A. Experimental Setup

Implementation details We adopt PyTorch [31] to implement FedACG and the other baselines. We follow the evaluation protocol of [1] and [45]. For local updates, we use the SGD optimizer with a learning rate of 0.1 for all approaches on the three benchmarks. We apply no momentum to the local SGD, but incorporate the weight decay of 0.001 to prevent overfitting. We also employ gradient clipping to increase the stability of the algorithms.

For the experiments on CIFAR-10 and CIFAR-100, we choose 5 as the number of local training epochs (50 iterations). We set the batch size of the local update to 50 and 10 for the 100 and 500 client participation, respectively. The learning rate decay parameter of each algorithm is selected from $\{0.995, 0.998, 1\}$ to achieve the best performance. The global learning rate is set to 1, except for FedAdam, which is set to 0.01.

For the experiments on Tiny-ImageNet, we match the total local iterations of local updates with other benchmarks by setting the batch size of local updates as 100 and 20 for the 100 and 500 client participation, respectively.

Hyperparameter selection To reproduce other compared algorithms, we primarily follow the configurations outlined in the original papers, adjusting the parameters only when it leads to improved performance. Specifically, α is chosen from $\{0.1, 0.3, 0.5\}$ in FedCM, $\{0.001, 0.01, 0.1\}$ in FedDyn, and is set to 0.01 in FedDC. τ in FedADAM is fixed at 0.001, while μ in MOON is set to 1. For β, in FedAvgM, choices are from $\{0.4, 0.6, 0.8\}$; in FedProx and FedACG, from $\{0.1, 0.01, 0.001\}$. In FedProx, FedNTD, and FedDecorr, β is set to 0.001, 0.3, and 0.01, respectively. Finally, λ in FedACG is selected from $\{0.8, 0.85, 0.9\}$.

B. Additional Experiments

B.1. Additional analysis for the effect of accelerated client gradient

FedACG uses a lookahead model, $\theta^{t-1} + \lambda m^{t-1}$, to start local training. This helps clients match their local solutions with the global loss, ensuring consistent updates. We observe more empirical evidence that supports our claim.

Figure A shows the convergence curves of FedACG and FedAvgM on CIFAR-10 in the moderate-scale setting without smoothing. For the experiments, we set the momentum coefficient to 0.85 for both algorithms. We observe that FedACG consistently outperforms FedAvgM and has a smaller accuracy variation throughout the training procedure. Specifically, when we compute the average squared difference between the accuracy at time step t without smoothing (Acc^t) and the accuracy given by the simple moving average ($\text{Acc}^t_{\text{SMA}}$) over 1,000 rounds of communication, *i.e.*, $\frac{1}{T}\sum_{t=0}^{T-1}(\text{Acc}^t - \text{Acc}^t_{\text{SMA}})^2$, the differences are 2.26 and 10.30 for FedACG and FedAvgM, respectively. We believe that this is partly because the proposed accelerated gradient allows each client's update to compensate for the potential noise in momentum, which is possible because the local updates start from the anticipated point, $\theta^{t-1} + \lambda m^{t-1}$.

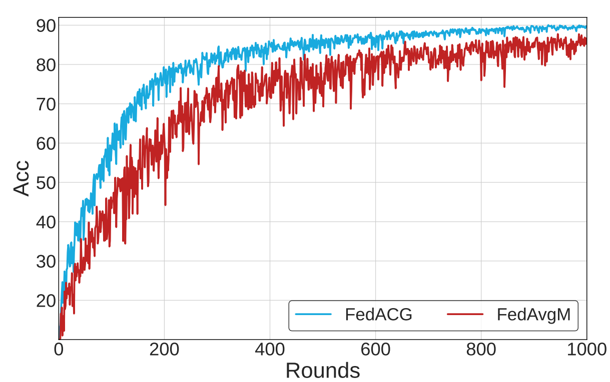


Figure A. Training curves of FedACG and FedAvgM on CIFAR-10 in a moderate-scale setting without smoothing.

B.2. FedACG with other local objectives

In Table A, we incorporate accelerated client gradient into a client-side optimization technique, FedMLB [19], FedLC [47], and FedDecorr [36] to test its benefits. "+ACG" means adopting the proposed accelerated client gradient. It shows that the momentum-integrated initialization helps client-side optimization approaches achieve significant improvements without any additional communication costs.

Table A. Results of incorporating accelerated client gradient (ACG) into client-side optimization techniques on CIFAR-100 and Tiny-ImageNet under non-*i.i.d.* settings.

(a) 100 clients, 5% participation, Dirichlet (0.3)

Method	CIFAR-100				Tiny-ImageNet			
	Acc. (%, ↑)		Rounds (↓)		Acc. (%, ↑)		Rounds (↓)	
	500R	1000R	47%	55%	500R	1000R	35%	38%
FedMLB [19]	47.39	54.58	488	1000+	37.20	40.16	414	539
FedMLB + ACG	**61.32**	**65.67**	**216**	**316**	**46.11**	**50.54**	**205**	**260**
FedLC [47]	42.74	47.23	980	1000+	35.03	35.95	500	1000+
FedLC + ACG	**57.18**	**62.09**	**239**	**420**	**43.43**	**44.57**	**187**	**268**
FedDecorr [36]	43.52	49.17	767	1000+	33.40	34.86	1000+	1000+
FedDecorr + ACG	**57.95**	**63.02**	**218**	**380**	**43.09**	**44.52**	**241**	**304**

(b) 500 clients, 2% participation, Dirichlet (0.3)

Method	CIFAR-100				Tiny-ImageNet			
	Acc. (%, ↑)		Rounds (↓)		Acc. (%, ↑)		Rounds (↓)	
	500R	1000R	36%	40%	500R	1000R	24%	30%
FedMLB [19]	32.30	42.61	642	800	28.39	33.67	384	489
FedMLB + ACG	**41.10**	**55.27**	**402**	**479**	**35.92**	**43.85**	**209**	**313**
FedLC [47]	29.58	36.78	936	1000+	22.14	26.83	676	1000+
FedLC + ACG	**35.87**	**43.51**	**503**	**675**	**29.13**	**33.17**	**263**	**557**
FedDecorr [36]	30.56	38.20	850	1000+	24.34	30.28	499	959
FedDecorr + ACG	**41.18**	**49.93**	**367**	**473**	**29.24**	**34.71**	**290**	**540**

B.3. Evaluation on Various Data Heterogeneity

Tables B and C show that FedACG also matches or outperforms the performance of competitive methods when data heterogeneity is not severe (Dirichlet 0.6) or very low (*i.i.d.*) on CIFAR-10 and CIFAR-100 in most cases.

Table B. Results with Dirichlet (0.6) data on CIFAR-10 and CIFAR-100 for two different settings.

