

A PROOFS

A.1 PROOF OF THEOREM 1

In this section, we give the proofs in detail. Due to the smoothness in Assumption equation 1, taking expectation of $f(w_{t+1})$ over the randomness in round t , we have

$$\mathbb{E}_t[f(w_{t+1})] \quad (13)$$

$$\leq f(w_t) + \langle \nabla f(w_t), \mathbb{E}_t[w_{t+1} - w_t] \rangle + \frac{L}{2} \mathbb{E}_t[\|w_{t+1} - w_t\|^2] \quad (14)$$

$$= f(w_t) + \langle \nabla f(w_t), \mathbb{E}_t[\eta \eta_L \Delta_t + \eta \eta_L E \nabla f(w_t) - \eta \eta_L E \nabla f(w_t)] \rangle + \frac{L}{2} \eta^2 \eta_L^2 \mathbb{E}_t[\|\Delta_t\|^2] \quad (15)$$

$$= f(w_t) - \eta \eta_L E \|\nabla f(w_t)\|^2 + \eta \underbrace{\langle \nabla f(w_t), \mathbb{E}[\eta_L \Delta_t + \eta_L E \nabla f(w_t)] \rangle}_{A_1} + \frac{L}{2} \eta^2 \eta_L^2 \underbrace{\mathbb{E}_t[\|\Delta_t\|^2]}_{A_2} \quad (16)$$

Note that the term A_1 can be bounded as follows:

$$A_1 = \langle \nabla f(w_t), \mathbb{E}_t[\eta_L \Delta_t + \eta_L E \nabla f(w_t)] \rangle \quad (17)$$

$$= \langle \nabla f(w_t), \mathbb{E}_t[\eta_L \bar{\Delta}_t + \eta_L e_t + \eta_L E \nabla f(w_t)] \rangle \quad (18)$$

$$= \left\langle \nabla f(w_t), \mathbb{E}_t \left[-\frac{1}{K} \sum_{k=1}^K \sum_{\tau=0}^{E-1} \eta_L \nabla F^k(w_{t,\tau}^k) + \eta_L e_t + \eta_L E \frac{1}{K} \sum_{k=1}^K \nabla F^k(w_t) \right] \right\rangle \quad (19)$$

$$= \left\langle \sqrt{\eta_L E} \nabla f(w_t), -\frac{\sqrt{\eta_L}}{K \sqrt{E}} \mathbb{E}_t \left[\sum_{k=1}^K \sum_{\tau=0}^{E-1} (\nabla F^k(w_{t,\tau}^k) - \nabla F^k(w_t)) - K e_t \right] \right\rangle \quad (20)$$

$$\stackrel{(a_1)}{=} \frac{\eta_L E}{2} \|\nabla f(w_t)\|^2 + \frac{\eta_L}{2EK^2} \mathbb{E}_t \left\| \sum_{k=1}^K \sum_{\tau=0}^{E-1} (\nabla F^k(w_{t,\tau}^k) - \nabla F^k(w_t)) - K e_t \right\|^2 \\ - \frac{\eta_L}{2EK^2} \mathbb{E}_t \left\| \sum_{k=1}^K \sum_{\tau=0}^{E-1} \nabla F^k(w_{t,\tau}^k) - K e_t \right\|^2 \quad (21)$$

$$\stackrel{(a_2)}{\leq} \frac{\eta_L E}{2} \|\nabla f(w_t)\|^2 + \frac{\eta_L}{EK^2} \mathbb{E}_t \left\| \sum_{k=1}^K \sum_{\tau=0}^{E-1} (\nabla F^k(w_{t,\tau}^k) - \nabla F^k(w_t)) \right\|^2 \\ - \frac{\eta_L}{2EK^2} \mathbb{E}_t \left\| \sum_{k=1}^K \sum_{\tau=0}^{E-1} \nabla F^k(w_{t,\tau}^k) - K e_t \right\|^2 + \frac{\eta_L \mathbb{E}_t \|e_t\|^2}{E} \quad (22)$$

$$\stackrel{(a_3)}{\leq} \frac{\eta_L E}{2} \|\nabla f(w_t)\|^2 + \frac{\eta_L}{K} \sum_{k=1}^K \sum_{\tau=0}^{E-1} \mathbb{E}_t \|\nabla F^k(w_{t,\tau}^k) - \nabla F^k(w_t)\|^2 \\ - \frac{\eta_L}{2EK^2} \mathbb{E}_t \left\| \sum_{k=1}^K \sum_{\tau=0}^{E-1} \nabla F^k(w_{t,\tau}^k) - K e_t \right\|^2 + \frac{\eta_L \mathbb{E}_t \|e_t\|^2}{E} \quad (23)$$

