ICLR 2022

On Bridging Generic and Personalized Federated Learning for Image Classification

Hong-You Chen, Wei-Lun Chao

Biao Mei
July 5, 2023

- Federated learning is promising for its capability to collaboratively train models, but vulnerable when clients' data diverge from each other.
- Generic FL

$$\min_{w} \mathcal{L}(w) = \sum_{m=1}^{M} \frac{|\mathcal{D}_m|}{|\mathcal{D}|} \mathcal{L}_m(w) \quad \text{where} \quad \mathcal{L}_m(w) = \frac{1}{|\mathcal{D}_m|} \sum_{i} \ell(x_i, y_i; w).$$

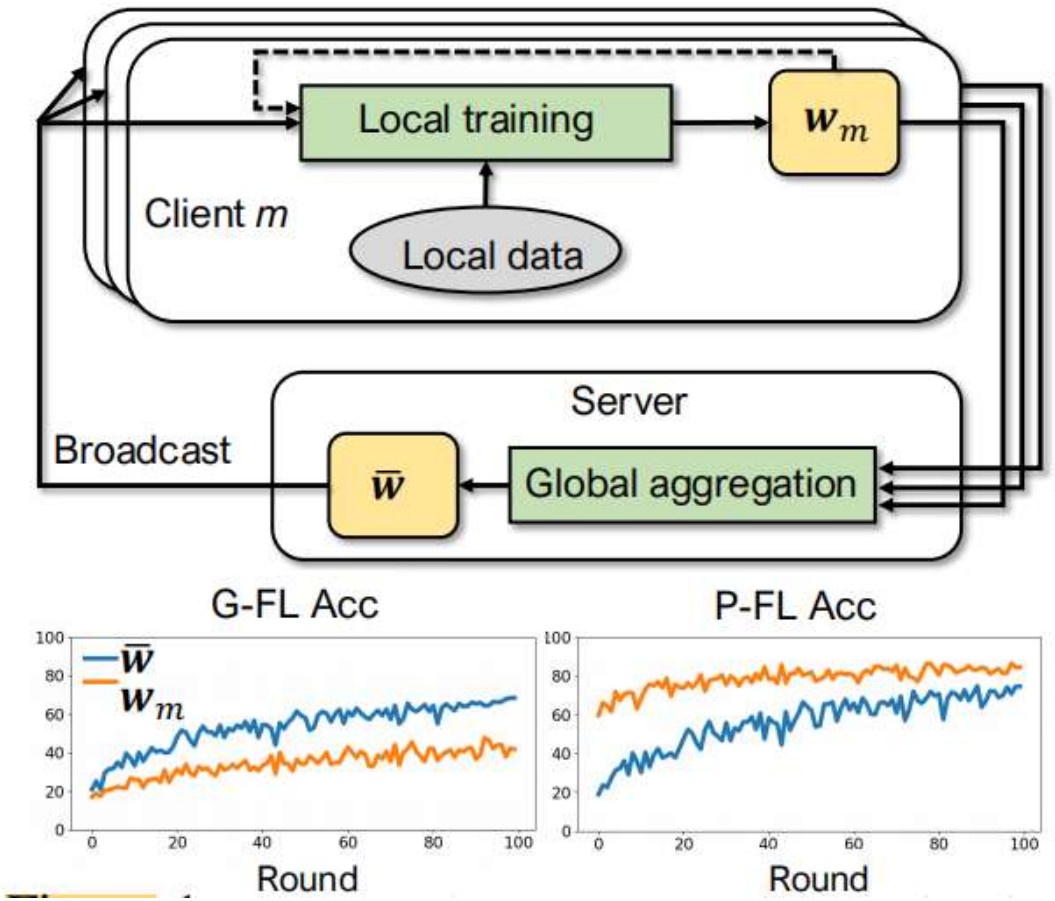
- The local objective function are quite different when local distribution $\mathcal{D}_m$ vary from each other, which thus results in drastic performance drop.

Personalized FL

Seek to construct a “personalized ” model for each client to acknowledge the heterogeneity among clients, which is shown to outperform the Generic FL at the local dataset.

Problem: Should we prioritize the learned model’s generic performance or its personalized performance?

