Federated Learning over Connected Modes

Dennis Grinwald, Philipp Wiesner, Shinichi Nakajima

Paper PDF

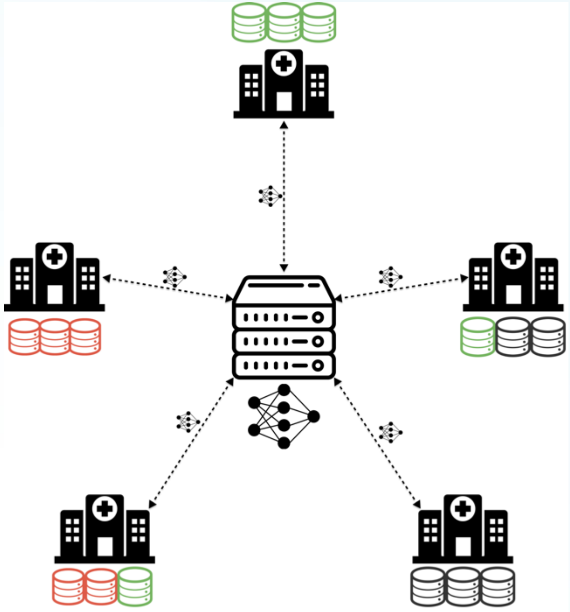
GitHub repo

www.bifold.berlin

Technische Universität Berlin

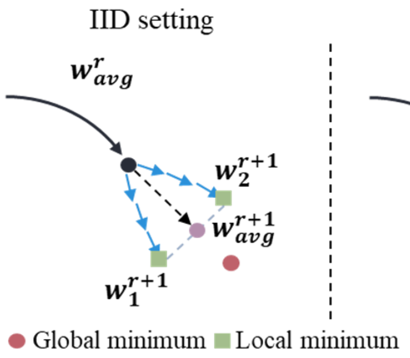
Learning on decentralized data

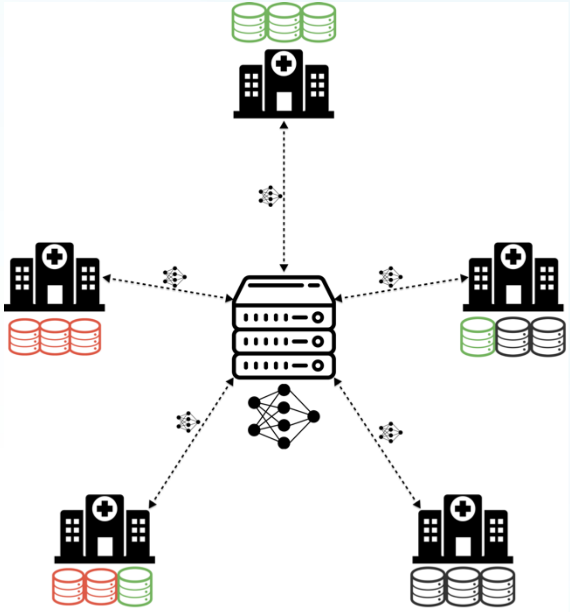
- Collaborative training of a common model on decentralized data (clients) [McMahan'17]



Learning on decentralized data

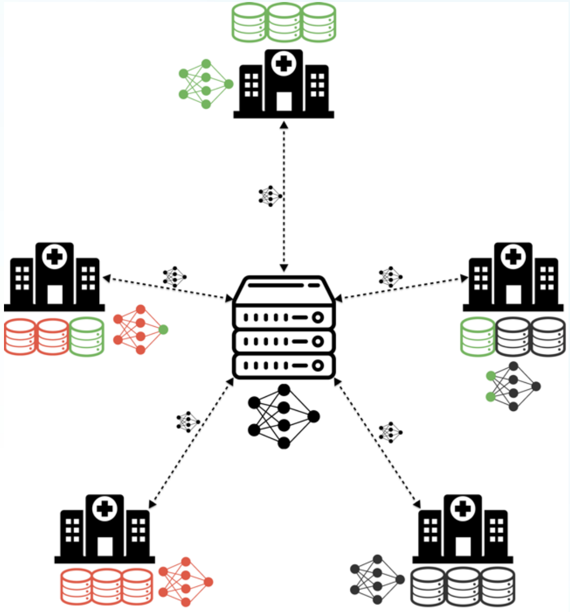
- Collaborative training of a common model on decentralized data (clients) [McMahan'17]
- **Challenge:** Communication-efficiency and statistically heterogeneous (Non-IID) client data [Zhao'20]





