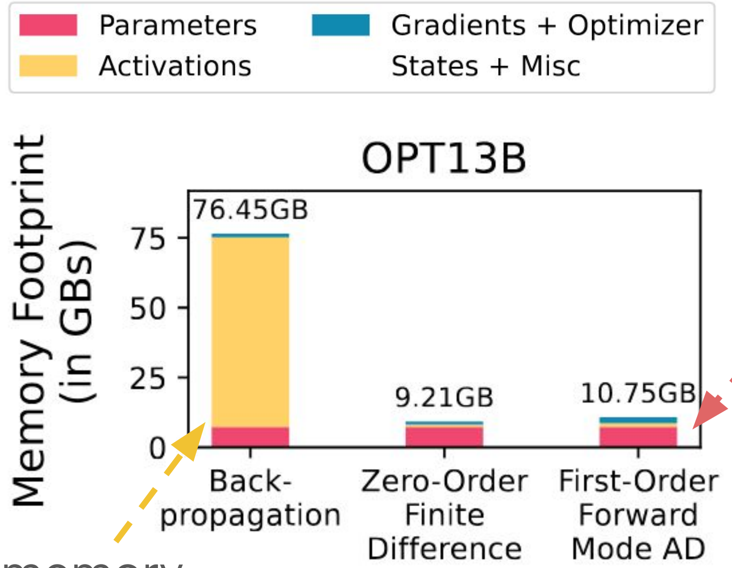
Thinking Forward: Memory-Efficient Federated Finetuning of Language Models (NeurIPS 2024)

by Kunjal Panchal[1] (kpanchal@umass.edu), Nisarg Parikh[1], Sunav Choudhary[2], Lijun Zhang[1], Yuriy Brun[1], Hui Guan[1]

[1]University of Massachusetts – Amherst, [2]Adobe Research

Backpropagation is Memory-Expensive

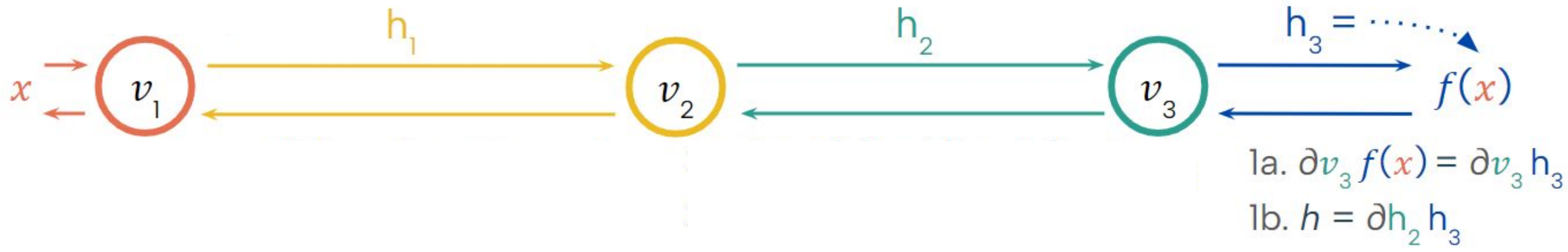


High memory footprint due to activations

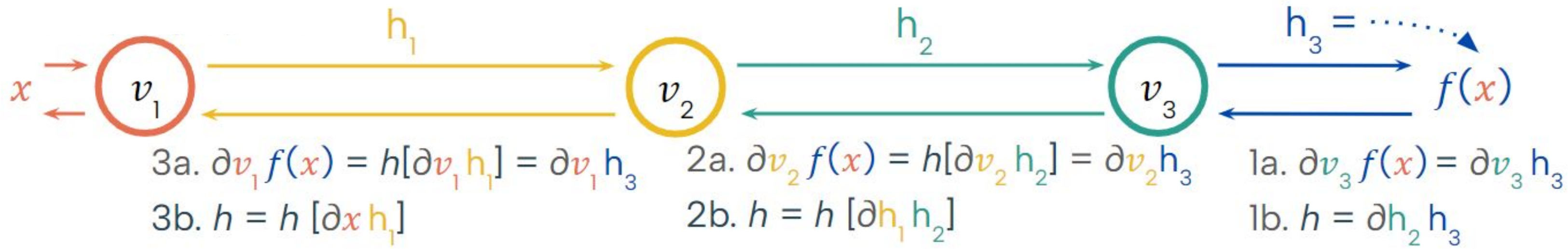
There are methods to derive gradients, with no overhead of intermediate activations.

