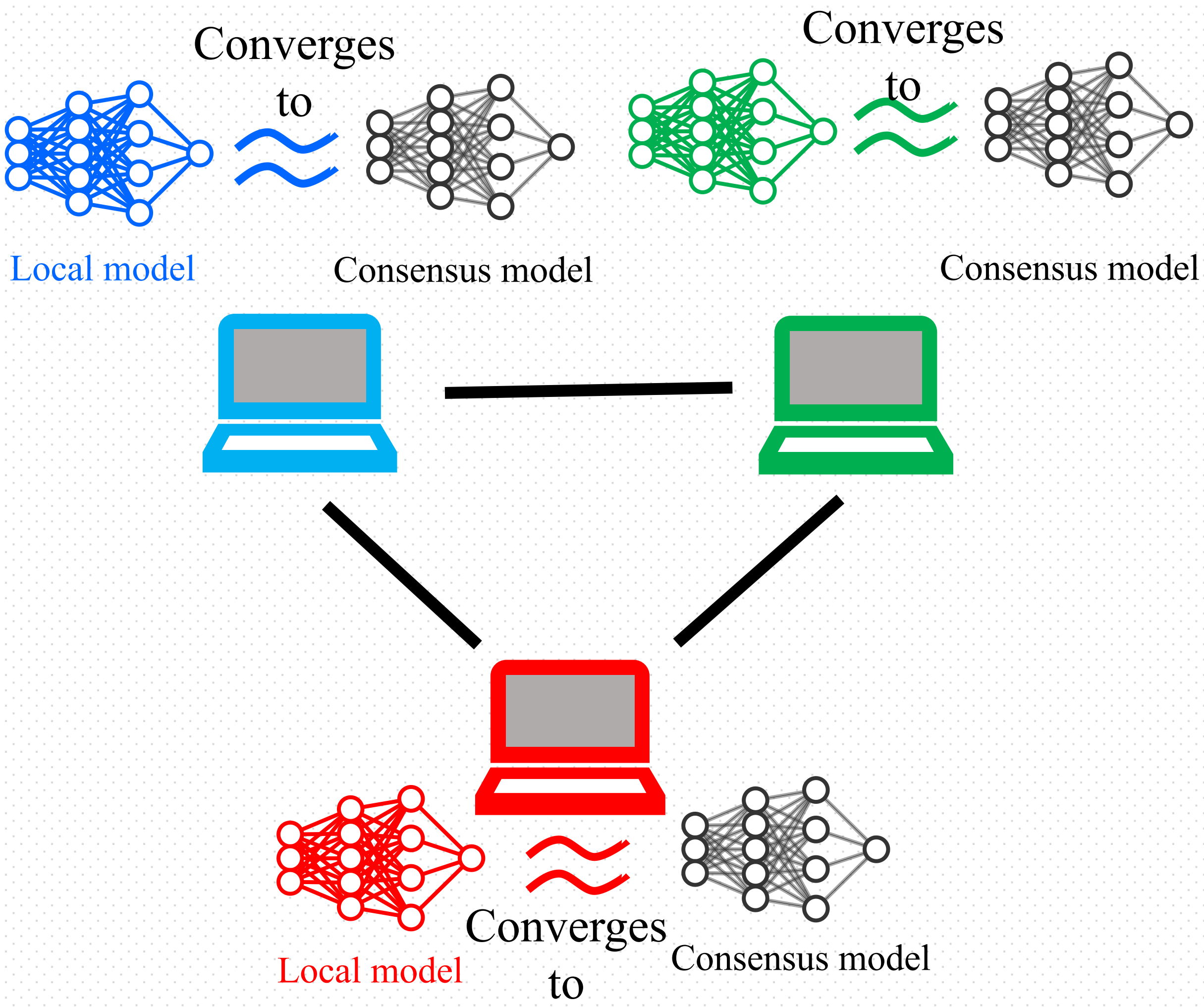
Structured Cooperative Learning with Graphical Model Priors

