泛在计算与智能系统研究中心
Research Center of Ubiquitous Computing and Intelligent Systems

哈爾濱工業大學
Harbin Institute of Technology

Rethinking the Flat Minima Searching in Federated Learning

Taehwan Lee [1] Sung Whan Yoon [1 2]

• Published as a conference paper at ICML 2024

汇报人：游先耀

2024.10.15

背景：搜索平滑最小Loss平面以提高泛化性

- **平滑最小Loss平面搜索：**将loss最小值附近的平坦度整合到成本函数中，有助于在训练期间找到更平坦的局部模型。该方法已被证实可以增强联邦学习算法的性能，尤其是在异构环境中。

SAM optimizer: The SAM optimizer transforms a loss function $f(w)$ into a min-max cost function as follows:

$$\min_{w} \max_{\|\delta\| \leq \rho} F(w+\delta), \tag{4}$$

where ρ is a positive real number and $\|\delta\|$ is L2-norm of δ.

- **FedSAM**：该方法将 SAM 优化器直接应用于联邦学习中的本地训练，旨在通过优化局部目标来找到更扁平的局部模型，当局部模型聚合到全局模型中时，有助于提高性能。

FedSAM: By adopting the min-max problem of Eq. (4) in local training, FedSAM perturbs local model $w_{i,k}^r$:

$$\tilde{w}_{i,k}^r = w_{i,k}^r + \delta = w_{i,k}^r + \rho g_{i,k}^r / \|g_{i,k}^r\| \tag{5}$$

$$w_{i,k+1}^r = w_{i,k}^r - \eta_l \tilde{g}_{i,k}^r, \tag{6}$$

论文贡献

- **平坦度差异：**本文指出了一个称为平坦度差异的关键问题，该问题是指观察到在局部训练期间搜索最小平面并不能保证全局模型也将位于损失格局的平坦区域。这种差异可能导致聚合的全局模型的性能不理想，即使局部模型经过有效训练。

$$\Delta_{\mathcal{F}} := \left| \max_{||\delta|| \leq \rho} F(w+\delta) - F(w) - \left[\sum_{i=1}^{N} \frac{m_i}{m} \max_{||\delta_i|| \leq \rho} F_i(w_i+\delta_i) - F_i(w_i) \right] \right|.$$

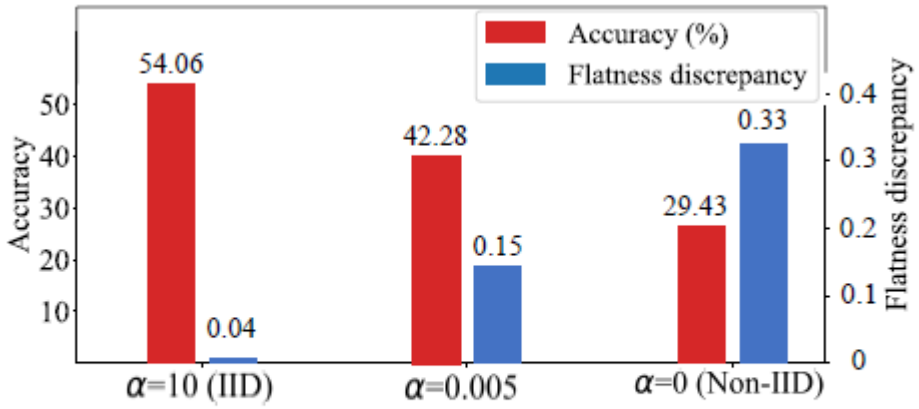
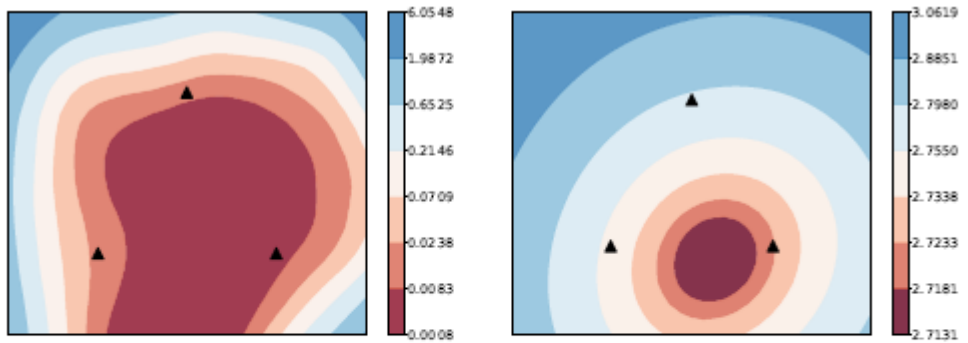


Figure 1: The performance and the flatness discrepancy ($\Delta_{\mathcal{F}}$) of FedSAM for the CIFAR-100 experiment



(a) Local model, w_i (b) Global model, w

Figure 2: Visualization of the loss surface of FedSAM for the CIFAR-100 case ($\alpha = 0$).

