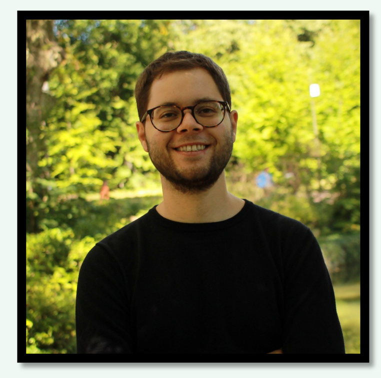
The multiple facets of Variance Reduction in Federated Learning

Angelo Rodio

Inria, Université Côte d'Azur

angelo.rodio@inria.fr

Paper Code

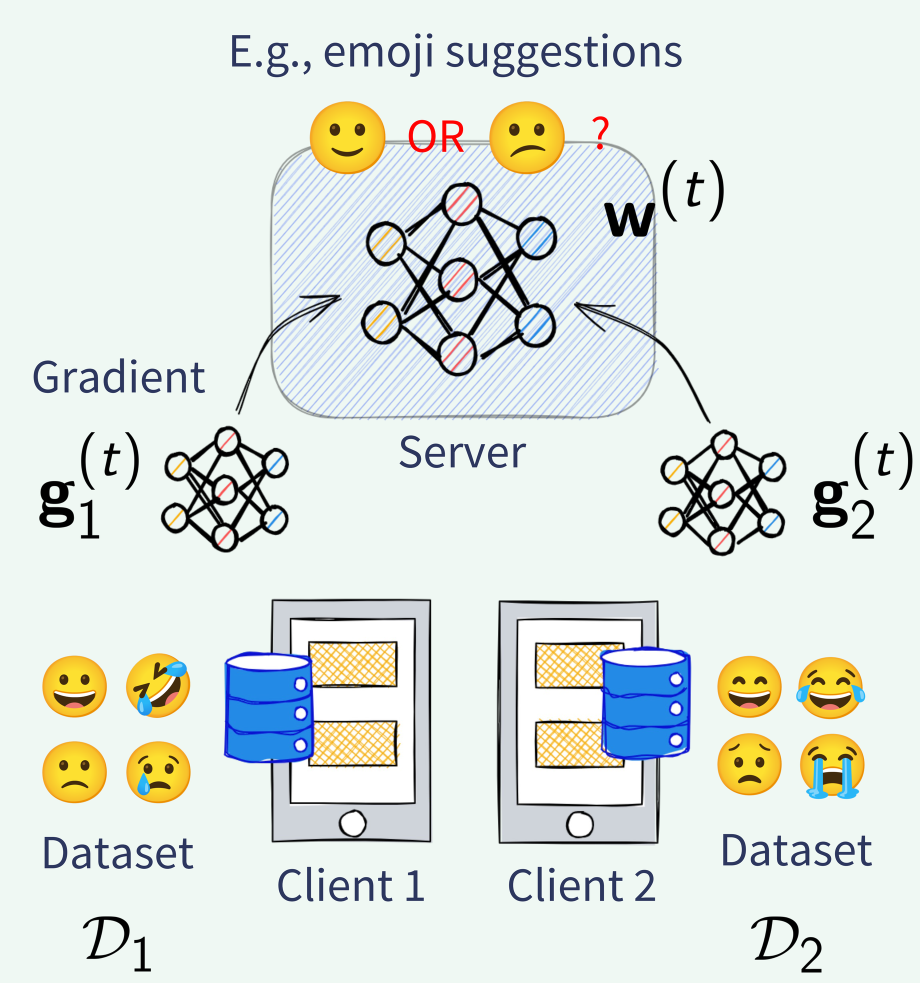


Context

Federated Learning (FL) allows decentralized machine learning model training on client devices (e.g., smartphones)

$$\min_{\mathbf{w}\in\mathbb{R}^d} F(\mathbf{w}) = \frac{1}{N}\sum_{i=1}^{N} F_i(\mathbf{w})$$

$$\underbrace{F_i(\mathbf{w})}_{\text{local objective}} = \mathbb{E}_{z_i\sim\mathcal{D}_i}\,[\underbrace{f}_{\text{loss function}}(\mathbf{w}; \underbrace{z_i}_{\text{data sample}})]$$



