APPENDIX A
PROOF OF PROPOSITION 1

First notice $v_t^{k,r}$ is essentially the average short-term label distribution of periods $t-M+1$ through t, thus

$$\mathbb{E}[(v_t^{k,r}-\pi^{k,r})^2]=\mathbb{E}[(\frac{1}{M}\sum_{m=1}^{M}u_{t-m+1}^{k,r}-\pi^{k,r})^2] \tag{23}$$

$$=\frac{1}{M^2}\sum_{m=1}^{M}\mathbb{E}[(u_{t-m+1}^{k,r}-\pi^{k,r})\sum_{m'=1}^{M}(u_{t-m'+1}^{k,r}-\pi^{k,r})] \tag{24}$$

$$\leq\frac{1}{M}(\min\{2\Gamma+1,M\})\delta^2 \tag{25}$$

where Eq. (25) is derived based on Assumption 1, taking into account the two scenarios: whether $M<\Gamma$ or not.

APPENDIX B
PROOF OF PROPOSITION 2

The discrepancy can be bounded as follows.

$$\mathbb{E}[(v_t^{k,r}-\pi^{k,r})^2]=\mathbb{E}[(\frac{\theta}{M}\sum_{\tau=0}^{t}(1-\frac{\theta}{M})^{\tau}u_{t-\tau}^{k,r}-\pi^{k,r})^2] \tag{26}$$

$$=\mathbb{E}[(\frac{\theta}{M}\sum_{\tau=0}^{t}(1-\frac{\theta}{M})^{\tau}(u_{t-\tau}^{k,r}-\pi^{k,r})+(1-\frac{\theta}{M})^{t}(-\pi^{k,r}))^2] \tag{27}$$

$$\leq 2\mathbb{E}[(\frac{\theta}{M}\sum_{\tau=0}^{t}(1-\frac{\theta}{M})^{\tau}(u_{t-\tau}^{k,r}-\pi^{k,r}))^2]+2(1-\frac{\theta}{M})^{2t}(\pi^{k,r})^2 \tag{28}$$

$$\leq\frac{2\theta^2}{M^2}\sum_{\tau=0}^{t}(1-\frac{\theta}{M})^{\tau}\mathbb{E}[(u_{t-\tau}^{k,r}-\pi^{k,r})\sum_{\tau'=0}^{t}(1-\frac{\theta}{M})^{\tau'}(u_{t-\tau'}^{k,r}-\pi^{k,r})]+2(1-\frac{\theta}{M})^{2t}(\pi^{k,r})^2 \tag{29}$$

$$\leq\frac{2\theta^2}{M^2}\sum_{\tau=0}^{t}(1-\frac{\theta}{M})^{\tau}(1-\frac{\theta}{M})^{\tau}\sum_{i=-\Gamma}^{\Gamma}(1-\frac{\theta}{M})^{i}\delta^2+2(1-\frac{\theta}{M})^{2t}(\pi^{k,r})^2 \tag{30}$$

$$=2\frac{\theta}{M}\frac{1-(1-\frac{\theta}{M})^{2t}}{1-(1-\frac{\theta}{M})^2}\left((1-\frac{\theta}{M})^{-\Gamma}-(1-\frac{\theta}{M})^{\Gamma+1}\right)\delta^2+2(1-\frac{\theta}{M})^{2t}(\pi^{k,r})^2 \tag{31}$$

$$=2(1-(1-\frac{\theta}{M})^{2t})\frac{(1-\frac{\theta}{M})^{-\Gamma}-(1-\frac{\theta}{M})^{\Gamma+1}}{2-\frac{\theta}{M}}\delta^2+2(1-\frac{\theta}{M})^{2t}(\pi^{k,r})^2 \tag{32}$$

where the Eq.(27) uses $\frac{\theta}{M}\sum_{\tau=0}^{t}(1-\frac{\theta}{M})^{\tau}+(1-\frac{\theta}{M})^{t}=1$; Eq.(28) is a result of the triangle inequality; Eq.(30) uses Assumption 1.

