Convergence and Linear Speed-Up in Stochastic Federated Learning

Paul Mangold (CMAP, École Polytechnique)

Séminaire — Université Paris–Dauphine

May 27th, 2025

Me and my research

... about me

My research journey

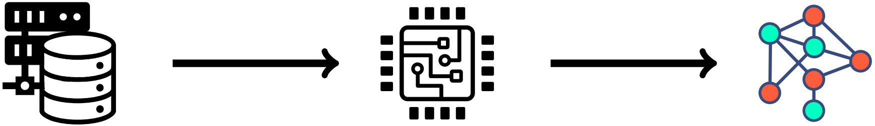
- Research themes: stochastic optimization, privacy, fairness, federated learning, reinforcement learning...
- 2023–present Post-doctoral Researcher (CMAP, École Polytechnique, Paris):
  - Federated (reinforcement) learning
- 2020–2023 PhD (MAGNET team, Inria Lille):
  - Differentially private optimization and fairness
- before: studied at ENS de Lyon

Me and my research

... about my (past) research

Optimization for Machine Learning

Data Optimization Model



$$\min_{x \in \mathbb{R}^d} f(x)$$

Overview of My Research

I. Differentially Private Optimization and Fairness } PhD

II. Federated Stochastic Optimization

III. Federated Reinforcement Learning

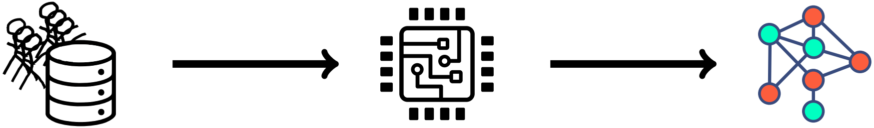
} Post-doc

I. Differentially Private Optimization

Sensitive Data

Optimization

Sensitive Model

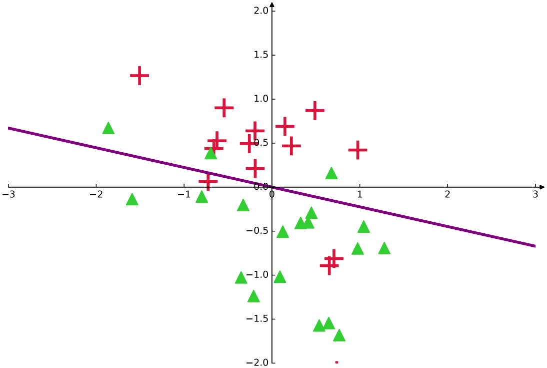


$$\min_{x \in \mathbb{R}^d} f(x)$$

Why is the model sensitive?

Membership Inference:

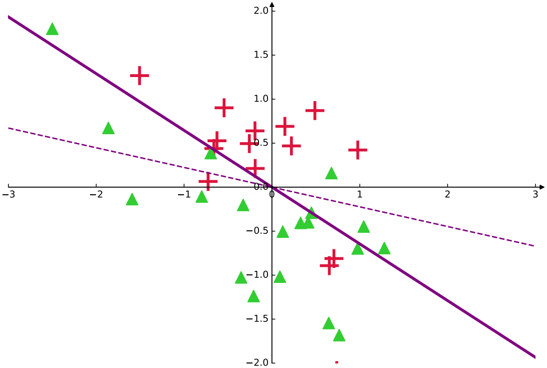
“ guess if an individual was in the training data ”



Why is the model sensitive?

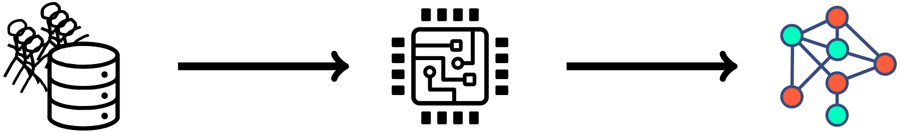
Membership Inference:

“ guess if an individual was in the training data ”



I. Differentially Private Optimization

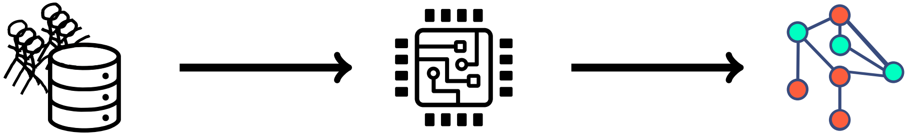
Sensitive Data — Optimization — **Sensitive** Model



$$\min_{x \in \mathbb{R}^d} f(x)$$

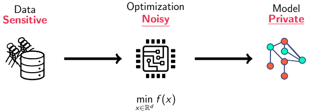
I. Differentially Private Optimization

Sensitive Data → **Noisy** Optimization → **Private** Model



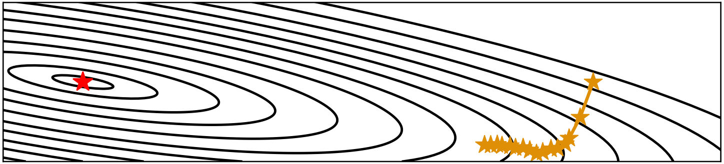
$$\min_{x \in \mathbb{R}^d} f(x)$$

I. Differentially Private Optimization

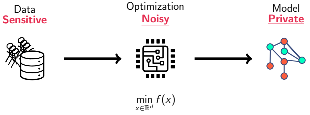


Private gradient descent

$$x^{(t+1)} = x^{(t)} - \gamma\left(\nabla f(x^{(t)}) + \mathcal{N}(0; \sigma^2 I)\right)$$

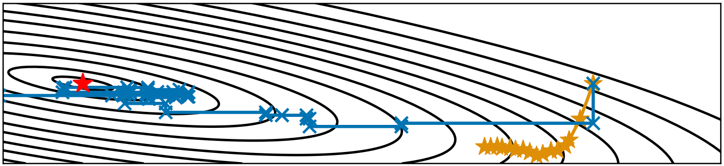


I. Differentially Private Optimization



Private **coordinate descent**[1,2]

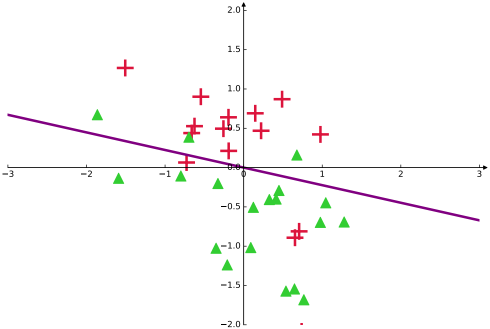
$$x_j^{(t+1)} = x_j^{(t)} - \gamma_j \left(\nabla_j f(x^{(t)}) + \mathcal{N}(0; \sigma_j^2) \right)$$



[1] P. Mangold et al. "Differentially private coordinate descent for composite empirical risk minimization". In: **ICML**. 2022.