(a) Dirichlet (0.6), 100 clients, 5% participation

Method	CIFAR-10				CIFAR-100			
	Acc. (%, ↑)		Rounds (↓)		Acc. (%, ↑)		Rounds (↓)	
	500R	1000R	81%	87%	500R	1000R	50%	56%
FedAvg [27]	80.56	85.97	520	1000+	43.91	49.18	1000+	1000+
FedProx [25]	80.39	85.53	524	1000+	43.15	48.45	1000+	1000+
FedAvgM [14]	84.65	87.96	355	811	46.66	52.49	735	1000+
FedADAM [33]	80.25	83.52	526	1000+	45.95	51.63	778	1000+
FedDyn [1]	87.23	89.49	310	487	50.51	56.78	488	886
MOON [23]	84.95	87.99	272	728	55.76	61.42	338	527
FedCM† [45]	82.84	86.64	385	1000+	53.75	60.48	331	468
FedMLB [19]	79.85	85.98	574	1000+	49.31	56.70	526	925
FedLC [47]	80.40	85.48	559	1000+	43.99	48.92	1000+	1000+
FedNTD [22]	81.2	86.44	498	1000+	44.26	50.34	916	1000+
FedDC‡ [10]	**88.05**	89.58	270	**437**	56.00	60.58	347	491
FedDecorr [36]	81.01	85.19	500	1000+	43.64	49.03	1000+	1000+
FedACG (ours)	87.57	**90.56**	**218**	453	**58.82**	**63.88**	**243**	**396**

(b) Dirichlet (0.6), 500 clients, 2% participation

Method	CIFAR-10				CIFAR-100			
	Acc. (%, ↑)		Rounds (↓)		Acc. (%, ↑)		Rounds (↓)	
	500R	1000R	69%	80%	500R	1000R	32%	41%
FedAvg [27]	62.79	75.17	671	1000+	29.41	36.62	648	1000+
FedProx [25]	62.48	75.10	688	1000+	29.62	36.70	647	1000+
FedAvgM [14]	69.10	80.26	498	981	32.78	41.93	468	942
FedADAM [33]	68.48	78.92	535	1000+	37.57	48.29	341	624
FedDyn [1]	68.53	80.33	513	983	32.06	43.28	498	917
MOON [23]	74.29	80.66	368	921	31.64	41.61	515	931
FedCM† [45]	71.42	78.94	429	1000+	26.82	39.78	714	1000+
FedMLB [19]	62.60	74.36	729	1000+	33.79	43.52	432	831
FedLC [47]	62.77	73.56	694	1000+	30.07	36.97	620	1000+
FedNTD [22]	61.9	74.38	717	1000+	28.85	35.88	691	1000+
FedDC‡ [10]	77.74	**86.32**	324	596	34.24	44.69	444	825
FedDecorr [36]	63.63	74.89	658	1000+	29.99	37.72	615	1000+
FedACG (ours)	**78.49**	85.28	**289**	**565**	**39.61**	**49.70**	**304**	**540**

Table C. Results with *i.i.d.* data on CIFAR-10 and CIFAR-100 for two different settings.

(a) *i.i.d.*, 100 clients, 5% participation

Method	CIFAR-10				CIFAR-100			
	Acc. (%, ↑)		Rounds (↓)		Acc. (%, ↑)		Rounds (↓)	
	500R	1000R	82%	89%	500R	1000R	52%	58%
FedAvg [27]	85.28	88.69	372	1000+	43.96	48.20	1000+	1000+
FedProx [25]	84.79	87.99	384	1000+	43.57	47.75	1000+	1000+
FedAvgM [14]	87.67	89.96	258	375	47.43	52.83	880	1000+
FedADAM [33]	85.29	87.97	286	1000+	52.23	57.73	496	1000+
FedDyn [1]	89.19	90.70	269	492	50.37	56.88	592	898
MOON [23]	88.24	89.96	207	628	58.50	**64.73**	311	484
FedCM† [45]	87.38	89.65	182	782	57.10	62.48	266	466
FedMLB [19]	86.32	89.65	359	784	50.12	56.40	586	1000+
FedLC [47]	84.48	88.26	393	1000+	43.84	46.70	1000+	1000+
FedNTD [22]	85.68	89.43	367	870	44.93	50.51	1000+	1000+
FedDC‡ [10]	90.07	90.80	194	425	55.17	61.00	400	633
FedDecorr [36]	85.21	88.17	364	1000+	45.16	49.16	1000+	1000+
FedACG (ours)	**90.57**	**92.29**	**157**	**354**	**59.82**	64.08	**244**	**342**

(b) *i.i.d.*, 500 clients, 2% participation

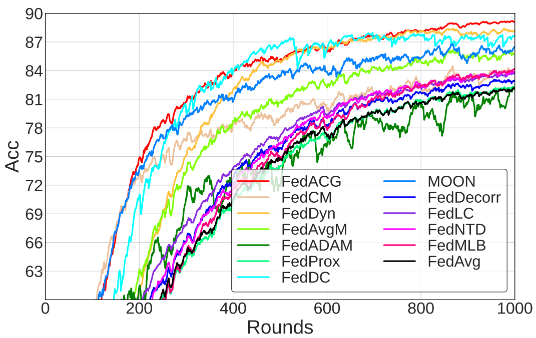
Method	CIFAR-10				CIFAR-100			
	Acc. (%, ↑)		Rounds (↓)		Acc. (%, ↑)		Rounds (↓)	
	500R	1000R	75%	83%	500R	1000R	33%	42%
FedAvg [27]	68.70	78.21	652	1000+	30.71	37.85	664	1000+
FedProx [25]	68.74	77.96	643	1000+	30.11	37.13	685	1000+
FedAvgM [14]	74.34	83.64	523	943	33.54	42.55	479	971
FedADAM [33]	75.32	84.01	491	915	38.74	48.94	328	636
FedDyn [1]	74.81	84.71	398	823	33.20	42.91	492	936
MOON [23]	69.86	81.89	586	1000+	28.82	41.26	649	1000+
FedCM† [45]	77.84	83.26	491	959	29.59	42.04	653	991
FedMLB [19]	62.60	80.16	729	1000+	34.56	44.95	440	817
FedLC [47]	68.92	79.09	727	1000+	29.91	37.18	677	1000+
FedNTD [22]	68.61	80.65	706	1000+	30.04	36.63	706	1000+
FedDC‡ [10]	**80.87**	**87.53**	358	**574**	33.93	45.80	476	817
FedDecorr [36]	68.12	77.39	802	1000+	30.41	37.53	585	1000+
FedACG (ours)	80.15	87.47	**316**	578	**41.16**	**49.10**	**299**	**525**

C. Convergence Plot

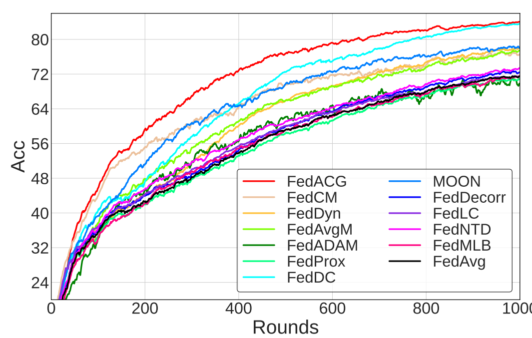
C.1. Evaluation on various federated learning scenarios

Figure B to Figure D show the convergence of FedACG and the compared algorithms on CIFAR-10, CIFAR-100, and Tiny-ImageNet for various federated learning settings: varying the number of total clients, participation rates, data heterogeneity. FedACG continuously matches or exceeds the performance of the most powerful of our competitors in most learning sections.