APPENDIX C
PROOF OF COROLLARY 1

It is easy to see that $\mathbb{E}[(v_t^{k,r}-\pi^{k,r})^2]$ is a weighted sum of $(\pi^{k,r})^2$ and $\frac{(1-\frac{\theta}{M})^{-\Gamma}-(1-\frac{\theta}{M})^{\Gamma+1}}{2-\frac{\theta}{M}}\delta^2$ where the weight $(1-\frac{\theta}{M})^{2t}$ decreases with t. Moreover, it is easy to prove that $\frac{(1-\frac{\theta}{M})^{-\Gamma}-(1-\frac{\theta}{M})^{\Gamma+1}}{2-\frac{\theta}{M}}\delta^2$ is increasing in θ/M. Thus, by choosing θ sufficiently small, $\frac{(1-\frac{\theta}{M})^{-\Gamma}-(1-\frac{\theta}{M})^{\Gamma+1}}{2-\frac{\theta}{M}}\delta^2$ can be made smaller than $(\pi^{k,r})^2$. Therefore, the weighted sum decreases with time and approaches $\frac{(1-\frac{\theta}{M})^{-\Gamma}-(1-\frac{\theta}{M})^{\Gamma+1}}{2-\frac{\theta}{M}}\delta^2$ in the limit.

APPENDIX D
PROOF OF PROPOSITION 3

We bound the discrepancy as follows. First plugging $\theta_t=\frac{B}{B_s t}$ into Eq.(14), we have

$$\tilde{n}_r=\frac{t-1}{t}n_r(\mathcal{L}_{t-1}^k)+\frac{B}{B_s t}n_r(\mathcal{S}_t^k) \tag{33}$$

$$=\frac{t-1}{t}\left(\frac{t-2}{t-1}n_r(\mathcal{L}_{t-2}^k)+\frac{B}{B_s(t-1)}n_r(\mathcal{S}_{t-1}^k)\right)+\frac{B}{B_s t}n_r(\mathcal{S}_t^k) \tag{34}$$

$$=\frac{B}{B_s t}\sum_{\tau=1}^{t}n_r(\mathcal{S}_\tau^k) \tag{35}$$

so we can obtain $v_t^{k,r}=\frac{1}{t}\sum_{\tau=1}^{t}u_t^{k,r}$, then we will get

$$\mathbb{E}[(v_t^{k,r}-\pi^{k,r})^2]=\mathbb{E}[(\frac{1}{t}\sum_{\tau=1}^{t}u_\tau^{k,r}-\pi^{k,r})^2]=\mathbb{E}[(\frac{1}{t}\sum_{\tau=1}^{t}(u_\tau^{k,r}-\pi^{k,r}))^2] \tag{36}$$

$$=\frac{1}{t^2}\sum_{\tau=1}^{t}\mathbb{E}\left[(u_\tau^{k,r}-\pi^{k,r})\sum_{\tau'=1}^{t}(u_{\tau'}^{k,r}-\pi^{k,r})\right]\leq\frac{2\Gamma+1}{t}\delta^2 \tag{37}$$

where the last inequality uses Assumption 1.

APPENDIX E
PROOF OF LEMMA 1

The difference between the *real local gradient* and *virtual local gradient* can be bounded as follows:

$$\mathbb{E}[|g_{t,\tau}^k-\hat{g}_{t,\tau}^k|^2] \tag{38}$$

$$=\mathbb{E}[|\sum_{r=1}^{R}v_t^{k,r}\nabla F^k(w_{t,\tau}^k;\mathcal{L}_t^{k,r})-\sum_{r=1}^{R}\pi^{k,r}\nabla F^k(w_{t,\tau}^k;\mathcal{L}_t^{k,r})|^2] \tag{39}$$

$$=\mathbb{E}[|\sum_{r=1}^{R}(v_t^{k,r}-\pi^{k,r})\nabla F^k(w_{t,\tau}^k;\mathcal{L}_t^{k,r})|^2] \tag{40}$$