[2] P. Mangold et al. "High-dimensional private empirical risk minimization by greedy coordinate descent". In: **AISTATS**. 2023.

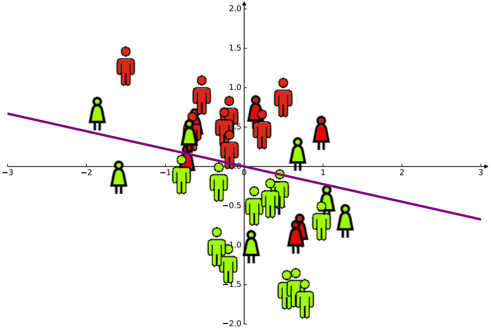
I. What about fairness?



Group Fairness:

All groups must be treated similarly

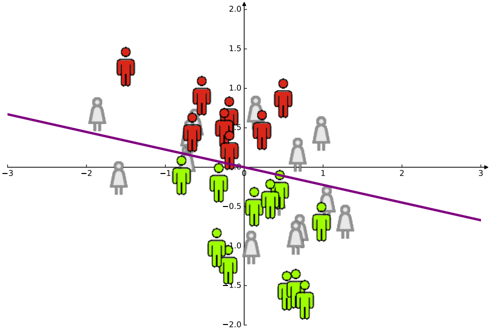
I. What about fairness?



Group Fairness:

All groups must be treated similarly

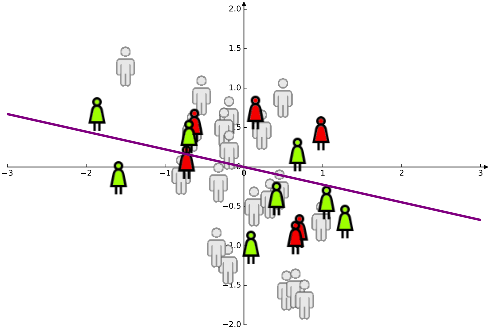
I. What about fairness?



Group Fairness:

All groups must be treated similarly

I. What about fairness?



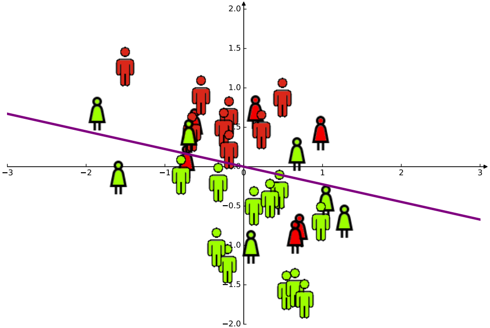
Group Fairness:

All groups must be treated similarly

I. Fairness... and Privacy?

Group Fairness and Privacy:

Perturbing the model can have a disparate impact[1]

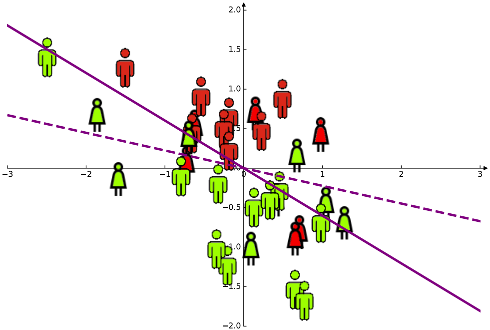


→ but, under some assumptions, this impact remains bounded[2]

[1] E. Bagdasaryan, O. Poursaeed, and V. Shmatikov. "Differential privacy has disparate impact on model accuracy". In: *NeurIPS* (2019).

[2] P. Mangold et al. "Differential privacy has bounded impact on fairness in classification". In: **ICML**. 2023.

I. Fairness… and Privacy?



Group Fairness and Privacy:

Perturbing the model can have a disparate impact[1]

→ but, under some assumptions, this impact remains bounded[2]

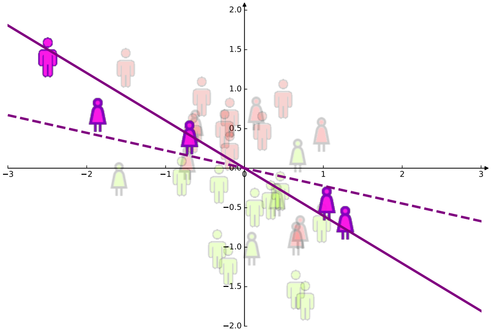
[1] E. Bagdasaryan, O. Poursaeed, and V. Shmatikov. "Differential privacy has disparate impact on model accuracy". In: *NeurIPS* (2019).

[2] P. Mangold et al. "Differential privacy has bounded impact on fairness in classification". In: **ICML**. 2023.

I. Fairness... and Privacy?

Group Fairness and Privacy:

Perturbing the model can have a disparate impact[1]



→ but, under some assumptions, this impact remains bounded[2]

[1] E. Bagdasaryan, O. Poursaeed, and V. Shmatikov. "Differential privacy has disparate impact on model accuracy". In: *NeurIPS* (2019).

[2] P. Mangold et al. "Differential privacy has bounded impact on fairness in classification". In: **ICML**. 2023.

Overview of My Research

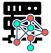
I. Differentially Private Optimization and Fairness } PhD

II. Federated Stochastic Optimization

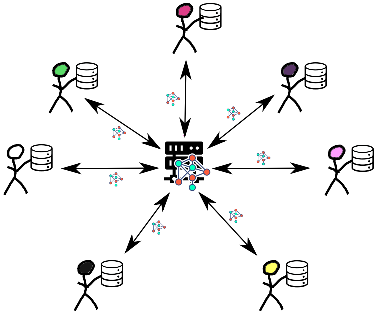
III. Federated Reinforcement Learning

} Post-doc

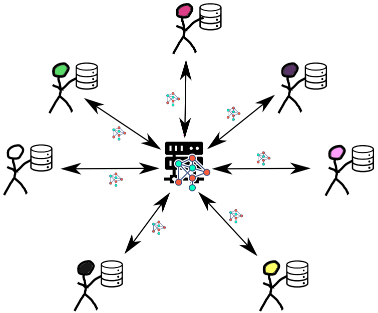
II. Federated Optimization



II. Federated Optimization



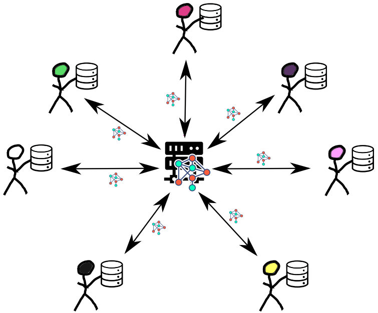
II. Federated Optimization



Collaborative Optimization

$$\min_{x\in\mathbb{R}^d} \frac{1}{N}\sum_{c=1}^{N} f_c(x) \quad , \quad f_c(x) = \mathbb{E}_Z[F_c(x;Z)]$$

II. Federated Optimization



Collaborative Optimization

$$\min_{x \in \mathbb{R}^d} \frac{1}{N} \sum_{c=1}^{N} f_c(x) \quad , \quad f_c(x) = \mathbb{E}_Z[F_c(x; Z)]$$

Central Challenges: data and computational heterogeneity

+ slow and difficult-to-establish communication

II. Federated Optimization

- Theoretical analysis of Federated Averaging[1] and Scaffold[2]
  - First proof showing linear acceleration with the number of clients!
  - Federated methods that correct bias... are still biased!