Recall the traditional pipline of FedAvg



Algorithm 1 `FederatedAveraging`. The K clients are indexed by k; B is the local minibatch size, E is the number of local epochs, and η is the learning rate.

Server executes:

initialize w_0

for each round $t = 1, 2, \dots$ **do**

$m \leftarrow \max(C \cdot K, 1)$

$S_t \leftarrow$ (random set of m clients)

for each client $k \in S_t$ **in parallel do**

$w_{t+1}^k \leftarrow \text{ClientUpdate}(k, w_t)$

$w_{t+1} \leftarrow \sum_{k=1}^{K} \frac{n_k}{n} w_{t+1}^k$

ClientUpdate(k, w): *// Run on client k*

$\mathcal{B} \leftarrow$ (split $\mathcal{P}_k$ into batches of size B)

for each local epoch i from 1 to E **do**

for batch $b \in \mathcal{B}$ **do**

$w \leftarrow w - \eta \nabla \ell(w; b)$

return w to server

The Local model of Generic FL are often discarded after aggregation. w_m outperforms Global model on the personalized accuracy.

Intuition:

- Personalized models seem to come for free from the local training step of generic FL.
- Global aggregation indeed acts like a regularizer for local models.
- Initialization with weight average equal to impose an $\frac{\lambda}{2}||w-\overline{w}||_2(\lambda \to \infty)$ regularizer.

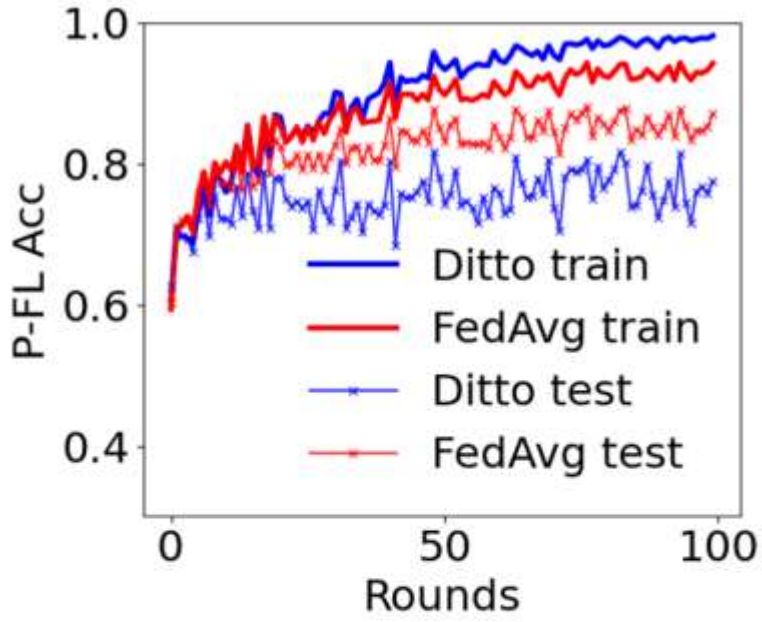


Figure 3: Comparison of the training and test accuracy in the P-FL setup. FEDAVG's local models achieve lower training accuracy but higher test accuracy.

- Strong personalized model emerge from generic FL algorithm provide us motive to focus on how to improve the Generic FL.
- One way to mitigate statistic heterogeneity influences is to make the local training objective aligned among clients.
- Suppose that data instance (x, y) of client m is sampled from distribution

$$\mathcal{P}_m(\boldsymbol{x}, y) = \mathcal{P}_m(\boldsymbol{x}|y)\mathcal{P}_m(y)$$

There are two type of causes result in heterogeneity.

- Considering the scenario that clients have different $P_m(y)$. We can indeed design a consistent local training objective that the learned local models should classify all the classes well.
- Proposing to treat client's local training as balanced classification tasks.

$$\mathcal{L}_m^{BR}(\boldsymbol{w}) \propto \sum_i q_{y_i} \ell(\boldsymbol{x}_i, y_i; \boldsymbol{w}) \qquad \text{where } q_{y_i} \text{ is usually set as } \frac{1}{N_{m,y_i}} \text{ or } \frac{1}{\sqrt{N_{m,y_i}}}.$$

- Many class-imbalanced work proposed to replace the instance loss(cross entropy) with a class balanced loss. Such as balanced softmax loss.

