Connecting Low-Loss Subspace for Personalized Federated Learning

Seok-Ju Hahn, Minwoo Jeong, Junghye Lee

Ulsan National Institute of Science and Technology (UNIST) & Kakao Enterprise

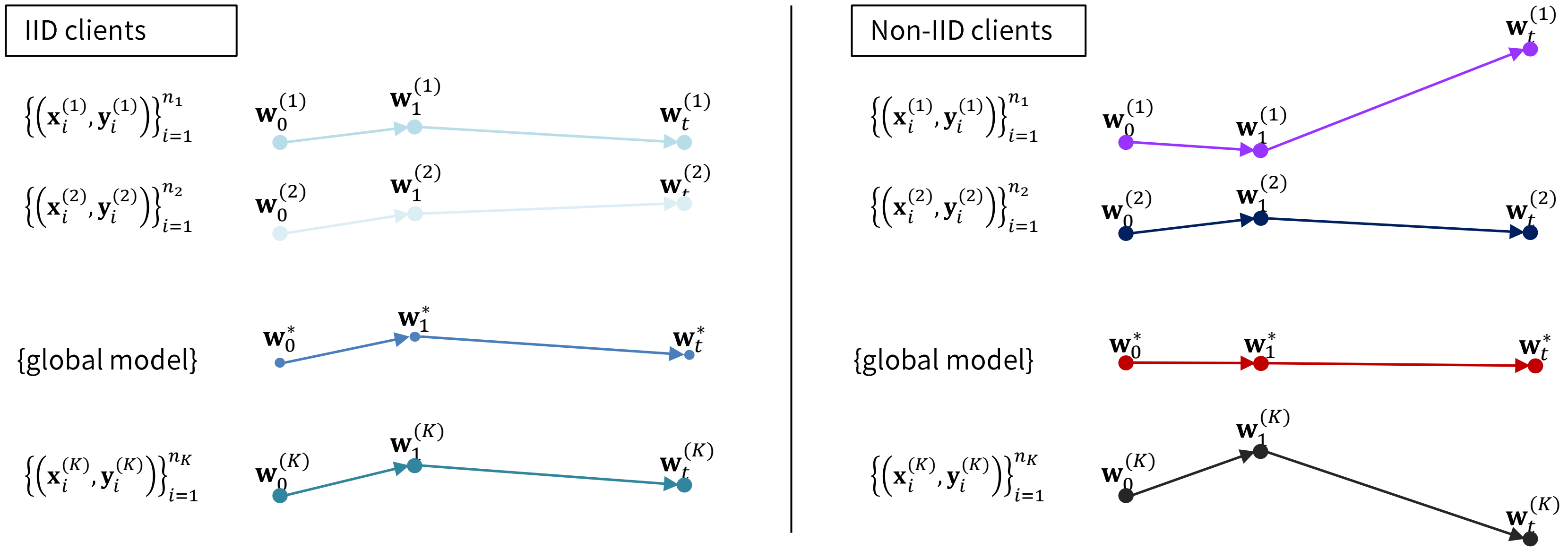
github.com/vaseline555/SuPerFed

Motivation

- **Problem in Federated Learning (FL)**
  - A main motivation of the participation in FL is an expectation to acquire a good global model that implicitly learned other clients' knowledge, in return for the training of a global model locally with client's own dataset.
  - When if the performance of the global model from FL is worse than a model trained solely with local dataset, there is NO need to participate in FL.
  - Early FL researches usually aimed to yield a decent single global model.
    - The main issue is to deal with the statistical heterogeneity (i.e., non-IIDness) across clients.

Motivation

- **Statistical Heterogeneity (i.e., non-IIDness) in FL**
  - Definition) Non-IIDness in FL – different clients have their own distinct data distribution.
    - $\left\{\left(\mathbf{x}_i^{(p)}, \mathbf{y}_i^{(p)}\right)\right\}_{i=1}^{n_p} \sim P_p, \left\{\left(\mathbf{x}_i^{(q)}, \mathbf{y}_i^{(q)}\right)\right\}_{i=1}^{n_q} \sim P_q \rightarrow P_p \neq P_q$
    - Divergence of optimization trajectories of locally updated models toward different directions
    - Hard to guarantee the convergence of a global model
    - Poor adaptation of the global model (i.e., no benefit of participation in federated learning)



Preliminaries

- **Personalized FL (PFL): more than a single global model**

  - Each client c_k has its own dataset $\mathcal{D}_k = \left\{\left(\mathbf{x}_i^{(k)}, \mathbf{y}_i^{(k)}\right)\right\}_{i=1}^{n_k} \sim \mathcal{P}_k$ with a model $\mathcal{W}_k$.
  - Assume a hypothesis $h \in \mathcal{H}$ can be learned by the objective $l: \mathcal{H} \times (\mathcal{X} \times \mathcal{Y}) \to \mathbb{R}^+$.
  - The expected loss of each client is $\mathcal{L}_{\mathcal{P}_k}(h_k, \mathcal{W}_k) = \mathbb{E}_{(\mathbf{x}^{(k)}, \mathbf{y}^{(k)}) \sim \mathcal{P}_k}\left[l\left(h_k\left(\mathbf{x}^{(k)}; \mathcal{W}_k\right), \mathbf{y}^{(k)}\right)\right]$, which is minimized by its empirical estimation $\hat{\mathcal{L}}_{\mathcal{D}_k}(h_k, \mathcal{W}_k) = \frac{1}{n_k} \sum_{i=1}^{n_k} l\left(h_k\left(\mathbf{x}_i^{(k)}; \mathcal{W}_k\right), \mathbf{y}_i^{(k)}\right)$.
  - Finally, the global objective of PFL is:

$$\min_{h_k \in \mathcal{H}} \frac{1}{\mathrm{K}} \sum_{k=1}^{\mathrm{K}} \mathcal{L}_{\mathcal{P}_k}(h_k, \mathcal{W}_k)$$

  - Through structural risk minimization, the global objective can be minimized through:

$$\min_{\mathcal{W}_1, \ldots, \mathcal{W}_\mathrm{K}} \frac{1}{\mathrm{K}} \sum_{k=1}^{\mathrm{K}} \hat{\mathcal{L}}_{\mathcal{D}_k}(h_k, \mathcal{W}_k) = \frac{1}{\mathrm{K}} \sum_{k=1}^{\mathrm{K}} \left\{ \frac{1}{n_k} \sum_{i=1}^{n_k} l\left(h_k\left(\mathbf{x}_i^{(k)}; \mathcal{W}_k\right), \mathbf{y}_i^{(k)}\right) + \Omega(\mathcal{W}_k) \right\}$$

where $\Omega(\cdot)$ is a regularization term.

Preliminaries

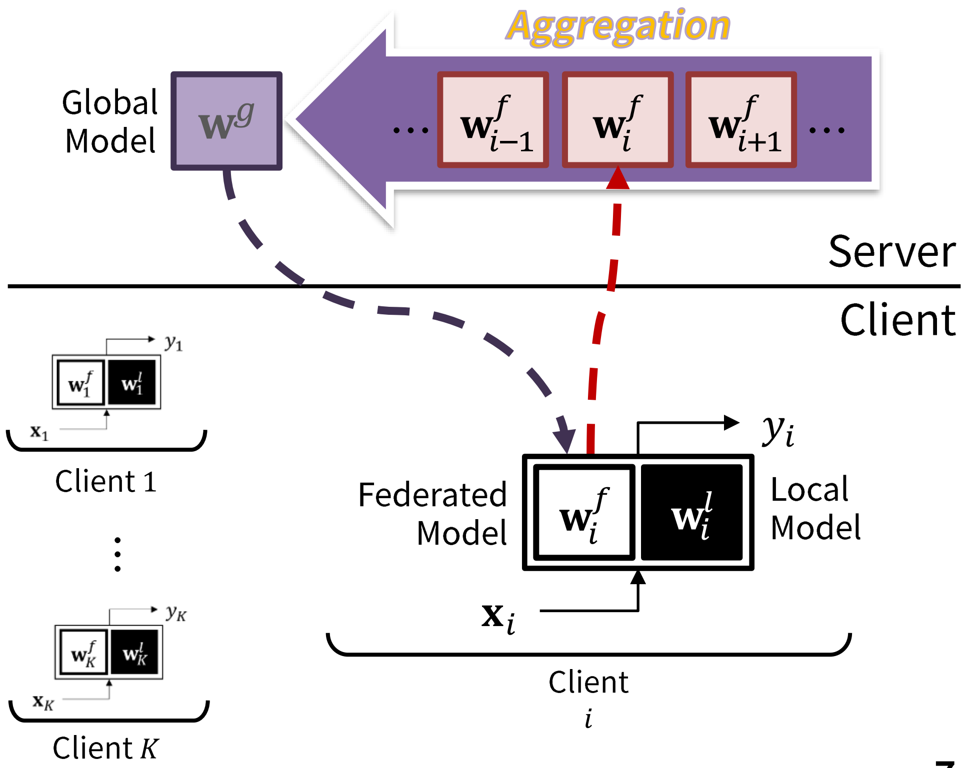
- **Personalized FL (PFL): more than a single global model**
  - Multi-task learning) training multiple models for tackling multiple target distributions
  - Model mixture) mixing local models (or local model's layers) with a global model (or global model's layers)
  - Meta learning) optimizing global model for fast adaptation as a local model in each client
  - Clustering) applying FL within the same cluster constructed by implicit information in client's model
  - Knowledge distillation) distilling local model's knowledge to the global model
  - Optimization variants) regularize the local model not to be far away from the global model, removing harmful oscillation when using momentum, adopting dampening variables, etc.

