ICLR

Understanding the Stability-based Generalization of Personalized Federated Learning

Yingqi Liu, Qinglun Li, Tan Jie, Yifan shi, Li Shen*, Xiaochun Cao

中山大學
SUN YAT-SEN UNIVERSITY

国防科技大学
National University of Defense Technology

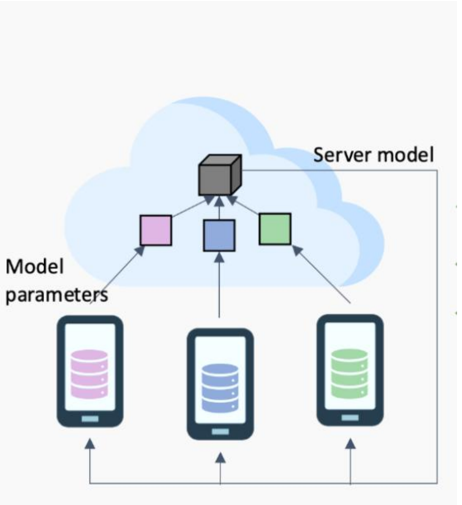
April, 2025

Outline

- Background
- Contributions
- Problem Setup
- Assumptions
- Theoretical Analysis
- Experiments
- Future Works

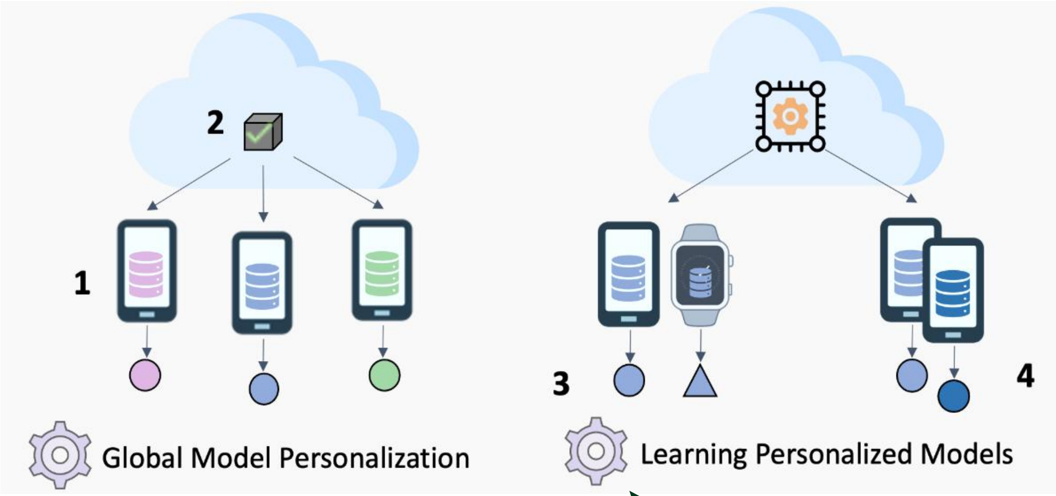
Background

Federated Learning



- ✓ Generalization
- ✓ Privacy
- ✓ Communication
- ✗ Data Heterogeneity
- ✗ Solution Personalization

Personalized Federated Learning



- ✓ Generalization
- ✓ Privacy
- ✓ Communication
- ✓ Data Heterogeneity
- ✓ Solution Personalization

Generalization Analysis?

- The figures are cited from *Towards Personalized Federated Learning*.

Background

Generalization Analysis for PFL

- high-probability generalization bounds with concentration inequalities based on the PAC hypothesis complexity (VC dimension complexity, Rademacher complexity)

- information-theoretical distances between the output hypothesis and the prior from PAC-Bayes generalization

- the upper privacy-preserving ability of the change in output hypothesis when the algorithm is exposed to attacks

✖ can not apply to the commonly used nonconvex functions such as neural networks
✖ weak at evaluating the effectiveness of algorithm design and hyperparameter selection
✖ can not reflect the personalized iteration performance

Contributions

➢ New framework of generalization analysis for PFL under non-convex conditions.

- We build up the first algorithm-dependent generalization analysis framework for PFL with the biased gradient from multi-local updates.
- It is consistent with the personalized training progress and bridges PFL, FL and Pure Local Training with the clever heterogeneity analysis, which reveals the effective-ness of PFL for personalized aims.
- We also extend it to the decentralized scenarios with different communication topologies.

➢ New results for upper generalization bounds and excess risks for PFL.

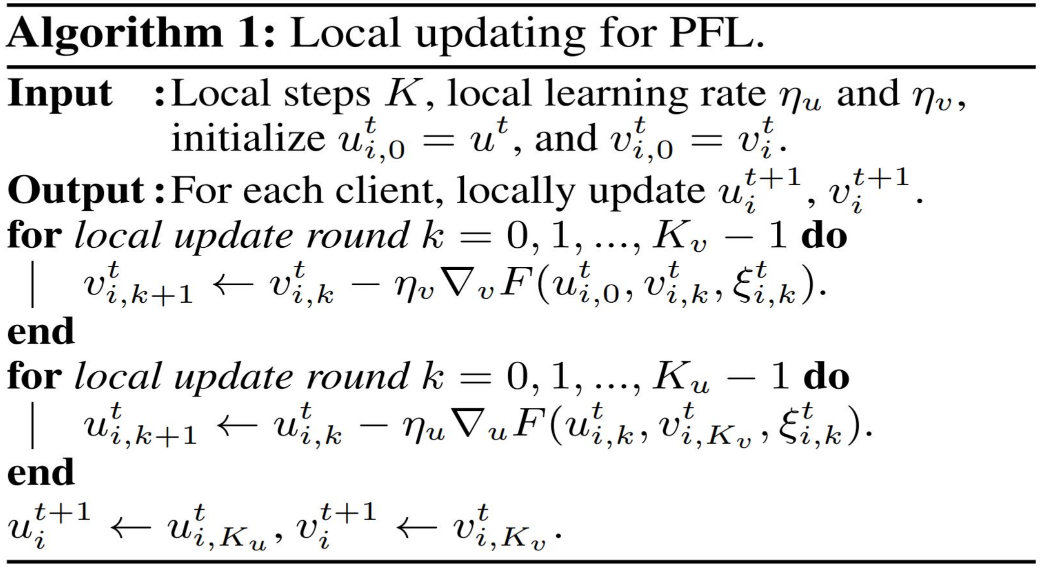
- Our algorithm-dependent results achieve comparable bounds and reflect the iteration nature, effectiveness of algorithm design as well as the hyperparameters selection of PFL.
- Then combined with the optimization errors, we obtain the excess risk analysis and find that the better performance is the trade-off between optimization and generalization.

➢ Massive experiments verify theoretical findings.

- Our experiments on CIFAR datasets with different models under non-convex conditions strongly support our theoretical insights.

