A-FedPD: Aligning Dual-Drift is All Federated Primal-Dual Learning Needs

Yan Sun[1], Li Shen[2], Dacheng Tao[3]

[1]The University of Sydney [2]Sun Yat-sen University [3]Nanyang Technological University

Introduction

As a popular paradigm for juggling data privacy and collaborative training, federated learning (FL) is flourishing to distributively process the large scale of heterogeneous datasets on edged clients. From an optimization perspective, two mainstream frameworks have gained widespread applications.

	Federated Primal Methods	Federated Primal-dual Methods
Basis	mandatory aggregation	constraints optimization
Sub-problems splitting	natural operator splitting	Lagrangian operator splitting
Advantages	1. easy to implement	1. support longer local training 2. fewer communications

Federated ADMM (Primal-dual) Family

The consensus finite-sum minimization:

$$\min_{\theta,\theta_i} \frac{1}{C} \sum_{i\in[C]} f_i(\theta_i), \qquad s.t.\, \theta_i = \theta \;\; for\, \forall\, i.$$

Augmented Lagrangian function:

$$L = \frac{1}{C} \sum_{i\in[C]} f_i(\theta_i) + \langle \lambda_i, \theta_i - \theta \rangle + \frac{\rho}{2} \|\theta_i - \theta\|^2 .$$

Alternating multiplier optimization: $\{\theta i \rightarrow \theta \rightarrow \lambda i \rightarrow \ldots \ldots\}$

When Primal Dual meets Partial Participation

Partial participation has led to the incompleteness of local problem solutions, and we have summarized the following three frameworks.

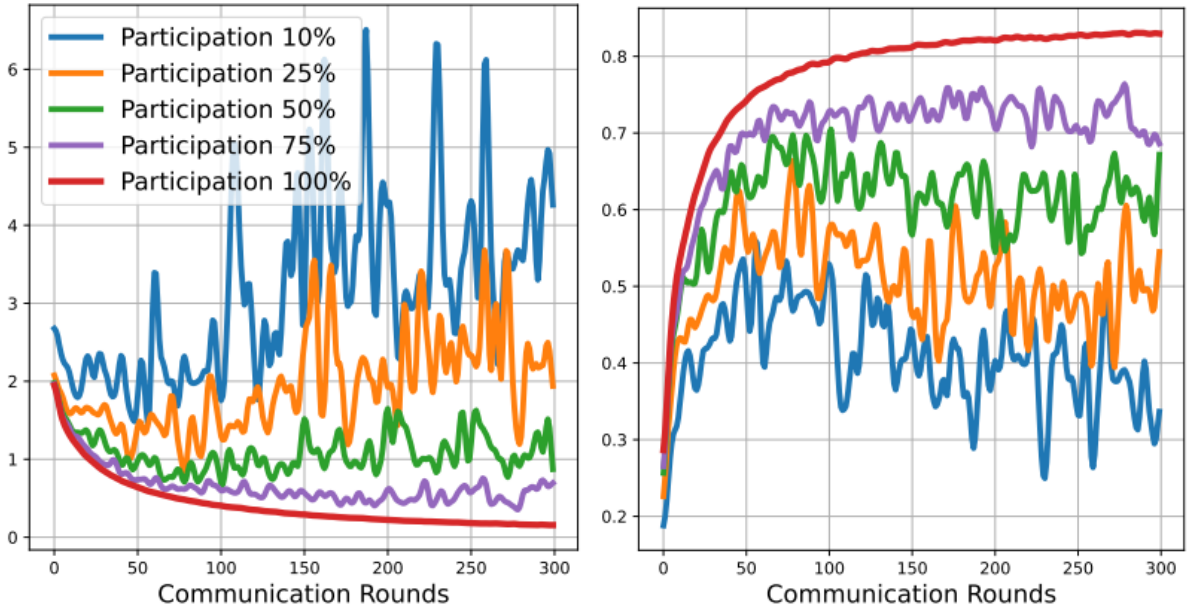
	FedPD	FedADMM	FedDyn
Global consensus	Average in [C]	Average in [P]	Average in [P]
Global dual	Average in [C]	Average in [P]	Average in [C]
Local primal	[C]	[P]	[P]