Preliminaries

- **Personalized FL (PFL): more than a single global model**
  - Multi-task learning) training multiple models for tackling multiple target distributions
  - Model mixture) mixing local models (or local model's layers) with a global model (or global model's layers)
  - Meta learning) optimizing global model for fast adaptation as a local model in each client
  - Clustering) applying FL within the same cluster constructed by implicit information in client's model
  - Knowledge distillation) distilling local model's knowledge to the global model
  - Optimization variants) regularize the local model not to be far away from the global model, removing harmful oscillation when using momentum, adopting dampening variables, etc.

Preliminaries

- **Model mixture-based PFL**
  - Global model ($\mathbf{w}^g$): a model aggregated at the central server from updated federated models.
  - Local model ($\mathbf{w}_i^l$): a local model for personalization thereby never be uploaded during the whole learning process.
  - Federated model ($\mathbf{w}_i^f$): a global model transmitted to and updated in the client.
  - $\mathcal{W}_i = G(\mathbf{w}_i^l, \mathbf{w}_i^f)$ in model-mixture based PFL.
  - $G(\cdot)$ is a grouping operator which can be:
    - A simple concatenation (e.g., FedPer, LG-FedAvg, FedRep)
    - A simple enumeration (e.g., pFedMe, Ditto)
    - A convex combination $G(\mathbf{w}_i^l, \mathbf{w}_i^f) = (1-\lambda)\mathbf{w}_i^f + \lambda \mathbf{w}_i^l, \lambda \in \mathbb{R}^{[0,1]}$ (e.g., APFL)
  - Let $\mathcal{W}_i(\lambda) \triangleq (1-\lambda)\mathbf{w}_i^f + \lambda \mathbf{w}_i^l$.



Preliminaries

- **Mode connectivity: existence of the connected path between two deep networks**
  - Why and when ensembled deep networks are working well?
    (Garipov et al., 2018; Draxler et al., 2018; Fort et al., 2019; Frankle et al., 2020)
    - Two deep networks having the same structure trained on the same dataset with different configurations (e.g., different initialization) may reach at different local optimum.
    - It is empirically studied that the two different local optimum can be connected in a path, and all solutions on the path have low-loss!
      (i.e., different optima can be connected to be a flat minima; a.k.a. mode connectivity)
    - If the optimization trajectory is similar (i.e., trained from the same initialization), cosine similarity between model weights are high.
    - When the optimization trajectory is dissimilar, the ensemble performance is higher.
    - Such a low-loss connected path can be discovered in various ways.
      (e.g., Fast Geometric Ensemble, Stochastic Weight Averaging, etc.)

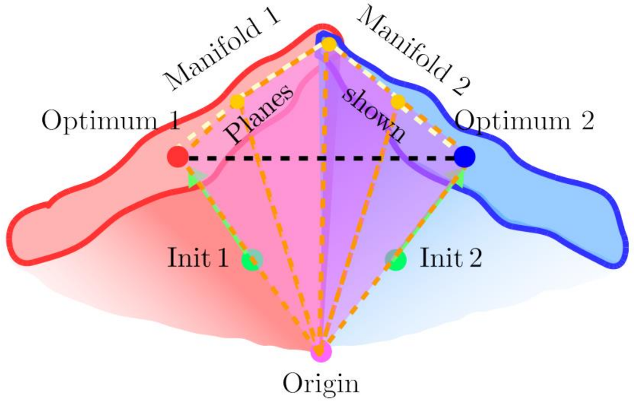


Figure adapted from Fort et al., 2019

Preliminaries

- **Inducing a connected path between two deep networks**
  - Learning Neural Network Subspaces (Wortsman et al., 2021)
    - Existing works inducing a mode connectivity requires extra trainings of the model or extra spaces for saving trained copies of the model.
    - This work proposed a simple add-on that inducing mode connectivity between weight spaces of two different deep networks.
      - Just minimizing the cosine similarity between two deep networks to be zero!
    - Thereby, a wide and flat minima can be recovered.

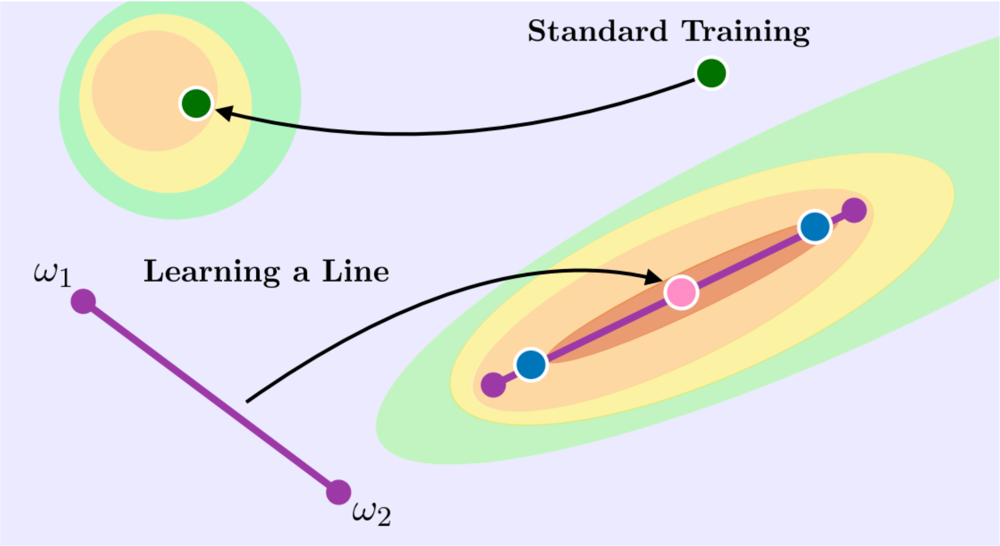


Figure adapted from Wortsman et al., 2021

Proposed Method

- **SuPerFed**
  - (Connected low-loss) **Su**bspace learning for **Per**sonalized **Fed**erated Learning (**SuPerFed**)
    - Challenge 1) Per model-mixture based PFL, can we also induce mode connectivity between a local and a global model?
      - Different from the ensemble learning, FL requires small steps of model update and not all data samples are accessible in the training time.
    - Challenge 2) If so, can this improve the performance of model-mixture based personalized federated learning?
      - Different from the ensemble learning, one of deep networks (i.e., global model) participating in the ensemble can be implicitly trained on other unseen distributions via aggregation in the central server, while the other (i.e., local model) can only see the local distribution as a personalized model.
    - Challenge 3) Can other benefits in the ensemble learning be equivalently adopted to PFL?
      - Ensembled deep networks usually show robustness to the label noise, while plain FL (including PFL) methods are not.

Proposed Method

- **Overview**

Algorithm 1 LocalUpdate

Inputs: global model from the server, w^g, batch size B, number of local epochs E, current round r, start round of personalization L, learning rate η, regularization constants μ, ν, client dataset $\mathcal{D}$.
Start: set the federated model as: $\mathrm{w}^f \leftarrow \mathrm{w}^g$.
if the local model w^l does not exist **then**
  Set the local model as: $\mathrm{w}^l \leftarrow \mathrm{w}^g$.
end if
for $e = 0, ..., \mathrm{E}-1$ **do**
  $\mathcal{B}_e \leftarrow$Split the client dataset $\mathcal{D}$ into batches of size B.
  for a local batch $(\mathrm{x}, \mathrm{y}) \in \mathcal{B}_e$ **do**
    if $r < \mathrm{L}$ **then**
      Set $\lambda = 0$
    else
      Sample $\lambda \sim \mathrm{Unif}(0, 1)$.
    end if
    Mix models $\mathcal{W}(\lambda) = (1-\lambda)\mathrm{w}^f + \lambda \mathrm{w}^l$.
    Set the local objective $\mathcal{L}$ using (3) and (4).
    Minimize $\mathcal{L}$ in terms of w^f (5) and w^l (6) each through $\mathcal{W}(\lambda)$ using SGD with the learning rate η.
  end for
end for
Return: updated federated model w^f.