$$\stackrel{(a_4)}{\leq} \frac{\eta_L E}{2} \|\nabla f(w_t)\|^2 + \frac{\eta_L L^2}{K} \sum_{k=1}^K \sum_{\tau=0}^{E-1} \mathbb{E}_t \|w_{t,\tau}^k - w_t\|^2 \\ - \frac{\eta_L}{2EK^2} \mathbb{E}_t \left\| \sum_{k=1}^K \sum_{\tau=0}^{E-1} \nabla F^k(w_{t,\tau}^k) - K e_t \right\|^2 + \frac{\eta_L \mathbb{E}_t \|e_t\|^2}{E} \quad (24)$$

$$\stackrel{(a_5)}{\leq} \eta_L E \left(\frac{1}{2} + 30\eta_L^2 E^2 L^2 \right) \|\nabla f(w_t)\|^2 + 5\eta_L^3 E^2 L^2 (\rho_L^2 + 6E\rho_G^2) \\ - \frac{\eta_L}{2EK^2} \mathbb{E}_t \left\| \sum_{k=1}^K \sum_{\tau=0}^{E-1} \nabla F^k(w_{t,\tau}^k) - K e_t \right\|^2 + \frac{\eta_L \mathbb{E}_t \|e_t\|^2}{E} \quad (25)$$

where (a_1) follows from that $\langle \mathbf{x}, \mathbf{y} \rangle = \frac{1}{2}[\|\mathbf{x}\|^2 + \|\mathbf{y}\|^2 - \|\mathbf{x} - \mathbf{y}\|^2]$, (a_2) is due to that $\mathbb{E}\|x_1 + x_2\|^2 \leq 2\mathbb{E}[\|x_1\|^2 + \|x_2\|^2]$, (a_3) is due to that $\mathbb{E}\|x_1 + \dots + x_n\|^2 \leq n\mathbb{E}[\|x_1\|^2 + \dots + \|x_n\|^2]$, (a_4) is due to Assumption equation 1 and (a_5) follows from Lemma 2.

The term A_2 can be bounded as

$$A_2 = \mathbb{E}_t[\|\Delta_t\|^2] = \mathbb{E}_t[\|\bar{\Delta}_t + e_t\|^2] \quad (26)$$

$$\stackrel{(a_6)}{\leq} 2\mathbb{E}_t\|\bar{\Delta}_t\|^2 + 2\mathbb{E}_t\|e_t\|^2$$

$$\leq \frac{2}{K^2}\mathbb{E}_t\left[\left\|\sum_{k=1}^K \sum_{\tau=0}^{E-1} g_{t,\tau}^k\right\|^2\right] + 2\mathbb{E}_t\|e_t\|^2 \quad (27)$$

$$\stackrel{(a_7)}{\leq} \frac{2}{K^2}\mathbb{E}_t\left[\left\|\sum_{k=1}^K \sum_{\tau=0}^{E-1} (g_{t,\tau}^k - \nabla F^k(w_{t,\tau}^k))\right\|^2\right] + \frac{2}{K^2}\mathbb{E}_t\left[\left\|\sum_{k=1}^K \sum_{\tau=0}^{E-1} \nabla F^k(w_{t,\tau}^k)\right\|^2\right] + 2\mathbb{E}_t\|e_t\|^2 \quad (28)$$

$$\stackrel{(a_8)}{\leq} \frac{2E}{K}\rho_L^2 + \frac{4}{K^2}\mathbb{E}_t\left[\left\|\sum_{k=1}^K \sum_{\tau=0}^{E-1} \nabla F^k(w_{t,\tau}^k) - Ke_t\right\|^2\right] + \frac{4}{K^2}\mathbb{E}_t\|Ke_t\|^2 + 2\mathbb{E}_t\|e_t\|^2$$

$$= \frac{2E}{K}\rho_L^2 + \frac{4}{K^2}\mathbb{E}_t\left[\left\|\sum_{k=1}^K \sum_{\tau=0}^{E-1} \nabla F^k(w_{t,\tau}^k) - Ke_t\right\|^2\right] + 6\mathbb{E}_t\|e_t\|^2 \quad (29)$$

where both (a_6) is due to that $\mathbb{E}\|x_1 + x_2\|^2 \leq 2\mathbb{E}[\|x_1\|^2 + \|x_2\|^2]$, (a_7) follows the fact that $\mathbb{E}[\|\mathbf{x}\|^2] = \mathbb{E}[\|\mathbf{x} - \mathbb{E}\mathbf{x}\|^2] + \|\mathbb{E}\mathbf{x}\|^2$, and (a_8) is due to Assumption equation 3

