^{M}\mathbb{E}\left\|y_i^{r,t-1} - x^r\right\|^2 &\leq \left(6T\beta\eta^2\left(f\left(z^r\right) - f\left(x^\star\right)\right) + 6T\eta^2 C_{r-1}\right)\left(\sum_{\tau=0}^{t-1}\left(1+\frac{1}{T-1}\right)^r\right)\\ &\leq \left(6T\beta\eta^2\left(f\left(z^r\right) - f\left(x^\star\right)\right) + 6T\eta^2 C_{r-1}\right)(T-1)\left(\left(1+\frac{1}{T-1}\right)^T - 1\right)\\ &\leq \left(6T\beta\eta^2\left(f\left(z^r\right) - f\left(x^\star\right)\right) + 6T\eta^2 C_{r-1}\right)3T\\ &\leq 18T^2\beta\eta^2\left(f\left(z^r\right) - f\left(x^\star\right)\right) + 18T^2\eta^2 C_{r-1}.\end{aligned} \tag{85}$$

The inequality $(T-1)\left(\left(1+\frac{1}{T-1}\right)^T - 1\right) \leq 3T$ can be verified for $T = 2, 3$ manually. For $T \geq 4$, $(T-1)\left(\left(1+\frac{1}{T-1}\right)^T - 1\right) < T\left(\exp\left(\frac{T}{T-1}\right) - 1\right) \leq T\left(\exp\left(\frac{4}{3}\right) - 1\right) < 3T$.

$$C_r = \frac{1}{M}\sum_{i=1}^{M}\left\|\mathbb{E}\left[c_i^r\right] - \nabla f_i\left(x^\star\right)\right\|^2. \tag{86}$$

Again averaging over t,

$$\frac{1}{TM}\sum_{t=1}^{T}\sum_{i=1}^{M}\mathbb{E}\left\|y_i^{r,t} - x^r\right\|^2 \leq 18T^2\beta\eta^2\left(f\left(z^r\right) - f\left(x^\star\right)\right) + 18T^2\eta^2 C_{r-1}. \tag{87}$$

Lemma 14. *For updates of FedBCGD+ with the control update and Assumptions 3-4, the following holds true for any $\tilde{\eta} \in [0, 1/\beta]$:*

$$\mathbb{E}\left[C_r\right] \leq \left(1 - \frac{K}{M}\right)C_{r-1} + \frac{K}{M}\left(4\beta\left(\mathbb{E}\left[f\left(z^{r-1}\right)\right] - f\left(x^\star\right)\right)\right). \tag{88}$$

Proof. We define client-drift to be how much the clients move from their starting point:

$$\mathcal{E}_r := \frac{1}{TM}\sum_{t=1}^{T}\sum_{i=1}^{M}\mathbb{E}\left\|y_i^{r,t} - z^r\right\|^2. \tag{89}$$

Plugging the above expression in the definition ofC_r we get

$$\begin{aligned}C_r &= \frac{1}{M}\sum_{i=1}^{M}\left\|\mathbb{E}\left[\mathbf{c}_i^r\right] - \nabla f_i\left(x^\star\right)\right\|^2\\ &= \frac{1}{M}\sum_{i=1}^{M}\left\|\left(1-\frac{K}{M}\right)\left(\mathbb{E}\left[\mathbf{c}_i^{r-1}\right] - \nabla f_i\left(x^\star\right)\right) + \frac{K}{M}\left(\left[\nabla f_i\left(z^r\right)\right] - \nabla f_i\left(x^\star\right)\right)\right\|^2\\ &\leq \left(1-\frac{K}{M}\right)C_{r-1} + \frac{K}{M^2}\sum_{i=1}^{M}\mathbb{E}\left\|\nabla f_i\left(z^r\right) - \nabla f_i\left(x^\star\right)\right\|^2.\end{aligned} \tag{90}$$

The final step applied Jensen's inequality twice. We can then further simplify using the relaxed triangle inequality as follows:

$$\begin{aligned}
\mathbb{E}\left[C_r\right] &\leq \left(1-\frac{K}{M}\right) C_{r-1}+\frac{K}{M^2} \sum_{i=1}^M \mathbb{E}\left\|\nabla f_i\left(z^r\right)-\nabla f_i\left(x^{\star}\right)\right\|^2 \\
&\leq \left(1-\frac{K}{M}\right) C_{r-1}+\frac{K}{M^2} \sum_{i=1}^M \mathbb{E}\left\|\nabla f_i\left(z^{r-1}\right)-\nabla f_i\left(x^{\star}\right)\right\|^2 \\
&\leq \left(1-\frac{K}{M}\right) C_{r-1}+\frac{K}{M^2} \sum_{i=1}^M \mathbb{E}\left\|\nabla f_i\left(z^{r-1}\right)-\nabla f_k\left(x^{\star}\right)\right\|^2 \\
&\leq \left(1-\frac{K}{M}\right) C_{r-1}+\frac{K}{M}\left(4 \beta\left(\mathbb{E}\left[f\left(z^{r-1}\right)\right]-f\left(x^{\star}\right)\right)\right).
\end{aligned} \tag{91}$$

The last two inequalities follow from smoothness of $\{f_i\}$ and the definition

$$\mathcal{E}_r:=\frac{1}{TM} \sum_{t=1}^T \sum_{i=1}^M \mathbb{E}\left\|y_i^{r,t}-z^r\right\|^2. \tag{92}$$

Lemma 15.

$$\left\|z^{r+1}-x^{\star}\right\|^2+9 \tilde{\eta}^2 \frac{M}{K} C_r \leq\left(1-\frac{\tilde{\eta} \mu}{2}\right)\left\|z^r-x^{\star}\right\|^2+\left(1-\frac{\mu \tilde{\eta}}{2}\right) 9 \tilde{\eta}^2 \frac{M}{K} C_{r-1} \tag{93}$$