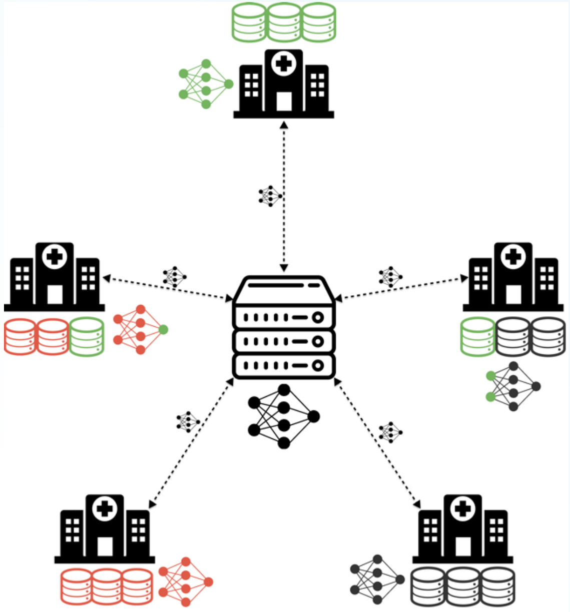
Global vs. Personalized FL (pFL)

- Each client owns and trains personalized model [Tan‘22]



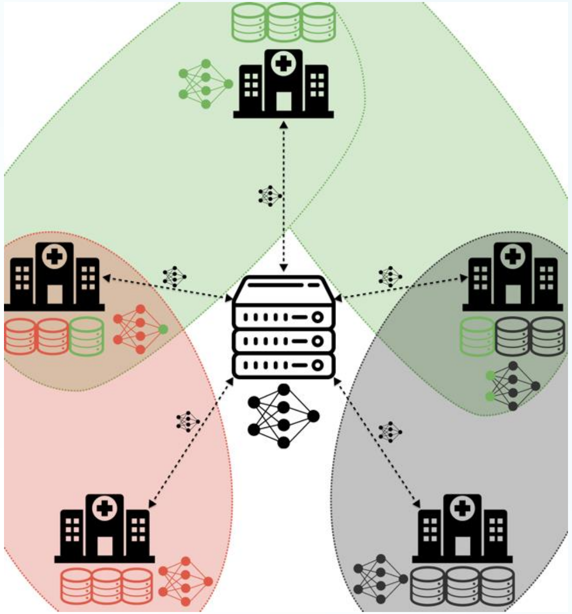
Global vs. Personalized FL (pFL)

- Each client owns and trains personalized model [3]
- **Problem 1**:
  pFL approaches typically do not benefit and can even harm global model performance

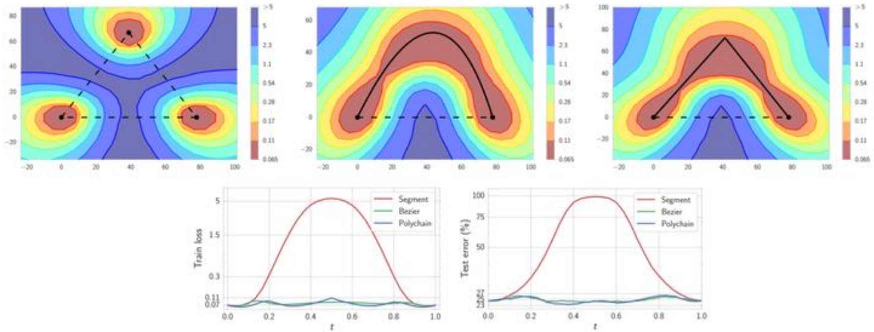


Global vs. Personalized FL (pFL)