Balanced Softmax Loss

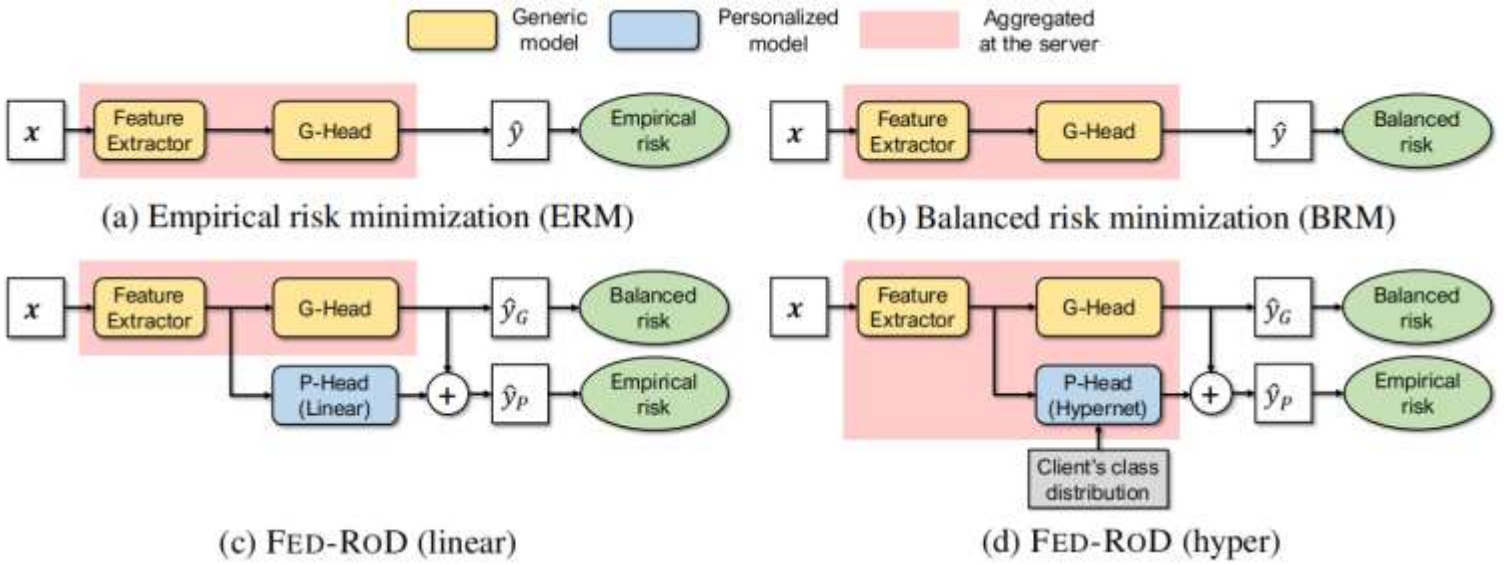
$$\mathcal{L}_m^{BR}(\boldsymbol{w}) \propto \sum_i \ell^{\text{BSM}}(\boldsymbol{x}_i, y_i; \boldsymbol{w}), \quad \text{where} \quad \ell^{\text{BSM}}(\boldsymbol{x}, y; \boldsymbol{w}) = -\log \frac{N_{m,y}^{\gamma} \exp(g_y(\boldsymbol{x}; \boldsymbol{w}))}{\sum_{c \in \mathbb{C}} N_{m,c}^{\gamma} \exp(g_c(\boldsymbol{x}; \boldsymbol{w}))}.$$

Theorem 1. *Assume ϕ to be the desired conditional probability of the balanced dataset, with the form $\phi_j = p(y = j|x) = \frac{p(x|y=j)}{p(x)} \frac{1}{k}$, and $\hat{\phi}$ to be the desired conditional probability of the imbalanced training set, with the form $\hat{\phi}_j = \hat{p}(y = j|x) = \frac{p(x|y=j)}{\hat{p}(x)} \frac{n_j}{\sum_{i=1}^{k} n_i}$. If ϕ is expressed by the standard Softmax function of model output η, then $\hat{\phi}$ can be expressed as*

$$\hat{\phi}_j = \frac{n_j e^{\eta_j}}{\sum_{i=1}^{k} n_i e^{\eta_i}}. \tag{3}$$

- The use of balanced risk result in better global model $\overline{w}$, but hurt the local model w_m's personalized performance(no longer optimize towards empirical risk $\mathcal{L}_m$)
- In order to realize personalized, the author propose to decouple the local model with global part and personalized part based on the shared feature extractor.

