Rethinking the Flat Minima Searching in Federated Learning

Taehwan Lee, Sung Whan Yoon

Ulsan National Institute of Science and Technology (UNIST)

taehwan@unist.ac.kr, shyoon8@unist.ac.kr

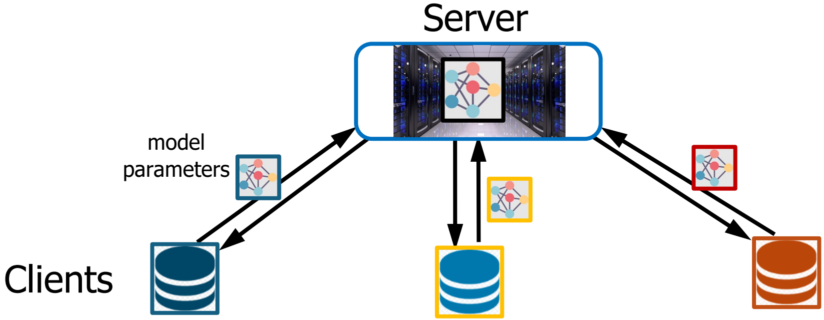
Jul. 23, 2024 @ICML 2024, Vienna

Outline

1. Preliminaries
2. Motivation
3. Proposed Method (FedGF)
4. Convergence Analysis
5. Experiment

Preliminaries – Federated Learning (FL)

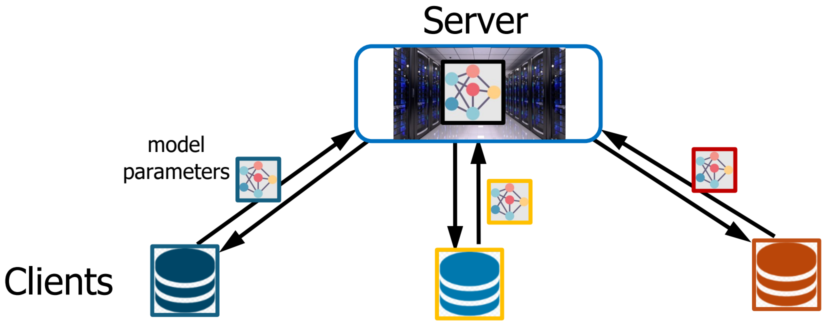
- Federated Learning (FL) is a framework of distributed learning.
- The **server** and **clients** communicate the model parameters to each other.



Preliminaries – Federated Learning (FL)

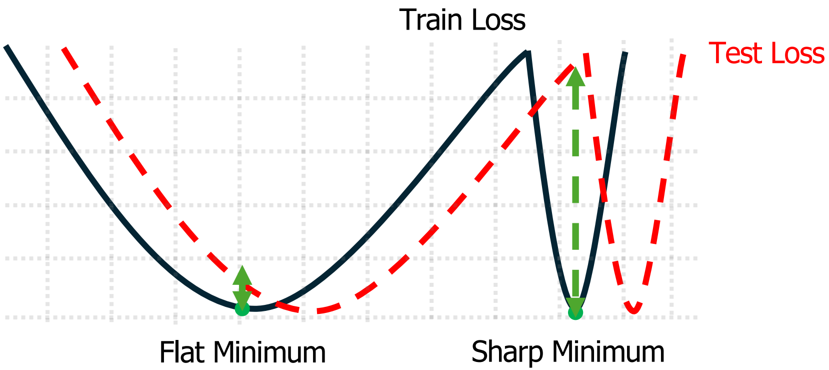
- The server can not access the clients' data samples.

It keeps the data privacy of clients.