Algorithm

At each round t:

- Each client computes $\mathbf{g}_i^{(t)} \leftarrow \texttt{ClientOpt}(\mathbf{w}^{(t)}, \mathcal{D}_i)$
- Server computes $\Delta^{(t)} = \frac{1}{N}\sum_i \mathbf{g}_i^{(t)}$, $\mathbf{w}^{(t+1)} \leftarrow \texttt{ServerOpt}(\mathbf{w}^{(t)}, \Delta^{(t)})$

👍 Communication efficiency 👍 Data privacy

Problem

Data heterogeneity

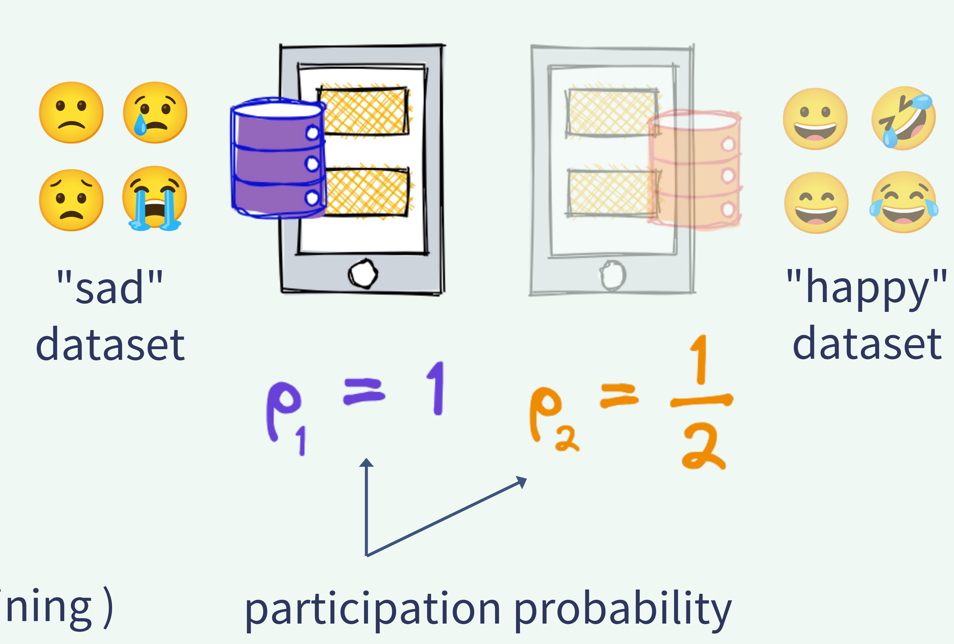
(Client 1: "sad", Client 2: "happy")

+

Client participation heterogeneity

(The "happy" client partakes less in training)

=



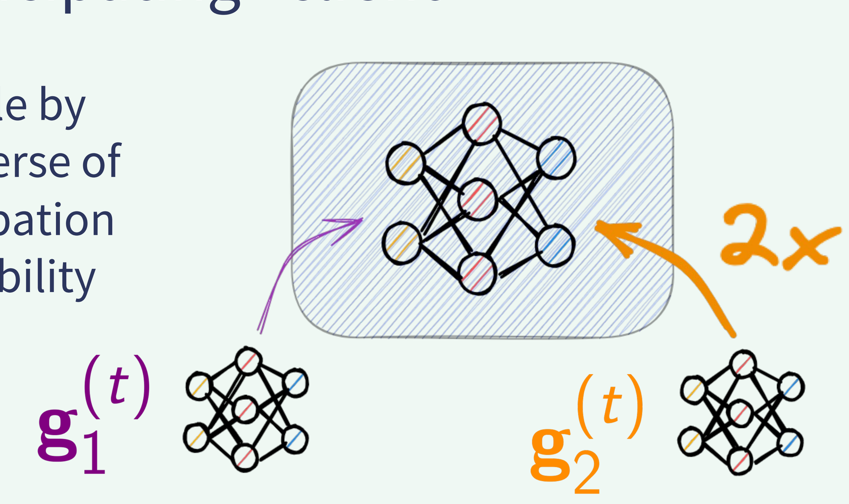
1) Bias in favor of the "more participating" client