- Each client owns and trains personalized model [3]
- **Problem 1**:
  pFL approaches typically do not benefit and can even harm global model performance
- **Problem 2**:
  Personalized models do not directly benefit from one another but through global model

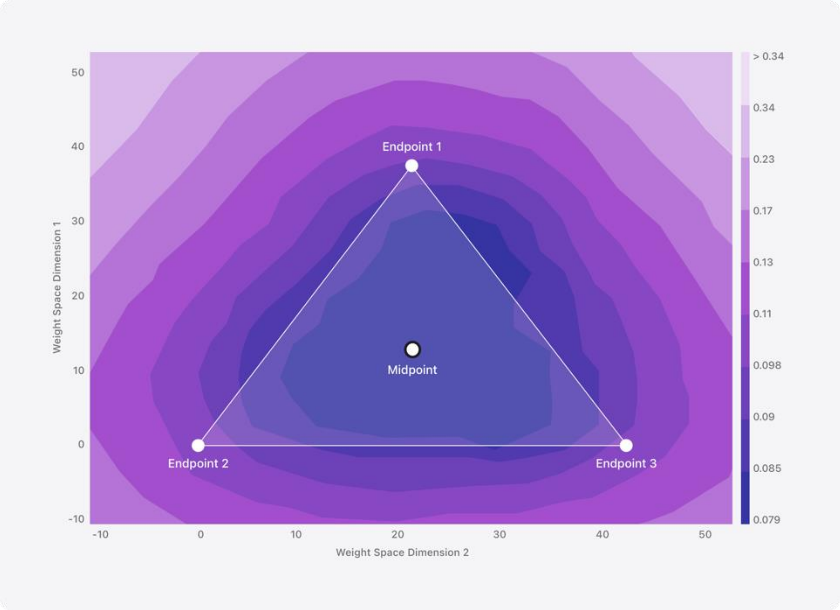


(Linear) Mode Connectivity



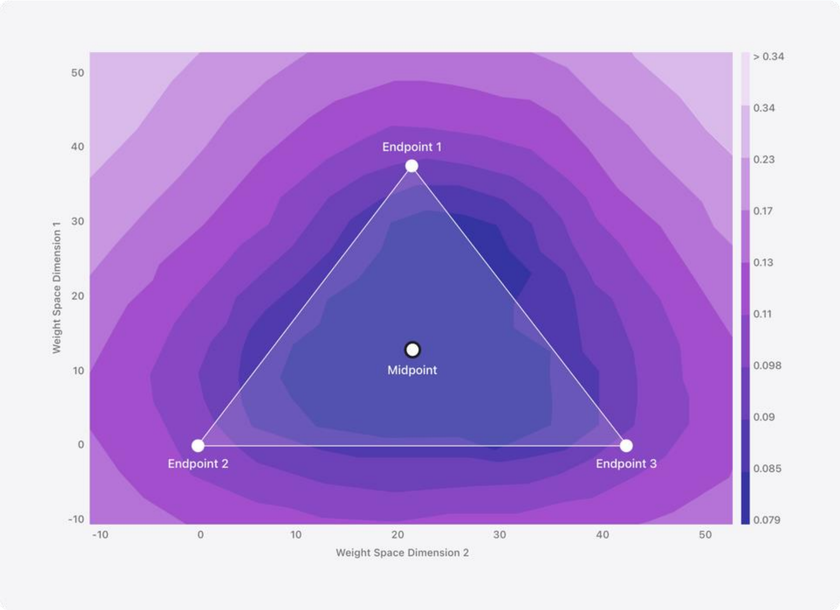
- **Observation:** Neural network solutions (modes) that started from different random initializations are connected by simple paths [Garipov'18]
- Models along these paths in parameter space exhibit low loss and functional diversity