$$\leq R\sum_{r=1}^{R}\mathbb{E}[|(v_t^{k,r}-\pi^{k,r})\nabla F^k(w_{t,\tau}^k;\mathcal{L}_t^{k,r})|^2] \tag{41}$$

$$\leq R\sum_{r=1}^{R}\mathbb{E}[|(v_t^{k,r}-\pi^{k,r})|^2]\mathbb{E}[|\nabla F^k(w_{t,\tau}^k;\mathcal{L}_t^{k,r})|^2] \tag{42}$$

$$\leq R^2\lambda_t^2\sigma_M^2 \tag{43}$$

where Eq. (41) is based on Cauchy Schwarz inequality, Eq. (42) is based on Holder inequality and Eq. (43) is based on Assumption 5.

Then the proof of the second inequality is as follows:

$$\mathbb{E}[|\hat{g}_{t,\tau}^k - \nabla f^k(w_{t,\tau}^k)|^2] \tag{44}$$
$$= \mathbb{E}[|\sum_{r=1}^{R}\pi^{k,r}\nabla F^k(w_{t,\tau}^k;\mathcal{L}_t^{k,r}) - \sum_{r=1}^{R}\pi^{k,r}\nabla f^{k,r}(w_{t,\tau}^k)|^2] \tag{45}$$
$$= \mathbb{E}[|\sum_{r=1}^{R}\pi^{k,r}(\nabla F^k(w_{t,\tau}^k;\mathcal{L}_t^{k,r}) - \nabla f^{k,r}(w_{t,\tau}^k))|^2] \tag{46}$$
$$\leq R\overline{\pi}^2\sum_{r=1}^{R}\mathbb{E}[|(\nabla F^k(w_{t,\tau}^k;\mathcal{L}_t^{k,r}) - \nabla f^{k,r}(w_{t,\tau}^k))|^2] \tag{47}$$
$$\leq 2R^2\overline{\pi}^2\sigma_M^2 \tag{48}$$

where $\overline{\pi} = \max_{k,r}\pi^{k,r}$ is the maximum ratio of the long-term label distribution and Eq. (48) is based on Assumption 5.

Appendix F
Proof of Lemma 2

In this subsection, we will get the local updates bound,

$$\mathbb{E}[\|w_{t,\tau}^k - w_t\|^2] \tag{49}$$
$$=\mathbb{E}[\|w_{t,\tau-1}^k - w_t - \eta_L g_{t,\tau-1}^k\|^2] \tag{50}$$
$$=\mathbb{E}[\|w_{t,\tau-1}^k - w_t - \eta_L(g_{t,\tau-1}^k - \hat{g}_{t,\tau-1}^k + \hat{g}_{t,\tau-1}^k - \nabla f^k(w_{t,\tau-1}^k) + \nabla f^k(w_{t,\tau-1}^k) - \nabla f^k(w_t) + \nabla f^k(w_t))\|^2] \tag{51}$$
$$\leq \left(1+\frac{1}{2E-1}\right)\mathbb{E}\|w_{t,\tau-1}^k - w_t\|^2 + \eta_L^2\mathbb{E}\|g_{t,\tau-1}^k - \hat{g}_{t,\tau-1}^k\|^2 + 6E\eta_L^2\mathbb{E}\|\hat{g}_{t,\tau-1}^k - \nabla f^k(w_{t,\tau-1}^k)\|^2 + 6E\eta_L^2\mathbb{E}\|\nabla f^k(w_{t,\tau-1}^k) - \nabla f^k(w_t)\|^2 + 6E\eta_L^2\mathbb{E}\|\nabla f^k(w_t)\|^2 \tag{52}$$
$$\leq \left(1+\frac{1}{2E-1}\right)\mathbb{E}\|w_{t,\tau-1}^k - w_t\|^2 + \eta_L^2R^2\lambda_t^2\sigma_M^2 + 12E\eta_L^2R^2\overline{\pi}^2\sigma_M^2 + 6E\eta_L^2L^2\mathbb{E}\|w_{t,\tau-1}^k - w_t\|^2 + 6E\eta_L^2\sigma_G^2 + 6E\eta_L^2(A^2+1)\|\nabla f(x)\|^2 \tag{53}$$
$$\leq \left(1+\frac{1}{E-1}\right)\mathbb{E}\|w_{t,\tau-1}^k - w_t\|^2 + \eta_L^2R^2\lambda_t^2\sigma_M^2 + 12E\eta_L^2R^2\overline{\pi}^2\sigma_M^2 + 6E\eta_L^2\sigma_G^2 + 6E\eta_L^2(A^2+1)\|\nabla f(x)\|^2 \tag{54}$$

