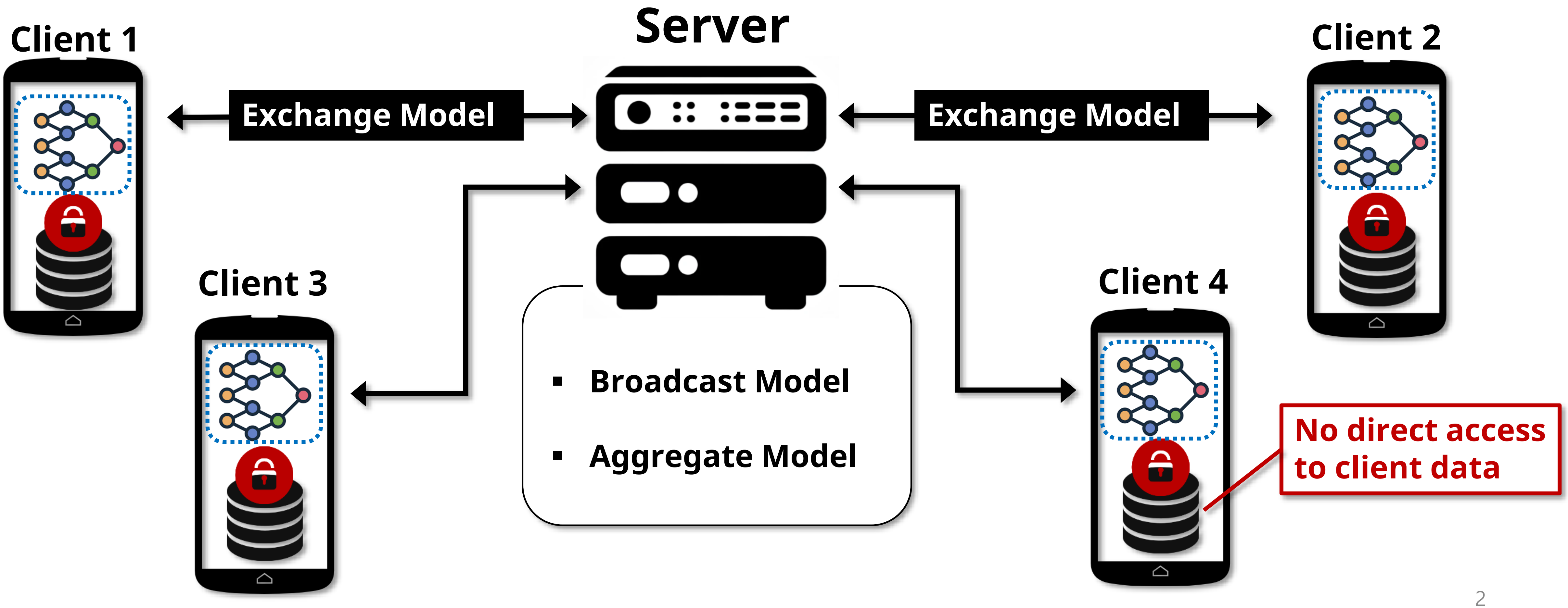
FedSOL: Stabilized Orthogonal Learning with Proximal Restrictions in Federated Learning

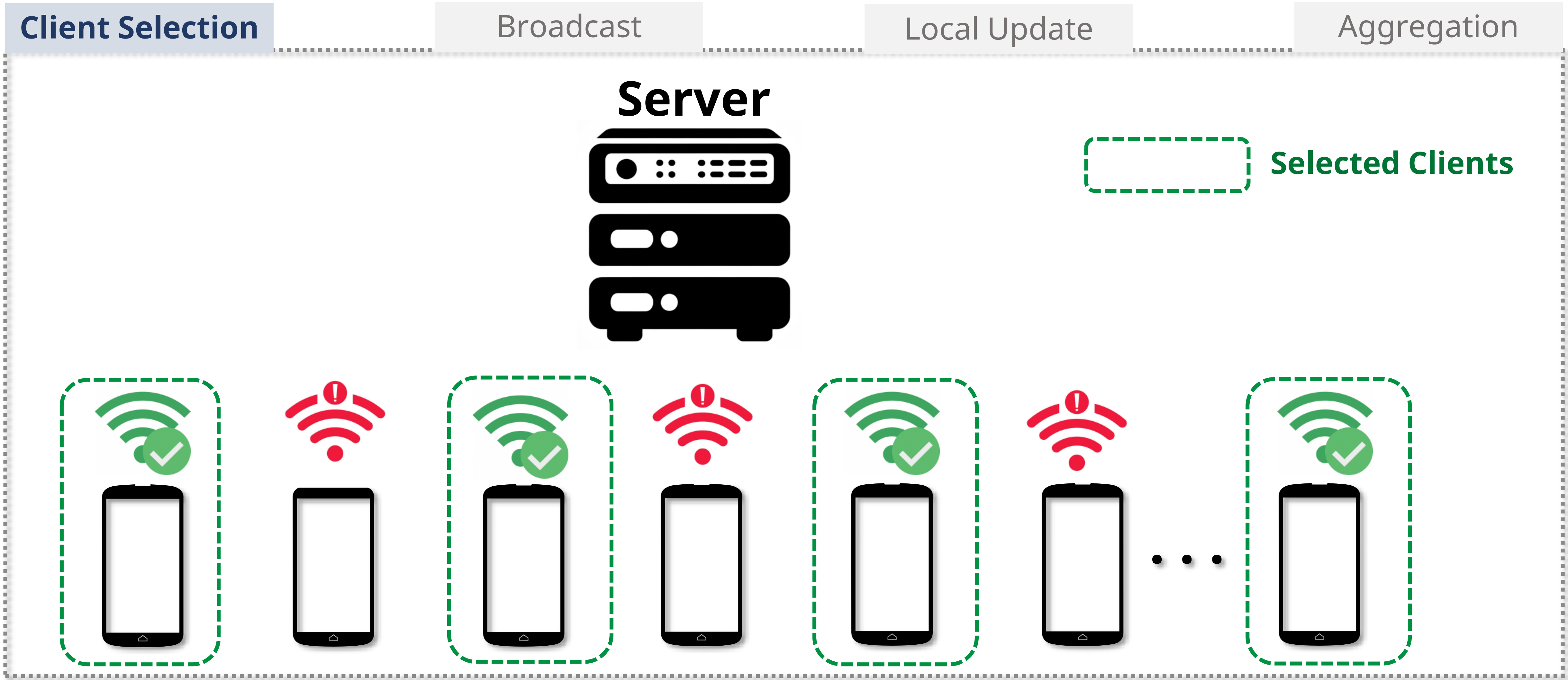
Gihun Lee, Minchan Jeong, Sangmook Kim, Jaehoon Oh,
and Se-Young Yun

