

1 **Supplementary Materials: FedBCGD: Communication-Efficient** 59
2 **Accelerated Block Coordinate Gradient Descent for Federated** 60
3 **Learning** 61
4 62
5 63
6 64
7 65
8 66
9 67
10 68
11 69
12 70
13 71
14 72
15 73
16 74
17 75
18 76
19 77
20 78
21 79
22 80
23 81
24 82
25 83
26 84
27 85
28 86
29 87
30 88
31 89
32 90
33 91
34 92
35 93
36 94
37 95
38 96
39 97
40 98
41 99
42 100
43 101
44 102
45 103
46 104
47 105
48 106
49 107
50 108
51 109
52 110
53 111
54 112
55 113
56 114
57 115
58 116

Anonymous Authors

117 1 APPENDIX A: BASIC ASSUMPTIONS AND NOTATIONS 175

118 1.1 Basic Assumptions 176

119 Before giving our theoretical results, we first present the common assumptions. 177

120 ASSUMPTION 1 (CONVEXITY). f_i is μ -strongly-convex for all $i \in [M]$, i.e., 180

$$123 f_i(\mathbf{y}) \geq f_i(\mathbf{x}) + \langle \nabla f_i(\mathbf{x}), \mathbf{y} - \mathbf{x} \rangle + \frac{\mu}{2} \|\mathbf{y} - \mathbf{x}\|^2 \quad (1) \\ 124$$

125 for all \mathbf{x}, \mathbf{y} in its domain and $i \in [M]$. We allow $\mu = 0$, which corresponds to general convex functions. 182

126 ASSUMPTION 2 (SMOOTHNESS). The gradient of the loss function is Lipschitz continuous with constant β , for all $\mathbf{x}_1, \mathbf{x}_2 \in \mathbb{R}^d$ 184

$$128 \|\nabla f(\mathbf{x}_1) - \nabla f(\mathbf{x}_2)\| \leq \beta \|\mathbf{x}_1 - \mathbf{x}_2\|. \quad (2) \\ 129$$

130 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 187
131 $\mathbf{x} \in \mathbb{R}^d$ 188

$$132 \mathbb{E}_{\zeta|\mathbf{x}} [\nabla f_i(\mathbf{x}; \zeta)] = \nabla f_i(\mathbf{x}). \quad (3) \\ 133$$

134 We also can get: 191

$$135 \mathbb{E}_{\zeta|\mathbf{x}} [\|\nabla f_i(\mathbf{x}; \zeta) - \nabla f_i(\mathbf{x})\|^2] \leq \sigma^2. \quad (4) \\ 136$$

137 ASSUMPTION 4 (BOUNDED HETEROGENEITY). The dissimilarity of $f_i(\mathbf{x})$ and $f(\mathbf{x})$ is bounded as follows: 194

$$138 \frac{1}{M} \sum_{i=1}^M \|\nabla f_i(\mathbf{x}) - \nabla f(\mathbf{x})\|^2 \leq G^2. \quad (5) \\ 139$$

140 ASSUMPTION 5 (STOCHASTIC GRADIENT SMOOTHNESS). The gradient of the loss function is Lipschitz continuous with constant β , for all 198
141 $\mathbf{x}_1, \mathbf{x}_2 \in \mathbb{R}^d$ 199

$$143 \|\nabla f(\mathbf{x}_1; \zeta) - \nabla f(\mathbf{x}_2; \zeta)\| \leq \beta \|\mathbf{x}_1 - \mathbf{x}_2\|. \quad (6) \\ 144$$

145 Assumption 2 bounds the variance of stochastic gradients, which is common in stochastic optimization analysis [?]. Assumption 3 bounds 203
146 the gradient difference between global and local loss functions, which is a widely-used approach to characterize client heterogeneity in 204
147 federated optimization literature [? ?]. Assumption 5 is a necessary assumption in stochastic gradient noise reduction, an assumption that is 205
148 used only in the proof of the convergence speed of the FedBCGD+ algorithm. 206

149 1.2 Notation 207

150 We first define the notations to be used in analyzing the convergence properties of our algorithms. 208

151 1. \mathbf{x}^r is the r communication rounds global model. 209

152 2. $\mathbf{x}_{(j)}^r$ is the j -th block of \mathbf{x} , so that $\mathbf{x}^r = [\mathbf{x}_{(1)}^{r\top}, \dots, \mathbf{x}_{(N)}^{r\top}]^\top$. Note that $\mathbf{x}_{(j)}^r$ is a virtual vector. It is realized at a hub j every r iterations, 211
153 but we will study the evolution of this virtual vector in every iteration. 212

154 3. $\mathbf{x}_{k,j}^r \in \mathbb{R}^d$ are the local versions of the coordinates of the weight vector $\mathbf{x}_{(j)}^r$ that each client k if hub j updates. 214

155 4. \mathbf{x}^* is the minimum value of the function $f(\mathbf{x})$. 215

156 5. $\mathbf{x}_{k,j,(j)}$ is the j -th block of $\mathbf{x}_{k,j}$ at client k in silo j , so that $\mathbf{x}_{(j)} = \frac{1}{K} \sum_{k=1}^K \mathbf{x}_{k,j,(j)}$. 216

157 6. $\mathbf{y}_{k,j}^{r,t}$ is the local parameter vector that client j in silo k at iteration t . 217

158 7. $\nabla_{(j)} f_{k,j} (\mathbf{y}_{k,j}; \zeta)$ is the partial derivative of $f(\mathbf{x})$ with respect to coordinate block j , computed at client k in silo j using the coordinates 219
159 and rows at client k corresponding to minibatch ζ . 220

160 8. $\mathbf{G}^r = \left[\left(\mathbf{G}_{(1)}^r \right)^\top, \dots, \left(\mathbf{G}_{(N)}^r \right)^\top \right]^\top$, where $\mathbf{G}_{(j)}^r = \frac{1}{K} \sum_{k=1}^K \sum_{t=1}^T \nabla_{(j)} f_{k,j} (\mathbf{y}_{k,j}^{r,t}; \zeta)$. 221

161 It should be noted that components on \mathbf{x} , i.e., $\mathbf{x}_{(j)}$ are realized every T iterations when the hubs communicate with clients and with other 223
162 hubs, but we will study the evolution of these virtual vectors at each iteration. Therefore, based on the above definitions, assumptions and 224
163 our algorithms, we can express the evolution of the virtual global parameter/weight vector in the following forms: 225

$$164 \mathbf{x}^r = \begin{bmatrix} \mathbf{x}_{(1)}^r \\ \mathbf{x}_{(2)}^r \\ \vdots \\ \mathbf{x}_{(N)}^r \end{bmatrix} = \frac{1}{K} \begin{bmatrix} \sum_{k=1}^K \mathbf{x}_{k,1,(1)}^r \\ \sum_{k=1}^K \mathbf{x}_{k,2,(2)}^r \\ \vdots \\ \sum_{k=1}^K \mathbf{x}_{k,N,(N)}^r \end{bmatrix} \quad (7) \\ 165$$

$$\mathbf{x}^{r+1} = \mathbf{x}^r - \frac{\eta}{K} \begin{bmatrix} \sum_{k=1}^K \sum_{t=1}^T \nabla_{(1)} f_{k,1} \left(y_{k,1}^{r,t}; \zeta \right) \\ \sum_{k=1}^K \sum_{t=1}^T \nabla_{(2)} f_{k,2} \left(y_{k,2}^{r,t}; \zeta \right) \\ \vdots \\ \sum_{k=1}^K \sum_{t=1}^T \nabla_{(N)} f_{k,N} \left(y_{k,N}^{r,t}; \zeta \right). \end{bmatrix} \quad (8)$$

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

$$\mathbf{G}^r = \frac{1}{K} \begin{bmatrix} \sum_{k=1}^K \sum_{t=1}^T \nabla_{(1)} f_{k,1} \left(y_{k,1}^{r,t}; \zeta \right) \\ \sum_{k=1}^K \sum_{t=1}^T \nabla_{(2)} f_{k,2} \left(y_{k,2}^{r,t}; \zeta \right) \\ \vdots \\ \sum_{k=1}^K \sum_{t=1}^T \nabla_{(N)} f_{k,N} \left(y_{k,N}^{r,t}; \zeta \right), \end{bmatrix} \quad (9)$$

$$\mathbb{E}_{\mathcal{S}} [\mathbf{G}^r] = \frac{1}{M} \begin{bmatrix} \sum_{i=1}^M \sum_{t=1}^T \nabla_{(1)} f_i \left(y_i^{r,t} \right) \\ \sum_{i=1}^M \sum_{t=1}^T \nabla_{(2)} f_i \left(y_i^{r,t} \right) \\ \vdots \\ \sum_{i=1}^M \sum_{t=1}^T \nabla_{(N)} f_i \left(y_i^{r,t} \right). \end{bmatrix} \quad (10)$$

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

$$\mathbf{x}^r = \left[\mathbf{x}_{(1)}^{r\top}, \dots, \mathbf{x}_{(N)}^{r\top} \right]^\top \quad (11)$$

where each $\mathbf{x}_{(j)}^r$ is the block of \mathbf{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. \bar{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

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

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

$$\begin{aligned} \mathbb{E} [f(\bar{z}^R)] - f(x^*) &\leq \frac{\beta^{\frac{3}{2}} d_0}{\alpha R} + \frac{\left(6144\beta^2 G^2 + \frac{3072}{T}\beta^2\sigma^2\right)}{\alpha^2 R} + \frac{\left[32 \left[\left(1 - \frac{K}{M}\right) \frac{1}{K} \right] G^2 + 32 \frac{\sigma^2}{KT}\right]^{\frac{1}{2}} d_0^{\frac{1}{2}}}{\sqrt{R}} \\ &\quad + \frac{\left(384\beta G^2 + \frac{192\beta}{T}\sigma^2\right)^{\frac{1}{3}} d_0^{\frac{2}{3}}}{\alpha^{\frac{2}{3}} R^{\frac{2}{3}}} \end{aligned} \quad (13)$$