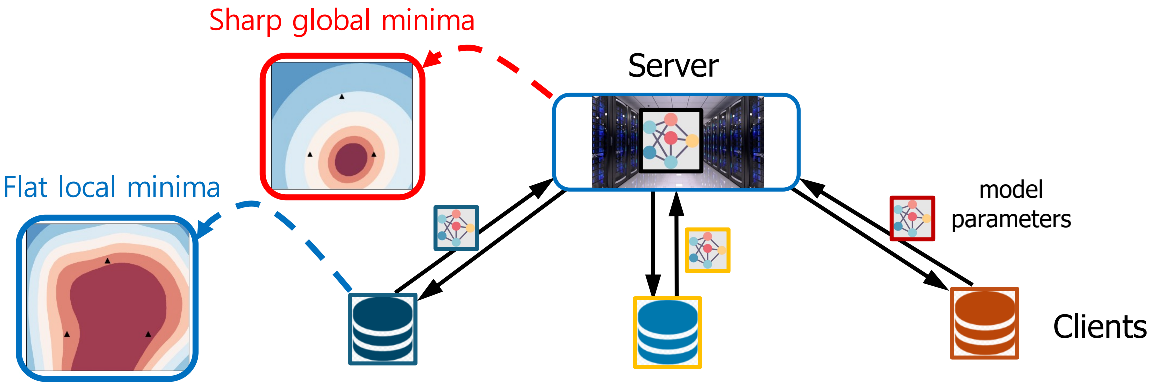
Preliminaries – Flat Minima

- Loss values at flat minima change more smoothly than sharp ones.
- Models at flat minima are robust to data distribution shifts.



Motivation

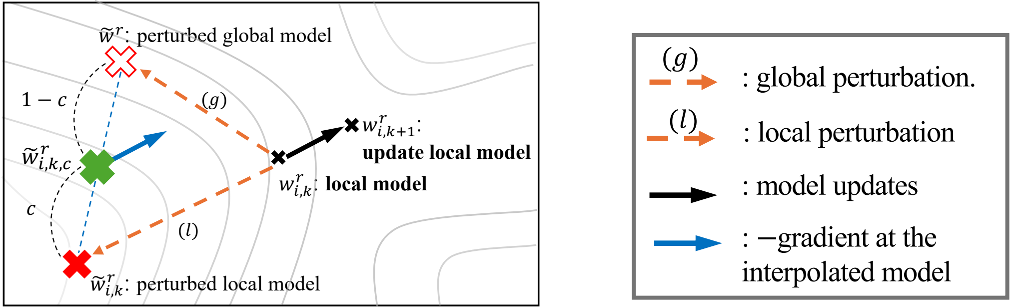
- FedSAM [ECCV'22, ICML'22] finds the local flatness of each client.
- However, FedSAM does not guarantee the **global flatness**.



[ECCV'22] D. Caldarola et al., "Improving Generalization in Federated Leaning by Seeking Flat Minima," ECCV 2022.
[ICML'22] Z. Qu et al., "Generalized Federated Learning via Sharpness Aware Minimization," ICML 2022.

Federated Learning for Global Flatness (FedGF)

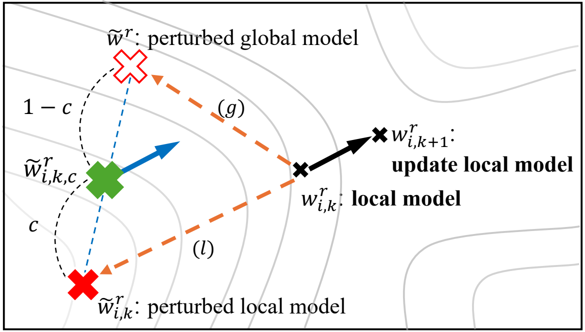
- We propose FedGF, which aims for global flatness.
- Key factors of FedGF
  - Perturbed local and global models, i.e., $\tilde{w}^r_{i,k}, \tilde{w}^r$, denoted by (,).
  - The interpolated model, i.e., $\tilde{w}^r_{i,k,c} = (1-c)\tilde{w}^r_{i,k} + c\tilde{w}^r$, denoted by ( ).



Global Perturbation in FedGF

- How could we find the global perturbation?