1-1. Federated Learning

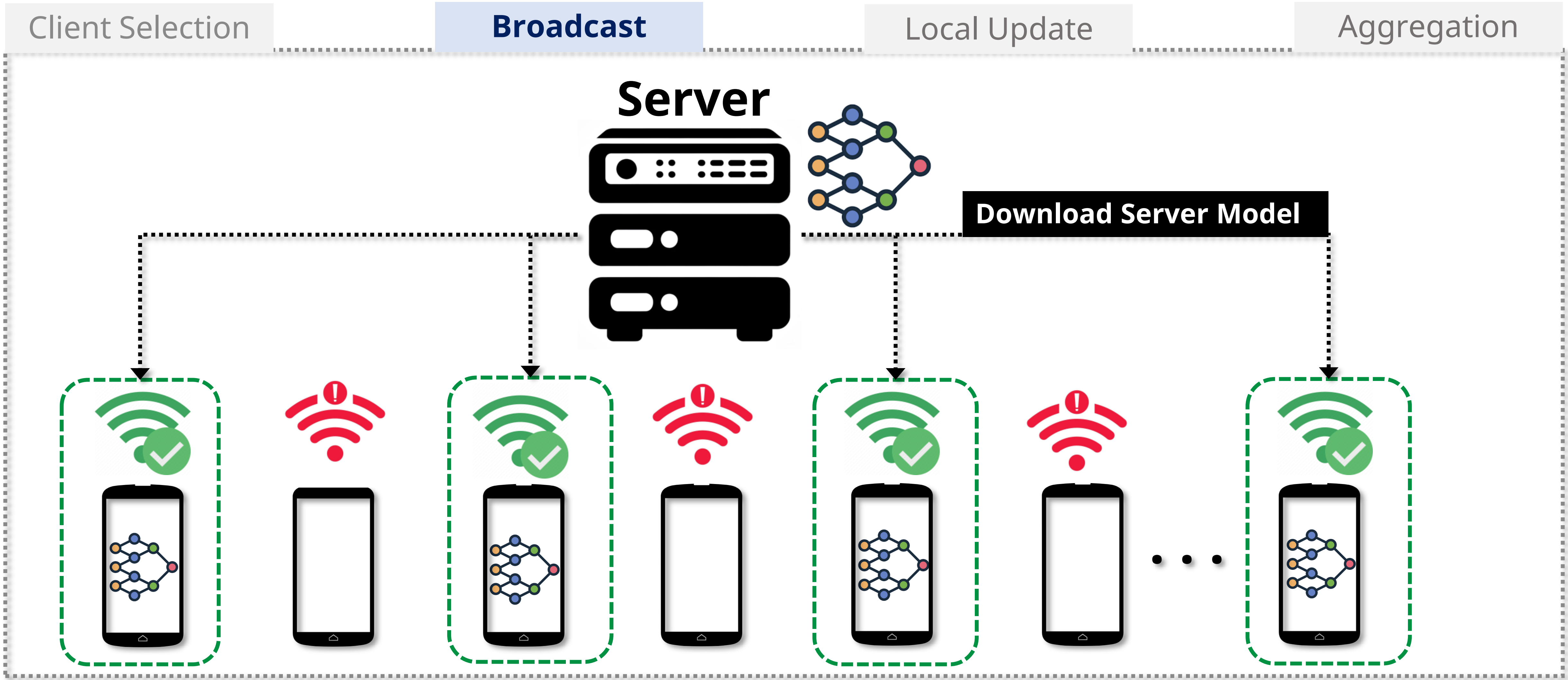
Federated Learning (FL) is a **Distributed Learning** paradigm in which multiple clients collaboratively learn a machine learning model while preserving their **Data Privacy**.