349 **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 407

$$\begin{aligned} 350 \quad & \frac{1}{R} \sum_{r=1}^R \|\nabla f(x^r)\|^2 \leq \frac{16\beta d_0}{TK\alpha R} + \frac{2\sqrt{d_0}}{\sqrt{RTM}} \left(\frac{8\beta}{K} \left(1 - \frac{K}{M}\right) G^2 + \frac{8\beta\sigma^2}{TK} \left(1 - \frac{K}{M}\right) + \frac{8\beta}{TM} \sigma^2 \right)^{\frac{1}{2}} \\ 351 \quad & + 2 \left(\frac{d_0}{R} \right)^{\frac{2}{3}} \left[\frac{384\beta^2}{\alpha^2} G^2 + \frac{92\beta^2}{T} \frac{\sigma^2}{\alpha^2} + \frac{(16\gamma^2\beta^2)}{TM} \sigma^2 + \frac{(16\gamma^2\beta^2)}{TK} \sigma^2 \left(1 - \frac{K}{M}\right) + \frac{16\gamma^2\beta^2}{K} \left(1 - \frac{K}{M}\right) G^2 \right]^{\frac{1}{3}} \\ 352 \quad & + 2 \left(\frac{d_0}{R} \right)^{\frac{3}{4}} \left[\frac{4608}{\alpha^2} \frac{\beta^3}{K} \left(1 - \frac{K}{M}\right) G^2 + \frac{1152\beta^3}{KT\alpha^2} \left(1 - \frac{K}{M}\right) \sigma^2 \right]^{\frac{1}{4}} \\ 353 \quad & + 2 \left(\frac{d_0}{R} \right)^{\frac{4}{5}} \left[\frac{9216}{\alpha^2} \frac{\gamma^2\beta^4}{K} \left(1 - \frac{K}{M}\right) G^2 + \frac{2304\beta^4}{K} \frac{\gamma^2}{\alpha^2 T} \left(1 - \frac{K}{M}\right) \sigma^2 \right]^{\frac{1}{5}}. \end{aligned} \quad (14)$$

362 **THEOREM 2 (CONVERGENCE RATES OF FEDBCGD+).** Suppose that each function $\{f_i\}$ satisfies Assumptions 1, 2, and 3. Then, in each of the 420
363 following cases, there exist weights $\{w_r\}$ and local step-sizes η , the output of FedBCGD+ (i.e., \bar{z}^R) satisfies the following inequalities. 421

364 **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 422

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

365 **2. Case of General convex:** Each f_i satisfies Assumption 1 for $\mu = 0$, $\tilde{\eta} \leq \frac{1}{\beta}$ then 423

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

374 **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 432

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

380 where $\tilde{D}^2 := \left(\|x^0 - x^*\|^2 + \frac{1}{2N\beta^2} \sum_{i=1}^N \|c_i^0 - \nabla f_i(x^*)\|^2 \right)$ and $F := (f(x_0) - f(x^*))$. 438

3 APPENDIX C: MAIN LEMMAS

383 In this section, we prove some main lemmas, which play key roles for the proofs of Theorems 1-3. 441

384 **LEMMA 1.** The following holds for any β -smooth and μ -strongly convex function h , and any x, y, z in the domain of h : 442

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

388 *Proof.* Given any x, y , and z , we get the following two inequalities using smoothness and strong convexity of h : 446

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

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

393 Furthermore, applying the relaxed triangle inequality, we can get 451

$$\frac{\mu}{2} \|y - x\|^2 \geq \frac{\mu}{4} \|y - z\|^2 - \frac{\mu}{2} \|x - z\|^2. \quad (21)$$

397 Combining all the inequalities together, we have 455

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

400 The lemma follows since $\beta \geq \mu$. 458

402 **LEMMA 2 (BOUNDRING HETEROGENEITY).** Recall our bound on the gradient dissimilarity: 460

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

465 If $\{f_i\}$ are convex, we can relax the assumption to 523

$$466 \quad 467 \quad 468 \quad \frac{1}{M} \sum_{i=1}^M \|\nabla f_i(x)\|^2 \leq G^2 + 2\beta (f(x) - f^\star). \quad (24)$$

469 *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 $\mathbf{a}_i \in \mathbb{R}^d$, $\bar{\mathbf{a}} = \frac{1}{n} \sum_{i=1}^n \mathbf{a}_i$, 527

$$470 \quad 471 \quad 472 \quad \frac{1}{M} \sum_{i=1}^M \|\nabla f_i(x) - \nabla f(x)\|^2 \leq G^2, \quad (25)$$

$$473 \quad 474 \quad 475 \quad \begin{aligned} 476 \quad 477 \quad 478 \quad & \frac{1}{M} \sum_{i=1}^M \|\nabla f_i(x)\|^2 \leq \|\nabla f(x)\|^2 + G^2 \\ 479 \quad 480 \quad 481 \quad & \leq \|\nabla f(x) - \nabla f(x^\star)\|^2 + G^2 \\ 482 \quad 483 \quad 484 \quad & \leq \frac{1}{M} \sum_{i=1}^M \|\nabla f_i(x) - \nabla f_i(x^\star)\|^2 + G^2 \\ 485 \quad 486 \quad 487 \quad & \leq 2\beta (f(x) - f^\star) + G^2. \end{aligned} \quad (26)$$

488 LEMMA 3. (**Relaxed triangle inequality**). Let $\{v_1, \dots, v_\tau\}$ be τ vectors in \mathbb{R}^d . Then the following inequalities are true: 542

- 488 1. $\|v_i + v_j\|^2 \leq (1+a) \|v_i\|^2 + \left(1 + \frac{1}{a}\right) \|v_j\|^2$ for any $a > 0$, and 543
- 488 2. $\|\sum_{i=1}^\tau v_i\|^2 \leq \tau \sum_{i=1}^\tau \|v_i\|^2$. 544

489 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. 546

$$490 \quad 491 \quad 492 \quad \mathbb{E} \left\| \frac{1}{K} \sum_{i=1}^K \nabla f_i(x) \right\|^2 \leq \mathbb{E} \|\nabla f(x)\|^2 + \mathbb{E} \left(1 - \frac{K}{M} \right) \frac{1}{KM} \sum_{i=1}^M \|\nabla f_i(x)\|^2, \quad (27)$$

$$493 \quad 494 \quad 495 \quad \mathbb{E} \left\| \frac{1}{K} \sum_{i=1}^K \nabla f_i(x) \right\|^2 \leq \frac{1}{M} \sum_{i=1}^M \|\nabla f_i(x)\|^2. \quad (28)$$

496 *Proof.* Define \mathbb{I}_i as the random variable which indicates client i is selected in the r -th global epoch. 555

$$\begin{aligned} 497 \quad 498 \quad 499 \quad & \mathbb{E} \left\| \frac{1}{K} \sum_{i=1}^K \nabla f_i(x) \right\|^2 \\ 500 \quad 501 \quad 502 \quad & = \mathbb{E} \left\| \frac{1}{K} \sum_{i=1}^M \nabla f_i(x) \mathbb{I}_i \right\|^2 \\ 503 \quad 504 \quad 505 \quad & = \mathbb{E} \left\langle \frac{1}{K} \sum_{i=1}^M \nabla f_i(x) \mathbb{I}_i, \frac{1}{K} \sum_{i=1}^M \nabla f_i(x) \mathbb{I}_i \right\rangle \\ 506 \quad 507 \quad 508 \quad & = \mathbb{E} \left\langle \frac{1}{K^2} \sum_{i,j \in [M], i \neq j} \langle \nabla f_i(x), \nabla f_j(x) \rangle \mathbb{E} [\mathbb{I}_i \mathbb{I}_j] + \sum_{i \in [M]} \langle \nabla f_i(x), \nabla f_i(x) \rangle \mathbb{E} [\mathbb{I}_i] \right\rangle \\ 509 \quad 510 \quad 511 \quad & = \mathbb{E} \left\langle \frac{1}{K^2} \sum_{i,j \in [M], i \neq j} \frac{K(K-1)}{M(M-1)} \langle \nabla f_i(x), \nabla f_j(x) \rangle + \sum_{i \in [M]} \frac{K}{M} \langle \nabla f_i(x), \nabla f_i(x) \rangle \right\rangle \\ 512 \quad 513 \quad 514 \quad & = \mathbb{E} \left\langle \frac{1}{K^2} \sum_{i,j \in [M]} \frac{K(K-1)}{M(M-1)} \langle \nabla f_i(x), \nabla f_j(x) \rangle + \sum_{i \in [M]} \frac{K(M-K)}{M(M-1)} \langle \nabla f_i(x), \nabla f_i(x) \rangle \right\rangle \\ 515 \quad 516 \quad 517 \quad & \leq \mathbb{E} \|\nabla f(x)\|^2 + \mathbb{E} \left(1 - \frac{K}{M} \right) \frac{1}{KM} \sum_{i \in [M]} \|\nabla f_i(x)\|^2 \\ 518 \quad 519 \quad 520 \quad & \leq \frac{1}{M} \sum_{i \in [M]} \|\nabla f_i(x)\|^2. \end{aligned} \quad (29)$$

581 We will now proceed to the second part of our lemma's exposition. 639

$$582 \quad 640 \\ 583 \quad \mathbb{E} \left\| \frac{1}{K} \sum_{i=1}^K \nabla f_i(x) \right\|^2 \leq \frac{1}{M} \sum_{i=1}^M \|\nabla f_i(x)\|^2. 641 \\ 584 \quad 642 \\ 585$$

586 *Proof :* 643

$$587 \quad \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\|^2 644 \\ 588 \quad 645 \\ 589 \quad = \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 646 \\ 590 \quad 647 \\ 591 \quad = \mathbb{E} \frac{1}{K^2} \left[\sum_{i,j \in [M], i \neq j} \langle \nabla f_i(x), \nabla f_j(x) \rangle \mathbb{E} [\mathbb{I}_i \mathbb{I}_j] + \sum_{i \in [M]} \langle \nabla f_i(x), \nabla f_i(x) \rangle \mathbb{E} [\mathbb{I}_i] \right] 648 \\ 592 \quad 649 \\ 593 \quad = \mathbb{E} \frac{1}{K^2} \left[\sum_{i,j \in [M], i \neq j} \frac{K(K-1)}{M(M-1)} \langle \nabla f_i(x), \nabla f_j(x) \rangle + \sum_{i \in [M]} \frac{K}{M} \langle \nabla f_i(x), \nabla f_i(x) \rangle \right] 650 \\ 594 \quad 651 \\ 595 \quad = \mathbb{E} \frac{1}{K^2} \left[\sum_{i,j \in [M]} \frac{K(K-1)}{M(M-1)} \langle \nabla f_i(x), \nabla f_j(x) \rangle + \sum_{i \in [M]} \frac{K(M-K)}{M(M-1)} \langle \nabla f_i(x), \nabla f_i(x) \rangle \right] 652 \\ 596 \quad 653 \\ 597 \quad = \mathbb{E} \frac{1}{K^2} \left[\sum_{i,j \in [M]} \frac{K(K-1)}{M(M-1)} \langle \nabla f_i(x), \nabla f_j(x) \rangle + \sum_{i \in [M]} \frac{K(M-K)}{M(M-1)} \langle \nabla f_i(x), \nabla f_i(x) \rangle \right] 654 \\ 598 \quad 655 \\ 599 \quad = \mathbb{E} \frac{1}{K^2} \left[\sum_{i,j \in [M]} \frac{K(K-1)}{M(M-1)} \langle \nabla f_i(x), \nabla f_j(x) \rangle + \sum_{i \in [M]} \frac{K(M-K)}{M(M-1)} \langle \nabla f_i(x), \nabla f_i(x) \rangle \right] 656 \\ 600 \quad 657 \\ 601 \quad \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)\|^2 658 \\ 602 \quad 659 \\ 603 \quad \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)\|^2 660 \\ 604 \quad 661 \\ 605 \quad \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)\|^2 662 \\ 606 \quad 663 \\ 607 \quad = \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)\|^2 664 \\ 608 \quad 665 \\ 609 \quad = \frac{1}{M} \sum_{i \in [M]} \mathbb{E} \|\nabla f_i(x)\|^2. 666 \\ 610 \quad 667 \\ 611 \quad 668 \\ 612 \quad 669 \\ 613 \quad 670 \\ 614 \quad 671 \\ 615 \quad 672 \\ 616 \quad 673$$