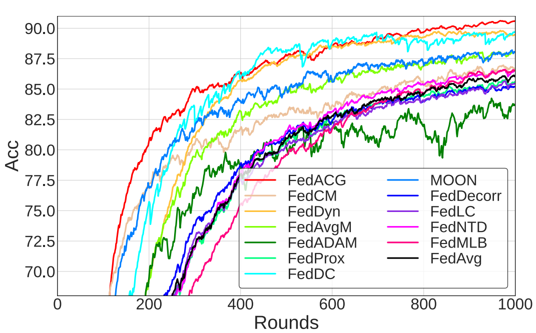
Figure E shows the convergence plots under massive clients with lower participation rates. The result shows that FedACG takes the lead in most learning sections, which also demonstrates the effectiveness of FedACG.



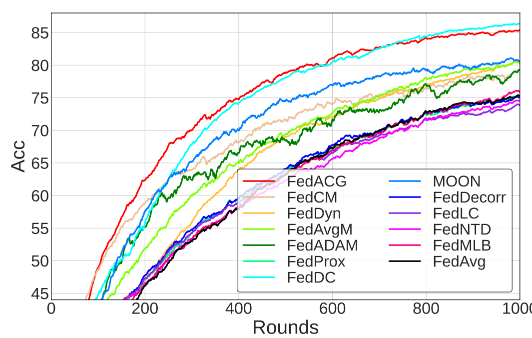
(a) Dirichlet (0.3), 100 clients, 5% participation



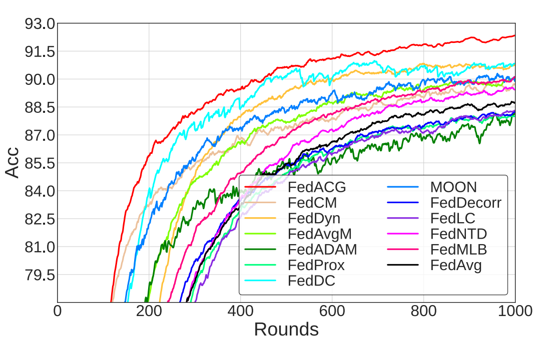
(b) Dirichlet (0.3), 500 clients, 2% participation



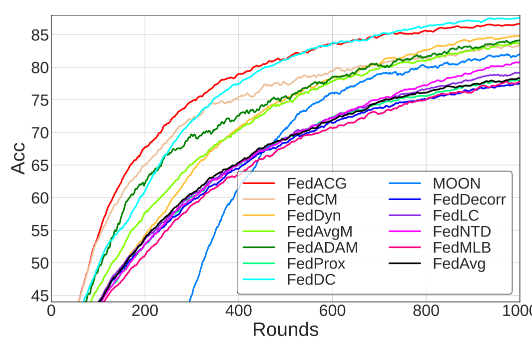
(c) Dirichlet (0.6), 100 clients, 5% participation



(d) Dirichlet (0.6), 500 clients, 2% participation

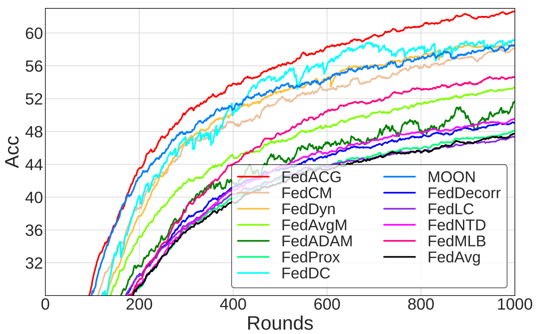


(e) *i.i.d.*, 100 clients, 5% participation

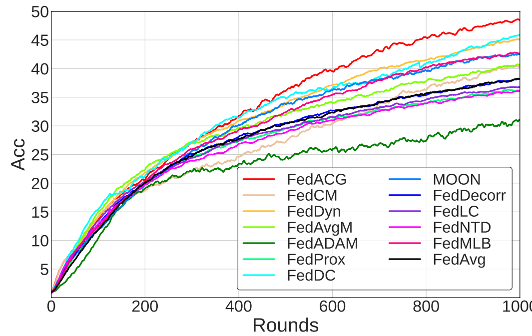


(f) *i.i.d.*, 500 clients, 2% participation

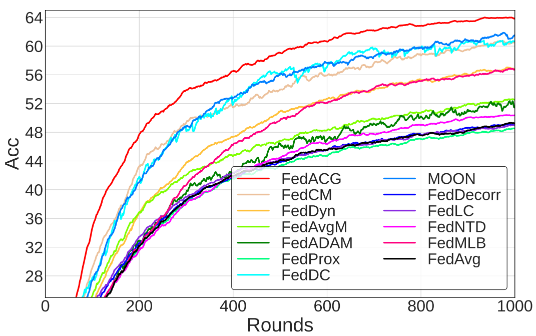
Figure B. The convergence plots of FedACG and the baselines on CIFAR-10 with different federated learning scenarios.



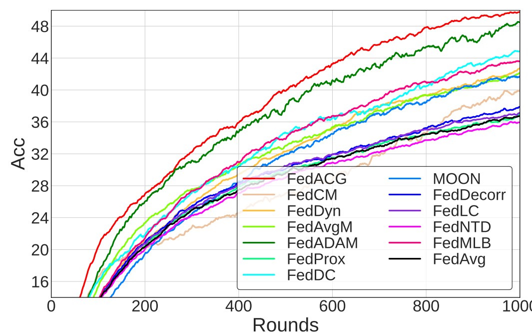
(a) Dirichlet (0.3), 100 clients, 5% participation



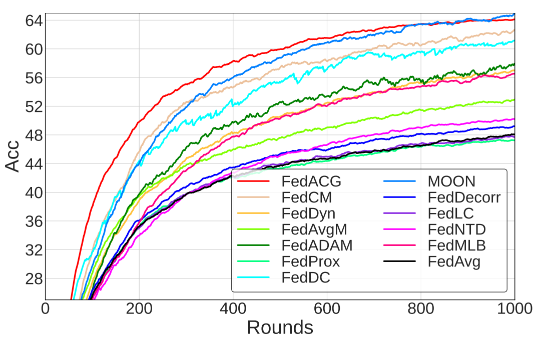
(b) Dirichlet (0.3), 500 clients, 2% participation



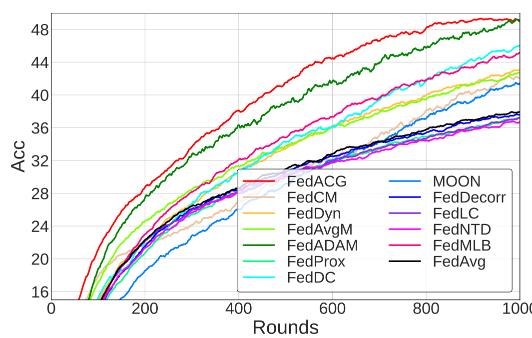
(c) Dirichlet (0.6), 100 clients, 5% participation



(d) Dirichlet (0.6), 500 clients, 2% participation

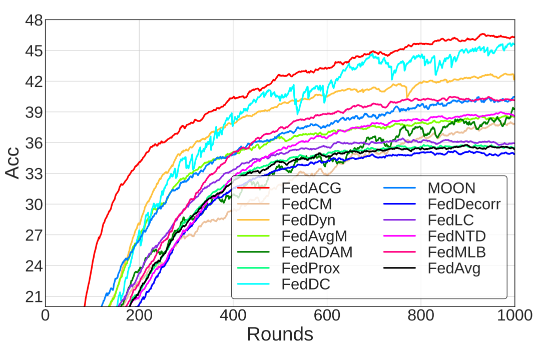


(e) *i.i.d.*, 100 clients, 5% participation

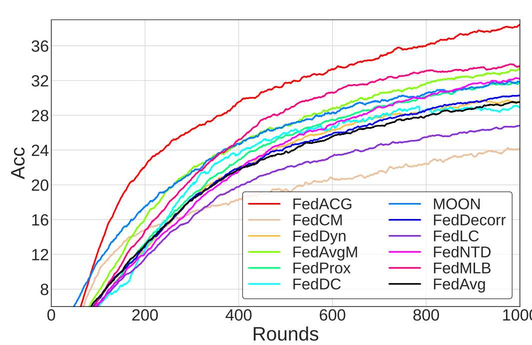


(f) *i.i.d.*, 500 clients, 2% participation

Figure C. The convergence plots of FedACG and the baselines on CIFAR-100 with different federated learning scenarios.