- Our method (Spry) is based on Forward-mode AD.
- Spry consumes 27.90–86.26% less memory than backpropagation.

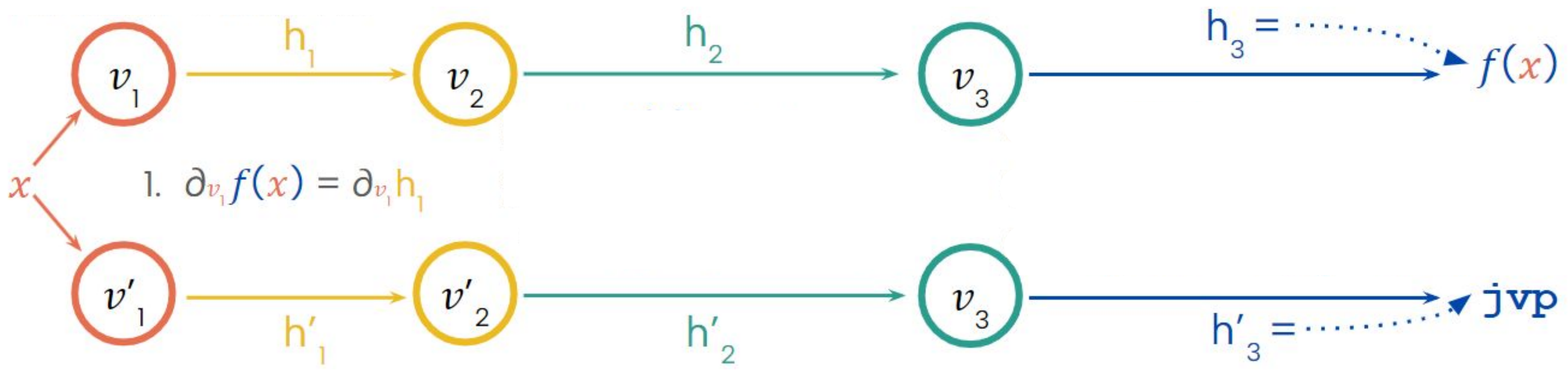
Backpropagation



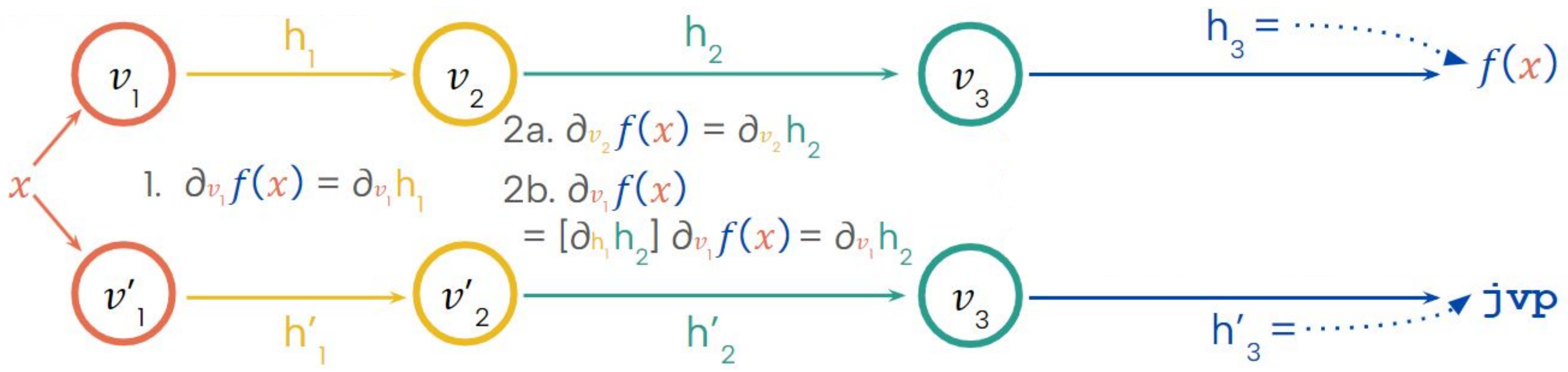
Backpropagation (Continued)



