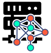
Convergence and Linear Speed-Up in Stochastic Federated Learning

Paul Mangold (CMAP, École polytechnique)

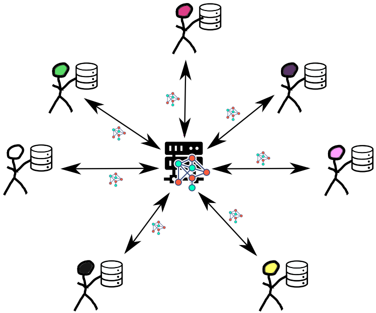
Workshop “Fondements Mathématiques de l’IA”

March 25th, 2025

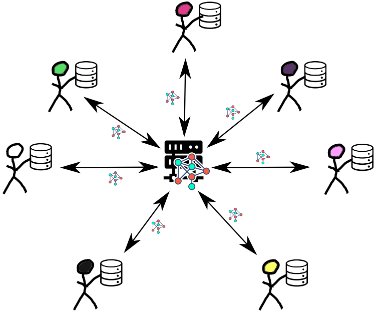
Federated Learning



Federated Learning



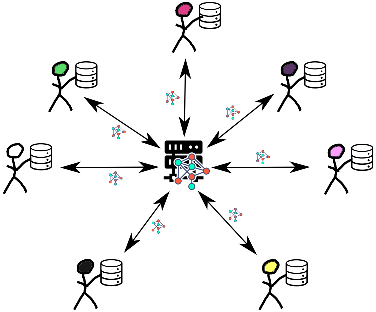
Federated Learning



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

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

Problem: data is heterogeneous, communication is expensive

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

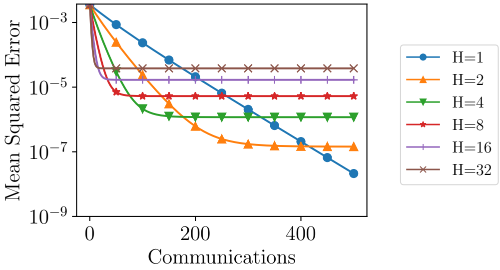
- For $c = 1$ à N in parallel
  - Receive $x^{(t)}$, set $x_c^{(t,0)} = x^{(t)}$
  - For $h = 0$ to $H-1$

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

- 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

- 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* (2022).

[3] J. Wang et al. "On the Unreasonable Effectiveness of Federated Averaging with Heterogeneous Data". In: *TMLR* (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]: $\zeta = \left\| \frac{1}{NH}\sum_{c=1}^{N}\sum_{h=0}^{H-1} \nabla f(x_c^{(h)}) - \nabla f(x^\star) \right\|^2$

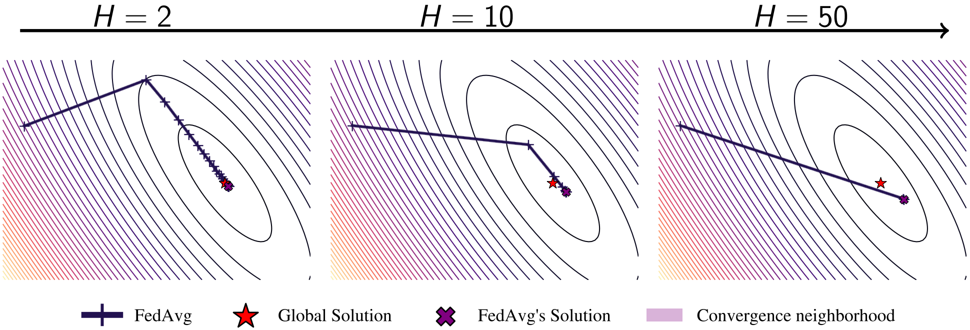
Show **convergence to a neighborhood** of $x^\star$