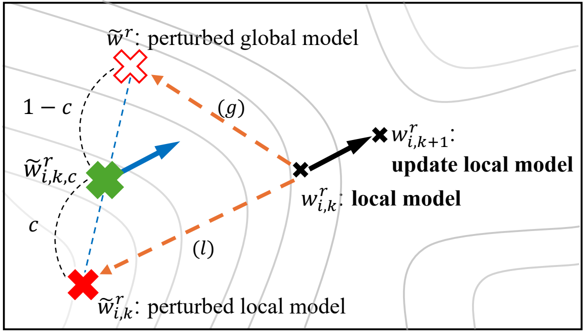
  - FedGF approximates the global gradient with the update of global model.
    - It approximates $\nabla F(w^r)$ as $\Delta^r \approx w^{r-1} - w^r$
    - Then, it calculates the pertuebed global model with Δ^r, i.e., $\tilde{w}^r = w^r + \rho \Delta^r / \|\Delta^r\|$.



Global Perturbation in FedGF

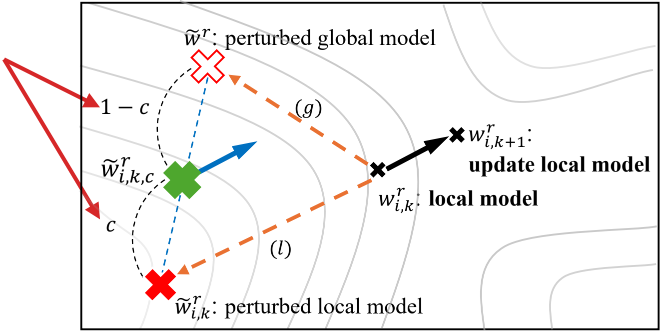
- How could we find the global perturbation?
  - The error caused by Δ^r is $\epsilon := \|\Delta^r/\|\Delta^r\| - \nabla F(w^r)/\|\nabla F(w^r)\|\|$.
  - The effect of ϵ will be discussed in **Convergence Analysis**.

(g) : global perturbation



Interpolated Model $\tilde{w}^r_{i,k,c}$ in FedGF

- From the perturbed models, i.e., $\tilde{w}^r_{i,k}, \tilde{w}^r$, we calculate $\tilde{w}^r_{i,k,c}$ (✖).
  - $\tilde{w}^r_{i,k,c} = (1-c)\tilde{w}^r_{i,k} + c\tilde{w}^r$
  - As c is close to 0, it finds local flatness. Otherwise, it aims for global flatness.
  - How do we control the c value (interpolation coefficient)?

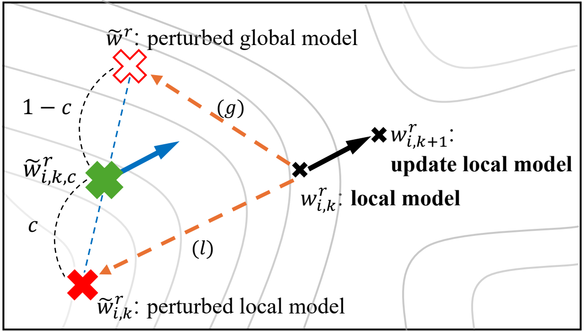


Interpolated Model $\tilde{w}^r_{i,k,c}$ in FedGF

- How do we control the c value (interpolation coefficient)?
  - FedGF controls c with the divergence D^r between global and local models.
  - As the heterogeneity gets severe, D^r increases, and FedGF pushes c to 1.

- Calculation Process of c
  - $D^r = \frac{1}{|S^r|}\sum_{i\in S^r}\|w^r - w^r_{i,k}\|$
  - $I^r = \mathbf{I}[D^r > T_D]$
  - $c = \frac{1}{W}\sum_{i=r-W+1}^{r} I^r$

$\mathbf{I}$: Indicator function
T_D: hyperparameter for threshold



Convergence of FedGF (Full client participation)

- The average of the norm of the gradient generated by the iterative rounds of FedGF satisfies:

$$\mathcal{O}\left(\frac{FL}{\sqrt{RKN}} + \frac{(1-c)^2}{R}\sigma_g^2 + \frac{L^2(1-c)^2}{R^{3/2}\sqrt{KN}}\sigma_l^2 + \frac{L^2c^2\epsilon^2}{R}\right),$$

where $F = F(\widetilde{w}^0) - F(\widetilde{w}^*)$ and $F(\widetilde{w}^*) = \min_{\widetilde{w}} F(\widetilde{w})$.