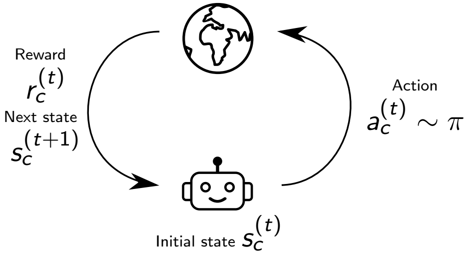
$\Rightarrow$ more details in the second part of this presentation

[1] P. Mangold et al. "Refined Analysis of Federated Averaging's Bias and Federated Richardson-Romberg Extrapolation". In: **AISTATS**. 2025.

[2] P. Mangold et al. "Scaffold with Stochastic Gradients: New Analysis with Linear Speed-Up". In: **ICML**. 2025.

III. Federated Reinforcement Learning

Each agent c operates in its environment independently of others:



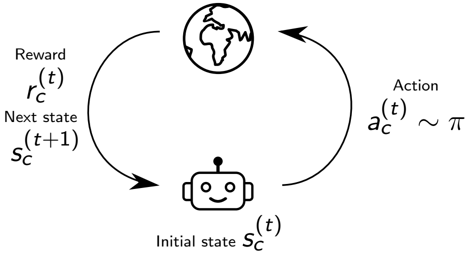
[1] P. Mangold et al. "Scafflsa: Taming heterogeneity in federated linear stochastic approximation and td learning". In: **NeurIPS**. 2024.

[2] S. Labbi et al. "Federated UCBVI: Communication-Efficient Federated Regret Minimization with Heterogeneous Agents". In: **AISTATS**. 2025.

[3] L. Mancini et al. "Integrating FedDRL for Efficient Vehicular Communication in Smart Cities". In: **Internet of Vehicles and Computer Vision Solutions for Smart City Transformations**. Springer, 2025.

III. Federated Reinforcement Learning

Each agent c operates in its environment independently of others:



Some of my work in this area:

- federated TD Learning[1]
- federated value iteration[2]
- federated deep RL for vehicular communications[3]

---

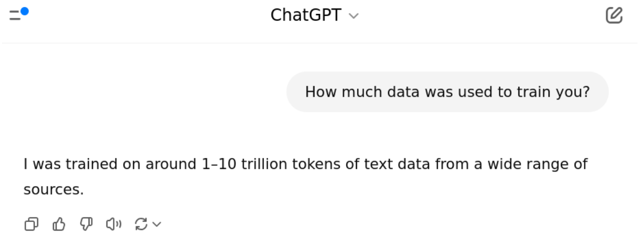
[1] P. Mangold et al. "Scafflsa: Taming heterogeneity in federated linear stochastic approximation and td learning". In: **NeurIPS**. 2024.

[2] S. Labbi et al. "Federated UCBVI: Communication-Efficient Federated Regret Minimization with Heterogeneous Agents". In: **AISTATS**. 2025.

[3] L. Mancini et al. "Integrating FedDRL for Efficient Vehicular Communication in Smart Cities". In: **Internet of Vehicles and Computer Vision Solutions for Smart City Transformations**. Springer, 2025.

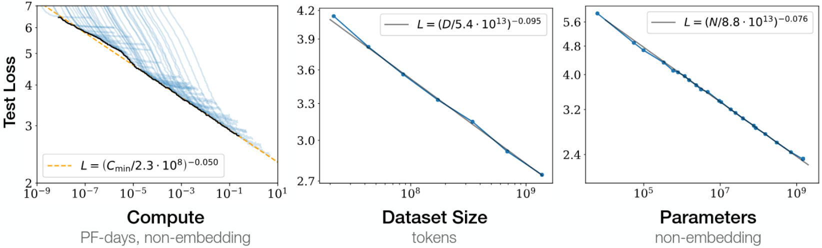
Why federated learning?

Modern Machine Learning Uses Lots of Data



Scaling Laws in Machine Learning

More data and compute give better models (plots from Kaplan et al., 2020[1])

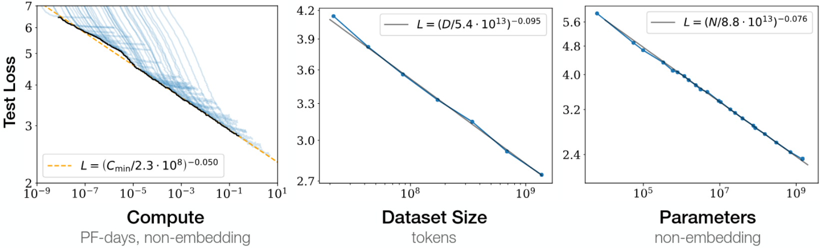


Question: how to collect enough data and compute...

[1] J. Kaplan et al. "Scaling laws for neural language models". In: *arXiv preprint arXiv:2001.08361* (2020).

Scaling Laws in Machine Learning

More data and compute give better models (plots from Kaplan et al., 2020[1])

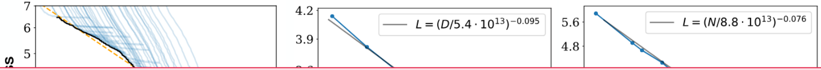


Question: how to collect enough data and compute... when you are not OpenAI?

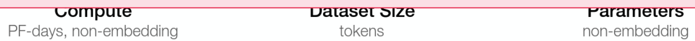
[1] J. Kaplan et al. "Scaling laws for neural language models". In: *arXiv preprint arXiv:2001.08361* (2020).

Scaling Laws in Machine Learning

More data and compute give better models (plots from Kaplan et al., 2020[1])



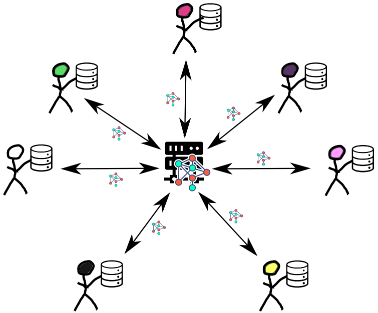
Tentative solution: train a model together!



Question: how to collect enough data and compute... when you are not OpenAI?

[1] J. Kaplan et al. "Scaling laws for neural language models". In: *arXiv preprint arXiv:2001.08361* (2020).

