Where to begin?

On the Impact of Pre-training and Initialization in Federated Learning

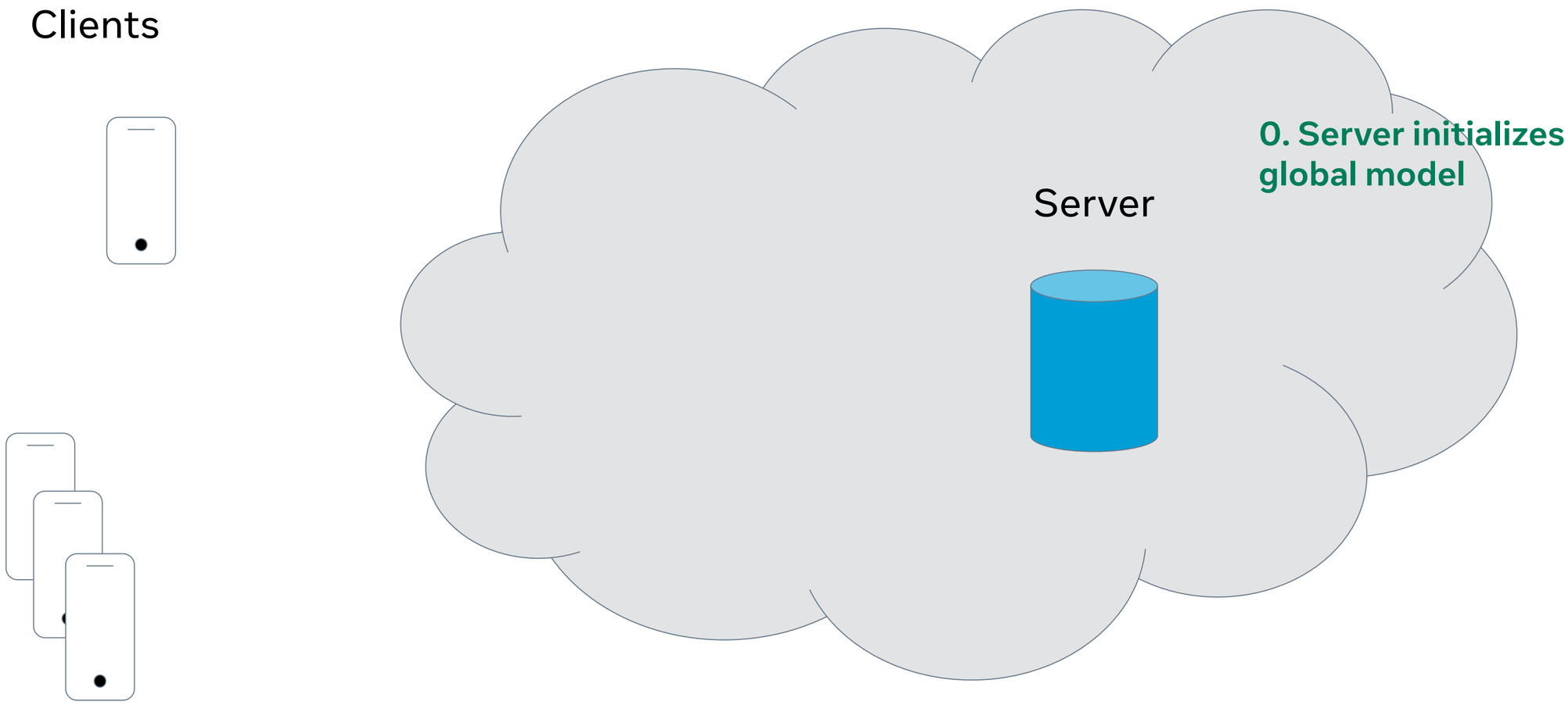
Joint work with: John Nguyen, Jianyu Wang, Kshitiz Malik, & Maziar Sanjabi