With E_1 and E_2,we can get

$$\begin{aligned}
&\mathbb{E}\left\|z^{r+1}-x^{\star}\right\|^2 \leq \mathbb{E}\left\|z^r-x^{\star}\right\|^2+2 \eta \alpha \underbrace{\mathbb{E}\left\langle-\mathbf{G}^r, z^r-x^{\star}\right\rangle}_{E_1}+\eta^2 \alpha^2 \underbrace{\mathbb{E}\left\|\mathbf{G}^r\right\|^2}_{E_2} \\
&\leq\left(1-\frac{\eta \alpha \mu}{2}\right) T\left\|z^r-x^{\star}\right\|^2+2 \eta \alpha T\left(-f\left(z^r\right)+f\left(x^{\star}\right)\right) \\
&+\left[\frac{2 \eta \alpha \beta}{M}+\eta^2 \alpha^2\left(\frac{4 T}{M}\right)\right] \sum_{i=1}^M \sum_{t=1}^T\left(\left\|y_k^{r,t}-z^r\right\|^2\right)+\eta^2 \alpha^2\left(\frac{8 T^2}{M}\right) \sum_{i=1}^M\left\|\mathbb{E} c_i^r-\nabla f_i\left(z^r\right)\right\|^2 \\
&\leq\left(1-\frac{\eta \alpha \mu T}{2}\right)\left\|z^r-x^{\star}\right\|^2+2 \eta \alpha T\left(-f\left(z^r\right)+f\left(x^{\star}\right)\right)+\left[2 \eta \alpha T \beta+4 T^2 \eta^2 \alpha^2\right] \frac{1}{MT} \sum_{i=1}^M \sum_{t=1}^T\left(\left\|y_k^{r,t}-z^r\right\|^2\right) \\
&+8 T^2 \eta^2 \alpha^2\left(\frac{1}{M}\right) \sum_{i=1}^M\left\|\mathbb{E} c_i^r-\nabla f_i\left(z^r\right)\right\|^2,
\end{aligned} \tag{94}$$

with $\tilde{\eta}=\alpha \eta T$, we have

$$\begin{aligned}
&\leq\left(1-\frac{\tilde{\eta} \mu}{2}\right)\left\|z^r-x^{\star}\right\|^2+2 \tilde{\eta}\left(-f\left(z^r\right)+f\left(x^{\star}\right)\right)+\left[2 \tilde{\eta} \beta+4 \tilde{\eta}^2\right] \frac{1}{MT} \sum_{i=1}^M \sum_{t=1}^T\left(\left\|y_i^{r,t}-z^r\right\|^2\right) \\
&+8 \tilde{\eta}^2\left(\frac{1}{M}\right) \sum_{i=1}^M\left\|\mathbb{E} c_i^r-\nabla f_i\left(z^r\right)\right\|^2 \\
&\leq\left(1-\frac{\tilde{\eta} \mu}{2}\right)\left\|z^r-x^{\star}\right\|^2+2 \tilde{\eta}\left(-f\left(z^r\right)+f\left(x^{\star}\right)\right)+\left[2 \tilde{\eta} \beta+4 \tilde{\eta}^2\right] \mathcal{E}_r+8 \tilde{\eta}^2 \mathbb{E}\left[C_r\right].
\end{aligned} \tag{95}$$

We can use Lemma 13 (scaled by $9 \tilde{\eta}^2 \frac{N}{S}$) to bound the control-lag

$$3 \beta \tilde{\eta} \mathcal{E}_r \leq \frac{54 \tilde{\eta}^3 \beta^2}{\alpha^2}\left(f\left(z^r\right)-f\left(x^{\star}\right)\right)+\frac{54 \tilde{\eta}^3 \beta}{\alpha^2} C_{r-1}. \tag{96}$$

Now recall that Lemma 14 bounds the client-drift:

$$\begin{aligned}
9 \tilde{\eta}^2 \frac{M}{K} C_r &\leq\left(1-\frac{\mu \tilde{\eta}}{2}\right) 9 \tilde{\eta}^2 \frac{M}{K} C_{r-1}+9\left(\frac{\mu \tilde{\eta} M}{2 K}-1\right) \tilde{\eta}^2 C_{r-1} \\
&+9 \tilde{\eta}^2\left(4 \beta\left(\mathbb{E}\left[f\left(z^{r-1}\right)\right]-f\left(x^{\star}\right)\right)+2 \beta^2 \mathcal{E}\right).
\end{aligned} \tag{97}$$

Adding all three inequalities together, we have

$$\begin{aligned}
&\left\|z^{r+1}-x^{\star}\right\|^2+9\tilde{\eta}^2\frac{M}{K}C_r \\
&\le\left(1-\frac{\tilde{\eta}\mu}{2}\right)\left\|z^{r}-x^{\star}\right\|^2+\left(1-\frac{\mu\tilde{\eta}}{2}\right)9\tilde{\eta}^2\frac{M}{K}C_{r-1}-\left(2\tilde{\eta}-36\tilde{\eta}^2\beta-54\tilde{\eta}^3\beta^2\right)\left(f\left(z^r\right)-f\left(x^{\star}\right)\right) \\
&+\left[-\tilde{\eta}\beta+4\tilde{\eta}^2\beta^2\right]\mathcal{E}_r+\left(\frac{9\mu\tilde{\eta}M}{2S}-9+8+54\tilde{\eta}\right)\tilde{\eta}^2C_{r-1}.
\end{aligned} \tag{98}$$

Finally, with $\tilde{\eta}\le\frac{1}{81\beta}$ and $\tilde{\eta}\le\frac{K}{15\mu M}$ the lemma follows from noting that

$$-54\beta^2\tilde{\eta}^2-36\beta\tilde{\eta}+2\ge 0, \tag{99}$$

$$-\tilde{\eta}\beta+4\tilde{\eta}^2\beta^2\le 0, \tag{100}$$

$$\frac{9\mu\tilde{\eta}N}{2S}-9+8+54\tilde{\eta}\le 0. \tag{101}$$

The final rate for the case of strongly convex follows simply by unrolling the recursive bound and using Lemma 7,

$$\left\|z^{r+1}-x^{\star}\right\|^2+9\tilde{\eta}^2\frac{M}{K}C_r\le\left(1-\frac{\tilde{\eta}\mu}{2}\right)\left\|z^{r}-x^{\star}\right\|^2+\left(1-\frac{\mu\tilde{\eta}}{2}\right)9\tilde{\eta}^2\frac{M}{K}C_{r-1}, \tag{102}$$

$$\mathbb{E}\left[f\left(\bar{z}^R\right)\right]-f\left(x^{\star}\right)\le\tilde{O}\left(\frac{M\mu}{K}\tilde{D}^2\exp\left(-\min\left\{\frac{K}{30M},\frac{\mu}{162\beta}\right\}R\right)\right). \tag{103}$$