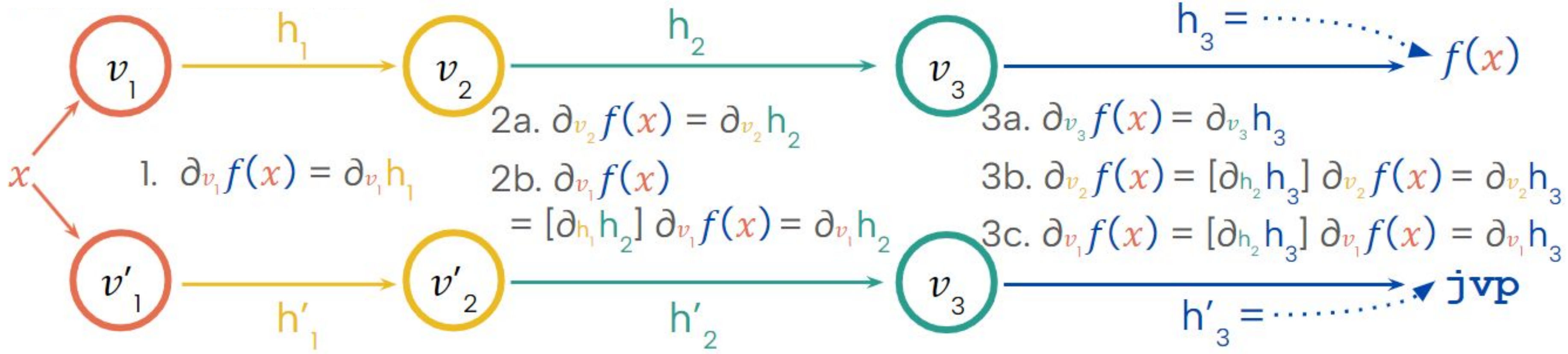
Forward-mode AD



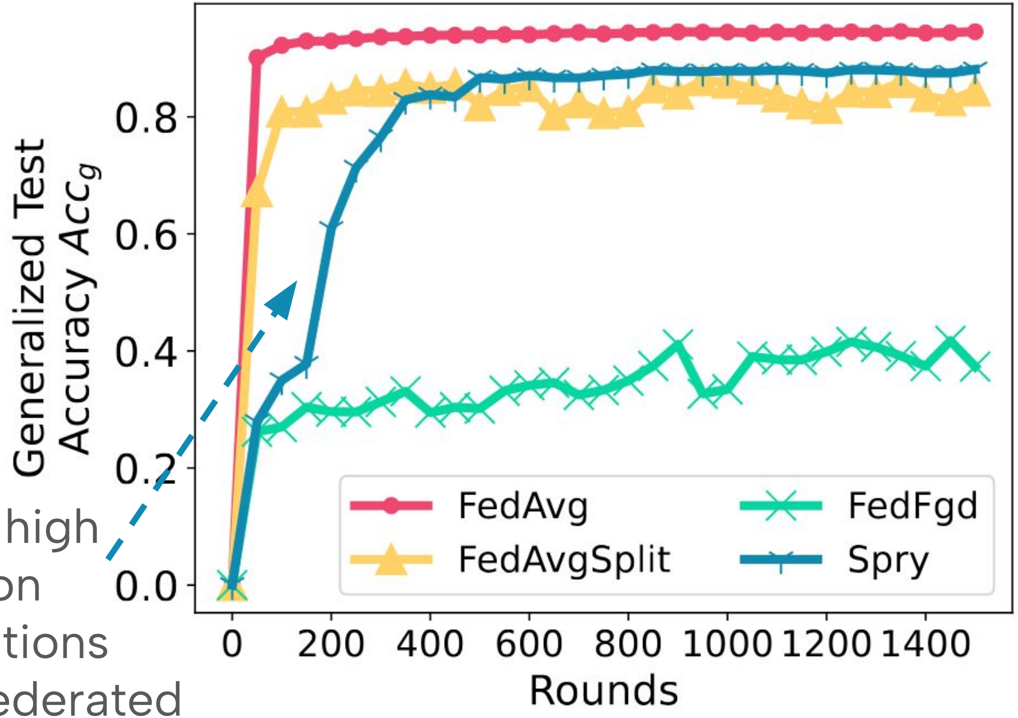
Forward-mode AD (Continued)



