

451 A Supplementary Materials for Section 3 - Methodology

452 This supplementary section contains missing proofs from Section 3

453 **Theorem 1.** When a piecewise function L_t is defined for every value of $K_0 \in [K]$ on l , such that
 454 $0.0 \leq l < 1.0$, we claim, under Assumption 2, that the following is a subgradient of $f(x_t)$ at
 455 $K_t = K_0$:

$$\frac{\partial L_t}{\partial l} = \nabla f(x_t) \cdot \left(-\eta_{L,t} \sum_{i=1}^m g_i(x_{t-1}^{i,K_t-1}) \frac{\nu_i}{\nu} \right) \quad (6)$$

456 where l represents the marginal fraction of local steps beyond K_0 . We leave the proof (with an
 457 illustration in Figure 2) in the Appendix section beginning in eq (20).

458 *Proof.* We clarify this quantity by re-writing $L_{K_0}(l) = L_t$ and its definition (with some abuse of
 459 notation):

$$L_{K_0}(l) = f(x_t) \Big|_{K=K_0+l} = f\left(x_{t-1} - \eta_{L,t} \sum_{i=1}^m \frac{\nu_i}{\nu} \sum_{k=0}^{K_0-2} (g_i(x_{t-1}^{i,k}) + l g_i(x_{t-1}^{i,K_0-1}))\right) \quad (20)$$

460 for $K_0 \geq 2$, and $L_{K_0}(l) = f(x_{t-1} - \eta_{L,t} \sum_{i=1}^m l g_i(x_{t-1}^{i,K_0-1}) \frac{\nu_i}{\nu})$ for $K_0 = 1$, where l represents the
 461 marginal fraction of local steps K_0 . Then, by convexity from Assumption 2, and by recalling that l
 462 represents the marginal fraction of local steps, we claim that:

$$\frac{\partial L_{K_0}(l)}{\partial l} \leq \left[\frac{f(x_t) \Big|_{K=K_0+l} - f(x_t) \Big|_{K=K_0}}{l} \right] \quad (21)$$

463 From this result, we conclude that $\frac{\partial L_{K_0}(l)}{\partial l}$ is a subgradient of $f(x_t)$ at $K = K_0$. With abuse of
 464 notation by setting $K_0 = K_t$ and therefore $L_t = L_{K_0}(l)$, we further derive $\frac{\partial L_t}{\partial l}$ by breaking it down
 465 similarly as eq (4), leads us to:

$$\frac{\partial L_t}{\partial l} = \nabla f(x_t) \cdot \frac{\partial \left(-\eta_{L,t} \sum_{i=1}^m \frac{\nu_i}{\nu} \sum_{k=0}^{K_t-2} l g_i(x_{t-1}^{i,K_t-1}) \right)}{\partial l} \quad (22)$$

$$= \nabla f(x_t) \cdot \left(-\frac{\eta_{L,t}}{m} \sum_{i=1}^m g_i(x_{t-1}^{i,K_t-1}) \right) \quad (23)$$

466 The above result concludes the proof for Theorem 1

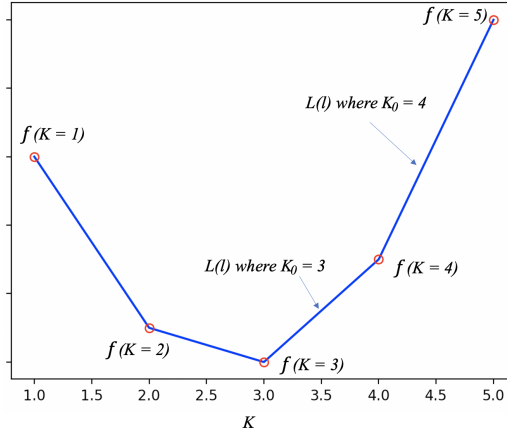


Figure 2: Illustration of piecewise function $L_{K_0}(l)$ and $f(K)$, where $f(K) = f(x, c)$ from eq 3, and where we ignore η_L and x_t for this discussion. Notice K is only defined on $\{K \mid K \geq 1\}$. Hence $f(K)$ are illustrated as red dots. However, since $0.0 \leq l < 1.0$, $L_{K_0}(l)$ extends from $f(K_0)$ to $f(K_0 + 1)$.

Theorem 2. When a piecewise function J_t is defined for every value of $K_0 \in [K]$ on l , such that $0.0 \leq l < 1.0$, we claim, under Assumption 2 that the following is a subgradient of $\sum_{i=1}^m f_i(x_t^{i,K_t})$ at $K_t = K_0$:

$$\frac{\partial J_t}{\partial l} = -\eta_{L,t} \sum_{i=1}^m \frac{\nu_i}{\nu} \mathbb{E}[g_i(x_t^{i,K_0-1})] \cdot g_i(x_t^{i,K_t}) \approx -\eta_{L,t} \sum_{i=1}^m \frac{\nu_i}{\nu} \sum_{k=0}^{K_t-1} g_i(x_t^{i,k}) \cdot g_i(x_t^{i,K_t}) \quad (10)$$

where l represents the marginal fraction of local steps beyond K_0 . We leave the proof in the Appendix section beginning in eq(24).