5.2 2: The convergence rate of general convex and smooth case:

$$\left\|z^{r+1}-x^{\star}\right\|^2+9\tilde{\eta}^2\frac{M}{K}C_r\le\left(1-\frac{\tilde{\eta}\mu}{2}\right)\left\|z^{r}-x^{\star}\right\|^2+\left(1-\frac{\mu\tilde{\eta}}{2}\right)9\tilde{\eta}^2\frac{M}{K}C_{r-1}. \tag{104}$$

For general convex case, we have $\mu=0$, then the following inequality holds:

$$\left\|z^{r+1}-x^{\star}\right\|^2+9\tilde{\eta}^2\frac{M}{K}C_r\le\left\|z^{r}-x^{\star}\right\|^2+\left(1-\frac{\mu\tilde{\eta}}{2}\right)9\tilde{\eta}^2\frac{M}{K}C_{r-1}. \tag{105}$$

For the general convex setting, averaging over r in Lemma 8,

$$\mathbb{E}\left[f\left(\bar{z}^R\right)\right]-f\left(x^{\star}\right)\le O\left(\sqrt{\frac{M}{K}}\frac{\beta\tilde{D}^2}{R}\right). \tag{106}$$

5.3 3. The convergence rate of non-convex and smooth case:

Recall that in round r, we update the control variate

$$\boldsymbol{c}_i^r=\begin{cases}\nabla f_i\left(\boldsymbol{x}^r\right) & \text{if } i\in\mathcal{S}^r\\ \boldsymbol{c}_i^{r-1} & \text{otherwise}\end{cases}. \tag{107}$$

We introduce the following notation to keep track of the lag in the update of the control variate: define a sequence of parameters $\left\{\boldsymbol{\alpha}_i^{r,t}\right\}$ such that for any $i\in[M]$ and $t\in[T]$ we have $\boldsymbol{\alpha}_i^{0,t}:=\boldsymbol{x}^0$ and for $r\ge 1$,

$$\boldsymbol{\alpha}_i^{r,t}:=\begin{cases}\boldsymbol{y}_i^{r,t} & \text{if } i\in\mathcal{S}^r\\ \boldsymbol{\alpha}_i^{r-1,t} & \text{otherwise}.\end{cases} \tag{108}$$

By the update rule for control variates (19) and the definition of $\left\{\boldsymbol{\alpha}_i^{r,t}\right\}$ above, the following property always holds:

$$\boldsymbol{c}_{k,j}^r=\nabla f_{k,j}\left(\boldsymbol{x}^r\right). \tag{109}$$

We can then define the following Ξ_r to be the error in control variate for round r:

$$\Xi_r:=\frac{1}{TM}\sum_{t=1}^{T}\sum_{i=1}^{M}\mathbb{E}\left\|\boldsymbol{\alpha}_i^{r,t}-\boldsymbol{z}^r\right\|^2. \tag{110}$$

Also recall the closely related definition of client drift caused by local updates:

$$\mathcal{E}_r := \frac{1}{TM}\sum_{t=1}^{T}\sum_{i=1}^{M}\mathbb{E}\left[\left\|\boldsymbol{y}_i^{r,t}-\boldsymbol{z}^r\right\|^2\right]. \tag{111}$$

From the smoothness of the function, we can obtain

$$\begin{aligned}
&\mathbb{E}f\left(z^{r+1}\right) \le \mathbb{E}f\left(z^r\right)+\mathbb{E}\left\langle \nabla f\left(z^r\right), z^{r+1}-z^r\right\rangle+\frac{\beta}{2}\mathbb{E}\left\|z^{r+1}-z^r\right\|^2\\
&\le \mathbb{E}f\left(z^r\right)+\alpha\eta\underbrace{\mathbb{E}\left\langle \nabla f\left(z^r\right), -\mathbf{G}^r\right\rangle}_{F_1}+\frac{\beta}{2}\eta^2\alpha^2\underbrace{\mathbb{E}\left\|\mathbf{G}^r\right\|^2}_{F_2}
\end{aligned}. \tag{112}$$

We will first calculate the upper bound limit for F_2,Let us analyze how the control variates effect the variance of the aggregate server update.

$$\begin{aligned}
&F_2=\mathbb{E}\left\|\mathbf{G}^r\right\|^2\\
&=\sum_{j=1}^{N}\mathbb{E}\left\|\frac{1}{K}\sum_{k=1}^{K}\sum_{t=1}^{T}\nabla_{(j)}f_{k,j}\left(y_{k,j}^{r,t};\zeta\right)+c_{(j)}^r-c_{k,j,(j)}^r+\nabla_{(j)}f_{k,j}\left(z^r\right)-\nabla_{(j)}f_{k,j}\left(z^r;\zeta\right)\right\|^2\\
&\le \frac{1}{M^2}\sum_{j=1}^{N}\sum_{i=1}^{M}\sum_{t=1}^{T}\mathbb{E}\left\|\nabla_{(j)}f_i\left(y_i^{r,t};\zeta\right)-\nabla_{(j)}f_i\left(z^r;\zeta\right)+c_{(j)}^r-c_{i,(j)}^r+\nabla_{(j)}f_i\left(z^r\right)\right\|^2\\
&\le \frac{1}{M^2}\sum_{j=1}^{N}\sum_{i=1}^{M}\sum_{t=1}^{T}\mathbb{E}\left\|\nabla f_i\left(y_i^{r,t};\zeta\right)-\nabla f_i\left(z^r;\zeta\right)+c^r-c_i^r+\nabla f_i\left(z^r\right)\right\|^2\\
&\le \frac{1}{M^2}\sum_{i=1}^{M}\sum_{t=1}^{T}\mathbb{E}\left\|\nabla f_i\left(y_i^{r,t};\zeta\right)-\nabla f_i\left(z^r;\zeta\right)+c^r-c_i^r+\nabla f_i\left(z^r\right)+\nabla f\left(z^r\right)-\nabla f_i\left(z^r\right)\right\|^2\\
&\le \left(\frac{T^2\beta^2}{MT}\right)\sum_{i=1}^{M}\sum_{t=1}^{T}\mathbb{E}\left\|y_i^{r,t}-z^r\right\|^2+\frac{4T^2}{M^2}\sum_{i=1}^{M}\mathbb{E}\left\|c_i^r-\nabla f_i\left(z^r\right)\right\|^2+4T^2\mathbb{E}\left\|c^r-\nabla f\left(z^r\right)\right\|^2\\
&+\left(\frac{4T^2}{M^2}\right)\sum_{i=1}^{M}\mathbb{E}\left\|\nabla f\left(z^r\right)\right\|^2\\
&\le \frac{4T}{M}\sum_{i=1}^{M}\sum_{t=1}^{T}\mathbb{E}\left\|y_i^{r,t}-z^r\right\|^2+\frac{4T^2}{M^2}\sum_{i=1}^{M}\mathbb{E}\left\|c_i^r-\nabla f_i\left(z^r\right)\right\|^2\\
&+4T^2\mathbb{E}\left\|\frac{1}{M}\sum_{i=1}^{M}\left[c_i^r-\nabla f_i\left(z^r\right)\right]\right\|^2+\frac{4T^2}{M^2}\sum_{i=1}^{M}\mathbb{E}\|\nabla f\left(z^r\right)\|^2.\\
&\le 4T^2\beta^2\mathcal{E}_r+8\beta^2T^2\Xi_{r-1}+4T^2\mathbb{E}\left\|\nabla f\left(z^r\right)\right\|^2.
\end{aligned} \tag{113}$$