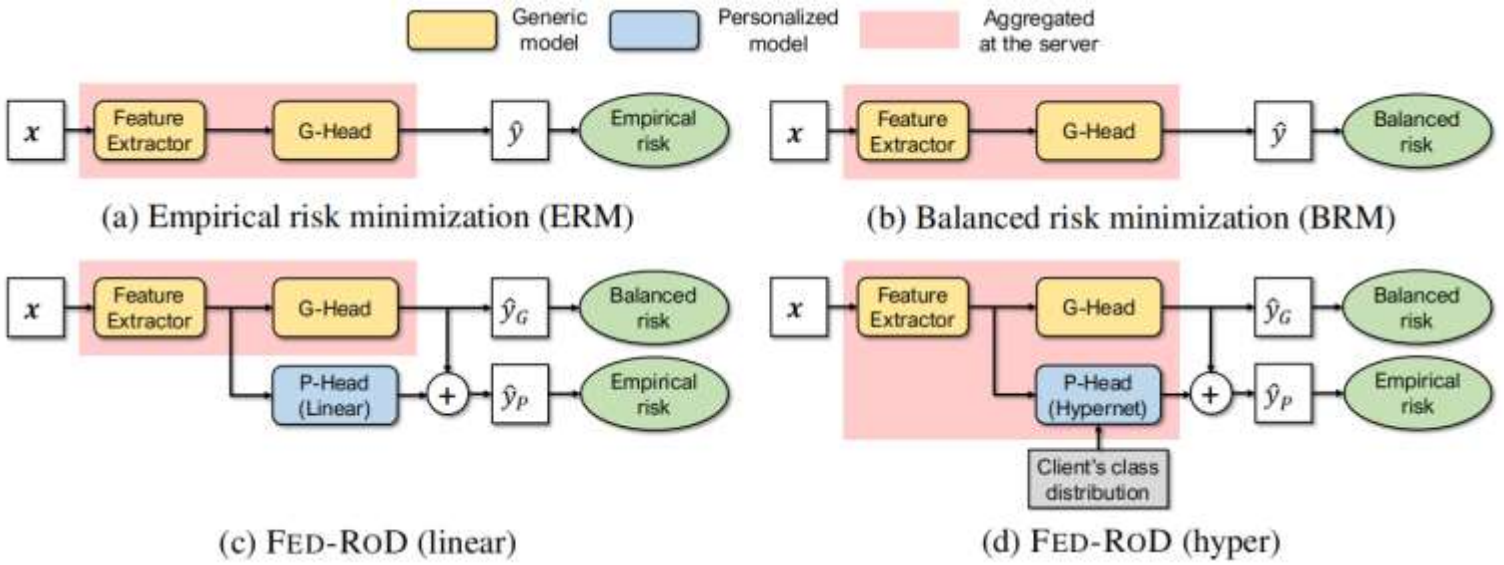
FedRoD



Feature extractor: $z = f(x;\theta)$ parameterized by θ.

Heads: $h^G(z;\psi)$ and $h^P(z;\phi_m)$ represent generic and personalized prediction heads(both are FC).

FedRoD



Denote $z = f(x;\theta)$, then $\hat{y}_G = h^G(z;\psi)$ and $\hat{y}_P = h^G(z;\psi) + h^P(z;\phi_m)$

Loss:

$$\min_{\boldsymbol{\theta},\boldsymbol{\psi}} \mathcal{L}(\{\boldsymbol{\theta},\boldsymbol{\psi}\}) = \sum_{m=1}^{M} \frac{|\mathcal{D}_m|}{|\mathcal{D}|} \mathcal{L}_m^{\mathrm{BR}}(\{\boldsymbol{\theta},\boldsymbol{\psi}\}) \quad \text{and} \quad \min_{\boldsymbol{\phi}_m} \mathcal{L}_m(\{\boldsymbol{\theta},\boldsymbol{\psi},\boldsymbol{\phi}_m\}), \forall m \in [M].$$