1-1. Federated Learning

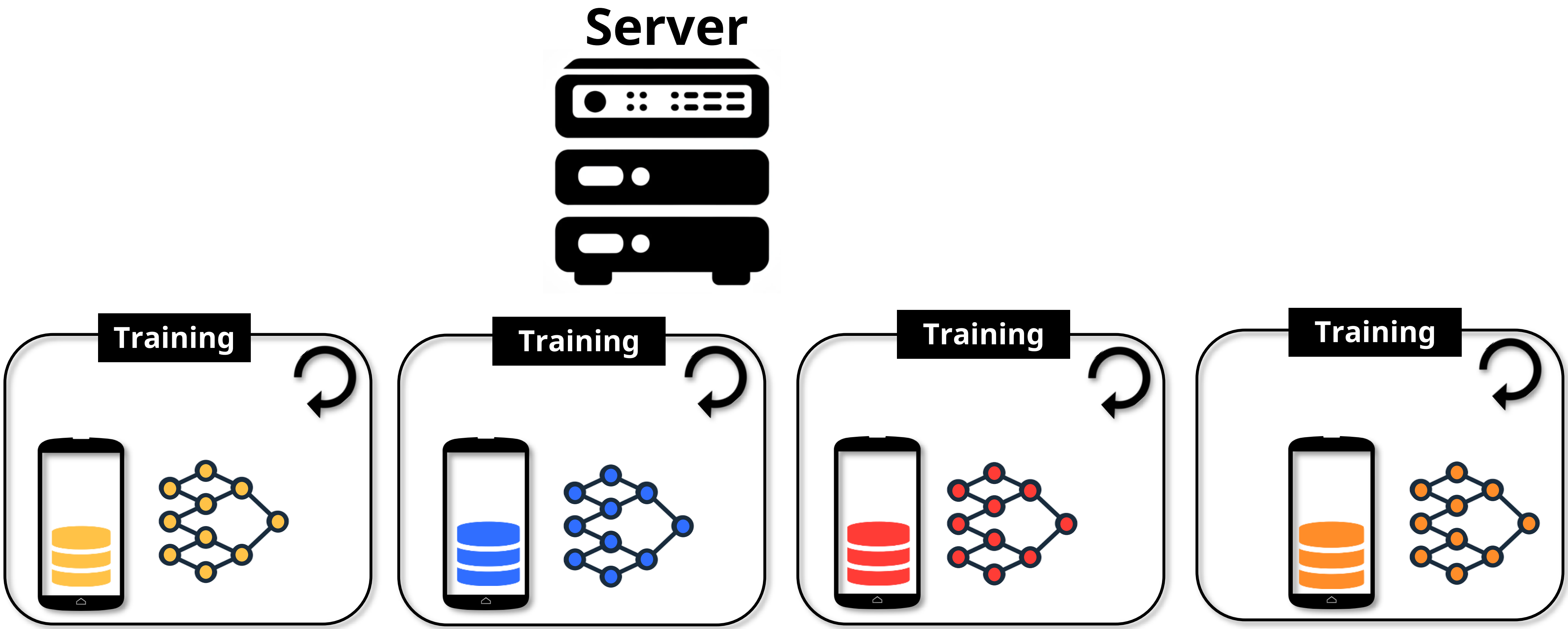


1-1. Federated Learning



1-1. Federated Learning

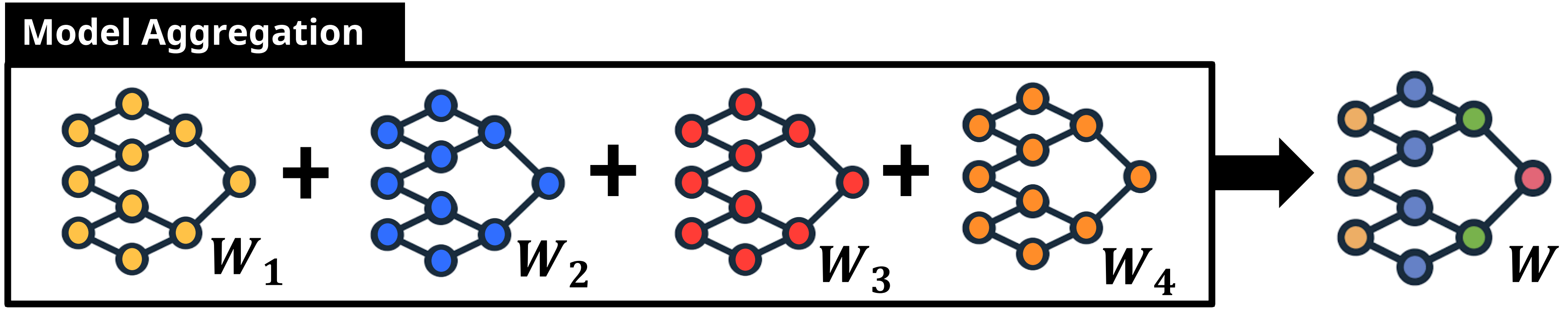
Client Selection | Broadcast | **Local Update** | Aggregation



1-1. Federated Learning

Client Selection | Broadcast | Local Update | **Aggregation**

Server



FedAvg (McMahan et al., 2018)

$$W^t = \frac{1}{K}\sum_{i=1}^{K} p_i W_i^{t-1}$$

Weighted Average of Local Models

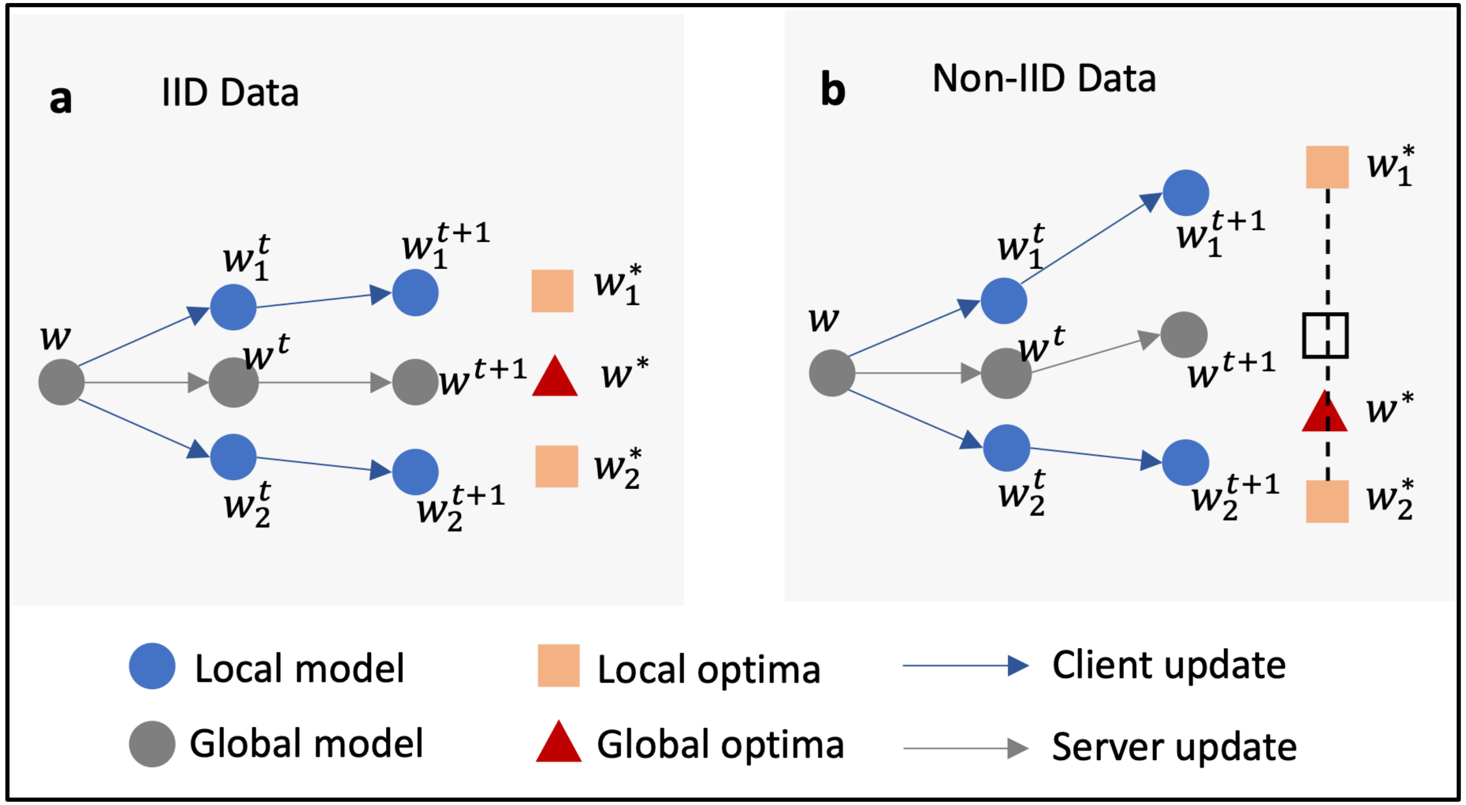
Konečný, Jakub, et al. "Federated learning: Strategies for improving communication efficiency." arXiv preprint arXiv:1610.05492 (2016).

Konečný, Jakub, et al. "Federated optimization: Distributed machine learning for on-device intelligence." arXiv preprint arXiv:1610.02527 (2016).

1-2. Data Heterogeneity

When client data is **heterogeneous**, FL suffers from "***Client Drift***".

→ How to accumulate knowledge from heterogeneous clients in a single **Global Model**?



Tan, Alysa Ziying, et al. "Towards personalized federated learning." IEEE Transactions on Neural Networks and Learning Systems (2022).

1-2. Data Heterogeneity

- **Sharding**: Uniform Size, Different Distributions
  (Higher Heterogeneity at **lower s**)

- **Latent Dirichlet Allocation** (LDA): Varying Size & Distributions
  (Higher Heterogeneity at **lower α**)