Forward-mode AD (Continued)



Forward-mode AD Fails for Large Models

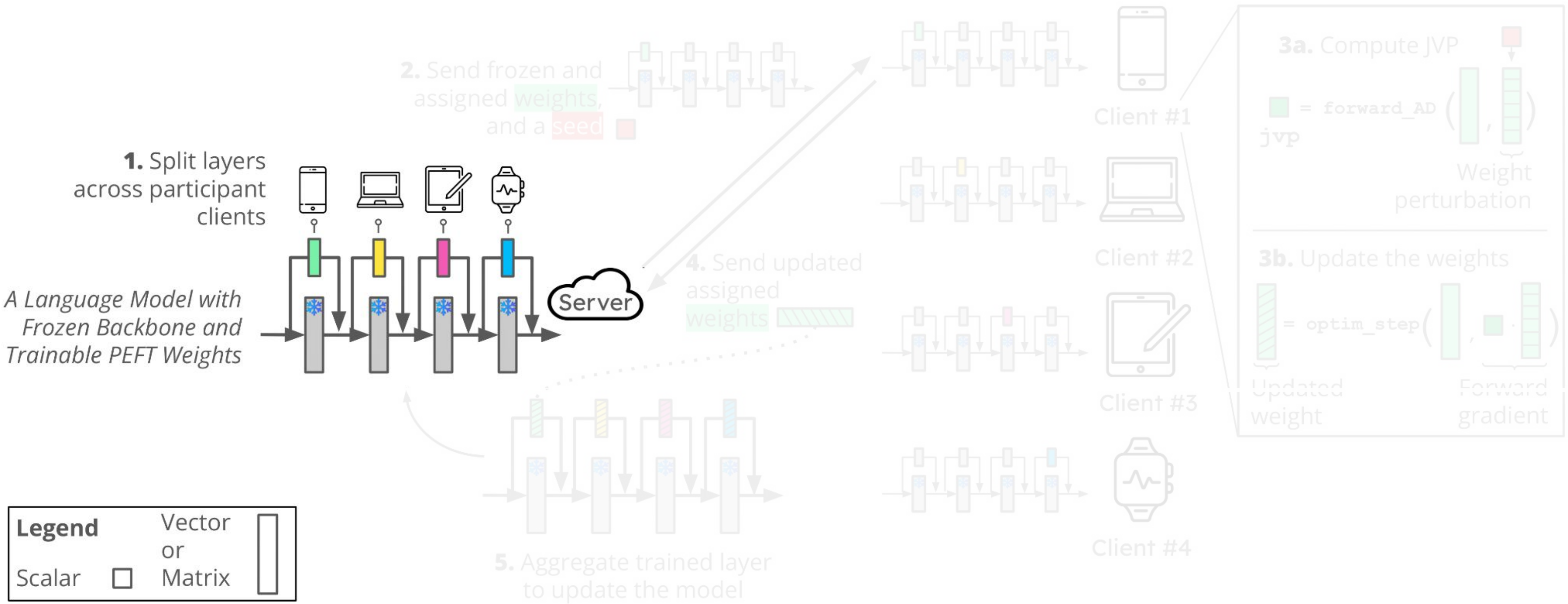


We split high dimension perturbations across federated clients with Spry

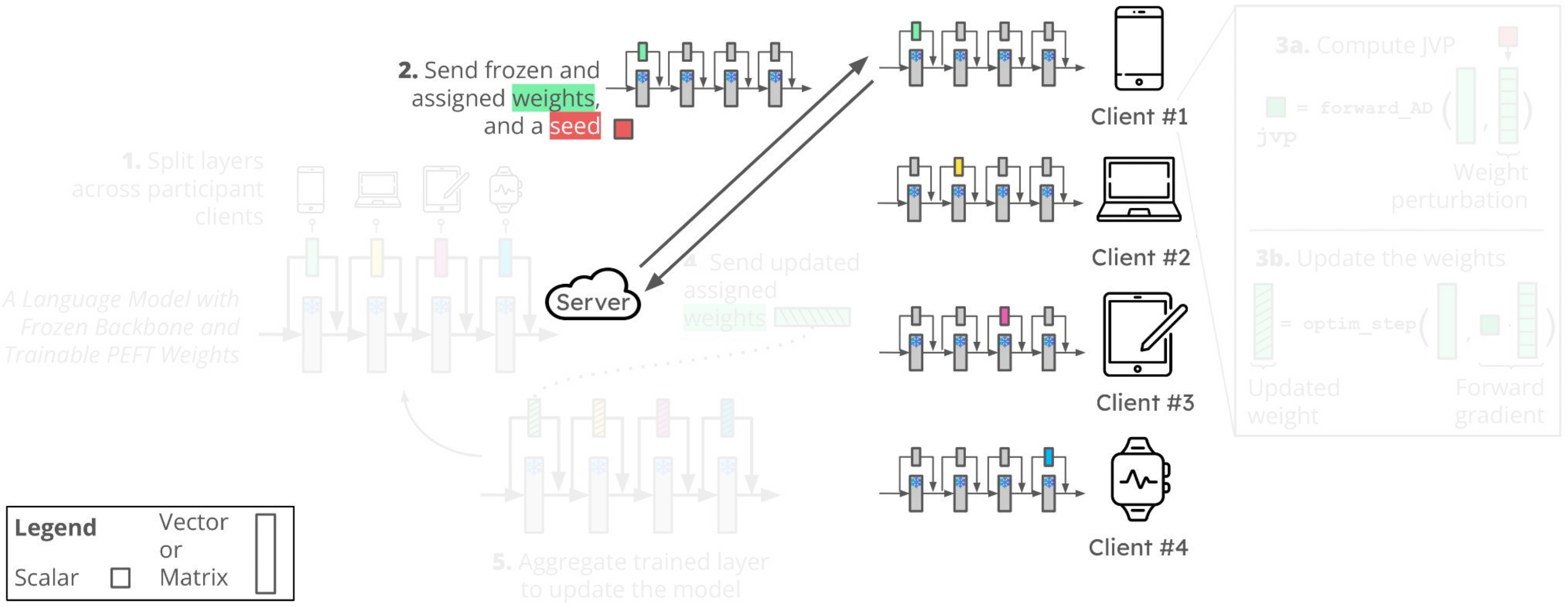
A federated version of FGD

It fails due to high-dimensional random perturbations

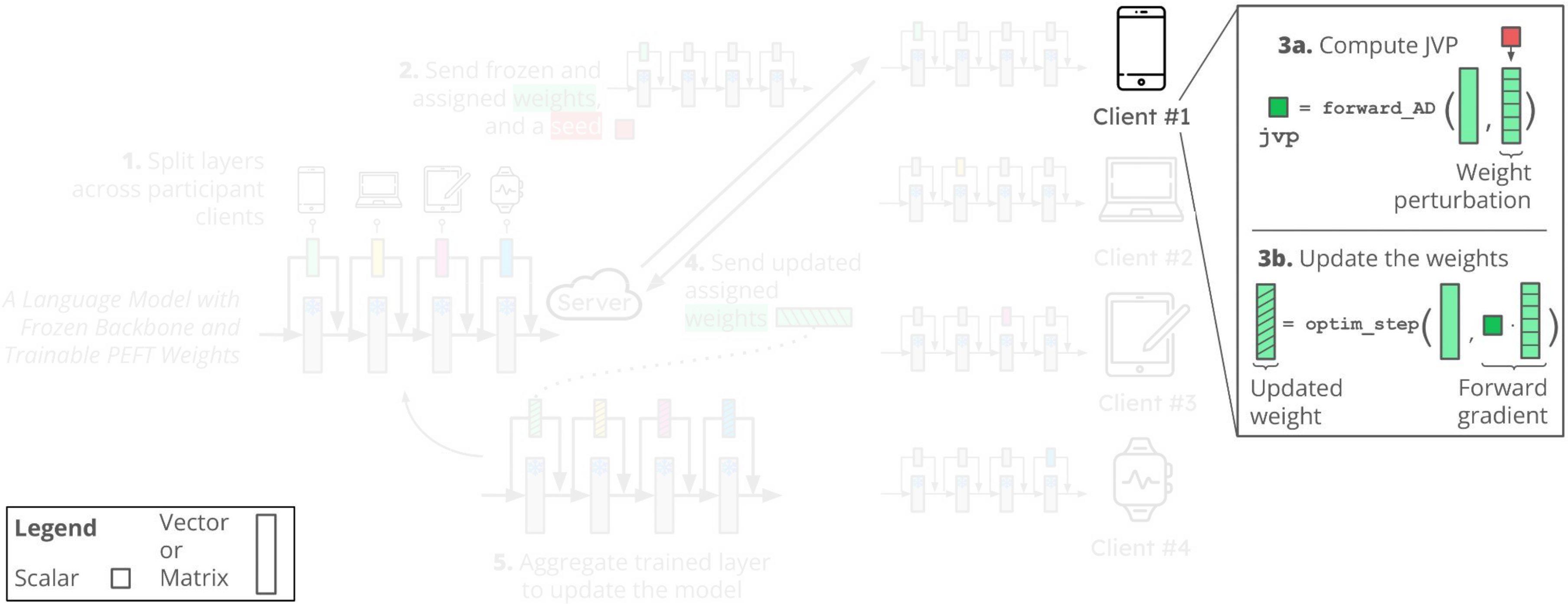
Spry: Making Forward-mode AD Feasible



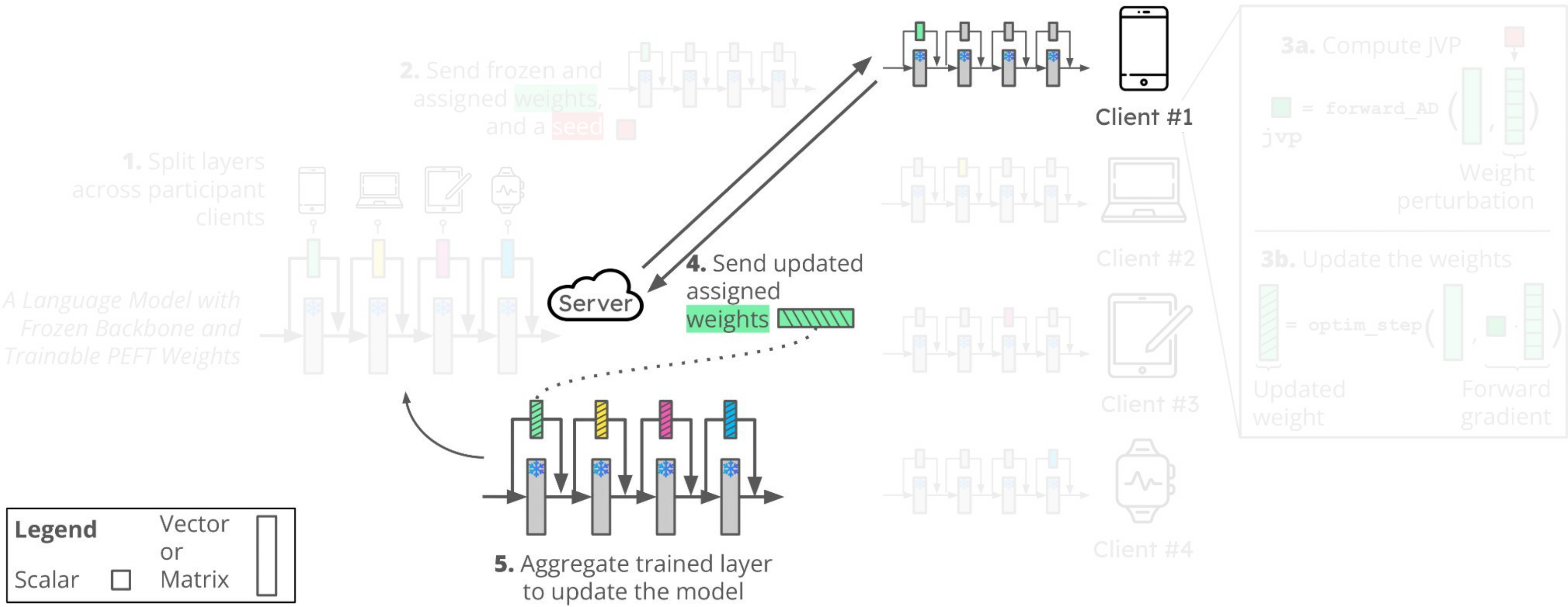
Spry: Making Forward-mode AD Feasible



Spry: Making Forward-mode AD Feasible



Spry: Making Forward-mode AD Feasible