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

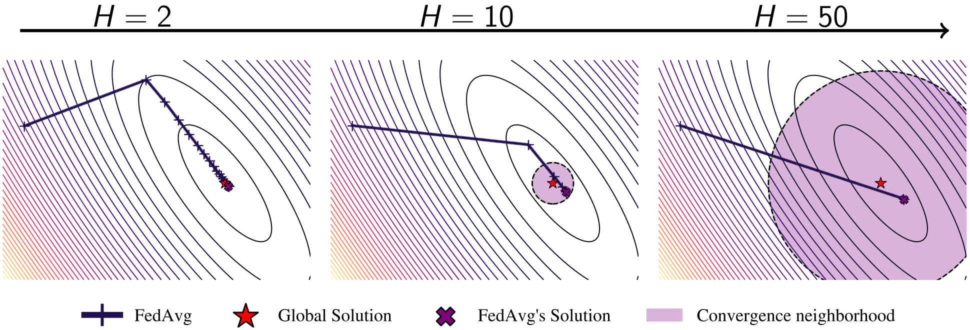
[1] X. Lian et al. "Can decentralized algorithms outperform centralized algorithms? a case study for decentralized parallel SGD". In: *NeurIPS* (2017).

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

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

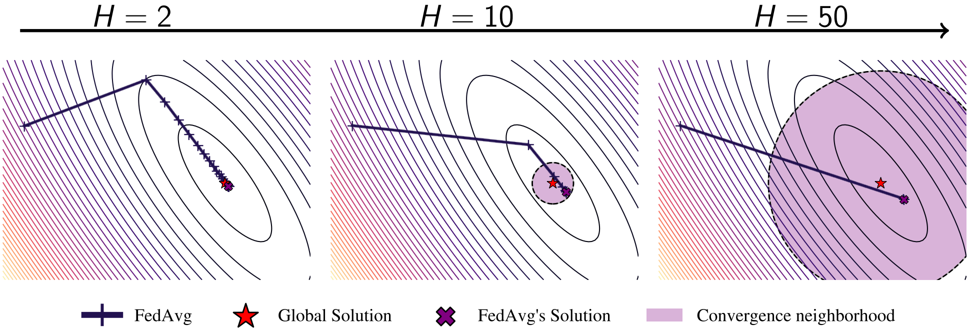


When the number of local iterations increases, bias incrases



When the number of local iterations increases, bias incrases

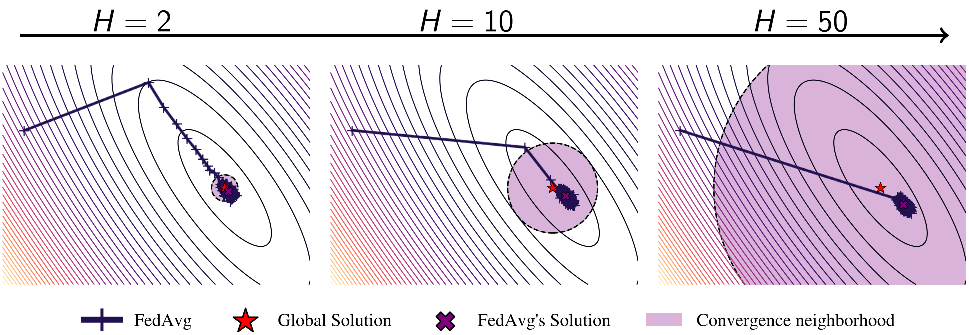
... but the bound is oblivious to problem's geometry



When the number of local iterations increases, bias incrases

... but the bound is oblivious to problem's geometry

Remark: It seems that iterates converge in some way?



When the number of local iterations increases, bias incrases

... but the bound is oblivious to problem's geometry

Remark: It seems that iterates converge in some way?

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 (radients) converges![1]

(For t trongly convex functions)

Linear speed-up !

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

- FedAvg conver on $\pi^{(\gamma,H)}$
- 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) converges![1]

Heterogeneity bias

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

th, μ-s

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 $x^\star$

- tributi
- FedAvg's iterates covariance is
- We can now give an **exact expansion of the bias**

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

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

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

II. Correcting heterogeneity: 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: **International conference on machine learning**. PMLR. 2020, pp. 5132–5143.