where Eq. (53) depends on the result of Eq. (48) and Assumption 4.

Unrolling the recursion, we get:

$$\frac{1}{K}\sum_{k=1}^{K}\mathbb{E}[\|w_{t,\tau}^k - w_t\|^2] \tag{55}$$
$$\leq\sum_{p=0}^{\tau-1}\left(1+\frac{1}{E-1}\right)^p[\eta_L^2R^2\lambda_t^2\sigma_M^2 + 12E\eta_L^2R^2\overline{\pi}^2\sigma_M^2 + 6E\eta_L^2\sigma_G^2 + 6E\eta_L^2(A^2+1)\|\nabla f(x)\|^2] \tag{56}$$
$$\leq(E-1)\left[(1+\frac{1}{E-1})^E - 1\right][\eta_L^2R^2\lambda_t^2\sigma_M^2 + 12E\eta_L^2R^2\overline{\pi}^2\sigma_M^2 + 6E\eta_L^2\sigma_G^2 + 6E\eta_L^2(A^2+1)\|\nabla f(x)\|^2] \tag{57}$$
$$\leq 5E\eta_L^2R^2\lambda_t^2\sigma_M^2 + 60E^2\eta_L^2R^2\overline{\pi}^2\sigma_M^2 + 30E^2\eta_L^2\sigma_G^2 + 30E^2\eta_L^2(A^2+1)\|\nabla f(x)\|^2 \tag{58}$$

This completes the proof of lemma 2.

Appendix G
Proof of Theorem 1

In this section, we give the proofs in detail. Due to the smoothness in Assumption (2), taking expectation of $f(w_{t+1})$ over the randomness in round t, we have

$$\mathbb{E}_t[f(w_{t+1})] \tag{59}$$
$$\leq f(w_t) + \langle\nabla f(w_t), \mathbb{E}_t[w_{t+1} - w_t]\rangle + \frac{L}{2}\mathbb{E}_t[\|w_{t+1} - w_t\|^2] \tag{60}$$
$$= f(w_t) + \langle\nabla f(w_t), \mathbb{E}_t[\eta\eta_L\Delta_t + \eta\eta_L E\nabla f(w_t) - \eta\eta_L E\nabla f(w_t)]\rangle + \frac{L}{2}\eta^2\eta_L^2\mathbb{E}_t[\|\Delta_t\|^2] \tag{61}$$
$$= f(w_t) - \eta\eta_L E\|\nabla f(w_t)\|^2 + \eta\underbrace{\langle\nabla f(w_t), \mathbb{E}[\eta_L\Delta_t + \eta_L E\nabla f(w_t)]\rangle}_{A_1} + \frac{L}{2}\eta^2\eta_L^2\underbrace{\mathbb{E}_t[\|\Delta_t\|^2]}_{A_2} \tag{62}$$