LEMMA 5 (BOUNDED DRIFT). 674

$$617 \quad \sum_{i=1}^M \sum_{t=1}^T \mathbb{E} \|y_i^{r,t} - x^r\|^2 \leq 6T^3 \eta^2 \sum_{i=1}^M \|\nabla f_i(x^r)\|^2 + 3MT^2 \eta^2 \sigma^2. 675 \\ 618 \quad 676 \\ 619 \quad 677$$

620 *Proof.* 678

621 679

$$622 \quad \mathbb{E} \|y_i^{r,t-1} - x^r - \eta \nabla f_i(y_i^{r,t-1}; \zeta)\|^2 680 \\ 623 \quad 681 \\ 624 \quad \leq \mathbb{E} \|y_i^{r,t-1} - x^r - \eta \nabla f_i(y_i^{r,t-1})\|^2 + \eta^2 \sigma^2 682 \\ 625 \quad 683 \\ 626 \quad \stackrel{a}{\leq} \left(1 + \frac{1}{T-1} \right) \mathbb{E} \|y_i^{r,t-1} - x^r\|^2 + T\eta^2 \|\nabla f_i(y_i^{r,t-1})\|^2 + \eta^2 \sigma^2 684 \\ 627 \quad 685 \\ 628 \quad = \left(1 + \frac{1}{T-1} \right) \mathbb{E} \|y_i^{r,t-1} - x^r\|^2 + T\eta^2 \|\nabla f_i(y_i^{r,t-1}) - \nabla f_i(x^r) + \nabla f_i(x^r)\|^2 + \eta^2 \sigma^2 686 \\ 629 \quad 687 \\ 630 \quad \leq \left(1 + \frac{1}{T-1} \right) \mathbb{E} \|y_i^{r,t-1} - x^r\|^2 + 2T\eta^2 \|\nabla f_i(y_i^{r,t-1}) - \nabla f_i(x^r)\|^2 + 2T\eta^2 \|\nabla f_i(x^r)\|^2 + \eta^2 \sigma^2 688 \\ 631 \quad 689 \\ 632 \quad \leq \left(1 + \frac{1}{T-1} + 2T\eta^2 \beta^2 \right) \mathbb{E}^2 \|y_i^{r,t-1} - x^r\|^2 + 2T\eta^2 \|\nabla f_i(x^r)\|^2 + \eta^2 \sigma^2 690 \\ 633 \quad 691 \\ 634 \quad \leq \left(1 + \frac{2}{(T-1)} \right) \mathbb{E}^2 \|y_i^{r,t-1} - x^r\|^2 + 2T\eta^2 \|\nabla f_i(x^r)\|^2 + \eta^2 \sigma^2, 692 \\ 635 \quad 693 \\ 636 \quad 694 \\ 637 \quad 695 \\ 638$$

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

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

So, we can get

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

LEMMA 6. *The variance of G^r can be bounded by the following inequality*

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

Proof.

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

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 $W_R := \sum_{r=1}^{R+1} w_r$*

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

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 $w_r = \frac{1}{r}$ such that,*

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

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

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

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

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

813 4 APPENDIX D: PROOF OF THEOREM 1 871

814 4.1 1. The rate of strongly convex and smooth convergence: 872

815 We outline the FEDBCGD algorithm in Algorithm 1. In round r , we perform the following updates: 873

$$817 \quad v^r = \lambda v^{r-1} + \Delta x^{r-1}, \Delta x^{r-1} = \eta \mathbf{G}^r, \quad (42) \quad 875$$

$$818 \quad x^r = x^{r-1} + v^{r-1}. \quad (43) \quad 876$$

819 Before giving the convergence analysis of Theorem 3, we first present the following lemma. 877

820 LEMMA 10. Let $z^r = x^r + \gamma (x^r - x^{r-1})$, $\gamma = \frac{\lambda}{1-\lambda}$, we can get 878

$$824 \quad z^{r+1} = z^r - \frac{1}{1-\lambda} \eta \mathbf{G}^r. \quad (44) \quad 882$$

826 Proof.

$$\begin{aligned} 829 \quad z^{r+1} &= x^{r+1} + \gamma (x^{r+1} - x^r) \\ 830 &= x^r + v^{r+1} + \gamma (v^{r+1}) \\ 831 &= z^r - \gamma (v^r) + v^{r+1} + \gamma (v^{r+1}) \\ 832 &= z^r - \gamma v^r + (1+\gamma)v^{r+1} \\ 833 &= z^r - \gamma v^r + (1+\gamma)(\lambda v^r - \eta \mathbf{G}^r) \\ 834 &\stackrel{a}{=} z^r + (-\gamma + (1+\gamma)\beta)v^r + (1+\gamma)(-\eta \mathbf{G}^r) \\ 835 &= z^r - \eta(1+\gamma)\mathbf{G}^r \\ 836 &= z^r - \frac{1}{1-\lambda} \eta \mathbf{G}^r, \end{aligned} \quad (45) \quad 893$$

842 with the equality $\stackrel{a}{=}$, we let $(-\gamma + (1+\gamma)\lambda) = 0$, $\gamma = \frac{\lambda}{1-\lambda}$. We complete the proof. 900

844 4.2 The proof of Theorem 1 902

845 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 903

$$\begin{aligned} 847 \quad \mathbb{E} \|z^{r+1} - x^\star\|^2 &= \mathbb{E} \|z^{r+1} - x^\star\|^2 \\ 848 &\leq \mathbb{E} \|z^r - x^\star\|^2 + \eta \alpha \underbrace{\mathbb{E} \langle -\mathbf{G}^r, z^r - x^\star \rangle}_{C_1} + \eta^2 \alpha^2 \underbrace{\mathbb{E} \|\mathbf{G}^r\|^2}_{C_2}. \end{aligned} \quad (46) \quad 904$$

852 The term C_1 can be bounded by using perturbed strong-convexity (Lemma 1) with $h = f_k$, $\mathbf{x} = y_k^{r,t}$, $\mathbf{y} = x^\star$, and $\mathbf{z} = z^r$ to get 910

$$\begin{aligned} 854 \quad C_1 &= -\mathbb{E} \langle \mathbf{G}^r, z^r - x^\star \rangle \\ 855 &= -\sum_{j=1}^N \left\langle \frac{1}{M} \sum_{i=1}^M \sum_{t=1}^T \nabla_{(j)} f_i(y_i^{r,t}), z_{(j)}^r - x_{(j)}^\star \right\rangle \\ 856 &= -\left\langle \frac{1}{M} \sum_{k=1}^M \sum_{t=1}^T \nabla f_i(y_i^{r,t}), z^r - x^\star \right\rangle \\ 857 &\leq -\frac{1}{M} \sum_{k=1}^M \sum_{t=1}^T (f_i(z^r) - f_i(x^\star) - \beta \|y_i^{r,t} - z^r\|^2 + \frac{\mu}{4} \|z^r - x^\star\|^2) \\ 858 &\leq T \left(-f(z^r) + f(x^\star) - \frac{\mu}{4} \|z^r - x^\star\|^2 \right) + \frac{\beta}{M} \sum_{i=1}^M \sum_{t=1}^T \|y_i^{r,t} - z^r\|^2. \end{aligned} \quad (47) \quad 919$$

869 The term C_2 can be bounded by using Lemma 6 in $\stackrel{a}{\leq}$, Lemma 3 in $\stackrel{b}{\leq}$, Lemma 4 and Lemma 2 in $\stackrel{c}{\leq}$. 926