Algorithm 2 SuPerFed

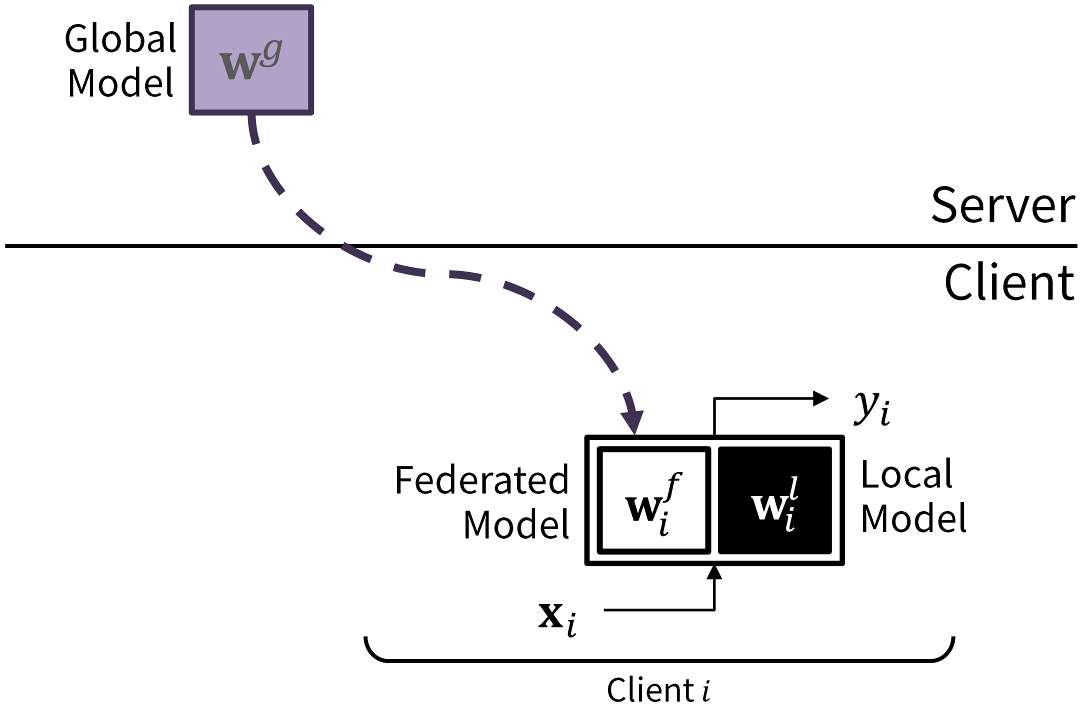
Inputs: batch size B, number of local epochs E, total communication rounds R, start round of personalization L, learning rate η, regularization constants μ, ν, number of clients K, fraction of clients to be sampled C, clients c_i having own dataset $\mathcal{D}_i = \{(\mathrm{x}_i^j, \mathrm{y}_i^j)\}_{j=1}^{n_i}, i \in [\mathrm{K}]$
Start: Server initializes a global model $\mathrm{w}^{g,0}$.
for $r = 0, .., \mathrm{R}-1$ **do**
  Server randomly selects $\max(\mathrm{C} \cdot \mathrm{K}, 1)$ clients as S_r.
  Server broadcasts the current global model $\mathrm{w}^{g,r}$ to S_r.
  for each client $c_i \in S_r$ **in parallel do**
    $\mathrm{w}_i^{f,r} \leftarrow \mathrm{LocalUpdate}(\mathrm{w}^{g,r}, \mathrm{B}, \mathrm{E}, r, \mathrm{L}, \eta, \mu, \nu)$
  end for
  Update a global model:
  $\mathrm{w}^{g,r+1} \leftarrow \frac{1}{\sum_{i \in S_r} n_i} \sum_{i \in S_r} n_i \mathrm{w}_i^{f,r}$.
end for

Proposed Method

- **Process**
  - Initialize the global model at the central server.

Proposed Method

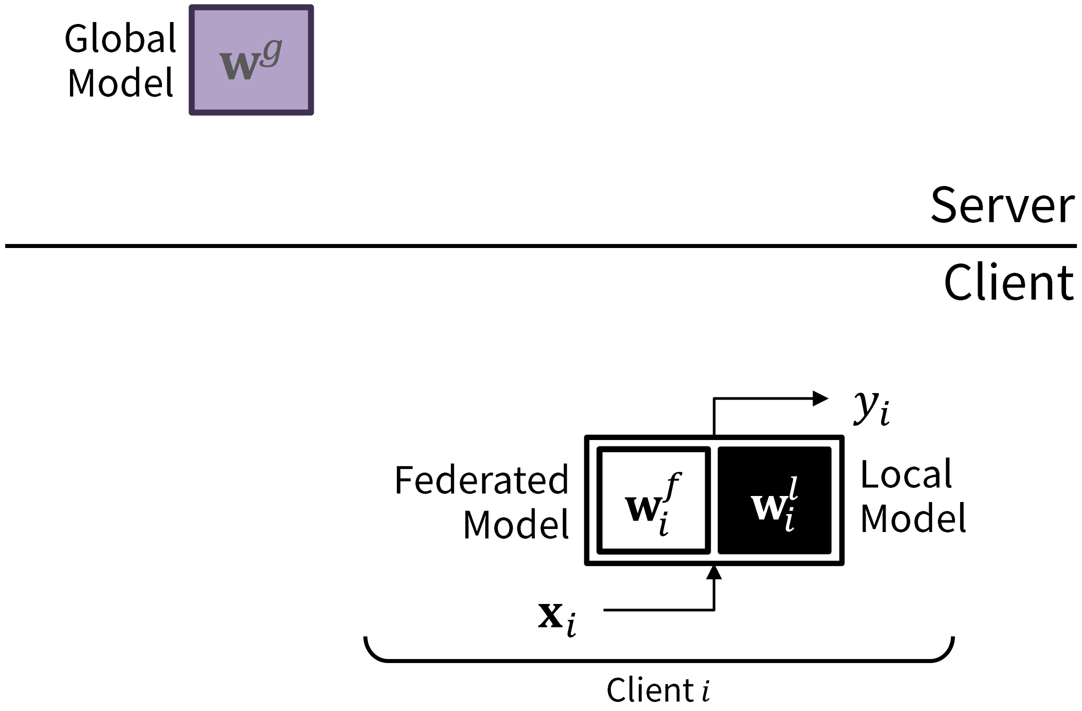
- **Process**
  - Transmit the global model to participating clients and request them to update the model with their own dataset.



* Note that the federated model and the local model are the SAME copy of the global model.

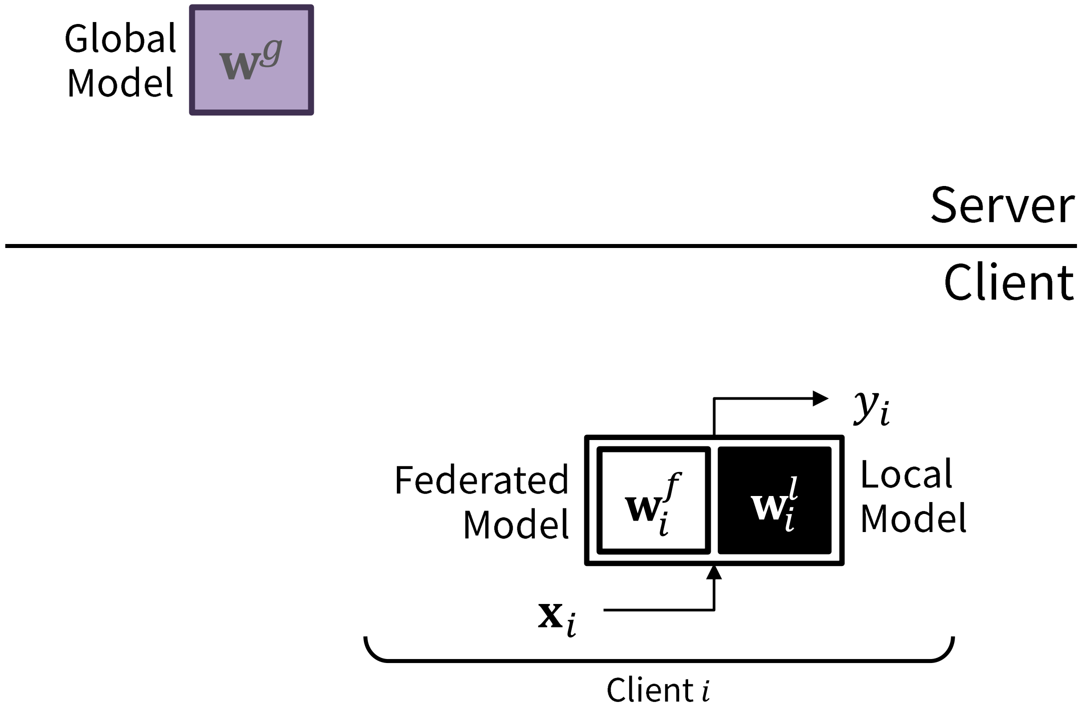
Proposed Method

- **Process**
  - Local update



Proposed Method

- **Process**
  - Local update

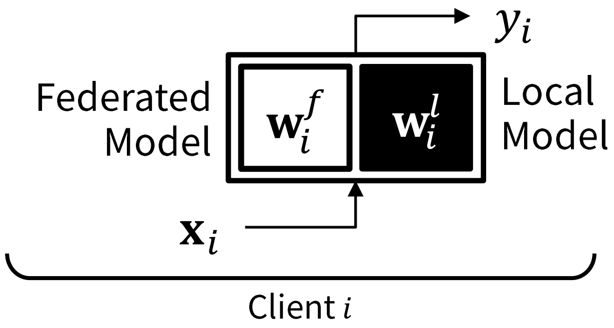


Proposed Method

- **Process**
  - Local update: randomly sample λ from the uniform distribution in every batch update.