Average in [C]: average the parameters of all clients

Average in [P]: average the parameters of participated clients

Instability in Federated Primal Dual Family

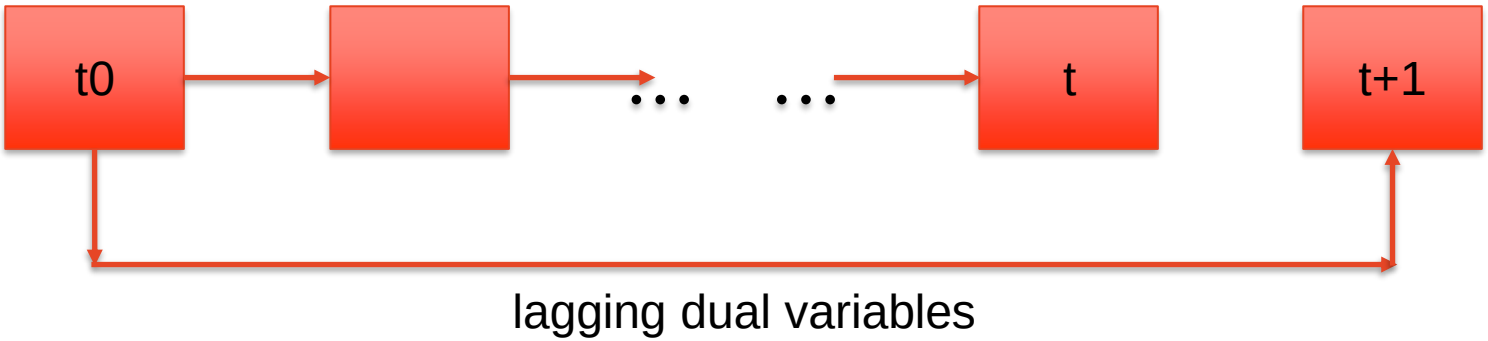
- **significant biases caused by partial participation**



(a) Train loss. (b) Test accuracy.

"Dual Drift"

- At round t, $\theta^{t+1} = argmin_{\theta} L(\theta^t, \lambda_i^t)$ for $i \in [P^t]$. At round $t+1$, when a client $i \notin [P^\tau]_{\tau=t_0+1}^{t}$ ($t_0 \ll t$) is activated, the local sub-problem $L(\theta^t, \lambda_i^{t_0})$ will fall into a very unstable state due to the mismatch between the primal and dual variables.



Our Method: virtual dual update

Algorithm 1 A-FedPD Algorithm

Input: $\theta^0, \theta_i^0, T, K, \lambda_i^0, \rho$
Output: global average model
1: **Initialization** : $\theta_i^0 = \theta^0, \lambda_i^0 = 0$.
2: **for** $t = 0, 1, 2, \cdots, T-1$ **do**
3: randomly select active clients set $\mathcal{P}^t$ from $\mathcal{C}$
4: **for** client $i \in \mathcal{P}^t$ **in parallel do**
5: receive λ_i^t, θ^t from the global server
6: $\theta_i^{t+1} = LocalTrain(\lambda_i^t, \theta^t, \eta^t, K)$
7: send θ_i^{t+1} to the global server
8: **end for**
9: $\overline{\theta}^{t+1} = \frac{1}{P}\sum_{i\in\mathcal{P}^t}\theta_i^{t+1}$
10: $\lambda_i^{t+1} = D\text{-}Update(\lambda_i^t, \theta^t, \theta_i^{t+1}, \overline{\theta}^{t+1}, \mathcal{P}^t)$
11: $\overline{\lambda}^{t+1} = \frac{1}{C}\sum_{i\in\mathcal{C}}\lambda_i^{t+1}$
12: $\theta^{t+1} = \overline{\theta}^{t+1} + \frac{1}{\rho}\overline{\lambda}^{t+1}$
13: **end for**
14: return global average model

◇ *LocalTrain*: (Optimize Eq.(4))
Input: $\lambda_i^t, \theta^t, \eta^t, K$
Output: $\theta_{i,K}^t$
1: **for** $k = 0, 1, 2, \cdots, K-1$ **do**
2: calculate the stochastic gradient $g_{i,k}^t$
3: $\theta_{i,k+1}^t = \theta_{i,k}^t - \eta^t(g_{i,k}^t + \lambda_i^t + \rho(\theta_{i,k}^t - \theta^t))$
4: **end for**

◇ *D-Update*: (update dual variables)
Input: $\lambda_i^t, \theta^t, \theta_i^{t+1}, \overline{\theta}^{t+1}, \mathcal{P}^t$
Output: λ_i^{t+1}
1: **if** $i \in \mathcal{P}^t$ **then**
2: $\lambda_i^{t+1} = \lambda_i^t + \rho_t(\theta_i^{t+1} - \theta^t)$
3: **else**
4: $\lambda_i^{t+1} = \lambda_i^t + \rho_t(\overline{\theta}^{t+1} - \theta^t)$
5: **end if**