$$\begin{aligned}
C_2 &= \mathbb{E} \|\mathbf{G}^r\|^2 && 997 \\
&= \sum_{j=1}^N \mathbb{E} \left\| \frac{1}{K} \sum_{k=1}^K \sum_{t=1}^T \nabla_{(j)} f_{k,j} (y_{k,j}^{r,t}; \zeta) \right\|^2 && 999 \\
&= \sum_{j=1}^N \mathbb{E} \left\| \frac{1}{K} \sum_{k=1}^K \sum_{t=1}^T \nabla_{(j)} f_{k,j} (y_{k,j}^{r,t}; \zeta) - \frac{1}{K} \sum_{k=1}^K \sum_{t=1}^T \nabla_{(j)} f_{k,j} (x^r) + \frac{1}{K} \sum_{k=1}^K \sum_{t=1}^T \nabla_{(j)} f_{k,j} (x^r) \right\|^2 && 1000 \\
&\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} (y_{k,j}^{r,t}; \zeta) - \frac{1}{K} \sum_{k=1}^K \sum_{t=1}^T \nabla_{(j)} f_{k,j} (x^r) \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} (x^r) \right\|^2 && 1002 \\
&\stackrel{\text{a}}{\leq} 2 \frac{T}{M} \sum_{j=1}^N \sum_{i=1}^M \sum_{t=1}^T \left\| \nabla_{(j)} f_i (y_i^{r,t}; \zeta) - \nabla_{(j)} f_i (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} (x^r) \right\|^2 && 1005 \\
&\leq 4 \frac{T}{M} \sum_{i=1}^M \sum_{t=1}^T \left\| \nabla f_i (y_i^{r,t}) - \nabla f_i (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} (x^r) \right\|^2 + 4 \frac{T}{K} \sigma^2 && 1006 \\
&\leq 4 \frac{T}{M} \sum_{i=1}^M \sum_{t=1}^T \left\| \nabla_{(j)} f_i (y_i^{r,t}) - \nabla_{(j)} f_i (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} (x^r) \right\|^2 + 4 \frac{T}{K} \sigma^2 && 1007 \\
&\leq 4 \frac{T\beta^2}{M} \sum_{i=1}^M \sum_{t=1}^T \|y_i^{r,t} - x^r\|^2 + \sum_{j=1}^N 2T^2 \mathbb{E} \left\| \frac{1}{K} \sum_{k=1}^K \nabla_{(j)} f_{k,j} (x^r) \right\|^2 + 4 \frac{T}{K} \sigma^2 && 1008 \\
&\leq 4 \frac{\beta^2}{M} \sum_{i=1}^M \sum_{t=1}^T \|y_i^{r,t} - x^r\|^2 + \sum_{j=1}^N 2\mathbb{E} \left\| \frac{T}{K} \sum_{k=1}^K \nabla_{(j)} f_{k,j} (x^r) \right\|^2 + 4 \frac{T}{K} \sigma^2 && 1009 \\
&\leq 4 \frac{T\beta^2}{M} \sum_{i=1}^M \sum_{t=1}^T \|y_i^{r,t} - x^r\|^2 + 2T^2 \sum_{j=1}^N \mathbb{E} \left\| \frac{1}{K} \sum_{k=1}^K \nabla_{(j)} f_{k,j} (x^r) - \nabla_{(j)} f (x^r) + \nabla_{(j)} f (x^r) \right\|^2 + 4 \frac{T}{K} \sigma^2 && 1010 \\
&\stackrel{\text{b}}{\leq} 4 \frac{T\beta^2}{M} \sum_{k=1}^M \sum_{t=1}^T \|y_i^{r,t} - x^r\|^2 + 2T^2 \|\nabla f (x^r)\|^2 + 2 \left(1 - \frac{K}{M}\right) T^2 \frac{1}{KM} \sum_{i=1}^M \|\nabla f_i (x^r)\|^2 + 4 \frac{T}{K} \sigma^2 && 1011 \\
&\stackrel{\text{c}}{\leq} 4 \frac{T\beta^2}{M} \sum_{k=1}^M \sum_{t=1}^T \|y_i^{r,t} - x^r\|^2 + 4T^2 \beta (f(x^r) - f(x^*)) + 2 \left(1 - \frac{K}{M}\right) \frac{T^2}{K} \left(G^2 + 2\beta (f(x^r) - f(x^*))\right) + 4 \frac{T}{K} \sigma^2, && 1012 \\
\end{aligned} \tag{48}$$

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

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

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

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

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

where the inequality $\stackrel{a}{\leq}$ follows $\left[\alpha \eta T - 6T^3 \alpha^2 \eta^2 2\beta (\eta \beta + 4T \beta^2 \eta^2) - \left(4T^2 \beta \alpha^2 \eta^2 + \left(1 - \frac{K}{M}\right) 4T^2 \alpha^2 \eta^2 \frac{1}{K} \beta\right)\right] \geq \frac{1}{4} \alpha \eta T$. 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

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

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:

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

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

$$\begin{aligned}
& \mathbb{E} [f(\bar{z}^R)] - f(x^*) \\
& \leq \frac{\beta^{\frac{3}{2}} d_0}{\alpha R} + \frac{\left(6144\beta^2 G^2 + \frac{3072}{T} \beta^2 \sigma^2\right)}{\alpha^2 R} + \frac{\left[32 \left[\left(1 - \frac{K}{M}\right) \frac{1}{K} \right] G^2 + 64 \frac{\sigma^2}{KT}\right]^{\frac{1}{2}} d_0^{\frac{1}{2}}}{\sqrt{R}} \\
& + \frac{\left(\frac{384}{\alpha^2} \beta G^2 + \frac{192\beta}{T\alpha^2} \sigma^2\right)^{\frac{1}{3}} d_0^{\frac{2}{3}}}{\alpha^{\frac{2}{3}} R^{\frac{2}{3}}}.
\end{aligned} \tag{54}$$

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

From the smoothness of the function, we can obtain,

$$\begin{aligned}
& \mathbb{E} f(z^{r+1}) \leq \mathbb{E} f(z^r) + \mathbb{E} \langle \nabla f(z^r), z^{r+1} - z^r \rangle + \frac{\beta}{2} \mathbb{E} \|z^{r+1} - z^r\|^2 \\
& \leq \mathbb{E} f(z^r) + \underbrace{\alpha \eta \mathbb{E} \langle \nabla f(z^r), -G^r \rangle}_{D_1} + \underbrace{\frac{\beta}{2} \eta^2 \alpha^2 \mathbb{E} \|G^r\|^2}_{D_2}.
\end{aligned} \tag{55}$$

Next we will perform an upper bound analysis on D_1 ,

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

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

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

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

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

Finally, we can get,

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

and

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

Next we will perform an upper bound analysis on D_2 ,

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

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

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

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

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

let

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

$$\begin{aligned}
& \leq \mathbb{E}f(z^0) + \left(\frac{\alpha\eta\beta^2}{2M} + \frac{3T\beta^2C_1}{KM} \left(1 - \frac{K}{M} \right) \right) \sum_{r=1}^R \sum_{i=1}^M \sum_{t=1}^T \mathbb{E} \|y_i^{r,t} - x\|^2 \\
& + \frac{RC_1T^2G^2}{K} \left(1 - \frac{K}{M} \right) + \left(\left(1 - \frac{K}{M} \right) \frac{C_1T^2}{K} - \frac{\alpha\eta T}{2} \right) \sum_{r=1}^R \|\nabla f(x^r)\|^2 \\
& + \left(\frac{C_1}{M^2} - \alpha\eta \left(\frac{1}{4T} - \frac{a}{2} \right) \right) \sum_{r=1}^R \|\mathbb{E}[\mathbf{G}^r]\|^2 + C_1 \frac{TR}{M} \sigma^2 + C_1 \frac{TR\sigma^2}{K} \left(1 - \frac{K}{M} \right) \\
& \leq \mathbb{E}f(z^0) + \left(\frac{\alpha\eta\beta^2}{2} + \frac{3T\beta^2C_1}{K} \left(1 - \frac{K}{M} \right) \right) \sum_{r=1}^R \left[6T^3\eta^2 \frac{1}{M} \sum_{i=1}^M \|\nabla f_i(x^r)\|^2 + 3T^2\eta^2\sigma^2 \right] \\
& + \frac{RC_1T^2G^2}{K} \left(1 - \frac{K}{M} \right) + \left(\frac{C_1T^2}{K} \left(1 - \frac{K}{M} \right) - \frac{\alpha\eta T}{2} \right) \sum_{r=1}^R \|\nabla f(x^r)\|^2 \\
& + \left(\frac{C_1}{M^2} - \alpha\eta \left(\frac{1}{4T} - \frac{a}{2} \right) \right) \sum_{r=1}^R \|\mathbb{E}[\mathbf{G}^r]\|^2 + C_1 \frac{TR}{M} \sigma^2 + C_1 \frac{TR\sigma^2}{K} \\
& \leq \mathbb{E}f(z^0) + \left(\frac{\alpha\eta\beta^2}{2} + \frac{3T\beta^2C_1}{K} \left(1 - \frac{K}{M} \right) \right) \sum_{r=1}^R \left[12T^3\eta^2G^2 + 12T^3\eta^2 \|\nabla f(x^r)\|^2 + 3T^2\eta^2\sigma^2 \right] \\
& + \frac{RC_1T^2G^2}{K} \left(1 - \frac{K}{M} \right) + \left(\frac{C_1T^2}{K} \left(1 - \frac{K}{M} \right) - \frac{\alpha\eta T}{2} \right) \sum_{r=1}^R \|\nabla f(x^r)\|^2 \\
& + \left(\frac{C_1}{M^2} - \alpha\eta \left(\frac{1}{4T} - \frac{a}{2} \right) \right) \sum_{r=1}^R \|\mathbb{E}[\mathbf{G}^r]\|^2 + C_1 \frac{TR}{M} \sigma^2 + C_1 \frac{TR\sigma^2}{K} \left(1 - \frac{K}{M} \right).
\end{aligned} \tag{65}$$

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

$$\begin{aligned}
& \left(\frac{\alpha\eta T}{2} - \frac{C_1T^2}{K} - 12T^3\eta^2 \left(\frac{\alpha\eta\beta^2}{2} + \frac{3T\beta^2C_1}{K} \left(1 - \frac{K}{M} \right) \right) \right) \sum_{r=1}^R \|\nabla f(x^r)\|^2 \\
& \leq \mathbb{E}f(z^0) + 12T^3\eta^2G^2 \left(\frac{\alpha\eta\beta^2}{2} + \frac{3T\beta^2C_1}{K} \left(1 - \frac{K}{M} \right) \right) R + \frac{RC_1T^2G^2}{K} \left(1 - \frac{K}{M} \right) \\
& + 3T^2\eta^2\sigma^2 \left(\frac{\alpha\eta\beta^2}{2} + \frac{3T\beta^2C_1}{K} \left(1 - \frac{K}{M} \right) \right) R + C_1 \frac{TR}{M} \sigma^2 + C_1 \frac{TR\sigma^2}{K} \left(1 - \frac{K}{M} \right).
\end{aligned} \tag{66}$$

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,

$$\begin{aligned}
& \tilde{\eta} \frac{1}{R} \sum_{r=1}^R \|\nabla f(x^r)\|^2 \leq \frac{\mathbb{E}f(z^0)}{R} + 12T^3\eta^2G^2 \left(\frac{\alpha\eta\beta^2}{2} + \frac{3T\beta^2C_1}{K} \left(1 - \frac{K}{M} \right) \right) + \frac{C_1T^2G^2}{K} \left(1 - \frac{K}{M} \right) \\
& + 3T^2\eta^2\sigma^2 \left(\frac{\alpha\eta\beta^2}{2} + \frac{3T\beta^2C_1}{K} \left(1 - \frac{K}{M} \right) \right) + C_1 \frac{TR}{M} \sigma^2 + C_1 \frac{TR\sigma^2}{K} \left(1 - \frac{K}{M} \right).
\end{aligned} \tag{67}$$

1625 Moving $\tilde{\eta}$ to the left, we can obtain

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

1659 With Lemma 8, $\tilde{\eta} \leq \frac{1}{16\beta}$, we can get

$$\begin{aligned}
 & \frac{1}{R} \sum_{r=1}^R \|\nabla f(x^r)\|^2 \leq \frac{16\beta d_0}{TK\alpha R} + \frac{2\sqrt{d_0}}{\sqrt{RTM}} \left(\frac{8\beta}{K} \left(1 - \frac{K}{M} \right) G^2 + \frac{8\beta\sigma^2}{TK} \left(1 - \frac{K}{M} \right) + \frac{8\beta}{TM} \sigma^2 \right)^{\frac{1}{2}} \\
 & + 2 \left(\frac{d_0}{R} \right)^{\frac{2}{3}} \left[\frac{384\beta^2}{\alpha^2} G^2 + \frac{92\beta^2}{T} \frac{\sigma^2}{\alpha^2} + \frac{(16\gamma^2\beta^2)}{TM} \sigma^2 + \frac{(16\gamma^2\beta^2)\sigma^2}{TK} \left(1 - \frac{K}{M} \right) + \frac{16\gamma^2\beta^2}{K} \left(1 - \frac{K}{M} \right) G^2 \right]^{\frac{1}{3}} \\
 & + 2 \left(\frac{d_0}{R} \right)^{\frac{3}{4}} \left[\frac{4608}{\alpha^2} \frac{\beta^3}{K} \left(1 - \frac{K}{M} \right) G^2 + \frac{1152\beta^3}{KT\alpha^2} \left(1 - \frac{K}{M} \right) \sigma^2 \right]^{\frac{1}{4}} \\
 & + 2 \left(\frac{d_0}{R} \right)^{\frac{4}{5}} \left[\frac{9216}{\alpha^2} \frac{\gamma^2\beta^4}{K} \left(1 - \frac{K}{M} \right) G^2 + \frac{2304\beta^4}{K} \frac{\gamma^2}{\alpha^2 T} \left(1 - \frac{K}{M} \right) \sigma^2 \right]^{\frac{1}{5}}.
 \end{aligned} \tag{69}$$

1673 5 APPENDIX E: PROOF OF THEOREM 2.