Substituting the inequalities of A_1 and A_2 into the original inequality, we have:

$$\mathbb{E}_t[f(w_{t+1})] \quad (30)$$

$$\leq f(w_t) - \eta\eta_L E \|\nabla f(w_t)\|^2 + \underbrace{\eta \langle \nabla f(w_t), \mathbb{E}[\eta_L \Delta_t + \eta_L E \nabla f(w_t)] \rangle}_{A_1} + \underbrace{\frac{L}{2}\eta^2 \eta_L^2 \mathbb{E}_t[\|\Delta_t\|^2]}_{A_2} \quad (31)$$

$$\leq f(w_t) - \eta\eta_L E \|\nabla f(w_t)\|^2$$

$$+ \eta\eta_L E \left(\frac{1}{2} + 30\eta_L^2 E^2 L^2 \right) \|\nabla f(w_t)\|^2 + 5\eta\eta_L^3 E^2 L^2 (\rho_L^2 + 6E\rho_G^2)$$

$$- \frac{\eta\eta_L}{2EK^2} \mathbb{E}_t \left\| \sum_{k=1}^K \sum_{\tau=0}^{E-1} \nabla F^k(w_{t,\tau}^k) - Ke_t \right\|^2 + \frac{\eta\eta_L \mathbb{E}_t\|e_t\|^2}{E}$$

$$+ \frac{EL\eta^2 \eta_L^2}{K} \rho_L^2 + \frac{2L\eta^2 \eta_L^2}{K^2} \mathbb{E}_t \left[\left\| \sum_{k=1}^K \sum_{\tau=0}^{E-1} \nabla F^k(w_{t,\tau}^k) - Ke_t \right\|^2 \right] + 3\eta^2 \eta_L^2 L \mathbb{E}_t\|e_t\|^2 \quad (32)$$

$$= f(w_t) - \eta\eta_L E \left(\frac{1}{2} - 30\eta_L^2 E^2 L^2 \right) \|\nabla f(w_t)\|^2$$

$$+ 5\eta\eta_L^3 E^2 L^2 (\rho_L^2 + 6E\rho_G^2) + \frac{EL\eta^2 \eta_L^2}{K} \rho_L^2 + \left(\frac{\eta\eta_L}{E} + 3\eta^2 \eta_L^2 L \right) \mathbb{E}_t\|e_t\|^2$$

$$- \left(\frac{\eta\eta_L}{2EK^2} - \frac{2L\eta^2 \eta_L^2}{K^2} \right) \mathbb{E}_t \left\| \sum_{k=1}^K \sum_{\tau=0}^{E-1} \nabla F^k(w_{t,\tau}^k) - Ke_t \right\|^2 \quad (33)$$

$$\stackrel{(a_9)}{\leq} f(w_t) - c\eta\eta_L E \|\nabla f(w_t)\|^2 + 5\eta\eta_L^3 E^2 L^2 (\rho_L^2 + 6E\rho_G^2) + \frac{EL\eta^2 \eta_L^2}{K} \rho_L^2 + \left(\frac{\eta\eta_L}{E} + 3\eta^2 \eta_L^2 L \right) \mathbb{E}_t\|e_t\|^2 \quad (34)$$

where (a_9) follows from $\left(\frac{\eta\eta_L}{2EK^2} - \frac{2L\eta^2 \eta_L^2}{K^2} \right) < 0$ if $\eta\eta_L \leq \frac{1}{4EL}$, and that there exists a constant $c > 0$ satisfying $\left(\frac{1}{2} - 30\eta_L^2 E^2 L^2 \right) > c > 0$ if $\eta_L < \frac{1}{\sqrt{60EL}}$.