Mike Rabbat

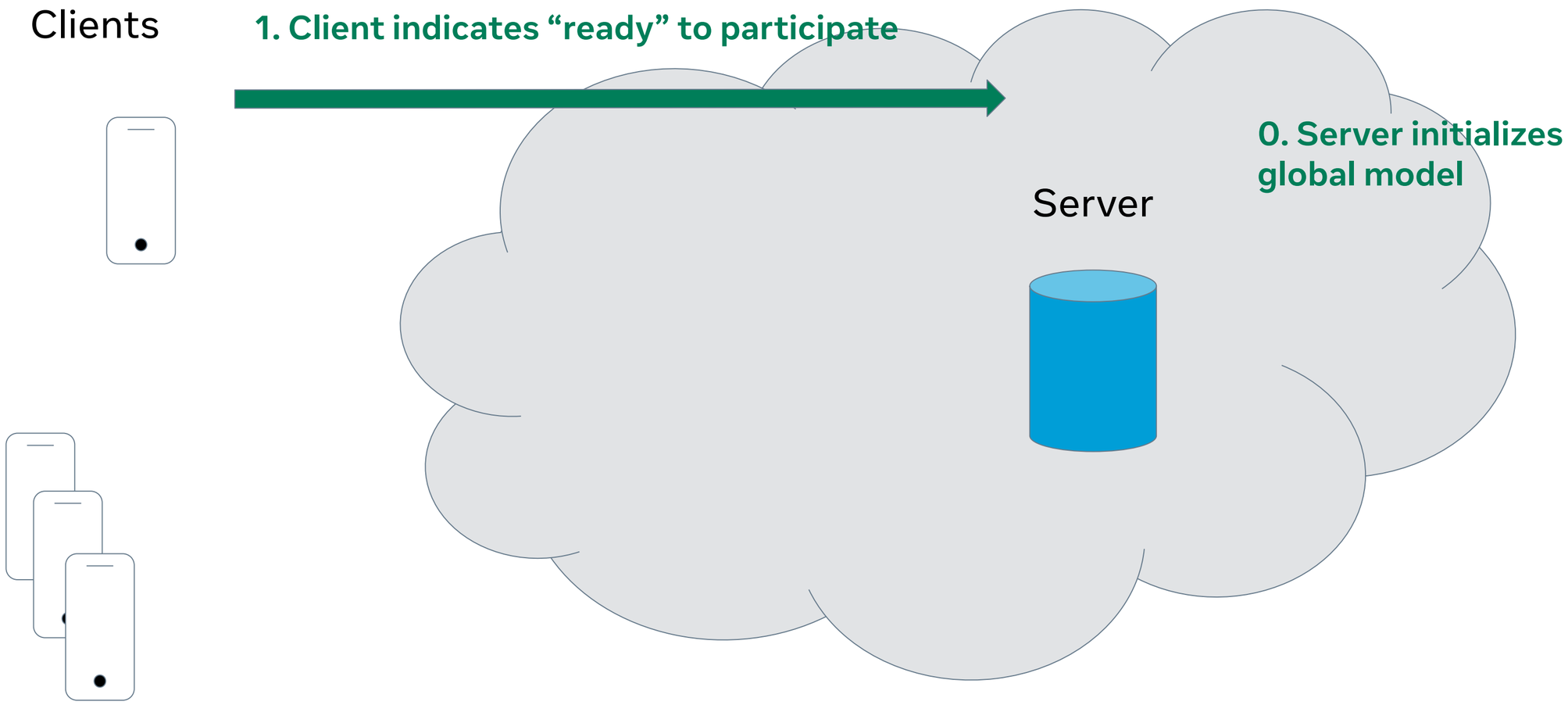
FedVision Workshop @CVPR 2023

Meta AI

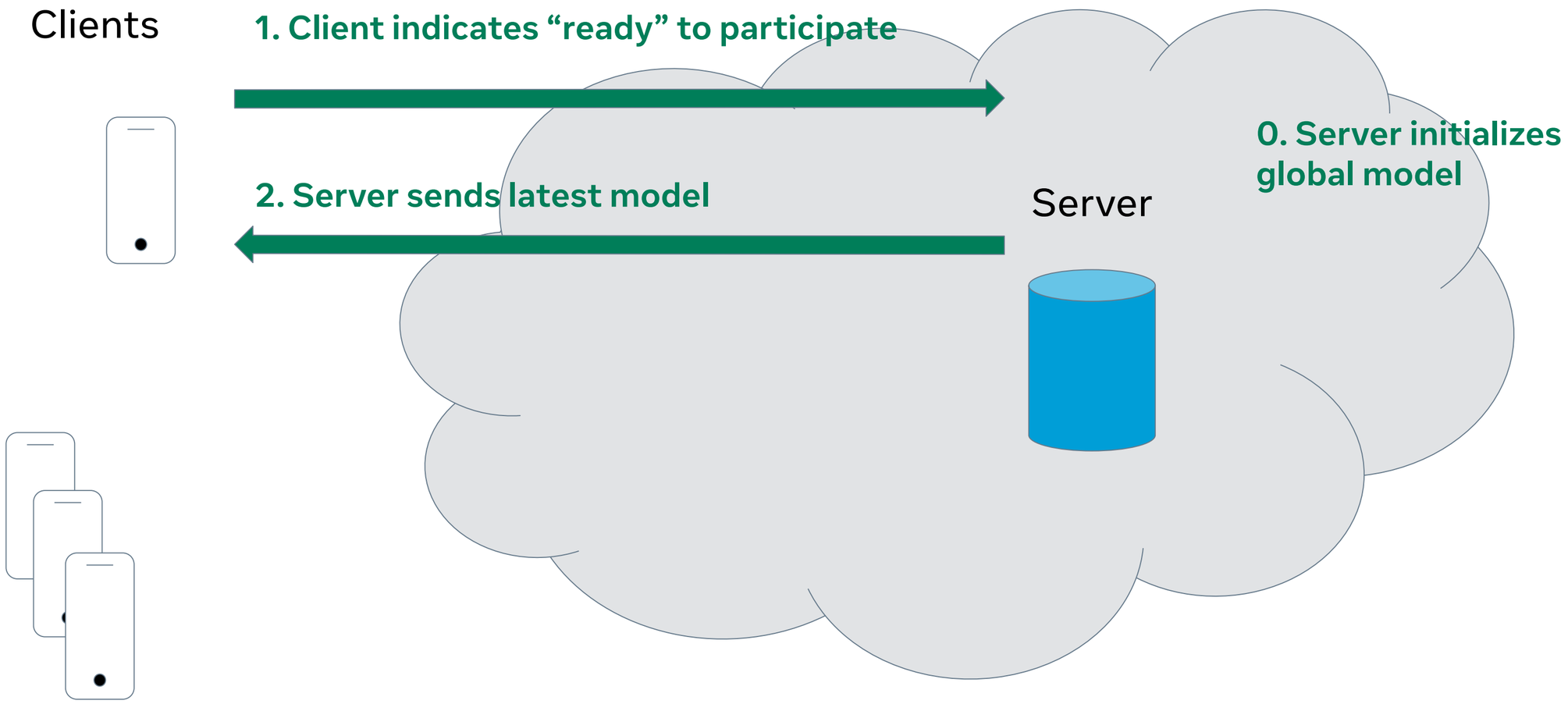
FedOpt Framework (Reddi et al., 2021)



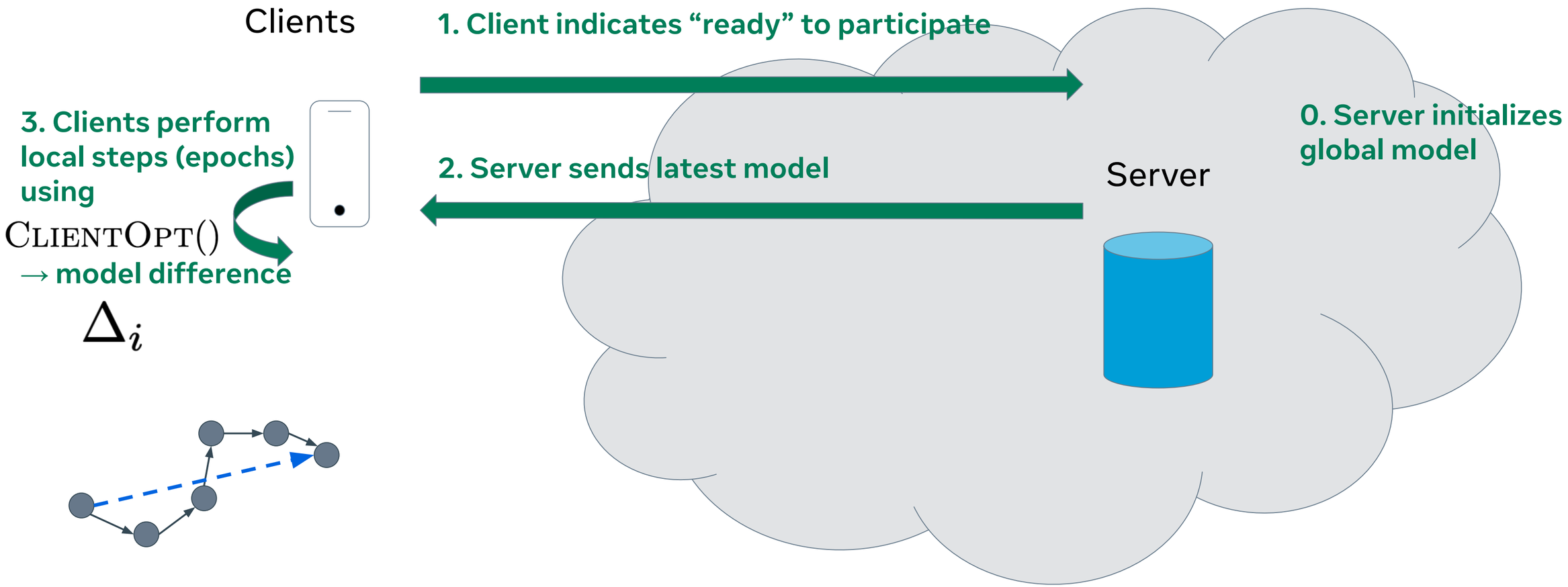
FedOpt Framework (Reddi et al., 2021)



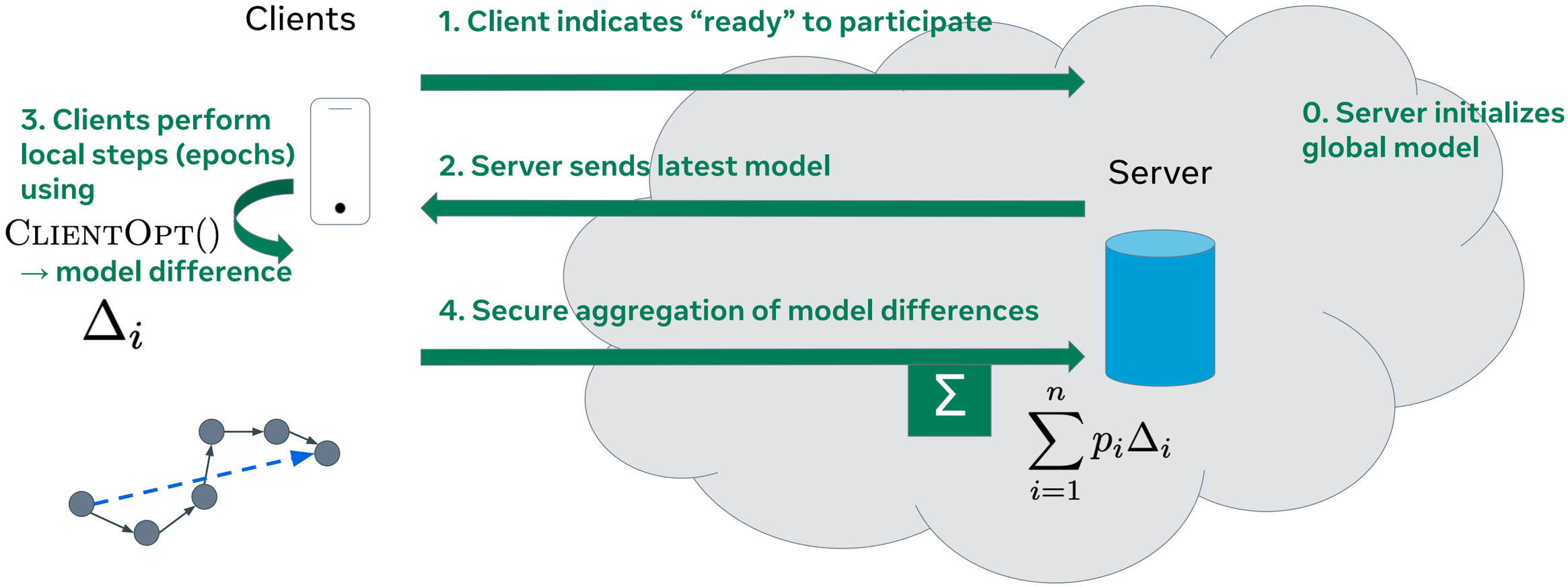
FedOpt Framework (Reddi et al., 2021)