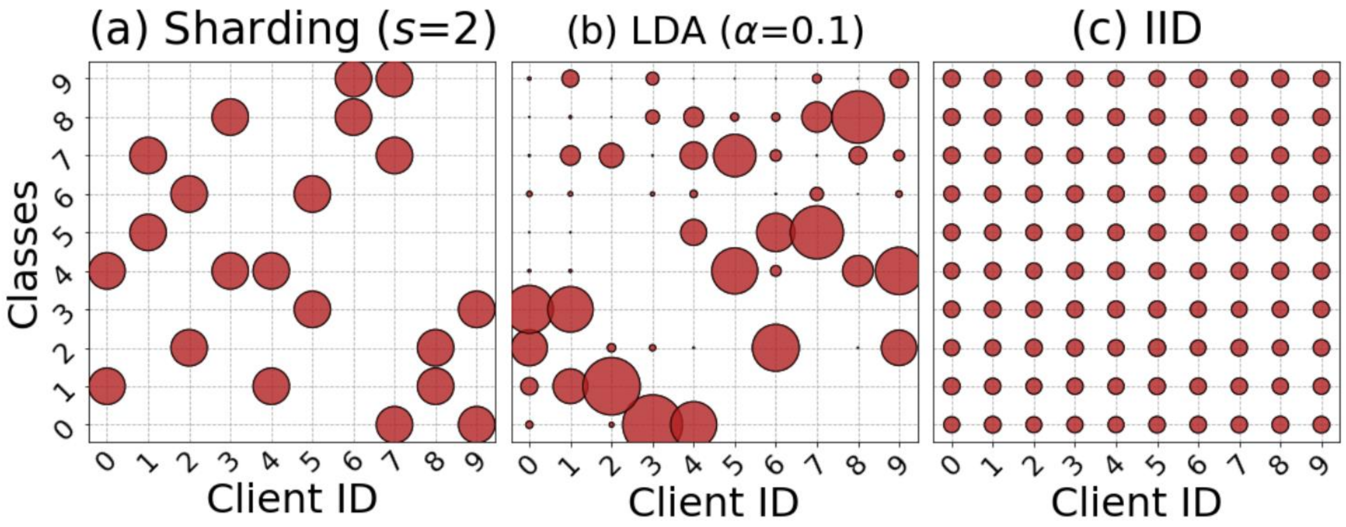


Figure 2. CIFAR-10 partition examples across 10 clients.

2. Motivation – Conflicting Objectives

Typical Local Objective:

$$\mathcal{L}^k(\boldsymbol{w}_k) = \mathcal{L}^k_{\text{local}}(\boldsymbol{w}_k) + \beta \cdot \mathcal{L}^k_p(\boldsymbol{w}_k; \boldsymbol{w}_g)$$

Original Objective
(e.g., CE Loss)

Proximal Objective
(e.g., FedProx, SCAFFOLD, FedKD…)

However, those two losses **conflicts** each other.

Knowledge Acquisition ⟷ **Knowledge Preservation**

2. Motivation – Proximal Objectives

Examples of **Proximal Objectives**:

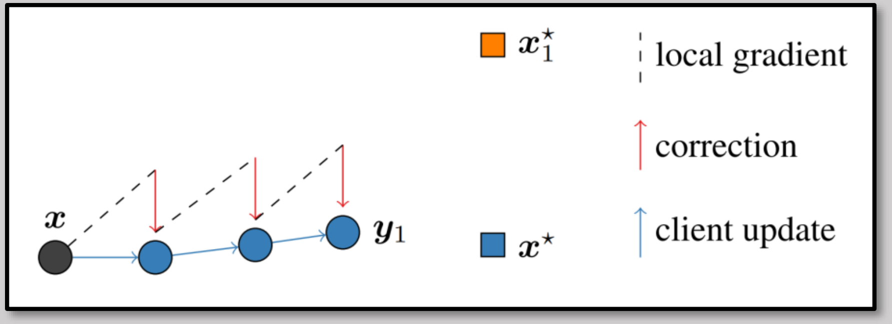
FedGKT (NeurIPS 2020) $\mathcal{L}_c = \mathcal{L}_{CE} + D_{KL}(p_s || p_k)$

FedProx (ICML 2019) $F_k(w) + \frac{\mu}{2}||w - w^t||^2$

SCAFFOLD (ICML 2020) $y_i \leftarrow y_i - \eta_l(g_i(y_i) + c - c_i)$

FedDyn (ICLR 2021) $\mathcal{L}_k(\theta) - \langle \nabla \mathcal{L}_k(\theta_k^{t-1}), \theta \rangle + \frac{\alpha}{2}||\theta - \theta^{t-1}||^2$

MOON (CVPR 2021) $\ell_{sup}\left(w_i^t; (x, y)\right) + \mu \ell_{con}(w_i^t; w_i^{t-1}; w_t; x)$

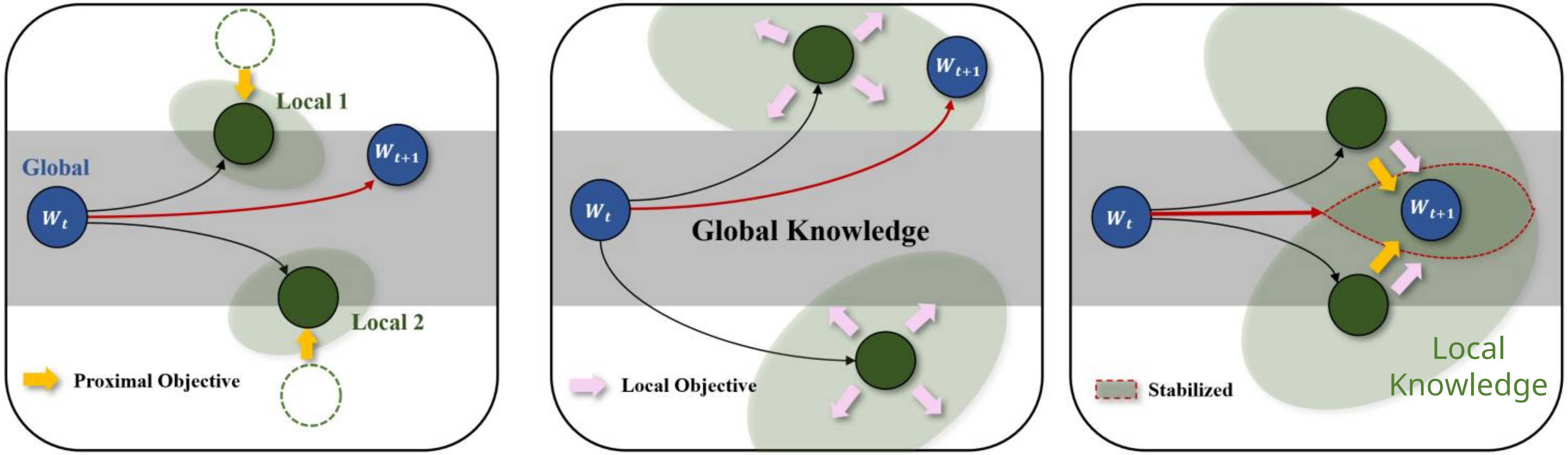


$$\mathcal{L}^k(w_k) = \mathcal{L}^k_{\text{local}}(w_k) + \beta \cdot \mathcal{L}^k_p(w_k; w_g)$$