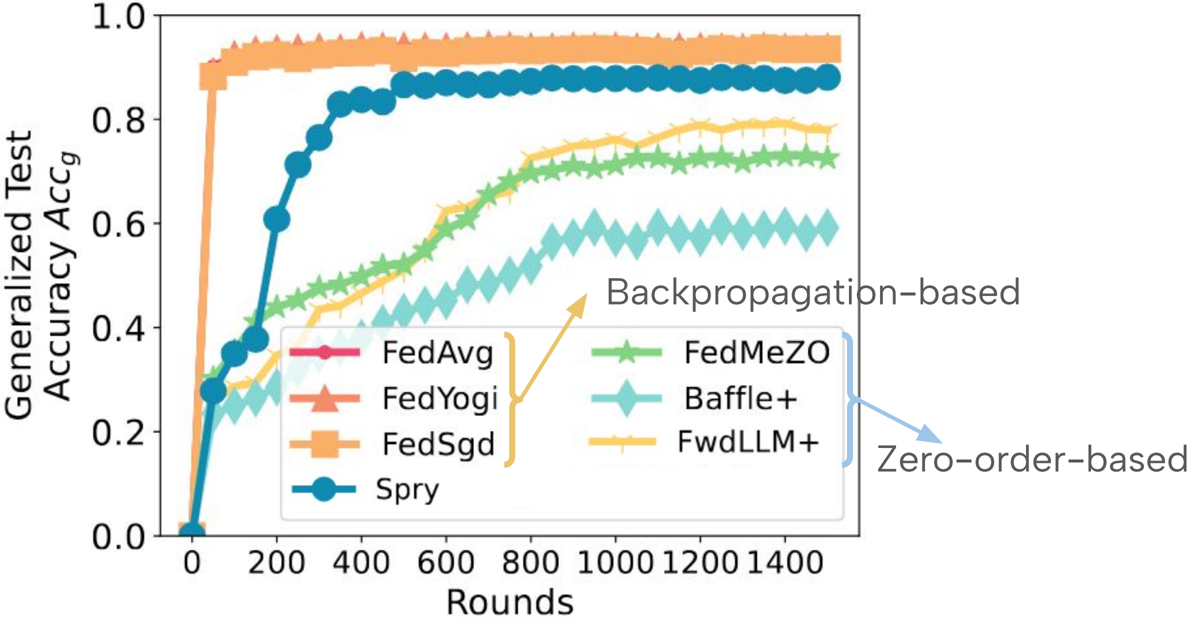


Results: Comparable Accuracy to Backpropagation

Spry achieves

5.15–13.50% higher accuracy than zero-order methods

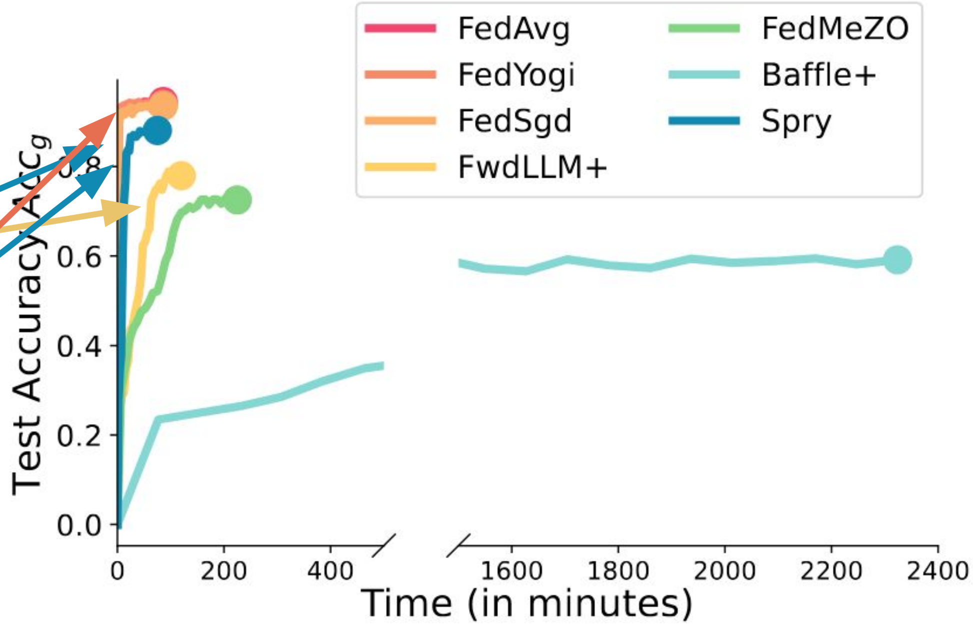
and is within 0.60–6.16% of backprop methods.



Results: Faster Convergence than Zero-order

Spry is 1.15–20.28× faster than the zero-order methods,

although still slower than backprop- based methods.



Takeaway

- Spry is a federated learning algorithm that enables finetuning LLMs using Forward-mode Auto Differentiation.