Proof. We clarify this quantity by re-writing $J_{K_0}(l) = J_t$ and its definition (with some abuse of notation):

$$J_{K_0}(l) = \sum_{i=1}^m \frac{\nu_i}{\nu} f_i(x_t^{i,K_0-1}) \Big|_{K=K_0+l} = \sum_{i=1}^m \frac{\nu_i}{\nu} f_i(x_t^{i,K_0-1} - l(\eta_{L,t} g_i(x_t^{i,K_0}))) \quad (24)$$

for $K_0 \geq 1$, where l represents the marginal fraction of local steps K_0 . Then, by convexity from Assumption 2 and by recalling that l represents the marginal fraction of local steps, we claim that:

$$\frac{\partial J_{K_0}(l)}{\partial l} \leq \left[\frac{\sum_{i=1}^m \frac{\nu_i}{\nu} (f_i(x_t^{i,K_0-1})|_{K=K_0+l} - f_i(x_t^{i,K_0-1})|_{K=K_0})}{l} \right] \quad (25)$$

From this result, we conclude that $\frac{\partial J_{K_0}(l)}{\partial l}$ is a subgradient of $f(x_t)$ at $K = K_0$. With abuse of notation by setting $K_0 = K_t$ and therefore $L_t = L_{K_0}(l)$, we further derive $\frac{\partial J_t}{\partial l}$ by breaking it down similarly as eq(4), leads us to:

$$\frac{\partial J_t}{\partial l} = \sum_{i=1}^m \frac{\nu_i}{\nu} \nabla f_i(x_t^{i,K_t-1}) \cdot \frac{\partial(-l\eta_{L,t}g_i(x_t^{i,K_t}))}{\partial l} \quad (26)$$

$$\stackrel{(J1)}{=} -\eta_{L,t} \sum_{i=1}^m \frac{\nu_i}{\nu} \mathbb{E}[g_i(x_t^{i,K_0-1})] \cdot g_i(x_t^{i,K_t}) \quad (27)$$

$$\stackrel{(J2)}{\approx} -\eta_{L,t} \sum_{i=1}^m \frac{\nu_i}{\nu} \sum_{k=0}^{K_t-1} g_i(x_t^{i,k}) \cdot g_i(x_t^{i,K_t}) \quad (28)$$

where J1 follows from Assumption 1, and J2 crudely assumes $\nabla f_i(x_t^{i,K_t-1}) \approx \sum_{k=0}^{K_t-1} g_i(x_t^{i,k})$ from the additional averaging. The above result concludes the proof for Theorem 2. \square

B Supplementary Materials for Section 4 - Theoretical Convergence

This supplementary section contains all the missing proofs from Section 4.

Theorem 3. Under Assumptions 1-5 and with full client participation, when FATHOM as shown in Algorithm 1 is used to find a solution x_* to the unconstrained problem defined in eq(1), the sequence of outputs $\{x_t\}$ satisfies the following upper-bound, where, with slight abuse of notation, $\mathcal{E} = \min_{t \in [T]} \mathbb{E}_t \|\nabla f(x_t)\|_2^2$:

$$\mathcal{E}_{fathom} = \mathcal{O}\left(\sqrt{\frac{\sigma_L^2 + G^2}{m\overline{K}T}} + \sqrt[3]{\frac{\sigma_L^2}{\overline{K}T^2}} + \sqrt[3]{\frac{G^2}{T^2}}\right) \quad (16)$$

with the following conditions: $\bar{\eta}_L = \min\left(\sqrt{\frac{2\beta_0 m D}{\beta_1 \overline{K} L T (\sigma_L^2 + G^2)}}, \sqrt[3]{\frac{\beta_0 D}{2.5\beta_2 \overline{K}^2 L^2 \sigma_L^2 T}}, \sqrt[3]{\frac{\beta_0 D}{2.5\beta_3 \overline{K}^3 L^2 G^2 T}}\right)$

and $\eta_{L,t} \leq 1/L$ for all t , where

$$\bar{\eta}_L \triangleq \frac{1}{T} \sum_{t=1}^T \eta_{L,t} \quad \text{and} \quad \overline{K} \triangleq \frac{1}{T} \sum_{t=1}^T K_t \quad (17)$$