论文贡献

- **FedGF:** 为了解决这些问题，本文提出了一种名为全球平坦度联邦学习（FedGF）的新方法。该方法旨在明确地为全局模型追求更平坦的最小值，从而缓解平坦度差异并在异构 FL 基准测试中实现显著的性能改进。

$$g_{i,k}^r = \nabla F_i(w_{i,k}^r, \zeta_{i,k}) \tag{8}$$

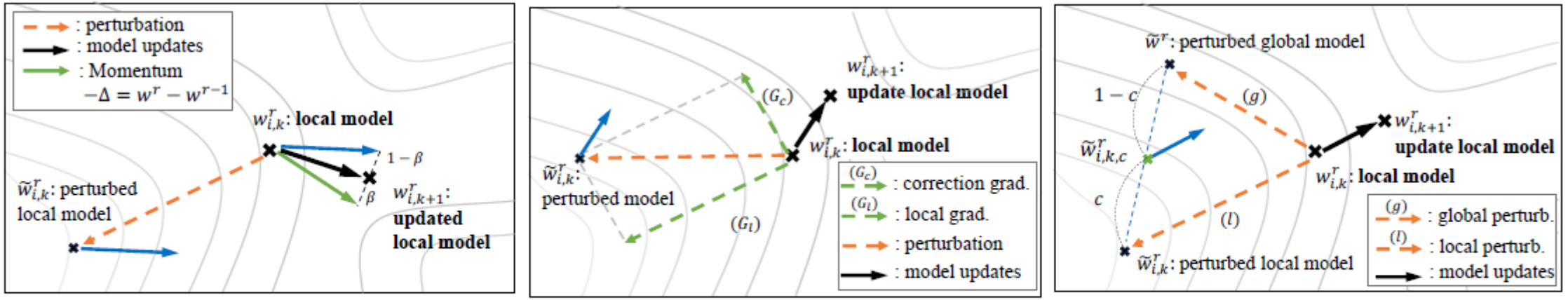
$$\tilde{w}_{i,k}^r = w_{i,k}^r + \rho g_{i,k}^r / \|g_{i,k}^r\| \quad \text{(perturbed local model)} \tag{9}$$

$$\triangle^r = w^{r-1} - w^r \tag{10}$$

$$\tilde{w}^r = w^r + \rho \triangle^r / \|\triangle^r\| \quad \text{(perturbed global model)} \tag{11}$$

$$\tilde{w}_{i,k,c}^r = c\tilde{w}^r + (1-c)\tilde{w}_{i,k}^r, \tag{12}$$

$$w_{i,k+1}^r = w_{i,k}^r - \eta_l \nabla F_i(\tilde{w}_{i,k,c}^r, \zeta_{i,k}). \tag{13}$$



(a) MoFedSAM (Qu et al., 2022) (b) FedSMOO (Sun et al., 2023a) (c) **FedGF** (ours)

Figure 3: Schematic of MoFedSAM, FedSMOO, and FedGF. The gray line illustrates the loss landscape of local distribution.

论文贡献

- **权重c的选择：**权重c 控制局部和全局扰动之间的扰动模型，基于以下策略：当局部模型与全局模型偏差很大时（non-IID），为了确保全局模型的平坦度，使用较大的 c；否则（IID），使用较小的 c。

$$D^r = \frac{1}{|S^r|} \sum_{i \in \mathcal{S}^r} \|w^r - w^r_{i,K}\|_2. \tag{15}$$