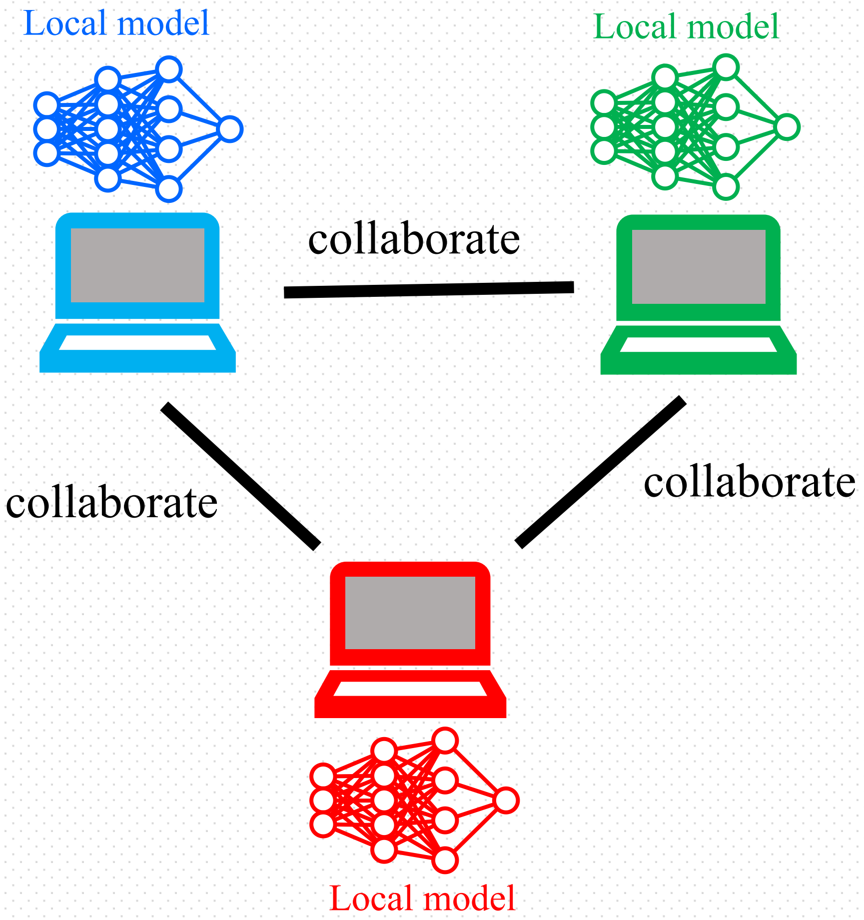
Shuangtong Li[1], Tianyi Zhou[2], Xinmei Tian[1 3], Dacheng Tao[4]

[1]University of Science and Technology of China
[2]University of Maryland, College Park
[3]Institute of Artificial Intelligence, Hefei Comprehensive National Science Center
[4]The University of Sydney

Background



Traditional decentralized Learning



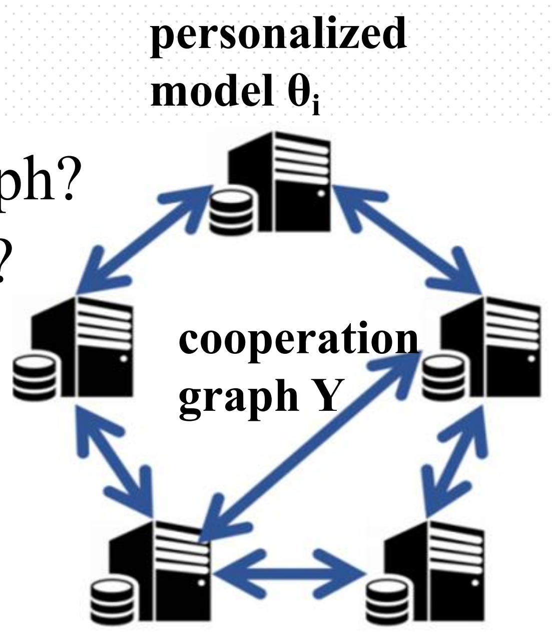
Decentralized Learning of Personalized Models (DLPM) [1]

[1] Shuangtong Li, Tianyi Zhou, Xinmei Tian, and Dacheng Tao. Learning to collaborate in decentralized learning of personalized models. In IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022.