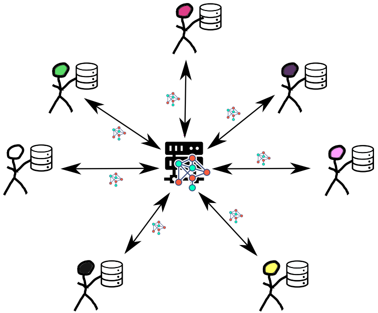
Federated Learning



Collaborative optimization problem

$$\min_{x\in\mathbb{R}^d} \frac{1}{N}\sum_{c=1}^{N} f_c(x) \quad , \quad f_c(x) = \mathbb{E}_Z[F_c(x;Z)]$$

Federated Learning



Collaborative optimization problem

$$\min_{x \in \mathbb{R}^d} \frac{1}{N} \sum_{c=1}^{N} f_c(x) \quad , \quad f_c(x) = \mathbb{E}_Z[F_c(x; Z)]$$

Central Challenges: data and computational heterogeneity

$+$ slow and difficult-to-establish communication

I. Federated Averaging

Federated Averaging[1]

$$x^\star \in \arg\min_{x\in\mathbb{R}^d} \frac{1}{N}\sum_{c=1}^N \mathbb{E}_Z[F_c(x;Z)]$$

At each global iteration

- For $c = 1$ à N in parallel
  - Receive $x^{(t)}$, set $x_c^{(t,0)} = x^{(t)}$
  - For $h = 0$ to $H-1$

$$x_c^{(t,h+1)} = x_c^{(t,h)} - \gamma \nabla F_c(x_c^{(t,h)}; Z_c^{(t,h+1)})$$

- Aggregate local models

$$x^{(t+1)} = \frac{1}{N}\sum_{c=1}^N x_c^{(t,H)}$$

[1] B. McMahan et al. "Communication-efficient learning of deep networks from decentralized data". In: **AISTATS**. 2017.

Federated Averaging[1]

$$x^\star \in \arg\min_{x\in\mathbb{R}^d} \frac{1}{N}\sum_{c=1}^N \mathbb{E}_Z[F_c(x;Z)]$$

At each global iteration

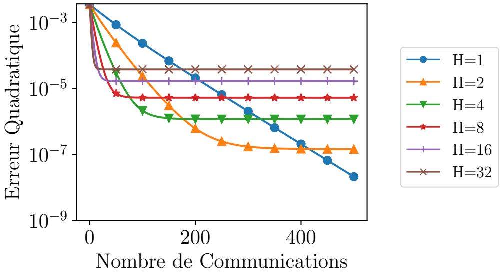
- For $c = 1$ à N in parallel
  - Receive $x^{(t)}$, set $x_c^{(t,0)} = x^{(t)}$
  - For $h = 0$ to $H-1$

$$x_c^{(t,h+1)} = x_c^{(t,h)} - \gamma \nabla F_c(x_c^{(t,h)}; Z_c^{(t,h+1)})$$

- Aggregate local models

$$x^{(t+1)} = \frac{1}{N}\sum_{c=1}^N x_c^{(t,H)}$$

With deterministic gradients:



[1] B. McMahan et al. "Communication-efficient learning of deep networks from decentralized data". In: **AISTATS**. 2017.

Classical analyses of this algorithm

(For L-smooth, μ-strongly convex functions)

Choose your favorite heterogeneity measure

[1] X. Lian et al. “Can decentralized algorithms outperform centralized algorithms? a case study for decentralized parallel SGD”. In: *NeurIPS* (2017).

[2] A. Khaled and C. Jin. “Faster federated optimization under second-order similarity”. In: *arXiv preprint arXiv:2209.02257* (2022).

[3] J. Wang et al. “On the Unreasonable Effectiveness of Federated Averaging with Heterogeneous Data”. In: *TMLR* 2024 (2024).

Classical analyses of this algorithm

(For L-smooth, μ-strongly convex functions)

Choose your favorite heterogeneity measure

- first-order[1]: $\zeta = \frac{1}{N}\sum_{c=1}^{N} \left\|\nabla f_c(x^\star) - \nabla f(x^\star)\right\|^2$

[1] X. Lian et al. "Can decentralized algorithms outperform centralized algorithms? a case study for decentralized parallel SGD". In: *NeurIPS* (2017).

[2] A. Khaled and C. Jin. "Faster federated optimization under second-order similarity". In: *arXiv preprint arXiv:2209.02257* (2022).

[3] J. Wang et al. "On the Unreasonable Effectiveness of Federated Averaging with Heterogeneous Data". In: *TMLR* 2024 (2024).

Classical analyses of this algorithm

(For L-smooth, μ-strongly convex functions)

Choose your favorite heterogeneity measure

- first-order[1]: $\zeta = \frac{1}{N}\sum_{c=1}^{N} \left\| \nabla f_c(x^\star) - \nabla f(x^\star) \right\|^2$
- second-order[2]: $\zeta = \frac{1}{N}\sum_{c=1}^{N} \left\| \nabla_c^2 f(x^\star) - \nabla^2 f(x^\star) \right\|^2$

[1] X. Lian et al. "Can decentralized algorithms outperform centralized algorithms? a case study for decentralized parallel SGD". In: *NeurIPS* (2017).

[2] A. Khaled and C. Jin. "Faster federated optimization under second-order similarity". In: *arXiv preprint arXiv:2209.02257* (2022).

[3] J. Wang et al. "On the Unreasonable Effectiveness of Federated Averaging with Heterogeneous Data". In: *TMLR* 2024 (2024).

Classical analyses of this algorithm

(For L-smooth, μ-strongly convex functions)

Choose your favorite heterogeneity measure

- first-order[1]: $\zeta = \frac{1}{N}\sum_{c=1}^{N} \left\|\nabla f_c(x^\star) - \nabla f(x^\star)\right\|^2$
- second-order[2]: $\zeta = \frac{1}{N}\sum_{c=1}^{N} \left\|\nabla_c^2 f(x^\star) - \nabla^2 f(x^\star)\right\|^2$
- average drift[3]: $\zeta = \left\|\frac{1}{NH}\sum_{c=1}^{N}\sum_{h=0}^{H-1} \nabla f(x_c^{(h)}) - \nabla f(x^\star)\right\|^2$

[1] X. Lian et al. "Can decentralized algorithms outperform centralized algorithms? a case study for decentralized parallel SGD". In: *NeurIPS* (2017).

[2] A. Khaled and C. Jin. "Faster federated optimization under second-order similarity". In: *arXiv preprint arXiv:2209.02257* (2022).

[3] J. Wang et al. "On the Unreasonable Effectiveness of Federated Averaging with Heterogeneous Data". In: *TMLR* 2024 (2024).