$\lambda \sim \mathrm{Unif}(0,1)$

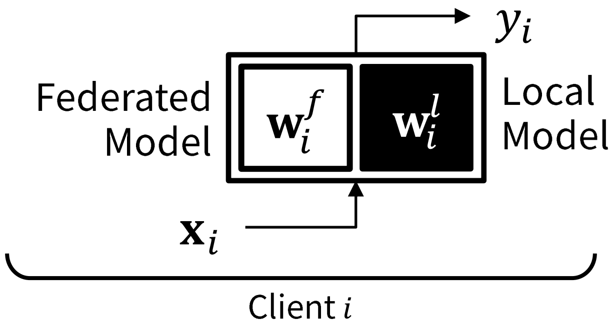
Client

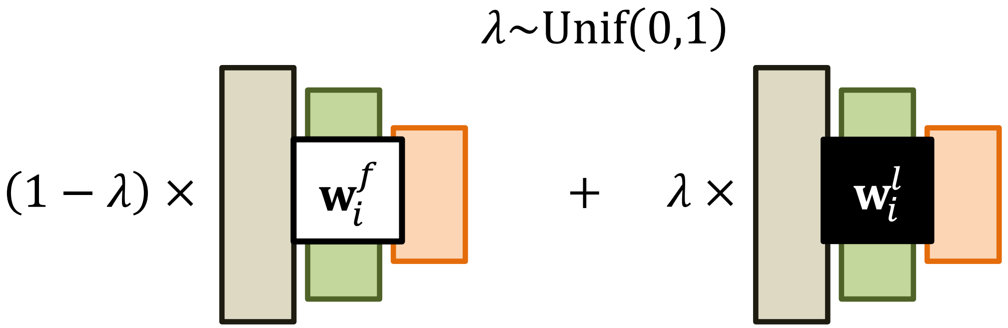


Proposed Method

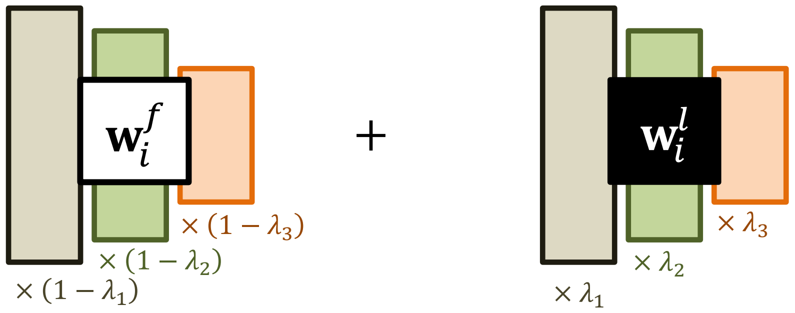
- **Process**
  - Local update: mix the federated model and the local model using λ.
    (i.e., convex combination of both models)

Client





Model-mixing (SuPerFed-MM)

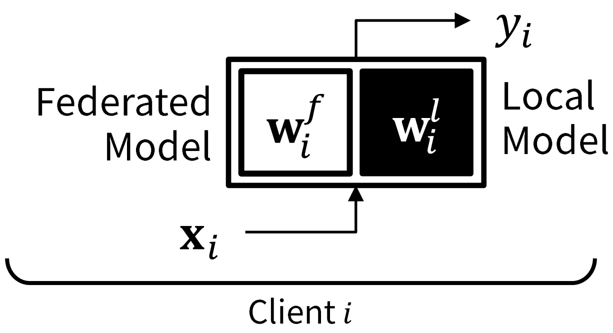


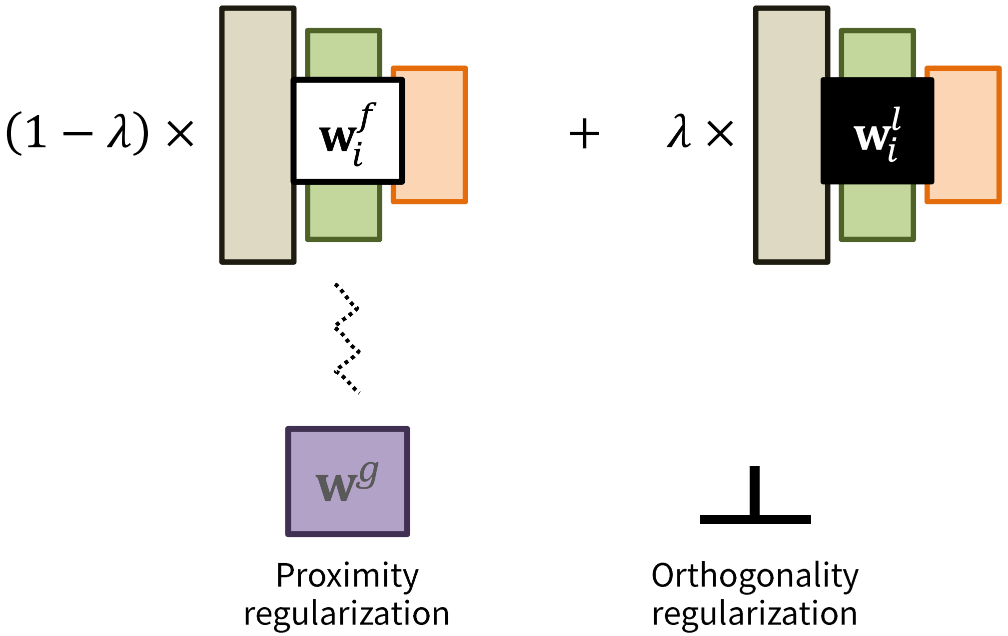
Layer-mixing (SuPerFed-LM)

Proposed Method

- **Process**
  - Local update: update with proximity (towards previous round's global model; controlled by μ) and orthogonality (for inducing mode connectivity; controlled by ν) regularization terms.

Client





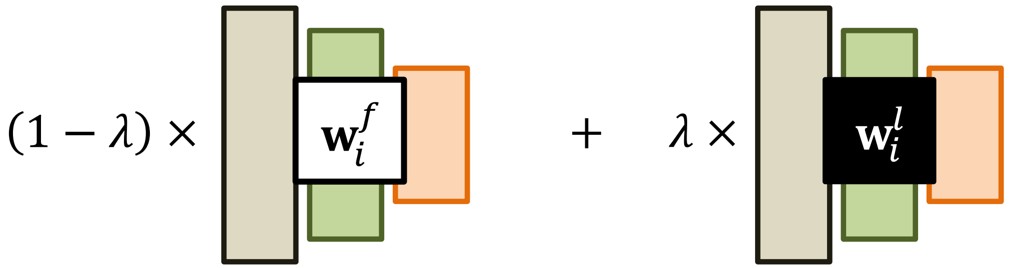
$\mu\|\mathbf{w}^f - \mathbf{w}^g\|^2$ $\quad$ $\nu\cos^2(\mathbf{w}^f, \mathbf{w}^l)$

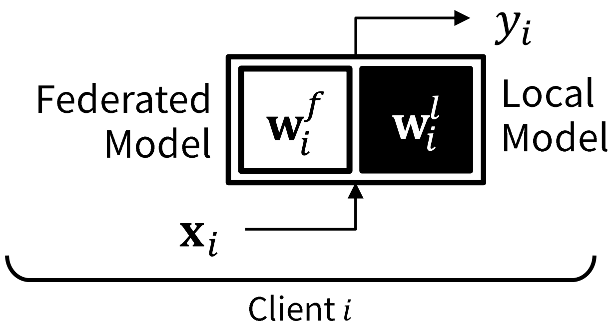
$$\mathcal{L}(h, \mathcal{W}) = l\big(h(\mathbf{x}), \mathbf{y}; \mathcal{W}(\lambda)\big) + \mu\|\mathbf{w}^f - \mathbf{w}^g\|^2 + \nu\cos^2(\mathbf{w}^f, \mathbf{w}^l)$$

