Understanding How Consistency Works in Federated Learning via Stage-wise Relaxed Initialization

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

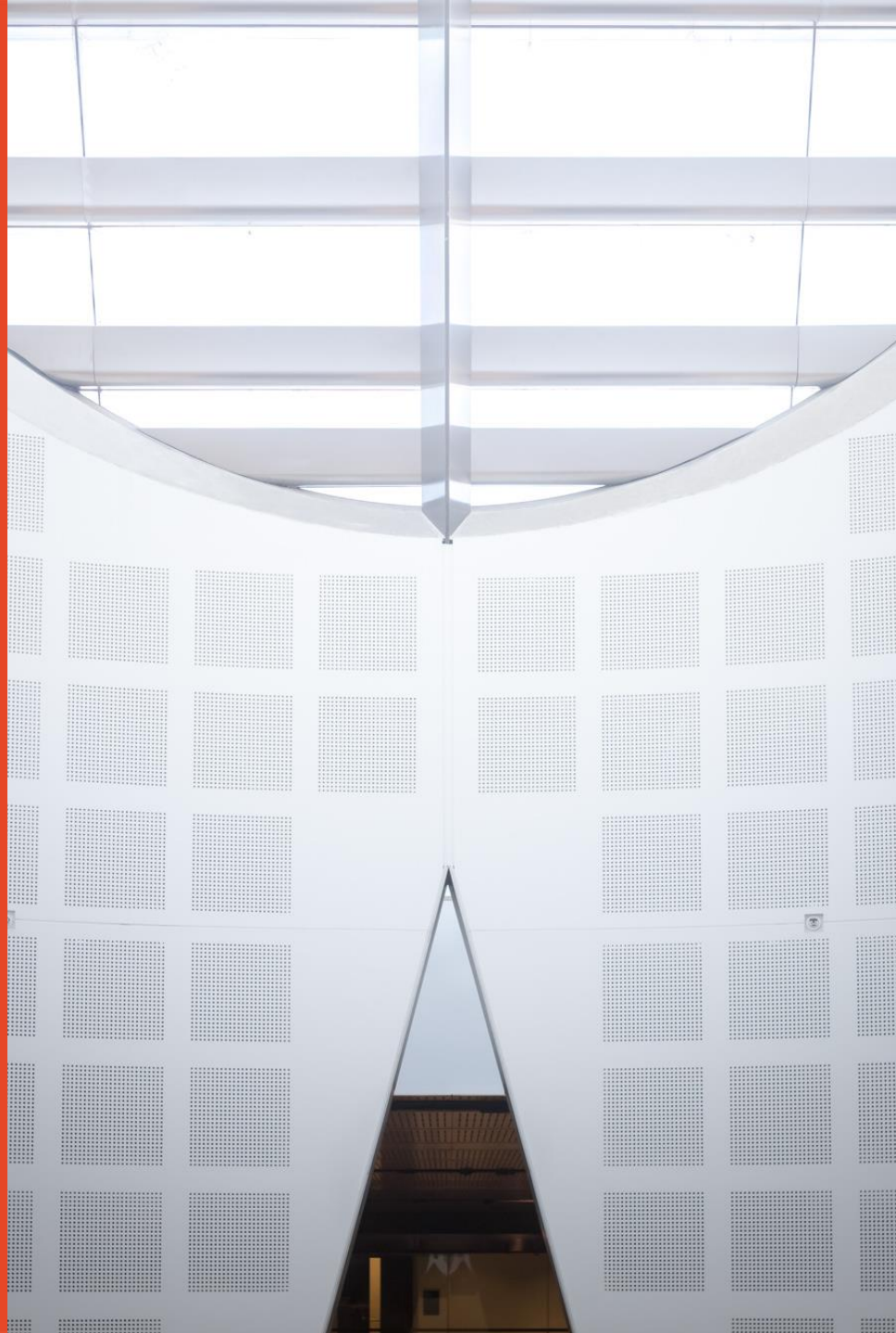
[1] the University of Sydney

[2] JD Explore Academy

* Corresponding Author

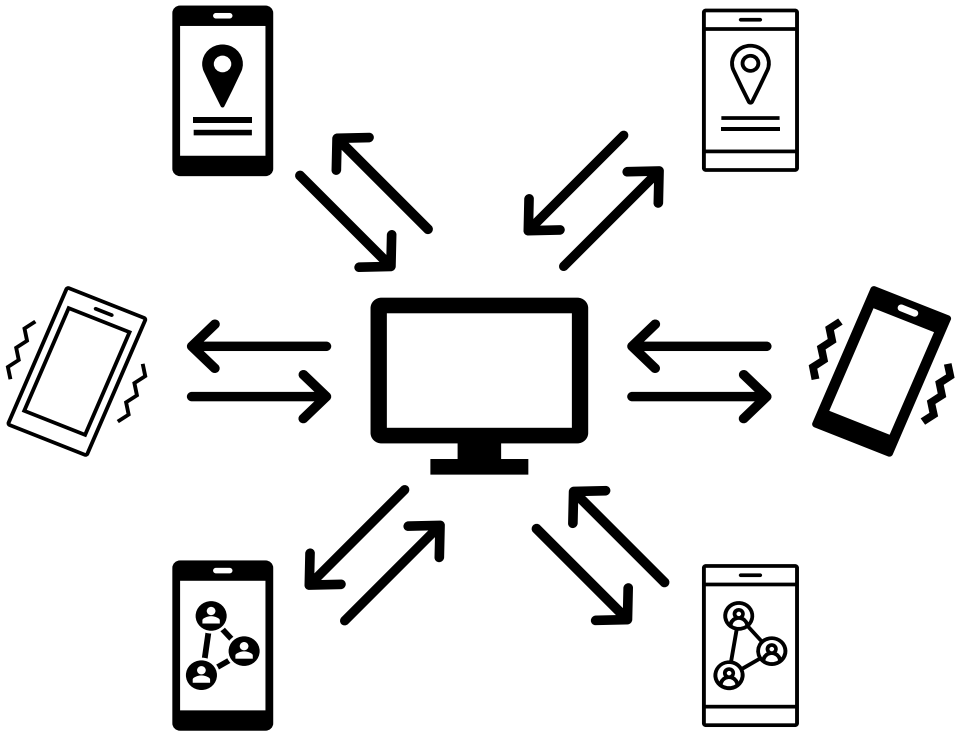
THE UNIVERSITY OF SYDNEY

NEURAL INFORMATION PROCESSING SYSTEMS



Introduction: Federated Learning

Federated learning (FL) is a distributed paradigm which coordinates massive local clients to collaboratively train global model via stage-wise local training processes on the heterogeneous dataset by a global central server.



Main Challenges in Federated Learning

- Low Computing Power on Edge-devices [Low power devices]
- Dataset Privacy [No direct accesses across devices]