Rearranging and summing from $t = 0, \dots, T - 1$, we have:

$$\sum_{t=0}^{T-1} c\eta\eta_L E \mathbb{E} \|\nabla f(w_t)\|^2 \quad (35)$$

$$\leq f(w_0) - f(w_T) + TE\eta\eta_L \left[5\eta_L^2 EL^2(\rho_L^2 + 6E\rho_G^2) + \frac{\eta\eta_L L}{K} \rho_L^2 \right] + \left(\frac{\eta\eta_L}{E} + 3\eta^2\eta_L^2 L \right) \sum_{t=0}^{T-1} \mathbb{E}_t \|e_t\|^2 \quad (36)$$

which implies,

$$\min_{t=0, \dots, T-1} \mathbb{E} \|\nabla f(w_t)\|^2 \leq \frac{f_0 - f_*}{c\eta\eta_L ET} + \Phi + \Psi(e_0, \dots, e_{T-1}) \quad (37)$$

where

$$\Phi = \frac{1}{c} \left[5\eta_L^2 EL^2(\rho_L^2 + 6E\rho_G^2) + \frac{\eta\eta_L L}{K} \sigma_L^2 \right] \quad (38)$$

$$\Psi(e_0, \dots, e_{T-1}) = \frac{1 + 3\eta\eta_L LE}{cE^2 T} \sum_{t=0}^{T-1} \mathbb{E}_t \|e_t\|^2 \quad (39)$$

This completes the proof.

A.2 PROOF OF LEMMA 2

Consider a dropout client k and a non-friend client $i \notin \mathcal{B}_k$. We analyze the probability that it is selected by the algorithm. According to our selection rule, i is selected only if it has the highest similarity score with client k so far. Hence, $R_t^{i,k}$ must be greater than $R_t^{j,k}$ for at least one $j \in \mathcal{B}_k \cap \mathcal{S}_t$. Thus, the following inequality holds

$$\Pr\{\phi_t(k) = i\} \leq \Pr\{R_t^{i,k} \geq R_t^{j,k}, \text{ for some } j \in \mathcal{B}_k \cap \mathcal{S}_t\} \quad (40)$$

$$\leq \Pr\{R_t^{i,k} \geq \mu^{i,k} + \frac{\delta}{2}\} + \Pr\{R_t^{j,k} \leq \mu^{j,k} - \frac{\delta}{2}\} \quad (41)$$

$$\leq \exp\left(\frac{-N_t^{i,k}\delta^2}{2}\right) + \exp\left(\frac{-N_t^{j,k}\delta^2}{2}\right) \quad (42)$$

$$\leq 2 \exp\left(\frac{-\beta\delta_k^2 t}{2}\right) \quad (43)$$

where $\delta = \mu^{j,k} - \mu^{i,k} \geq \delta_k$. Because the number of non-friend clients of a client k is at most K , the probability of selecting a non-friend client is thus upper-bounded by $2K \exp\left(\frac{-\beta\delta_k^2 t}{2}\right)$. Taking into account $\delta_{\min} = \min_k \delta_k$ completes the proof.

A.3 PROOF OF THEOREM 2

A sufficient condition for the bound to hold is that after T FL rounds, no friend of client k was eliminated from \mathcal{C}_k^t by running our algorithm. Thus, we are interested in bounding the probability that any particular friend client i is eliminated in a particular round t before T .

$$\Pr(i \text{ is eliminated in round } t) \quad (44)$$

$$\leq \Pr(R_t^{k,j} - R_t^{k,i} \geq \Theta_t, \text{ for some } j \neq i) \quad (45)$$

$$\leq \sum_{j \neq i} \Pr(R_t^{k,j} - R_t^{k,i} \geq \Theta_t) \quad (46)$$

$$\leq K \left(\Pr(R_t^{k,i} \leq \mu^{k,i} - \frac{\Theta_t - \delta_f}{2}) + \Pr(R_t^{k,j^*} \geq \mu^{k,j^*} + \frac{\Theta_t - \delta_f}{2}) \right) \quad (47)$$

$$\leq 2K \exp\left(\frac{-\beta(\Theta_t - \delta_f)^2 t}{2}\right) = q \quad (48)$$

where j^* is the best friend of client k . The last equality holds by letting

$$\Theta_t = \sqrt{\frac{2 \ln(2K) - 2 \ln q}{\beta t}} + \delta_f \quad (49)$$

Next, the probability that a friend client i is eliminated in any round up to round T is bounded as follows

$$\Pr(i \text{ is eliminated up to round } T) \leq \sum_{t \leq T-1} \Pr(i \text{ is eliminated in round } t) \leq Tq \quad (50)$$

Thus,

$$\Pr(\text{any friend of client } k \text{ is eliminated up to round } T) \leq |\mathcal{B}_k|Tq \quad (51)$$

Furthermore,

$$\Pr(\text{any friend of any client is eliminated up to round } T) \leq K|\mathcal{B}_k|Tq \quad (52)$$

Therefore, by letting $p = KB_{\max}Tq$ and

$$\Theta_t = \sqrt{\frac{2 \ln(2K^2TB_{\max}) - 2 \ln p}{\beta t}} + \delta_f \quad (53)$$

we ensure that the probability that no friend of any client was eliminated from the corresponding candidate set by T is at least $1 - p$. This concludes the proof.