Classical analyses of this algorithm

(For L-smooth, μ-strongly convex functions)

Choose your favorite heterogeneity measure

- first-order[1]: $\zeta = \frac{1}{N}\sum_{c=1}^{N} \left\| \nabla f_c(x^\star) - \nabla f(x^\star) \right\|^2$
- second-order[2]: $\zeta = \frac{1}{N}\sum_{c=1}^{N} \left\| \nabla_c^2 f(x^\star) - \nabla^2 f(x^\star) \right\|^2$
- average drift[3]: $\zeta = \left\| \frac{1}{NH}\sum_{c=1}^{N}\sum_{h=0}^{H-1} \nabla f(x_c^{(h)}) - \nabla f(x^\star) \right\|^2$

Show **convergence to a neighborhood** of $x^\star$

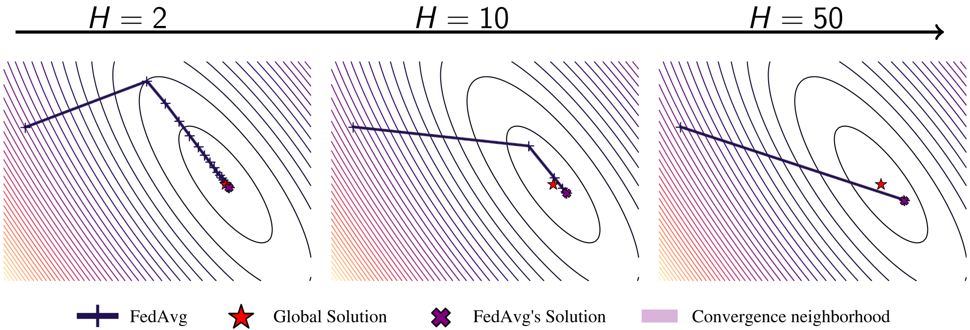
$$\|x^{(T)} - x^\star\|^2 \lesssim (1-\gamma\mu)^{HT}\|x^{(0)} - x^\star\|^2 + \boldsymbol{\chi(\gamma, H, \zeta)} \qquad \text{(for some function } \chi\text{)}$$

[1] X. Lian et al. "Can decentralized algorithms outperform centralized algorithms? a case study for decentralized parallel SGD". In: *NeurIPS* (2017).

[2] A. Khaled and C. Jin. "Faster federated optimization under second-order similarity". In: *arXiv preprint arXiv:2209.02257* (2022).

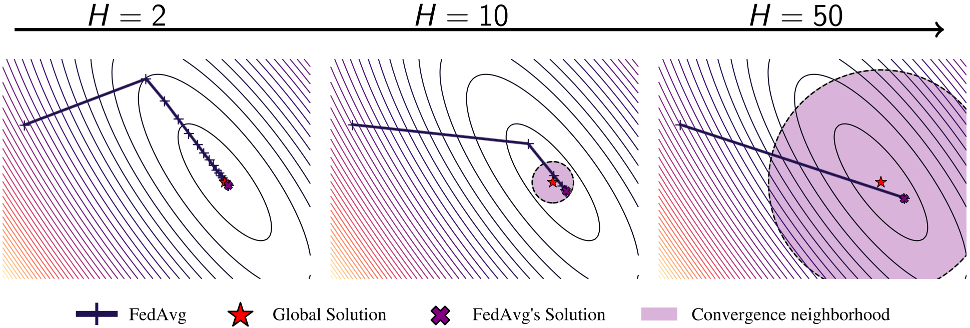
[3] J. Wang et al. "On the Unreasonable Effectiveness of Federated Averaging with Heterogeneous Data". In: *TMLR* 2024 (2024).

$$\|x^{(T)} - x^\star\|^2 \lesssim (1-\gamma\mu)^{HT}\|x^{(0)} - x^\star\|^2 + \chi(\gamma, H, \zeta) \qquad \text{(for some function } \chi\text{)}$$



When the number of local iterations increases, bias incrases

$$\|x^{(T)} - x^\star\|^2 \lesssim (1-\gamma\mu)^{HT}\|x^{(0)} - x^\star\|^2 + \chi(\gamma, H, \zeta) \qquad \text{(for some function } \chi)$$



When the number of local iterations increases, bias incrases

Remark: It seems that iterates converge in some way?

Federated Averaging as Fixed Point Iteration

Remark that

$$x_c^{(t,h+1)} - y_c^{(t,h+1)} = x_c^{(t,h)} - y_c^{(t,h)} - \gamma(\nabla f_c(x_c^{(t,h)}) - \nabla f_c(y_c^{(t,h)}))$$

Thus

$$\|x_c^{(t+1)} - y_c^{(t+1)}\| \leq (1-\gamma\mu)^H \|x_c^{(t)} - y_c^{(t)}\|$$

[1] G. Malinovskiy et al. "From local SGD to local fixed-point methods for federated learning". In: **ICML**. 2020.

Federated Averaging as Fixed Point Iteration

Remark that

$$x_c^{(t,h+1)} - y_c^{(t,h+1)} = x_c^{(t,h)} - y_c^{(t,h)} - \gamma(\nabla f_c(x_c^{(t,h)}) - \nabla f_c(y_c^{(t,h)}))$$

Thus

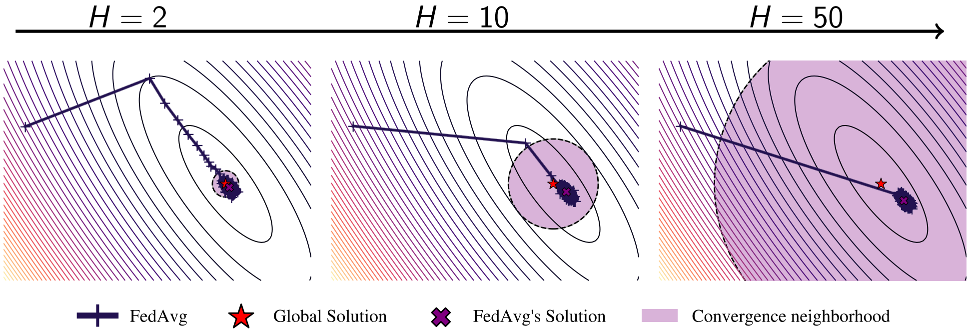
$$\|x_c^{(t+1)} - y_c^{(t+1)}\| \leq (1 - \gamma\mu)^H \|x_c^{(t)} - y_c^{(t)}\|$$

$\Rightarrow$ deterministic FedAvg converges to a unique point[1]

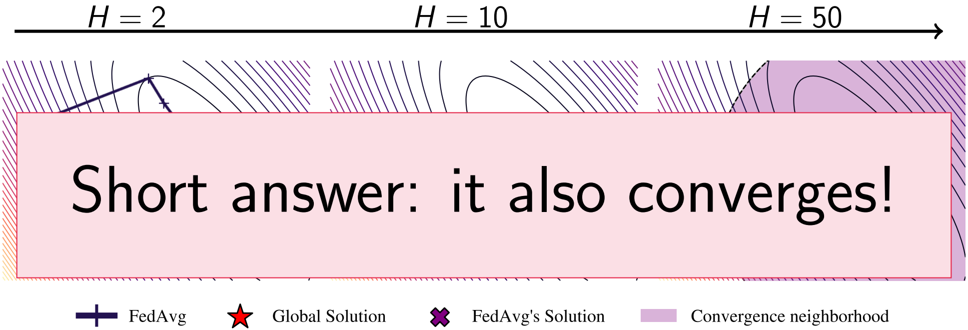
[1] G. Malinovskiy et al. "From local SGD to local fixed-point methods for federated learning". In: **ICML**. 2020.