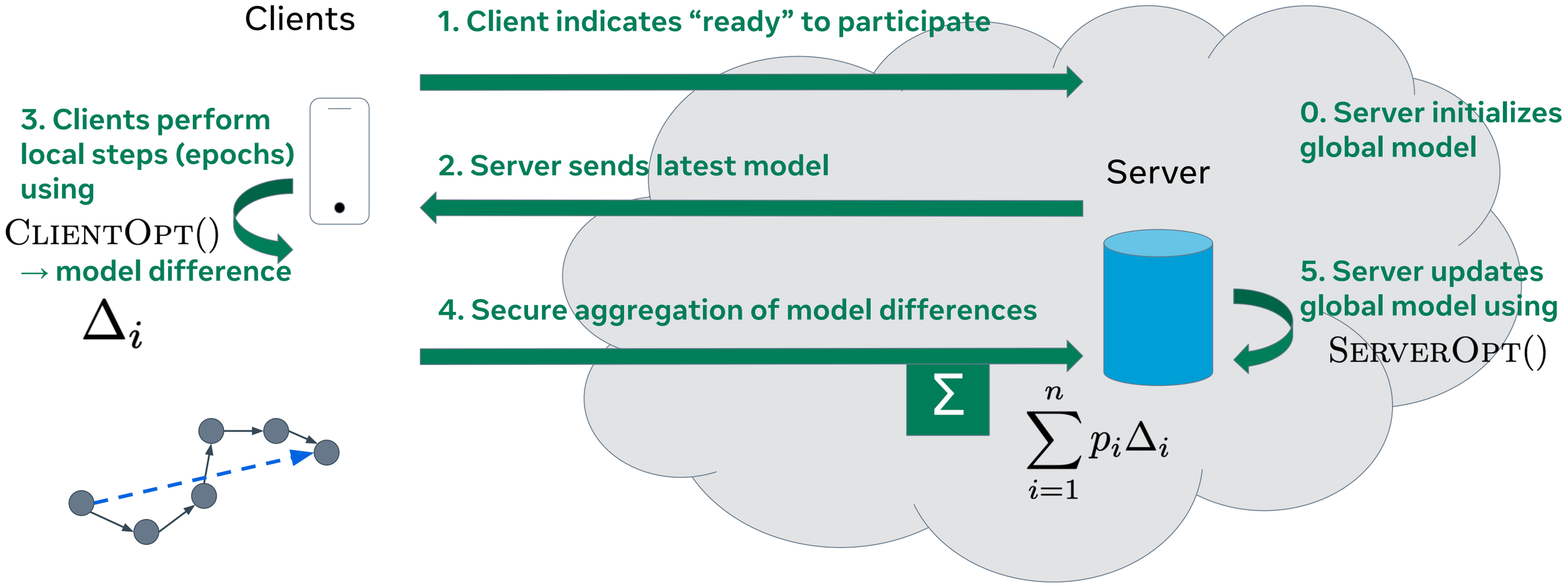
FedOpt Framework (Reddi et al., 2021)



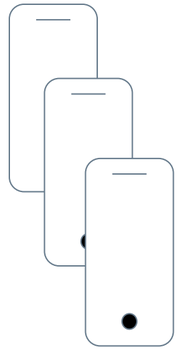
FedOpt Framework (Reddi et al., 2021)



FedOpt Framework (Reddi et al., 2021)



FL Challenges due to Heterogeneity



Data heterogeneity

- Different client devices hold data from different distributions
- Local update steps may cause model drift

System heterogeneity & data imbalance

- Different client devices have different capabilities, different amounts of data
- Can lead to “objective inconsistency” problem (Wang et al., 2020)

Previous Work Addressing Heterogeneity

FedProx (Li et al., 2018)

- Penalize local updates for drifting too far from last server model

FedAvgM (Hsu et al., 2019) and **FedAdam** (Reddi et al., 2020)

- Incorporate momentum/adaptivity at the server

Scaffold (Karimireddy et al., 2020) and **Mime** (Karimireddy et al., 2021)

- Control variates and server-provided momentum buffer to implicitly discourage drift

FedNova (Wang et al., 2020)

- Scale local step-size/update to address objective inconsistency

FedPD (Zhang et al., 2020) and **FedDyn** (Acar et al., 2021)

- Adjust local training objective to address objective inconsistency

All modify training algorithm and dynamics

How does initialization impact federated training?

Proxy data may be available in many practical applications of FL
→ Initialize from a pre-trained model rather than from random

Converge to better quality solution (starting from a better init?)

Converge faster (starting closer?)

Diminish the effects of heterogeneity?

Related work

Transfer learning widely used in centralized machine/deep learning

Evidence that fine-tuning problems may be better behaved/conditioned than pre-training

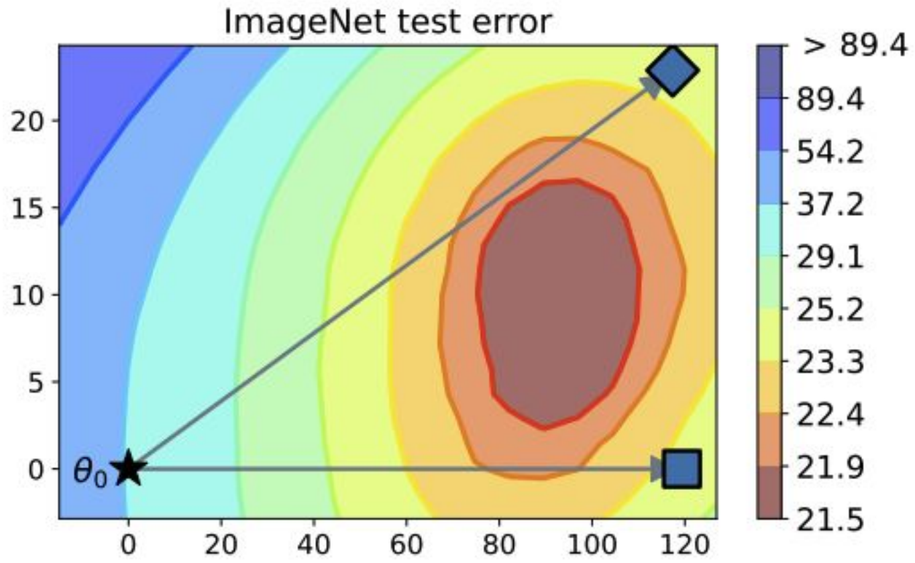


Fig source: Wortsman et al., (2022)

B Neyshabur, H Sedghi, & C Zhang, “What is being transferred in transfer learning?” NeurIPS 2020

M Wortsman et al., “Model soups: Averaging weights of multiple fine-tuned models improves accuracy without increasing inference time,” ICML 2022

Related work

Pre-training in FL closes the accuracy gap to centralized/IID training

HY Chen, CH Tu, Z Li, HW Shen, & WL Chao, “On pre-training for federated learning” arxiv:2206.11488 (2022)

O Weller, M Marone, V Braverman, D Lawrie, & B Van Durme, “Pre-trained models for multilingual federated learning,” arxiv:2206.02291 (2022)

Y Tan, G Long, J Ma, L Liu, T Zhou, & J Jiang, “FL from Pre-trained models: A contrastive learning approach,” NeurIPS 2022

Experiment Setup

Federated Dataset	Model	Pre-training/Initialization
CIFAR-10 (Dirichlet, α=0.1)	SqueezeNet ResNet-18 (with GroupNorm)	ImageNet (TorchVision) ImageNet (TorchVision)
FEMNIST	SqueezeNet ResNet-18 (with GroupNorm)	ImageNet (TorchVision) ImageNet (TorchVision)
Stack Overflow	DistilGPT-2 CharLM	Openwebtext (HuggingFace) WikiText-103
pushshift.io Reddit	CharLM	WikiText-103

Standard, supervised pre-training.
Randomly initialize last layer when transferring to federated training.