Algorithm:

Learning. Equation 7 cannot be solved directly in federated learning, so FED-ROD follows FEDAVG to learn iteratively between the local training and global aggregation steps

$$\textbf{Local:}\quad \boldsymbol{\theta}_m^\star, \boldsymbol{\psi}_m^\star = \arg\min_{\boldsymbol{\theta},\boldsymbol{\psi}} \mathcal{L}_m^{\text{BR}}(\{\boldsymbol{\theta}, \boldsymbol{\psi}\}), \quad \text{initialized with } \bar{\boldsymbol{\theta}}, \bar{\boldsymbol{\psi}}, \tag{8}$$

$$\boldsymbol{\phi}_m^\star = \arg\min_{\boldsymbol{\phi}_m} \mathcal{L}_m(\{\boldsymbol{\theta}, \boldsymbol{\psi}, \boldsymbol{\phi}_m\}), \quad \text{initialized with } \boldsymbol{\phi}'_m, \tag{9}$$

$$\textbf{Global:}\quad \bar{\boldsymbol{\theta}} \leftarrow \sum_{m=1}^{M} \frac{|\mathcal{D}_m|}{|\mathcal{D}|} \boldsymbol{\theta}_m^\star, \quad \bar{\boldsymbol{\psi}} \leftarrow \sum_{m=1}^{M} \frac{|\mathcal{D}_m|}{|\mathcal{D}|} \boldsymbol{\psi}_m^\star, \tag{10}$$

Algorithm 1: FED-ROD (linear) (**Fed**erated **Ro**bust **D**ecoupling)

Server input : initial global model parameter $\bar{\boldsymbol{\theta}}$ and $\bar{\boldsymbol{\psi}}$;
Client m's input : initial local model parameter $\boldsymbol{\phi}_m^\star$, local step size η, local labeled data $\mathcal{D}_m$;
for $r \leftarrow 1$ **to** R **do**
  Sample clients $\mathcal{S} \subseteq \{1, \cdots, N\}$;
  Communicate $\bar{\boldsymbol{\theta}}$ and $\bar{\boldsymbol{\psi}}$ to all clients $m \in \mathcal{S}$;
  for *each client* $m \in \mathcal{S}$ *in parallel* **do**
    Initialize $\boldsymbol{\theta} \leftarrow \bar{\boldsymbol{\theta}}$, $\boldsymbol{\psi} \leftarrow \bar{\boldsymbol{\psi}}$, and $\boldsymbol{\phi}_m \leftarrow \boldsymbol{\phi}_m^\star$;
    $\{\boldsymbol{\theta}_m^\star, \boldsymbol{\psi}_m^\star, \boldsymbol{\phi}_m^\star\} \leftarrow$ **Client local training**$(\{\boldsymbol{\theta}, \boldsymbol{\psi}, \boldsymbol{\phi}_m\}, \mathcal{D}_m, \eta)$; [Equation 8 and Equation 9]
    Communicate $\boldsymbol{\theta}_m^\star$ and $\boldsymbol{\psi}_m^\star$ to the server;
  end
  Construct $\bar{\boldsymbol{\theta}} = \sum_{m\in\mathcal{S}} \frac{|\mathcal{D}_m|}{\sum_{m'\in\mathcal{S}} |\mathcal{D}_{m'}|} \boldsymbol{\theta}_m^\star$;
  Construct $\bar{\boldsymbol{\psi}} = \sum_{m\in\mathcal{S}} \frac{|\mathcal{D}_m|}{\sum_{m'\in\mathcal{S}} |\mathcal{D}_{m'}|} \boldsymbol{\psi}_m^\star$;
end
Server output : $\bar{\boldsymbol{\theta}}$ and $\bar{\boldsymbol{\psi}}$;
Client m's output : $\{\boldsymbol{\theta}_m^\star, \boldsymbol{\psi}_m^\star, \boldsymbol{\phi}_m^\star\}$.

Adaptive Personalized Predictors Via Hypernetwokers

- Motivation: The personalized parameter ϕ_m is learned independently for each client, and thus FedRoD can only offer the global model for new client.
- Here the author propose to learn a meta-model which can generate ϕ_m for a client given the client's class distribution.
- Formally Hypernetworks: $\phi_m = H^P(a_m; \nu)$ parameterized by ν (two FC).