Original Objective **Proximal Objective**

2. Motivation – Global vs Local Knowledge

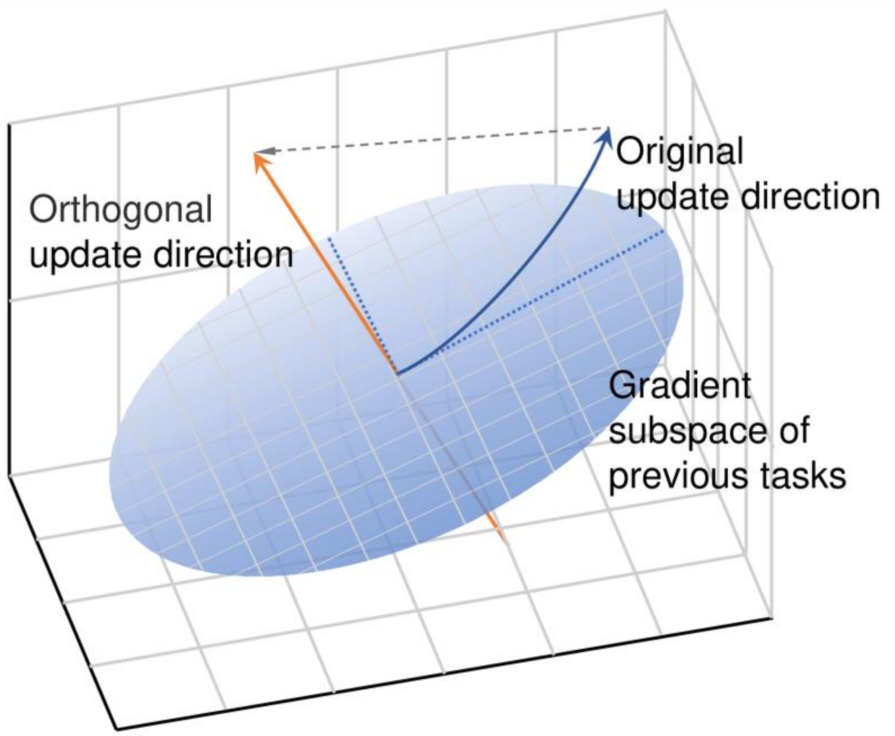
FL must navigate the **balance between those two conflicting objectives**.



(a) Global Knowledge Preservation (b) Local Knowledge Acquisition (c) FedSOL (Ours)

"How should the local learning **acquire local knowledge** while *minimizing its conflicts* with **global knowledge**?"

2. Motivation – Orthogonal Learning



Orthogonal Learning in CL

Find update direction that is orthogonal to the previous task.

Challenges to implement to FL

- Cannot retain past data or gradients for reference for global knowledge.
- The global distribution overlaps with individual local distributions.

2. Motivation – Gradient Projection

Proximal Gradient Projection

$$\boldsymbol{g}_u^{\text{Proj}} = \boldsymbol{g}_l - \frac{{\boldsymbol{g}_l}^T \boldsymbol{g}_p}{\boldsymbol{g}_p^T \boldsymbol{g}_p} \boldsymbol{g}_p \qquad \text{if} \quad \boldsymbol{g}_l^T \boldsymbol{g}_p < 0$$

g_l: Gradient on **Original Local Loss**

g_p: Gradient on **Proximal Loss**

g_u: Gradient for Parameter Update

Experiment: CIFAR-10 ($\alpha = 0.1$)

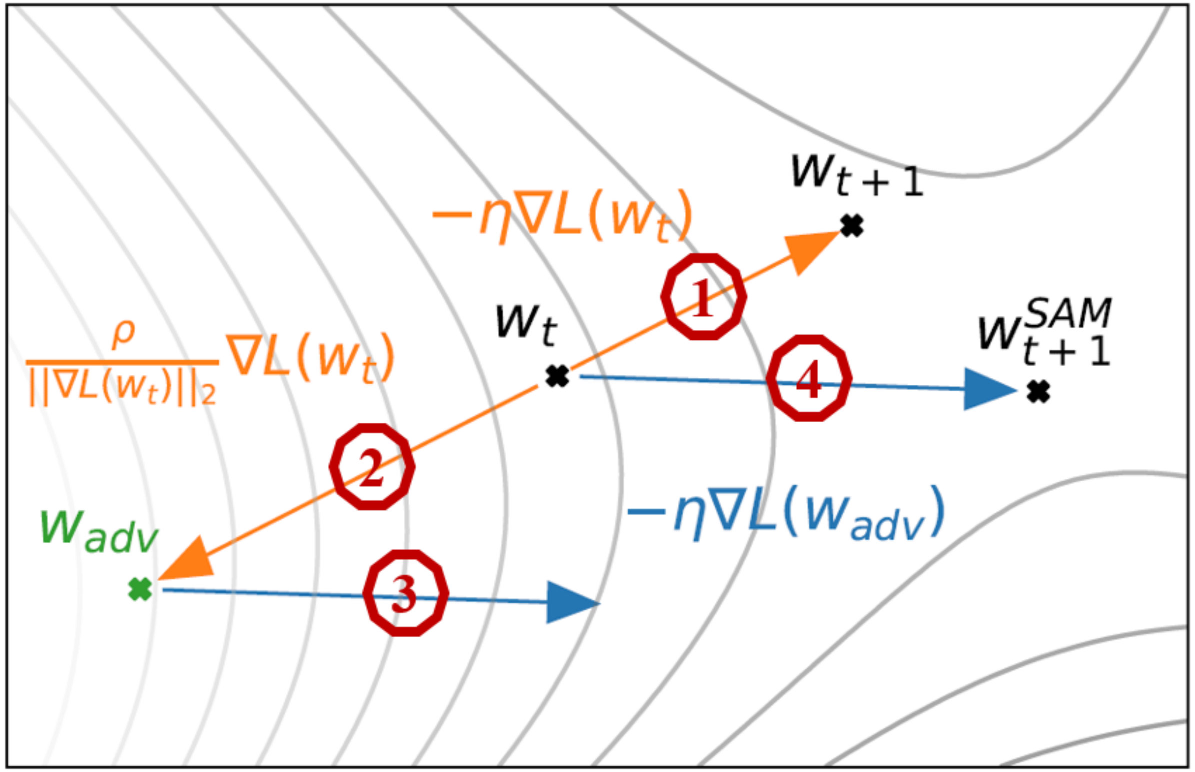
Proximal Loss	**Usage** Base	Projection
None (FedAvg)	$56.13_{\pm 0.78}$	
L2 Distance	**59.80**$_{\pm 1.12}$	$56.35_{\pm 2.85}$ (- 3.45)
KL-Divergence	**60.31**$_{\pm 2.07}$	$50.88_{\pm 3.55}$ (- 9.43)

Directly negating the **proximal gradient (g_p)** rather *undermines* the local learning.

3. Main Approach - Preliminary: SAM

3 FedSOL

Sharpness-Aware Minimization (SAM)



Perturb & **Update** *(using same $\mathcal{L}$)*

$$\min_{\boldsymbol{w}} \max_{||\boldsymbol{\epsilon}||_2<\rho} \mathcal{L}(\boldsymbol{w}+\boldsymbol{\epsilon})$$

Use single gradient step.

$$\epsilon^* = \rho \frac{\nabla_w L(w)}{||\nabla_w L(w)||_2}$$

3. Main Approach - Stabilized Orthogonal Learning

3 FedSOL

[Main Idea]: Perturb using "**Global Knowledge**" & Update to acquire "**Local Knowledge**"

Step 1: Weight Perturbation

$$\epsilon_p^* = \underset{\|\epsilon\|_2 \leq \rho}{\operatorname{argmax}} \mathcal{L}_p^k(\boldsymbol{w}_k + \epsilon; \boldsymbol{w}_g) \approx \rho \frac{\boldsymbol{g}_p}{\|\boldsymbol{g}_p\|_2}$$

Perturbation Strength

Adversarial perturb using **Proximal Loss L_p^k**

Step 2: Parameter Update

$$\boldsymbol{w}_k \leftarrow \boldsymbol{w}_k - \gamma \cdot \nabla_{\boldsymbol{w}_k} \mathcal{L}_{\text{local}}^k(\boldsymbol{w}_k + \epsilon_p^*)$$

Update the model with the **Original Local Loss L_{local}^k**.