A.4 BOUNDS ON $\mathbb{E}\|e_t\|^2$

The error bound with client dropout:

$$\mathbb{E}[\|e_t\|^2] = \mathbb{E}\left[\left\|\frac{1}{K} \sum_{k \in \mathcal{K} \setminus \mathcal{S}_t} (\tilde{\Delta}_t^k - \Delta_t^k)\right\|^2\right] = \mathbb{E}\left[\left\|\frac{1}{K} \sum_{k \in \mathcal{K} \setminus \mathcal{S}_t} \frac{1}{S_t} \sum_{k' \in \mathcal{S}_t} (\Delta_t^{k'} - \Delta_t^k)\right\|^2\right] \quad (54)$$

$$\leq \frac{(K - S_t)^2}{K^2} \sigma_P^2 \leq \alpha^2 \sigma_P^2 \quad (55)$$

The error bound with friend model substitution (full information) :

$$\mathbb{E}[\|e_t\|^2] = \mathbb{E}\left[\left\|\frac{1}{K} \sum_{k \in \mathcal{K} \setminus \mathcal{S}_t} (\tilde{\Delta}_t^k - \Delta_t^k)\right\|^2\right] = \mathbb{E}\left[\left\|\frac{1}{K} \sum_{k \in \mathcal{K} \setminus \mathcal{S}_t} (\Delta_t^{\phi_t(k)} - \Delta_t^k)\right\|^2\right] \quad (56)$$

$$\leq \frac{(K - S_t)^2}{K^2} \sigma_F^2 \leq \alpha^2 \sigma_F^2 \quad (57)$$

where $\phi_t(k)$ is a friend of k that does not dropout in round t .

The error bound with friend model substitution (learning):

$$\mathbb{E}\|e_t\|^2 = \mathbb{E}\left[\left\|\frac{1}{K} \sum_{k \in \mathcal{K} \setminus \mathcal{S}_t} (\tilde{\Delta}_t^k - \Delta_t^k)\right\|^2\right] \leq \frac{K - S_t}{K^2} \sum_{k \in \mathcal{K} \setminus \mathcal{S}_t} \mathbb{E}\|\tilde{\Delta}_t^k - \Delta_t^k\|^2 \quad (58)$$

$$\leq \frac{K - S_t}{K^2} \sum_{k \in \mathcal{K} \setminus \mathcal{S}_t} \left(2K \exp\left(\frac{-\beta \delta_k^2 t}{2}\right) \sigma_P^2 + (1 - 2K \exp\left(\frac{-\beta \delta_k^2 t}{2}\right)) \sigma_F^2\right) \quad (59)$$

$$\leq \frac{(K - S_t)^2}{K^2} \left(\sigma_F^2 + 2K \exp\left(\frac{-\beta \delta_{\min}^2 t}{2}\right) (\sigma_P^2 - \sigma_F^2)\right) \quad (60)$$

$$\leq \alpha^2 \left(\sigma_F^2 + 2K \exp\left(\frac{-\beta \delta_{\min}^2 t}{2}\right) (\sigma_P^2 - \sigma_F^2)\right) \quad (61)$$

A.5 BOUNDS ON $\Psi(e_0, \dots, e_{T-1})$ WITH FRIEND MODEL SUBSTITUTION (LEARNING)

$$\Psi(e_0, \dots, e_{T-1}) \quad (62)$$

$$= \frac{1 + 3\eta\eta_L LE}{cE^2 T} \sum_{t=0}^{T-1} \mathbb{E}_t[\|e_t\|^2] \quad (63)$$

$$\leq \frac{\alpha^2 \sigma_F^2 (1 + 3\eta\eta_L LE)}{cE^2} + 2K \frac{\alpha^2 (\sigma_P^2 - \sigma_F^2) (1 + 3\eta\eta_L LE)}{cE^2 T} \sum_{t=0}^{T-1} \exp\left(\frac{-\beta\delta_{\min}^2 t}{2}\right) \quad (64)$$