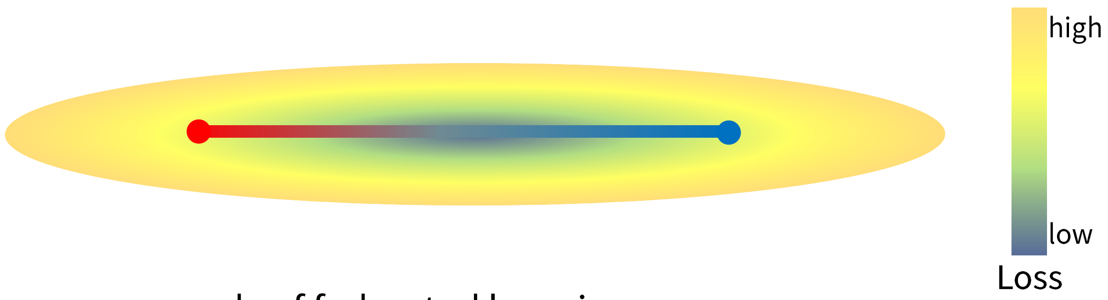
Proposed Method

- **Process**
  - As a result of the local update, the mode connectivity between the federated and the local model is induced!

Client







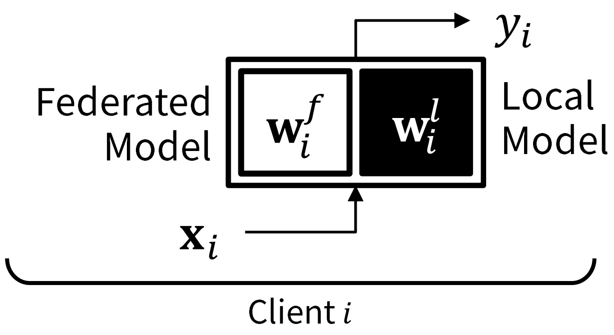
* Note that the mixing process can be started after some rounds of federated learning.

(i.e., Phase I: learning only a global model / Phase II: learning both global and local model while inducing the mode connectivity)

Proposed Method

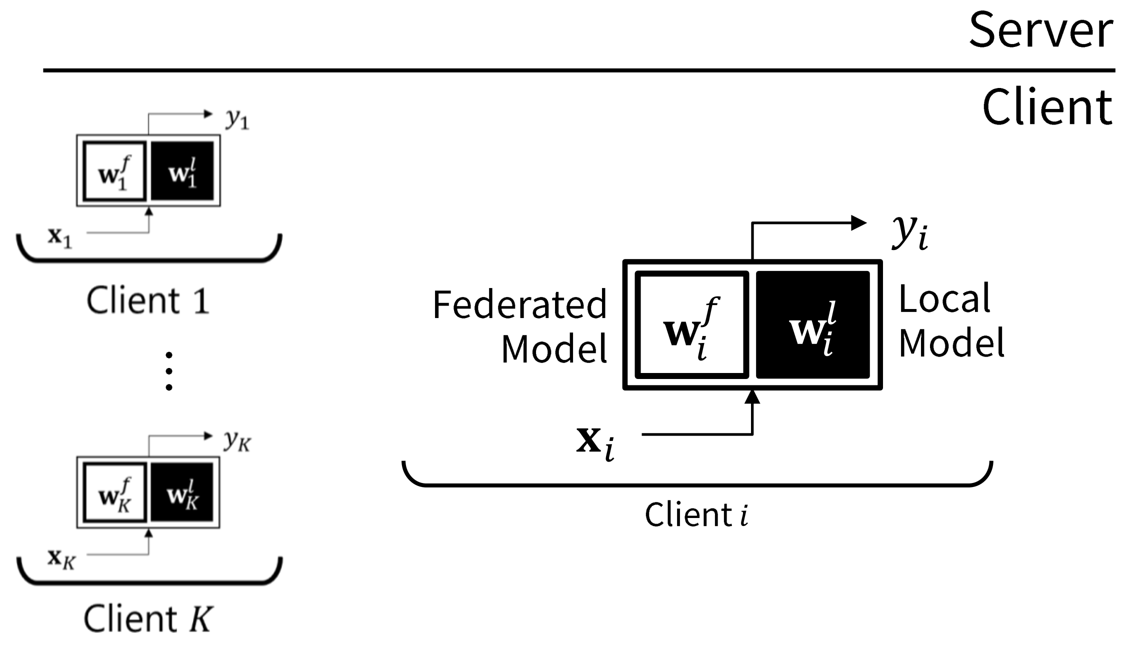
- **Process**
  - Per each round, selected clients are updated in parallel.

Client



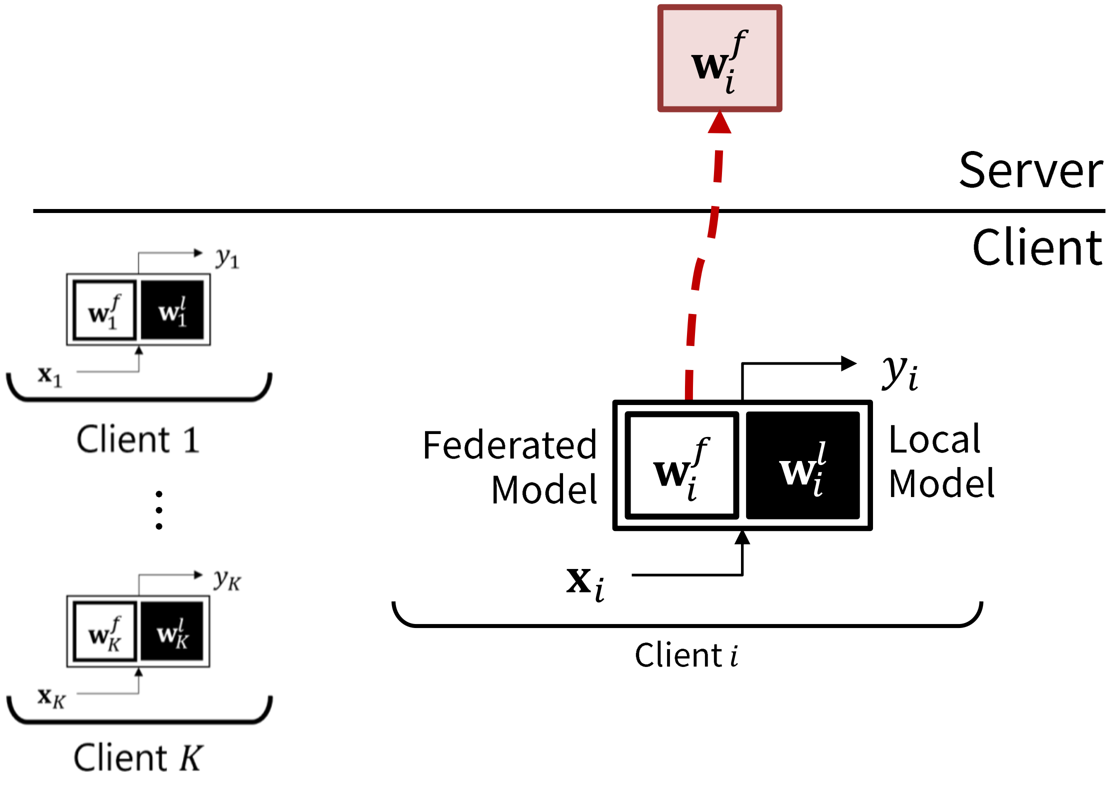
Proposed Method

- **Process**
  - Per each round, selected clients are updated in parallel.