Federated Optimization Methods

ClientOpt variations

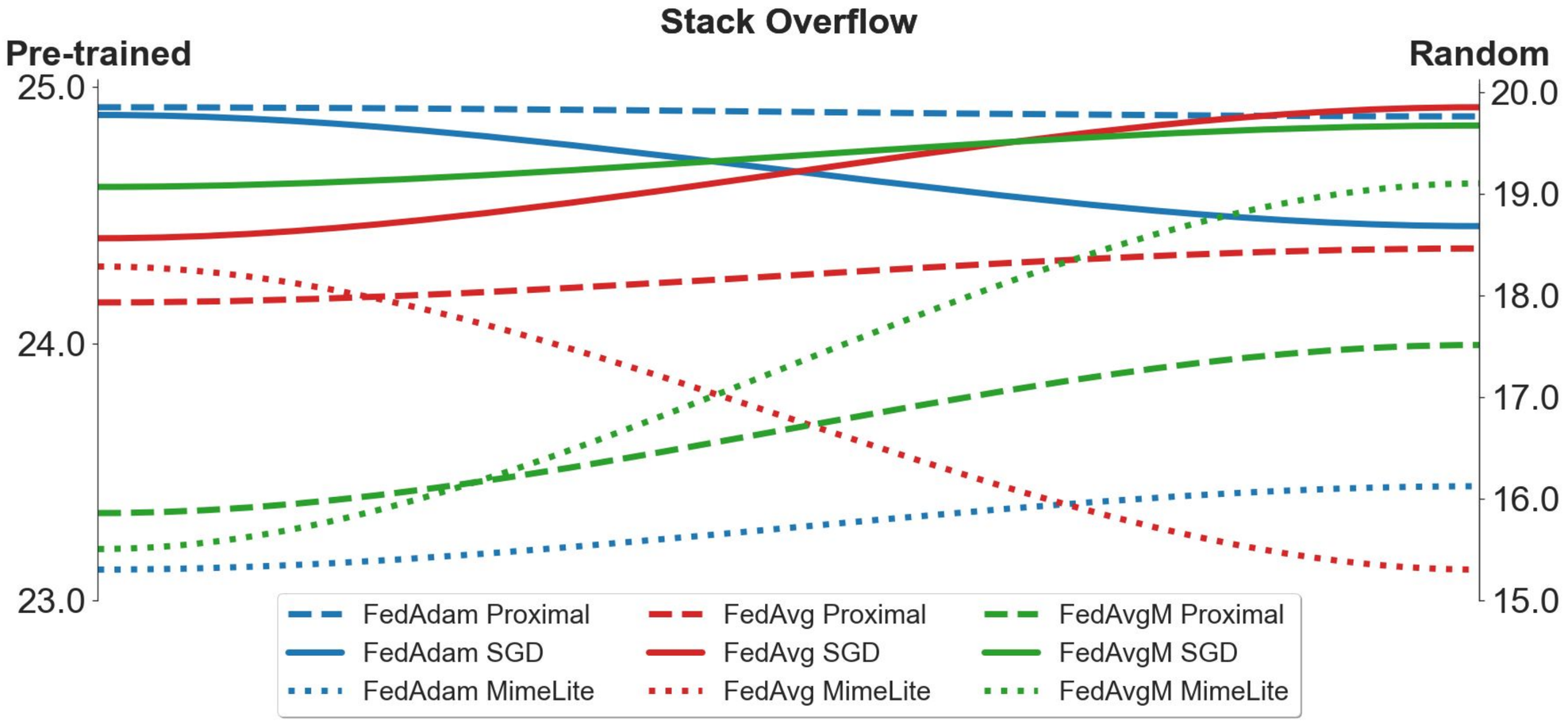
1. SGD
2. Proximal (FedProx; Li et al., 2018)
3. Normalized Averaging (FedNova; Wang et al., 2020)
4. MimeLite (Karimireddy et al., 2021)
5. Full-batch Gradient Descent