Open Question: What about the Stochastic Case?

Open Question: What about the Stochastic Case?



Open Question: What about the Stochastic Case?



Short answer: it also converges!

FedAvg (with stochastic gradients) converges![1]

(For thrice derivable, L-smooth, μ-strongly convex functions)

- FedAvg converges to a stationary distribution $\pi^{(\gamma,H)}$
  - denoting $x^{(t)} \sim \psi_{x^{(t)}}$, we have

$$\mathcal{W}_2(\psi_{x^{(t)}}; \pi^{(\gamma,H)}) \leq (1 - \gamma\mu)^{Ht} \mathcal{W}_2(\psi_{x^{(0)}}; \pi^{(\gamma,H)})$$

  - where $\mathcal{W}_2$ is the second order Wasserstein distance

[1] P. Mangold et al. "Refined Analysis of Federated Averaging's Bias and Federated Richardson-Romberg Extrapolation". In: **AISTATS**. 2025.

FedAvg (with stochastic gradients) converges![1]

(For thrice derivable, L-smooth, μ-strongly convex functions)

- FedAvg converges to a stationary distribution $\pi^{(\gamma,H)}$
- FedAvg's iterates covariance is

$$\int (x - x^\star)(x - x^\star)^\top \pi^{(\gamma,H)}(\mathrm{d}x) = \boxed{\frac{\gamma}{N} C(x^\star)} + O(\gamma^{3/2} H)$$

[1] P. Mangold et al. "Refined Analysis of Federated Averaging's Bias and Federated Richardson-Romberg Extrapolation". In: **AISTATS**. 2025.

FedAvg … dients) converges![1]

(F … ngly convex functions)

Linear speed-up !

variance decreases in $1/N$

$C(x^\star)$ is ∇F^Z's covariance at $x^\star$

- FedAvg converges to a stationary distribution $\pi^{(\gamma,H)}$
- FedAvg's iterates covariance is

$$\int (x - x^\star)(x - x^\star)^\top \pi^{(\gamma,H)}(\mathrm{d}x) = \boxed{\frac{\gamma}{N} C(x^\star)} + O(\gamma^{3/2} H)$$

[1] P. Mangold et al. "Refined Analysis of Federated Averaging's Bias and Federated Richardson-Romberg Extrapolation". In: **AISTATS**. 2025.

FedAvg (with stochastic gradients) converges![1]

(For thrice derivable, L-smooth, μ-strongly convex functions)

- FedAvg converges to a stationary distribution $\pi^{(\gamma,H)}$
- FedAvg's iterates covariance is
- We can now give an **exact expansion of the bias**

$$\int x\pi^{(\gamma,H)}(\mathrm{d}x) = x^\star + \boxed{\frac{\gamma(H-1)}{2N}\sum_{c=1}^{N}\nabla^2 f(x^\star)^{-1}\big(\nabla^2 f_c(x^\star) - \nabla^2 f(x^\star)\big)\nabla f_c(x^\star)}$$
$$- \boxed{\frac{\gamma}{2N}\nabla^2 f(x^\star)^{-1}\nabla^3 f(x^\star)A^{-1}C(x^\star)} + O(\gamma^{3/2}H)$$

[1] P. Mangold et al. "Refined Analysis of Federated Averaging's Bias and Federated Richardson-Romberg Extrapolation". In: **AISTATS**. 2025.

FedAvg (with stochastic gradients)[1]

oth, μ-st

Heterogeneity bias

vanishes when $\nabla^2 f_c(x^\star) = \nabla^2 f(x^\star)$

or when $\nabla f_c(x^\star) = \nabla f(x^\star)$

Stochasticity bias

$A = I \otimes \nabla^2 f(x^\star) + \nabla^2 f(x^\star) \otimes I$

$C(x^\star)$ is ∇F^Z's covariance at $x^\star$

- FedAvg converges to a stationary distribution $\pi^{(\gamma,H)}$
- FedAvg's iterates covariance is
- We can now give an **exact expansion of the bias**

$$\int x \pi^{(\gamma,H)}(\mathrm{d}x) = x^\star + \boxed{\frac{\gamma(H-1)}{2N} \sum_{c=1}^{N} \nabla^2 f(x^\star)^{-1} \left(\nabla^2 f_c(x^\star) - \nabla^2 f(x^\star)\right) \nabla f_c(x^\star)}$$

$$- \boxed{\frac{\gamma}{2N} \nabla^2 f(x^\star)^{-1} \nabla^3 f(x^\star) A^{-1} C(x^\star)} + O(\gamma^{3/2} H)$$

[1] P. Mangold et al. "Refined Analysis of Federated Averaging's Bias and Federated Richardson-Romberg Extrapolation". In: **AISTATS**. 2025.

Correcting the Bias

Novel Algorithm: Federated Richardson-Romberg Extrapolation

Run FedAvg twice:

- with step size γ: global iterates $x_\gamma^{(t)}$
- with step size 2γ: global iterates $x_{2\gamma}^{(t)}$

We can combine the iterates

$$\chi_{\mathrm{RR}}^{(t)} = 2x_\gamma^{(t)} - x_{2\gamma}^{(t)}$$

Correcting the Bias

Novel Algorithm: Federated Richardson-Romberg Extrapolation

Run FedAvg twice:

- with s
- with s

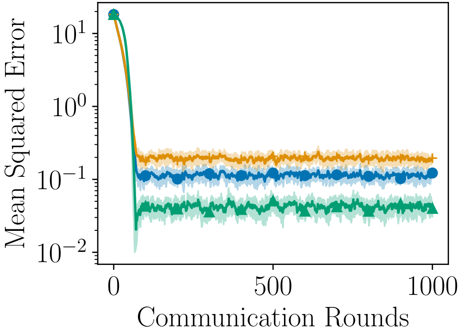
We can co

Theorem: $\mathbb{E}[\chi_{\mathrm{RR}}^{(t)}] = x_\star + O(\gamma^2 H^2 + \gamma^{3/2} H)$