Algorithm Modification

$$\textbf{Local:}\quad \boldsymbol{\nu}_m^\star = \arg\min_{\boldsymbol{\nu}} \mathcal{L}_m(\{\boldsymbol{\theta}, \boldsymbol{\psi}, \boldsymbol{\nu}\}),\ \text{initialized with } \bar{\boldsymbol{\nu}};\quad \textbf{Global:}\quad \bar{\boldsymbol{\nu}} \leftarrow \sum_{m=1}^{M} \frac{|\mathcal{D}_m|}{|\mathcal{D}|} \boldsymbol{\nu}_m^\star.$$

Experiments

Dataset: CIFAR10/100, FMNIST, EMNIST

Number of clients: 20 for CIFAR10/100, 100 for FMNSIT and 2185 for EMNSIT

Sampling rate :40%, 20% and 5% respectively.

Generic / personalized performance:

$$\textbf{G-FL accuracy}: \quad \frac{1}{|\mathcal{D}_{\text{test}}|} \sum_i \mathbf{1}(y_i = \hat{y}_{i,G}),$$

$$\textbf{P-FL accuracy}: \quad \frac{1}{M} \sum_m \frac{\sum_i \mathcal{P}_m(y_i) \mathbf{1}(y_i = \hat{y}_{i,P})}{\sum_i \mathcal{P}_m(y_i)}.$$

Table 1: Results in G-FL accuracy and P-FL accuracy (%). ⋆: methods with no G-FL models and we combine their P-FL models. §: official implementation. Blue/**bold** fonts highlight the best baseline/our approach.

Dataset	EMNIST			FMNIST			FMNIST			CIFAR-10			CIFAR-10			CIFAR-100			CIFAR-100		
Non-IID	Writers			Dir(0.1)			Dir(0.3)			Dir(0.1)			Dir(0.3)			Dir(0.1)			Dir(0.3)		
Test Set	G-FL	P-FL		G-FL	P-FL		G-FL	P-FL		G-FL	P-FL		G-FL	P-FL		G-FL	P-FL		G-FL	P-FL	
Method / Model	GM	GM	PM	GM	GM	PM	GM	GM	PM	GM	GM	PM	GM	GM	PM	GM	GM	PM	GM	GM	PM
FedAvg	97.0	96.9	97.2	81.1	81.0	91.5	83.4	83.2	90.5	57.6	57.1	90.5	68.6	69.4	85.1	41.8	41.6	70.2	46.4	46.2	61.7
FedProx	97.0	97.0	97.0	82.2	82.3	91.4	84.5	84.5	89.7	58.7	58.9	89.7	69.9	69.8	84.7	41.7	41.6	70.4	46.5	46.4	61.5
SCAFFOLD	97.1	97.0	97.1	83.1	83.0	89.0	85.1	85.0	90.4	61.2	60.8	90.1	71.1	71.5	84.8	42.3	42.1	70.4	46.5	46.5	61.7
FedDyn §	97.3	97.3	97.3	83.2	83.2	90.7	86.1	86.1	91.5	63.4	63.9	92.4	72.5	73.2	85.4	43.0	43.0	72.0	47.5	47.4	62.5
MTL ⋆	75.4	75.0	85.6	36.1	36.0	87.3	53.1	53.4	78.3	12.1	12.7	90.6	13.5	13.7	80.2	9.5	9.3	60.7	10.8	10.7	49.9
LG-FedAvg ⋆§	80.1	80.0	95.6	54.8	54.5	89.5	66.8	66.8	84.4	29.5	28.8	90.8	46.7	46.2	82.4	23.5	23.4	66.7	34.5	33.9	55.4
FedPer ⋆	93.3	93.1	97.2	74.5	74.4	91.3	79.9	79.9	90.4	50.4	50.2	89.9	64.4	64.5	84.9	37.6	37.6	71.0	40.3	40.1	62.5
Per-FedAvg	95.1	-	97.0	80.5	-	82.8	84.1	-	86.7	60.7	-	82.7	70.5	-	80.7	39.0	-	66.6	44.5	-	58.9
pFedMe §	96.3	96.0	97.1	76.7	76.7	83.4	79.0	79.0	83.4	50.6	50.7	76.6	62.1	61.7	70.5	38.6	38.5	63.0	41.4	41.1	53.4
Ditto	97.0	97.0	97.4	81.5	81.5	89.4	83.3	83.2	90.1	58.1	58.3	86.8	69.7	69.8	81.5	41.7	41.8	68.5	46.4	46.4	58.8
FedFOMO ⋆	80.5	80.4	95.9	34.5	34.3	90.0	70.1	69.9	89.6	30.5	31.2	90.5	45.3	45.1	83.4	35.4	35.3	68.9	39.6	39.3	58.4
FedRep ⋆§	95.0	95.1	97.5	79.5	80.1	91.8	80.6	80.5	90.5	56.6	56.2	91.0	67.7	67.5	85.2	40.7	40.7	71.5	46.0	46.0	62.1
Local only	-	-	64.2	-	-	85.9	-	-	85.0	-	-	87.4	-	-	75.7	-	-	40.0	-	-	32.5
Fed-RoD (linear)	**97.3**	**97.3**	**97.5**	**83.9**	**83.9**	**92.7**	**86.3**	**86.3**	**94.5**	**68.5**	**68.5**	**92.7**	**76.9**	**76.8**	**86.4**	**45.9**	**45.8**	**72.2**	**48.5**	**48.5**	**62.3**
Fed-RoD (hyper)	**97.3**	**97.3**	**97.5**	**83.9**	**83.9**	**92.9**	**86.3**	**86.3**	**94.8**	**68.5**	**68.5**	**92.5**	**76.9**	**76.8**	**86.8**	**45.9**	**45.8**	**72.3**	**48.5**	**48.5**	**62.5**
+ FedDyn	**97.4**	**97.4**	**97.5**	**85.9**	**85.7**	**95.3**	**87.5**	**87.5**	**94.6**	**68.2**	**68.2**	**92.7**	**74.6**	**74.6**	**85.6**	**46.2**	**46.2**	**72.5**	**48.4**	**48.4**	**62.5**