Motivation

❑DLPM Challenges:

- How to determine when and which clients should cooperate?
- How to cooperate when personal tasks and data cannot be shared?
- To save communication cost, how to discover a sparse cooperation graph?
- How to adjust the graph adaptive to model changes in training process?



❑Structured Cooperative Learning (SCooL):

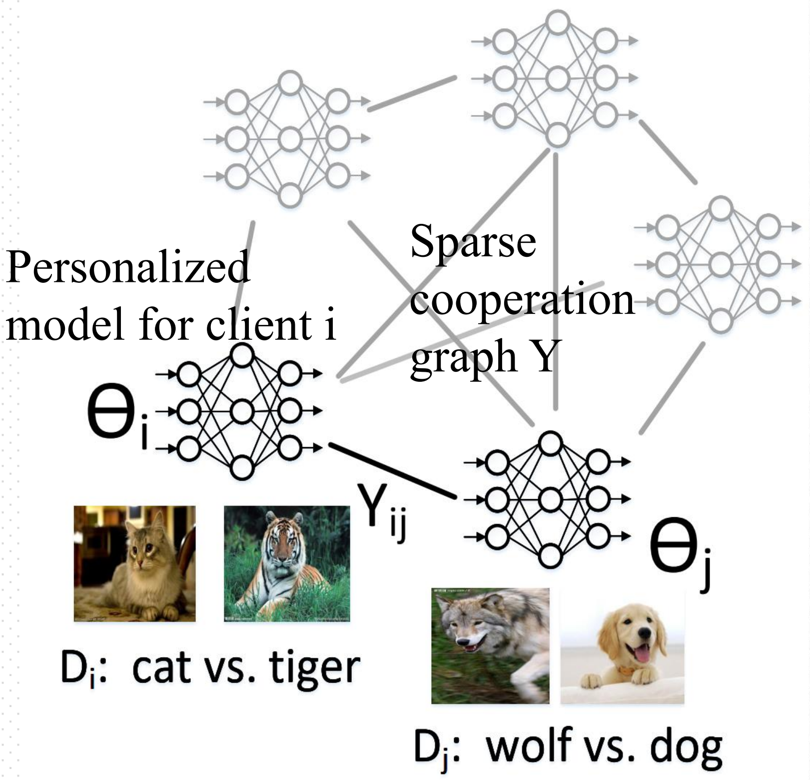
❖A **general** probabilistic modelling framework.

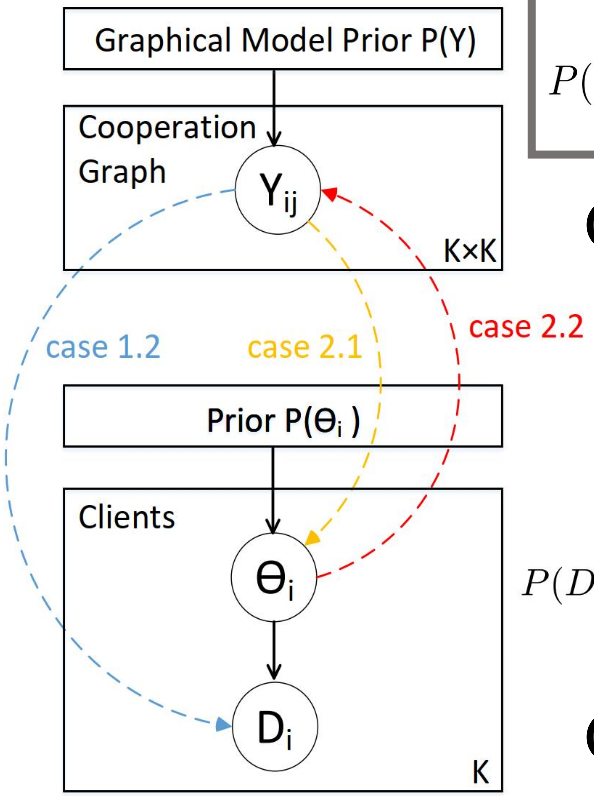
❖Jointly optimize **personalized models $\theta_{1:K}$** and **cooperation graph Y**.