$$\leq \frac{\alpha^2 \sigma_F^2 (1 + 3\eta\eta_L LE)}{cE^2} + 2K \frac{\alpha^2 \sigma_P^2 (1 + 3\eta\eta_L LE)}{cE^2 T} \sum_{t=0}^{T-1} \exp\left(\frac{-\beta\delta_{\min}^2 t}{2}\right) \quad (65)$$

$$\leq \frac{\alpha^2 \sigma_F^2 (1 + 3\eta\eta_L LE)}{cE^2} + 2K \frac{\alpha^2 \sigma_P^2 (1 + 3\eta\eta_L LE)}{cE^2} \frac{1 - \exp\left(\frac{-\beta\delta_{\min}^2 T}{2}\right)}{T \left(1 - \exp\left(\frac{-\beta\delta_{\min}^2}{2}\right)\right)} \quad (66)$$

$$\leq \Psi^* + 2K\bar{\Psi} \frac{1 - \exp\left(\frac{-\beta\delta_{\min}^2 T}{2}\right)}{T \left(1 - \exp\left(\frac{-\beta\delta_{\min}^2}{2}\right)\right)} \quad (67)$$

B THE RELAXED FRIEND PRESENCE CASE

In this section, we consider a relaxed case without the Assumption 5. Suppose that a friend of the dropout client is present in each round with a probability equals $1 - r$, then the probability that our algorithm selects a non-friend client for a dropout client in round t is upper bounded by $(1 - r)2K \exp\left(\frac{-\beta\delta_{\min}^2 t}{2}\right) + r$. Then we can get the bound on $\mathbb{E}\|e_t\|^2$:

$$\mathbb{E}\|e_t\|^2 \leq \alpha^2 \left(\sigma_F^2 + 2(1 - r)K \exp\left(\frac{-\beta\delta_{\min}^2 t}{2}\right) \sigma_P^2 + r\sigma_P^2 \right) \quad (68)$$

Plugging this bound in $\Psi(e_0, \dots, e_{T-1})$, we can get the accumulate substitution error

$$\Psi(e_0, \dots, e_{T-1}) \leq \Psi^* + r\bar{\Psi} + 2(1 - r)K\bar{\Psi} \frac{1 - \exp\left(\frac{-\beta\delta_{\min}^2 T}{2}\right)}{T \left(1 - \exp\left(\frac{-\beta\delta_{\min}^2}{2}\right)\right)} \quad (69)$$

The convergence bound without the Assumption 5 has an additional constant term resulting from friend absence, and the additional constant term cannot be eliminated with time or batch size increase.

Proof. The error bound with friend model substitution (learning) under relaxed friend presence case equals:

$$\mathbb{E}\|e_t\|^2 = \mathbb{E} \left\| \frac{1}{K} \sum_{k \in \mathcal{K} \setminus \mathcal{S}_t} (\tilde{\Delta}_t^k - \Delta_t^k) \right\|^2 \leq \frac{K - S_t}{K^2} \sum_{k \in \mathcal{K} \setminus \mathcal{S}_t} \mathbb{E} \|\tilde{\Delta}_t^k - \Delta_t^k\|^2 \quad (70)$$

$$\leq \alpha^2 \left((1 - r)2K \exp\left(\frac{-\beta\delta_{\min}^2 t}{2}\right) + r \right) \sigma_P^2 + \alpha^2 \left(1 - \left((1 - r)2K \exp\left(\frac{-\beta\delta_{\min}^2 t}{2}\right) + r \right) \right) \sigma_F^2 \quad (71)$$

$$\leq \alpha^2 \sigma_F^2 + \alpha^2 \left((1 - r)2K \exp\left(\frac{-\beta\delta_{\min}^2 t}{2}\right) + r \right) \sigma_P^2 \quad (72)$$

$$\leq \alpha^2 \left(\sigma_F^2 + 2(1 - r)K \exp\left(\frac{-\beta\delta_{\min}^2 t}{2}\right) \sigma_P^2 + r\sigma_P^2 \right) \quad (73)$$