3. Main Approach - Stabilized Orthogonal Learning

3 FedSOL

[Main Idea]: Regularize **Local Loss Surface** using "**Global Perspective**"

Local gradient and its direction

$$g_l = \nabla_w \mathcal{L}^k_{local}(w_k) \qquad \widehat{g_l} = \frac{g_l}{\|g_l\|_2}$$

Proximal gradient and its direction

$$g_p = \nabla_w \mathcal{L}^k_p(w_k) \qquad \widehat{g_p} = \frac{g_p}{\|g_p\|_2}$$

FedSAM

$$g_u^{FedSAM} = \nabla_{w_k} \mathcal{L}^k_{local}(w_k + \epsilon^*)$$

$$\approx \boldsymbol{g_l} + \rho\, \nabla^2_{w_k} \mathcal{L}^k_{local}(w_k)\, \widehat{\boldsymbol{g_l}}$$

Regularize Loss Surface w.r.t. "**Local Loss**"

FedSOL

$$g_u^{FedSOL} = \nabla_{w_k} \mathcal{L}^k_{local}(w_k + \epsilon^*_p)$$

$$\approx \boldsymbol{g_l} + \rho\, \nabla^2_{w_k} \mathcal{L}^k_{local}(w_k)\, \widehat{\boldsymbol{g_p}}$$

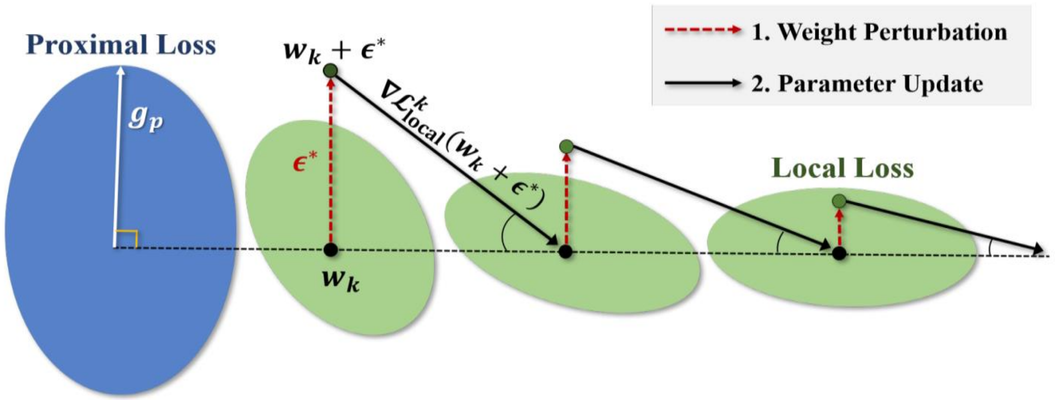
Regularize Loss Surface w.r.t. "**Proximal Loss**"

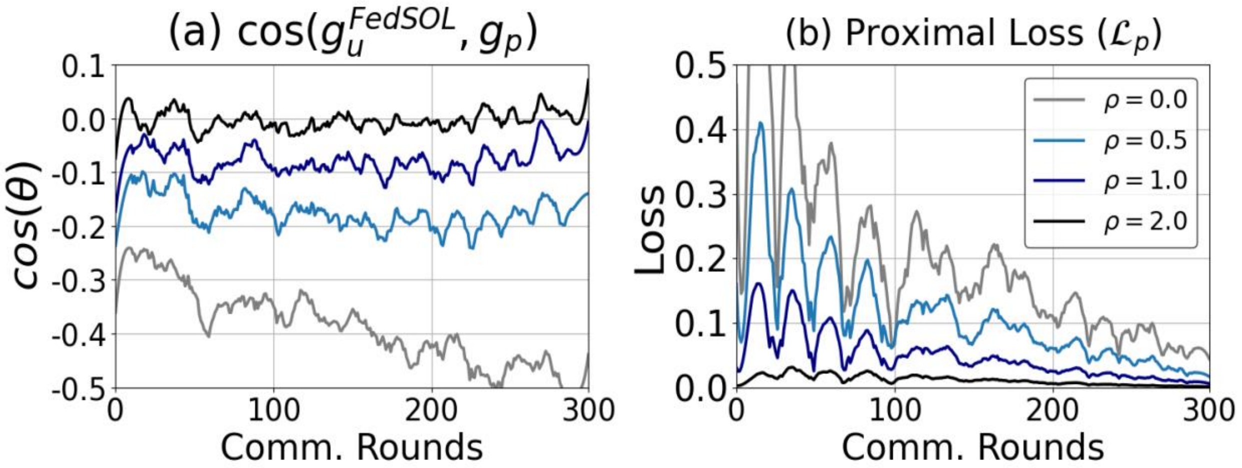
3. Main Approach - Stabilized Orthogonal Learning

3 FedSOL

[Main Idea]: Identify the **local gradient is minimally affected by proximal gradient** finds the *orthogonal direction* of local knowledge acquisition.

Experiment: CIFAR-10 ($\alpha = 0.1$)