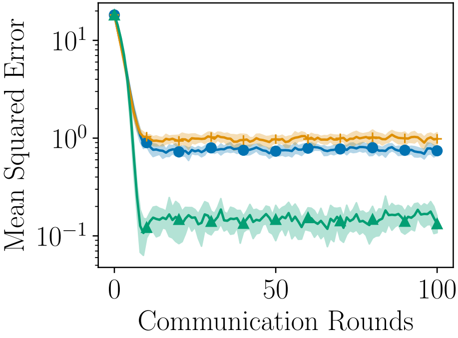
$\rightarrow$ bias is effectively reduced!!

$$\chi_{\mathrm{RR}}^{(t)} = 2x_\gamma^{(t)} - x_{2\gamma}^{(t)}$$

Numerical Illustration: FedAvg



(a) $H = 10$



(b) $H = 100$

Blue: FedAvg, Orange: Scaffold, Green: Federated Richardson-Romberg

II. Correcting heterogeneity with Scaffold

Scaffold[1]

$$x^\star \in \arg\min_{x\in\mathbb{R}^d} \frac{1}{N}\sum_{c=1}^N \mathbb{E}_Z[F_c(x;Z)]$$

At each global iteration

- For $c = 1$ to N in parallel
  - Receive $x^{(t)}$, set $x_c^{(t,0)} = x^{(t)}$
  - For $h = 0$ to $H-1$

$$x_c^{(t,h+1)} = x_c^{(t,h)} - \gamma\big(\nabla F_c(x_c^{(t,h)}; Z_c^{(t,h+1)}) + \boldsymbol{\xi_c^{(t)}}\big)$$

- Aggregate models, update control variates

$$x^{(t+1)} = \frac{1}{N}\sum_{c=1}^N x_c^{(t,H)}$$

$$\boldsymbol{\xi_c^{(t+1)} = \xi_c^{(t)} + \frac{1}{\gamma H}(\theta_c^{t,H} - \theta^{(t+1)})}$$

[1] S. P. Karimireddy et al. “Scaffold: Stochastic controlled averaging for federated learning”. In: ICML. 2020.

Scaffold[1]

$x^\star \in \arg\min_{x\in\mathbb{R}^d} \frac{1}{N}\sum_{c=1}^N \mathbb{E}_Z[F_c(x;Z)]$

At each global iteration

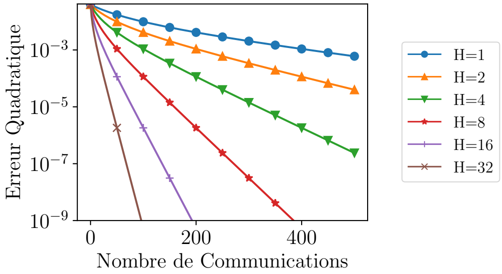
- For $c = 1$ to N in parallel
  - Receive $x^{(t)}$, set $x_c^{(t,0)} = x^{(t)}$
  - For $h = 0$ to $H-1$

$$x_c^{(t,h+1)} = x_c^{(t,h)} - \gamma\big(\nabla F_c(x_c^{(t,h)}; Z_c^{(t,h+1)}) + \boldsymbol{\xi_c^{(t)}}\big)$$

- Aggregate models, update control variates

$$x^{(t+1)} = \frac{1}{N}\sum_{c=1}^N x_c^{(t,H)}$$

$$\boldsymbol{\xi_c^{(t+1)} = \xi_c^{(t)} + \frac{1}{\gamma H}(\theta_c^{t,H} - \theta^{(t+1)})}$$



$\rightarrow$ No more heterogeneity bias!

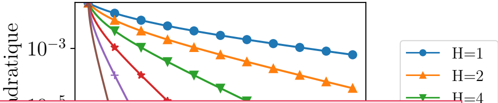
[1] S. P. Karimireddy et al. "Scaffold: Stochastic controlled averaging for federated learning". In: **ICML**. 2020.

Scaffold[1]

$x^\star \in \arg\min_{x\in\mathbb{R}^d} \frac{1}{N}\sum_{c=1}^{N} \mathbb{E}_Z[F_c(x;Z)]$

At each global iteration

- For $c = 1$ to N in parallel
  - Receive $x^{(t)}$, set $x_c^{(t,0)} = x^{(t)}$



Open Question:

Does linear speed-up remain with control variates?

$\xi_c^{(t+1)} = \xi_c^{(t)} + \frac{1}{\gamma H}(\theta_c^{t,H} - \theta^{(t+1)})$

$\rightarrow$ No more heterogeneity bias!

[1] S. P. Karimireddy et al. "Scaffold: Stochastic controlled averaging for federated learning". In: **ICML**. 2020.

Scaffold also converges ![1]

(For L-smooth, μ-strongly convex functions with $\nabla^3 f(x)$ bounded by Q)

- Scaffold converges if $\gamma HL \leq 1$, towards a distribution $\pi^{(\gamma,H)}$
  - denoting $x^{(t)} \sim \psi_{x^{(t)}}$, we have

$$\mathcal{W}_2(\psi_{x^{(t)}}; \pi^{(\gamma,H)}) \leq (1-\gamma\mu)^{Ht}\mathcal{W}_2(\psi_{x^{(0)}}; \pi^{(\gamma,H)})$$

  - where $\mathcal{W}_2$ is the second order Wasserstein distance

[1] P. Mangold et al. "Scaffold with Stochastic Gradients: New Analysis with Linear Speed-Up". In: ICML. 2025.

Scaffold also converges ![1]

(For L-smooth, μ-strongly convex functions with $\nabla^3 f(x)$ bounded by Q)

- Scaffold converges if $\gamma HL \leq 1$, towards a distribution $\pi^{(\gamma,H)}$
- Scaffold's variance is close to FedAvg's variance

$$\int (x - x^\star)(x - x^\star)^\top \pi^{(\gamma,H)}(\mathrm{d}x) = \boxed{\frac{\gamma}{N} C(x^\star)} + O(\gamma^{3/2})$$

[1] P. Mangold et al. "Scaffold with Stochastic Gradients: New Analysis with Linear Speed-Up". In: **ICML**. 2025.

Scaffold also converges ![1]

(For L-smoo ns with $\nabla^3 f(x)$ bounded by Q)

Linear speed-up !

variance decreases in $1/N$
variance scales in γ

- Scaffold conver listribution $\pi^{(\gamma,H)}$
- Scaffold's variance is close to FedAvg's variance

$$\int (x - x^\star)(x - x^\star)^\top \pi^{(\gamma,H)}(\mathrm{d}x) = \boxed{\frac{\gamma}{N} C(x^\star)} + O(\gamma^{3/2})$$

[1] P. Mangold et al. "Scaffold with Stochastic Gradients: New Analysis with Linear Speed-Up". In: ICML. 2025.