Problem Setup

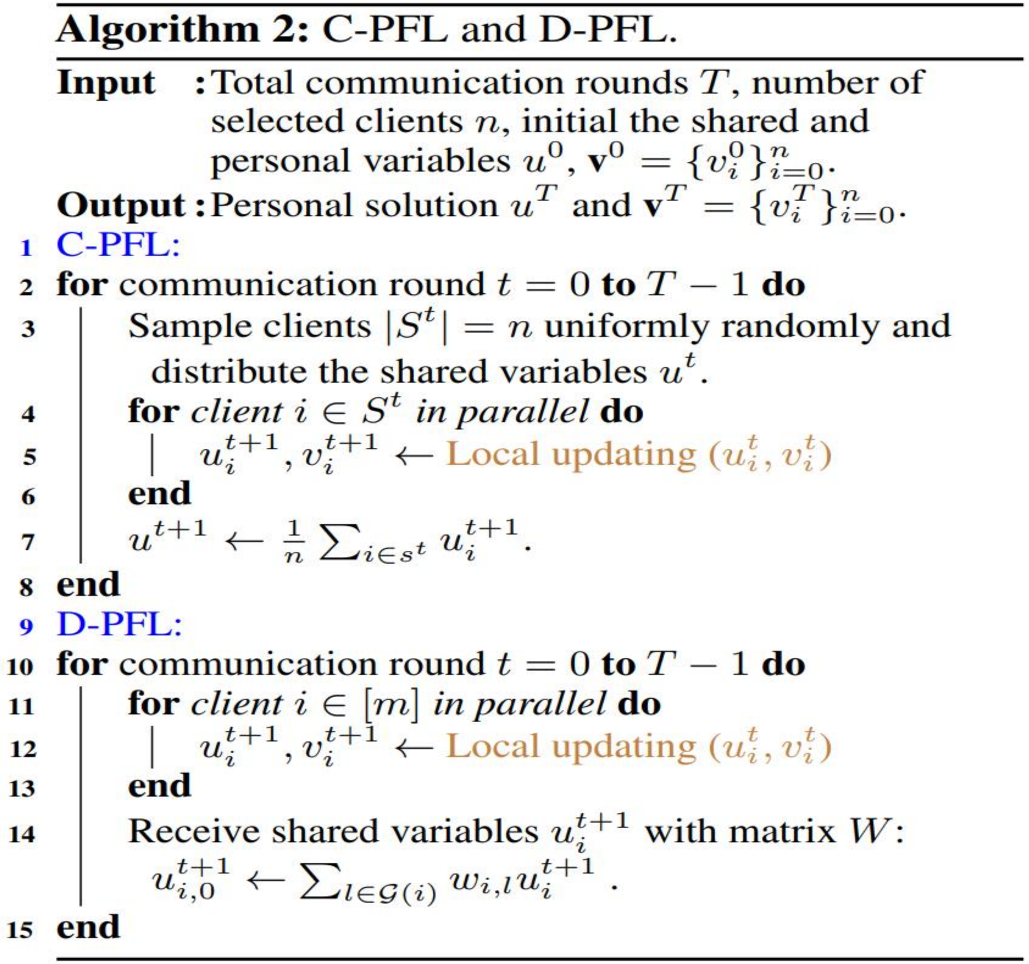
Personalized Federated Learning.



Algorithm 1: Local updating for PFL.

```
Input  : Local steps $K$, local learning rate $\eta_u$ and $\eta_v$,
         initialize $u_{i,0}^t = u^t$, and $v_{i,0}^t = v_i^t$.
Output : For each client, locally update $u_i^{t+1}$, $v_i^{t+1}$.
for local update round $k = 0, 1, ..., K_v - 1$ do
    $v_{i,k+1}^t \leftarrow v_{i,k}^t - \eta_v \nabla_v F(u_{i,0}^t, v_{i,k}^t, \xi_{i,k}^t)$.
end
for local update round $k = 0, 1, ..., K_u - 1$ do
    $u_{i,k+1}^t \leftarrow u_{i,k}^t - \eta_u \nabla_u F(u_{i,k}^t, v_{i,K_v}^t, \xi_{i,k}^t)$.
end
$u_i^{t+1} \leftarrow u_{i,K_u}^t$, $v_i^{t+1} \leftarrow v_{i,K_v}^t$.
```

C-PFL and D-PFL.



Algorithm 2: C-PFL and D-PFL.

```
Input  : Total communication rounds $T$, number of
         selected clients $n$, initial the shared and
         personal variables $u^0$, $\mathbf{v}^0 = \{v_i^0\}_{i=0}^n$.
Output : Personal solution $u^T$ and $\mathbf{v}^T = \{v_i^T\}_{i=0}^n$.
C-PFL:
for communication round $t = 0$ to $T - 1$ do
    Sample clients $|S^t| = n$ uniformly randomly and
     distribute the shared variables $u^t$.
    for client $i \in S^t$ in parallel do
        $u_i^{t+1}, v_i^{t+1} \leftarrow$ Local updating $(u_i^t, v_i^t)$
    end
    $u^{t+1} \leftarrow \frac{1}{n} \sum_{i \in s^t} u_i^{t+1}$.
end
D-PFL:
for communication round $t = 0$ to $T - 1$ do
    for client $i \in [m]$ in parallel do
        $u_i^{t+1}, v_i^{t+1} \leftarrow$ Local updating $(u_i^t, v_i^t)$
    end
    Receive shared variables $u_i^{t+1}$ with matrix $W$:
     $u_{i,0}^{t+1} \leftarrow \sum_{l \in \mathcal{G}(i)} w_{i,l} u_i^{t+1}$ .
end
```

Problem Setup

Generalization Stability. The generalization error between the **population risk** in (1) and **empirical risk** in (2) can be defined as:

$$\min_{u,V} \quad F(u,V) := \frac{1}{m}\sum_{i=1}^{m} F_i\left(u, v_i\right), \quad where \quad F_i\left(u, v_i\right) = \mathbb{E}_{\xi_i \sim \mathcal{D}_i} f_i\left(u, v_i; \xi_i\right). \qquad (1)$$

$$\min_{u,V} \quad f(u,V) := \frac{1}{m}\sum_{i=1}^{m} f_i\left(u, v_i\right), \quad where \quad f_i\left(u, v_i\right) = \frac{1}{\mathcal{S}}\sum_{\xi_i \in \mathcal{S}_i}\left[f_i\left(u, v_i; \xi_i\right)\right]. \qquad (2)$$

$$\varepsilon_G = \mathbb{E}_{\mathcal{S},\mathcal{A}}[F(\mathcal{A}(\mathcal{S})) - f(\mathcal{A}(\mathcal{S}))]$$