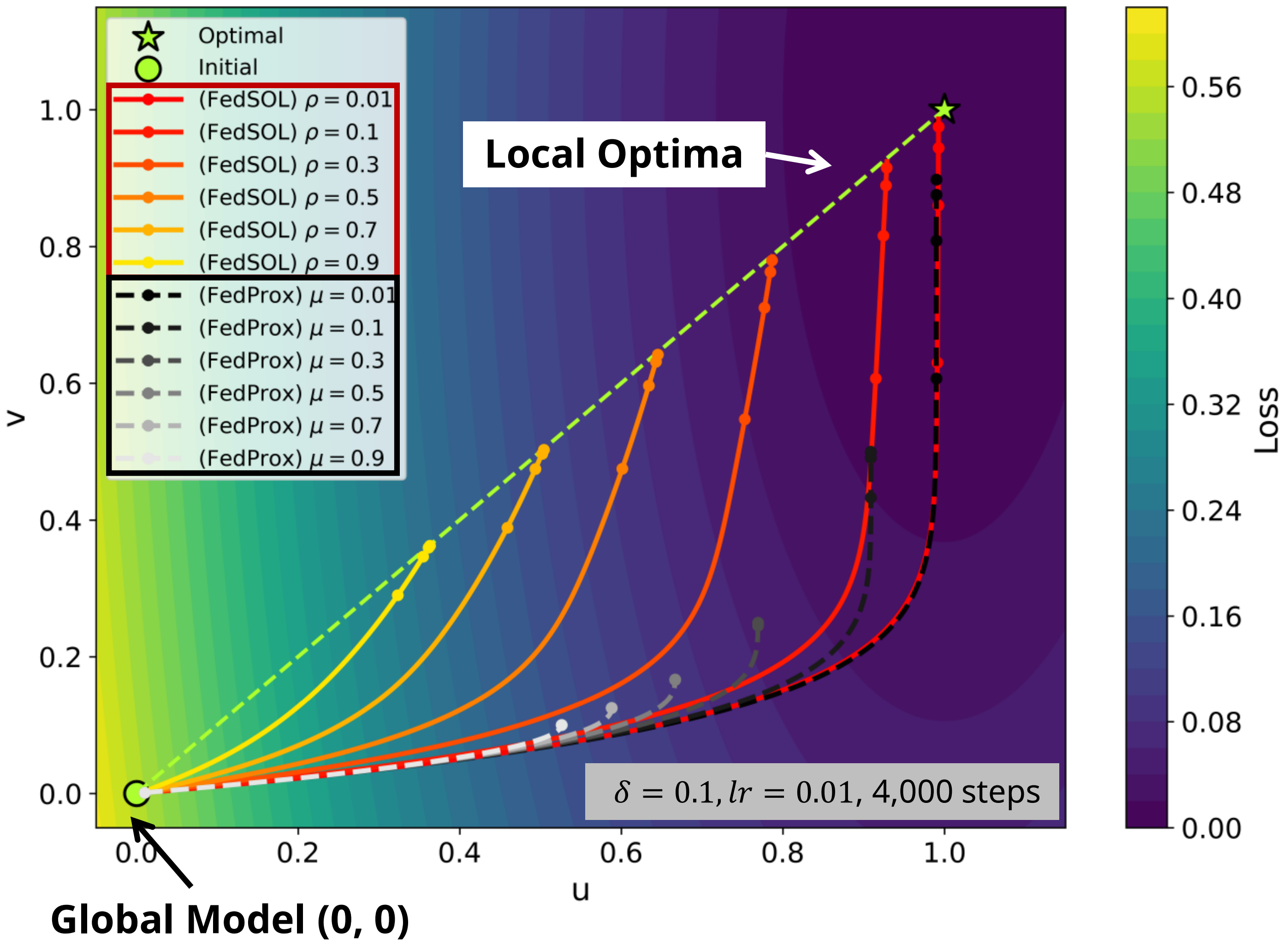




3. Main Approach – Toy Example (2-dimensional)

3 FedSOL

Optimization Path for FedProx and FedSOL



Local Loss

Distribution coefficient

$$\mathcal{L}_{local}(\boldsymbol{u},\boldsymbol{v}) = \frac{1}{2}(u-u_l)^2 + \frac{\delta}{2}(v-v_l)$$

Proximal Loss

$$\mathcal{L}_{p}(\boldsymbol{u},\boldsymbol{v}) = \frac{\mu}{2}(u^2+v^2)$$

- **FedProx** results in *skewed* update that influenced by local distribution.
- **FedSOL** *aligns* parallel to the local optima regardless of δ.

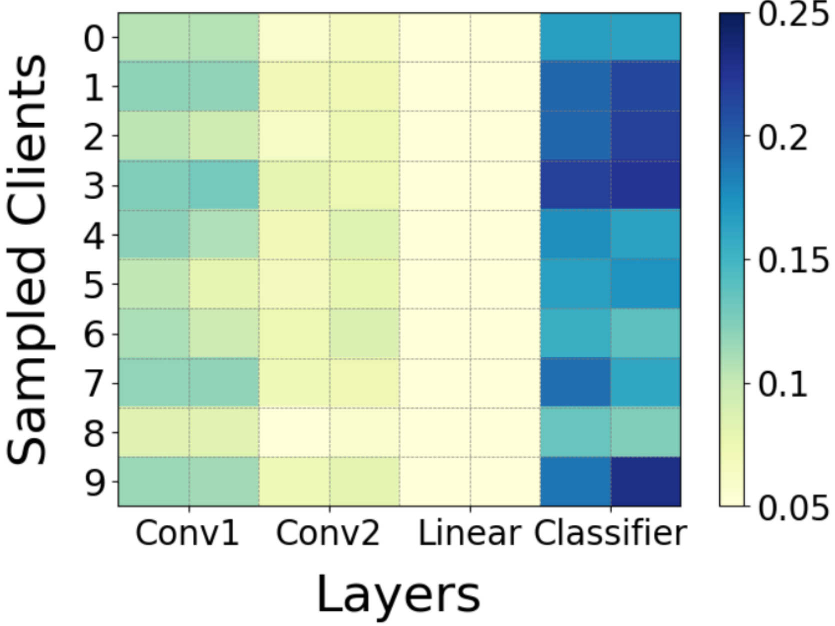
3. Main Approach – C. Partial Perturbation

3 FedSOL

[Main Idea]: Perturbing only the last classifier layer is sufficient for FedSOL. The performance reaches as high as full-model perturbation.

Experiment: CIFAR-10 ($\alpha = 0.1$)

Deviation: $\|w_k - w_g\|$



Target Position	Perturbation (ρ) 0.0	0.5	1.0	1.5	2.0	FLOPs
All (*full*)		61.17	**64.16**	**64.38**	**63.94**	$2\times +\delta$
Body (*partial*)	56.13	60.98	62.95	63.94	63.80	$1.96\times +\delta$
Head (*partial*)		**62.65**	63.62	64.13	63.25	$1.33\times +\delta$

- δ : Computation for the proximal loss.
- **FLOPs** shows relative computation w.r.t. **FedAvg.**

4. Main Experiment – Data Heterogeneity

4 Results

Experiment: Test Accuracy @1 (%). The values in ( ) are the standard deviations.