490 and where

$$\beta_0 = \frac{\sum_t \eta_{L,t} K_t}{T[\frac{1}{T} \sum_t \eta_{L,t}][\frac{1}{T} \sum_t K_t]}, \quad \beta_1 = \frac{\sum_t \eta_{L,t} K_t [\frac{1}{T} \sum_t \eta_{L,t}]}{\sum_t \eta_{L,t}^2 K_t} \quad (18)$$

$$\beta_2 = \frac{\sum_t \eta_{L,t} K_t [\frac{1}{T} \sum_t \eta_{L,t}]^2 [\frac{1}{T} \sum_t K_t]}{\sum_t \eta_{L,t}^3 K_t^2}, \quad \beta_3 = \frac{\sum_t \eta_{L,t} K_t [\frac{1}{T} \sum_t \eta_{L,t}]^2 [\frac{1}{T} \sum_t K_t]^2}{\sum_t \eta_{L,t}^3 K_t^3} \quad (19)$$

491 We leave the proof in the Appendix beginning in eq(29).

492 *Proof.* We begin by first defining the following:

493 By re-writing eq(38) from Lemma 1 with adaptive $\eta_{L,t}$ and K_t , we end up with:

$$\sum_{t=0}^{T-1} \frac{\eta_{L,t} K_t}{2} \mathbb{E}_t \|\nabla f(x_t)\|^2 \leq f(x_0) - f(x_T) + \sum_{t=0}^{T-1} \eta_{L,t} K_t \left[\frac{\eta_{L,t} L}{2m} (\sigma_L^2 + G^2) + \frac{5\eta_{L,t}^2 K_t L^2}{2} (\sigma_L^2 + K_t G^2) \right] \quad (29)$$

494 After re-arranging:

$$\min_{t \in [T]} \mathbb{E}_t \|\nabla f(x_t)\|^2 \leq \underbrace{\frac{2D}{\sum_t \eta_{L,t} K_t}}_{\text{progress}} + \underbrace{\frac{L \sum_t \eta_{L,t}^2 K_t}{m \sum_t \eta_{L,t} K_t} (\sigma_L^2 + G^2)}_{\text{deviation 1}} + \underbrace{\frac{5L^2 \sum_t \eta_{L,t}^3 K_t^2}{\sum_t \eta_{L,t} K_t} \sigma_L^2}_{\text{deviation 2}} + \underbrace{\frac{5L^2 \sum_t \eta_{L,t}^3 K_t^3}{\sum_t \eta_{L,t} K_t} G^2}_{\text{deviation 3}} \quad (30)$$

495 which is followed by:

$$\mathcal{E}_{fathom} \leq \underbrace{\frac{2\beta_0 D}{\bar{\eta}_L \bar{K} T}}_{\text{progress}} + \underbrace{\frac{\beta_1 \bar{\eta}_L L}{m} (\sigma_L^2 + G^2)}_{\text{deviation 1}} + \underbrace{5\beta_2 \bar{\eta}_L^2 \bar{K} L^2 \sigma_L^2}_{\text{deviation 2}} + \underbrace{5\beta_3 \bar{\eta}_L^2 \bar{K}^2 L^2 G^2}_{\text{deviation 3}} \quad (31)$$

496 We have one progress term, and three deviation terms, similar to the labeling scheme in the convex
 497 result from Wang et al. [2021]. Typically, one of these terms dominates during the course of the
 498 optimization process, where it is desirable to never let one of the deviation terms to become dominant.
 499 When we set each of the deviation terms to be equal to the progress term, we recover the bound
 500 shown in eq(16) when the conditions are met. This concludes the proof for Theorem 3. \square

501 **Lemma 1.** Under Assumptions 1-5 and with full client participation, when FedAvg with constant
 502 hyperparameters is used to find a solution x_* to the unconstrained problem defined in eq(7), the
 503 sequence of outputs $\{x_t\}$ satisfies the following upper-bound, where, with slight abuse of notation,
 504 $\mathcal{E} = \min_{t \in [T]} \mathbb{E}_t \|\nabla f(x_t)\|_2^2$:

$$\min_{t \in [T]} \mathbb{E}_t \|\nabla f(x_t)\|^2 \leq \underbrace{\frac{2D}{\eta_L K T}}_{\text{progress}} + \underbrace{\frac{\eta_L L}{m} (\sigma_L^2 + G^2)}_{\text{deviation 1}} + \underbrace{5\eta_L^2 K L^2 \sigma_L^2}_{\text{deviation 2}} + \underbrace{5\eta_L^2 K^2 L^2 G^2}_{\text{deviation 3}} \quad (32)$$

505 where $D = f(x_0) - f(x_T) = f(x_0) - f(x_*)$ with x_* being the fixed point solution discussed in
 506 Section 4.1. Eq(32) has one progress term and three deviation terms, where $\eta_L \leq \frac{1}{L}$ needs to hold
 507 for client local gradient descent to guarantee local progress.