General updates via the true model θi

Virtual updates via the constructed θ

Optimization and Generalization

Theorem 1 *Let non-convex objective f satisfies Assumption 1, let ρ be selected as a non-zero positive constant, $\{\overline{\theta}^t\}_{t=0}^{T}$ sequence generated by algorithm 1 satisfies:*

$$\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}\|\nabla f(\overline{\theta}^t)\|^2 \leq \frac{\rho\left[f(\overline{\theta}^1)-f^\star\right]+R_0}{T}+\mathcal{O}(\epsilon),$$

where $f^\star$ is the optimum and $R_0=\frac{1}{C}\sum_{i\in\mathcal{C}}\mathbb{E}_t\|\theta_i^1-\theta^0\|^2$ is the first local training volumes.

Theorem 2 *Let non-convex objective f satisfies Assumption 1 and 2 and $H=\sup_{\theta,\xi}f(\theta,\xi)$, after T communication rounds training with Algorithm 1, the generalization error bound achieves:*

$$\mathbb{E}\left[F(\theta^T)-f(\theta^T)\right]\leq\frac{\kappa_c}{CS}(HPT)^{\frac{\mu L}{1+\mu L}},$$

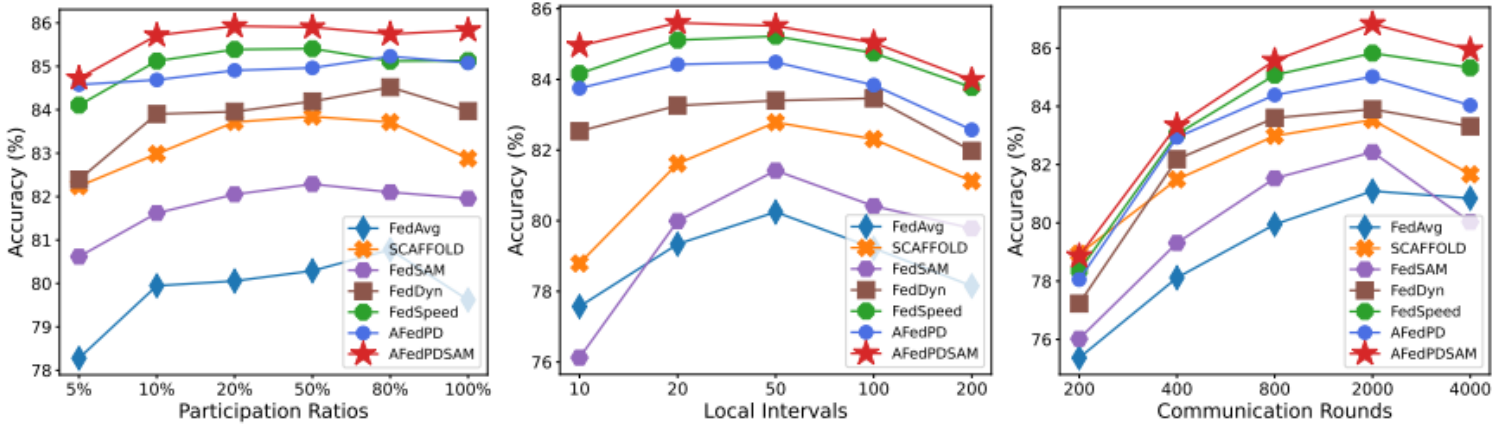
where μ is a constant related to the learning rate and $\kappa_c=4\left(G^2/L\right)^{\frac{1}{1+\mu L}}$ is a constant.

Experimental Results

Table 2: Test accuracy on the CIFAR-10 / 100 dataset. We fix the total client $C = 100$ and $P = 10$ under training local 50 iterations. We test 3 setups of IID, Dir-1.0, and Dir-0.1 on each dataset. Each group is tested on LeNet (upper portion) and ResNet-18 (lower portion) models. Each results are tested with 4 different random seeds. "–" means can not stably converge. "Family" distinguishes whether the algorithm is a primal method (P) or a primal dual method (PD) and "Local Opt" distinguishes whether the algorithm adopts SGD-based or SAM-based local optimizer.