$$c = \frac{1}{W} \sum_{i=r-W+1}^{r} I^i. \tag{16}$$

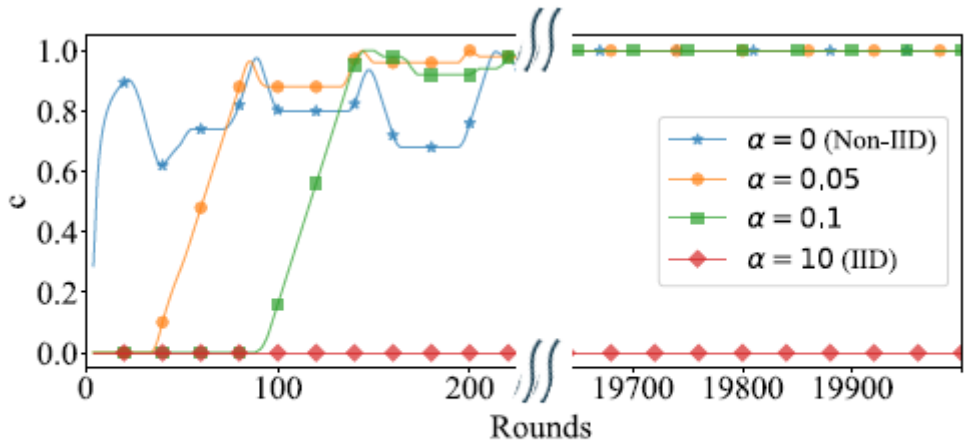


Figure 7: Behavior of c for CIFAR-100

实验：BENCHMARK RESULTS

Table 1: Test accuracies with of the FL algorithms on the CIFAR-10 and CIFAR-100 benchmarks

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

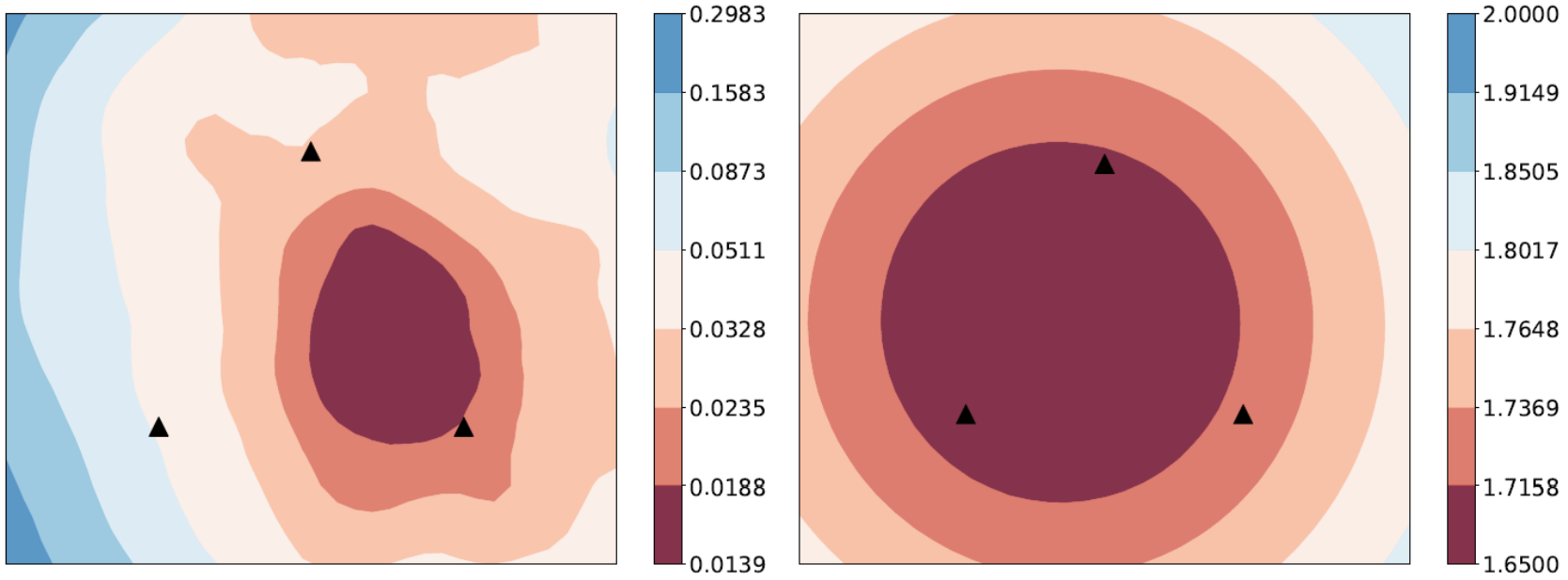
(✗) indicates that the method fails to train, so the results remain at the same level as the random prediction.

实验：固定权重c vs 动态计算权重c

Table 5: Static c vs. FedGF (adaptive c)

Dataset	IIDness	$c = 0$	$c = 0.5$	$c = 1$	**FedGF**
CIFAR-10	Non-IID	68.11	71.24	78.05	**78.41**
	IID	83.78	82.95	81.94	**84.71**
CIFAR-100	Non-IID	29.43	26.64	44.39	**45.37**
	IID	54.06	52.47	46.68	**54.16**

实验：Loss平面



(a) Local model, w_i (b) Global model, w

Figure 6: Loss surface of FedGF for CIFAR-100 ($\alpha = 0$).