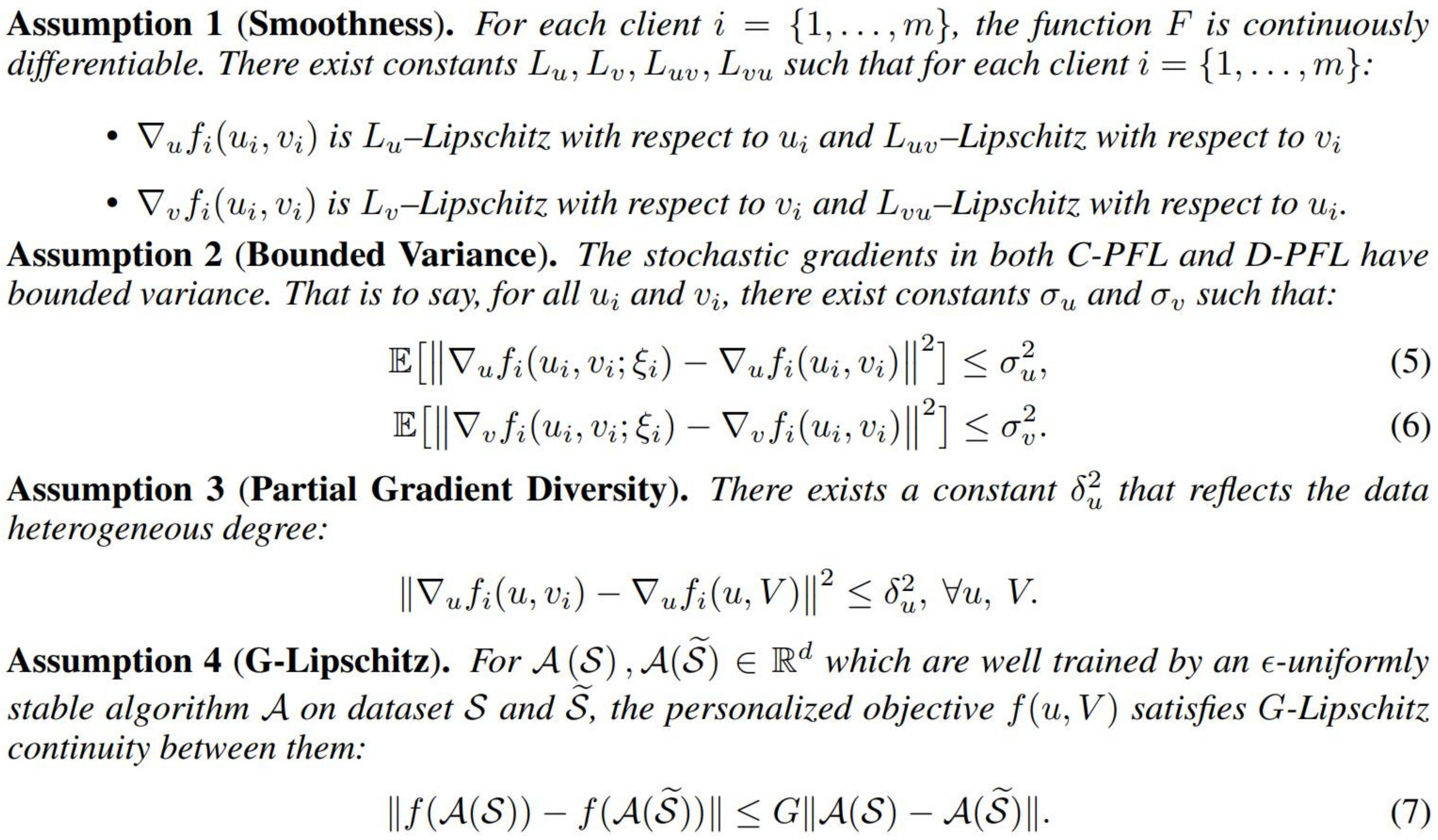
Uniform Stability. The ε-uniformly stability for algorithm is defined as below:

$$\sup_{z_j \sim \{\mathcal{D}_i\}} \mathbb{E}[f(u, V; z_j) - f(\widetilde{u}, \widetilde{V}; z_j)] \leq \epsilon.$$

Excess Risk. Considering *(u* , V *)* as the optimal model that can be achieved by the algorithm *A* on the dataset *S*, the real test performance *E[F(A(S))]* can be measured by the excess risk as follows:

$$\begin{aligned}\mathcal{E}_E &= \mathbb{E}[F(\mathcal{A}(\mathcal{S}))] - \mathbb{E}[f(u^*, V^*)] \\ &\leq \underbrace{\mathbb{E}[F(u,V) - f(u,V)]}_{\mathcal{E}_G:\ generalization\ error} + \underbrace{\mathbb{E}[f(u,V) - f(u^*, V^*)]}_{\mathcal{E}_O:\ optimization\ error}.\end{aligned}$$

Assumptions



Assumption 1 (Smoothness). *For each client* $i = \{1, \dots, m\}$*, the function* F *is continuously differentiable. There exist constants* L_u, L_v, L_{uv}, L_{vu} *such that for each client* $i = \{1, \dots, m\}$*:*

- $\nabla_u f_i(u_i, v_i)$ *is* L_u*–Lipschitz with respect to* u_i *and* L_{uv}*–Lipschitz with respect to* v_i
- $\nabla_v f_i(u_i, v_i)$ *is* L_v*–Lipschitz with respect to* v_i *and* L_{vu}*–Lipschitz with respect to* u_i*.*

Assumption 2 (Bounded Variance). *The stochastic gradients in both C-PFL and D-PFL have bounded variance. That is to say, for all* u_i *and* v_i*, there exist constants* σ_u *and* σ_v *such that:*

$$\mathbb{E}\left[\left\|\nabla_u f_i(u_i, v_i; \xi_i) - \nabla_u f_i(u_i, v_i)\right\|^2\right] \leq \sigma_u^2, \tag{5}$$

$$\mathbb{E}\left[\left\|\nabla_v f_i(u_i, v_i; \xi_i) - \nabla_v f_i(u_i, v_i)\right\|^2\right] \leq \sigma_v^2. \tag{6}$$

Assumption 3 (Partial Gradient Diversity). *There exists a constant* δ_u^2 *that reflects the data heterogeneous degree:*

$$\|\nabla_u f_i(u, v_i) - \nabla_u f_i(u, V)\|^2 \leq \delta_u^2, \ \forall u, \ V.$$

Assumption 4 (G-Lipschitz). *For* $\mathcal{A}(\mathcal{S}), \mathcal{A}(\widetilde{\mathcal{S}}) \in \mathbb{R}^d$ *which are well trained by an* ϵ*-uniformly stable algorithm* $\mathcal{A}$ *on dataset* $\mathcal{S}$ *and* $\widetilde{\mathcal{S}}$*, the personalized objective* $f(u, V)$ *satisfies G-Lipschitz continuity between them:*

$$\|f(\mathcal{A}(\mathcal{S})) - f(\mathcal{A}(\widetilde{\mathcal{S}}))\| \leq G\|\mathcal{A}(\mathcal{S}) - \mathcal{A}(\widetilde{\mathcal{S}})\|. \tag{7}$$