❖Different graphical model priors of Y ⟶ various novel DLPM algorithms.

❖A systematic optimization method: **variational inference**.

SCooL Framework





Probabilistic Modeling with Cooperation Graph

$$P(\theta_{1:K}|D_{1:K}) \propto P(\theta_{1:K}, D_{1:K}) = \int \boxed{P(D_{1:K}|\theta_{1:K}, Y)} \boxed{P(\theta_{1:K}, Y)} dY.$$

(1) Joint Likelihood $\boxed{P(D_{1:K}|\theta_{1:K}, Y)}$

case 1.1 Y does not affect data distribution.

$$P(D_{1:K}|\theta_{1:K}) = \prod_{i=1}^{K} P(D_i|\theta_i)$$

case 1.2 Y coordinates the training process.

$$P(D_{1:K}|\theta_{1:K}, Y) = \prod_{i=1}^{K} P(D_{1:K}|\theta_i, Y) = \prod_{i=1}^{k} \left(P(D_i|\theta_i) \prod_{j\neq i, Y_{ij}=1} P(D_j|\theta_i) \right)$$

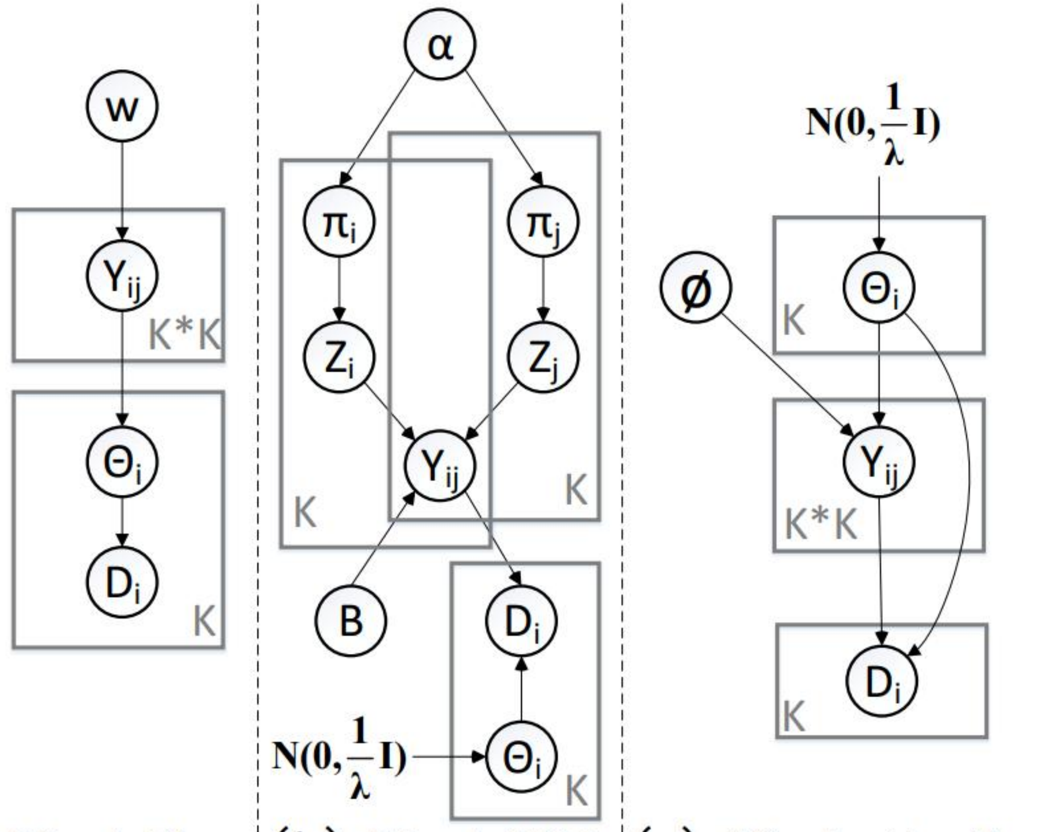
(2) Joint Priors $\boxed{P(\theta_{1:K}, Y)}$

case 2.1 : $P(\theta_{1:K}|Y)P(Y)$, $\theta_{1:K}$ is derived from Y.

case 2.2 : $P(Y|\theta_{1:K})P(\theta_{1:K})$, $\theta_{1:K}$ determines Y.

case 2.3 : $P(\theta_{1:K})P(Y)$, $\theta_{1:K}$ is independent to Y.