Neural Network Simplex Learning

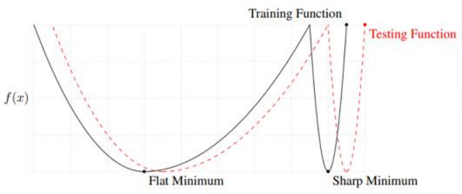


- Linear connectivity can be enforced during training with extra computational cost [Wortsman'21]
- Midpoint exhibits good generalization performance
- Midpoint per design lies in **flat minimum**

Neural Network Simplex Learning



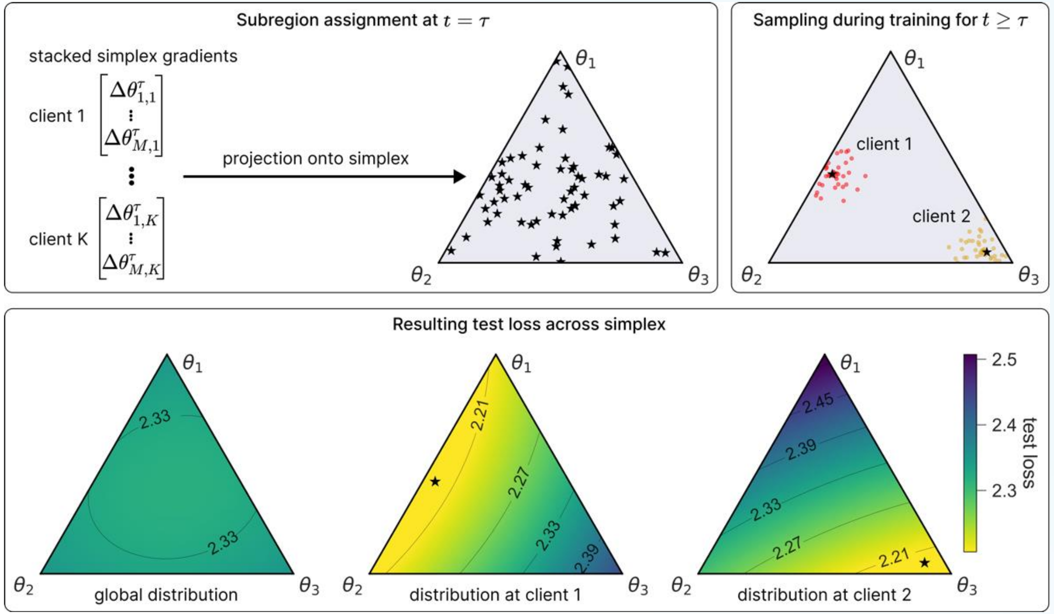
- Linear connectivity can be enforced during training with extra computational cost [Wortsman'21]
- Midpoint per design lies in **flat minimum**
- **Connection to Hochreiter et al. (1997)** : Flat minima and tend to be more robust to gap between empirical (training) loss and population loss (test loss) and thus generalize better.



[Foret'21]

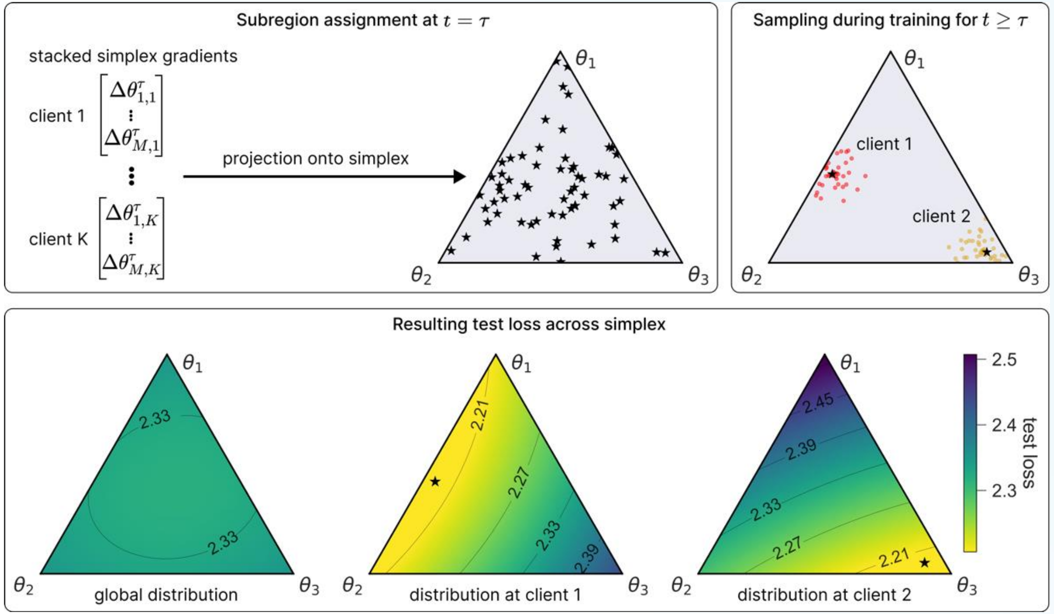
Federated Learning over Connected Modes (Floco)