1674 5.1 1. The rate of strongly convex and smooth convergence:

1675 We update the local control variates only for clients $i \in \mathcal{S}^r$

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

1676 Compute the new global parameters and global control variate using only updates from the clients $i \in \mathcal{K}_j^r$:

1677

1678

1679

1680

1681

1682

1683

1684

1685

1686

1687

1688

1689

1690

1691

1692

1693

1694

1695

1696

1697

1698

1699

1700

1701

1702

1703

1704

1705

1706

1707

1708

1709

1710

1711

1712

1713

1714

1715

1716

1717

1718

1719

1720

1721

1722

1723

1724

1725

1726

1727

1728

1729

1730

1731

1732

1733

1734

1735

1736

1737

1738

1739

1740

1741
1742
1743
1744

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

1745

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

1749

$$\mathbf{c}^r = \left[\mathbf{c}_{(1)}^{r\top}, \dots, \mathbf{c}_{(N)}^{r\top} \right]^\top. \quad (73)$$

1751

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

1753
1754
1755

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

1756
1757

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:

1758

$$C_r = \frac{1}{M} \sum_{i=1}^M \|\mathbb{E}[c_i^r] - \nabla f_i(z^r)\|^2. \quad (75)$$

1761

With Lemma 10, we have

1762
1763
1764
1765

$$\mathbb{E}\|z^{r+1} - x^*\|^2 \leq \mathbb{E}\|z^r - x^*\|^2 + 2\eta\alpha \underbrace{\mathbb{E}\langle -\mathbf{G}^r, z^r - x^* \rangle}_{E_1} + \eta^2\alpha^2 \underbrace{\mathbb{E}\|\mathbf{G}^r\|^2}_{E_2}. \quad (76)$$

1766 Before giving the convergence analysis of Theorem 1, we first present the following lemma.

1767

LEMMA 11. We can get the bound of E_2

1768
1769
1770
1771

$$\mathbb{E}\|\mathbf{G}^r\|^2 \leq \left(\frac{4T^2}{MT} \right) \sum_{i=1}^M \sum_{t=1}^T \mathbb{E}\|y_i^{r,t} - z^r\|^2 + \left(\frac{8T^2}{M} \right) \sum_{i=1}^M \|\mathbb{E}c_i^r - \nabla f_i(x^r)\|^2 + \left(\frac{4T^2}{M} \right) 2\beta(f(z^r) - f^*). \quad (77)$$

1772 Proof.

1773
1774

$$\begin{aligned} E_2 &= \mathbb{E}\|\mathbf{G}^r\|^2 \\ &= \sum_{j=1}^N \mathbb{E} \left\| \frac{1}{K} \sum_{k=1}^K \sum_{t=1}^T \nabla_{(j)} f_{k,j} \left(y_{k,j}^{r,t}; \zeta \right) + c_{(j)}^r - c_{k,j,(j)}^r + \nabla_{(j)} f_{k,j}(x^r) - \nabla_{(j)} f_{k,j}(x^r; \zeta) \right\|^2 \\ &= \sum_{j=1}^N \mathbb{E} \left\| \frac{1}{K} \sum_{k=1}^K \sum_{t=1}^T \nabla_{(j)} f_{k,j} \left(y_{k,j}^{r,t}; \zeta \right) + c_{(j)}^r - c_{k,j,(j)}^r + \nabla_{(j)} f_{k,j}(x^r) - \nabla_{(j)} f_{k,j}(x^r; \zeta) \right\|^2 \\ &\leq \frac{T}{M} \sum_{i=1}^M \sum_{t=1}^T \mathbb{E} \left\| \nabla f_i(y_i^{r,t}; \zeta) - \nabla f_i(z^r; \zeta) + c^r - c_i^r + \nabla f_i(z^r) \right\|^2 \\ &\leq \frac{T}{M} \sum_{k=1}^M \sum_{t=1}^T \mathbb{E} \left\| \nabla f_i(y_i^{r,t}; \zeta) - \nabla f_i(z^r; \zeta) + c^r - c_i^r + \nabla f_i(z^r) + \nabla f_i(x^*) - \nabla f_i(x^*) \right\|^2 \\ &\leq \left(\frac{4T}{M} \right) \sum_{i=1}^M \sum_{t=1}^T \mathbb{E}\|y_i^{r,t} - z^r\|^2 + \left(\frac{4T^2}{M} \right) \sum_{i=1}^M \|\mathbb{E}c_i^r - \nabla f_i(x^r)\|^2 \\ &\quad + \left(4T^2 \right) \mathbb{E}\|c^r\|^2 + \left(\frac{4T^2}{M} \right) \sum_{i=1}^M \|\nabla f_i(z^r) - \nabla f_i(x^*)\|^2 \\ &\leq \left(\frac{4T}{M} \right) \sum_{i=1}^M \sum_{t=1}^T \mathbb{E}\|y_i^{r,t} - z^r\|^2 + \left(\frac{4T^2}{M} \right) \sum_{i=1}^M \|\mathbb{E}c_i^r - \nabla f_i(z^r)\|^2 \\ &\quad + \left(4T^2 \right) \|\mathbb{E}c^r\|^2 + \left(\frac{4T^2}{M} \right) 2\beta(f(z^r) - f^*). \end{aligned} \quad (78)$$

1798

1857 LEMMA 12. We can get the bound of E_1

1915

$$1858 \\ 1859 \quad E_1 \leq -f(z^r) + f(x^\star) + \frac{\beta}{M} \sum_{i=1}^M \sum_{t=1}^T (\|y_i^{r,t} - z^r\|^2) - \frac{\mu}{4} T \|z^r - x^\star\|^2. \\ 1860 \\ 1861$$

1916

1917

1918

1919

1862 Proof. The term E_1 can be bounded by using perturbed strong-convexity (Lemma 1) with $h = f_i, \mathbf{x} = y_i^{r,t}, \mathbf{y} = x^\star$, and $\mathbf{z} = z^r$. Next we
1863 will calculate the upper bound for E_1 .

1920

1921

1922

$$1864 \\ 1865 \quad E_1 = -\mathbb{E} \langle \mathbf{G}^r, z^r - x^\star \rangle = -\sum_{j=1}^N \left\langle \frac{1}{M} \sum_{i=1}^M \sum_{t=1}^T \nabla_{(j)} f_i(y_i^{r,t}), z_{(j)}^r - x_{(j)}^\star \right\rangle \\ 1866 \\ 1867 \\ 1868 \\ 1869 \\ 1870 \\ 1871 \\ 1872 \\ 1873 \\ 1874 \\ 1875 \\ 1876 \\ 1877 \\ 1878 \\ 1879 \\ 1880 \\ 1881 \\ 1882 \\ 1883 \\ 1884 \\ 1885 \\ 1886 \\ 1887 \\ 1888 \\ 1889 \\ 1890 \\ 1891 \\ 1892 \\ 1893 \\ 1894 \\ 1895 \\ 1896 \\ 1897 \\ 1898 \\ 1899 \\ 1900 \\ 1901 \\ 1902 \\ 1903 \\ 1904 \\ 1905 \\ 1906 \\ 1907 \\ 1908 \\ 1909 \\ 1910 \\ 1911 \\ 1912 \\ 1913 \\ 1914 \\ 1915 \\ 1916 \\ 1917 \\ 1918 \\ 1919 \\ 1920 \\ 1921 \\ 1922 \\ 1923 \\ 1924 \\ 1925 \\ 1926 \\ 1927 \\ 1928 \\ 1929 \\ 1930 \\ 1931 \\ 1932 \\ 1933 \\ 1934 \\ 1935 \\ 1936 \\ 1937 \\ 1938 \\ 1939 \\ 1940 \\ 1941 \\ 1942 \\ 1943 \\ 1944 \\ 1945 \\ 1946 \\ 1947 \\ 1948 \\ 1949 \\ 1950 \\ 1951 \\ 1952 \\ 1953 \\ 1954 \\ 1955 \\ 1956 \\ 1957 \\ 1958 \\ 1959 \\ 1960 \\ 1961 \\ 1962 \\ 1963 \\ 1964 \\ 1965 \\ 1966 \\ 1967 \\ 1968 \\ 1969 \\ 1970 \\ 1971 \\ 1972$$

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

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

1935

1936

1937

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

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

1938

1939

1940

1941

Proof. First, we observe that if $T = 1$, $\mathcal{E}_r = 0$ since $y_i^{r,0} = z^r$ 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 $t \in [T]$

$$1887 \\ 1888 \quad \frac{1}{M} \sum_{i \in M} \mathbb{E} \|y_i^{r,t} - z^r\|^2 \\ 1889 \\ 1890 \\ 1891 \\ 1892 \\ 1893 \\ 1894 \\ 1895 \\ 1896 \\ 1897 \\ 1898 \\ 1899 \\ 1900 \\ 1901 \\ 1902 \\ 1903 \\ 1904 \\ 1905 \\ 1906 \\ 1907 \\ 1908 \\ 1909 \\ 1910 \\ 1911 \\ 1912 \\ 1913 \\ 1914 \\ 1915 \\ 1916 \\ 1917 \\ 1918 \\ 1919 \\ 1920 \\ 1921 \\ 1922 \\ 1923 \\ 1924 \\ 1925 \\ 1926 \\ 1927 \\ 1928 \\ 1929 \\ 1930 \\ 1931 \\ 1932 \\ 1933 \\ 1934 \\ 1935 \\ 1936 \\ 1937 \\ 1938 \\ 1939 \\ 1940 \\ 1941 \\ 1942 \\ 1943 \\ 1944 \\ 1945 \\ 1946 \\ 1947 \\ 1948 \\ 1949 \\ 1950 \\ 1951 \\ 1952 \\ 1953 \\ 1954 \\ 1955 \\ 1956 \\ 1957 \\ 1958 \\ 1959 \\ 1960 \\ 1961 \\ 1962 \\ 1963 \\ 1964 \\ 1965 \\ 1966 \\ 1967 \\ 1968 \\ 1969 \\ 1970 \\ 1971 \\ 1972$$