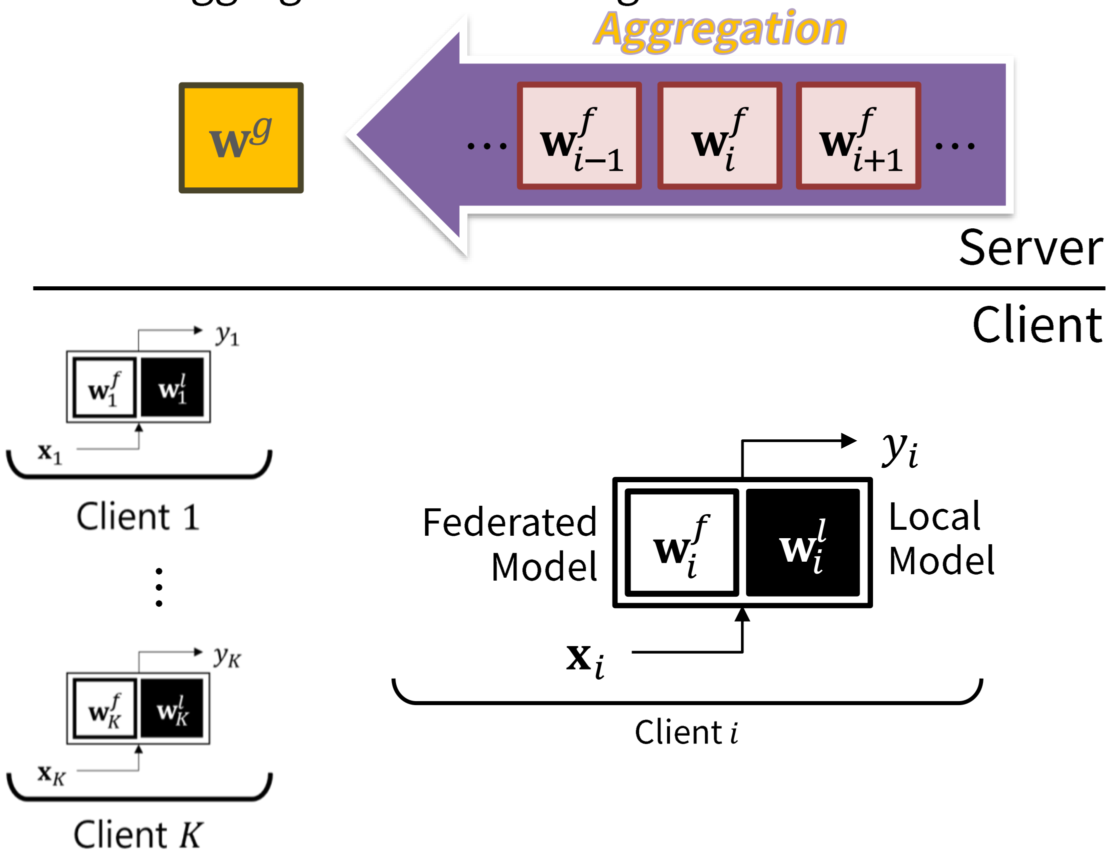
Proposed Method

- **Process**
  - Clients upload the federated model only to the server.



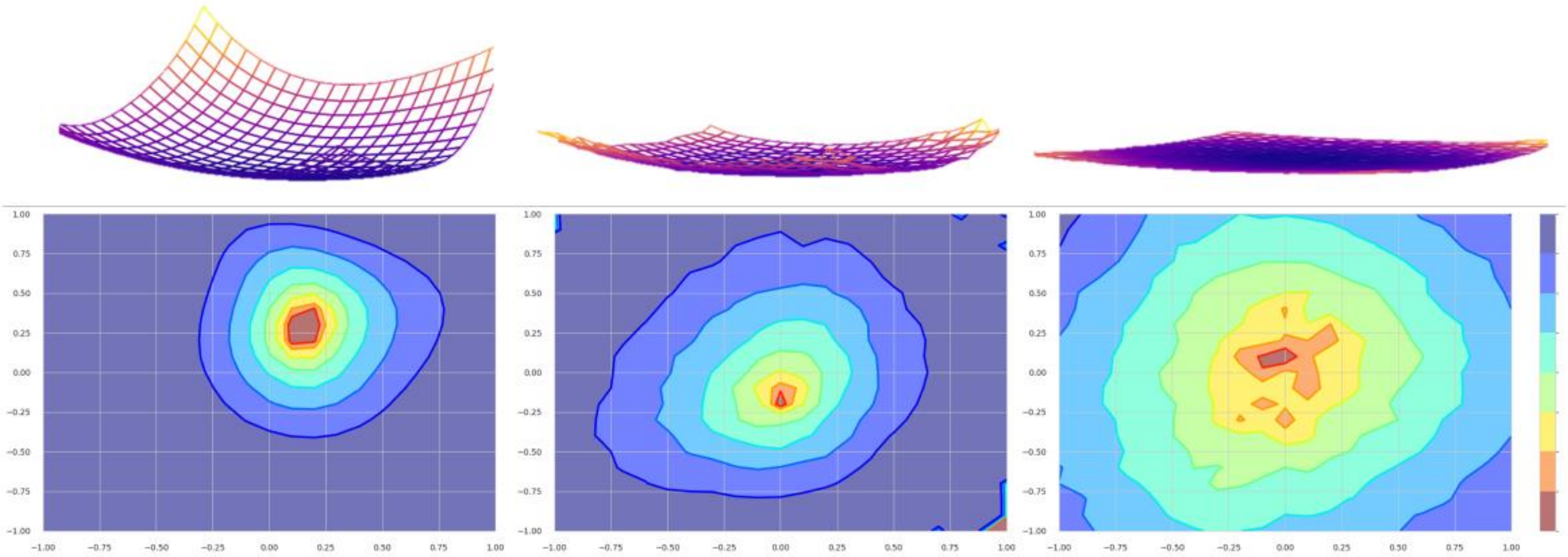
Proposed Method

- **Process**
  - Uploaded federated models are aggregated into a new global model.



Experimental Results

- **Challenge 1**
  - ➢ Per model-mixture based PFL, can we also induce mode connectivity between a local and a global model?
  - Visualization of the loss surface of a global model trained through FedAvg, SuPerFed-MM, and SuPerFed-LM. (CIFAR10 dataset in pathological non-IID setting with 100 clients)
    - Dataset used for this visualization has never been used for training, a separate test dataset at the central server.



FedAvg SuPerFed-MM SuPerFed-LM

Experimental Results

- **Challenge 2**
  - ➢ If so, can this improve the performance of model-mixture based personalized federated learning?
  - Evaluation of personalization performance in three settings (sampled 5 clients randomly per round)
    - Pathological non-IID setting (adapted from McMahan et al., 2016)
      - MNIST, CIFAR10
      - 50 / 100 / 500 clients
    - Dirichlet distribution-based non-IID setting (adapted from Hsu et al., 2019)
      - CIFAR100, TinyImageNet
      - 100 / 200 clients
      - α=1 / 10 / 100
    - LEAF benchmark (realistic scenario; adapted from Caldas et al., 2018)
      - FEMNIST, Shakespeare
      - 770 / 660 clients

Experimental Results

- **Personalization performance**
  - Pathological non-IID setting (adapted from McMahan et al., 2016)

Dataset	MNIST			CIFAR10		
# clients	50	100	500	50	100	500
# samples	960	480	96	800	400	80
FedAvg	95.69 ± 2.39	89.78 ± 11.30	96.04 ± 4.74	43.09 ± 24.56	36.19 ± 29.54	47.90 ± 25.05
FedProx	95.13 ± 2.67	93.25 ± 6.12	96.50 ± 4.52	49.01 ± 19.87	38.56 ± 28.11	48.60 ± 25.71
SCAFFOLD	95.50 ± 2.71	90.58 ± 10.13	96.60 ± 4.26	43.81 ± 24.30	36.31 ± 29.42	40.27 ± 26.90
LG-FedAvg	98.21 ± 1.28	97.52 ± 2.11	96.05 ± 5.02	89.03 ± 4.53	70.25 ± 35.66	78.52 ± 11.22
FedPer	99.23 ± 0.66	99.14 ± 0.93	98.67 ± 2.61	89.10 ± 5.41	87.99 ± 5.70	82.35 ± 9.85
APFL	99.40 ± 0.58	99.19 ± 0.92	98.98 ± 2.22	92.83 ± 3.47	91.73 ± 4.61	87.38 ± 9.39
pFedMe	81.10 ± 8.52	82.48 ± 7.62	81.96 ± 12.28	92.97 ± 3.07	92.07 ± 5.05	88.30 ± 8.53
Ditto	97.07 ± 1.38	97.13 ± 2.06	97.20 ± 3.72	85.53 ± 6.22	83.01 ± 5.62	84.45 ± 10.67
FedRep	99.11 ± 0.63	99.04 ± 1.02	97.94 ± 3.37	82.00 ± 5.41	81.27 ± 7.90	80.66 ± 11.00
SuPerFed-MM	99.45 ± 0.46	99.38 ± 0.93	99.24 ± 2.12	94.05 ± 3.18	93.25 ± 3.80	90.81 ± 9.35
SuPerFed-LM	99.48 ± 0.54	99.31 ± 1.09	98.83 ± 3.02	93.88 ± 3.55	93.20 ± 4.19	89.63 ± 11.11

Experimental Results

- **Personalization performance**
  - Dirichlet distribution-based non-IID setting (adapted from Hsu et al., 2019)