Note that the term A_1 can be bounded as follows:

$$A_1 = \langle\nabla f(w_t), \mathbb{E}_t[\eta_L\Delta_t + \eta_L E\nabla f(w_t)]\rangle \tag{63}$$
$$=\langle\nabla f(w_t), \mathbb{E}_t[\eta_L\bar{\Delta}_t + \eta_L e_t + \eta_L E\nabla f(w_t)]\rangle \tag{64}$$
$$=\langle\nabla f(w_t), \mathbb{E}_t[-\frac{1}{K}\sum_{k=1}^{K}\sum_{\tau=0}^{E-1}\eta_L\nabla F^k(w_{t,\tau}^k) + \eta_L e_t + \eta_L E\frac{1}{K}\sum_{k=1}^{K}\nabla F^k(w_t)]\rangle \tag{65}$$
$$=\langle\sqrt{\eta_L E}\nabla f(w_t), -\frac{\sqrt{\eta_L}}{K\sqrt{E}}\mathbb{E}_t[\sum_{k=1}^{K}\sum_{\tau=0}^{E-1}(\nabla F^k(w_{t,\tau}^k) - \nabla F^k(w_t)) - Ke_t]\rangle \tag{66}$$
$$\overset{(a_1)}{=}\frac{\eta_L E}{2}\|\nabla f(w_t)\|^2 + \frac{\eta_L}{2EK^2}\mathbb{E}_t\left\|\sum_{k=1}^{K}\sum_{\tau=0}^{E-1}(\nabla F^k(w_{t,\tau}^k) - \nabla F^k(w_t)) - Ke_t\right\|^2 - \frac{\eta_L}{2EK^2}\mathbb{E}_t\left\|\sum_{k=1}^{K}\sum_{\tau=0}^{E-1}\nabla F^k(w_{t,\tau}^k) - Ke_t\right\|^2 \tag{67}$$
$$\overset{(a_2)}{\leq}\frac{\eta_L E}{2}\|\nabla f(w_t)\|^2 + \frac{\eta_L}{EK^2}\mathbb{E}_t\left\|\sum_{k=1}^{K}\sum_{\tau=0}^{E-1}(\nabla F^k(w_{t,\tau}^k) - \nabla F^k(w_t))\right\|^2 - \frac{\eta_L}{2EK^2}\mathbb{E}_t\left\|\sum_{k=1}^{K}\sum_{\tau=0}^{E-1}\nabla F^k(w_{t,\tau}^k) - Ke_t\right\|^2 + \frac{\eta_L\mathbb{E}_t\|e_t\|^2}{E} \tag{68}$$

$$\overset{(a_3)}{\leq} \frac{\eta_L E}{2}\|\nabla f(w_t)\|^2 + \frac{\eta_L}{K}\sum_{k=1}^{K}\sum_{\tau=0}^{E-1}\mathbb{E}_t||\nabla F^k(w_{t,\tau}^k) - \nabla F^k(w_t)||^2 - \frac{\eta_L}{2EK^2}\mathbb{E}_t\left\|\sum_{k=1}^{K}\sum_{\tau=0}^{E-1}\nabla F^k(w_{t,\tau}^k) - Ke_t\right\|^2 + \frac{\eta_L \mathbb{E}_t\|e_t\|^2}{E} \tag{69}$$

$$\overset{(a_4)}{\leq} \frac{\eta_L E}{2}\|\nabla f(w_t)\|^2 + \frac{\eta_L L^2}{K}\sum_{k=1}^{K}\sum_{\tau=0}^{E-1}\mathbb{E}_t\left\|w_{t,\tau}^k - w_t\right\|^2 - \frac{\eta_L}{2EK^2}\mathbb{E}_t\left\|\sum_{k=1}^{K}\sum_{\tau=0}^{E-1}\nabla F^k(w_{t,\tau}^k) - Ke_t\right\|^2 + \frac{\eta_L \mathbb{E}_t\|e_t\|^2}{E} \tag{70}$$

$$\overset{(a_5)}{\leq} \eta_L E(\frac{1}{2} + 30(A^2+1)\eta_L^2 E^2 L^2)\|\nabla f(w_t)\|^2 + 5\eta_L^3 EL^2\left(R^2\lambda_t^2\sigma_M^2 + 12ER^2\overline{\pi}^2\sigma_M^2 + 6E\sigma_G^2\right) - \frac{\eta_L}{2EK^2}\mathbb{E}_t\left\|\sum_{k=1}^{K}\sum_{\tau=0}^{E-1}\nabla F^k(w_{t,\tau}^k) - Ke_t\right\|^2 + \frac{\eta_L \mathbb{E}_t\|e_t\|^2}{E} \tag{71}$$