BRM effectively reduces the variance of G-FL accuracy and local gradients

- Visualize the global model $\bar{w}$'s and local model $w_m's$ G-FL accuracy on CIAFR-10(Dir(0.3)).
- FED-ROD not only learns a better global model for GFL, but also has a smaller variance of accuracy.
- FED-ROD has a smaller variance of gradients.

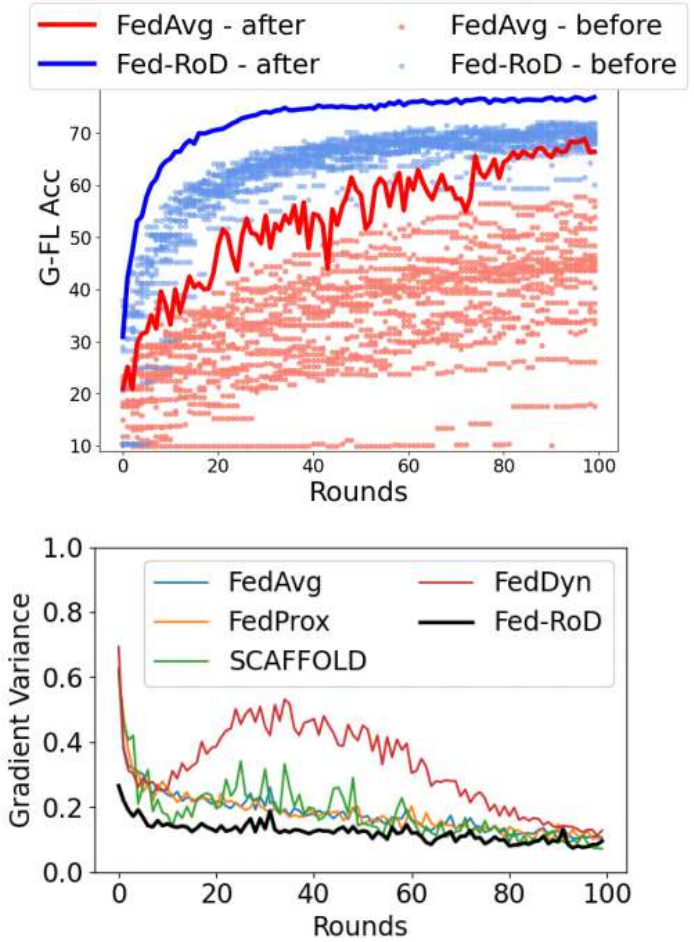


Figure 5: **Upper:** G-FL test accuracy along the training rounds before/after averaging the local models. **Lower:** variances of $\boldsymbol{w}_m - \bar{\boldsymbol{w}}$ across clients.