Lemma 16. *Suppose f_i satisfies Assumptions 4-5. We can bound the drift $\mathcal{E}_r \le \frac{1}{MT}\sum_{i\in M}\mathbb{E}\left\|y_i^{r,t}-z^r\right\|^2$ as*

$$\mathcal{E}_r \le 24T^2\eta^2\beta^2\mathbb{E}\left\|z^r-\boldsymbol{\alpha}^r\right\|^2+12T^2\eta^2\mathbb{E}\left\|\nabla f\left(z^r\right)\right\|^2 \tag{114}$$

Proof. First, we observe that if $T=1, \mathcal{E}_r=0$ since $\boldsymbol{y}_i^{r,0}=\boldsymbol{x}^r$ for all $i\in[M]$ and that Ξ_{r-1} and the right hand side are both positive. Thus the lemma is trivially true if $T=1$ and we will henceforth assume $T\geq 2$. Starting from the update rule (18) for $i\in[N]$ and $t\in[T]$

$$\begin{aligned}
&\frac{1}{M}\sum_{i\in M}\mathbb{E}\left\|y_i^{r,t}-z^r\right\|^2\\
&=\frac{1}{M}\sum_{i\in M}\mathbb{E}\left\|y_i^{r,t-1}+\eta\nabla f_i\left(y_i^{r,t};\zeta\right)-\eta\nabla f_i\left(z^r;\zeta\right)+\eta c^r-\eta c_i^r+\eta\nabla f_i\left(x^r\right)-z^r\right\|^2\\
&\leq(1+a)\frac{1}{M}\sum_{i\in M}\mathbb{E}\left\|y_i^{r,t-1}-z^r\right\|^2+\left(1+\frac{1}{a}\right)\eta^2\frac{1}{M}\sum_{i\in M}\mathbb{E}\left\|\nabla f_i\left(y_i^{r,t-1};\zeta\right)-\nabla f_i\left(z^r;\zeta\right)+c^r-c_i^r+\nabla f_i\left(z^r\right)\right\|^2\\
&\leq\left(1+\frac{1}{T-1}+4T\beta^2\eta^2\right)\frac{1}{M}\sum_{i\in M}\mathbb{E}\left\|y_i^{r,t-1}-z^r\right\|^2+4T\eta^2\frac{1}{M}\sum_{k\in M}\mathbb{E}\left\|c^r-\nabla f\left(z^r\right)\right\|^2\\
&+4T\eta^2\frac{1}{M}\sum_{i\in M}\mathbb{E}\left\|\nabla f_i\left(z^r\right)-c_i^r\right\|^2+4T\eta^2\mathbb{E}\left\|\nabla f\left(z^r\right)\right\|^2\\
&\leq\left(1+\frac{1}{T-1}+4T\beta^2\eta^2\right)\frac{1}{M}\sum_{i\in M}\mathbb{E}\left\|y_i^{r,t-1}-z^r\right\|^2+4T\eta^2\frac{1}{M}\sum_{i\in M}\mathbb{E}\left\|c^r-\nabla f\left(z^r\right)\right\|^2\\
&+4T\eta^2\frac{1}{M}\sum_{i\in M}\mathbb{E}\left\|\nabla f_i\left(z^r\right)-c_i^r\right\|^2+4T\eta^2\mathbb{E}\left\|\nabla f\left(z^r\right)\right\|^2\\
&\leq\left(1+\frac{1}{T-1}+4T\beta^2\eta^2\right)\frac{1}{M}\sum_{i\in M}\mathbb{E}\left\|y_i^{r,t-1}-z^r\right\|^2+8T\eta^2\beta^2\mathbb{E}\left\|z^r-\boldsymbol{\alpha}^r\right\|^2+4T\eta^2\mathbb{E}\left\|\nabla f\left(z^r\right)\right\|^2\\
&\leq 24T^2\eta^2\beta^2\mathbb{E}\left\|z^r-\boldsymbol{\alpha}^r\right\|^2+12T^2\eta^2\mathbb{E}\left\|\nabla f\left(z^r\right)\right\|^2.
\end{aligned}\tag{115}$$

Averaging the above over i, the definition of c and Ξ_{r-1}, we have

$$\frac{1}{MT}\sum_{i\in M}\mathbb{E}\left\|y_i^{r,t}-z^r\right\|^2\leq 24T^2\eta^2\beta^2\mathbb{E}\left\|z^r-\boldsymbol{\alpha}^r\right\|^2+12T^2\eta^2\mathbb{E}\left\|\nabla f\left(z^r\right)\right\|.\tag{116}$$

Lemma 17. *For updates of FedBCGD+ and Assumptions 3 and 4, the following holds true for any $\tilde{\eta}\leq\frac{1}{24\beta}\left(\frac{S}{N}\right)^a$ for $a\in\left[\frac{1}{2},1\right]$ where $\tilde{\eta}:=\alpha T\eta$:*

$$\Xi_r\leq\left(1-\frac{17K}{36M}\right)\Xi_{r-1}+\frac{1}{48\beta^2}\left(\frac{K}{M}\right)^{2a-1}\left\|\nabla f\left(z^r\right)\right\|^2+\frac{97}{48}\left(\frac{K}{M}\right)^{2a-1}\mathcal{E}_r.\tag{117}$$