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

The term A_2 can be bounded as

$$A_2 = \mathbb{E}_t[\|\Delta_t\|^2] = \mathbb{E}_t[\|\bar{\Delta}_t + e_t\|^2] \tag{72}$$

$$\overset{(a_6)}{\leq} 2\mathbb{E}_t\|\bar{\Delta}_t\|^2 + 2\mathbb{E}_t\|e_t\|^2 \tag{73}$$

$$\leq \frac{2}{K^2}\mathbb{E}_t\left[\left\|\sum_{k=1}^{K}\sum_{\tau=0}^{E-1}\hat{g}_{t,\tau}^k\right\|^2\right] + 2\mathbb{E}_t\|e_t\|^2 \tag{74}$$

$$\overset{(a_7)}{\leq} \frac{2}{K^2}\mathbb{E}_t\left[\left\|\sum_{k=1}^{K}\sum_{\tau=0}^{E-1}(\hat{g}_{t,\tau}^k - \nabla F^k(w_{t,\tau}^k))\right\|^2\right] + \frac{2}{K^2}\mathbb{E}_t\left[\left\|\sum_{k=1}^{K}\sum_{\tau=0}^{E-1}\nabla F^k(w_{t,\tau}^k)\right\|^2\right] + 2\mathbb{E}_t\|e_t\|^2 \tag{75}$$

$$\overset{(a_8)}{\leq} \frac{4E}{K}R^2\overline{\pi}^2\sigma_M^2 + \frac{4}{K^2}\mathbb{E}_t\left[\left\|\sum_{k=1}^{K}\sum_{\tau=0}^{E-1}\nabla F^k(w_{t,\tau}^k) - Ke_t\right\|^2\right] + \frac{4}{K^2}\mathbb{E}_t\left\|Ke_t\right\|^2 + 2\mathbb{E}_t\|e_t\|^2 \tag{76}$$

$$\leq \frac{4E}{K}R^2\overline{\pi}^2\sigma_M^2 + \frac{4}{K^2}\mathbb{E}_t\left[\left\|\sum_{k=1}^{K}\sum_{\tau=0}^{E-1}\nabla F^k(w_{t,\tau}^k) - Ke_t\right\|^2\right] + 6\mathbb{E}_t\|e_t\|^2 \tag{77}$$

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 $\mathbb{E}[\|\boldsymbol{x}\|^2] = \mathbb{E}[\|\boldsymbol{x} - \mathbb{E}\boldsymbol{x}\|^2] + \|\mathbb{E}\boldsymbol{x}\|^2$, and (a_8) is due to Assumption (3).

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

$$\mathbb{E}_t[f(w_{t+1})] \tag{78}$$

$$\leq f(w_t) - \eta\eta_L E\|\nabla f(w_t)\|^2 + \eta\underbrace{\langle\nabla f(w_t), \mathbb{E}[\eta_L\Delta_t + \eta_L E\nabla f(w_t)]\rangle}_{A_1} + \frac{L}{2}\eta^2\eta_L^2\underbrace{\mathbb{E}_t[\|\Delta_t\|^2]}_{A_2} \tag{79}$$