- **Idea:** Train neural network solution simplex within which similar clients are grouped together [Grinwald'24]
- Each point in the simplex correspond to one model realization
- Sufficient to train solution simplex over last layer parameters only

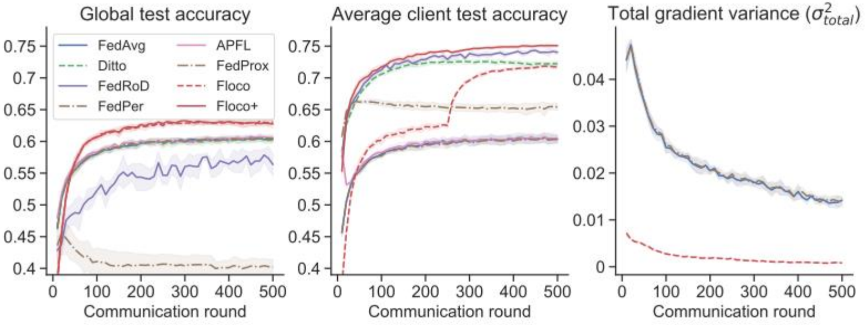


Federated Learning over Connected Modes (Floco)

- **Idea:** Train neural network solution simplex within which similar clients are grouped together
- Each point in the simplex correspond to one model realization
- Sufficient to train solution simplex over last layer parameters only
- **Result:**
  - Flat region in loss surface
  - SOTA personalized models that benefit each other (proj. points)
  - Robust and well-performing global model (midpoint)



Evaluation



Reduced gradient variance

Table 1: Average global and local test accuracy.

	CIFAR-10								FEMNIST			
	CifarCNN				pre-trained ResNet-18				FemnistCNN		pre-trained SqueezeNet	
	5-Fold		Dir(0.3)		5-Fold		Dir(0.3)					
FedAvg	60.36	*60.38*	60.74	*60.78*	75.33	*76.94*	68.59	*59.27*	78.83	*79.84*	75.13	*75.51*
FedProx	60.68	*60.36*	60.40	*60.27*	76.93	*77.46*	62.27	*60.26*	78.84	*80.15*	75.47	*75.99*
FedPer	40.23	*65.42*	33.90	*67.86*	68.64	*84.06*	50.84	*85.05*	50.76	*73.83*	64.03	*74.43*
APFL	60.56	*60.33*	60.55	*60.65*	53.25	*46.46*	50.97	*44.57*	4.95	*4.98*	38.21	*58.86*
Ditto	60.36	*72.22*	60.74	*73.90*	75.33	*69.18*	68.59	*76.23*	78.83	*82.02*	57.89	*65.06*
FedRoD	56.36	*74.03*	46.12	*76.42*	17.46	*31.82*	10.27	*33.85*	4.95	*4.99*	4.95	*4.95*
FLOCO	**62.93**	*71.78*	**62.57**	*71.04*	**77.15**	***85.90***	**73.62**	*80.38*	**78.99**	*84.09*	**75.86**	*77.00*
FLOCO$^+$	**62.93**	***75.08***	**62.57**	***76.50***	**77.15**	*84.88*	**73.62**	***85.89***	**78.99**	***84.75***	**75.86**	***82.41***

Table 2: Average global and local expected test calibration error.