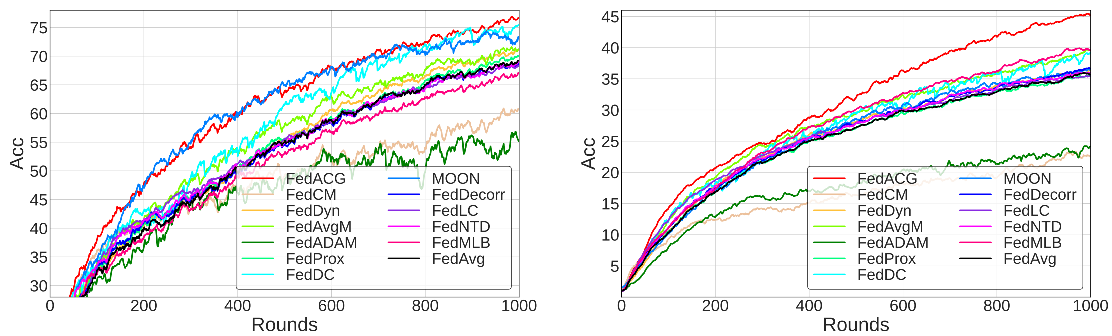


(a) Dirichlet (0.3), 100 clients, 5% participation



(b) Dirichlet (0.3), 500 clients, 2% participation

Figure D. The convergence plots of FedACG and the baselines on Tiny-ImageNet with different federated learning scenarios.



(a) CIFAR-10 (b) CIFAR-100

Figure E. The convergence plots of FedACG and the baselines when the participation rate is low (1%) for 500 clients on CIFAR-10 and CIFAR-100. The Dirichlet parameter is commonly set to 0.3 for the experiments.

C.2. Evaluation on dynamic client set

Figure F shows a convergence plot when the entire client's pool changes during training. The result shows that FedACG outperforms the baselines in most learning sections. Note that FedDyn shows worse performance than FedACG in the overall section of learning. This is partly because it needs to store local states for local training in each client, which requires a kind of warm-up period for newly participating clients to contain useful information. In contrast, FedACG, which is free from these restrictions, shows strength in a realistic federated learning scenario where the pool of entire clients changes during training.

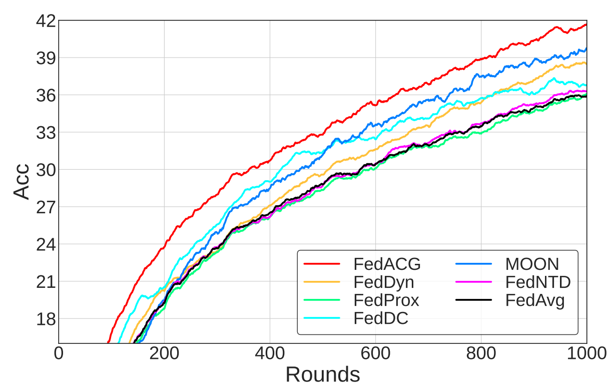


Figure F. The convergence plots of FedACG and other compared methods on CIFAR-100 when the client set changes over dynamically: we sample 250 clients out of 500 clients as a candidate clients set at every 100 rounds over 10 stages on Dirichlet (0.3) split. 10 clients out of the sampled client set participate for the local training for each communication round.

D. Convergence of FedACG

We now present the theoretical convergence result of FedACG. We first state a few assumptions for the local loss functions $\mathcal{F}_i(\cdot)$, which are commonly used in several previous works on federated optimization [15, 33, 45]. First, the local function $\mathcal{F}_i(\cdot)$ is assumed to be L-smooth for all $C_i \in \{C_1, \dots, C_N\}$, *i.e.*,

$$\|\nabla \mathcal{F}_i(x) - \nabla \mathcal{F}_i(y)\| \leq L\|x - y\| \;\; \forall x, y. \tag{2}$$

This also implies

$$\mathcal{F}_i(y) \leq \mathcal{F}_i(x) + \langle \nabla \mathcal{F}_i(x), y - x\rangle + \frac{L}{2}\|y - x\|^2. \tag{3}$$

Second, we assume the stochastic gradient of the local loss function $\nabla f_i(x) := \nabla \mathcal{F}_i(x; \mathcal{D}_i)$ is unbiased and possesses a bounded variance, *i.e.* $\mathbb{E}_{\mathcal{D}_i}[\|\nabla f_i(x) - \nabla \mathcal{F}_i(x)\|^2] \leq \sigma^2$. Third, we assume the average norm of local gradients is bounded by a function of the global gradient magnitude as $\frac{1}{N}\sum_{i=1}^{N}\|\nabla\mathcal{F}_i(x)\|^2 \leq \sigma_g^2 + B^2\|\nabla\mathcal{F}(x)\|^2$, where $\sigma_g \geq 0$ and $B \geq 1$. Based on the above assumptions, we derive the following asymptotic convergence bound of FedACG.

D.1. Preliminary Lemmas

We present several technical lemmas that are useful for subsequent proofs.

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

Lemma 2 (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 $w_r = 1$ such that,*

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

Proof. Unrolling the sum, we can simplify

$$\Psi_R \leq \frac{d_0}{\eta(R+1)} + c_1\eta + c_2\eta^2\,.$$

The lemma can be established through the adjustment of η. We consider the following two cases based on the magnitudes of R and $\eta_{\max}$:

- When $R+1 \leq \frac{d_0}{c_1\eta_{\max}^2}$ and $R+1 \leq \frac{d_0}{c_2\eta_{\max}^3}$, selecting $\eta = \eta_{\max}$ satisfies

$$\Psi_R \leq \frac{d_0}{\eta_{\max}(R+1)} + c_1\eta_{\max} + c_2\eta_{\max}^2 \leq \frac{d_0}{\eta_{\max}(R+1)} + \frac{\sqrt{c_1 d_0}}{\sqrt{R+1}} + \left(\frac{d_0}{R+1}\right)^{\frac{2}{3}} c_2^{\frac{1}{3}}\,.$$

- In the other case, we have $\eta_{\max}^2 \geq \frac{d_0}{c_1(R+1)}$ or $\eta_{\max}^3 \geq \frac{d_0}{c_2(R+1)}$. Choosing $\eta = \min\left\{\sqrt{\frac{d_0}{c_1(R+1)}}, \sqrt[3]{\frac{d_0}{c_2(R+1)}}\right\}$ satisfies

$$\Psi_R \leq \frac{d_0}{\eta(R+1)} + c\eta = \frac{2\sqrt{c_1 d_0}}{\sqrt{R+1}} + 2\sqrt[3]{\frac{d_0^2 c_2}{(R+1)^2}}\,.$$