And the corresponding accumulate substitution error equals:

$$\Psi(e_0, \dots, e_{T-1}) \quad (74)$$

$$= \frac{1 + 3\eta\eta_L LE}{cE^2 T} \sum_{t=0}^{T-1} \mathbb{E}_t[\|e_t\|^2] \quad (75)$$

$$\leq \frac{\alpha^2(\sigma_F^2 + r\sigma_P^2)(1 + 3\eta\eta_L LE)}{cE^2} + 2K \frac{\alpha^2(\sigma_P^2 - \sigma_F^2)(1 + 3\eta\eta_L LE)}{cE^2 T} \sum_{t=0}^{T-1} \exp\left(\frac{-\beta\delta_{\min}^2 t}{2}\right) \quad (76)$$

$$\leq \frac{\alpha^2(\sigma_F^2 + r\sigma_P^2)(1 + 3\eta\eta_L LE)}{cE^2} + 2K \frac{\alpha^2\sigma_P^2(1 + 3\eta\eta_L LE)}{cE^2 T} \sum_{t=0}^{T-1} \exp\left(\frac{-\beta\delta_{\min}^2 t}{2}\right) \quad (77)$$

$$\leq \frac{\alpha^2(\sigma_F^2 + r\sigma_P^2)(1 + 3\eta\eta_L LE)}{cE^2} + 2K \frac{\alpha^2\sigma_P^2(1 + 3\eta\eta_L LE)}{cE^2} \frac{1 - \exp\left(\frac{-\beta\delta_{\min}^2 T}{2}\right)}{T \left(1 - \exp\left(\frac{-\beta\delta_{\min}^2}{2}\right)\right)} \quad (78)$$

$$\leq \Psi^* + r\bar{\Psi} + 2K\bar{\Psi} \frac{1 - \exp\left(\frac{-\beta\delta_{\min}^2 T}{2}\right)}{T \left(1 - \exp\left(\frac{-\beta\delta_{\min}^2}{2}\right)\right)} \quad (79)$$

□

C EXPERIMENT SETTINGS

We use Python3 and the Pytorch library, and our code is adapted from Jadhav (2020), which is under the MIT License. The experiments were run on an Ubuntu 18.04 machine with an Intel Core i7-10700KF 3.8GHz CPU and GeForce RTX 3070 GPU. All experiments results are averaged over 10 repeats.

We perform experiments on two standard public datasets, namely MNIST and CIFAR-10, which are widely used in FL experiments, in a clustered setting as well as a general setting. In the clustered settings (one on MNIST and one on CIFAR-10), we artificially create 5 client clusters where clients in the same cluster possess data samples with the same labels. Thus, clients in the same cluster are naturally regarded as friends. However, the clustering structure is *unknown* to our algorithm. Such a clustering setting provides a controlled environment for us to evaluate the friend discovery performance of FL-FDMS. In the general setting (on CIFAR-10), 20 clients receive a random subset of the whole dataset using a common way of generating non-iid FL datasets that is widely used in existing works.

C.1 FL DATASET

Clustered Setting - MNIST: The MNIST dataset has 60000 training data samples with 10 classes. The training dataset is first split into 10 sub-datasets with samples in the same sub-dataset having the same label. There are 20 clients which are grouped into 5 client clusters with an equal number of clients. Each client cluster is associated with 2 randomly drawn sub-datasets. Then each client randomly draws 200 samples from its corresponding two sub-datasets. This approach to creating the FL dataset was introduced in a recent clustered FL work Ghosh et al. (2020).

Clustered Setting - CIFAR-10: The CIFAR-10 dataset has 50000 training data samples with 10 classes. The training dataset is first split into 10 sub-datasets with samples in the same sub-dataset having the same label. There are 20 clients which are grouped into 5 client clusters with an equal number of clients. Each client cluster is associated with 2 randomly drawn sub-datasets. Then each client randomly draws 1000 samples from its corresponding two sub-datasets.

General Setting - CIFAR-10: The CIFAR-10 dataset has 50000 training data samples. After shuffling the samples in label order, all samples are divided into 250 partitions with each partition having 200 samples. There are 20 clients. Each client then randomly picks 2 partitions. This method is a common way of generating non-i.i.d. FL dataset, which is widely used in the existing works McMahan et al. (2017); Li et al. (2021)

C.2 FL MODELS

MNIST: The CNN model has two 5×5 convolution layers, a fully connected layer with 320 units and ReLU activation, and a final output layer with softmax. The first convolution layer has 10 channels while the second one has 20 channels. Both layers are followed by 2×2 max pooling. The following parameters are used for training: the local batch size $BS = 5$, the number of local iterations $E = 2$, the local learning rate $\eta_L = 0.1$ and the global learning rate $\eta = 0.1$.

CIFAR-10: The CNN model has two 5×5 convolution layers, three fully connected layers and ReLU activation, and a final output layer with softmax. The following parameters are used for training: the local batch size $BS = 20$, the number of local iterations $E = 2$, the local learning rate $\eta_L = 0.1$ and the global learning rate $\eta = 0.1$.