Proof. The proof proceeds similar to that of Lemma 13 except that we cannot rely on convexity. Recall that after round r, the definition of $\boldsymbol{\alpha}_i^{r,t}$ implies that

$$\mathbb{E}\left[\boldsymbol{\alpha}^r\right]=\left(1-\frac{K}{M}\right)\boldsymbol{\alpha}^{r-1}+\frac{K}{M}z^{r-1},\tag{118}$$

$$\begin{aligned}
&\Xi_r=\mathbb{E}\left\|\alpha^r-z^r\right\|^2=\left(1-\frac{K}{M}\right)\cdot\mathbb{E}\left\|\alpha^{r-1}-z^r\right\|^2+\frac{K}{M}\cdot\mathbb{E}\left\|z^{r-1}-z^r\right\|^2\\
&\leq\left(1-\frac{K}{M}\right)\mathbb{E}\left(\left\|\alpha^{r-1}-z^{r-1}\right\|^2+\left\|z^r-z^{r-1}\right\|^2+2\left\langle z^r-z^{r-1},z^{r-1}-\alpha^{r-1}\right\rangle\right)+\frac{K}{M}\cdot\mathbb{E}\left\|z^{r-1}-z^r\right\|^2\\
&\leq\left(1-\frac{K}{M}\right)\mathbb{E}\left(\left\|\alpha^{r-1}-z^{r-1}\right\|^2+\left\|z^r-z^{r-1}\right\|^2+\frac{1}{b}\left(2\tilde{\eta}^2\beta^2\mathcal{E}_r+2\tilde{\eta}^2\mathbb{E}\left\|\nabla f\left(z^{r-1}\right)\right\|^2\right)+b\left\|\alpha^{r-1}-z^{r-1}\right\|^2\right)\\
&+\frac{K}{M}\cdot\mathbb{E}\left\|z^{r-1}-z^r\right\|^2\\
&\leq\left(1-\frac{K}{M}\right)(1+b)\mathbb{E}\left\|\alpha^{r-1}-z^{r-1}\right\|^2+\left\|z^r-z^{r-1}\right\|^2+\left(1-\frac{K}{M}\right)\frac{1}{b}\left(2\tilde{\eta}^2\beta^2\mathcal{E}_r+2\tilde{\eta}^2\mathbb{E}\left\|\nabla f\left(z^r\right)\right\|^2\right)\\
&\leq\left[\left(1-\frac{K}{M}\right)(1+b)+8\tilde{\eta}^2\beta^2\right]\Xi_{r-1}+\left(4\tilde{\eta}^2\beta^2+2\left(1-\frac{K}{M}\right)\frac{1}{b}\tilde{\eta}^2\beta^2\right)\mathcal{E}_r\\
&+\left(4+2\left(1-\frac{K}{M}\right)\frac{1}{b}\right)\tilde{\eta}^2\mathbb{E}\left\|\nabla f\left(z^r\right)\right\|^2.
\end{aligned}\tag{119}$$

The last inequality applied Lemma 15. Verify that with choice of $b=\frac{K}{2(M-K)}$, we have $\left(1-\frac{K}{M}\right)(1+b)\leq\left(1-\frac{K}{2M}\right)$ and $\frac{1}{b}\leq\frac{2M}{K}$. Plugging these values along with the bound on the step-size $8\beta^2\tilde{\eta}^2\leq\frac{1}{36}\left(\frac{K}{M}\right)^{2a}\leq\frac{K}{36M},\tilde{\eta}\leq\frac{1}{24\beta}\left(\frac{K}{M}\right)^{a}$ for $a\in\left[\frac{1}{2},1\right]$ completes the lemma.

$$\Xi_r\leq\left(1-\frac{17K}{36M}\right)\Xi_{r-1}+\frac{1}{48\beta^2}\left(\frac{K}{M}\right)^{2a-1}\left\|\nabla f\left(z^r\right)\right\|^2+\frac{97}{48}\left(\frac{K}{M}\right)^{2a-1}\mathcal{E}_r. \tag{120}$$

Lemma 18. *Suppose the updates of FedBCGD+ satisfy Assumptions 2-4. For any effective step-size $\tilde{\eta}$ satisfying $\tilde{\eta}\leq\frac{1}{24\beta}\left(\frac{K}{M}\right)^{\frac{2}{3}}$*

$$\left(\mathbb{E}\left[f\left(z^r\right)\right]+12\beta^3\tilde{\eta}^2\frac{M}{K}\Xi_r\right)\leq\left(\mathbb{E}\left[f\left(z^{r-1}\right)\right]+12\beta^3\tilde{\eta}^2\frac{M}{K}\Xi_{r-1}\right)-\frac{\tilde{\eta}}{14}\mathbb{E}\left\|\nabla f\left(z^{r-1}\right)\right\|^2 \tag{121}$$

Proof. Applying the upper bounds of F_1 and F_2,

$$\begin{aligned}
\mathbb{E}f\left(z^{r+1}\right)&\leq\mathbb{E}f\left(z^r\right)+\mathbb{E}\left\langle\nabla f\left(z^r\right),z^{r+1}-z^r\right\rangle+\frac{\beta}{2}\mathbb{E}\left\|z^{r+1}-z^r\right\|^2\\
&\leq\mathbb{E}f\left(z^r\right)+\alpha\eta\underbrace{\mathbb{E}\left\langle\nabla f\left(z^r\right),-\mathbf{G}^r\right\rangle}_{F_1}+\frac{\beta}{2}\eta^2\alpha^2\underbrace{\mathbb{E}\left\|\mathbf{G}^r\right\|^2}_{F_2}\\
&\leq\mathbb{E}f\left(z^r\right)-\frac{\tilde{\eta}}{2}\left\|\nabla f\left(z^r\right)\right\|^2+\frac{\tilde{\eta}\beta^2}{2}\mathcal{E}_r+\frac{\beta}{2}\eta^2\alpha^2\left[4T^2\beta^2\mathcal{E}_r+8\beta^2T^2\Xi_{r-1}+4T^2\mathbb{E}\left\|\nabla f\left(z^r\right)\right\|^2\right]\\
&\leq\mathbb{E}f\left(z^r\right)-\frac{\tilde{\eta}}{2}\left\|\nabla f\left(z^r\right)\right\|^2+\left(\frac{\tilde{\eta}\beta^2}{2}+2\beta^3\tilde{\eta}^2\right)\mathcal{E}_r+4\beta^3\tilde{\eta}^2\Xi_{r-1}+2\beta\tilde{\eta}^2\mathbb{E}\left\|\nabla f\left(z^r\right)\right\|^2\\
&\leq\mathbb{E}f\left(z^r\right)-\left(\frac{\tilde{\eta}}{2}-2\beta\tilde{\eta}^2\right)\left\|\nabla f\left(z^r\right)\right\|^2+\left(\frac{\tilde{\eta}\beta^2}{2}+2\beta^3\tilde{\eta}^2\right)\mathcal{E}_r+4\beta^3\tilde{\eta}^2\Xi_{r-1}.
\end{aligned} \tag{122}$$