□

Lemma 3 (separating mean and variance)**.** *Given a set of τ random variables $\{\mathbf{x}_1, \ldots, \mathbf{x}_\tau\}$ in $\mathbb{R}^d$, where $\mathbb{E}[\mathbf{x}_i|\mathbf{x}_{i-1}, \ldots \mathbf{x}_1] = \xi_i$ and $\mathbb{E}[\|\mathbf{x}_i - \xi_i\|^2] \leq \sigma^2$ represent their conditional mean and variance, respectively, the variables $\{\mathbf{x}_i - \xi_i\}$ form a martingale difference sequence. Based on this setup, the following holds*

$$\mathbb{E}[\|\sum_{i=1}^{\tau}\mathbf{x}_i\|^2] \leq 2\|\sum_{i=1}^{\tau}\xi_i\|^2 + 2\tau\sigma^2\,.$$

Proof.

$$\begin{aligned}\mathbb{E}[\|\sum_{i=1}^{\tau}\mathbf{x}_i\|^2] &\leq 2\|\sum_{i=1}^{\tau}\xi_i\|^2 + 2\mathbb{E}[\|\sum_{i=1}^{\tau}\mathbf{x}_i - \xi_i\|^2] \\ &= 2\|\sum_{i=1}^{\tau}\xi_i\|^2 + 2\sum_{i}\mathbb{E}[\|\mathbf{x}_i - \xi_i\|^2] \\ &\leq 2\|\sum_{i=1}^{\tau}\xi_i\|^2 + 2\tau\sigma^2\,. \end{aligned} \tag{4}$$

The first inequality comes from the relaxed triangle inequality and the following equality holds because $\{\mathbf{x}_i - \xi_i\}$ forms a martingale difference sequence. □

D.2. Convergence of FedACG for non-convex functions

Theorem 1. ***(Convergence for non-convex functions)*** *Suppose that local functions $\{\mathcal{F}_i\}_{i=1}^N$ are non-convex and L-smooth. By setting $\eta \leq \frac{(1-\lambda)^2}{64KL(B^2+1)}$, FedACG satisfies*

$$\min_{t=1,\ldots,T} \mathbb{E}\left\|\nabla\mathcal{F}\left(\theta^{t-1}+\lambda m^{t-1}\right)\right\|^2 \leq \mathcal{O}\left(\frac{M_1\sqrt{LD}}{\sqrt{TK|S_t|}} + \frac{\left(LD(1-\lambda)^2\right)^{\frac{2}{3}} M_2^{\frac{1}{3}}}{(T+1)^{\frac{2}{3}}} + \frac{B^2LD}{T}\right),$$

where $M_1^2 := \sigma^2 + K\left(1-\frac{|S_t|}{N}\right)\sigma_g^2$, $M_2 := \frac{\sigma^2}{K}+\sigma_g^2$, and $D := \frac{\mathcal{F}(\theta^0)-\mathcal{F}(\theta^*)}{1-\lambda}$.

Proof. Let $z^t = \theta^t + \frac{\lambda}{1-\lambda}m^t$ and $\Phi^t = \theta^t + \lambda m^t$. Note that $z^0 = \theta^0$ and $z^t - z^{t-1} = \frac{1}{1-\lambda}\Delta^t$. By the smoothness of the function $\mathcal{F}(\mathbf{x})$, we have

$$\mathcal{F}(z^t) \leq \mathcal{F}(z^{t-1}) + \langle\nabla\mathcal{F}(z^{t-1}), z^t - z^{t-1}\rangle + \frac{L}{2}\|z^t - z^{t-1}\|^2.$$

By taking the expectation on both sides, we have

$$\begin{aligned}\mathbb{E}[\mathcal{F}(z^t)] &\leq \mathbb{E}[\mathcal{F}(z^{t-1})] + \frac{1}{1-\lambda}\mathbb{E}[\langle\nabla\mathcal{F}(z^{t-1}),\Delta^t\rangle] + \frac{L}{2}\mathbb{E}[\|z^t-z^{t-1}\|^2]\\ &= \mathbb{E}[\mathcal{F}(z^{t-1})] + \frac{1}{1-\lambda}\mathbb{E}[\langle\nabla\mathcal{F}(z^{t-1})-\nabla\mathcal{F}(\Phi^{t-1}),\Delta^t\rangle] + \frac{1}{1-\lambda}\mathbb{E}[\langle\nabla\mathcal{F}(\Phi^{t-1}),\Delta^t\rangle] + \frac{L}{2(1-\lambda)^2}\mathbb{E}[\|\Delta^t\|^2].\end{aligned} \tag{5}$$

We note that

$$\begin{aligned}\frac{1}{1-\lambda}\mathbb{E}[\langle\nabla\mathcal{F}(z^{t-1})-\nabla\mathcal{F}(\Phi^{t-1}),\Delta^t\rangle] &\leq \frac{1-\lambda}{2\lambda^3L}\mathbb{E}[\|\nabla\mathcal{F}(z^{t-1})-\nabla\mathcal{F}(\Phi^{t-1})\|^2] + \frac{\lambda^3L}{2(1-\lambda)^3}\mathbb{E}[\|\Delta^t\|^2]\\ &\leq \frac{(1-\lambda)L}{2\lambda^3}\mathbb{E}[\|z^{t-1}-\Phi^{t-1}\|^2] + \frac{\lambda^3L}{2(1-\lambda)^3}\mathbb{E}[\|\Delta^t\|^2]\\ &\leq \frac{L}{2(1-\lambda)}\mathbb{E}[\|m^{t-1}\|^2] + \frac{L}{2(1-\lambda)^3}\mathbb{E}[\|\Delta^t\|^2],\end{aligned} \tag{6}$$

where the first inequality holds because $\langle a,b\rangle \leq \frac{1}{2}(\|a\|^2+\|b\|^2)$, while the second inequality follows from the L-smoothness. The third inequality follows because $z^t - \Phi^t = \frac{\lambda^2}{1-\lambda}m^t$ and $0\leq\lambda<1$.

We also note that

$$\begin{aligned}\frac{1}{1-\lambda}\mathbb{E}[\langle\nabla\mathcal{F}(\Phi^{t-1}),\Delta^t\rangle] &= \frac{1}{1-\lambda}\mathbb{E}[\langle\nabla\mathcal{F}(\Phi^{t-1}), \frac{-\eta K}{KN}\sum_{k,C_i}\nabla\mathcal{F}_i(\theta^t_{i,k-1})\rangle]\\ &\leq \frac{\eta K}{2(1-\lambda)}\left(\mathbb{E}[\|\nabla\mathcal{F}(\Phi^{t-1}) - \frac{1}{KN}\sum_{k,C_i}\nabla\mathcal{F}_i(\theta^t_{i,k-1})\|^2] - \mathbb{E}[\|\nabla\mathcal{F}(\Phi^{t-1})\|^2]\right)\\ &\leq \frac{\eta K}{2(1-\lambda)}\left(\frac{L^2}{KN}\sum_{k,C_i}\mathbb{E}[\|\theta^t_{i,k-1}-\theta^t_{i,0}\|^2] - \mathbb{E}[\|\nabla\mathcal{F}(\Phi^{t-1})\|^2]\right),\end{aligned} \tag{7}$$

where the first inequality holds because $\langle a,b\rangle \leq \frac{1}{2}\|a+b\|^2 - \frac{1}{2}\|a\|^2$.