508 *Proof.* We start proving convergence of the non-convex problem by bounding the progress made
 509 in the loss function within a single round, loosely following the beginning steps from the Proof of
 510 Theorem 1 in Yang et al [2021]:

$$\mathbb{E}_t[f(x_{t+1})] \leq \mathbb{E}_t[f(x_t) + \langle \nabla f(x_t), \mathbb{E}_t(x_{t+1} - x_t) \rangle] + \frac{L}{2} \mathbb{E}_t \|x_{t+1} - x_t\|^2 \quad (33)$$

$$= \mathbb{E}_t[f(x_t) + \langle \nabla f(x_t), \mathbb{E}_t[\bar{\Delta}_t + \eta_L K \nabla f(x_t) - \eta_L K \nabla f(x_t)] \rangle] + \frac{L}{2} \mathbb{E}_t \|\bar{\Delta}_t\|^2 \quad (34)$$

$$= \mathbb{E}_t[f(x_t) - \eta_L K \|\nabla f(x_t)\|^2] + \underbrace{\langle \nabla f(x_t), \mathbb{E}_t[\bar{\Delta}_t + \eta_L K \nabla f(x_t)] \rangle}_{A_1} + \frac{L}{2} \underbrace{\mathbb{E}_t \|\bar{\Delta}_t\|^2}_{A_2} \quad (35)$$

By using results from Lemma 2 and Lemma 3, we have what follows:

$$\mathbb{E}_t[f(x_{t+1})] \leq \mathbb{E}_t[f(x_t)] - \frac{\eta_L K}{2} \|\nabla f(x_t)\|^2 + \frac{5\eta_L^3 K^2 L^2}{2} (\sigma_L^2 + KG^2) + \frac{L\eta_L^2 K}{2m} (\sigma_L^2 + G^2) \quad (36)$$

Therefore, in order to guarantee progress in each round, the following condition is required to hold true:

$$\|\nabla f(x_t)\|^2 \geq 5\eta_L^2 KL^2 (\sigma_L^2 + KG^2) + \frac{L\eta_L}{m} (\sigma_L^2 + G^2) \quad (37)$$

Continuing from eq(36), and summing telescopically, we end up with:

$$\sum_{t=0}^{T-1} \frac{\eta_L K}{2} \mathbb{E}_t \|\nabla f(x_t)\|^2 \leq f(x_0) - f(x_T) + T\eta_L K \left[\frac{\eta_L L}{2m} (\sigma_L^2 + G^2) + \frac{5\eta_L^2 KL^2}{2} (\sigma_L^2 + KG^2) \right] \quad (38)$$

where $D = f(x_0) - f(x_T) = f(x_0) - f(x_*)$ with x_* being the fixed point solution discussed in Section 4.1. This concludes the proof of Lemma 1. \square

Lemma 2. Under Assumptions 1 and 5 and with full client participation, we claim the following is true:

$$A_1 \leq \frac{\eta_L K}{2} \|\nabla f(x)\|^2 + \frac{5K^2 \eta_L^3 L^2}{2} (\sigma_L^2 + KG^2) \quad (39)$$

Proof. We start by following most of the initial steps from the Proof of Theorem 1 in Yang et al [2021]:

$$A_1 = \langle \nabla f(x_t), \mathbb{E}_t[\bar{\Delta}_t + \eta_L K \nabla f(x_t)] \rangle \quad (40)$$

$$= \left\langle \nabla f(x_t), \mathbb{E}_t \left[-\frac{1}{m} \sum_{i=1}^m \sum_{k=0}^{K-1} \eta_L g_i(x_t^{i,k}) + \eta_L K \nabla f(x_t) \right] \right\rangle \quad (41)$$

$$= \left\langle \nabla f(x_t), \mathbb{E}_t \left[-\frac{1}{m} \sum_{i=1}^m \sum_{k=0}^{K-1} \eta_L \nabla f_i(x_t^{i,k}) + \eta_L K \frac{1}{m} \sum_{i=1}^m \nabla f_i(x_t) \right] \right\rangle \quad (42)$$

$$= \left\langle \sqrt{\eta_L K} \nabla f(x_t), -\frac{\sqrt{\eta_L}}{m\sqrt{K}} \mathbb{E}_t \sum_{i=1}^m \sum_{k=0}^{K-1} (\nabla f_i(x_t^{i,k}) - \nabla f_i(x_t)) \right\rangle \quad (43)$$

$$\stackrel{(a1)}{=} \frac{\eta_L K}{2} \|\nabla f(x_t)\|^2 + \frac{\eta_L}{2Km^2} \mathbb{E}_t \left\| \sum_{i=1}^m \sum_{k=0}^{K-1} \nabla f_i(x_t^{i,k}) - \nabla f_i(x_t) \right\|^2 - \frac{\eta_L}{2Km^2} \mathbb{E}_t \left\| \sum_{i=1}^m \sum_{k=0}^{K-1} \nabla f_i(x_t^{i,k}) \right\|^2 \quad (44)$$