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

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

$$1903 \\ 1904 \\ 1905 \\ 1906 \\ 1907 \\ 1908 \\ 1909 \\ 1910 \\ 1911 \\ 1912 \\ 1913 \\ 1914 \\ 1915 \\ 1916 \\ 1917 \\ 1918 \\ 1919 \\ 1920 \\ 1921 \\ 1922 \\ 1923 \\ 1924 \\ 1925 \\ 1926 \\ 1927 \\ 1928 \\ 1929 \\ 1930 \\ 1931 \\ 1932 \\ 1933 \\ 1934 \\ 1935 \\ 1936 \\ 1937 \\ 1938 \\ 1939 \\ 1940 \\ 1941 \\ 1942 \\ 1943 \\ 1944 \\ 1945 \\ 1946 \\ 1947 \\ 1948 \\ 1949 \\ 1950 \\ 1951 \\ 1952 \\ 1953 \\ 1954 \\ 1955 \\ 1956 \\ 1957 \\ 1958 \\ 1959 \\ 1960 \\ 1961 \\ 1962 \\ 1963 \\ 1964 \\ 1965 \\ 1966 \\ 1967 \\ 1968 \\ 1969 \\ 1970 \\ 1971 \\ 1972$$

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

1973 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
 1974
 1975
 1976

$$\frac{1}{M} \sum_{i=1}^M \mathbb{E} \|y_i^{r,t-1} - z^r\|^2 \leq \frac{\left(1 + \frac{1}{T-1}\right)}{M} \sum_{i=1}^M \mathbb{E} \|y_i^{r,t-1} - z^r\|^2 + 6\eta^2 T \beta (f(z^r) - f(x^\star))$$

 1977
 1978
 1979

$$+ \frac{6T\eta^2}{M} \sum_{i=1}^M \|\mathbb{E} e_i^r - \nabla f_i(x^\star)\|^2.$$

 1980
 1981
 1982
 1983
 1984
 1985
 1986
 1987
 1988
 1989
 1990
 1991
 1992
 1993
 1994
 1995
 1996
 1997
 1998
 1999
 2000
 2001
 2002
 2003
 2004
 2005
 2006
 2007
 2008
 2009
 2010
 2011
 2012
 2013
 2014
 2015
 2016
 2017
 2018
 2019
 2020
 2021
 2022
 2023
 2024
 2025
 2026
 2027
 2028
 2029
 2030
 2031
 2032
 2033
 2034
 2035
 2036
 2037
 2038
 2039
 2040
 2041
 2042
 2043
 2044
 2045
 2046
 2047
 2048
 2049
 2050
 2051
 2052
 2053
 2054
 2055
 2056
 2057
 2058
 2059
 2060
 2061
 2062
 2063
 2064
 2065
 2066
 2067
 2068
 2069
 2070
 2071
 2072
 2073
 2074
 2075
 2076
 2077
 2078
 2079
 2080
 2081
 2082
 2083
 2084
 2085
 2086
 2087
 2088

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

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

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 $T \geq 4$, $(T-1) \left(\left(1 + \frac{1}{T-1}\right)^T - 1\right) < T \left(\exp\left(\frac{T}{T-1}\right) - 1\right) \leq T \left(\exp\left(\frac{4}{3}\right) - 1\right) < 3T$.

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

Again averaging over t ,

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

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

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

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

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

Plugging the above expression in the definition of C_r we get

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

2089 The final step applied Jensen's inequality twice. We can then further simplify using the relaxed triangle inequality as follows: 2147

$$\begin{aligned}
 2090 \quad & \mathbb{E}[C_r] \leq \left(1 - \frac{K}{M}\right) C_{r-1} + \frac{K}{M^2} \sum_{i=1}^M \mathbb{E} \|\nabla f_i(z^r) - \nabla f_i(x^*)\|^2 & 2148 \\
 2091 \quad & \leq \left(1 - \frac{K}{M}\right) C_{r-1} + \frac{K}{M^2} \sum_{i=1}^M \mathbb{E} \|\nabla f_i(z^{r-1}) - \nabla f_i(x^*)\|^2 & 2149 \\
 2092 \quad & \leq \left(1 - \frac{K}{M}\right) C_{r-1} + \frac{K}{M^2} \sum_{i=1}^M \mathbb{E} \|\nabla f_i(z^{r-1}) - \nabla f_k(x^*)\|^2 & 2150 \\
 2093 \quad & \leq \left(1 - \frac{K}{M}\right) C_{r-1} + \frac{K}{M} \left(4\beta (\mathbb{E}[f(z^{r-1})] - f(x^*))\right). & 2151 \\
 2094 \quad & & 2152 \\
 2095 \quad & & 2153 \\
 2096 \quad & & 2154 \\
 2097 \quad & & 2155 \\
 2098 \quad & & 2156 \\
 2099 \quad & & 2157 \\
 2100 \quad & & 2158 \\
 2101 \quad & & 2159 \\
 2102 \quad & & 2160 \\
 2103 \quad & & 2161 \\
 2104 \quad & & 2162 \\
 2105 \quad & \mathcal{E}_r := \frac{1}{TM} \sum_{t=1}^T \sum_{i=1}^M \mathbb{E} \|y_i^{r,t} - z^r\|^2. & 2163 \\
 2106 \quad & & 2164 \\
 2107 \quad & \text{LEMMA 15.} & 2165 \\
 2108 \quad & \|\mathbf{z}^{r+1} - \mathbf{x}^*\|^2 + 9\tilde{\eta}^2 \frac{M}{K} C_r \leq \left(1 - \frac{\tilde{\eta}\mu}{2}\right) \|\mathbf{z}^r - \mathbf{x}^*\|^2 + \left(1 - \frac{\mu\tilde{\eta}}{2}\right) 9\tilde{\eta}^2 \frac{M}{K} C_{r-1} & 2166 \\
 2109 \quad & & 2167 \\
 2110 \quad & \text{With } E_1 \text{ and } E_2, \text{ we can get} & 2168 \\
 2111 \quad & & 2169 \\
 2112 \quad & \mathbb{E} \|\mathbf{z}^{r+1} - \mathbf{x}^*\|^2 \leq \mathbb{E} \|\mathbf{z}^r - \mathbf{x}^*\|^2 + \underbrace{2\eta\alpha \mathbb{E} \langle -\mathbf{G}^r, \mathbf{z}^r - \mathbf{x}^* \rangle}_{E_1} + \underbrace{\eta^2\alpha^2 \mathbb{E} \|\mathbf{G}^r\|^2}_{E_2} & 2170 \\
 2113 \quad & & 2171 \\
 2114 \quad & \leq \left(1 - \frac{\eta\alpha\mu}{2}\right) T \|\mathbf{z}^r - \mathbf{x}^*\|^2 + 2\eta\alpha T (-f(\mathbf{z}^r) + f(\mathbf{x}^*)) & 2172 \\
 2115 \quad & + \left[\frac{2\eta\alpha\beta}{M} + \eta^2\alpha^2 \left(\frac{4T}{M}\right)\right] \sum_{i=1}^M \sum_{t=1}^T \left(\|y_i^{r,t} - z^r\|^2\right) + \eta^2\alpha^2 \left(\frac{8T^2}{M}\right) \sum_{i=1}^M \|\mathbb{E} c_i^r - \nabla f_i(z^r)\|^2 & 2173 \\
 2116 \quad & & 2174 \\
 2117 \quad & \leq \left(1 - \frac{\eta\alpha\mu T}{2}\right) \|\mathbf{z}^r - \mathbf{x}^*\|^2 + 2\eta\alpha T (-f(\mathbf{z}^r) + f(\mathbf{x}^*)) + [2\eta\alpha T\beta + 4T^2\eta^2\alpha^2] \frac{1}{MT} \sum_{i=1}^M \sum_{t=1}^T \left(\|y_i^{r,t} - z^r\|^2\right) & 2175 \\
 2118 \quad & & 2176 \\
 2119 \quad & + 8T^2\eta^2\alpha^2 \left(\frac{1}{M}\right) \sum_{i=1}^M \|\mathbb{E} c_i^r - \nabla f_i(z^r)\|^2, & 2177 \\
 2120 \quad & & 2178 \\
 2121 \quad & & 2179 \\
 2122 \quad & & 2180 \\
 2123 \quad & & 2181 \\
 2124 \quad & & 2182 \\
 2125 \quad & & 2183 \\
 2126 \quad & \text{with } \tilde{\eta} = \alpha\eta T, \text{ we have} & 2184 \\
 2127 \quad & & 2185 \\
 2128 \quad & \leq \left(1 - \frac{\tilde{\eta}\mu}{2}\right) \|\mathbf{z}^r - \mathbf{x}^*\|^2 + 2\tilde{\eta} (-f(\mathbf{z}^r) + f(\mathbf{x}^*)) + [2\tilde{\eta}\beta + 4\tilde{\eta}^2] \frac{1}{MT} \sum_{i=1}^M \sum_{t=1}^T \left(\|y_i^{r,t} - z^r\|^2\right) & 2186 \\
 2129 \quad & & 2187 \\
 2130 \quad & + 8\tilde{\eta}^2 \left(\frac{1}{M}\right) \sum_{i=1}^M \|\mathbb{E} c_i^r - \nabla f_i(z^r)\|^2 & 2188 \\
 2131 \quad & & 2189 \\
 2132 \quad & \leq \left(1 - \frac{\tilde{\eta}\mu}{2}\right) \|\mathbf{z}^r - \mathbf{x}^*\|^2 + 2\tilde{\eta} (-f(\mathbf{z}^r) + f(\mathbf{x}^*)) + [2\tilde{\eta}\beta + 4\tilde{\eta}^2] \mathcal{E}_r + 8\tilde{\eta}^2 \mathbb{E}[C_r]. & 2190 \\
 2133 \quad & & 2191 \\
 2134 \quad & & 2192 \\
 2135 \quad & & 2193 \\
 2136 \quad & \text{We can use Lemma 13 (scaled by } 9\tilde{\eta}^2 \frac{N}{S} \text{) to bound the control-lag} & 2194 \\
 2137 \quad & & 2195 \\
 2138 \quad & 3\beta\tilde{\eta}\mathcal{E}_r \leq \frac{54\tilde{\eta}^3\beta^2}{\alpha^2} (f(\mathbf{z}^r) - f(\mathbf{x}^*)) + \frac{54\tilde{\eta}^3\beta}{\alpha^2} C_{r-1}. & 2196 \\
 2139 \quad & & 2197 \\
 2140 \quad & \text{Now recall that Lemma 14 bounds the client-drift:} & 2198 \\
 2141 \quad & & 2199 \\
 2142 \quad & 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}{2K} - 1\right) \tilde{\eta}^2 C_{r-1} & 2200 \\
 2143 \quad & + 9\tilde{\eta}^2 \left(4\beta (\mathbb{E}[f(z^{r-1})] - f(x^*)) + 2\beta^2 \mathcal{E}\right). & 2201 \\
 2144 \quad & & 2202 \\
 2145 \quad & & 2203 \\
 2146 \quad & & 2204$$