Also recall that Lemmas 16 and 17 state that

$$12\beta^3\tilde{\eta}^2\frac{M}{K}\Xi_r\leq12\beta^3\tilde{\eta}^2\frac{M}{K}\left(\left(1-\frac{17K}{36M}\right)\Xi_{r-1}+\frac{1}{48\beta^2}\left(\frac{K}{M}\right)^{2a-1}\left\|\nabla f\left(z^r\right)\right\|^2+\frac{97}{48}\left(\frac{K}{M}\right)^{2a-1}\mathcal{E}_r\right) \tag{123}$$

$$\frac{5}{3}\beta^2\tilde{\eta}\mathcal{E}_r\leq\frac{5}{3\alpha^2}\beta^3\tilde{\eta}^2\Xi_{r-1}+\frac{\tilde{\eta}}{24\alpha^2}\mathbb{E}\left\|\nabla f\left(z^r\right)\right\|^2. \tag{124}$$

Adding these bounds on Ξ_r and $\mathcal{E}_r$ to that of $\mathbb{E}[f(z^{r+1})]$ gives

$$\left(\mathbb{E}\left[f\left(z^{r+1}\right)\right]+12\beta^3\tilde{\eta}^2\frac{M}{K}\Xi_r\right)\leq\left(\mathbb{E}\left[f\left(z^r\right)\right]+12\beta^3\tilde{\eta}^2\frac{M}{K}\Xi_{r-1}\right)+\left(4+\frac{5}{3\alpha^2}-\frac{17}{3}\right)\beta^3\tilde{\eta}^2\Xi_{r-1} \tag{125}$$

$$-\left(\frac{\tilde{\eta}}{2}-2\beta\tilde{\eta}^2-\frac{1}{4}\beta\tilde{\eta}^2\left(\frac{N}{S}\right)^{2-2a}-\frac{\tilde{\eta}}{24\alpha^2}\right)\left\|\nabla f\left(z^r\right)\right\|^2+\left(\frac{\tilde{\eta}}{2}-\frac{5\tilde{\eta}}{3}+2\beta\tilde{\eta}^2+\frac{97}{4}\beta\tilde{\eta}^2\left(\frac{M}{K}\right)^{2-2a}\right)\beta^2\mathcal{E}_r. \tag{126}$$

$$\left(\mathbb{E}\left[f\left(z^r\right)\right]+12\beta^3\tilde{\eta}^2\frac{M}{K}\Xi_r\right)\leq\left(\mathbb{E}\left[f\left(z^{r-1}\right)\right]+12\beta^3\tilde{\eta}^2\frac{M}{K}\Xi_{r-1}\right)-\frac{\tilde{\eta}}{14}\mathbb{E}\left\|\nabla f\left(z^{r-1}\right)\right\|^2. \tag{127}$$

By our choice of $a=\frac{2}{3}$ and plugging in the bound on step-size $\beta\tilde{\eta}\left(\frac{N}{S}\right)^{2-2a}\leq\frac{1}{24}$ proves the lemma. The non-convex rate of convergence now follows by unrolling the recursion in Lemma 18 and selecting an appropriate step-size $\tilde{\eta}$ as in Lemma 8. Finally, note that if we initialize $\boldsymbol{c}_i^0=\nabla f_i\left(\boldsymbol{x}^0\right)$ then we have $\Xi_0=0$. We can get

$$\mathbb{E}\left[\left\|\nabla f\left(\bar{z}^R\right)\right\|^2\right]\leq O\left(\frac{\beta F}{R}\left(\frac{M}{K}\right)^{\frac{2}{3}}\right). \tag{128}$$

6 APPENDIX F: MORE EXPERIMENTAL DETAILS

In this section, we give some experimental results:

6.1 Methods

We also demonstrate the robustness of FedBCGD and FedBCGD+ in different settings. For comparison, we use FedAvg [?], SCAFFOLD [?], FedAvgM [?], FedDC [?] , FedAdam [?] FL baselines. The following is a detailed introduction to the experimental setup, model and dataset, and comparison methods.

6.2 Dataset processing

We evaluate FL on world datasets of image classification tasks including CIFAR-10 dataset, CIFAR-100 dataset, Tiny ImageNet dataset, mnist dataset in our study.Both CIFAR10 and CIFAR100 datasets contain 60000 sheets of 3 × 32 × 32 images. For CIFAR10, there are 10 categories, while there are 100 categories on CIFAR100. For CIFAR10 and CIFAR100, the sample size in the training set is 50000, and the sample size in the test set is 10000. In the experiment, we set up 100 clients with 500 images per client.

Tiny ImageNet Challenge is the default course project for Stanford CS231N. Tiny Imagenet has 200 classes. Each class has 500 training images, 50 validation images, and 50 test images. In the experiment, we set up 100 clients with 1000 images per client. We adjusted the size to 256 × 256 and crop to 224 × 224 to preprocess each image

6.3 Model

To test the robustness of our algorithms, we use standard classifiers (including LeNet-5 [?], VGG-11, VGG-19 [?], and ResNet-18 [?]), Vision Transformer (ViT-Base) [?], Logistic regression Model [?]. We divided the parameters of the model into 5 blocks or more blocks and provide the detailed parameter block division of the model in the Appendix.

6.4 Hyper-parameter setting

We provide hyperparameter settings for different datasets. For all real-world datasets in the convolutional network, including CIFAR10 and CIFAR100, set the sampling rate to 10% for 100 clients. We set the batch size to 50, the number of local epochs for one round of communication to 5, and the initial learning rate is searched in $\{0.01, 0.03, 0.05, 0.1, 0.2, 0.3\}$. The learning rate decay for each round is 0.998, and the weight decay is 0.001. We searched for FedBCGD and FedAvgM α in $\{0.4, 0.5, 0.6, 0.7, 0.8, 0.9\}$, FedDC settings $\alpha = 0.01$, FedAdam setting $\alpha = 0.9$.