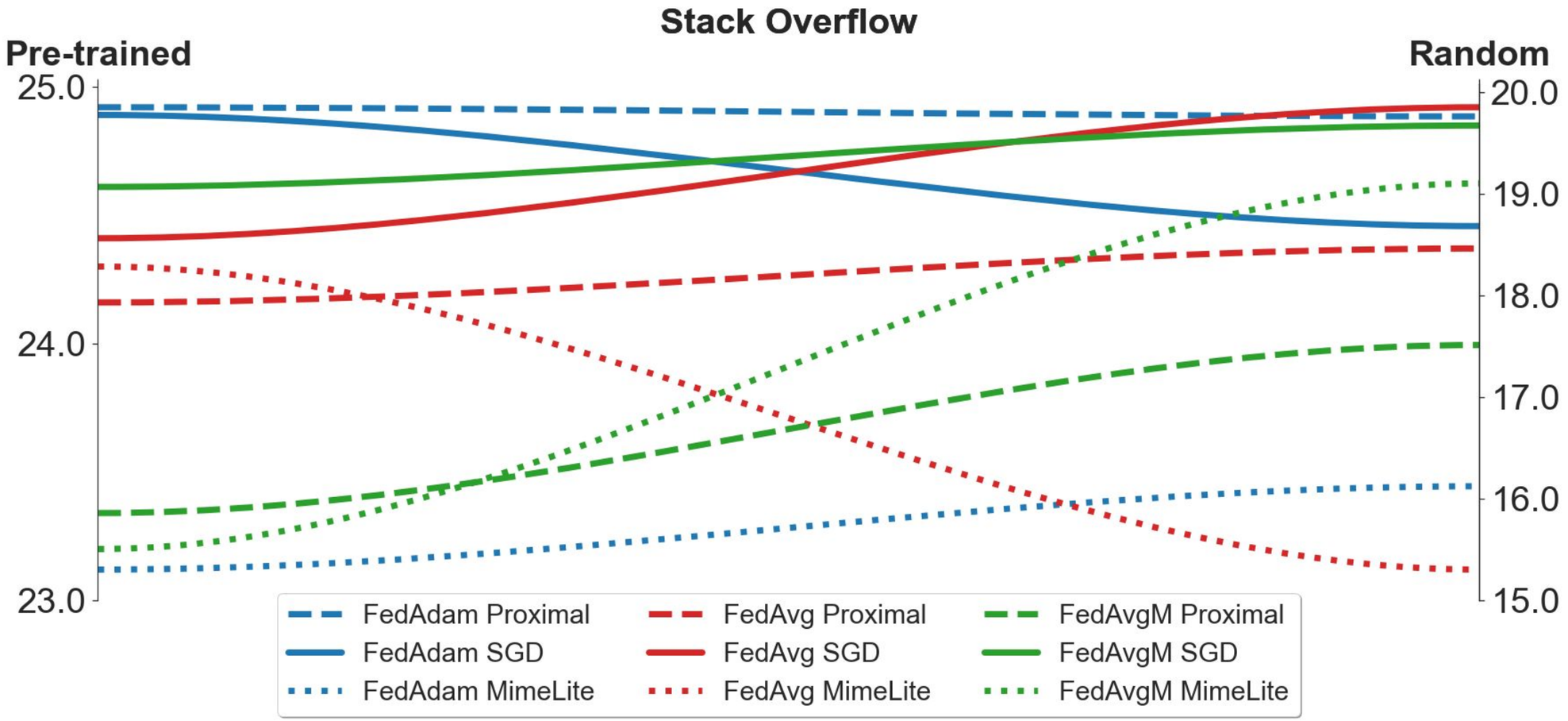
ServerOpt variations

1. SGD
2. SGD with momentum
3. Adam

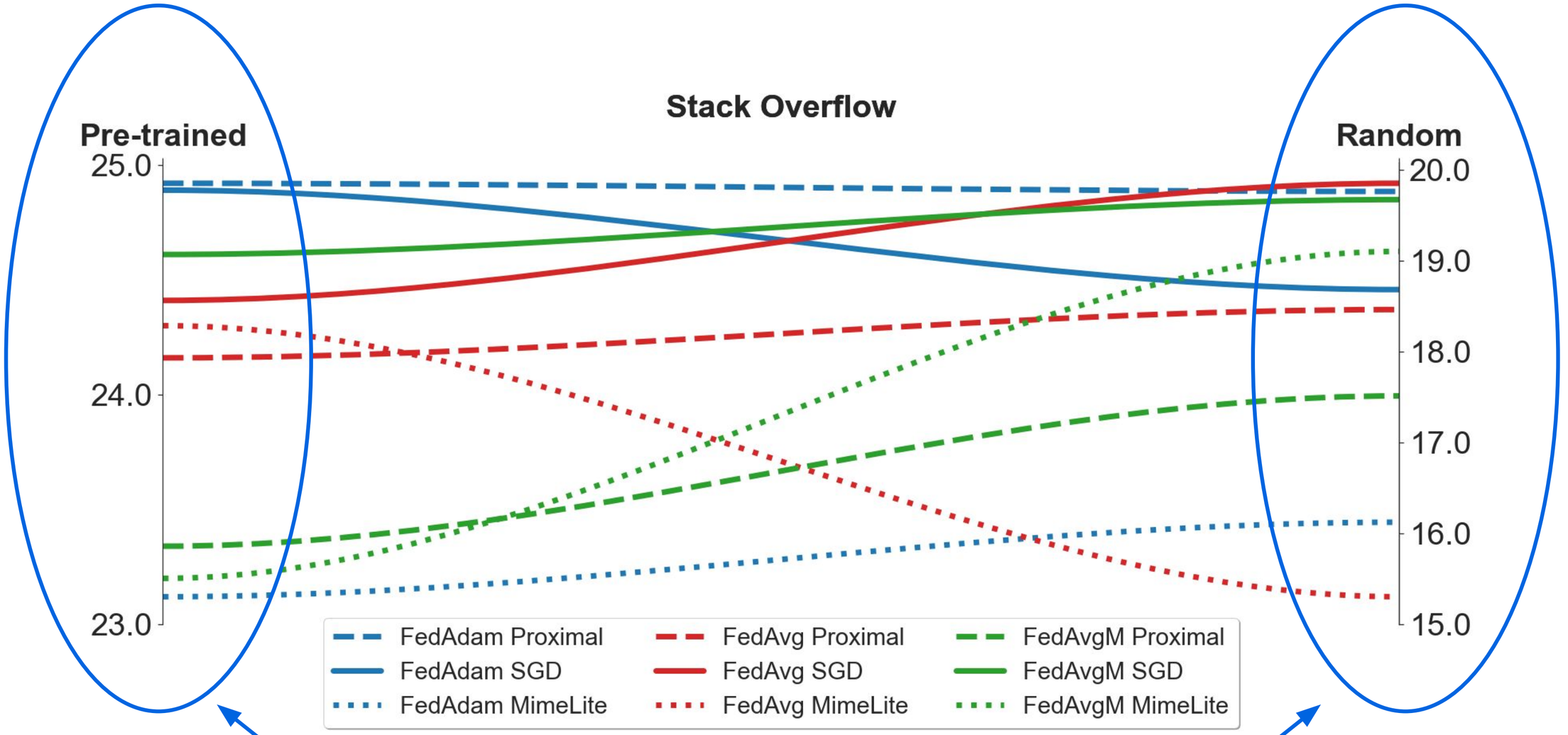
Same number of rounds, local batch size, and cohort size for all methods.
Tune other hyperparameters for each method and each workload.

Does initialization impact all FL algorithms equally?

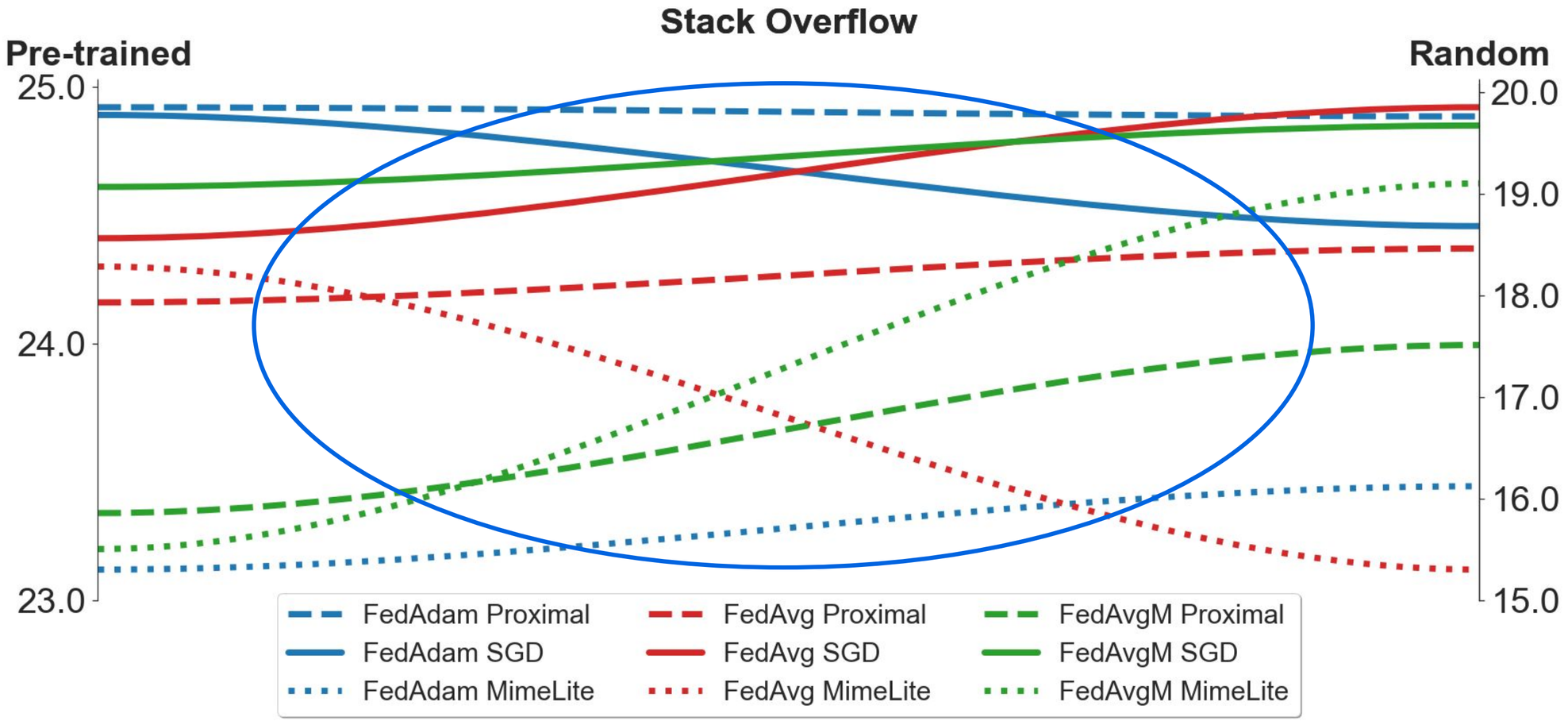

Stack Overflow
Pre-trained
Random
25.0
24.0
23.0
20.0
19.0
18.0
17.0
16.0
15.0
FedAdam Proximal
FedAdam SGD
FedAdam MimeLite
FedAvg Proximal
FedAvg SGD
FedAvg MimeLite
FedAvgM Proximal
FedAvgM SGD
FedAvgM MimeLite