Substituting Eq. (6) and Eq. (7) into Eq. (5) yields

$$\begin{aligned}\mathbb{E}[\mathcal{F}(z^t)] \leq \mathbb{E}[\mathcal{F}(z^{t-1})] &+ \frac{\eta K}{2(1-\lambda)}\left(\frac{L^2}{KN}\sum_{k,C_i}\mathbb{E}[\|\theta^t_{i,k-1}-\theta^t_{i,0}\|^2] - \mathbb{E}[\|\nabla\mathcal{F}(\Phi^{t-1})\|^2]\right)\\ &+ \frac{L}{2(1-\lambda)}\mathbb{E}[\|m^{t-1}\|^2] + (\frac{L}{2(1-\lambda)^3}+\frac{L}{2(1-\lambda)^2})\mathbb{E}[\|\Delta^t\|^2].\end{aligned}$$

By rearranging the inequality above, we have

$$\frac{\eta K}{2(1-\lambda)}\mathbb{E}[\|\nabla\mathcal{F}(\Phi^{t-1})\|^2] \leq (\mathbb{E}[\mathcal{F}(z^{t-1})] - \mathbb{E}[\mathcal{F}(z^t)]) + \frac{\eta K L^2}{2(1-\lambda)}\frac{1}{KN}\sum_{k,C_i}\mathbb{E}[\|\theta_{i,k-1}^t - \theta_{i,0}^t\|^2]$$
$$+ \frac{L}{2(1-\lambda)}\mathbb{E}[\|m^{t-1}\|^2] + (\frac{L}{2(1-\lambda)^3} + \frac{L}{2(1-\lambda)^2})\mathbb{E}[\|\Delta^t\|^2].$$

Summing the above inequality for $t \in \{1, \ldots, T\}$ yields

$$\frac{\eta K}{2(1-\lambda)}\sum_{t=1}^{T}\mathbb{E}[\|\nabla\mathcal{F}(\Phi^{t-1})\|^2] \leq (\mathbb{E}[\mathcal{F}(z^0)] - \mathbb{E}[\mathcal{F}(z^T)]) + \frac{\eta K L^2}{2(1-\lambda)}\sum_{t=1}^{T}\frac{1}{KN}\sum_{k,C_i}\mathbb{E}[\|\theta_{i,k-1}^t - \theta_{i,0}^t\|^2]$$
$$+ \frac{L}{2(1-\lambda)}\sum_{t=1}^{T}\mathbb{E}[\|m^{t-1}\|^2] + (\frac{L}{2(1-\lambda)^3} + \frac{L}{2(1-\lambda)^2})\sum_{t=1}^{T}\mathbb{E}[\|\Delta^t\|^2].$$

By applying Lemma 4, Lemma 5, and Lemma 6, we have

$$\begin{aligned}
\frac{\eta K}{2(1-\lambda)}\sum_{t=1}^{T}\mathbb{E}[\|\nabla\mathcal{F}(\Phi^{t-1})\|^2] &\leq (\mathbb{E}[\mathcal{F}(z^0)] - \mathbb{E}[\mathcal{F}(z^T)]) + \frac{\eta K L^2}{2(1-\lambda)}\sum_{t=1}^{T}\frac{1}{KN}\sum_{k,C_i}\mathbb{E}[\|\theta_{i,k-1}^t - \theta_{i,0}^t\|^2] \\
&\quad + (\frac{2L}{2(1-\lambda)^3} + \frac{L}{2(1-\lambda)^2})\sum_{t=1}^{T}\mathbb{E}[\|\Delta^t\|^2] \\
&\leq (\mathbb{E}[\mathcal{F}(z^0)] - \mathbb{E}[\mathcal{F}(z^T)]) \\
&\quad + \frac{\eta K}{2(1-\lambda)}L^2(\frac{8\eta KL}{(1-\lambda)^2} + \frac{4\eta KL}{1-\lambda} + 1)\sum_{t=1}^{T}\frac{1}{KN}\sum_{k,C_i}\mathbb{E}[\|\theta_{i,k-1}^t - \theta_{i,0}^t\|^2] \\
&\quad + \frac{\eta K}{2(1-\lambda)}(\frac{4\eta KL}{(1-\lambda)^2} + \frac{2\eta KL}{1-\lambda})\sum_{t=1}^{T}\left(4(B^2+1)\|\nabla\mathcal{F}(\Phi^{t-1})\|^2 + \frac{4(1-\frac{|S_t|}{N})}{|S_t|}\sigma_g^2 + \frac{\sigma^2}{K|S_t|}\right) \\
&\leq (\mathbb{E}[\mathcal{F}(z^0)] - \mathbb{E}[\mathcal{F}(z^T)]) \\
&\quad + \frac{\eta K}{2(1-\lambda)}L^2(\frac{8\eta KL}{(1-\lambda)^2} + \frac{4\eta KL}{1-\lambda} + 1)\sum_{t=1}^{T}\left(6\eta^2K^2\left(\delta_g^2 + B^2\mathbb{E}[\|\nabla\mathcal{F}\left(\Phi^{t-1}\right)\|^2]\right) + 3\eta^2K\sigma^2\right) \\
&\quad + \frac{\eta K}{2(1-\lambda)}(\frac{4\eta KL}{(1-\lambda)^2} + \frac{2\eta KL}{1-\lambda})\sum_{t=1}^{T}\left(4(B^2+1)\|\nabla\mathcal{F}(\Phi^{t-1})\|^2 + \frac{4(1-\frac{|S_t|}{N})}{|S_t|}\sigma_g^2 + \frac{\sigma^2}{K|S_t|}\right).
\end{aligned}$$

If $\eta \leq \frac{(1-\lambda)^2}{64KL(B^2+1)}$, we can rewrite the above inequality as follows

$$\begin{aligned}
\frac{\eta K}{4(1-\lambda)}\sum_{t=1}^{T}\mathbb{E}[\|\nabla\mathcal{F}(\Phi^{t-1})\|^2] &\leq (\mathbb{E}[\mathcal{F}(z^0)] - \mathbb{E}[\mathcal{F}(z^T)]) + \frac{\eta K}{2(1-\lambda)}L^2(\frac{8\eta KL}{(1-\lambda)^2} + \frac{4\eta KL}{1-\lambda} + 1)\sum_{t=1}^{T}\left(6\eta^2K^2\delta_g^2 + 3\eta^2K\sigma^2\right) \\
&\quad + \frac{\eta K}{2(1-\lambda)}(\frac{4\eta KL}{(1-\lambda)^2} + \frac{2\eta KL}{1-\lambda})\sum_{t=1}^{T}\left(\frac{4(1-\frac{|S_t|}{N})}{|S_t|}\sigma_g^2 + \frac{\sigma^2}{K|S_t|}\right) \\
&\leq (\mathbb{E}[\mathcal{F}(z^0)] - \mathbb{E}[\mathcal{F}(z^T)]) + 35L^2(1-\lambda)^5\left(\frac{\eta K}{4(1-\lambda)^2}\right)^3\sum_{t=1}^{T}\left(6\delta_g^2 + \frac{3}{K}\sigma^2\right) \\
&\quad + 8L(1-\lambda)\left(\frac{\eta K}{4(1-\lambda)^2}\right)^2\sum_{t=1}^{T}\left(\frac{4(1-\frac{|S_t|}{N})}{|S_t|}\sigma_g^2 + \frac{\sigma^2}{K|S_t|}\right).
\end{aligned}$$

