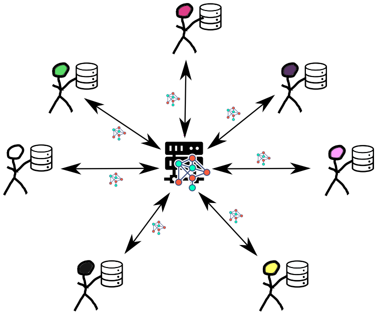
Convergence and Linear Speed-Up in Stochastic Federated Learning

Paul Mangold (CMAP, École polytechnique)

ICSP 2025 – Mini-Symposium
Communication-efficient methods for distributed optimization and federated learning

July 28th, 2025

Federated Learning



Collaborative optimization problem

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

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

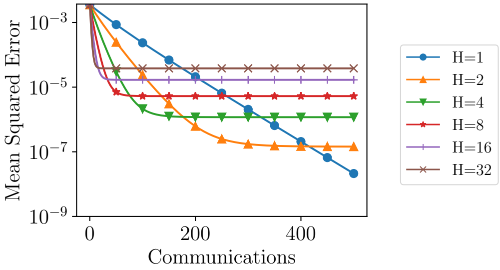
- For $c = 1$ à N in parallel
  - Receive $x^{(t)}$, set $x_c^{(t,0)} = x^{(t)}$
  - For $h = 0$ to $H-1$

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

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

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

[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]: $\zeta = \left\|\frac{1}{NH}\sum_{c=1}^{N}\sum_{h=0}^{H-1} \nabla f(x_c^{(h)}) - \nabla f(x^\star)\right\|^2$

[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 $x^\star$

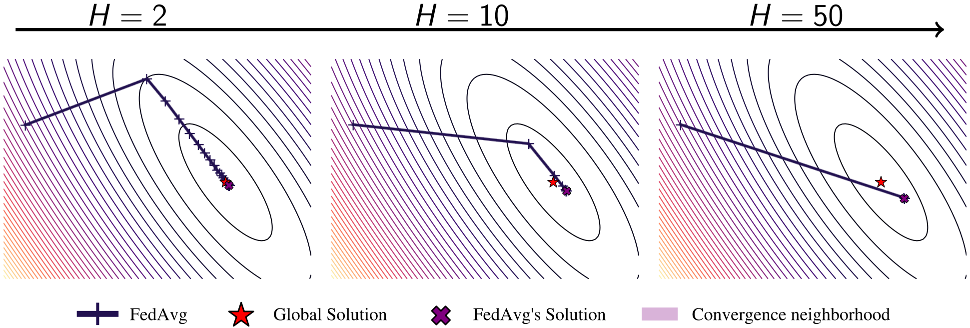
$$\|x^{(T)} - x^\star\|^2 \lesssim (1-\gamma\mu)^{HT}\|x^{(0)} - x^\star\|^2 + \boldsymbol{\chi(\gamma, H, \zeta)} \qquad \text{(for some function } \chi)$$

[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).

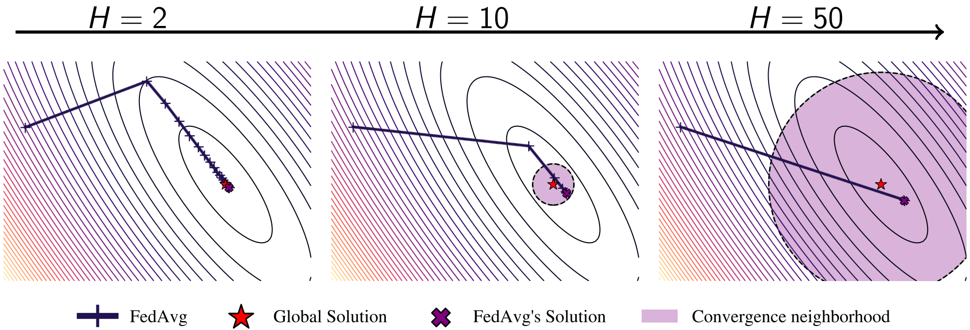
[3] J. Wang et al. "On the Unreasonable Effectiveness of Federated Averaging with Heterogeneous Data". In: *TMLR* (2024).