$$\stackrel{(a2)}{\leq} \frac{\eta_L K}{2} \|\nabla f(x_t)\|^2 + \frac{\eta_L}{2m} \sum_{i=1}^m \sum_{k=0}^{K-1} \mathbb{E}_t \|\nabla f_i(x_t^{i,k}) - \nabla f_i(x_t)\|^2 - \frac{\eta_L}{2Km^2} \sum_{i=1}^m \sum_{k=0}^{K-1} \mathbb{E}_t \|\nabla f_i(x_t^{i,k})\|^2 \quad (45)$$

$$\stackrel{(a3)}{\leq} \frac{\eta_L K}{2} \|\nabla f(x_t)\|^2 + \frac{\eta_L L^2}{2m} \sum_{i=1}^m \sum_{k=0}^{K-1} \mathbb{E}_t \|x_t^{i,k} - x_t\|^2 - \frac{\eta_L}{2m} G^2 \quad (46)$$

$$\stackrel{(a4)}{\leq} \frac{\eta_L K}{2} \|\nabla f(x_t)\|^2 + \frac{5K^2 \eta_L^3 L^2}{2} (\sigma_L^2 + KG^2) \quad (47)$$

where, from Yang et al [2021], (a1) follows from that $\langle x, y \rangle = \frac{1}{2} [\|x\|^2 + \|y\|^2 - \|x - y\|^2]$ for $x = \sqrt{\eta_L K} \nabla f(x_t)$ and $y = -\frac{\sqrt{\eta_L}}{m\sqrt{K}} \sum_{i=1}^m \sum_{k=0}^{K-1} (\nabla f_i(x_t^{i,k}) - \nabla f_i(x_t))$, (a2) is due to $\mathbb{E} \|x_1 + x_2 + \dots + x_n\|^2 \leq n\mathbb{E} [\|x_1\|^2 + \|x_2\|^2 + \dots + \|x_n\|^2]$ and $\mathbb{E} [\|x_1\|^2 + \|x_2\|^2 + \dots + \|x_n\|^2] \leq \mathbb{E} \|x_1 + x_2 + \dots + x_n\|^2$, (a3) is due to Assumption 3, which is where we start to diverge from Yang et al [2021]. Our result from Lemma 4 by using Assumption 5, combined with removal of the last term, justifies (a4) above, and thus concludes the proof for Lemma 2. The last term of eq(46) could have remained for a tighter final bound in the theorems, but would require to restrict K such that $\eta_L K \leq \frac{1}{L}$ which we try to avoid.

\square

529 **Lemma 3.** Under Assumptions 1 and 5 and with full client participation, we claim the following is true:

$$A_2 \leq \frac{\eta_L K^2}{m} [\sigma_L^2 + G^2] \quad (48)$$

530 *Proof.* We start with the following definition of $\bar{\Delta}_t = \frac{1}{m} \sum_{i=1}^m \Delta_t^i = (-\frac{\eta_L}{m} \sum_{i=1}^m \sum_{k=0}^{K-1} g_i(x_t^{i,k}))$:

$$\mathbb{E}_t \|\bar{\Delta}_t\|^2 = \mathbb{E}_t \left\| \frac{1}{m} \sum_{i=1}^m \Delta_t^i \right\|^2 \quad (49)$$

$$= \frac{1}{m^2} \mathbb{E}_t \left\| \sum_{i=1}^m \Delta_t^i \right\|^2 = \frac{\eta_L^2}{m^2} \mathbb{E}_t \left\| \sum_{i=1}^m \sum_{k=0}^{K-1} g_i(x_t^{i,k}) \right\|^2 \quad (50)$$

$$= \frac{\eta_L^2}{m^2} \mathbb{E}_t \left\| \sum_{i=1}^m \sum_{k=0}^{K-1} (g_i(x_t^{i,k}) - \nabla f(x_t^{i,k})) \right\|^2 + \frac{\eta_L^2}{m^2} \mathbb{E}_t \left\| \sum_{i=1}^m \sum_{k=0}^{K-1} \nabla f(x_t^{i,k}) \right\|^2 \quad (51)$$

$$\leq \frac{\eta_L K^2}{m} [\sigma_L^2 + G^2] \quad (52)$$

531 which completes the proof of Lemma 3. \square

532 **Lemma 4.** Under Assumptions 3 and 5 and with full client participation, we claim the following is true:

$$\frac{1}{m} \sum_{i=1}^m \mathbb{E}_t \|x_t^{i,k} - x_t\|^2 \leq 5K [K\eta_L^2 G^2 + \eta_L^2 \sigma_L^2] \quad (53)$$