Dataset	CIFAR100			TinyImageNet		
# clients	100			200		
concentration (α)	1	10	100	1	10	100
FedAvg	58.12 ± 7.06	59.04 ± 7.19	58.49 ± 5.27	46.61 ± 5.64	48.90 ± 5.50	48.90 ± 5.40
FedProx	57.71 ± 6.79	58.24 ± 5.94	58.75 ± 5.56	47.37 ± 5.94	47.73 ± 5.94	48.97 ± 5.02
SCAFFOLD	51.16 ± 6.79	51.40 ± 5.22	52.90 ± 4.89	46.54 ± 5.49	48.77 ± 5.49	48.27 ± 5.32
LG-FedAvg	28.88 ± 5.64	21.25 ± 4.64	20.05 ± 4.61	14.70 ± 3.84	9.86 ± 3.13	9.25 ± 2.89
FedPer	46.78 ± 7.63	35.73 ± 6.80	35.52 ± 6.58	21.90 ± 4.71	11.10 ± 3.19	9.63 ± 3.12
APFL	61.13 ± 6.86	56.90 ± 7.05	55.43 ± 5.45	41.98 ± 5.94	34.74 ± 5.14	34.23 ± 5.07
pFedMe	19.00 ± 5.37	17.94 ± 4.72	18.28 ± 3.41	6.05 ± 2.84	8.01 ± 2.92	7.69 ± 2.41
Ditto	60.04 ± 6.82	58.55 ± 7.12	58.73 ± 5.39	46.36 ± 5.44	43.84 ± 5.44	43.11 ± 5.35
FedRep	38.49 ± 6.65	26.61 ± 5.20	24.50 ± 4.21	18.67 ± 4.66	9.23 ± 2.84	8.09 ± 2.83
SuPerFed-MM	60.14 ± 6.24	58.32 ± 6.25	59.08 ± 5.12	50.07 ± 5.73	49.86 ± 5.03	49.73 ± 4.84
SuPerFed-LM	62.50 ± 6.34	61.64 ± 6.23	59.05 ± 5.59	47.28 ± 5.19	48.98 ± 4.79	49.29 ± 4.82

Experimental Results

- **Personalization performance**
  - LEAF benchmark (realistic scenario; adapted from Caldas et al., 2018)

Dataset	FEMNIST		Shakespeare	
# clients	730		660	
Accuracy	Top-1	Top-5	Top-1	Top-5
FedAvg	80.12 ± 12.01	98.74 ± 2.97	50.90 ± 7.85	80.15 ± 7.87
FedProx	80.23 ± 11.88	98.73 ± 2.94	51.33 ± 7.54	80.31 ± 6.95
SCAFFOLD	80.03 ± 11.78	98.85 ± 2.77	50.76 ± 8.01	80.43 ± 7.09
LG-FedAvg	50.84 ± 20.97	75.11 ± 21.49	33.88 ± 10.28	62.84 ± 13.16
FedPer	73.79 ± 14.10	86.39 ± 14.70	45.82 ± 8.10	75.68 ± 9.25
APFL	84.85 ± 8.83	98.83 ± 2.73	54.08 ± 8.31	83.32 ± 6.22
pFedMe	5.98 ± 4.55	24.64 ± 9.43	32.29 ± 6.64	63.12 ± 8.00
Ditto	64.61 ± 31.49	81.14 ± 28.56	49.04 ± 10.22	78.14 ± 12.61
FedRep	59.27 ± 15.72	70.42 ± 15.82	38.15 ± 9.54	68.65 ± 12.50
SuPerFed-MM	85.20 ± 8.40	99.16 ± 2.13	54.52 ± 7.54	84.27 ± 6.00
SuPerFed-LM	83.36 ± 9.61	98.81 ± 2.58	54.52 ± 7.54	83.97 ± 5.72

Experimental Results

- **Challenge 3**
  - ➢ Can other benefits of the ensemble learning be equivalently adopted to PFL?
  - Evaluation of personalization performance with simulated label-noise
    - Pathological non-IID setting
      on MNIST and CIFAR10 dataset with 100 clients
    - Pairwise flipping for label noise (ratio = 0.1/0.4)
      and symmetric flipping label noise (ratio = 0.2/0.6)
    - Showed low expected calibration error (ECE) and
      low maximum calibration error (MCE).
      (Proposed in Guo et al., 2017)

Dataset	MNIST				CIFAR10			
Noise type	pair		symmetric		pair		symmetric	
Noise ratio	0.1	0.4	0.2	0.6	0.1	0.4	0.2	0.6
FedAvg	0.17 ± 0.03 0.58 ± 0.08 (82.40 ± 3.31)	0.38 ± 0.03 0.67 ± 0.04 (41.01 ± 4.64)	0.29 ± 0.04 0.66 ± 0.07 (66.94 ± 4.27)	0.42 ± 0.04 0.75 ± 0.05 (49.52 ± 5.42)	0.46 ± 0.04 0.80 ± 0.06 (45.08 ± 5.61)	0.57 ± 0.04 0.87 ± 0.04 (20.90 ± 4.66)	0.52 ± 0.05 0.81 ± 0.04 (38.62 ± 5.28)	0.59 ± 0.05 0.84 ± 0.04 (30.18 ± 5.53)
FedProx	0.17 ± 0.03 0.58 ± 0.07 (82.05 ± 3.98)	0.38 ± 0.03 0.78 ± 0.05 (41.39 ± 4.61)	0.29 ± 0.03 0.66 ± 0.07 (67.15 ± 4.60)	0.42 ± 0.05 0.74 ± 0.05 (49.98 ± 5.57)	0.47 ± 0.05 0.80 ± 0.05 (44.31 ± 6.20)	0.70 ± 0.05 0.87 ± 0.04 (21.58 ± 4.62)	0.53 ± 0.05 0.81 ± 0.05 (36.91 ± 5.68)	0.59 ± 0.05 0.84 ± 0.05 (29.50 ± 6.11)
SCAFFOLD	0.16 ± 0.03 0.58 ± 0.07 (60.86 ± 4.09)	0.45 ± 0.04 0.77 ± 0.04 (44.92 ± 5.07)	0.29 ± 0.04 0.59 ± 0.08 (70.51 ± 4.25)	0.44 ± 0.04 0.73 ± 0.05 (51.46 ± 5.12)	0.46 ± 0.05 0.76 ± 0.05 (47.54 ± 5.64)	0.65 ± 0.04 0.86 ± 0.03 (22.72 ± 4.01)	0.53 ± 0.04 0.79 ± 0.04 (38.85 ± 5.45)	0.61 ± 0.04 0.83 ± 0.04 (30.18 ± 4.63)
LG-FedAvg	0.23 ± 0.04 0.66 ± 0.08 (73.65 ± 5.32)	0.50 ± 0.05 0.81 ± 0.04 (37.69 ± 5.41)	0.34 ± 0.04 0.71 ± 0.07 (61.54 ± 4.96)	0.45 ± 0.05 0.75 ± 0.06 (47.79 ± 5.02)	0.59 ± 0.06 0.83 ± 0.05 (30.22 ± 6.49)	0.69 ± 0.05 0.89 ± 0.03 (17.44 ± 4.66)	0.63 ± 0.05 0.85 ± 0.04 (25.68 ± 5.18)	0.66 ± 0.05 0.87 ± 0.03 (22.04 ± 5.56)
FedPer	0.17 ± 0.03 0.57 ± 0.08 (82.43 ± 4.18)	0.40 ± 0.04 0.78 ± 0.05 (40.75 ± 5.49)	0.28 ± 0.04 0.66 ± 0.08 (68.44 ± 5.63)	0.40 ± 0.04 0.73 ± 0.07 (52.41 ± 5.56)	0.54 ± 0.05 0.81 ± 0.06 (39.81 ± 6.02)	0.70 ± 0.05 0.87 ± 0.04 (20.37 ± 5.43)	0.60 ± 0.06 0.82 ± 0.05 (32.82 ± 5.90)	0.65 ± 0.05 0.85 ± 0.04 (26.82 ± 5.38)
APFL	0.18 ± 0.03 0.60 ± 0.07 (80.18 ± 4.57)	0.42 ± .0.05 0.78 ± 0.06 (40.43 ± 6.06)	0.28 ± 0.05 0.67 ± 0.06 (67.83 ± 5.53)	0.40 ± 0.05 0.79 ± 0.07 (52.15 ± 5.63)	0.45 ± 0.06 0.78 ± 0.06 (45.27 ± 6.21)	0.60 ± 0.06 0.86 ± 0.04 (23.42 ± 5.49)	0.51 ± 0.06 0.81 ± 0.05 (37.85 ± 5.84)	0.56 ± 0.06 0.83 ± 0.05 (31.46 ± 5.73)
pFedMe	0.23 ± 0.04 0.66 ± 0.07 (72.05 ± 0.05)	0.46 ± 0.04 0.80 ± .0.05 (37.89 ± 5.77)	0.33 ± 0.05 0.72 ± 0.06 (59.80 ± 5.94)	0.44 ± 0.04 0.76 ± 0.05 (45.79 ± 5.26)	0.58 ± 0.06 0.83 ± 0.52 (29.62 ± 6.37)	0.66 ± 0.04 0.88 ± 0.03 (17.94 ± 3.97)	0.60 ± 0.05 0.85 ± 0.04 (26.01 ± 0.05)	0.64 ± 0.05 0.87 ± 0.03 (21.36 ± 4.42)
Ditto	0.22 ± 0.04 0.66 ± 0.07 (72.39 ± 5.25)	0.42 ± 0.05 0.79 ± 0.05 (38.20 ± 6.13)	0.31 ± 0.04 0.69 ± 0.07 (60.62 ± 5.67)	0.41 ± 0.04 0.77 ± 0.05 (45.11 ± 4.95)	0.54 ± 0.05 0.84 ± 0.05 (29.41 ± 5.15)	0.60 ± 0.06 0.88 ± 0.04 (18.17 ± 4.43)	0.56 ± 0.07 0.85 ± 0.04 (26.39 ± 5.83)	0.58 ± 0.06 0.87 ± 0.04 (21.72 ± 4.90)
FedRep	0.20 ± 0.04 0.62 ± 0.08 (77.95 ± 4.65)	0.53 ± 0.05 0.81 ± 0.05 (35.88 ± 5.15)	0.30 ± 0.05 0.68 ± 0.08 (66.58 ± 5.60)	0.43 ± 0.05 0.74 ± 0.06 (51.30 ± 5.27)	0.62 ± 0.05 0.82 ± 0.05 (33.10 ± 5.23)	0.76 ± 0.04 0.87 ± 0.03 (17.80 ± 4.43)	0.66 ± 0.05 0.83 ± 0.05 (29.22 ± 5.67)	0.71 ± 0.05 0.85 ± 0.04 (22.94 ± 5.20)
SuPerFed-MM	0.16 ± 0.03 0.53 ± 0.07 (83.67 ± 3.51)	0.28 ± 0.04 0.69 ± 0.06 (46.41 ± 5.14)	0.27 ± 0.03 0.64 ± 0.07 (71.99 ± 5.01)	0.30 ± 0.04 0.69 ± 0.08 (56.07 ± 4.32)	0.28 ± 0.06 0.69 ± 0.07 (48.79 ± 5.73)	0.27 ± 0.05 0.63 ± 0.09 (26.64 ± 4.34)	0.31 ± 0.06 0.71 ± 0.10 42.66 ± 5.61	0.28 ± 0.06 0.68 ± 0.10 (35.74 ± 5.21)
SuPerFed-LM	0.14 ± 0.03 0.49 ± 0.08 (84.78 ± 3.63)	0.35 ± 0.04 0.77 ± 0.04 (46.82 ± 5.39)	0.27 ± 0.04 0.66 ± 0.07 (69.23 ± 5.25)	0.32 ± 0.04 0.72 ± 0.06 (54.69 ± 4.20)	0.29 ± 0.04 0.68 ± 0.06 (51.44 ± 5.67)	0.40 ± 0.04 0.80 ± 0.04 (28.51 ± 4.66)	0.36 ± 0.05 0.70 ± 0.07 (42.81 ± 5.50)	0.37 ± 0.06 0.76 ± 0.05 (34.10 ± 5.18)