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



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, starting with $x_c^{(t)}, y_c^{(t)} \in \mathbb{R}^d$,

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

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, starting with $x_c^{(t)}, y_c^{(t)} \in \mathbb{R}^d$,

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

Thus

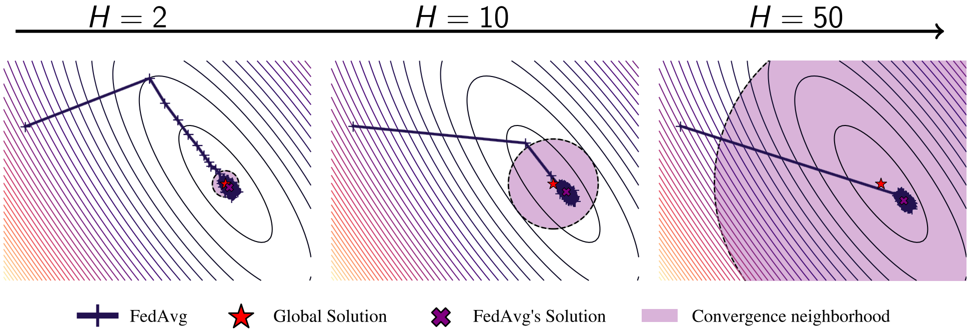
$$\|x_c^{(t+1)} - y_c^{(t+1)}\| \leq (1 - \gamma\mu)^H \|x_c^{(t)} - y_c^{(t)}\|$$

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

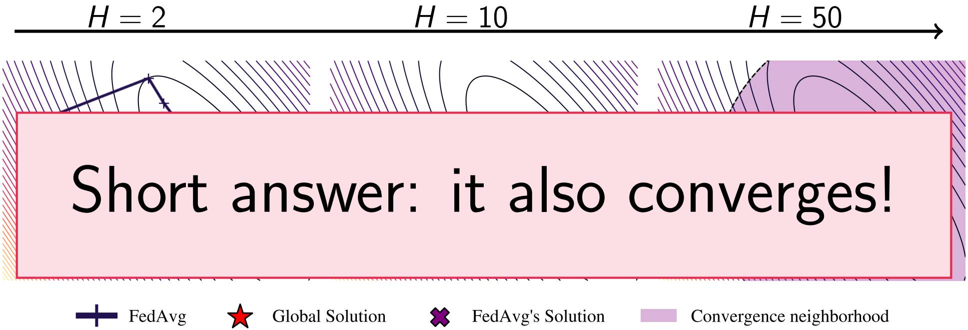
[1] G. Malinovskiy et al. "From local SGD to local fixed-point methods for federated learning". In: **ICML**. 2020.

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 Constant Step Size Federated Averaging 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

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

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

FedAvg (with stochastic gradients) converges![1]

(For t… trongly convex functions)

Linear speed-up !

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

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

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

[1] P. Mangold et al. "Refined Analysis of Constant Step Size Federated Averaging 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**

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

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

[1] P. Mangold et al. "Refined Analysis of Constant Step Size Federated Averaging 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 $\nabla f_c(x^\star) = \nabla f(x^\star)$

Stochasticity bias

$\boldsymbol{A} = (I \otimes \nabla^2 f(x^\star) + \nabla^2 f(x^\star) \otimes I)^{-1}$
$C(x^\star)$ is ∇F^Z's covariance at $x^\star$

- We can now give an **exact expansion of the bias**

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

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

[1] P. Mangold et al. "Refined Analysis of Constant Step Size Federated Averaging 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