533 *Proof.* We start by loosely following Lemma 3 from Reddi et al [2020]:

$$\mathbb{E}_t \|x_t^{i,k} - x_t\|^2 = \mathbb{E}_t \|x_t^{i,k-1} - x_t - \eta_L g_i(x_t^{i,k-1})\|^2 \quad (54)$$

$$= \mathbb{E}_t \|x_t^{i,k-1} - x_t - \eta_L (g_i(x_t^{i,k-1}) - \nabla f_i(x_t^{i,k-1}) + \nabla f_i(x_t^{i,k-1}))\|^2 \quad (55)$$

$$\leq (1 + \frac{1}{K-1}) \mathbb{E}_t \|x_t^{i,k-1} - x_t\|^2 + K\eta_L^2 \mathbb{E}_t \|\nabla f_i(x_t^{i,k-1})\|^2 + \eta_L^2 (g_i(x_t^{i,k-1}) - \nabla f_i(x_t^{i,k-1}))^2 \quad (56)$$

$$\leq (1 + \frac{1}{K-1}) \mathbb{E}_t \|x_t^{i,k-1} - x_t\|^2 + K\eta_L^2 G^2 + \eta_L^2 \sigma_L^2 \quad (57)$$

534 The last two inequalities follows Assumption 4 and Assumption 5 which yields a looser bound and
535 which diverges from Lemma 3 from Reddi et al [2020]. Unrolling the recursion over k and summing
536 over clients $i \in [m]$:

$$\frac{1}{m} \sum_{i=1}^m \mathbb{E}_t \|x_t^{i,k} - x_t\|^2 \leq \sum_{p=0}^{K-1} (1 + \frac{1}{K-1})^p [K\eta_L^2 G^2 + \eta_L^2 \sigma_L^2] \quad (58)$$

$$\leq K (1 + \frac{1}{K-1})^K [K\eta_L^2 G^2 + \eta_L^2 \sigma_L^2] \quad (59)$$

$$\leq 5K [K\eta_L^2 G^2 + \eta_L^2 \sigma_L^2] \quad (60)$$

537 where $(1 + \frac{1}{K-1})^K \leq 5$ for $K > 1$. This concludes the proof of Lemma 4. \square

538 C Supplementary Materials for Section 5 - Empirical Evaluation and 539 Numerical Results

540 This section summarizes the missing details from Section 5. As mentioned, the datasets, models
541 and tasks are exactly the same as the "EMNIST CR" task and the "SO NWP" task from Reddi et al
542 [2020], such that we can use their optimized FedAvg results as baseline. However, Reddi et al
543 [2020] implement their algorithms on the Tensorflow Federated framework (Ingberman et al. [2019]),
544 whereas for our work, we build our algorithms on the FedJAX framework (Ro et al. [2021]) which is
545 under the Apache License.

Table 2: EMNIST character recognition model architecture.

Layer	Output Shape	# of Trainable Parameters	Activation	Hyperparameters
Input	(28, 28, 1)	0		
Conv2d	(26, 26, 32)	320		kernel size = 3; strides = (1, 1)
Conv2d	(24, 24, 64)	18496	ReLU	kernel size = 3; strides = (1, 1)
MaxPool2d	(12, 12, 64)	0		pool size = (2, 2)
Dropout	(12, 12, 64)	0		p = 0.25
Flatten	9216	0		
Dense	128	1179776		
Dropout	128	0		p = 0.5
Dense	62	7998	softmax	

Table 3: Stack Overflow next word prediction model architecture.

Layer	Output Shape	# of Trainable Parameters
Input	20	0
Embedding	(20, 96)	960384
LSTM	(20, 670)	2055560
Dense	(20, 96)	64416
Dense	(20, 10004)	970388

C.1 Datasets, Models, and Tasks

We train a CNN to do character recognition (EMNIST CR) on the federated EMNIST-62 dataset (Cohen et al. [2017]). Next, we train a RNN to do next-word-prediction (SO NWP) on the federated Stack Overflow dataset (Authors [2019]).

Federated EMNIST-62 with CNN EMNIST consists of images of digits and upper and lower case English characters, with 62 total classes. The federated version of EMNIST (Caldas et al., 2018) partitions the digits by their author. The dataset has natural heterogeneity stemming from the writing style of each person. See Table 4 for more on the statistics of the federated EMNIST-62 dataset. On our select task of character recognition for this dataset (EMNIST CR), a Convolutional Neural Network (CNN) is used. The network has two convolutional layers (with 3×3 kernels), max pooling, and dropout, followed by a 128 unit dense layer. A full description of the model is in Table 2.