9 algorithms (ServerOpt - ClientOpt combos)



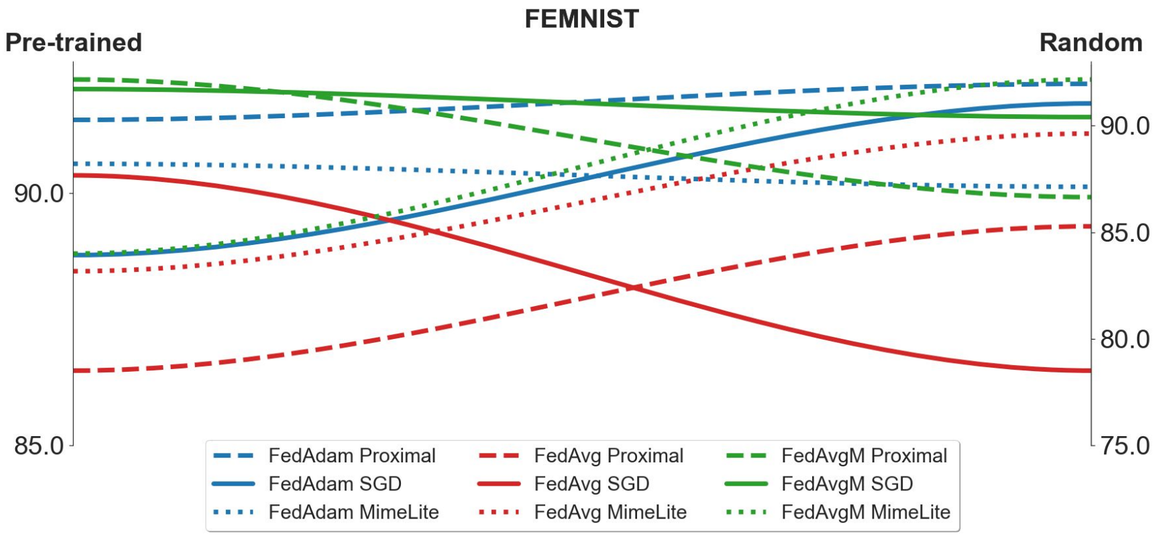
18 final accuracies (9 algs x 2 initializations)
(different ranges on each axis)

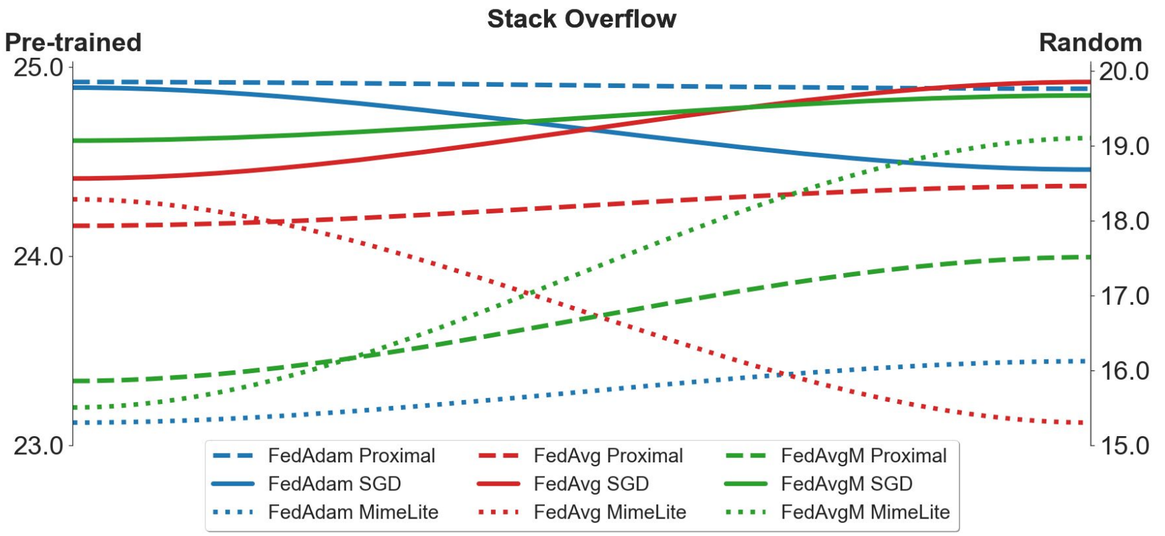


Interpolating lines cross

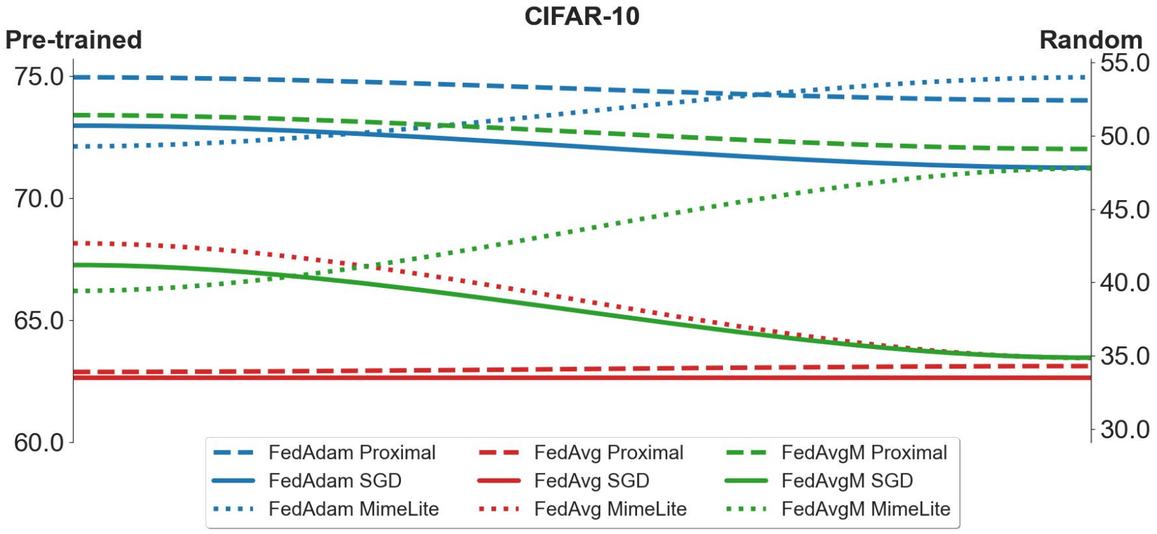
⇒

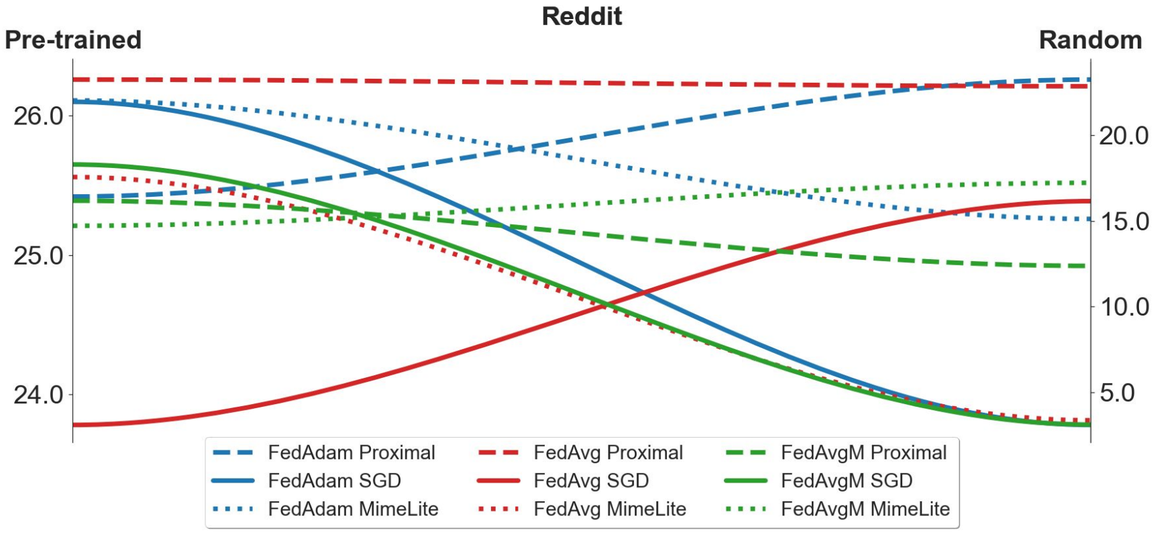
Initialization changes ranking of algorithms