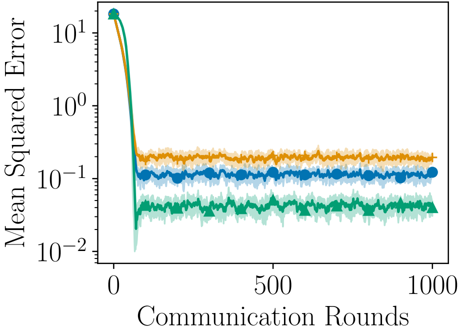
We can co

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

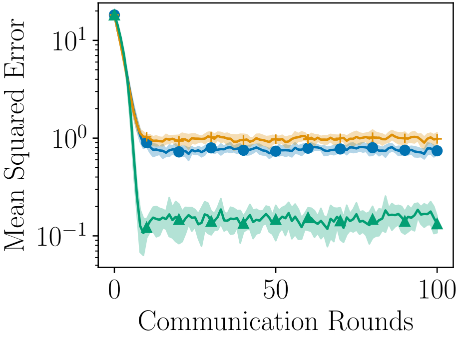
$\rightarrow$ bias is effectively reduced!!

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

Numerical Illustration: FedAvg



(a) $H = 10$



(b) $H = 100$

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

II. Correcting heterogeneity: Scaffold

Scaffold[1]

(*without global step size)

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

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

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

$$\boldsymbol{\xi_c^{(t+1)} = \xi_c^{(t)} + \frac{1}{\gamma H}\left(x_c^{t,H} - x^{(t+1)}\right)}$$

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

Scaffold[1]
(*without global step size)

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

At each global iteration

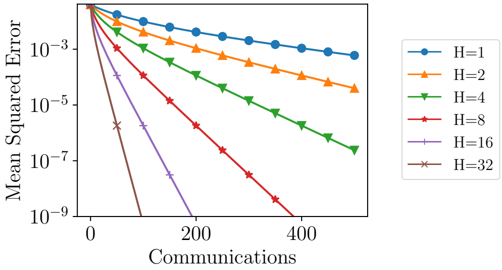
- For $c = 1$ to N in parallel
  - Receive $x^{(t)}$, set $x_c^{(t,0)} = x^{(t)}$
  - For $h = 0$ to $H-1$

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

- Aggregate models, update control variates

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

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



$\rightarrow$ 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 $(x^{(t)}, \xi^{(t)}_{1:N}) \sim \psi_{(x^{(t)},\xi^{(t)}_{1:N})}$, we have

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

  - 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

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

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

Scaffold also converges ![1]

(For L-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 variance is close to FedAvg's variance

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

[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 **removes heterogeneity bias**

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

$\Rightarrow$ **but it is still biased**

[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 funct

Stochasticity bias remains

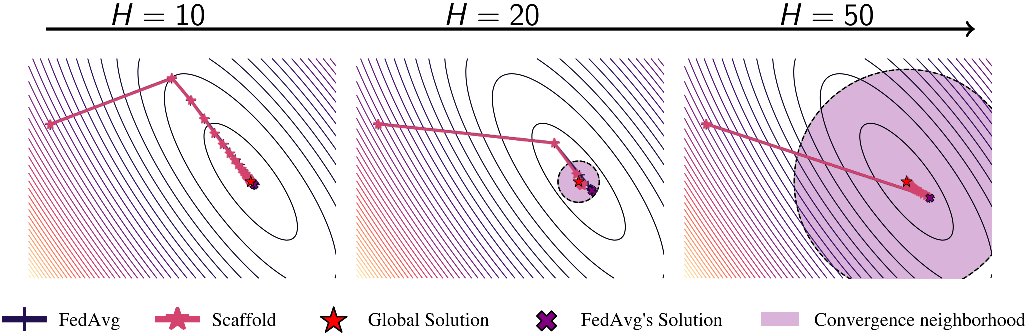
$A = (I \otimes \nabla^2 f(x^\star) + \nabla^2 f(x^\star) \otimes I)^{-1}$
$C(x^\star)$ is ∇F^Z's covariance at $x^\star$

- Scaffold converges if $\gamma HL \leq 1$, towards a
- Scaffold's variance is close to FedAvg's v
- Scaffold **removes heterogeneity bias**