Federated Stack Overflow with RNN Stack Overflow is a language modeling dataset consisting of question and answers from the question and answer site, Stack Overflow. The questions and answers also have associated metadata, including tags. The dataset contains 342,477 unique users which we use as clients. See Table 4 for more on the statistics of the federated Stack Overflow dataset. We perform next-word prediction (Stack Overflow NWP, SO NWP for short) on this dataset. We restrict the task to the 10,000 most frequently used words, and each client to the first 128 sentences in their dataset. We also perform padding and truncation to ensure that sentences have 20 words. We then represent the sentence as a sequence of indices corresponding to the 10,000 frequently used words, as well as indices representing padding, out-of-vocabulary words, beginning of sentence, and end of sentence. We perform next-word-prediction on these sequences using a Recurrent Neural Network (RNN) that embeds each word in a sentence into a learned 96-dimensional space. It then feeds the embedded words into a single LSTM layer of hidden dimension 670, followed by a densely connected softmax output layer. A full description of the model is in Table 3. The metric used in the main body is the top-1 accuracy over the proper 10,000-word vocabulary; that is, it does not include padding, out-of-vocab, or beginning or end of sentence tokens.

Table 4: Data statistics.

Dataset	Train Clients	Train Examples	Test Clients	Test Examples
EMNIST-62	3,400	671,585	3,400	77,483
STACKOVERFLOW	342,477	135,818,730	204,088	16,586,035

C.2 Client Sampling

In all our experiments, we do not include updates from all clients in each communication round. Instead, client sampling is done, where clients are sampled uniformly at random from all training clients, without replacement within a given round, but with replacement across rounds. In our EMNIST CR experiments, 10 out of a total of 3,400 clients are sampled in each communication round, and in our SO NWP experiments, 50 out of a total of 342,477 clients are sampled in each round.

C.3 Additional Results

Below, we provide additional results from our experiments conducted in Section 5 and whose test accuracy performance results shown in Figure 1. The baseline values were selected for best performance from Reddi et al. [2020].

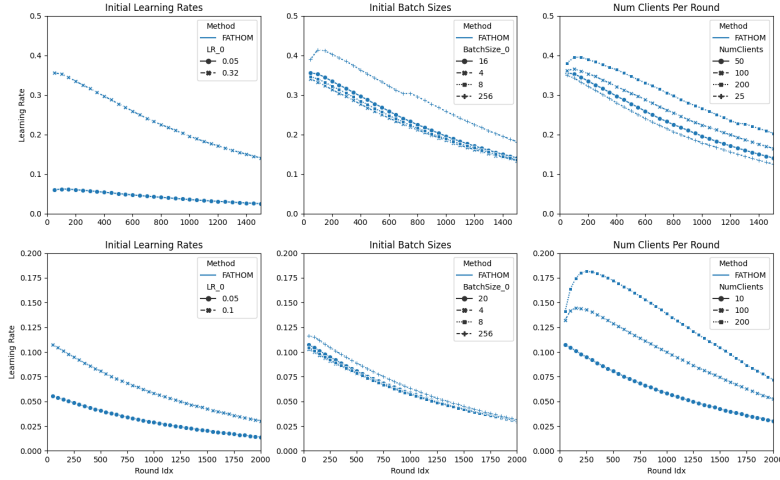


Figure 3: Adaptive client learning rate from the same experiments conducted in Section 5 and in Figure 1. Top row: FSO sims. Bottom row: FEMNIST sims. Baseline values for FEMNIST: LR_0=0.1, BatchSize_0=20, NumClients=10. Baseline values for FSO: LR_0=0.32, BatchSize_0=16, NumClients=50.

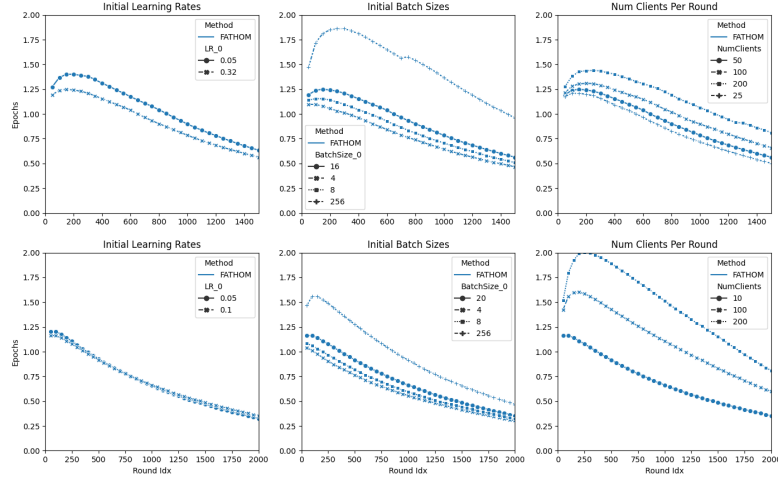


Figure 4: Adaptive number of epochs from the same experiments conducted in Section 5 and in Figures 1 and 3. Top row: FSO sims. Bottom row: FEMNIST sims. Baseline values for FEMNIST: LR_0=0.1, BatchSize_0=20, NumClients=10. Baseline values for FSO: LR_0=0.32, BatchSize_0=16, NumClients=50.