	CIFAR-10								FEMNIST			
	CifarCNN				pre-trained ResNet-18				FemnistCNN		pre-trained SqueezeNet	
	5-Fold		Dir(0.3)		5-Fold		Dir(0.3)					
FedAvg	24.08	*25.61*	22.95	*24.51*	13.77	*19.57*	13.48	*19.57*	12.40	*16.86*	15.54	*20.43*
FedProx	23.76	*25.56*	23.19	*24.89*	12.40	*12.41*	15.16	*19.83*	12.41	*16.93*	15.48	*20.04*
FedPer	47.75	*28.22*	56.39	*25.70*	19.73	*11.19*	38.48	***10.88***	38.44	*21.68*	28.28	*22.31*
APFL	23.30	*25.01*	22.19	*23.91*	28.39	*33.39*	20.02	*26.01*	4.95	*4.98*	**7.6**	*15.82*
Ditto	24.08	*19.13*	22.95	*17.64*	13.77	*16.43*	13.48	*14.50*	12.40	*14.65*	15.54	*18.06*
FedRoD	29.78	*18.40*	41.91	*17.45*	75.59	*64.07*	89.31	*64.07*	4.95	*4.99*	4.99	*4.99*
FLOCO	**21.82**	*18.44*	**20.06**	*18.75*	**11.48**	***9.44***	**10.30**	*11.28*	**10.28**	*13.94*	14.65	*19.15*
FLOCO$^+$	**21.82**	***17.69***	**20.06**	***16.50***	**11.48**	*12.42*	**10.30**	*11.98*	**10.28**	***13.87***	14.65	***15.35***

Summary

- Floco beats SOTA pFL baselines on local and global test accuracy and ECE.
- Applicable to both randomly initialized as well as pretrained models.
- Minimal computational overhead as compared to regular FedAvg.
- Promising future directions include cross-device FL settings and the more general model merging setting.

BIFOLD

Thank you!

Paper PDF

GitHub repo

www.bifold.berlin

Literature

[1] McMahan, Brendan, et al. "Communication-efficient learning of deep networks from decentralized data." Artificial intelligence and statistics. PMLR, 2017.

[2] Zhao, Yue, et al. "Federated learning with non-iid data." arXiv preprint arXiv:1806.00582 (2018).[3] Kulkarni, Viraj, Milind Kulkarni, and Aniruddha Pant. "Survey of personalization techniques for federated learning." 2020 fourth world conference on smart trends in systems, security and sustainability (WorldS4). IEEE, 2020.

[3] Tan, Alysa Ziying, et al. "Towards personalized federated learning." IEEE transactions on neural networks and learning systems 34.12 (2022): 9587-9603.

[4] Garipov, Timur, et al. "Loss surfaces, mode connectivity, and fast ensembling of dnns." Advances in neural information processing systems 31 (2018).

[5] Nagarajan, Vaishnavh, and J. Zico Kolter. "Uniform convergence may be unable to explain generalization in deep learning." Advances in Neural Information Processing Systems 32 (2019).

[6] Wortsman, Mitchell, et al. "Learning neural network subspaces." International Conference on Machine Learning. PMLR, 2021.

[7] Hochreiter, Sepp, and Jürgen Schmidhuber. "Flat minima." Neural computation 9.1 (1997): 1-42.

[8] Foret, Pierre, et al. "Sharpness-aware minimization for efficiently improving generalization." International Conference on Learning Representations. ICLR, 2021.

[9] Grinwald, Dennis, Philipp Wiesner, and Shinichi Nakajima. "Federated Learning over Connected Modes." The Thirty-eighth Annual Conference on Neural Information Processing Systems.