			CIFAR-10			CIFAR-100		
	Family	Local Opt	IID	Dir-1.0	Dir-0.1	IID	Dir-1.0	Dir-0.1
FedAvg	P	SGD	$81.87_{\pm.12}$	$80.58_{\pm.15}$	$75.57_{\pm.27}$	$40.11_{\pm.17}$	$39.65_{\pm.07}$	$38.37_{\pm.14}$
FedCM	P	SGD	$80.34_{\pm.14}$	$79.31_{\pm.33}$	$72.89_{\pm.37}$	$43.33_{\pm.13}$	$42.35_{\pm.25}$	$37.11_{\pm.51}$
SCAFFOLD	P	SGD	$84.25_{\pm.16}$	$83.61_{\pm.14}$	$78.66_{\pm.29}$	$49.65_{\pm.06}$	$49.11_{\pm.14}$	$46.36_{\pm.30}$
FedSAM	P	SAM	$83.22_{\pm.09}$	$81.94_{\pm.13}$	$77.41_{\pm.36}$	$43.02_{\pm.09}$	$42.83_{\pm.29}$	$42.29_{\pm.23}$
FedDyn	PD	SGD	$84.49_{\pm.22}$	$84.20_{\pm.14}$	$79.51_{\pm.13}$	$50.27_{\pm.11}$	$49.64_{\pm.21}$	$46.30_{\pm.26}$
FedSpeed	PD	SAM	$86.01_{\pm.18}$	$85.11_{\pm.21}$	$80.86_{\pm.18}$	$54.01_{\pm.15}$	$53.45_{\pm.23}$	$51.28_{\pm.18}$
A-FedPD	PD	SGD	$85.31_{\pm.14}$	$84.94_{\pm.13}$	$80.28_{\pm.20}$	$51.41_{\pm.15}$	$51.17_{\pm.17}$	$48.15_{\pm.28}$
A-FedPDSAM	PD	SAM	$\mathbf{86.47}_{\pm.18}$	$\mathbf{85.90}_{\pm.29}$	$\mathbf{81.96}_{\pm.19}$	$\mathbf{55.56}_{\pm.27}$	$\mathbf{54.62}_{\pm.16}$	$\mathbf{53.15}_{\pm.19}$
FedAvg	P	SGD	$81.67_{\pm.21}$	$80.94_{\pm.17}$	$76.24_{\pm.35}$	$44.68_{\pm.21}$	$44.27_{\pm.25}$	$41.64_{\pm.27}$
FedCM	P	SGD	$84.22_{\pm.17}$	$82.85_{\pm.21}$	$76.93_{\pm.32}$	$50.04_{\pm.16}$	$48.66_{\pm.28}$	$44.07_{\pm.30}$
SCAFFOLD	P	SGD	$84.31_{\pm.14}$	$83.70_{\pm.11}$	$78.70_{\pm.21}$	$50.69_{\pm.21}$	$50.28_{\pm.21}$	$47.12_{\pm.34}$
FedSAM	P	SAM	$83.79_{\pm.28}$	$82.58_{\pm.19}$	$77.83_{\pm.27}$	$48.66_{\pm.29}$	$48.42_{\pm.19}$	$45.03_{\pm.22}$
FedDyn	PD	SGD	$83.71_{\pm.26}$	$82.66_{\pm.15}$	$79.44_{\pm.25}$	–	–	–
FedSpeed	PD	SAM	$86.90_{\pm.18}$	$85.92_{\pm.24}$	$81.47_{\pm.19}$	$53.22_{\pm.28}$	$52.75_{\pm.16}$	$49.66_{\pm.13}$
A-FedPD	PD	SGD	$85.11_{\pm.12}$	$84.33_{\pm.16}$	$81.05_{\pm.28}$	$48.15_{\pm.22}$	$48.02_{\pm.29}$	$46.24_{\pm.26}$
A-FedPDSAM	PD	SAM	$\mathbf{87.44}_{\pm.13}$	$\mathbf{86.46}_{\pm.25}$	$\mathbf{82.48}_{\pm.21}$	$\mathbf{55.30}_{\pm.23}$	$\mathbf{53.49}_{\pm.17}$	$\mathbf{50.31}_{\pm.23}$

Ablation Study



(a) Different Participation Ratios. (b) Different Local Intervals. (c) Different Rounds.

Figure 2: Test of the proposed *A-FedPD* method on setups of different participation ratios, different local intervals, and different rounds. In these experiments, we fix the total training data samples and total training iterations and then learn their variation trends.

THE UNIVERSITY OF
SYDNEY

Thank you!