Theoretical Analysis

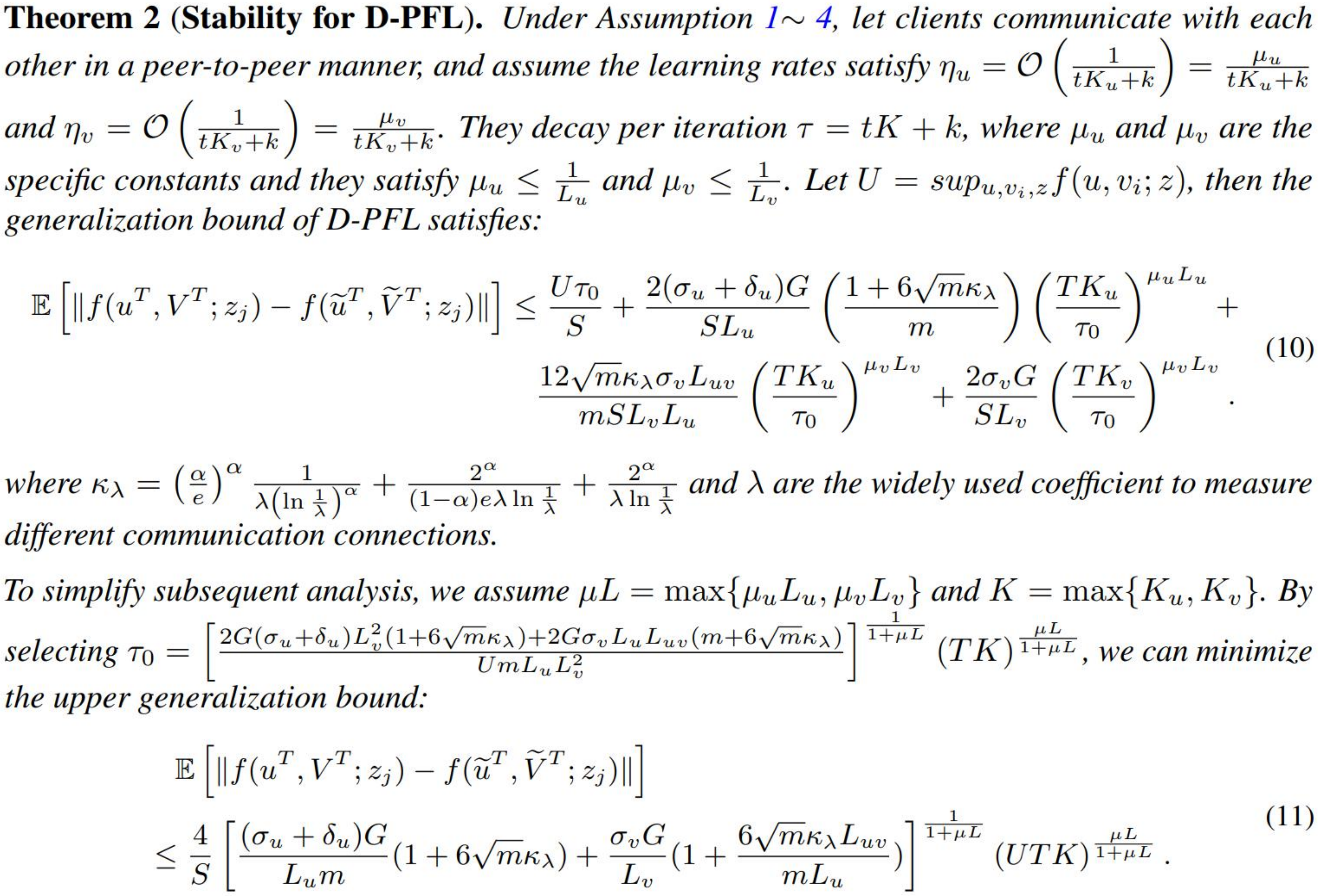
Theorem 1 (Stability of C-PFL). *Under Assumption 1∼ 4, let the active ratio per communication round be* n/m*, and assume the learning rates satisfy* $\eta_u = \mathcal{O}\left(\frac{1}{tK_u+k}\right) = \frac{\mu_u}{tK_u+k}$ *and* $\eta_v = \mathcal{O}\left(\frac{1}{tK_v+k}\right) = \frac{\mu_v}{tK_v+k}$*. They decay per iteration* $\tau = tK + k$*, where* μ_u *and* μ_v *are the specific constants and satisfy* $\mu_u \le \frac{1}{L_u}$ *and* $\mu_v \le \frac{1}{L_v}$*. Let* $U = sup_{u,v_i,z} f(u, v_i; z)$*, then the generalization bound of C-PFL satisfies:*

$$\begin{aligned}&\mathbb{E}\left[\|f(u^T,V^T;z_j)-f(\widetilde{u}^T,\widetilde{V}^T;z_j)\|\right]\\ &\le \frac{nU\tau_0}{mS}+\left(\frac{TK_u}{\tau_0}\right)^{\mu_u L_u}\frac{2G(\sigma_u+\delta_u)}{mSL_u}+\left(\frac{TK_v}{\tau_0}\right)^{\mu_v L_v}\left(1+\frac{L_{uv}}{L_u}(\frac{TK_u}{\tau_0})^{\mu_u L_u}\right)\frac{2G\sigma_v}{mSL_v}.\end{aligned} \tag{8}$$

To simplify subsequent analysis, we assume $\mu L = \max\{\mu_u L_u, \mu_v L_v\}$ *and* $K = \max\{K_u, K_v\}$*. By selecting* $\tau_0 = \left[\frac{2G((\sigma_u+\delta_u)L_v+\sigma_v L_u)}{nUL_uL_v}\right]^{\frac{1}{1+\mu L}}(TK)^{\frac{\mu L}{1+\mu L}}$*, we can minimize the bound with* τ_0*:*

$$\mathbb{E}\left[\|f(u^T,V^T;z_j)-f(\widetilde{u}^T,\widetilde{V}^T;z_j)\|\right] \le \frac{4}{mS}\left[\frac{G((\sigma_u+\delta_u)L_v+\sigma_v L_u)}{L_uL_v}\right]^{\frac{1}{1+\mu L}}(nUTK)^{\frac{\mu L}{1+\mu L}}. \tag{9}$$

Theoretical Analysis



Theorem 2 (Stability for D-PFL). *Under Assumption 1∼ 4, let clients communicate with each other in a peer-to-peer manner, and assume the learning rates satisfy* $\eta_u = \mathcal{O}\left(\frac{1}{tK_u+k}\right) = \frac{\mu_u}{tK_u+k}$ *and* $\eta_v = \mathcal{O}\left(\frac{1}{tK_v+k}\right) = \frac{\mu_v}{tK_v+k}$. *They decay per iteration* $\tau = tK + k$, *where* μ_u *and* μ_v *are the specific constants and they satisfy* $\mu_u \leq \frac{1}{L_u}$ *and* $\mu_v \leq \frac{1}{L_v}$. *Let* $U = sup_{u,v_i,z} f(u, v_i; z)$, *then the generalization bound of D-PFL satisfies:*

$$\mathbb{E}\left[\|f(u^T, V^T; z_j) - f(\widetilde{u}^T, \widetilde{V}^T; z_j)\|\right] \leq \frac{U\tau_0}{S} + \frac{2(\sigma_u+\delta_u)G}{SL_u}\left(\frac{1+6\sqrt{m}\kappa_\lambda}{m}\right)\left(\frac{TK_u}{\tau_0}\right)^{\mu_u L_u} + \frac{12\sqrt{m}\kappa_\lambda\sigma_v L_{uv}}{mSL_vL_u}\left(\frac{TK_u}{\tau_0}\right)^{\mu_v L_v} + \frac{2\sigma_v G}{SL_v}\left(\frac{TK_v}{\tau_0}\right)^{\mu_v L_v}. \tag{10}$$

where $\kappa_\lambda = \left(\frac{\alpha}{e}\right)^\alpha \frac{1}{\lambda\left(\ln\frac{1}{\lambda}\right)^\alpha} + \frac{2^\alpha}{(1-\alpha)e\lambda\ln\frac{1}{\lambda}} + \frac{2^\alpha}{\lambda\ln\frac{1}{\lambda}}$ *and* λ *are the widely used coefficient to measure different communication connections.*

To simplify subsequent analysis, we assume $\mu L = \max\{\mu_u L_u, \mu_v L_v\}$ *and* $K = \max\{K_u, K_v\}$. *By selecting* $\tau_0 = \left[\frac{2G(\sigma_u+\delta_u)L_v^2(1+6\sqrt{m}\kappa_\lambda)+2G\sigma_v L_u L_{uv}(m+6\sqrt{m}\kappa_\lambda)}{UmL_uL_v^2}\right]^{\frac{1}{1+\mu L}}(TK)^{\frac{\mu L}{1+\mu L}}$, *we can minimize the upper generalization bound:*

$$\mathbb{E}\left[\|f(u^T, V^T; z_j) - f(\widetilde{u}^T, \widetilde{V}^T; z_j)\|\right] \leq \frac{4}{S}\left[\frac{(\sigma_u+\delta_u)G}{L_u m}(1+6\sqrt{m}\kappa_\lambda) + \frac{\sigma_v G}{L_v}(1+\frac{6\sqrt{m}\kappa_\lambda L_{uv}}{mL_u})\right]^{\frac{1}{1+\mu L}}(UTK)^{\frac{\mu L}{1+\mu L}}. \tag{11}$$

Theoretical Analysis

- **Remark 1 (Influencal factors of PFL).** The stability of PFL is impacted by the number of samples *S*, total clients *m*, total iterations *TKu* and *TKv* , data heterogeneity *δu* in both C-PFL and D-PFL, and the number of the selected clients *n* in C-PFL and communication topologies κ_λ in D-PFL.