- Client Drifts [Heterogeneous local optimal]
- Communication Bottleneck [Limited Network Bandwidth]

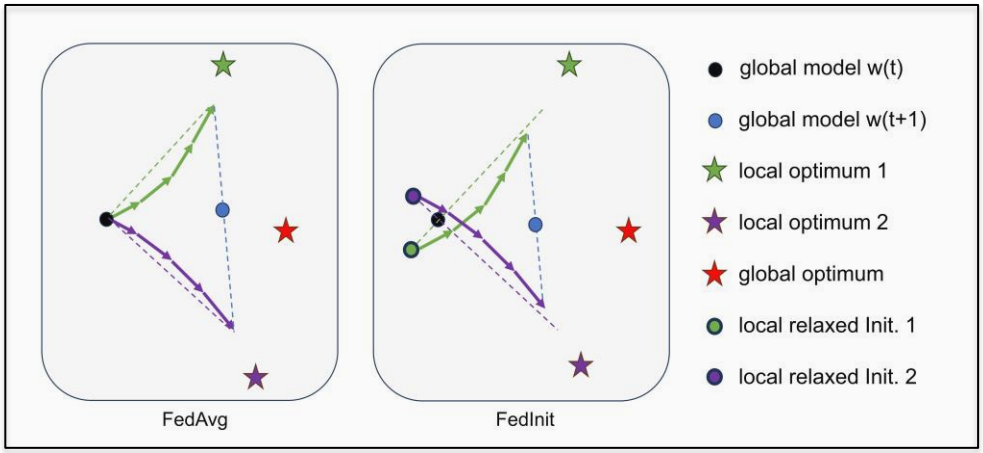
Our Contributions:

- To alleviate the client drifts, we propose the FedInit method to improve the local consistency, which employs personalized relaxed initialization (RI) on selected local clients at each communication round.

- To understand its benefits, we study the *Excess Risk* to jointly analyze the error of optimization and generalization. Different from the previous work,
  (1) *Excess risk* could be considered as the test error performance. Our analysis directly reflects why RI works in the FL paradigm.
  (2) We expand the analysis of the stability in FL.
  (3) We study the impacts of the inconsistency to FL.