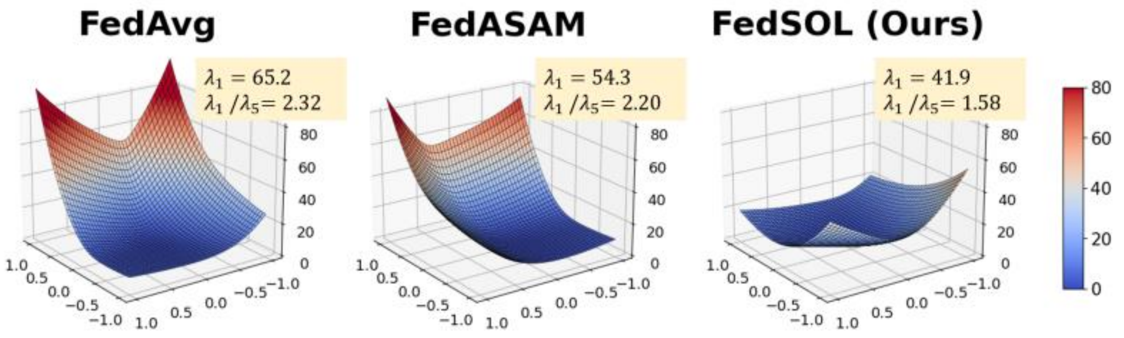
Non-IID Partition Strategy : LDA									
Method	**MNIST**	**CIFAR-10** $\alpha = 0.05$	**CIFAR-10** $\alpha = 0.1$	**CIFAR-10** $\alpha = 0.3$	**CIFAR-10** $\alpha = 0.5$	**SVHN**	**CINIC-10**	**PathMNIST**	**TissueMNIST**
FedAvg [46]	96.11 (0.19)	42.27 (1.34)	56.13 (0.78)	67.32 (0.94)	73.90 (0.66)	55.36 (4.85)	36.49 (4.37)	65.98 (4.76)	42.78 (2.03)
FedProx [38]	96.05 (0.13) ↓	50.58 (0.57) ↑	59.80 (1.12) ↑	68.39 (0.81) ↑	72.87 (0.55) ↑	72.40 (3.15) ↑	40.09 (3.97) ↑	70.44 (1.92) ↑	52.25 (1.40) ↑
FedNova [64]	88.24 (1.37) ↓	10.00 (*Failed*) ↓	10.00 (*Failed*) ↓	64.67 (0.77) ↓	70.04 (0.45) ↓	53.07 (3.30) ↓	21.89 (1.71) ↓	38.94 (2.34) ↓	15.03 (3.74) ↓
Scaffold [28]	94.18 (0.32) ↓	10.00 (*Failed*) ↓	10.00 (*Failed*) ↓	**71.92** (0.17) ↑	**75.49** (0.21) ↑	21.46 (1.75) ↓	16.89 (2.25) ↓	18.07 (0.04) ↓	32.04 (0.07) ↓
FedNTD [33]	96.97 (0.27) ↑	58.08 (0.48) ↑	**63.16** (1.02) ↑	71.56 (0.26) ↑	74.91 (0.33) ↑	79.25 (0.61) ↑	50.22 (3.71) ↑	74.26 (1.25) ↑	44.55 (1.95) ↑
FedSAM [55]	95.72 (0.43) ↓	36.14 (1.21) ↓	52.14 (0.94) ↓	64.83 (0.56) ↓	70.74 (0.40) ↓	13.27 (2.78) ↓	36.70 (4.28) ↑	66.64 (3.76) ↑	44.07 (3.02) ↑
FedASAM [6]	96.60 (0.10) ↑	43.12 (1.25) ↑	57.00 (0.30) ↑	67.45 (0.92) ↑	73.91 (0.51) ↑	60.25 (4.56) ↑	36.93 (4.60) ↑	69.45 (3.19) ↑	42.73 (2.35) ↑
FedSOL (Ours)	**97.44** (0.11) ↑	**60.01** (0.30) ↑	**64.13** (0.46) ↑	**71.94** (0.57) ↑	**75.60** (0.32) ↑	**83.92** (0.29) ↑	**55.07** (1.48) ↑	**78.88** (0.46) ↑	**53.40** (0.85) ↑

FedSOL achieves best results in most cases,
showing its robustness against data heterogeneity.

4. Main Experiment – Analysis

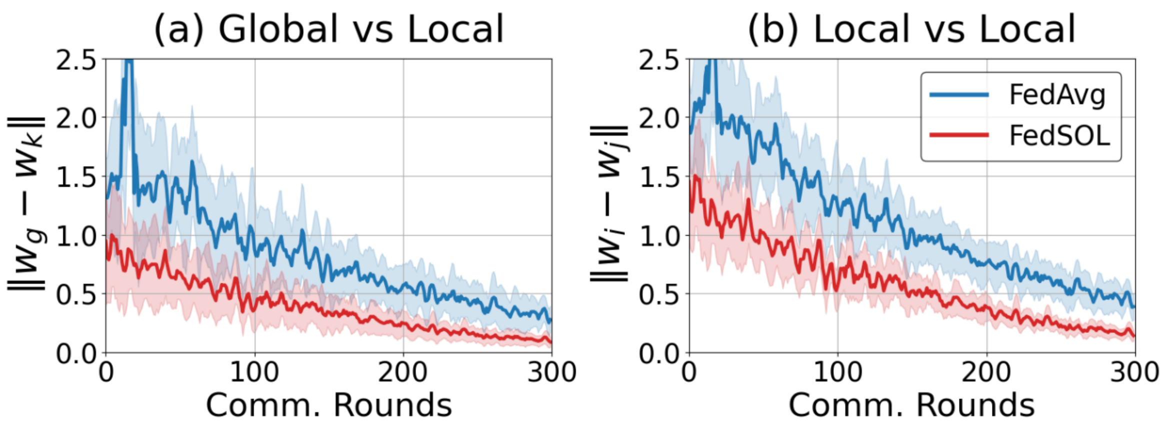
4 Results

Smoothness of Global Model

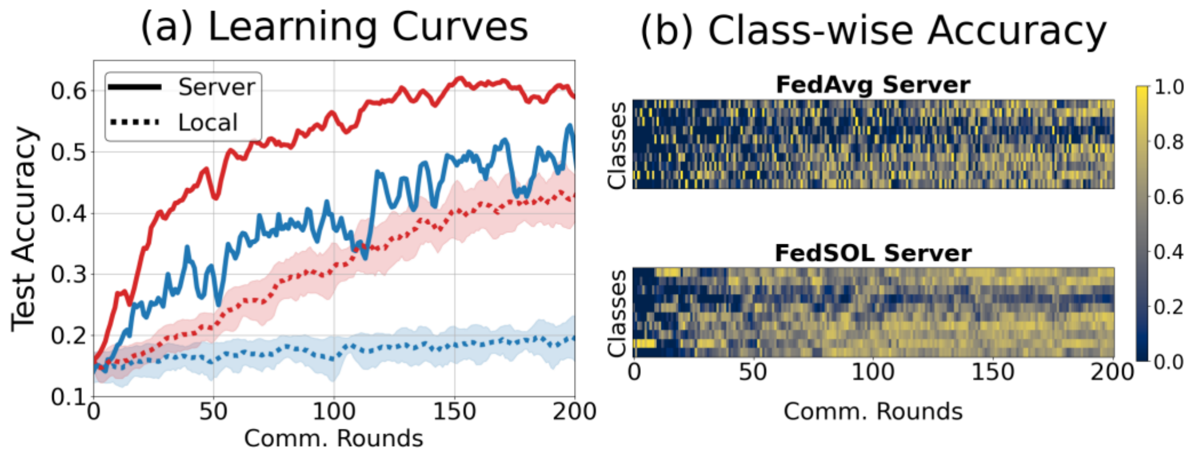


FedSOL stabilizes federated learning in various perspectives.

Weight Divergence



Knowledge Preservation



5. Conclusion

5 Takeaway

- FL must strike the balance between **global knowledge preservation** (Proximal Loss) and **local knowledge acquisition** (Original Local Loss).
- We suggest that **orthogonal learning in CL** could be an effective strategy in FL, by **resolving conflicts of local knowledge on global knowledge**.
- We propose **FedSOL**, which aims to obtain the **local gradient which is orthogonal to the proximal gradient**.
- **FedSOL** acquire local knowledge during local learning that less conflicts with the global knowledge.

Thank You!