- As c approaches 1, FedGF suppresses the effect of σ_g^2 and σ_l^2.
- As c gets closer to 0, the effect of ϵ^2 is minimized.

σ_g^2: heterogeneity
σ_l^2: stochastic variance
ϵ^2: approximation error

Experiments – Test Accuracy

Task	Algorithms	Dirichlet distribution parameter α								
		$Dir.(\alpha = 0$, non-IID)			$Dir.(\alpha = 0.005)$			$Dir.(\alpha = 10$, IID)		
		Number of participating clients per each round								
		5	10	20	5	10	20	5	10	20
CIFAR-10	FedAvg	63.63	65.83	68.33	67.85	71.37	73.03	82.90	82.96	82.93
	FedAvgM	62.73	65.61	68.57	67.56	71.32	75.53	82.72	83.60	83.30
	FedProx	63.13	65.95	67.98	68.06	71.42	72.87	82.72	83.19	82.92
	SCAFFOLD	(✗)	(✗)	(✗)	57.13	56.46	45.27	82.93	83.05	83.39
	FedDyn	66.84	71.01	69.45	70.74	73.78	75.43	83.07	83.58	83.67
	FedSAM	68.11	71.17	72.49	71.87	74.31	76.07	83.78	83.88	83.82
	FedASAM	73.32	74.5	75.49	74.96	75.59	76.57	83.11	83.28	82.89
	MoFedSAM	73.1	71.08	76.66	74.43	77.53	79.27	80.9	81.01	81.02
	FedGAMMA	45.32	47.55	35.07	46.99	48.44	35.58	74.99	66.12	54.85
	FedSMOO	68.82	71.59	72.48	71.9	74.46	75.44	83.72	83.67	83.79
	FedGF	**78.41**	**79.68**	**80.86**	**78.79**	**79.39**	**79.69**	**84.71**	**83.94**	**83.85**
CIFAR-100	FedAvg	29.35	33.79	36.62	38.15	40.58	41.27	50.41	50.20	49.98
	FedAvgM	29.94	30.07	39.35	38.64	40.72	**48.44**	50.37	51.2	50.57
	FedProx	29.19	33.16	36.41	38.54	40.52	40.77	50.10	49.98	49.96
	SCAFFOLD	(✗)	(✗)	(✗)	36.25	(✗)	(✗)	52.28	52.12	52.48
	FedDyn	(✗)	(✗)	(✗)	(✗)	(✗)	(✗)	51.74	52.41	52.59
	FedSAM	29.43	34.32	36.88	42.28	44.57	45.18	54.06	53.75	53.5
	FedASAM	34.43	37.09	38.93	44.36	45.76	46.94	**54.6**	54.42	**54.73**
	MoFedSAM	29.02	35.82	41.26	34.64	42.24	44.92	52.13	52.21	52.07
	FedGAMMA	(✗)	(✗)	(✗)	20.52	14.76	10.33	47.43	38.18	25.06
	FedSMOO	35.35	38.78	40.82	44.39	46.03	47.5	54.31	54.89	54.65
	FedGF	**45.37**	**46.86**	**47.77**	**46.48**	**46.70**	46.08	54.16	**54.62**	54.59

(✗) indicates that the method fails to train, so the results remain at the same level as the random prediction.

We evaluate classification tasks with various settings, such as heterogeneity and the number of participating clients per round.