For the VIT model, experiments were conducted on Tiny ImageNet and CIFAR100 datasets, and a pre trained model was adopted, with a sampling rate of 10% for 100 clients. We set the batch size for local training to 16, the number of local epochs for one round of communication to 1, and the initial learning rate to search in $\{0.01, 0.03, 0.05, 0.1, 0.2, 0.3\}$. The learning rate decay for each round is 0.998, and the weight decay is 0.001. We searched for FedBCGD and FedAvgM α in $\{0.4, 0.5, 0.6, 0.7, 0.8, 0.9\}$, FedDC settings $\alpha = 0.01$, FedAdam setting $\alpha = 0.9$.

For the logical classification model, we set the batch size to 50, the number of local epochs in one round of communication to 1 on EMNIST. The initial learning rate is searched in $\{0.01, 0.03, 0.05, 0.1, 0.2, 0.3\}$, with a learning rate decay of 0.998 and a weight decay of 0.001 for each round. We searched for FedBCGD and FedAvgM α in $\{0.4, 0.5, 0.6, 0.7, 0.8, 0.9\}$, FedDC settings $\alpha = 0.01$, FedAdam setting $\alpha = 0.9$.

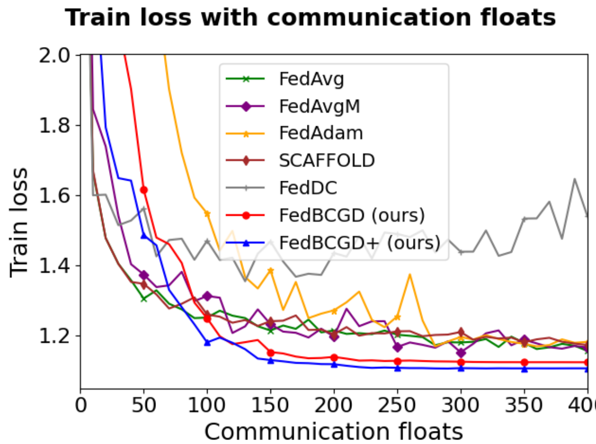
6.5 Results on Logistic Regression

We use a logistic regression model to verify the consistency between FedBCGD+'s practice and theory results. We conducted the classification tests on the EMNIST dataset by using strongly convex and non-convex loss function models. To test the performance of our algorithms, we use classical logistic regression problems, whose function has the following form:

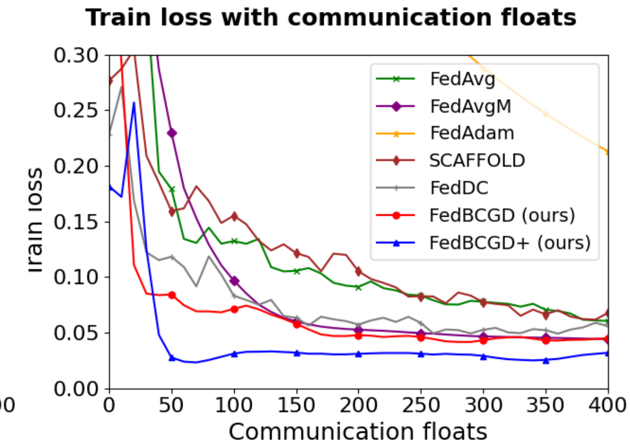
$$f(x) = \frac{1}{N}\sum_{i=1}^{N} \log\left(1 + \exp\left(-b_i a_i^\top x\right)\right) + \frac{\lambda}{2}\|x\|^2, \tag{129}$$

where $a_i \in \mathbb{R}^d$ and $b_i \in \{-1, +1\}$ are the data samples, and N is their total number. We set the regularization parameter $\lambda = 10^{-4}L$, where L is the smoothness constant.

From the results of logistic regression in Figure 8 (a), we observe that our FedBCGD and FedBCGD+ algorithms demonstrate faster convergence speed. Particularly, under the strong convexity condition with high client data heterogeneity, our FedBCGD+ algorithm exhibits even faster convergence compared to our FedBCGD, which aligns with our theoretical analysis.



(a) Strongly Convex



(b) Non-convex

Figure 1: (a) Logistic regression with $E=1$ and $\rho=0.1$. (b) The problem with non-convex loss, where $E=1$ and $\rho=0.1$. The number of blocks is set to $N=5$.

ERM with Non-Convex Loss: We also apply our algorithms to solve the regularized Empirical Risk Minimization (ERM) problem with non-convex sigmoid loss:

$$\min_{x\in\mathbb{R}^d} \frac{1}{n}\sum_{i=1}^{n} f_i(x) + \frac{\lambda}{2}\|x\|^2, \tag{130}$$

where $f_i(x) = 1/\left[1+\exp\left(b_i a_i^\top x\right)\right]$. Here, we consider binary classification on EMNIST. Note that we only consider classifying the first class in EMNIST.

From the results of the ERM problem in Figure 8 (b), we observe that our algorithms exhibit much faster convergence speeds than other algorithms. Moreover, in the case of high client data heterogeneity, FedBCGD+ demonstrates faster convergence than FedBCGD, which is consistent with our theoretical results.

6.6 Parameter Block Division

In this section we will show the parameter block division.

Parameter Block	Network Layers	Number of parameters
Block 1	conv3-64	4800
Block 2	conv3-64	102400
Block 3	FC-1600	614400
Block 4	FC-384	73728
Block 5	FC-192 (share)	1920

Table 1: The parameters block division of the LeNet-5 network.

Parameter Block	Network Layers	Number of parameters
Block 1	conv3-64	1728
Block 1	conv3-128	73728
Block 2	conv3-256	294912
Block 2	conv3-256	589824
Block 3	conv3-512	1179648
Block 3	conv3-512	2359296
Block 4	conv3-512	2359296
Block 4	conv3-512	2359296
Block 5	FC-2048	2359296
Block 5	FC-2048	2359296
Block share	FC-100	102400

Table 2: The parameter block division of the VGG-11 network.