Specially, for D-PFL, the impact of communication topologies κ_λ is as follows:

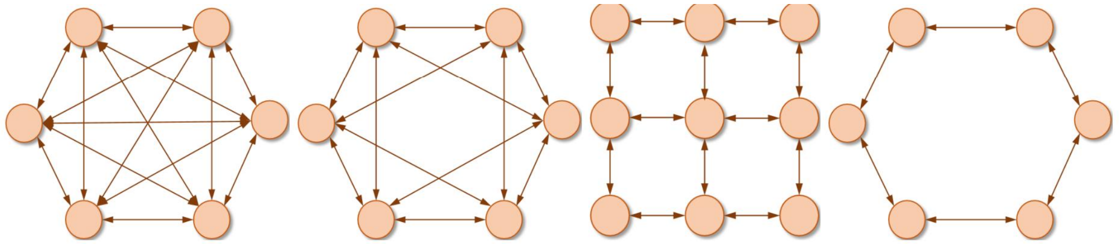
Table 3: κ_λ and Spectral Gap $1-\lambda$ of communication topologies (Sun et al., 2023c; Zhu et al., 2024).

Network Topology	κ_λ	Spectral Gap $1-\lambda$
Fully-connected	0	1
Disconnected	1	0
Ring	$\mathcal{O}(m^2)$	$\approx 3m^2/16\pi^2$
Grid	$\mathcal{O}(mlnm\ )$	$\mathcal{O}(mlog\_2(m))$
Exponential	$\mathcal{O}(lnm)$	$\mathcal{O}\left(1+log\_2(m)\right)$

$$\lambda \rightarrow 1 \qquad \kappa_\lambda \rightarrow \mathcal{O}\left(1/(\lambda\left(\ln\tfrac{1}{\lambda}\right))\right)$$

$$\lambda \rightarrow 0 \qquad \kappa_\lambda \rightarrow \mathcal{O}\left(1/(\lambda\left(\ln\tfrac{1}{\lambda}\right)^{\alpha})\right)$$

$$\kappa_\lambda = \left(\frac{\alpha}{e}\right)^{\alpha}\frac{1}{\lambda\left(\ln\frac{1}{\lambda}\right)^{\alpha}} + \frac{2^{\alpha}}{(1-\alpha)e\lambda\ln\frac{1}{\lambda}} + \frac{2^{\alpha}}{\lambda\ln\frac{1}{\lambda}},$$



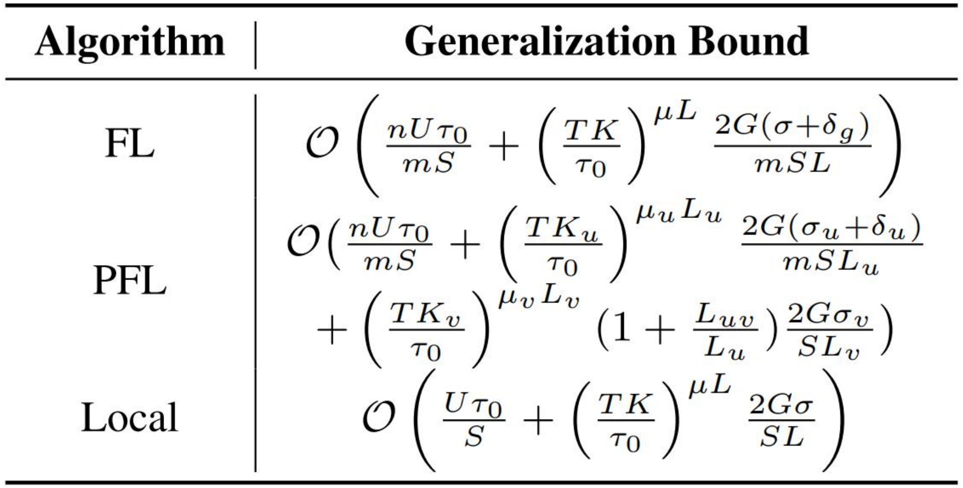
Fully-connected Exponential Grid Ring

Theoretical Analysis

- **Remark 2 (Special cases of PFL).** The degradation analysis of the shared variables *u* is as the same with vanilla SGD. The degradation analysis of personalized variables *v* is as the same with FedAvg and DFedAvg.

Specially, Personalization performs better than no personalization and Pure Local Training.

Table 2: Comparison with FL, PFL, Pure Local Training.



Algorithm	**Generalization Bound**
FL	$\mathcal{O}\left(\frac{nU\tau_0}{mS}+\left(\frac{TK}{\tau_0}\right)^{\mu L}\frac{2G(\sigma+\delta_g)}{mSL}\right)$
PFL	$\mathcal{O}\big(\frac{nU\tau_0}{mS}+\left(\frac{TK_u}{\tau_0}\right)^{\mu_u L_u}\frac{2G(\sigma_u+\delta_u)}{mSL_u}$ $+\left(\frac{TK_v}{\tau_0}\right)^{\mu_v L_v}(1+\frac{L_{uv}}{L_u})\frac{2G\sigma_v}{SL_v}\big)$
Local	$\mathcal{O}\left(\frac{U\tau_0}{S}+\left(\frac{TK}{\tau_0}\right)^{\mu L}\frac{2G\sigma}{SL}\right)$

FL

$$\frac{1}{m}\sum_{i=1}^{m}\left\|\nabla F_i\left(w_i\right)-\nabla F(w)\right\|^2 \leq \delta_g^2.$$

PFL

$$\delta_u^2 \leq \delta_g^2$$

$$\frac{1}{m}\sum_{i=1}^{m}\left\|\nabla_u F_i\left(u, v_i\right)-\nabla_u F(u, V)\right\|^2 \leq \delta_u^2$$

Theoretical Analysis

Corollary 1 (Excess risk of C-PFL.)

$$\mathcal{E}_E \le \mathbb{E}[F(\mathcal{A}(\mathcal{S}))] - \mathbb{E}[f(w^*)] \le \mathcal{O}\left(1/\sqrt{T} + (nKT)^{\frac{\mu L}{1+\mu L}}/m\right)$$

Corollary 2 (Excess risk of D-PFL.)

$$\mathcal{E}_E \le \mathbb{E}[F(\mathcal{A}(\mathcal{S}))] - \mathbb{E}[f(w^*)] \le \mathcal{O}\left(1/(1-\lambda)^2\sqrt{T} + (1+6\sqrt{m}\kappa_\lambda/m)^{\frac{1}{1+\mu L}}(KT)^{\frac{\mu L}{1+\mu L}}\right)$$

➢ **Remark 3 (Comparisions between the C-PFL and D-PFL).**

C-PFL always converges and generalizes better than D-PFL in theoretical analysis. C-PFL largely reduces the generalization error with the regular averaging on a global server, which leads to better consensus of the shared variables and better generalization.

Core Comparisons

- Compared with the current stability-based analysis, our work is the first to propose a personalized federated learning algorithm analysis with multiple local updates and hyperparameter analysis.