$$\leq f(w_t) - \eta\eta_L E\|\nabla f(w_t)\|^2 + \eta\eta_L E(\frac{1}{2} + 30(A^2+1)\eta_L^2E^2L^2)\|\nabla f(w_t)\|^2 + 5\eta\eta_L^3E^2L^2\left(R^2\lambda_t^2\sigma_M^2 + 12ER^2\overline{\pi}^2\sigma_M^2 + 6E\sigma_G^2\right) - \frac{\eta\eta_L}{2EK^2}\mathbb{E}_t\left\|\sum_{k=1}^{K}\sum_{\tau=0}^{E-1}\nabla F^k(w_{t,\tau}^k) - Ke_t\right\|^2 + \frac{\eta\eta_L\mathbb{E}_t\|e_t\|^2}{E} + \frac{2EL\eta^2\eta_L^2}{K}R^2\overline{\pi}^2\sigma_M^2 + \frac{2L\eta^2\eta_L^2}{K^2}\mathbb{E}_t\left[\left\|\sum_{k=1}^{K}\sum_{\tau=0}^{E-1}\nabla F^k(w_{t,\tau}^k) - Ke_t\right\|^2\right] + 3\eta^2\eta_L^2L\mathbb{E}_t\|e_t\|^2 \tag{80}$$

$$= f(w_t) - \eta\eta_L E(\frac{1}{2} - 30(A^2+1)\eta_L^2E^2L^2)\|\nabla f(w_t)\|^2 + 5\eta\eta_L^3E^2L^2R^2\lambda_t^2\sigma_M^2 + 60\eta\eta_L^3E^3L^2R^2\overline{\pi}^2\sigma_M^2 + 30\eta\eta_L^3E^3L^2\sigma_G^2 + \left(\frac{\eta\eta_L}{E} + 3\eta^2\eta_L^2L\right)\mathbb{E}_t\left\|\frac{1}{K}\sum_{k=1}^{K}\sum_{\tau=0}^{E-1}\left(\hat{g}_{t,\tau}^k - g_{t,\tau}^k\right)\right\|^2 - \left(\frac{\eta\eta_L}{2EK^2} - \frac{2L\eta^2\eta_L^2}{K^2}\right)\mathbb{E}_t\left\|\sum_{k=1}^{K}\sum_{\tau=0}^{E-1}\nabla F^k(w_{t,\tau}^k) - Ke_t\right\|^2 \tag{81}$$

$$\overset{(a_9)}{\leq} f(w_t) - c\eta\eta_L E\|\nabla f(w_t)\|^2 + 60\eta\eta_L^3E^3L^2R^2\overline{\pi}^2\sigma_M^2 + 30\eta\eta_L^3E^3L^2\sigma_G^2 + \left(5\eta\eta_L^3E^2L^2 + \eta\eta_L E + 3\eta^2\eta_L^2LE^2\right)R^2\lambda_t^2\sigma_M^2 \tag{82}$$

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 exits a constant $c > 0$ satisfying $(\frac{1}{2} - 30(A^2+1)\eta_L^2E^2L^2) > c > 0$ if $\eta_L < \frac{1}{\sqrt{60(A^2+1)}EL}$.

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

$$\sum_{t=0}^{T-1}c\eta\eta_L E\mathbb{E}\|\nabla f(w_t)\|^2 \tag{83}$$

$$\leq f(w_0) - f(w_T) + E\eta\eta_L\sum_{t=0}^{T-1}60\eta_L^2E^2L^2R^2\overline{\pi}^2\sigma_M^2 + E\eta\eta_L\sum_{t=0}^{T-1}30\eta_L^2E^2L^2\sigma_G^2 + E\eta\eta_L\sum_{t=0}^{T-1}\left(5\eta_L^2EL^2 + 3\eta\eta_LLE + 1\right)R^2\lambda_t^2\sigma_M^2 \tag{84}$$

which implies,

$$\min_{t=0,...,T-1}\mathbb{E}\|\nabla f(w_t)\|^2 \leq \frac{f_0 - f_*}{c\eta\eta_L ET} + \Phi_G + \Phi_M + \Phi_L \tag{85}$$

where

$$\Phi_G = \frac{30E^2\eta_L^2 L^2}{c}\sigma_G^2 \tag{86}$$

$$\Phi_M = \frac{60\eta_L^2 E^2 L^2 R^2 \overline{\pi}^2}{c}\sigma_M^2 \tag{87}$$

$$\Phi_L = \frac{\left(5\eta_L^2 EL^2 + 3\eta\eta_L LE + 1\right) R^2\sigma_M^2}{cT}\sum_{t=0}^{T-1}\lambda_t^2 \tag{88}$$

This completes the proof.