Parameter Block	Network Layers	Number of parameters
Block 1	conv3-64	1728
Block 1	conv3-64	36864
Block 1	conv3-64	36864
Block 1	conv3-64	36864
Block 1	conv3-64	36864
Block 2	conv3-128	73728
Block 2	conv3-128	147456
Block 2	conv3-128	147456
Block 2	conv3-128	147456
Block 3	conv3-256	294912
Block 3	conv3-256	589824
Block 3	conv3-256	589824
Block 3	conv3-256	589824
Block 4	conv3-512	1179648
Block 4	conv3-512	2359296
Block 5	conv3-512	2359296
Block 5	conv3-512	2359296
Block share	FC-512 (share)	5120

Table 3: The parameters block division of the ResNet-18 network.

Parameter Block	Network Layers	Number of parameters
Block 1	conv3-64	1728
Block 1	conv3-64	36864
Block 1	conv3-128	73728
Block 1	conv3-128	147456
Block 1	conv3-256	294912
Block 1	conv3-256	589824
Block 1	conv3-256	589824
Block 1	conv3-256	589824
Block 2	conv3-512	1179648
Block 2	conv3-512	2359296
Block 3	conv3-512	2359296
Block 3	conv3-512	2359296
Block 4	conv3-512	2359296
Block 4	conv3-512	2359296
Block 5	conv3-512	2359296
Block 5	conv3-512	2359296
Block 5	FC-2048	1048576
Block 5	FC-2048	131072
Block share	FC-100	25600

Table 4: The parameters block division of the VGG-19 network.

Parameter Block	Network Layers	Number of parameters
Block 1	ViT-Block 1	14299520
Block 2	ViT-Block 2	14299520
Block 3	ViT-Block 3	14299520
Block 4	ViT-Block 4	14299520
Block 5	ViT-Block 5	14299520
Block share	FC-100	153600

Table 5: The parameters block division of the VGG-19 network.

7 APPENDIX G: FEDBCGD AND FEDBCGD+ ALGORITHMS

The proposed FedBCGD+ and FedBCGD algorithms as shown in Algorithms 2 and 3, respectively.

Algorithm 1 FedBCGD+

Initialize $\boldsymbol{x}_i^{0,0} = \boldsymbol{x}^{init}, \forall i \in [M]$.
Divide the model parameters $\boldsymbol{x}$ into N blocks.
for $r = 0, ..., R$ **do**
  Client:
  Sample clients $\mathcal{S} \subseteq \{1, \ldots, M\}, |\mathcal{S}| = NK$;
  Divide the sampled clients into N blocks;
  Communicate $(\boldsymbol{x}, \boldsymbol{c})$ to all clients $i \in \mathcal{S}$;
  for $j = 1, \ldots, N$ client blocks in parallel **do**
    for $k = 1, \ldots, K$ clients in parallel **do**
      Compute full batch gradient $\nabla f_{k,j}(\boldsymbol{x}^r)$;
      for $t = 1, \ldots, T$ local update **do**
        Compute mini-batch gradient $\nabla f_{k,j}\left(\boldsymbol{x}_{k,j}^{r,t}; \zeta\right)$ and $\nabla f_{k,j}(\boldsymbol{x}^r; \zeta)$;
        $\boldsymbol{x}_{k,j}^{r,t+1} = \boldsymbol{x}_{k,j}^{r,t} - \eta \nabla f_{k,j}\left(\boldsymbol{x}_{k,j}^{r,t}; \zeta\right) + \eta \boldsymbol{c} - \eta \boldsymbol{c}_{k,j} + \eta \nabla f_{k,j}(\boldsymbol{x}^r) - \eta \nabla f_{k,j}(\boldsymbol{x}^r; \zeta)$;
      end for
      $\boldsymbol{c}_{k,j}^{+} \leftarrow \nabla f_{k,j}(\boldsymbol{x}^r)$;
      Send $\boldsymbol{x}_{k,j,(j)}^{r,T}, \boldsymbol{x}_{k,j,s}^{r,T}$ and $\Delta \boldsymbol{c}_{(j)} = \boldsymbol{c}_{k,j,(j)}^{+} - \boldsymbol{c}_{k,j,(j)}, \Delta \boldsymbol{c}_s = \boldsymbol{c}_{k,j,s}^{+} - \boldsymbol{c}_{k,j,s}$ to server;
      $\boldsymbol{c}_i \leftarrow \boldsymbol{c}_i^{+}$;
    end for
  end for
  Server:
  for $j = 1, \ldots, N$ Blocks in parallel **do**
    Block j computes,
    $\boldsymbol{x}_{(j)}^r = \frac{1}{K} \sum_{k=1}^{K} \boldsymbol{x}_{k,j,(j)}^{r,T}$;
    $v_{(j)}^r = \lambda v_{(j)}^{r-1} + \boldsymbol{x}_{(j)}^r - \boldsymbol{x}_{(j)}^{r-1}$;
    $\boldsymbol{x}_{(j)}^r = \boldsymbol{x}_{(j)}^r + v_{(j)}^r$,
    $\boldsymbol{c}_{(j)} = \boldsymbol{c}_{(j)} + \frac{1}{M} \sum_{k=1}^{K} \Delta \boldsymbol{c}_{k,j,(j)}$;
  end for
  $\boldsymbol{x}_s^r = \frac{1}{NK} \sum_{j=1}^{N} \sum_{k=1}^{K} \boldsymbol{x}_{k,j,s}^{r,T}$;
  $v_s^r = \lambda v_s^r + \boldsymbol{x}_s^{r-1} - \boldsymbol{x}_s^{r-1}$;
  $\boldsymbol{x}_s^r = \boldsymbol{x}_s^r + v_s^r$;
  $\boldsymbol{c}_s = \boldsymbol{c}_s + \frac{1}{MN} \sum_{j=1}^{N} \sum_{k=1}^{K} \Delta \boldsymbol{c}_{k,j,s}$;
  $\boldsymbol{x}^r = \left[\boldsymbol{x}_{(1)}^{r\top}, \ldots, \boldsymbol{x}_{(N)}^{r\top}, \boldsymbol{x}_s^{r\top}\right]^\top$;
  $\boldsymbol{v}^r = \left[\boldsymbol{v}_{(1)}^{r\top}, \ldots, \boldsymbol{v}_{(N)}^{r\top}, \boldsymbol{v}_s^{r\top}\right]^\top$;
  $\boldsymbol{c} = \left[\boldsymbol{c}_{(1)}^\top, \ldots, \boldsymbol{c}_{(N)}^\top, \boldsymbol{c}_s^\top\right]^\top$;
end for