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

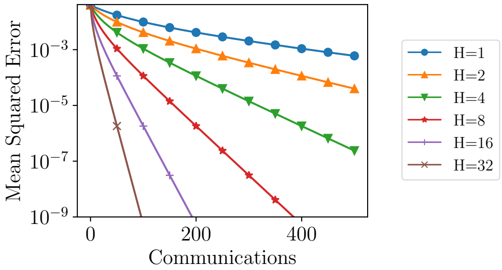
- For $c = 1$ to N in parallel
  - Receive $x^{(t)}$, set $x_c^{(t,0)} = x^{(t)}$
  - For $h = 0$ to $H-1$

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

- 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: **International conference on machine learning**. PMLR. 2020, pp. 5132–5143.

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: **arxiv preprint**. 2025.

Scaffold also converges ![1]

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

- Scaffold converges if $\gamma H L \leq 1$, 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: **arxiv preprint**. 2025.

Scaffold also converges ![1]

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

Linear speed-up !

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

- Scaffold conve ... istribution $\pi^{(\gamma,H)}$
- Scaffold's varia ... ance

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

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

Scaffold also converges ![1]

(For L-smooth, μ-strongly convex functio

Stochasticity bias remains!

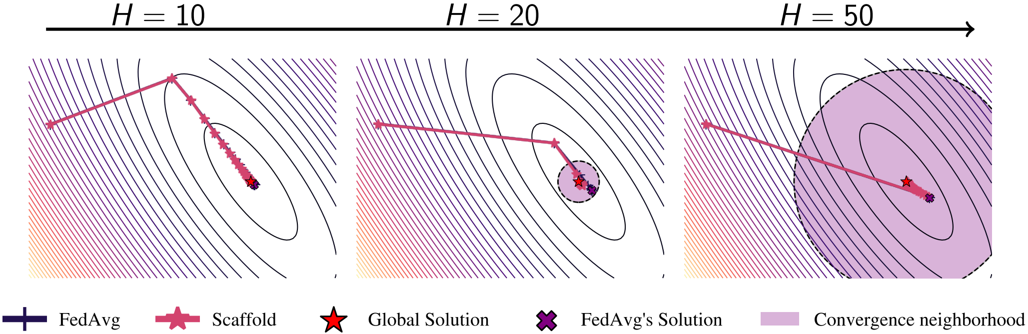
$$A = I \otimes \nabla^2 f(x^\star) + \nabla^2 f(x^\star) \otimes I$$

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

- Scaffold converges if $\gamma H L \leq 1$, towards a d
- Scaffold's variance is close to FedAvg's vari
- Scaffold still has some bias

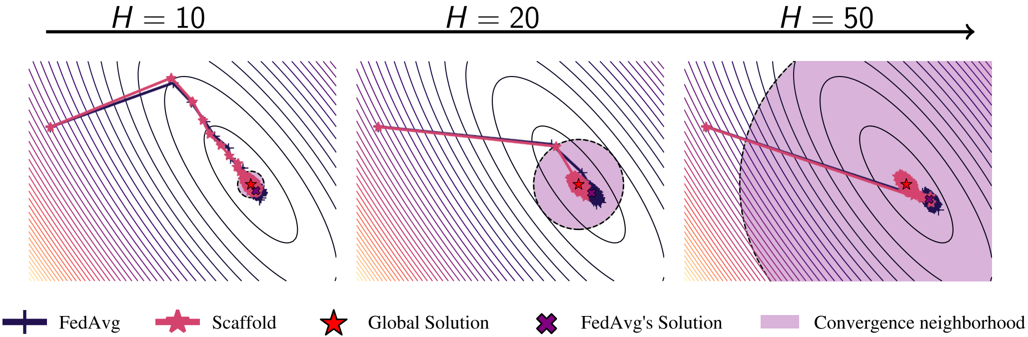
$$\int x \pi^{(\gamma, H)}(\mathrm{d}x) = x^\star - \boxed{\frac{\gamma}{2N} \nabla^2 f(x^\star)^{-1} \nabla^3 f(x^\star) A^{-1} C(x^\star)} + O(\gamma^{3/2})$$

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

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

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!!

Thank you!

Check the papers:

- 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: **arxiv preprint**. 2025

Find this presentation on my website:

- `https://pmangold.fr/research.php?page=talks`