$$\begin{aligned}
 2141 \quad & 9\tilde{\eta}^2 \frac{M}{K} C_r \leq \left(1 - \frac{\mu\tilde{\eta}}{2}\right) 9\tilde{\eta}^2 \frac{M}{K} C_{r-1} + 9 \left(\frac{\mu\tilde{\eta}M}{2K} - 1\right) \tilde{\eta}^2 C_{r-1} & 2199 \\
 2142 \quad & + 9\tilde{\eta}^2 \left(4\beta (\mathbb{E}[f(z^{r-1})] - f(x^*)) + 2\beta^2 \mathcal{E}\right). & 2200 \\
 2143 \quad & & 2201 \\
 2144 \quad & & 2202 \\
 2145 \quad & & 2203 \\
 2146 \quad & & 2204$$

2205 Adding all three inequalities together, we have

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

2213 Finally, with $\tilde{\eta} \leq \frac{1}{81\beta}$ and $\tilde{\eta} \leq \frac{K}{15\mu M}$ the lemma follows from noting that

$$-54\beta^2\tilde{\eta}^2 - 36\beta\tilde{\eta} + 2 \geq 0, \quad (99)$$

$$-\tilde{\eta}\beta + 4\tilde{\eta}^2\beta^2 \leq 0, \quad (100)$$

$$\frac{9\mu\tilde{\eta}N}{2S} - 9 + 8 + 54\tilde{\eta} \leq 0. \quad (101)$$

2219 The final rate for the case of strongly convex follows simply by unrolling the recursive bound and using Lemma 7,

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

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

2226 5.2 2: The convergence rate of general convex and smooth case:

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

2231 For general convex case, we have $\mu = 0$, then the following inequality holds:

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

2235 For the general convex setting, averaging over r in Lemma 8,

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

2240 5.3 3. The convergence rate of non-convex and smooth case:

2241 Recall that in round r , we update the control variate

$$c_i^r = \begin{cases} \nabla f_i(x^r) & \text{if } i \in \mathcal{S}^r \\ c_i^{r-1} & \text{otherwise} \end{cases}. \quad (107)$$

2246 We introduce the following notation to keep track of the lag in the update of the control variate: define a sequence of parameters $\{\alpha_i^{r,t}\}$
2247 such that for any $i \in [M]$ and $t \in [T]$ we have $\alpha_i^{0,t} := x^0$ and for $r \geq 1$,

$$\alpha_i^{r,t} := \begin{cases} y_i^{r,t} & \text{if } i \in \mathcal{S}^r \\ \alpha_i^{r-1,t} & \text{otherwise} \end{cases}. \quad (108)$$

2252 By the update rule for control variates (19) and the definition of $\{\alpha_i^{r,t}\}$ above, the following property always holds:

$$c_{k,j}^r = \nabla f_{k,j}(x^r). \quad (109)$$

2256 We can then define the following Ξ_r to be the error in control variate for round r :

$$\Xi_r := \frac{1}{TM} \sum_{t=1}^T \sum_{i=1}^M \mathbb{E} \|\alpha_i^{r,t} - z^r\|^2. \quad (110)$$

2261 Also recall the closely related definition of client drift caused by local updates:

$$\mathcal{E}_r := \frac{1}{TM} \sum_{t=1}^T \sum_{i=1}^M \mathbb{E} \left[\|y_i^{r,t} - z^r\|^2 \right]. \quad (111)$$

From the smoothness of the function, we can obtain

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

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.

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

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

$$\mathcal{E}_r \leq 24T^2 \eta^2 \beta^2 \mathbb{E} \|z^r - \alpha^r\|^2 + 12T^2 \eta^2 \mathbb{E} \|\nabla f(z^r)\|^2 \quad (114)$$

Proof. First, we observe that if $T = 1$, $\mathcal{E}_r = 0$ since $\mathbf{y}_i^{r,0} = \mathbf{x}^r$ 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 $t \in [T]$

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

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

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

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 $a \in [\frac{1}{2}, 1]$ where $\tilde{\eta} := \alpha T \eta$:

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

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

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

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

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

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

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

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

Proof. Applying the upper bounds of F_1 and F_2 ,

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

Also recall that Lemmas 16 and 17 state that

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

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

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

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

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

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

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 $c_i^0 = \nabla f_i(x^0)$ then we have $\Xi_0 = 0$. We can get

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

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.

2669 6.2 Dataset processing

2670 We evaluate FL on world datasets of image classification tasks including CIFAR-10 dataset, CIFAR-100 dataset, Tiny ImageNet dataset, mnist
 2671 dataset in our study. Both CIFAR10 and CIFAR100 datasets contain 60000 sheets of $3 \times 32 \times 32$ images. For CIFAR10, there are 10 categories,
 2672 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
 2673 the test set is 10000. In the experiment, we set up 100 clients with 500 images per client.

2674 Tiny ImageNet Challenge is the default course project for Stanford CS231N. Tiny Imagenet has 200 classes. Each class has 500 training
 2675 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
 2676 to 256×256 and crop to 224×224 to preprocess each image

2680 6.3 Model

2681 To test the robustness of our algorithms, we use standard classifiers (including LeNet-5 [?], VGG-11, VGG-19 [?], and ResNet-18 [?]), Vision
 2682 Transformer (ViT-Base) [?], Logistic regression Model [?]. We divided the parameters of the model into 5 blocks or more blocks and provide
 2683 the detailed parameter block division of the model in the Appendix.

2688 6.4 Hyper-parameter setting

2689 We provide hyperparameter settings for different datasets. For all real-world datasets in the convolutional network, including CIFAR10 and
 2690 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
 2691 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
 2692 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 $\alpha = 0.9$.

2693 For the ViT model, experiments were conducted on Tiny ImageNet and CIFAR100 datasets, and a pre trained model was adopted, with a
 2694 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
 2695 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
 2696 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 $\alpha = 0.9$.

2697 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.
 2698 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
 2699 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 $\alpha = 0.9$.

2704 6.5 Results on Logistic Regression

2705 We use a logistic regression model to verify the consistency between FedBCGD+'s practice and theory results. We conducted the classification
 2706 tests on the EMNIST dataset by using strongly convex and non-convex loss function models. To test the performance of our algorithms, we
 2707 use classical logistic regression problems, whose function has the following form:

$$2713 f(x) = \frac{1}{N} \sum_{i=1}^N \log(1 + \exp(-b_i a_i^\top x)) + \frac{\lambda}{2} \|x\|^2, \quad (129)$$

2721 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
 2722 is the smoothness constant.

2723 From the results of logistic regression in Figure 8 (a), we observe that our FedBCGD and FedBCGD+ algorithms demonstrate faster
 2724 convergence speed. Particularly, under the strong convexity condition with high client data heterogeneity, our FedBCGD+ algorithm exhibits
 2725 even faster convergence compared to our FedBCGD, which aligns with our theoretical analysis.

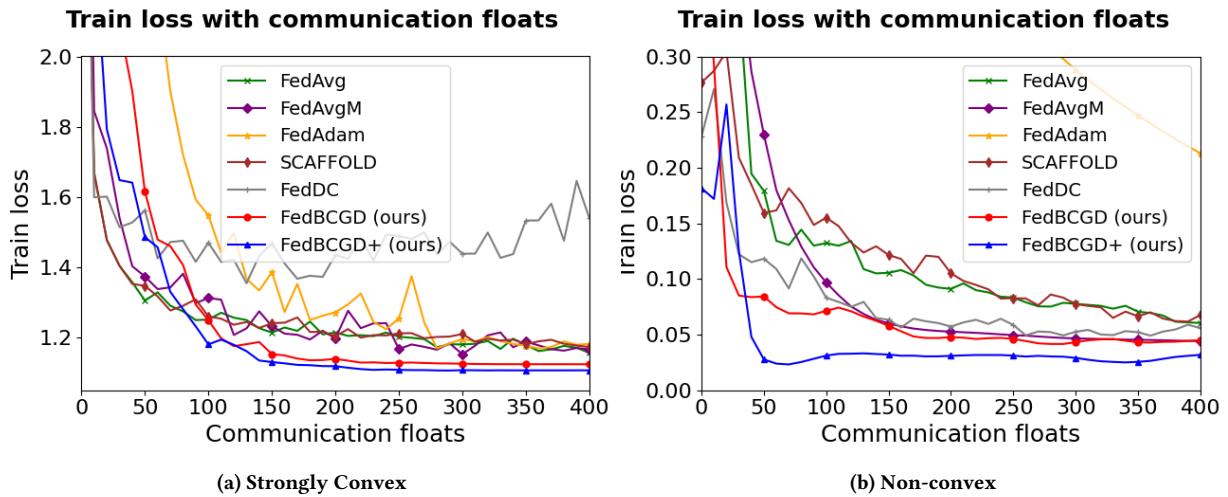


Figure 1: (a) Logistic regression with $E=1$ and $\rho=0.1$. (b) The problem with non-convex loss, where $E=1$ and $\rho=0.1$. The number of blocks is set to $N = 5$.