- We conduct extensive experiments on several general model backbones and the different FL setups to validate the efficiency of RI.

FedInit Methodology



Algorithm 1: *FedInit* Algorithm

Input: model w, local model w_i, T, K, β.
Output: model w^T.

Initialize states: initialize $w^{-1} = w_{i,0}^{-1} = w^0$.
for $t = 0, 1, ..., T-1$ **do**
  randomly select active clients set $\mathcal{N}$ from $\mathcal{C}$
  for $i \in \mathcal{N}$ ***in parallel*** **do**
    send the w^t to the active clients
    set the $w^t + \beta(w^t - w_{i,K}^{t-1})$ as $w_{i,0}^t$
    for $k = 0, 1, ..., K-1$ **do**
      compute gradient $g_{i,k}^t$ at $w_{i,k}^t$
      $w_{i,k+1}^t = w_{i,k}^t - \eta g_{i,k}^t$
    end
    send the $w_i^t = w_{i,K}^t$ to the server
  end
  $w^{t+1} = \frac{1}{N} \sum_{i \in \mathcal{N}} w_i^t$
end

FedInit begins the local training from a new personalized state, which moves away from the last global model towards the reverse direction from the latest local state:

$$w_{i,0}^t = w^t + \beta(w^t - w_{i,K}^{t-1}).$$

Though it changes the consistency around selected clients,

(1) Its global expectation is still unbiased, which is $\frac{1}{N}\sum_{i \in [N]} w_{i,0}^t = w^t$.

(2) It enlarge the distance to achieve local optimal.

Benefits from the Relaxed Initialization

- Relaxed initialization helps to avoid the local overfitting
- Relaxed initialization helps to enhance the local consistency
- Relaxed initialization could work as a plug-in to several advanced methods to further improve their efficiency without conflicts
- Relaxed initialization does not require additional costs of storage and communication across local clients

Different from "Lookahead"

The personalized relaxed initialization (RI) is partially motivated by the "lookahead" optimization which has shown significant improvements in the stochastic case, i.e., for one client,

Algorithm 1 Lookahead Optimizer:

Require: Initial parameters ϕ_0, objective function L
Require: Synchronization period k, slow weights step size α, optimizer A
for $t = 1, 2, \ldots$ **do**
  Synchronize parameters $\theta_{t,0} \leftarrow \phi_{t-1}$
  for $i = 1, 2, \ldots, k$ **do**
    sample minibatch of data $d \sim \mathcal{D}$
    $\theta_{t,i} \leftarrow \theta_{t,i-1} + A(L, \theta_{t,i-1}, d)$
  end for
  Perform outer update $\phi_t \leftarrow \phi_{t-1} + \alpha(\theta_{t,k} - \phi_{t-1})$
end for
return parameters ϕ

(1) k-step updates and 1-step lookahead
(2) interpolation of current and previous states

$$w_t = w_{t-1} + \alpha(\theta_t - w_{t-1})$$
$$= \alpha\theta_t + (1-\alpha)w_{t-1}$$

However, RI performs as $w_{i,0}^t = w^t + \beta\left(w^t - w_{i,K}^{t-1}\right) = (1+\beta)w^t - \beta w_{i,K}^t$, which is the extrapolation of the global and local states.

Furthermore, "Lookahead" acts on the end of each training stage, while RI acts on the beginning of each training stage.

Theoretical Analysis: Assumptions

Assumption 1. (for opt) $f_i(w)$ is L-smooth, $\|\nabla f_i(w_1) - \nabla f_i(w_2)\| \leq L\|w_1 - w_2\|$.

Assumption 2. (for opt) bounded gradients, $E[g_i] = \nabla f_i(w),\ E\|g_i - \nabla f_i(w)\|^2 \leq \sigma_l^2$.

Assumption 3. (for opt) bounded heterogeneity, $E\|\nabla f_i(w)\|^2 \leq G^2 + B^2 E\|\nabla f(w)\|^2$.

Assumption 4. (for gen) $f(w)$ is L_G-Lipschitz, $\|f(w_1) - f(w_2)\| \leq L_G\|w_1 - w_2\|$.

Assumption 5. (for gen) PL-condition, $2\mu\big(f(w) - f(w^\star)\big) \leq \|\nabla f(w)\|^2$.