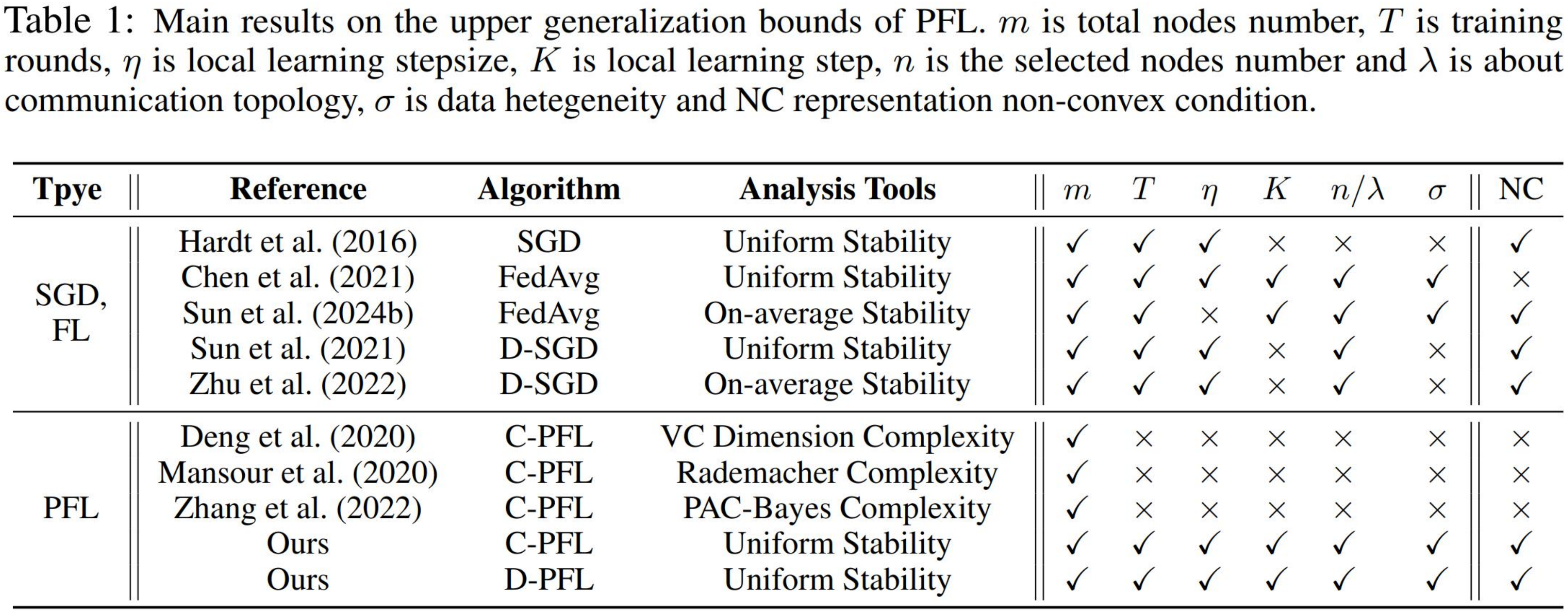


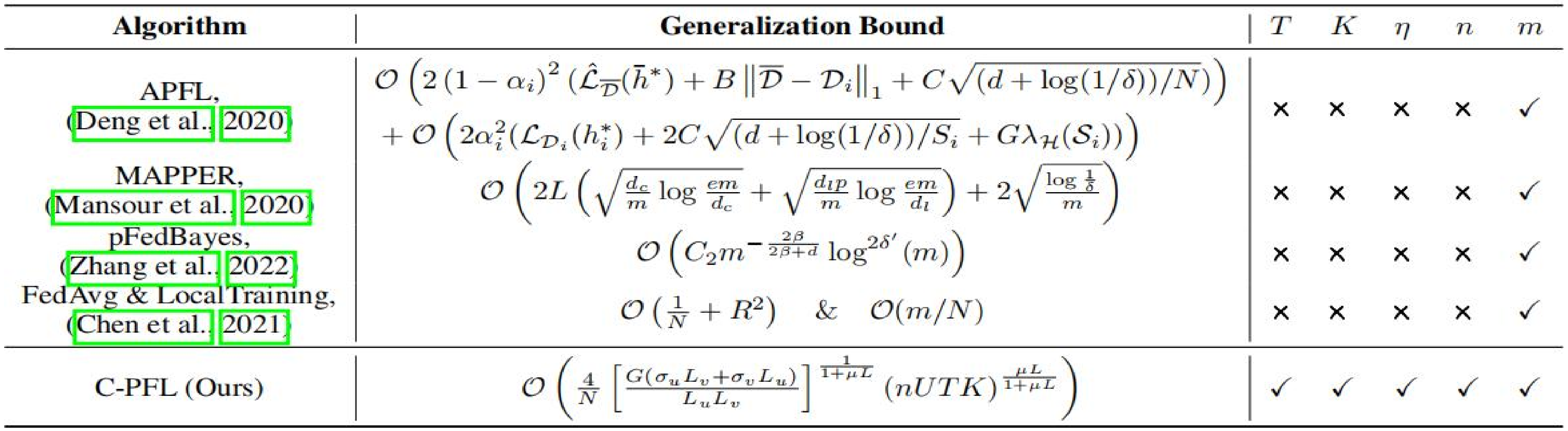
Table 1: Main results on the upper generalization bounds of PFL. m is total nodes number, T is training rounds, η is local learning stepsize, K is local learning step, n is the selected nodes number and λ is about communication topology, σ is data hetegeneity and NC representation non-convex condition.

Tpye	Reference	Algorithm	Analysis Tools	m	T	η	K	n/λ	σ	NC
SGD, FL	Hardt et al. (2016)	SGD	Uniform Stability	✓	✓	✓	×	×	×	✓
	Chen et al. (2021)	FedAvg	Uniform Stability	✓	✓	✓	✓	✓	✓	×
	Sun et al. (2024b)	FedAvg	On-average Stability	✓	✓	×	✓	✓	✓	✓
	Sun et al. (2021)	D-SGD	Uniform Stability	✓	✓	✓	×	✓	×	✓
	Zhu et al. (2022)	D-SGD	On-average Stability	✓	✓	✓	×	✓	×	✓
PFL	Deng et al. (2020)	C-PFL	VC Dimension Complexity	✓	×	×	×	×	×	×
	Mansour et al. (2020)	C-PFL	Rademacher Complexity	✓	×	×	×	×	×	×
	Zhang et al. (2022)	C-PFL	PAC-Bayes Complexity	✓	×	×	×	×	×	×
	Ours	C-PFL	Uniform Stability	✓	✓	✓	✓	✓	✓	✓
	Ours	D-PFL	Uniform Stability	✓	✓	✓	✓	✓	✓	✓

Core Comparisons

- As the first algorithm-dependent personalized generalization analysis, we introduce the discussion of data heterogeneity and for the first time explain the performance advantage of personalized algorithms over non-personalized algorithms from the perspective of generalization theory.