Appendix H
Experiments

We provide the full details of the experiments and present supplementary results derived from the Cifar100 dataset.

A. Experiments Details

FMNIST: For the FMNIST datasets, we transform the data into tensors and normalize them with a mean of 0.1307 and a standard deviation of 0.3081. We employ the LeNet architecture [32] as our primary model. This architecture consists of two convolutional layers, each followed by ReLU activations and max-pooling. Subsequently, there are three dense layers, with the first two incorporating ReLU activations and the last one producing outputs for 10 classes. The training parameters are set as follows: cache size $B = 300$, streaming data size $B_s = [30, 150]$, number of classes per client $C = 3$, local iteration count $E = 2$, local learning rate $\eta_L = 0.1$, and global learning rate $\eta = 1$.

NTC: For the NTC datasets, we transform the data into tensors and reshape the feature into a 3-dimensional array with dimensions $1 \times 39 \times 39$. We also employ the LeNet architecture as our model. The training parameters are set as follows: cache size $B = 300$, streaming data size $B_s = [30, 150]$, number of classes per client $C = 3$, local iteration count $E = 2$, local learning rate $\eta_L = 0.01$, and global learning rate $\eta = 1$.

CIFAR-100: For the CIFAR-100 datasets, we use 20 superclasses to reclassify the data samples in CIFAR-100. We employ the MobileNetV2 architecture [33] as our model. This architecture consists of an initial convolutional layer followed by seven stages of LinearBottleNeck blocks, with each stage potentially containing multiple such blocks. After these stages, there are two additional convolutional layers, with the final layer outputting 20 classes. The model employs ReLU6 activations and adaptive average pooling before the final output. The training parameters are set as follows: cache size $B = 500$, streaming data size $B_s = 250$, number of classes per client $C = 5$, parameter $\theta = \frac{2}{3}$, local iteration count $E = 2$, local learning rate $\eta_L = 0.01$, and global learning rate $\eta = 1$.

Considerations for choosing local cache size: We use the network traffic classification task as an example. The data samples of the network traffic classification dataset (NTC) are extracted from ISCXVPN2016, which is a representative dataset of real-world traffic generated at ISCX. The local cache size is set to 300 data samples for the majority of simulation settings. This size results in about 1.2MB of local storage space being used for each client. The selection of 300 samples aims to balance two considerations: it fits within the limited storage capacities mentioned in the previous comment, while also being large enough to maintain adequate learning performance.

B. Additional Experiment on the CIFAR-100 datasets

In our extended experiments on the CIFAR-100 datasets, we initially examine the label distribution among clients, particularly for the case where $C = 5$. Fig. 9 illustrates the long-term label distribution on clients. The performance comparison of the proposed update rules and benchmarks with full participation for CIFAR-100 datasets is shown in Fig. 10, which is consistent with the conclusions we have drawn from other datasets.

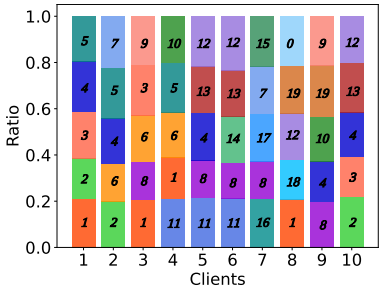

Fig. 9: Underlying Label Distribution on Clients ($C = 5$).

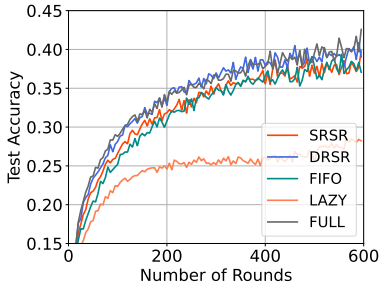

Fig. 10: Performance comparison of the proposed update rules and benchmarks with full participation (CIFAR-100).