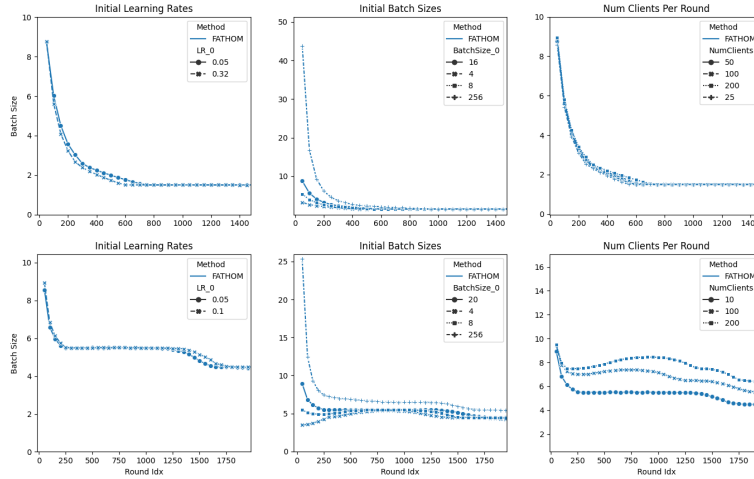


Figure 5: Adaptive batch size from the same experiments conducted in Section 5 and in Figures 1, 3 and 4. Top row: FSO sims. Bottom row: FEMNIST sims. Baseline values for FEMNIST: LR_0=0.1, BatchSize_0=20, NumClients=10. Baseline values for FSO: LR_0=0.32, BatchSize_0=16, NumClients=50.

583 References for the Appendix

- 584 Gregory Cohen, Saeed Afshar, Jonathan Tapson, and Andre Van Schaik. Emnist: Extending mnist to
585 handwritten letters. In *2017 international joint conference on neural networks (IJCNN)*, pages
586 2921–2926. IEEE, 2017.
- 587 Sebastian Caldas, Sai Meher Karthik Duddu, Peter Wu, Tian Li, Jakub Konečný, H. Brendan
588 McMahan, Virginia Smith, and Ameet Talwalkar. Leaf: A benchmark for federated settings, 2018.

589 Alex Ingerman and Krzys Ostrowski. Introducing tensorflow federated, 2019.

590 Sashank Reddi, Zachary Charles, Manzil Zaheer, Zachary Garrett, Keith Rush, Jakub Konečný,
591 Sanjiv Kumar, and H. Brendan McMahan. Adaptive federated optimization, 2020.

592 Jae Hun Ro, Ananda Theertha Suresh, and Ke Wu. Fedjax: Federated learning simulation with jax.
593 *arXiv preprint arXiv:2108.02117*, 2021.

594 TensorFlow-Federated-Authors. Tensorflow federated stack overflow dataset, 2019.

595 Jianyu Wang, Zachary Charles, Zheng Xu, Gauri Joshi, H. Brendan McMahan, Blaise Aguera y Arcas,
596 Maruan Al-Shedivat, Galen Andrew, Salman Avestimehr, Katharine Daly, Deepesh Data, Suhas
597 Diggavi, Hubert Eichner, Advait Gadhikar, Zachary Garrett, Antonious M. Girgis, Filip Hanzely,
598 Andrew Hard, Chaoyang He, Samuel Horvath, Zhouyuan Huo, Alex Ingerman, Martin Jaggi, Tara
599 Javidi, Peter Kairouz, Satyen Kale, Sai Praneeth Karimireddy, Jakub Konecny, Sanmi Koyejo,
600 Tian Li, Luyang Liu, Mehryar Mohri, Hang Qi, Sashank J. Reddi, Peter Richtarik, Karan Singhal,
601 Virginia Smith, Mahdi Soltanolkotabi, Weikang Song, Ananda Theertha Suresh, Sebastian U. Stich,
602 Ameet Talwalkar, Hongyi Wang, Blake Woodworth, Shanshan Wu, Felix X. Yu, Honglin Yuan,
603 Manzil Zaheer, Mi Zhang, Tong Zhang, Chunxiang Zheng, Chen Zhu, and Wennan Zhu. A field
604 guide to federated optimization, 2021.

605 Haibo Yang, Minghong Fang, and Jia Liu. Achieving linear speedup with partial worker participation
606 in non-iid federated learning, 2021.