FED-ROD benefits from decoupling(Ablation experiment)

Table 2: Ablation study on variants of FED-ROD. FT: fine-tuning

Test Set	|G-FL|	P-FL	
Method / Model	| GM	|GM	PM
Centralized	| 85.4	85.4	-
FEDAVG	68.6	69.4	85.1
FEDAVG (BRM)	76.8	76.7	76.1
FEDAVG (BRM, FT)	76.8	76.7	84.5
FED-ROD (linear, BRM)	76.9	76.8	86.4
FED-ROD (hyper, BRM)	76.9	76.8	86.8

Table 3: FED-ROD with different balanced losses. ⋆: BSM

Test Set	|G-FL|	P-FL	
Loss / Model	| GM	|GM	PM
Cross entropy	68.6	69.4	85.1
(Hsu et al., 2020)	65.8	65.8	80.1
(Cao et al., 2019)	75.7	75.9	83.3
(Ye et al., 2020)	75.2	75.0	85.1
(Ren et al., 2020)⋆	76.9	76.8	86.8

- FEDAVG *with BRM significantly improves G-FL but degrades in P-FL.*
- FED-ROD remedies this by training a decoupled personalized head.
- Advanced losses outperforms importance re-weighting(Hau er al.2020)

Fed-RoD benefits future clients(Zero-shot personalization).

- Split the training data into 100 clients (50 are in training; 50 are new). And then evaluate on the 50 new clients.
- Without fine-tuning, FED-ROD (hyper) can already generate personalized models

Table 4: P-FL accuracy on future non-IID clients (Dir(0.3) for all datasets). Each cell is before/after local fine-tuning.

Method	FMNIST	CIFAR-10	CIFAR-100
FedAvg	80.3/87.2	56.2/76.0	38.9/54.3
FedDyn	82.3/87.6	61.7/76.2	40.1/57.2
Per-FedAvg	82.1/89.6	60.0/79.8	37.6/55.6
Fed-RoD (linear)	83.5/91.3	62.4/80.2	40.0/58.2
Fed-RoD (hyper)	88.9/91.4	75.7/81.5	40.7/59.0
+FedDyn	89.2/91.3	77.1/83.5	41.4/59.5

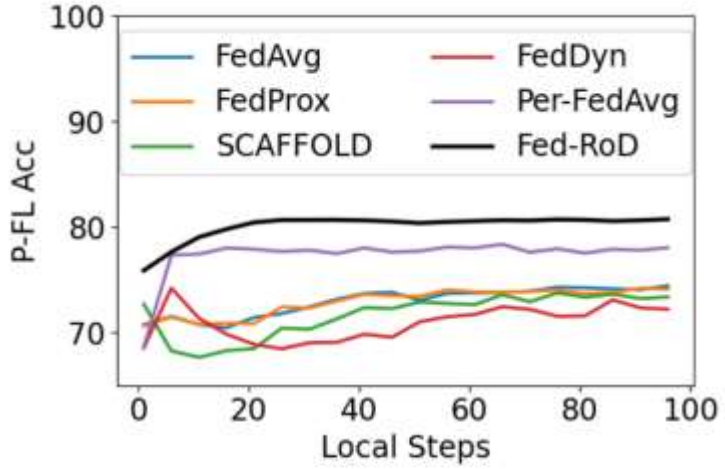


Figure 6: The average P-FL accuracy on future clients, with local training.

Summary/Main contributions

- Propose FedRoD to excel on both Generic FL and Personalized FL at the same time.
- Show a tuition that personalized model emerge from the local training step of Generic FL, due to implicit regularization.
- Enables zero-shot adaption and much effective fine-tuning for new clients(hypernetworks).

Advantages: Obtain a great GM, and it can be improved using are other advanced generic technics such as FedDyn. It has good compatibility with other methods.