Let $\tilde{\eta} = \frac{\eta K}{4(1-\lambda)^2}$. By dividing both sides by $1-\lambda$, we have

$$\tilde{\eta}\sum_{t=1}^{T}\mathbb{E}[\|\nabla\mathcal{F}(\Phi^{t-1})\|^2] \leq \frac{(\mathbb{E}[\mathcal{F}(z^0)] - \mathbb{E}[\mathcal{F}(z^T)])}{1-\lambda} + 35L^2(1-\lambda)^4\tilde{\eta}^3 T\left(6\delta_g^2 + \frac{3}{K}\sigma^2\right) + 8L\tilde{\eta}^2 T\left(\frac{4(1-\frac{|S_t|}{N})}{|S_t|}\sigma_g^2 + \frac{\sigma^2}{K|S_t|}\right).$$

Dividing both side by $\tilde{\eta}T$ yields

$$\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}[\|\nabla\mathcal{F}(\Phi^{t-1})\|^2] \leq \frac{(\mathbb{E}[\mathcal{F}(\theta^0)] - \mathbb{E}[\mathcal{F}(\theta^*)])}{\tilde{\eta}T(1-\lambda)} + 35L^2(1-\lambda)^4\tilde{\eta}^2\left(6\delta_g^2 + \frac{3}{K}\sigma^2\right) + 8L\tilde{\eta}\left(\frac{4(1-\frac{|S_t|}{N})}{|S_t|}\sigma_g^2 + \frac{\sigma^2}{K|S_t|}\right).$$

Now we get the desired rate by applying Lemma 2, which finishes the proof. □

Lemma 4. *Algorithm 1 satisfies*

$$\sum_{t=1}^{T}\mathbb{E}[\left\|m^t\right\|^2] \leq \frac{1}{(1-\lambda)^2}\sum_{t=1}^{T}\mathbb{E}[\|\Delta^t\|^2]$$

Proof. Unrolling the recursion of the momentum m^t, *i.e.*, $m^t = \sum_{r=1}^{t}\lambda^{t-r}\Delta^r$

$$\mathbb{E}[\|m^t\|^2] = \mathbb{E}[\|\sum_{r=1}^{t}\lambda^{t-r}\Delta^r\|^2].$$

Let $\Gamma_t = \sum_{r=0}^{t-1}\lambda^r = \frac{1-\lambda^t}{1-\lambda}$. Since $0 \leq \lambda < 1$, $\Gamma_t \leq \frac{1}{1-\lambda}$, we have

$$\begin{aligned}\mathbb{E}[\|\sum_{r=1}^{t}\lambda^{t-r}\Delta^r\|^2] &= \Gamma_t^2\mathbb{E}[\|\frac{1}{\Gamma_t}\sum_{r=1}^{t}\lambda^{t-r}\Delta^r\|^2] \\ &\leq \Gamma_t\sum_{r=1}^{t}\lambda^{t-r}\mathbb{E}[\|\Delta^r\|^2] \\ &\leq \frac{1}{1-\lambda}\sum_{r=1}^{t}\lambda^{t-r}\mathbb{E}[\|\Delta^r\|^2].\end{aligned}$$

By summing the above inequality for $t \in \{0, \dots, T-1\}$, we have

$$\begin{aligned}\sum_{t=1}^{T}\mathbb{E}\left[\left\|m^t\right\|^2\right] &\leq \sum_{t=1}^{T}\frac{1}{1-\lambda}\sum_{r=1}^{t}\lambda^{t-r}\mathbb{E}[\|\Delta^r\|^2] \\ &\leq \frac{1}{(1-\lambda)^2}\sum_{t=1}^{T}\mathbb{E}[\|\Delta^t\|^2],\end{aligned}$$

which finishes the proof. □

Lemma 5. *For all $t \geq 1$, Algorithm 1 satisfies*

$$\mathbb{E}[\|\Delta^t\|^2] \leq 2(\eta K)^2\left(\frac{2L^2}{KN}\sum_{k,C_i}\mathbb{E}[\|\theta_{i,k}^t - \theta_{i,0}^t\|^2] + 4(B^2+1)\|\nabla\mathcal{F}(\Phi^{t-1})\|^2 + \frac{4(1-\frac{|S_t|}{N})}{|S_t|}\sigma_g^2 + \frac{\sigma^2}{K|S_t|}\right),$$

where $\theta_{i,0}^t$ denotes the initial point for the local model of the i-th client, i.e., $\theta_{i,0}^t = \Phi^{t-1}$.

Proof. By applying Lemma 3, we have

$$\begin{aligned}
\mathbb{E}[\|\Delta^t\|^2] &= \mathbb{E}[\|\frac{\eta K}{K|S_t|}\sum_{k,C_i\in S_t}\nabla f_i(\theta_{i,k}^t)\|^2] \\
&\leq 2(\eta K)^2\big(\mathbb{E}[\|\frac{1}{K|S_t|}\sum_{k,C_i\in S_t}\nabla\mathcal{F}_i(\theta_{i,k}^t)\|^2] + \frac{\sigma^2}{K|S_t|}\big),
\end{aligned}$$

We note that

$$\begin{aligned}
&\mathbb{E}[\|\frac{1}{K|S_t|}\sum_{k,C_i\in S_t}\nabla\mathcal{F}_i(\theta_{i,k}^t)\|^2] \\
&= \mathbb{E}[\|\frac{1}{K|S_t|}\sum_{k,C_i\in S_t}\left(\nabla\mathcal{F}_i(\theta_{i,k}^t) - \nabla\mathcal{F}_i(\theta_{i,0}^t) + \nabla\mathcal{F}_i(\theta_{i,0}^t)\right)\|^2] \\
&\leq 2\mathbb{E}[\|\frac{1}{K|S_t|}\sum_{k,C_i\in S_t}\left(\nabla\mathcal{F}_i(\theta_{i,k}^t) - \nabla\mathcal{F}_i(\theta_{i,0}^t)\right)\|^2] + 2\mathbb{E}[\|\frac{1}{|S_t|}\sum_{C_i\in S_t}\nabla\mathcal{F}_i(\theta_{i,0}^t)\|^2] \\
&\leq \frac{2}{KN}\sum_{k,C_i}\mathbb{E}[\|\nabla\mathcal{F}_i(\theta_{i,k}^t) - \nabla\mathcal{F}_i(\theta_{i,0}^t)\|^2] + \mathbb{E}[\|\frac{2}{|S_t|}\sum_{C_i\in S_t}\left(\nabla\mathcal{F}_i(\theta_{i,0}^t) - \nabla\mathcal{F}(\Phi^{t-1}) + \nabla\mathcal{F}(\Phi^{t-1})\right)\|^2] \\
&\leq \frac{2L^2}{KN}\sum_{k,C_i}\mathbb{E}[\|\theta_{i,k}^t - \theta_{i,0}^t\|^2] + \mathbb{E}[\|\frac{2}{|S_t|}\sum_{C_i\in S_t}\left(\nabla\mathcal{F}_i(\theta_{i,0}^t) - \nabla\mathcal{F}(\Phi^{t-1}) + \nabla\mathcal{F}(\Phi^{t-1})\right)\|^2] \\
&\leq \frac{2L^2}{KN}\sum_{k,C_i}\mathbb{E}[\|\theta_{i,k}^t - \theta_{i,0}^t\|^2] + 4\|\nabla\mathcal{F}(\Phi^{t-1})\|^2 + \frac{4(1-\frac{|S_t|}{N})}{|S_t|N}\sum_{C_i}\|\nabla\mathcal{F}_i(\theta_{i,0}^t)\|^2 \\
&\leq \frac{2L^2}{KN}\sum_{k,C_i}\mathbb{E}[\|\theta_{i,k}^t - \theta_{i,0}^t\|^2] + 4(B^2+1)\|\nabla\mathcal{F}(\Phi^{t-1})\|^2 + \frac{4(1-\frac{|S_t|}{N})}{|S_t|}\sigma_g^2,
\end{aligned}$$