Experiments – Test Accuracy

Task	Algorithms	$Dir.(\alpha = 0, \text{non-IID})$			$Dir.(\alpha = 0.005)$			$Dir.(\alpha = 10, \text{IID})$		
		Dirichlet distribution parameter α								
		Number of participating clients per each round								
		5	10	20	5	10	20	5	10	20
CIFAR-10	FedAvg	63.63	65.83	68.33	67.85	71.37	73.03	82.90	82.96	82.93
	FedAvgM	62.73	65.61	68.57	67.56	71.32	75.53	82.72	83.60	83.30
	FedProx	63.13	65.95	67.98	68.06	71.42	72.87	82.72	83.19	82.92
	SCAFFOLD	(✗)	(✗)	(✗)	57.13	56.46	45.27	82.93	83.05	83.39
	FedDyn	66.84	71.01	69.45	70.74	73.78	75.43	83.07	83.58	83.67
	FedSAM	68.11	71.17	72.49	71.87	74.31	76.07	83.78	83.88	83.82
	FedASAM	73.32	74.5	75.49	74.96	75.59	76.57	83.11	83.28	82.89
	MoFedSAM	73.1	71.08	76.66	74.43	77.53	79.27	80.9	81.01	81.02
	FedGAMMA	45.32	47.55	35.07	46.99	48.44	35.58	74.99	66.12	54.85
	FedSMOO	68.82	71.59	72.48	71.9	74.46	75.44	83.72	83.67	83.79
	FedGF	**78.41**	**79.68**	**80.86**	**78.79**	**79.39**	**79.69**	**84.71**	**83.94**	**83.85**
CIFAR-100	FedAvg	29.35	33.79	36.62	38.15	40.58	41.27	50.41	50.20	49.98
	FedAvgM	29.94	30.07	39.35	38.64	40.72	**48.44**	50.37	51.2	50.57
	FedProx	29.19	33.16	36.41	38.54	40.52	40.77	50.10	49.98	49.96
	SCAFFOLD	(✗)	(✗)	(✗)	36.25	(✗)	(✗)	52.28	52.12	52.48
	FedDyn	(✗)	(✗)	(✗)	(✗)	(✗)	(✗)	51.74	52.41	52.59
	FedSAM	29.43	34.32	36.88	42.28	44.57	45.18	54.06	53.75	53.5
	FedASAM	34.43	37.09	38.93	44.36	45.76	46.94	**54.6**	54.42	**54.73**
	MoFedSAM	29.02	35.82	41.26	34.64	42.24	44.92	52.13	52.21	52.07
	FedGAMMA	(✗)	(✗)	(✗)	20.52	14.76	10.33	47.43	38.18	25.06
	FedSMOO	35.35	38.78	40.82	44.39	46.03	47.5	54.31	54.89	54.65
	FedGF	**45.37**	**46.86**	**47.77**	**46.48**	**46.70**	46.08	54.16	**54.62**	54.59

(✗) indicates that the method fails to train, so the results remain at the same level as the random prediction.

As heterogeneity gets severe ($\alpha = 10 \rightarrow \alpha = 0$),
FedGF shows more remarkable test accuracy than prior works.

Experiments – Test Accuracy

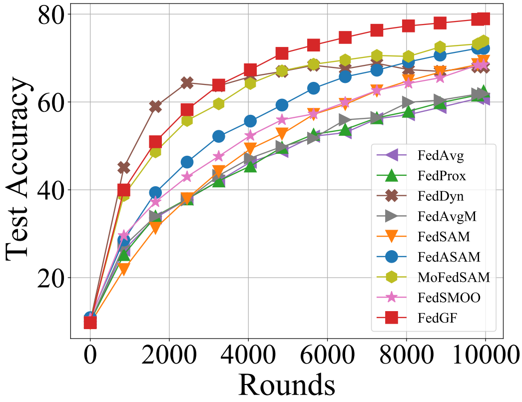
Task	Algorithms	Dirichlet distribution parameter α: $Dir.(\alpha=0, \text{non-IID})$			$Dir.(\alpha=0.005)$			$Dir.(\alpha=10, \text{IID})$		
		Number of participating clients per each round								
		5	10	20	5	10	20	5	10	20
CIFAR-10	FedAvg	63.63	65.83	68.33	67.85	71.37	73.03	82.90	82.96	82.93
	FedAvgM	62.73	65.61	68.57	67.56	71.32	75.53	82.72	83.60	83.30
	FedProx	63.13	65.95	67.98	68.06	71.42	72.87	82.72	83.19	82.92
	SCAFFOLD	(✗)	(✗)	(✗)	57.13	56.46	45.27	82.93	83.05	83.39
	FedDyn	66.84	71.01	69.45	70.74	73.78	75.43	83.07	83.58	83.67
	FedSAM	68.11	71.17	72.49	71.87	74.31	76.07	83.78	83.88	83.82
	FedASAM	73.32	74.5	75.49	74.96	75.59	76.57	83.11	83.28	82.89
	MoFedSAM	73.1	71.08	76.66	74.43	77.53	79.27	80.9	81.01	81.02
	FedGAMMA	45.32	47.55	35.07	46.99	48.44	35.58	74.99	66.12	54.85
	FedSMOO	68.82	71.59	72.48	71.9	74.46	75.44	83.72	83.67	83.79
	FedGF	**78.41**	**79.68**	**80.86**	**78.79**	**79.39**	**79.69**	**84.71**	**83.94**	**83.85**
CIFAR-100	FedAvg	29.35	33.79	36.62	38.15	40.58	41.27	50.41	50.20	49.98
	FedAvgM	29.94	30.07	39.35	38.64	40.72	**48.44**	50.37	51.2	50.57
	FedProx	29.19	33.16	36.41	38.54	40.52	40.77	50.10	49.98	49.96
	SCAFFOLD	(✗)	(✗)	(✗)	36.25	(✗)	(✗)	52.28	52.12	52.48
	FedDyn	(✗)	(✗)	(✗)	(✗)	(✗)	(✗)	51.74	52.41	52.59
	FedSAM	29.43	34.32	36.88	42.28	44.57	45.18	54.06	53.75	53.5
	FedASAM	34.43	37.09	38.93	44.36	45.76	46.94	**54.6**	54.42	**54.73**
	MoFedSAM	29.02	35.82	41.26	34.64	42.24	44.92	52.13	52.21	52.07
	FedGAMMA	(✗)	(✗)	(✗)	20.52	14.76	10.33	47.43	38.18	25.06
	FedSMOO	35.35	38.78	40.82	44.39	46.03	47.5	54.31	54.89	54.65
	FedGF	**45.37**	**46.86**	**47.77**	**46.48**	**46.70**	46.08	54.16	**54.62**	54.59