SCooL Instantiations



(a) SCooL-Dirac (b) SCooL-SBM (c) SCooL-attention

(a) SCooL-Dirac

$Y \sim \delta$ (w)

SCooL-Dirac is equivalent to DPSGD [2].

(b) SCooL-SBM

Y~Stochastic Block model (SBM) [3]

(c) SCooL-attention

$Y_{ij} \sim$ Attention(θi,θj)

$$p_{ij} = \frac{exp(f(\theta_i, \theta_j))}{\sum_l exp(f(\theta_i, \theta_l))}$$

$$\vec{Y} \sim Categorical(pi1, ... pik)$$

[2] Xiangru Lian, Ce Zhang, Huan Zhang, Cho-Jui Hsieh, Wei Zhang, and Ji Liu. Can decentralized algorithms outperform centralized algorithms? a case study for decentralized parallel stochastic gradient descent. In Advances in Neural Information Processing Systems, 2017.
[3] Paul W Holland, Kathryn Blackmond Laskey, and Samuel Leinhardt. Stochastic blockmodels: First steps. Social networks, 5(2):109–137, 1983.

EM Algorithm for SCooL

We derive EM algorithms for SCooL models via **variational inference** method.

ELBO:

$$\log p(X|\Phi) = \log \int p(X, Z|\Phi) dZ$$

$$\geq \int q(Z) \log \frac{p(X, Z|\Phi)}{q(Z)} dZ := H(q, \Phi).$$

E-step: update cooperation graph Y.

$$w_{ij} \leftarrow F\left(\log P(D_j|\theta_i), \beta, \Phi \right) \forall i, j \in [K]$$

M-step: optimize the local models $\theta_{1:K}$.

$$\theta_i \leftarrow \theta_i - \eta_1 \left(\sum_{j \neq i} w_{ij} \nabla L(D_j; \theta_i) + \nabla L(D_i; \theta_i) + G(\beta, \Phi) \right)$$

Experiments

Methodology	Algorithm	CIFAR-10	CIFAR-100	MiniImageNet
Local only	local SGD	87.5±7.02	55.47±5.20	41.59±7.71
Federated	FedAvg	70.65±10.64	40.15±7.25	34.26±6.01
	FOMO	88.72±5.41	52.44±5.09	44.56±4.31
	Ditto	87.32±6.42	54.28±5.31	42.73±5.19
Decentralized	D-PSGD(1s)	83.01±7.34	40.56±6.94	30.26±5.75
	D-PSGD(5e)	75.89±6.65	35.03±4.83	28.41±5.18
	CGA(1s)	65.65±12.66	30.81±10.79	27.65±11.78
	CGA(5e)	diverge	diverge	diverge
	SPDB(1s)	82.36±7.14	54.29±6.15	39.17±3.93
	SPDB(5e)	81.15±7.06	53.23±7.48	35.93±5.05
	Dada	85.65±6.36	57.61±5.45	37.81±7.15
	meta-L2C	92.10±4.71	58.28±3.09	48.80±4.17
SCooL (Ours)	SCool -SBM	91.37±5.03	58.76±4.30	48.69±5.21
	SCool -attention	**92.21±5.15**	**59.47±4.95**	**49.53±3.29**

Conclusion

we propose a **general probabilistic modelling framework**: Structured Cooperative Learning (SCooL), for DLPM problems.

❖SCooL jointly optimizes **personalized models $\boldsymbol{\theta}_{1:K}$** and **cooperation graph Y**.

❖Different graphical model priors of Y generate various novel DLPM algorithms.

❖SCooL uses a systematic optimization method: **variational inference**.

❖SCooL outperfroms previous federated / decentralized learning baselines in experiments.