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

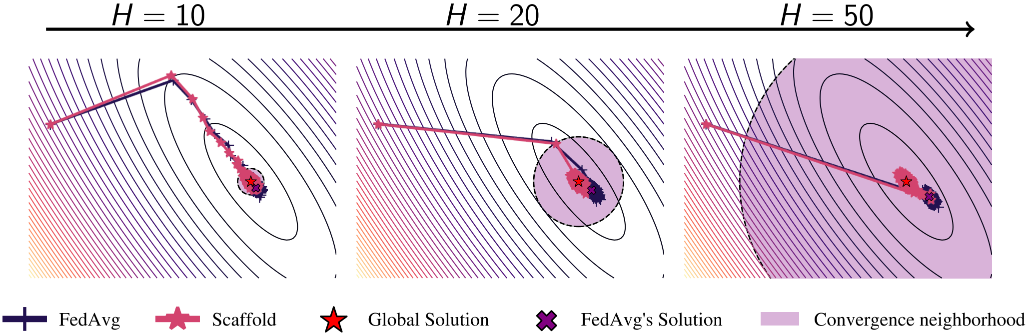
$\Rightarrow$ **but it is still biased**

[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!

Bounding the Covariance

Define covariance matrices

$$\bar{\Sigma}^{x} \triangleq \int \left(x - x_{\star}\right)^{\otimes 2} \pi^{(\gamma,H)}(\mathrm{d}x, \mathrm{d}\Xi)$$

$$\bar{\Sigma}^{\xi}_{(c,c')} \triangleq \int \left(\xi_c - \xi_c^{\star}\right)\left(\xi_{c'} - \xi_c^{\star}\right)^{\top} \pi^{(\gamma,H)}(\mathrm{d}x, \mathrm{d}\Xi)$$

$$\bar{\Sigma}^{x,\xi}_{(c)} \triangleq \int \left(x - x_{\star}\right)\left(\xi_c - \xi_c^{\star}\right)^{\top} \pi^{(\gamma,H)}(\mathrm{d}x, \mathrm{d}\Xi)$$

Expansion of Covariance

$$\bar{\Sigma}^x = \tfrac{\gamma}{N}\boldsymbol{A}\mathcal{C}(x_\star) + O(\gamma^2 H + \gamma^{3/2})$$

$$\bar{\Sigma}^{x,\xi}_{(c)} = \tfrac{\gamma}{N}\boldsymbol{A}\mathcal{C}(x_\star)(\nabla^2 f_c(c)x_\star - \nabla^2 f(x_\star)) + \tfrac{\gamma}{N}(\mathcal{C}_c(x_\star) - \mathcal{C}(x_\star)) + O(\gamma^2 H + \gamma^{3/2})$$

$$\bar{\Sigma}^{\xi}_{(c,c)} = (1 - \tfrac{2}{N})\tfrac{1}{H}\mathcal{C}_c(x_\star) + \tfrac{1}{NH}\mathcal{C}(x_\star) + O(\gamma)$$

$$\bar{\Sigma}^{\xi}_{(c,c')} = \tfrac{1}{NH}(\mathcal{C}(x_\star) - \mathcal{C}_c(x_\star) - \mathcal{C}_{c'}(x_\star)) + O(\gamma)$$

where

$$\boldsymbol{A} \triangleq (Id \otimes \nabla^2 f(x_\star) + \nabla^2 f(x_\star) \otimes Id)^{-1}$$

$$\mathcal{C}_c(x_\star) \triangleq \mathbb{E}\Big[\big(\nabla F_c^{Z_c}(x_\star) - \nabla f_c(x_\star)\big)^{\otimes 2}\Big] \quad \mathcal{C}(x_\star) \triangleq \frac{1}{N}\sum_{c=1}^{N}\mathcal{C}_c(x_\star)$$

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!!
- But: Scaffold is still biased
  $\Rightarrow$ **Need for algorithms tailored for FL <u>and</u> stochasticity!**

Thank you!

Check the papers:

- P. Mangold et al. "Refined Analysis of Constant Step Size Federated Averaging 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

Find this presentation on my website:

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