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

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

where $f_i(x) = 1/[1 + \exp(b_i a_i^\top x)]$. 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	
2901	Block 1	conv3-64	1728	2959
2902	Block 1	conv3-128	73728	2960
2903	Block 2	conv3-256	294912	2961
2904	Block 2	conv3-256	589824	2962
2905	Block 3	conv3-512	1179648	2963
2906	Block 3	conv3-512	2359296	2964
2907	Block 4	conv3-512	2359296	2965
2908	Block 4	conv3-512	2359296	2966
2909	Block 5	FC-2048	2359296	2967
2910	Block 5	FC-2048	2359296	2968
2911	Block share	FC-100	102400	2969
2912				2970
2913				2971
2914				2972
2915				2973
2916				2974
2917				2975
2918				2976
2919				2977
2920				2978
2921				2979
2922				2980
2923				2981
2924				2982
2925				2983
2926				2984
2927				2985
2928				2986
2929				2987
2930				2988
2931				2989
2932				2990
2933				2991
2934				2992
2935				2993
2936	Parameter Block	Network Layers	Number of parameters	2994
2937	Block 1	conv3-64	1728	2995
2938	Block 1	conv3-64	36864	2996
2939	Block 1	conv3-64	36864	2997
2940	Block 1	conv3-64	36864	2998
2941	Block 1	conv3-64	36864	2999
2942	Block 2	conv3-128	73728	3000
2943	Block 2	conv3-128	147456	3001
2944	Block 2	conv3-128	147456	3002
2945	Block 2	conv3-128	147456	3003
2946	Block 3	conv3-256	294912	3004
2947	Block 3	conv3-256	589824	3005
2948	Block 3	conv3-256	589824	3006
2949	Block 3	conv3-256	589824	3007
2950	Block 4	conv3-512	1179648	3008
2951	Block 4	conv3-512	2359296	3009
2952	Block 5	conv3-512	2359296	3010
2953	Block 5	conv3-512	2359296	3011
2954	Block share	FC-512 (share)	5120	3012
2955				3013
2956				3014
2957				3015
2958				3016

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

	Parameter Block	Network Layers	Number of parameters	
2937	Block 1	conv3-64	1728	2995
2938	Block 1	conv3-64	36864	2996
2939	Block 1	conv3-64	36864	2997
2940	Block 1	conv3-64	36864	2998
2941	Block 1	conv3-64	36864	2999
2942	Block 2	conv3-128	73728	3000
2943	Block 2	conv3-128	147456	3001
2944	Block 2	conv3-128	147456	3002
2945	Block 2	conv3-128	147456	3003
2946	Block 3	conv3-256	294912	3004
2947	Block 3	conv3-256	589824	3005
2948	Block 3	conv3-256	589824	3006
2949	Block 3	conv3-256	589824	3007
2950	Block 4	conv3-512	1179648	3008
2951	Block 4	conv3-512	2359296	3009
2952	Block 5	conv3-512	2359296	3010
2953	Block 5	conv3-512	2359296	3011
2954	Block share	FC-512 (share)	5120	3012
2955				3013
2956				3014
2957				3015
2958				3016

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

	Parameter Block	Network Layers	Number of parameters	
3017	Block 1	conv3-64	1728	3075
3018	Block 1	conv3-64	36864	3076
3019	Block 1	conv3-128	73728	3077
3020	Block 1	conv3-128	147456	3078
3021	Block 1	conv3-256	294912	3079
3022	Block 1	conv3-256	589824	3080
3023	Block 1	conv3-256	589824	3081
3024	Block 1	conv3-256	589824	3082
3025	Block 2	conv3-512	1179648	3083
3026	Block 2	conv3-512	2359296	3084
3027	Block 3	conv3-512	2359296	3085
3028	Block 3	conv3-512	2359296	3086
3029	Block 4	conv3-512	2359296	3087
3030	Block 4	conv3-512	2359296	3088
3031	Block 5	conv3-512	2359296	3089
3032	Block 5	conv3-512	2359296	3090
3033	Block 5	FC-2048	1048576	3091
3034	Block 5	FC-2048	131072	3092
3035	Block share	FC-100	25600	3093
3036				3094
3037	Table 4: The parameters block division of the VGG-19 network.			
3038				3095
3039				3096
3040				3097
3041				3098
3042				3099
3043				3100
3044				3101
3045				3102
3046				3103
3047				3104
3048				3105
3049				3106
3050				3107
3051				3108
3052				3109
3053				3110
3054				3111
3055	Parameter Block	Network Layers	Number of parameters	3112
3056	Block 1	ViT-Block 1	14299520	3113
3057	Block 2	ViT-Block 2	14299520	3114
3058	Block 3	ViT-Block 3	14299520	3115
3059	Block 4	ViT-Block 4	14299520	3116
3060	Block 5	ViT-Block 5	14299520	3117
3061	Block share	FC-100	153600	3118
3062	Table 5: The parameters block division of the VGG-19 network.			
3063				3119
3064				3120
3065				3121
3066				3122
3067				3123
3068				3124
3069				3125
3070				3126
3071	7 APPENDIX G: FEDBCGD AND FEDBCGD+ ALGORITHMS			
3072	The proposed FedBCGD+ and FedBCGD algorithms as shown in Algorithms 2 and 3, respectively.			
3073				3131
3074				3132

	Parameter Block	Network Layers	Number of parameters	
3055	Block 1	ViT-Block 1	14299520	3113
3056	Block 2	ViT-Block 2	14299520	3114
3057	Block 3	ViT-Block 3	14299520	3115
3058	Block 4	ViT-Block 4	14299520	3116
3059	Block 5	ViT-Block 5	14299520	3117
3060	Block share	FC-100	153600	3118
3061	Table 5: The parameters block division of the VGG-19 network.			
3062				3119
3063				3120
3064				3121
3065				3122
3066				3123
3067				3124
3068				3125
3069				3126
3070				3127
3071	7 APPENDIX G: FEDBCGD AND FEDBCGD+ ALGORITHMS			
3072	The proposed FedBCGD+ and FedBCGD algorithms as shown in Algorithms 2 and 3, respectively.			
3073				3131
3074				3132

Algorithm 1 FedBCGD+

```

3134 1: Initialize  $\mathbf{x}_i^{0,0} = \mathbf{x}^{init}, \forall i \in [M]$ . 3192
3135 2: Divide the model parameters  $\mathbf{x}$  into  $N$  blocks. 3193
3136 3: for  $r = 0, \dots, R$  do 3194
3137 4:   Client: 3195
3138 5:     Sample clients  $\mathcal{S} \subseteq \{1, \dots, M\}, |\mathcal{S}| = NK$ ; 3196
3139 6:     Divide the sampled clients into  $N$  blocks; 3197
3140 7:     Communicate  $(\mathbf{x}, \mathbf{c})$  to all clients  $i \in \mathcal{S}$ ; 3198
3141 8:     for  $j = 1, \dots, N$  client blocks in parallel do 3199
3142 9:       for  $k = 1, \dots, K$  clients in parallel do 3200
3143 10:      Compute full batch gradient  $\nabla f_{k,j}(\mathbf{x}^r)$ ; 3201
3144 11:      for  $t = 1, \dots, T$  local update do 3202
3145 12:        Compute mini-batch gradient  $\nabla f_{k,j}(\mathbf{x}_{k,j}^{r,t}; \zeta)$  and  $\nabla f_{k,j}(\mathbf{x}^r; \zeta)$ ; 3203
3146 13:         $\mathbf{x}_{k,j}^{r,t+1} = \mathbf{x}_{k,j}^{r,t} - \eta \nabla f_{k,j}(\mathbf{x}_{k,j}^{r,t}; \zeta) + \eta \mathbf{c} - \eta \mathbf{c}_{k,j} + \eta \nabla f_{k,j}(\mathbf{x}^r) - \eta \nabla f_{k,j}(\mathbf{x}^r; \zeta)$ ; 3204
3147 14:      end for 3205
3148 15:       $\mathbf{c}_{k,j}^+ \leftarrow \nabla f_{k,j}(\mathbf{x}^r)$ ; 3206
3149 16:      Send  $\mathbf{x}_{k,j,(j)}^{r,T}, \mathbf{x}_{k,j,s}^{r,T}$  and  $\Delta \mathbf{c}_{(j)} = \mathbf{c}_{k,j,(j)}^+ - \mathbf{c}_{k,j,(j)}$ ,  $\Delta \mathbf{c}_s = \mathbf{c}_{k,j,s}^+ - \mathbf{c}_{k,j,s}$  to server; 3207
3150 17:       $\mathbf{c}_i \leftarrow \mathbf{c}_i^+$ ; 3208
3151 18:    end for 3209
3152 19:  end for 3210
3153 20: Server: 3211
3154 21: for  $j = 1, \dots, N$  Blocks in parallel do 3212
3155 22:   Block  $j$  computes, 3213
3156 23:    $\mathbf{x}_{(j)}^r = \frac{1}{K} \sum_{k=1}^K \mathbf{x}_{k,j,(j)}^{r,T}$ ; 3214
3157 24:    $v_{(j)}^r = \lambda v_{(j)}^{r-1} + \mathbf{x}_{(j)}^r - \mathbf{x}_{(j)}^{r-1}$ ; 3215
3158 25:    $\mathbf{x}_{(j)}^r = \mathbf{x}_{(j)}^r + v_{(j)}^r$ ; 3216
3159 26:    $\mathbf{c}_{(j)} = \mathbf{c}_{(j)} + \frac{1}{M} \sum_{k=1}^K \Delta \mathbf{c}_{k,j,(j)}$ ; 3217
3160 27: end for 3218
3161 28:  $\mathbf{x}_s^r = \frac{1}{NK} \sum_{j=1}^N \sum_{k=1}^K \mathbf{x}_{k,j,s}^{r,T}$ ; 3219
3162 29:  $v_s^r = \lambda v_s^{r-1} + \mathbf{x}_s^{r-1} - \mathbf{x}_s^{r-1}$ ; 3220
3163 30:  $\mathbf{x}_s^r = \mathbf{x}_s^r + v_s^r$ ; 3221
3164 31:  $\mathbf{c}_s = \mathbf{c}_s + \frac{1}{MN} \sum_{j=1}^N \sum_{k=1}^K \Delta \mathbf{c}_{k,j,s}$ ; 3222
3165 32:  $\mathbf{x}^r = [\mathbf{x}_{(1)}^{r^\top}, \dots, \mathbf{x}_{(N)}^{r^\top}, \mathbf{x}_s^{r^\top}]^\top$ ; 3223
3166 33:  $\mathbf{v}^r = [v_{(1)}^{r^\top}, \dots, v_{(N)}^{r^\top}, v_s^{r^\top}]^\top$ ; 3224
3167 34:  $\mathbf{c} = [\mathbf{c}_{(1)}^\top, \dots, \mathbf{c}_{(N)}^\top, \mathbf{c}_s^\top]^\top$ ; 3225
3168 35: end for 3226
3169
3170
3171
3172
3173
3174
3175
3176
3177
3178
3179
3180
3181
3182
3183
3184
3185
3186
3187
3188
3189
3190

```