Scaffold also converges ![1]

(For L-smooth, μ-strongly convex functions with $\nabla^3 f(x)$ bounded by Q)

- Scaffold converges if $\gamma HL \leq 1$, towards a distribution $\pi^{(\gamma,H)}$
- Scaffold's variance is close to FedAvg's variance
- Scaffold still has some bias

$$\int x\pi^{(\gamma,H)}(\mathrm{d}x) = x^\star - \boxed{\frac{\gamma}{2N}\nabla^2 f(x^\star)^{-1}\nabla^3 f(x^\star)A^{-1}C(x^\star)} + O(\gamma^{3/2})$$

[1] P. Mangold et al. "Scaffold with Stochastic Gradients: New Analysis with Linear Speed-Up". In: **ICML**. 2025.

Scaffold also converges ![1]

(For L-smooth, μ-strongly convex function

Stochasticity bias remains

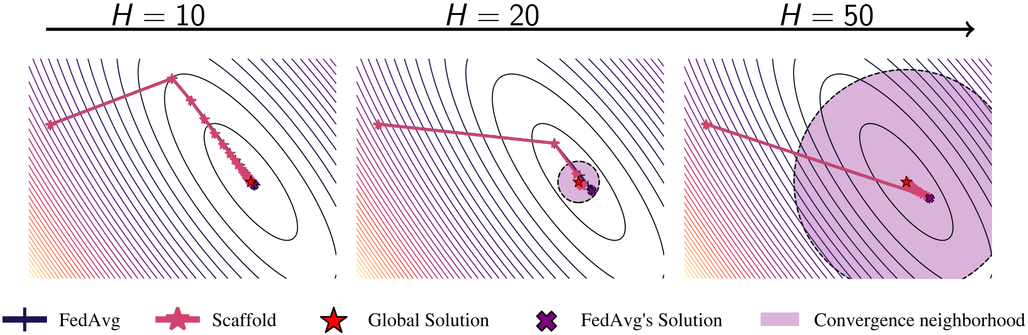
$$A = I \otimes \nabla^2 f(x^\star) + \nabla^2 f(x^\star) \otimes I$$

$C(x^\star)$ is ∇F^Z's covariance at $x^\star$

- Scaffold converges if $\gamma HL \leq 1$, towards a di
- Scaffold's variance is close to FedAvg's variance
- Scaffold still has some bias

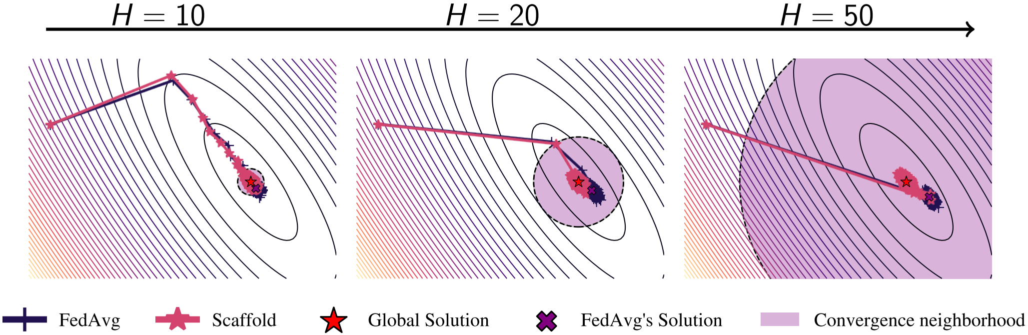
$$\int x \pi^{(\gamma,H)}(\mathrm{d}x) = x^\star - \boxed{\frac{\gamma}{2N} \nabla^2 f(x^\star)^{-1} \nabla^3 f(x^\star) A^{-1} C(x^\star)} + O(\gamma^{3/2})$$

[1] P. Mangold et al. "Scaffold with Stochastic Gradients: New Analysis with Linear Speed-Up". In: **ICML**. 2025.



Scaffold converges to the right point

... and its variance is similar to FedAvg!



Scaffold converges to the right point

... and its variance is similar to FedAvg!

New Convergence Rate for Scaffold

(For L-smooth, μ-strongly convex functions with $\nabla^3 f(x)$ bounded by Q)

$$\mathbb{E}\left[\|x^{(T)} - x^\star\|^2\right] \lesssim \left(1 - \frac{\gamma\mu}{4}\right)^{HT} \left\{\|x^{(0)} - x^\star\|^2 + 2\gamma^2 H^2 \zeta^2 + \frac{\sigma_\star^2}{L\mu}\right\} + \frac{\gamma}{\boldsymbol{N}\mu}\sigma_\star^2 + \frac{\gamma^{3/2} Q}{\mu^{5/2}}\sigma_\star^3 + \frac{\gamma^3 H Q^2}{\mu^3}\sigma_\star^4$$

where

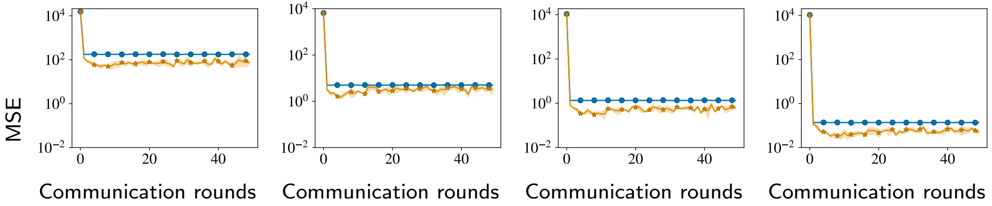
- $\sigma_\star^2 = \mathbb{E}[\frac{1}{N}\sum_{c=1}^N \|\nabla F_c^Z(x^\star) - \nabla f_c(x^\star)\|^2$ is the variance at $x^\star$
- $\zeta^2 = \frac{1}{N}\sum_{c=1}^N \|\nabla f_c^Z(x^\star)\|^2$ measures gradient heterogeneity

Linear Speed-Up!

As long as N is not too large, one can obtain $\mathbb{E}\left[\|x^{(T)} - x^\star\|^2\right] \leq \epsilon^2$ with

$$\#\text{grad per client} = \widetilde{O}\Big(\frac{\sigma_\star^2}{\boldsymbol{N}\mu^2\epsilon^2}\log\left(\frac{1}{\epsilon}\right)\Big)$$

Numerical Illustration: Speed-Up of Scaffold



(a) $N = 10$ (b) $N = 100$ (c) $N = 1000$ (d) $N = 10000$

Blue: FedAvg, Orange: Scaffold

Conclusion

- FedAvg and Scaffold converge (even with stochastic gradients)
- This allows to derive new analyses for these problems, with exact first-order expression for bias
- And we proved that Scaffold has:
  - variance similar to FedAvg's variance
  - *linear speed-up* in the number of clients!!

  But... Scaffold is still biased: some good directions for future. :)

Thank you!

Papers related to this presentation:

- P. Mangold et al. "Refined Analysis of Federated Averaging's Bias and Federated Richardson-Romberg Extrapolation". In: **AISTATS**. 2025
- P. Mangold et al. "Scaffold with Stochastic Gradients: New Analysis with Linear Speed-Up". In: **ICML**. 2025