where, in the fourth inequality, the improvement of $(1-\frac{|S_t|}{N})$ follows from sampling the active client set S_t without replacement at the t-th communication round. The last inequality holds because the average norm of local gradients is bounded as $\frac{1}{N}\sum_{i=1}^{N}\|\nabla\mathcal{F}_i(x)\|^2 \leq \sigma_g^2 + B^2\|\nabla\mathcal{F}(x)\|^2$, which concludes the proof. □

Lemma 6. *For all $t \geq 1$, we have*

$$\frac{1}{KN}\sum_{k,C_i}\mathbb{E}[\|\theta_{i,k}^t - \theta_{i,0}^t\|^2] \leq 6\eta^2K^2\left(\delta_g^2 + B^2\mathbb{E}[\|\nabla\mathcal{F}\left(\Phi^{t-1}\right)\|^2]\right) + 3\eta^2K\sigma^2.$$

Proof. We first define the following terms as

$$I_{i,k}^t = \mathbb{E}[\|\theta_{i,k}^t - \theta_{i,0}^t\|^2],\ I_i^t = \frac{1}{K}\sum_{k=1}^{K}I_{i,k}^t,\ I^t = \frac{1}{N}\sum_{C_i}I_i^t. \tag{8}$$

Initially, we commence by deriving an upper bound for the variable $I_{i,k}^t$ as

$$\begin{aligned}
I_{i,k}^t &= \mathbb{E}[\|\theta_{i,k}^t - \theta_{i,0}^t\|^2] \\
&= \mathbb{E}[\|\theta_{i,k-1}^t - \theta_{i,0}^t - \eta\nabla f_i\left(\theta_{i,k-1}^t\right)\|^2] \\
&= \mathbb{E}[\|\theta_{i,k-1}^t - \theta_{i,0}^t - \eta\nabla\mathcal{F}_i\left(\theta_{i,k-1}^t\right) + \eta\nabla\mathcal{F}_i\left(\theta_{i,k-1}^t\right) - \eta\nabla f_i\left(\theta_{i,k-1}^t\right)\|^2] \\
&\leq \mathbb{E}[\|\theta_{i,k-1}^t - \theta_{i,0}^t - \eta\nabla\mathcal{F}_i\left(\theta_{i,k-1}^t\right)\|^2] + \eta^2\sigma^2 \\
&\leq \left(1+\frac{1}{K-1}\right)\mathbb{E}[\|\theta_{i,k-1}^t - \theta_{i,0}^t\|^2] + K\eta^2\mathbb{E}[\|\nabla\mathcal{F}_i\left(\theta_{i,k-1}^t\right)\|^2] + \eta^2\sigma^2,
\end{aligned} \tag{9}$$

where the first inequality follows because the stochastic gradient possesses a bounded variance, while the second inequality follows from the Lemma 1.

We note that

$$\begin{aligned}\mathbb{E}[\|\nabla\mathcal{F}_i\left(\theta_{i,k-1}^t\right)\|^2] &= \mathbb{E}[\|\nabla\mathcal{F}_i\left(\theta_{i,k-1}^t\right) - \nabla\mathcal{F}_i\left(\theta_{i,0}^t\right) + \nabla\mathcal{F}_i\left(\theta_{i,0}^t\right)\|^2] \\ &\leq 2\mathbb{E}[\|\nabla\mathcal{F}_i\left(\theta_{i,k-1}^t\right) - \nabla\mathcal{F}_i\left(\theta_{i,0}^t\right)\|^2] + 2\mathbb{E}[\|\nabla\mathcal{F}_i\left(\theta_{i,0}^t\right)\|^2] \\ &\leq 2L^2\mathbb{E}[\|\theta_{i,k-1}^t - \theta_{i,0}^t\|^2] + 2\mathbb{E}[\|\nabla\mathcal{F}_i\left(\theta_{i,0}^t\right)\|^2]. \end{aligned} \tag{10}$$

By substituting Eq. (10) into Eq. (9), we have

$$\begin{aligned} I_{i,k}^t &\leq \left(1 + \frac{1}{K-1} + 2K\eta^2L^2\right)\mathbb{E}[\|\theta_{i,k-1}^t - \theta_{i,0}^t\|^2] + 2K\eta^2\mathbb{E}[\|\nabla\mathcal{F}_i\left(\theta_{i,0}^t\right)\|^2] + \eta^2\sigma^2 \\ &\leq \left(1 + \frac{1}{K-1} + 2K\eta^2L^2\right)I_{i,k-1}^t + 2K\eta^2\mathbb{E}[\|\nabla\mathcal{F}_i\left(\theta_{i,0}^t\right)\|^2] + \eta^2\sigma^2. \end{aligned}$$

By unrolling the recursion, we have

$$I_{i,k}^t \leq \sum_{r=0}^{k-1}\left(2K\eta^2\mathbb{E}[\|\nabla\mathcal{F}_i\left(\theta_{i,0}^t\right)\|^2] + \eta^2\sigma^2\right)\left(1 + \frac{2}{K-1}\right)^r \leq 3K\left(2K\eta^2\mathbb{E}[\|\nabla\mathcal{F}_i\left(\theta_{i,0}^t\right)\|^2] + \eta^2\sigma^2\right).$$

By the definitions in Eq. (8), we have

$$\begin{aligned} I_i^t = \frac{1}{K}\sum_{k=1}^{K} I_{i,k}^t &\leq 3K\left(2K\eta^2\mathbb{E}[\|\nabla\mathcal{F}_i\left(\theta_{i,0}^t\right)\|^2] + \eta^2\sigma^2\right) \\ &= 6\eta^2K^2\mathbb{E}[\|\nabla\mathcal{F}_i\left(\theta_{i,0}^t\right)\|^2] + 3\eta^2K\sigma^2. \end{aligned}$$

$$\begin{aligned} I^t &= 6\eta^2K^2\frac{1}{N}\sum_{C_i}\mathbb{E}[\|\nabla\mathcal{F}_i\left(\theta_{i,0}^t\right)\|^2] + 3\eta^2K\sigma^2 \\ &\leq 6\eta^2K^2\left(\delta_g^2 + B^2\mathbb{E}[\|\nabla\mathcal{F}\left(\Phi^{t-1}\right)\|^2]\right) + 3\eta^2K\sigma^2, \end{aligned}$$

where the inequality follows due to the assumption that the average norm of the local gradients is bounded, *i.e.*, $\frac{1}{N}\sum_{i=1}^{N}\|\nabla\mathcal{F}_i(x)\|^2 \leq \sigma_g^2 + B^2\|\nabla\mathcal{F}(x)\|^2$, which completes the proof. □