(E.g., "I am feeling awesome! 😭")

Solution [1, 2]:

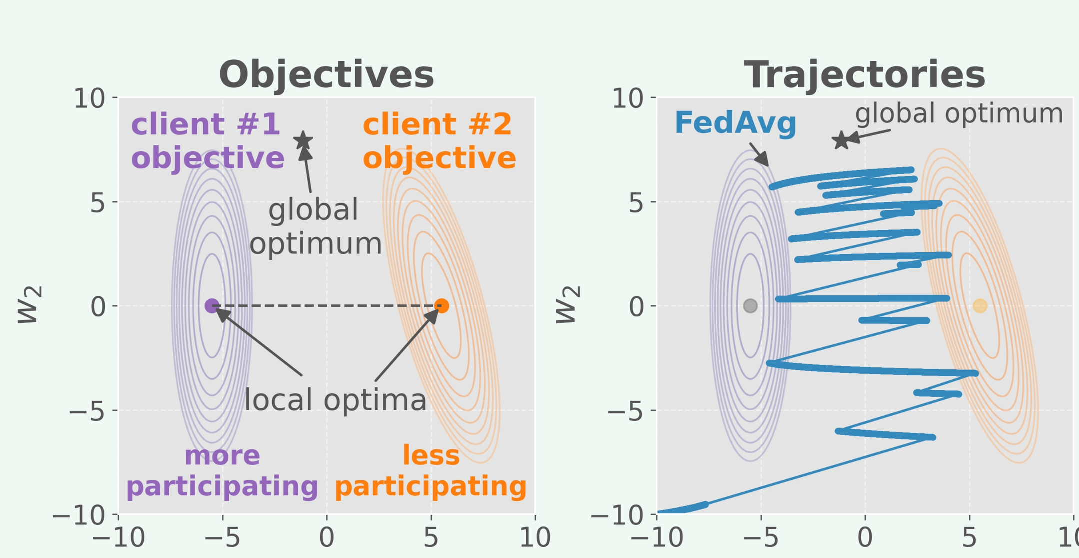
$$\Delta^{(t)}_{\text{FedAvg}} = \frac{1}{N}\sum_{i\in\mathcal{P}^{(t)}} \frac{\mathbf{g}_i^{(t)}}{p_i}$$