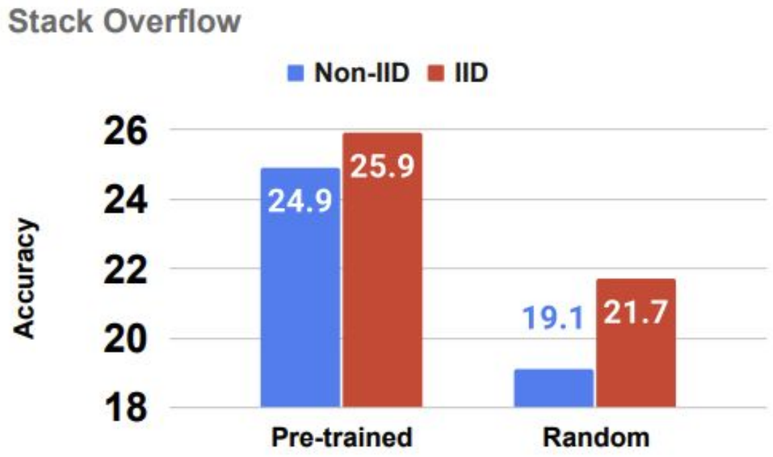


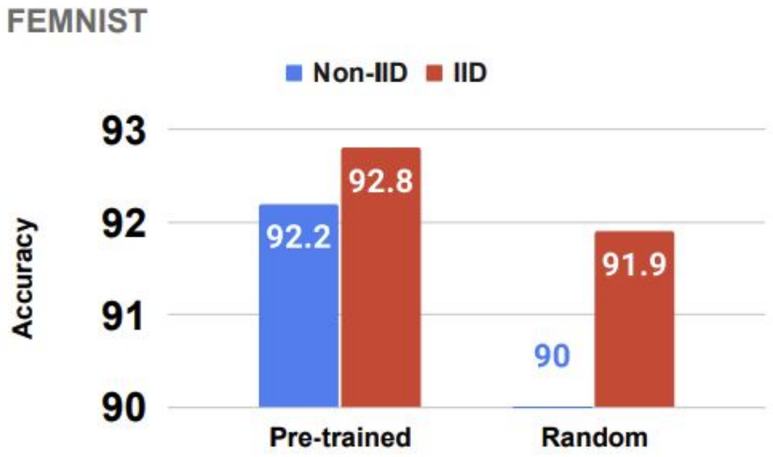
Same conclusion on multiple workloads

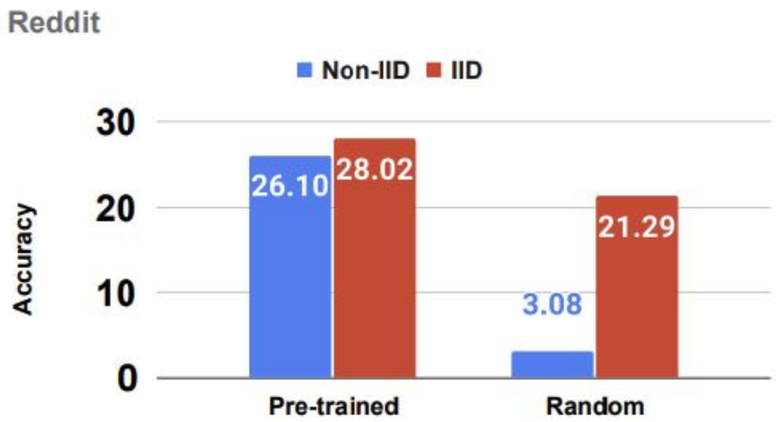




What if there wasn't data heterogeneity? (Shuffle data, then partition across clients)

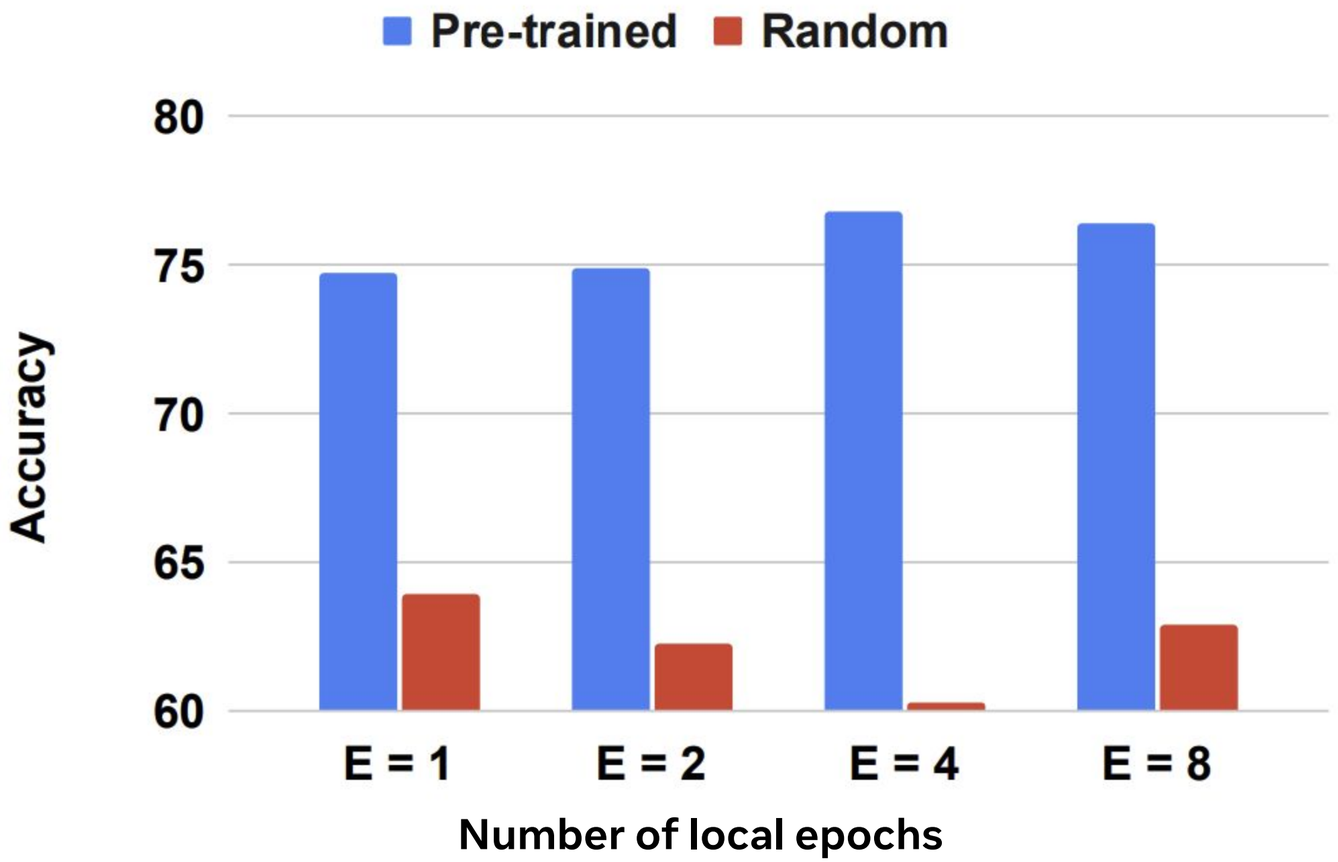






FedAdam less sensitive to data heterogeneity when starting from pre-trained weights

Pre-trained init allows for running more local epochs without degrading performance



FedAdam
CIFAR-10 (Dirichlet, α=0.1)
ResNet-18

Why does Pre-training help Federated Opt?