Our work focuses on understanding how the generalization performance changes in the training process. We consider the entire training process and adopt uniform stability to measure the global generality in FL. Assumption 1-3 are the general assumptions in convergence analysis. Assumption 4-5 are adopted to analyze the uniform stability and generalization.

Theoretical Analysis: Excess Risk and Test Error

	Opt. Term	Convergence Bound	Rate
A.1 2 3	$\frac{1}{T}\sum_{t=1}^{T-1} E\|\nabla f(w^t)\|^2$	$\frac{2D}{\lambda\eta KT} + \frac{\kappa_2 \eta L \sigma_l^2}{\lambda N} + \frac{3\kappa_1 \eta KLG^2}{\lambda N}$	$O\left(\frac{1}{\sqrt{NKT}}\right)$
A.1 2 3 5	$E[f(w^T) - f(w^\star)]$	$e^{-\lambda\mu\eta KT} E[f(w^0) - f(w^\star)] + \eta \frac{3\kappa_1 KLG^2}{2\mu\lambda N} + \eta \frac{\kappa_2 L \sigma_l^2}{2\mu\lambda N}$	$O\left(\frac{1}{NKT}\right)$
	Consistency	**Upper Bound**	**Rate**
A.1 2 3	$\frac{1}{CT}\sum_{i,t} E\left\|w_{i,K}^t - w^t\right\|^2$	–	$O\left(\frac{N}{T}\right)$
A.1 2 3 5	$\frac{1}{C}\sum_{i} E\left\|w_{i,K}^T - w^T\right\|^2$	–	$O\left(\frac{1}{T^2}\right)$
	Gen. Term	**Stability Bound**	**Rate**
A.1 2 4 5	$E[F(w^T) - f(w^T)]$	$\frac{1}{S}\left(\frac{(UTK)^{cL}}{L}\right)^{\frac{1}{1+cL}} + (1+\beta)^{\frac{1}{\beta cL}} \left(\frac{ULTK}{2(L_G^2 + SL_G\sigma_l)}\right)^{\frac{cL}{1+cL}} \frac{\sqrt{C_\Delta}}{T}$	$O\left(\frac{1}{S} + \left(\frac{K}{T}\right)^{\frac{1}{1+cL}}\right)$

Experiments

Table 1: Test accuracy (%) on the CIFAR-10/100 dataset. We test two participation ratios on each dataset. Under each setup, we test two Dirichlet splittings, and each result test for 3 times. This table reports results on ResNet-18-GN (upper part) and VGG-11 (lower part) respectively.