(✗) indicates that the method fails to train, so the results remain at the same level as the random prediction.

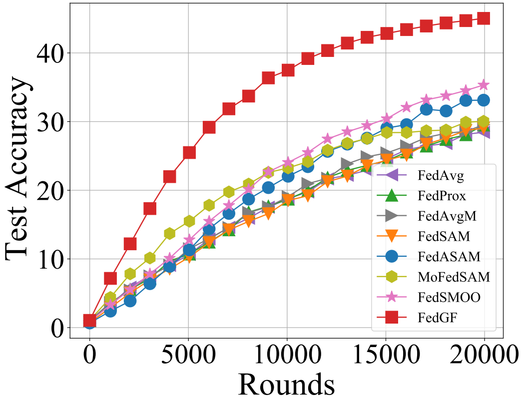
FedGF also shows robust performance
even as the number of clients decreases from 20 to 5.

Experiments – Convergence Behavior (non-IID)

CIFAR-10

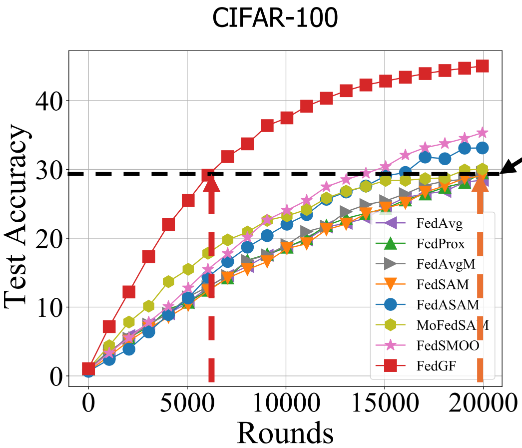


CIFAR-100



In convergence behavior, FedGF converges faster.

Experiments – Convergence Behavior (non-IID)



- Converged Accuracy of FedAvg: 29%
- Communication cost per round

	Upload	Download	Total
FedAvg	1	1	2
FedSAM	1	1	2
FedGF	1	2	3

When we compare the total communication cost to reach 29%,
FedGF takes 21k(7k*3), FedSAM takes 40k(20k*2).

Thank you

Our paper will be presented in the poster session at Hall C 4-9 #2306
on Tuesday, July 23rd, at 1:30 p.m. ~ 3 p.m.

Please visit our poster booth and have a discussion.