C.3 THE DRAWBACK OF STALE BENCHMARK

Two apparent drawbacks of the stale approach are, firstly, the local model updates can be very outdated if a client keeps dropping out, and secondly, the server has to keep a copy of the most recent local model update for every client, thereby incurring a large storage cost when the number of clients is large.

D ADDITIONAL EXPERIMENTS

The error bound of FL-FDMS $\mathbb{E}\|e_t\|^2$ in Eq.11 is influenced by the number of local iterations E and the number of clients K . Next we perform additional experiments to explore their impacts.

D.1 IMPACT OF NUMBER OF LOCAL ITERATIONS E

We present more results on the performance comparison in the MNIST clustered setting and the CIFAR-10 clustered setting with different E . We fix $\alpha = 0.5$ and $K = 20$ for all the following experiments. To investigate the impact of E , we consider two values $E = 1$ and $E = 5$.

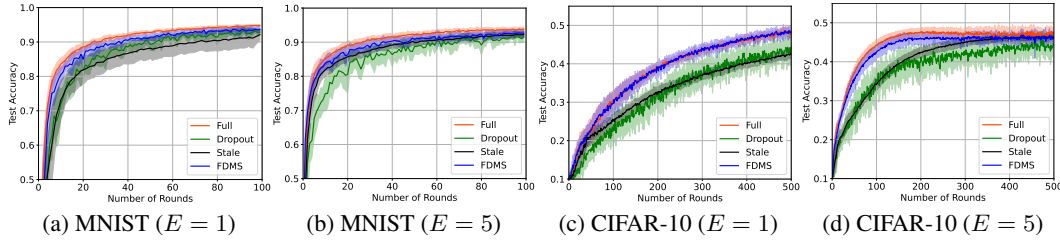


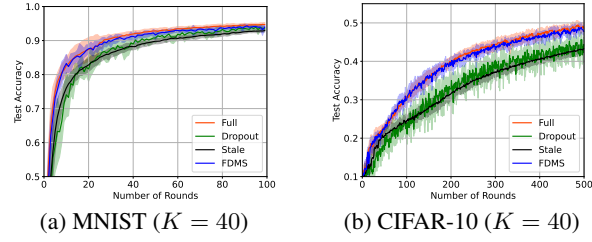
Figure 5: Performance comparison with $\alpha = 0.5$ and $K = 20$

In Fig.5, we find that the **FL-FDMS** still shows the superior performance in terms of test accuracy and convergence speed. However, **Dropout** and **Stale** show different trends for different E . For a larger E , using staled models tends to help the dropout situation better.

D.2 IMPACT OF NUMBER OF CLIENTS K

To investigate the impact of K , we fix $E = 2$ and increase the number of clients to $K = 40$. To keep the same total amount of data in the system, we adjust just the number of data samples on each client. For MNIST, each client now has 100 samples. For CIFAR-10, each client has 500 samples. Other settings are as described in Appendix C.

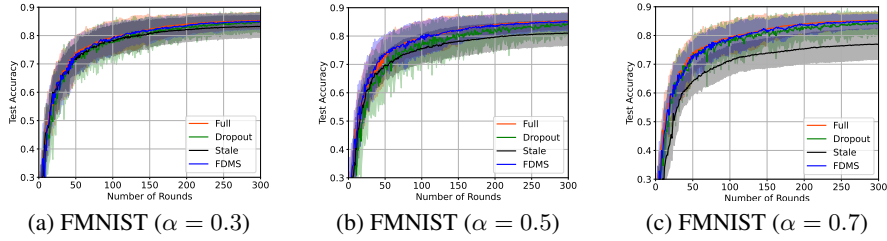
By comparing Fig.6 and the corresponding parts in Fig.1, we find that as K increases, the **FL-FDMS** outperforms **Dropout** and **Stale** even more. This is because as K increases, more clients dropout. If the model updates from dropout clients are not compensated, the global model can gradually deviate from the optimal value and eventually degrade the learning performance and affect

Figure 6: Performance comparison with $\alpha = 0.5$ and $E = 2$

the system stability. The additional experiments further verify that **FL-FDMS** can handle well the client dropout in FL.

D.3 ADDITIONAL EXPERIMENTS ON THE FMNIST DATASETS

We present additional performance comparison results in the FMNIST clustered setting, and the results are consistent with the conclusions we have drawn from prior other datasets.

Figure 7: Performance comparison on the FMNIST clustered setting with various α