Method	CIFAR-10				CIFAR-100			
	10%-100 clients		5%-200 clients		10%-100 clients		5%-200 clients	
	Dir-0.6	Dir-0.1	Dir-0.6	Dir-0.1	Dir-0.6	Dir-0.1	Dir-0.6	Dir-0.1
FedAvg	$78.77_{\pm.11}$	$72.53_{\pm.17}$	$74.81_{\pm.18}$	$70.65_{\pm.21}$	$46.35_{\pm.15}$	$42.62_{\pm.22}$	$44.70_{\pm.22}$	$40.41_{\pm.33}$
FedAdam	$76.52_{\pm.14}$	$70.44_{\pm.22}$	$73.28_{\pm.18}$	$68.87_{\pm.26}$	$48.35_{\pm.17}$	$40.77_{\pm.31}$	$44.33_{\pm.26}$	$38.04_{\pm.25}$
FedSAM	$79.23_{\pm.22}$	$72.89_{\pm.23}$	$75.45_{\pm.19}$	$71.23_{\pm.26}$	$47.51_{\pm.26}$	$43.43_{\pm.12}$	$45.98_{\pm.27}$	$40.22_{\pm.27}$
SCAFFOLD	$81.37_{\pm.17}$	$75.06_{\pm.16}$	$78.17_{\pm.28}$	$74.24_{\pm.22}$	$51.98_{\pm.23}$	**44.41**$_{\pm.15}$	$50.70_{\pm.29}$	$41.83_{\pm.29}$
FedDyn	$82.43_{\pm.16}$	$75.08_{\pm.19}$	$79.96_{\pm.13}$	$74.15_{\pm.34}$	$50.82_{\pm.19}$	$42.50_{\pm.28}$	$47.32_{\pm.21}$	$41.74_{\pm.21}$
FedCM	$81.67_{\pm.17}$	$73.93_{\pm.26}$	$79.49_{\pm.17}$	$73.12_{\pm.18}$	$51.56_{\pm.20}$	$43.03_{\pm.26}$	$50.93_{\pm.19}$	$42.33_{\pm.19}$
FedInit	**83.11**$_{\pm.29}$	**75.95**$_{\pm.19}$	**80.58**$_{\pm.20}$	**74.92**$_{\pm.17}$	**52.21**$_{\pm.09}$	$44.22_{\pm.21}$	**51.16**$_{\pm.18}$	**43.77**$_{\pm.36}$
FedAvg	$85.28_{\pm.12}$	$78.02_{\pm.22}$	$81.23_{\pm.14}$	$74.89_{\pm.25}$	$53.46_{\pm.25}$	$50.53_{\pm.20}$	$47.55_{\pm.13}$	$45.05_{\pm.33}$
FedAdam	$86.44_{\pm.13}$	$77.55_{\pm.28}$	$81.05_{\pm.23}$	$74.04_{\pm.17}$	$55.56_{\pm.29}$	$53.41_{\pm.18}$	$51.33_{\pm.25}$	$47.26_{\pm.21}$
FedSAM	$86.37_{\pm.22}$	$79.10_{\pm.07}$	$81.76_{\pm.26}$	$75.22_{\pm.13}$	$54.85_{\pm.31}$	$51.88_{\pm.27}$	$48.65_{\pm.21}$	$46.58_{\pm.28}$
SCAFFOLD	$87.73_{\pm.17}$	$81.98_{\pm.19}$	$84.81_{\pm.15}$	$79.04_{\pm.16}$	**59.45**$_{\pm.17}$	$56.67_{\pm.24}$	$53.73_{\pm.32}$	$50.08_{\pm.19}$
FedDyn	$87.35_{\pm.19}$	$82.70_{\pm.24}$	$84.84_{\pm.19}$	**80.01**$_{\pm.22}$	$56.13_{\pm.18}$	$53.97_{\pm.11}$	$51.74_{\pm.18}$	$48.16_{\pm.17}$
FedCM	$86.80_{\pm.33}$	$79.85_{\pm.29}$	$83.23_{\pm.31}$	$76.42_{\pm.36}$	$53.88_{\pm.22}$	$50.73_{\pm.35}$	$47.83_{\pm.19}$	$46.33_{\pm.25}$
FedInit	**88.47**$_{\pm.22}$	**83.51**$_{\pm.13}$	**85.36**$_{\pm.19}$	$79.73_{\pm.14}$	$58.84_{\pm.11}$	**57.22**$_{\pm.21}$	**54.12**$_{\pm.08}$	**50.27**$_{\pm.29}$

Experiments

Table 2: We incorporate the relaxed initialization (RI) into the benchmarks to test improvements on ResNet-18-GN on CIFAR-10 with the same hyperparameters and specific relaxed coefficient β.

Method	10%-100 clients				5%-200 clients			
	Dir-0.6		Dir-0.1		Dir-0.6		Dir-0.1	
	-	+RI	-	+RI	-	+RI	-	+RI
FedAvg	78.77	83.11	72.53	75.95	74.81	80.58	70.65	74.92
FedAdam	76.52	78.33	70.44	72.55	73.28	78.33	68.87	71.34
FedSAM	79.23	**83.36**	72.89	76.34	75.45	80.66	71.23	75.08
SCAFFOLD	81.37	83.27	75.06	**77.30**	78.17	**81.02**	74.24	**76.22**
FedDyn	82.43	81.91	75.08	75.11	79.96	79.88	74.15	74.34
FedCM	81.67	81.77	73.93	73.71	79.49	79.72	73.12	72.98

Method	communication	ratio	gradient calculation	ratio	total storage	ratio
FedAvg	Nd	$1\times$	NKd	$1\times$	Cd	$1\times$
FedAdam	Nd	$1\times$	NKd	$1\times$	Cd	$1\times$
FedSAM	Nd	$1\times$	$2NKd$	$2\times$	$2Cd$	$2\times$
SCAFFOLD	$2Nd$	$2\times$	NKd	$1\times$	$3Cd$	$3\times$
FedDyn	Nd	$1\times$	NKd	$1\times$	$3Cd$	$3\times$
FedCM	$2Nd$	$2\times$	NKd	$1\times$	$2Cd$	$3\times$
FedInit	Nd	$1\times$	NKd	$1\times$	Cd	$1\times$

Acknowledgements

This study is partially supported by Australian Research Council Project FL-170100117.

Thanks for your attentions.