rescale by the inverse of the participation probability



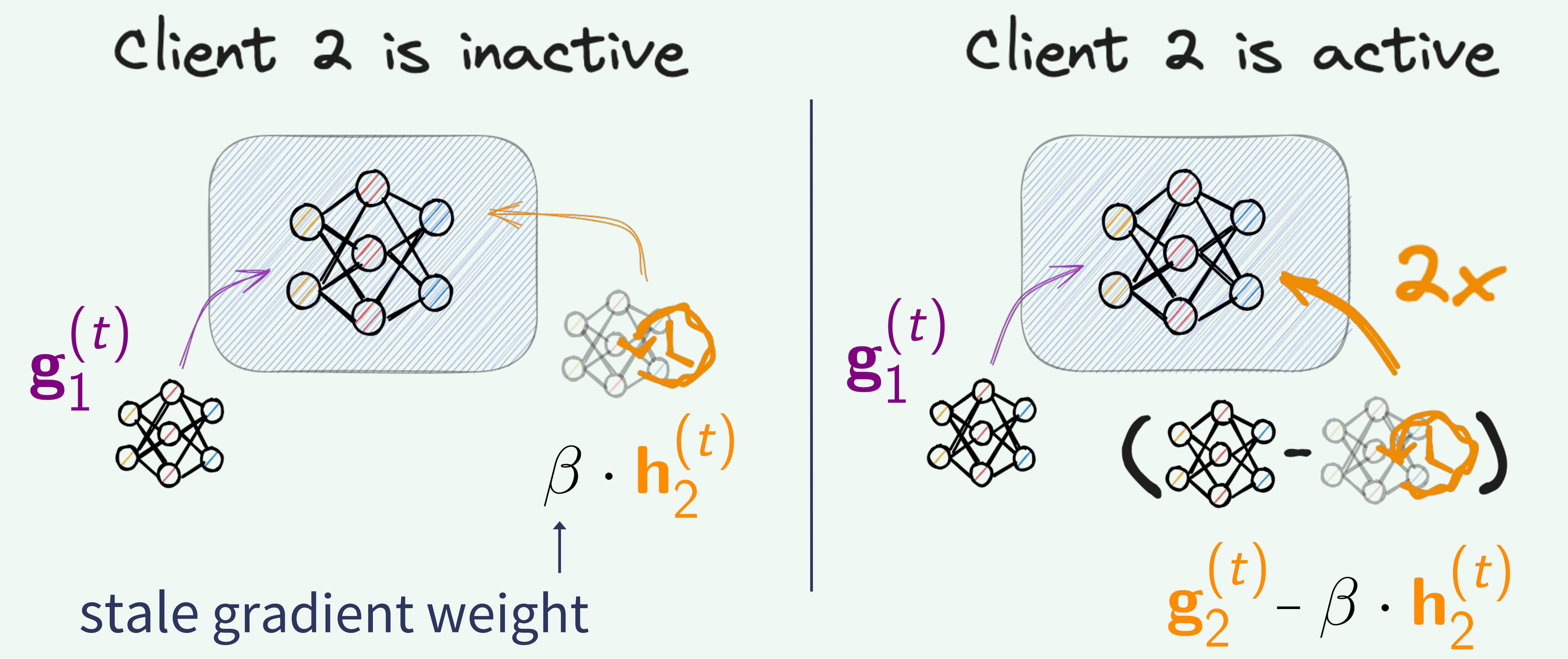
2) Variance in the training process

FedAvg [1, 2] exhibits sub-optimal trajectories and slow convergence



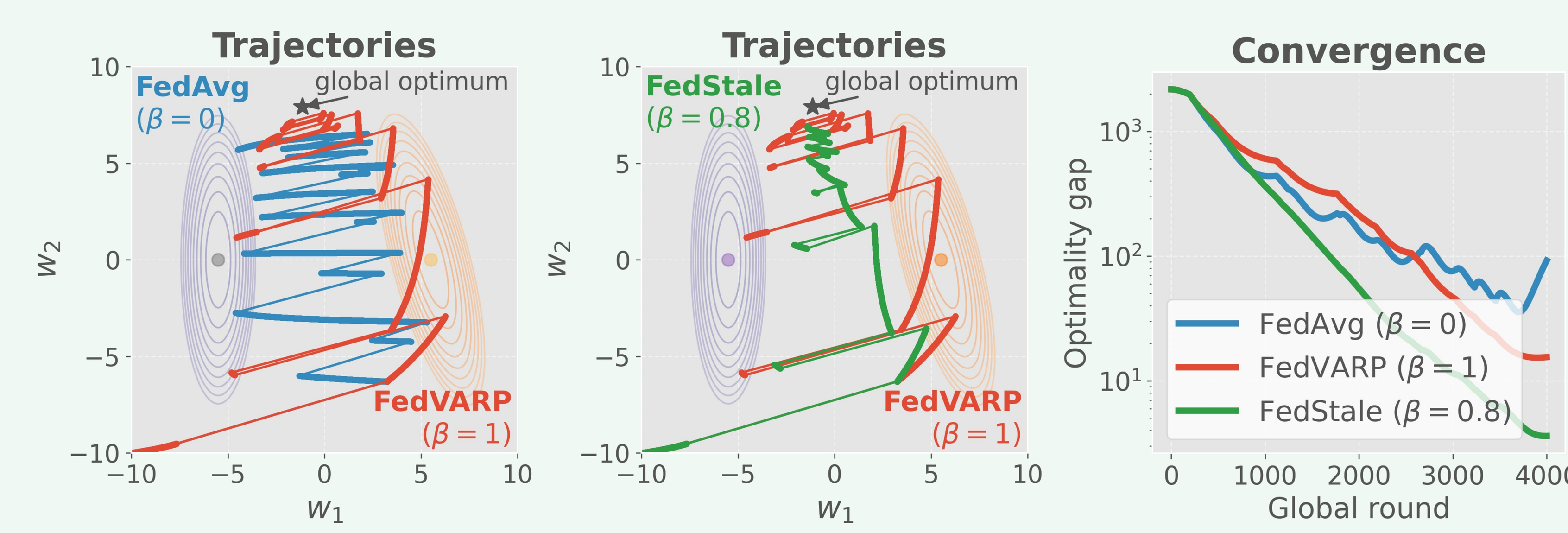
Proposed Solution

A convex combination of "fresh" gradients from participating clients and "stale" gradients from non-participating ones



By leveraging stale gradients for non-participating clients, *FedStale* acts as variance reduction method

$$\Delta^{(t)}_{\text{FedStale}} = \frac{1}{N}\sum_{i=1}^{N} \beta\mathbf{h}_i^{(t)} + \sum_{i\in\mathcal{P}^{(t)}} \frac{\mathbf{g}_i^{(t)} - \beta\mathbf{h}_i^{(t)}}{p_i}$$



Related Work

FedVARP et al. [3-5] assume homogeneous client participation with a fixed $\beta = 1$ (equal weight to fresh and stale gradients)

Theoretical Guarantee

Analyzing *FedStale* convergence, we find stale gradient weight depends on client data and participation heterogeneity

Guideline A: increase stale gradient weight (β) with higher data heterogeneity

Guideline B: decrease stale gradient weight (β) with higher participation heterogeneity ($p_{\text{avg}}/p_{\text{min}}$)