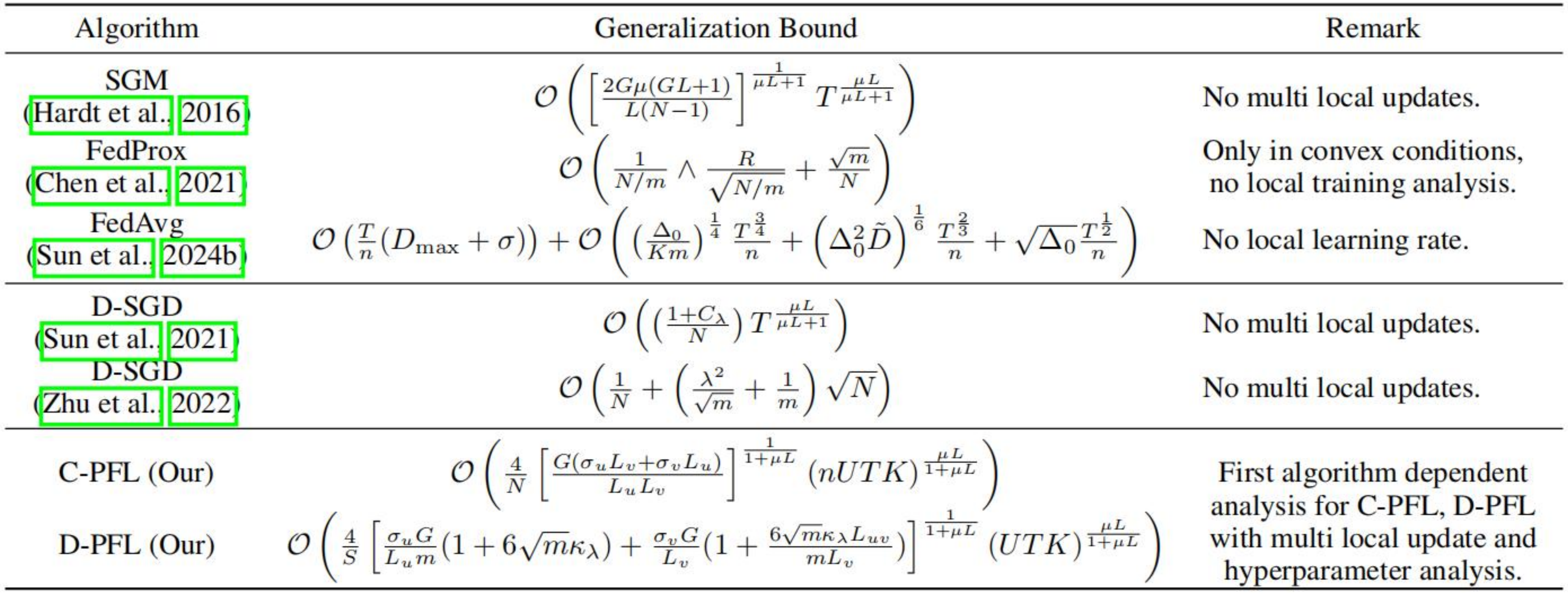
Table 3: Main results on the upper generalization bounds of PFL.



Algorithm	Generalization Bound	T	K	η	n	m
APFL, (Deng et al., 2020)	$\mathcal{O}\left(2\,(1-\alpha_i)^2\,(\hat{\mathcal{L}}_{\overline{\mathcal{D}}}(\bar{h}^*)+B\left\lVert\overline{\mathcal{D}}-\mathcal{D}_i\right\rVert_1+C\sqrt{(d+\log(1/\delta))/N})\right)$ $+\,\mathcal{O}\left(2\alpha_i^2(\mathcal{L}_{\mathcal{D}_i}(h_i^*)+2C\sqrt{(d+\log(1/\delta))/S_i}+G\lambda_{\mathcal{H}}(\mathcal{S}_i))\right)$	✗	✗	✗	✗	✓
MAPPER, (Mansour et al., 2020)	$\mathcal{O}\left(2L\left(\sqrt{\frac{d_c}{m}\log\frac{em}{d_c}}+\sqrt{\frac{d_l p}{m}\log\frac{em}{d_l}}\right)+2\sqrt{\frac{\log\frac{1}{\delta}}{m}}\right)$	✗	✗	✗	✗	✓
pFedBayes, (Zhang et al., 2022)	$\mathcal{O}\left(C_2 m^{-\frac{2\beta}{2\beta+d}}\log^{2\delta'}(m)\right)$	✗	✗	✗	✗	✓
FedAvg & Local Training, (Chen et al., 2021)	$\mathcal{O}\left(\frac{1}{N}+R^2\right)\quad\&\quad\mathcal{O}(m/N)$	✗	✗	✗	✗	✓
C-PFL (Ours)	$\mathcal{O}\left(\frac{4}{N}\left[\frac{G(\sigma_u L_v+\sigma_v L_u)}{L_u L_v}\right]^{\frac{1}{1+\mu L}}(nUTK)^{\frac{\mu L}{1+\mu L}}\right)$	✓	✓	✓	✓	✓

Core Comparisons

- It is the first work to analyze the generalization impact from personalized variables to shared variables, which uncovers the interaction mechanism between these two updating processes and provides valuable guidance for alternating personalized optimization.



Algorithm	Generalization Bound	Remark
SGM (Hardt et al., 2016)	$\mathcal{O}\left(\left[\frac{2G\mu(GL+1)}{L(N-1)}\right]^{\frac{1}{\mu L+1}} T^{\frac{\mu L}{\mu L+1}}\right)$	No multi local updates.
FedProx (Chen et al., 2021)	$\mathcal{O}\left(\frac{1}{N/m} \wedge \frac{R}{\sqrt{N/m}} + \frac{\sqrt{m}}{N}\right)$	Only in convex conditions, no local training analysis.
FedAvg (Sun et al., 2024b)	$\mathcal{O}\left(\frac{T}{n}(D_{\max}+\sigma)\right) + \mathcal{O}\left(\left(\frac{\Delta_0}{Km}\right)^{\frac{1}{4}} \frac{T^{\frac{3}{4}}}{n} + \left(\Delta_0^2 \tilde{D}\right)^{\frac{1}{6}} \frac{T^{\frac{2}{3}}}{n} + \sqrt{\Delta_0}\frac{T^{\frac{1}{2}}}{n}\right)$	No local learning rate.
D-SGD (Sun et al., 2021)	$\mathcal{O}\left(\left(\frac{1+C_\lambda}{N}\right) T^{\frac{\mu L}{\mu L+1}}\right)$	No multi local updates.
D-SGD (Zhu et al., 2022)	$\mathcal{O}\left(\frac{1}{N} + \left(\frac{\lambda^2}{\sqrt{m}} + \frac{1}{m}\right)\sqrt{N}\right)$	No multi local updates.
C-PFL (Our)	$\mathcal{O}\left(\frac{4}{N}\left[\frac{G(\sigma_u L_v+\sigma_v L_u)}{L_u L_v}\right]^{\frac{1}{1+\mu L}} (nUTK)^{\frac{\mu L}{1+\mu L}}\right)$	First algorithm dependent analysis for C-PFL, D-PFL with multi local update and hyperparameter analysis.
D-PFL (Our)	$\mathcal{O}\left(\frac{4}{S}\left[\frac{\sigma_u G}{L_u m}(1+6\sqrt{m}\kappa_\lambda) + \frac{\sigma_v G}{L_v}(1+\frac{6\sqrt{m}\kappa_\lambda L_{uv}}{mL_v})\right]^{\frac{1}{1+\mu L}} (UTK)^{\frac{\mu L}{1+\mu L}}\right)$	

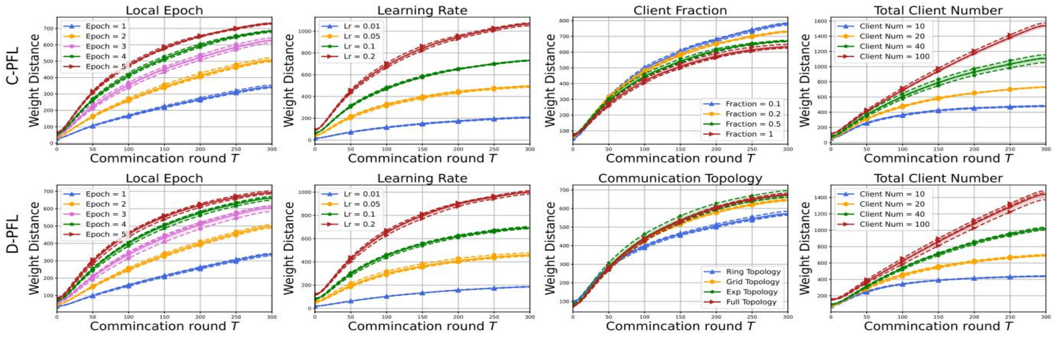
Experiments

We conduct the experiments on C-PFL and D-PFL:

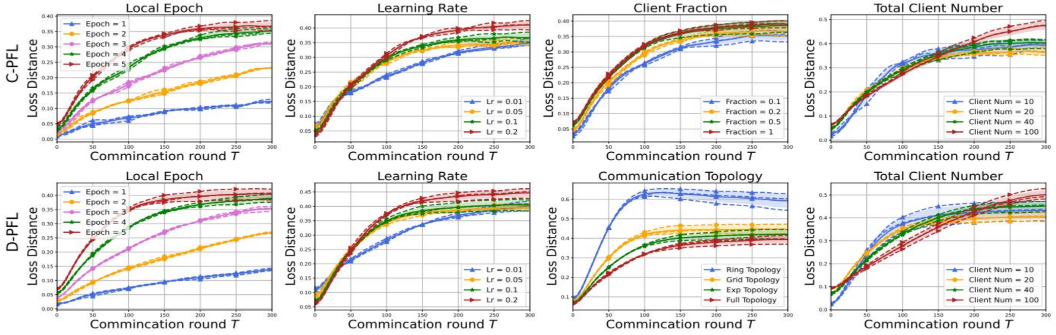
- CIFAR-10 datasets in the Dirichlet distribution (Non- IID α = 0.3) with ResNet-18,
- CIFAR-100 datasets in the Pathological distribution (Non-IID c = 20) with VGG-11 for

To verify the impacts of the key hyperparameters, we explore the impact of the four factors:

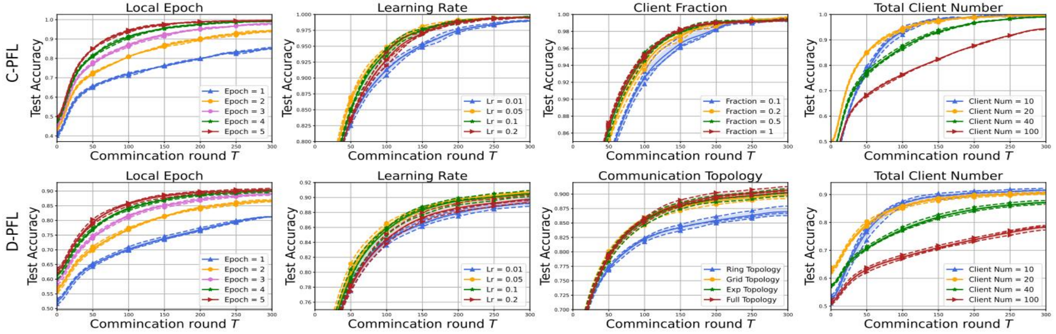
- Local Learning Epochs,
- Local Learning Rates,
- Client Fraction / Communication topology,
- Total Client Number.



(a) Disturbed loss distance of C-PFL and D-PFL on CIFAR-10.



(b) Testing and training loss distance of C-PFL and D-PFL on CIFAR-10.



(c) Testing accuracies of C-PFL and D-PFL on CIFAR-10.

CIFAR10 ResNet18

Experiments

CIFAR100 VGG

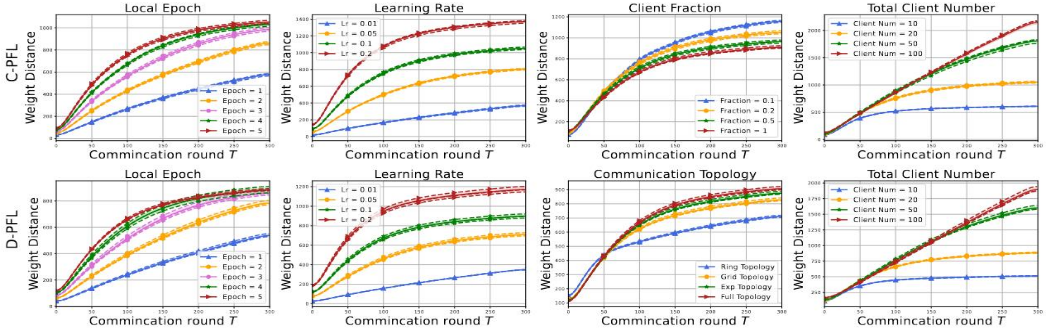


Figure 6: Loss distance of C-PFL and D-PFL on CIFAR-100.

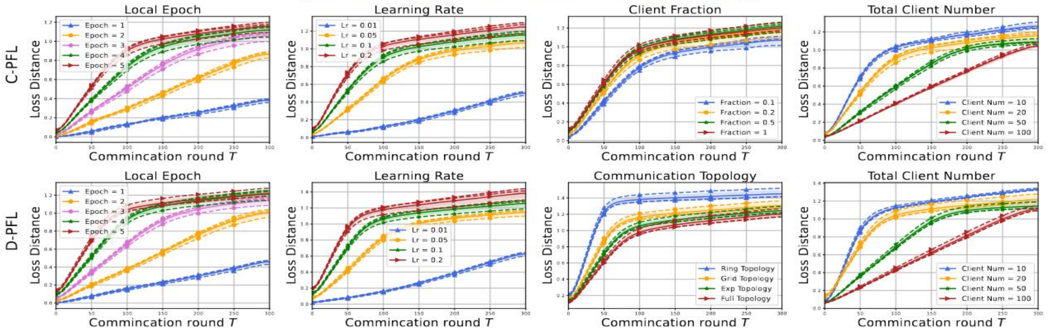


Figure 7: Training losses of C-PFL and D-PFL on CIFAR-100.

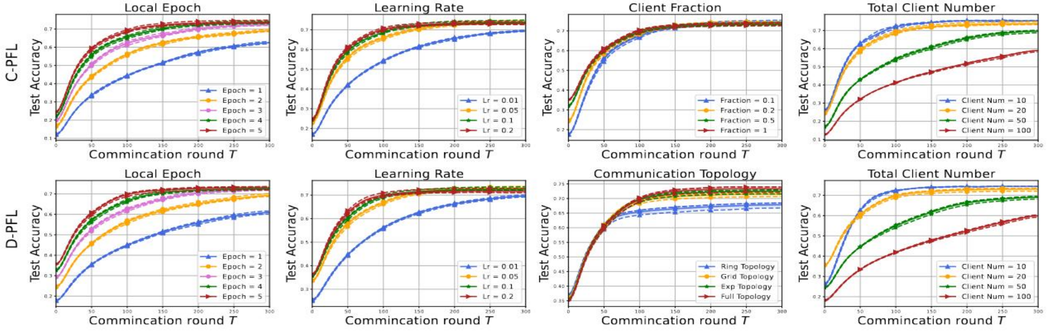


Figure 8: Testing accuracies of C-PFL and D-PFL on CIFAR-100.

- Both less local learning epochs and lower learning rates lead to better generalization performance, but they affect the convergence speed more seriously.

- More client participation and denser network connection in each communication round enlarge the generalization gap, but they speed up the convergence rate to the same extent.

- A larger total participation of clients and a smaller number of local training samples increase the generalization error and reduce the convergence speed simultaneously.

- C-PFL outperforms D-PFL in both generalization and convergence when their upper communication bandwidths are at the same level.

Future Works

- Improve the generalization bounds or less assumptions for PFL with the more advanced stability methods under convex, strongly-convex and non-convex conditions;

- Discuss the lower bound and tightness of the generalization of PFL to obtain the optimal training strategies for personalized training.

中山大學
SUN YAT-SEN UNIVERSITY

ICLR

Understanding the Stability-based Generalization of Personalized Federated Learning

Yingqi Liu, Qinglun Li, Tan Jie, Yifan Shi, Li Shen*, Xiaochun Cao

lyq@njust.edu.cn

Thank you!