Hypothesis: Better conditioned loss landscape

	CIFAR-10	FEMNIST	Stack Overflow	Reddit
Pre-trained	661.99	26.29	151.05	647.19
Random	4843.13	355.51	185.02	1309.68

Largest Hessian eigenvalue is consistently lower for pre-trained initialization

Additional results

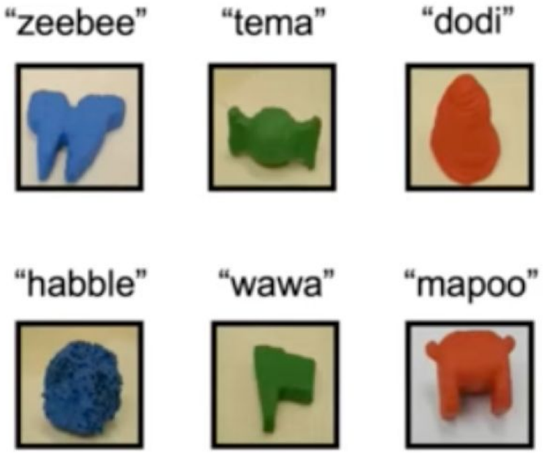
(See the paper for details)

- Pretrained init consistently achieves better final accuracy
- Training loss is not always lower with pretrained initialization
- Local updates not necessary when starting from pretrained init (FedAdam GD)
- Our observations are not completely explained by existing theory

Beyond Supervised Pretraining

(Ongoing & future work)

Novel Words & Objects



Parent-Toddler Free play

Word Learning Test

Pereira, A., Smith, L. B. & Yu, C. (2014) A Bottom-up View of Toddler Word Learning. **Psychonomic Bulletin & Review,** 21, 178-185.
Yu, C. & Smith, L. B. (2012) Embodied Attention and Word Learning by Toddlers. **Cognition**, 125, 244-262
Suanda, S. H., Barnhart, M., Smith, L. B. & Yu, C. (2018) The Signal in the Noise: The Visual Ecology of Parents' Object Naming. **Infancy**, 1-22.

Image Joint-Embedding Predictive Architecture

From a single context block, predict representations of various target blocks...

... no hand-crafted data augmentations!

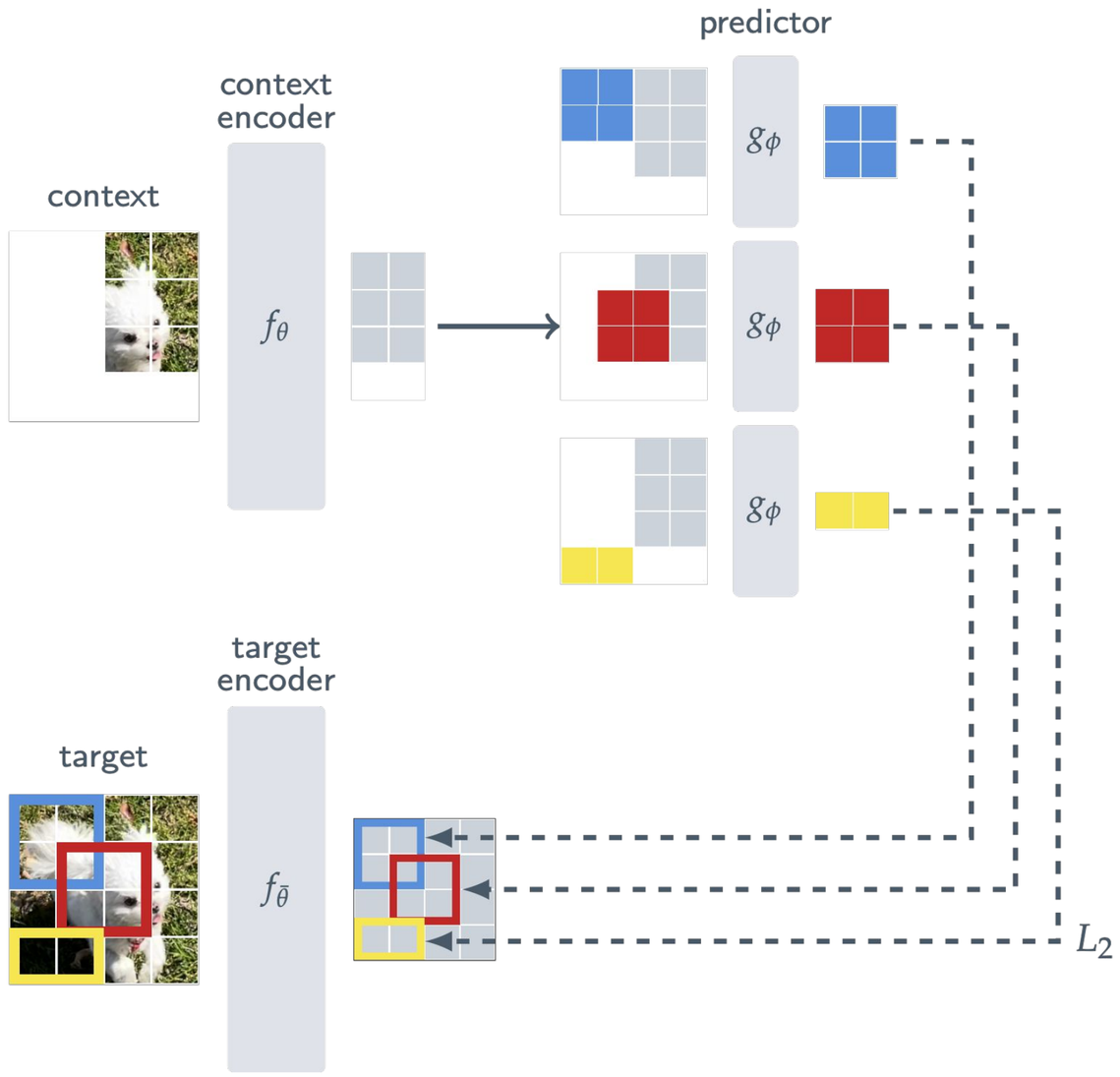


Image Joint-Embedding Predictive Architecture

New SoTA for ImageNet 1% semi-supervised evaluation (~10 labeled examples / class)

Method	Arch.	Epochs	Top-1
Methods without view data augmentations			
data2vec [7]	ViT-L/16	1600	73.3
MAE [34]	ViT-L/16	1600	67.1
	ViT-H/14	1600	71.5
I-JEPA	ViT-L/16	600	69.4
	ViT-H/14	300	73.3
	ViT-H/16_{448}	300	**77.3**
Methods using extra view data augmentations			
iBOT [74]	ViT-B/16	250	69.7
DINO [17]	ViT-B/8	300	70.0
SimCLR v2 [33]	RN151 (2×)	800	70.2
BYOL [33]	RN200 (2×)	800	71.2
MSN [3]	ViT-B/4	300	**75.7**

Come see the poster on Wednesday to learn more!

Poster

Self-Supervised Learning From Images With a Joint-Embedding Predictive Architecture

Mahmoud Assran · Quentin Duval · Ishan Misra · Piotr Bojanowski · Pascal Vincent · Michael Rabbat · Yann LeCun · Nicolas Ballas

West Building Exhibit Halls ABC 311

[Abstract]

[Paper PDF] [Slides] [Poster]

Wed 21 Jun 4:30 p.m. PDT — 6 p.m. PDT (Bookmark)

Conclusion

Pre-training init helps alleviate effects of heterogeneity in FL.

Recommendation: Evaluate methods on random and pre-trained initializations

Future work:

- Theory to better quantify effects of initialization/pre-training in FL (bias, distribution shift)
- Other forms of pre-training (meta-learning, self-supervised)

https://arxiv.org/abs/2210.08090
mikerabbat@meta.com