- It reduces memory footprint during training by 1.4–7.1× in contrast to backpropagation.
- It reduces the convergence time by 1.2–20.3× and achieves 5.2–13.5% higher accuracy against zero-order methods.
- Theoretical analysis shows how Spry's global gradients estimate true gradients based on the heterogeneity of FL clients.

Takeaway

- Spry is a federated learning algorithm that enables finetuning LLMs using Forward-mode Auto Differentiation.
- It reduces memory footprint during training by 1.4–7.1× in contrast to backpropagation.
- It reduces the convergence time by 1.2–20.3× and achieves 5.2–13.5% higher accuracy against zero-order methods.
- Theoretical analysis shows how Spry's global gradients estimate true gradients based on the heterogeneity of FL clients.

Takeaway

- Spry is a federated learning algorithm that enables finetuning LLMs using Forward-mode Auto Differentiation.
- It reduces memory footprint during training by 1.4–7.1× in contrast to backpropagation.
- It reduces the convergence time by 1.2–20.3× and achieves 5.2–13.5% higher accuracy against zero-order methods.
- Theoretical analysis shows how Spry's global gradients estimate true gradients based on the heterogeneity of FL clients.

Takeaway

- Spry is a federated learning algorithm that enables finetuning LLMs using Forward-mode Auto Differentiation.
- It reduces memory footprint during training by 1.4–7.1× in contrast to backpropagation.
- It reduces the convergence time by 1.2–20.3× and achieves 5.2–13.5% higher accuracy against zero-order methods.
- Theoretical analysis shows how Spry's global gradients estimate true gradients based on the heterogeneity of FL clients.