Experiments

Data heterogeneity \ Participation heterogeneity (p_{avg}/p_{min})	1.5	3	5	10	25	50	100	250
1.0	β=1.0	β=0.8	β=0.8	β=0.8	β=0.5	β=1.0	β=0.8	β=0.2
0.8	β=1.0	β=0.8	β=0.8	β=0.8	β=0.5	β=0.5	β=0.5	β=0.2
0.6	β=1.0	β=0.8	β=0.8	β=0.5	β=0.5	β=0.5	β=0.5	β=0.0
0.4	β=1.0	β=0.8	β=0.8	β=0.5	β=0.2	β=0.2	β=0.2	β=0.5
0.2	β=1.0	β=0.8	β=0.8	β=0.2	β=0.0	β=0.0	β=0.0	β=0.2
0.0	β=0.8	β=0.2	β=0.2	β=0.0	β=0.0	β=0.0	β=0.0	β=0.0

β-value yielding the highest test accuracy on the MNIST dataset

Stale gradients can improve or hurt performance based on client data and participation heterogeneity

References

[1] McMahan et al. "Communication-Efficient Learning of Deep Networks from Decentralized Data." In AISTATS, 2017.

[2] Wang et al. "A Unified Analysis of Federated Learning with Arbitrary Client Participation." In NeurIPS, 2022.

[3] Jhunjhunwala et al. "FedVARP: Tackling the Variance Due to Partial Client Participation in Federated Learning." In UAI, 2022.

[4] Gu et al. "Fast Federated Learning in the Presence of Arbitrary Device Unavailability." In NeurIPS, 2021.

[5] Yan et al. "Federated Optimization Under Intermittent Client Availability." In ACM INFORMS, 2024.