Experimental Results

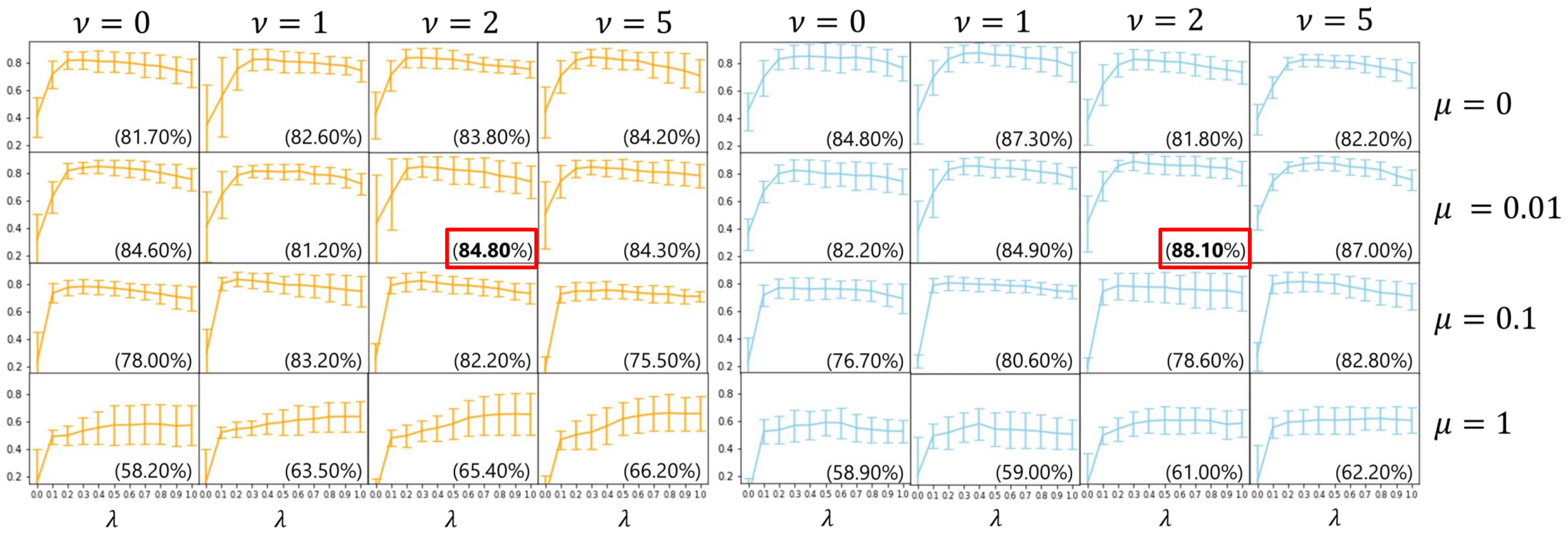
- **Ablation study – effects of regularization terms**

$$\mathcal{L}(h, \mathcal{W}) = l\big(h(\mathbf{x}), \mathbf{y}; \mathcal{W}(\lambda)\big) + \mu\left\|\mathbf{w}^f - \mathbf{w}^g\right\|^2 + \nu \cos^2\big(\mathbf{w}^f, \mathbf{w}^l\big)$$

- μ
  - Proximity regularization that penalizes a federated model when it deviates far from the global model constructed in the previous round.
- ν
  - Orthogonality regularization between the federated model and the local model which is critical for inducing the mode connectivity.

Experimental Results

- **Ablation study – effects of regularization terms**
  - Results when varying each constant μ and ν (L: SuPerFed-MM / R: SuPerFed-LM)
    - CIFAR10 dataset in pathological non-IID setting with 100 clients



Experimental Results

- **Ablation study – effects of phase-wise learning**
  - Phase-wise learning
    - Is holding some rounds for learning global knowledge helpful?
    - CIFAR10 dataset in pathological non-IID setting with 100 clients
      - R: total rounds of federated learning
      - L: rounds for Phase II (personalization phase)

L/R	0.0	0.2	0.4	0.6	0.8
SuPerFed-MM	88.10 ± 6.25	$\mathbf{92.13 \pm 4.95}$	91.99 ± 4.60	91.64 ± 6.78	86.84 ± 14.86
	0.91 ± 0.41	0.92 ± 0.47	$\mathbf{0.86 \pm 0.38}$	1.02 ± 0.44	2.79 ± 0.76
SuPerFed-LM	90.89 ± 4.58	91.78 ± 4.53	92.10 ± 4.41	$\mathbf{92.15 \pm 4.03}$	89.60 ± 11.56
	0.80 ± 0.43	0.68 ± 0.35	$\mathbf{0.66 \pm 0.32}$	1.88 ± 0.75	5.23 ± 2.04

Conclusion

- **SuPerFed: Connecting Low-Loss Subspace for Personalized Federated Learning**
  - We propose a novel model-mixture based personalized federated learning method that leverages benefits of the mode connectivity in terms of boosting personalization performance as well as securing robustness to the label noise.
  - While existing studies on the mode connectivity has been conducted assuming data-centralized setting, we adopt and exploit good properties of the mode connectivity by adjusting them to be suitable for the setting of federated learning.
  - We proposed a new personalized federated learning objective that ensures stable convergence through proximity regularization term and induces the mode connectivity through orthogonality regularization term.
  - Rigorous theoretical analyses are lacked, e.g., the optimal start round of Phase II, the optimal combination of constants for regularization terms, the convergence analysis of the proposed algorithm, and the optimal mixing ratio (i.e., λ); these should be further studied in the future work.

Thank You