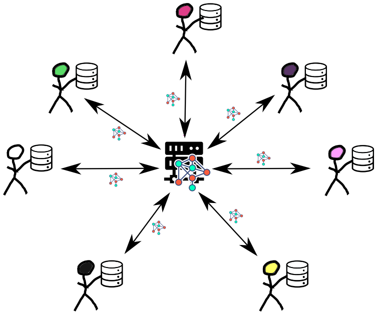
Analyse Raffinée de Federated Averaging et Extrapolation de Richardson-Romberg Fédérée

Paul Mangold (CMAP, École Polytechnique)

Journées de Statistique de la SFdS

Mercredi 4 juin 2025

Federated Learning



Collaborative optimization problem

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

Central Challenges: data and computational heterogeneity

+ slow and difficult-to-establish communication

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

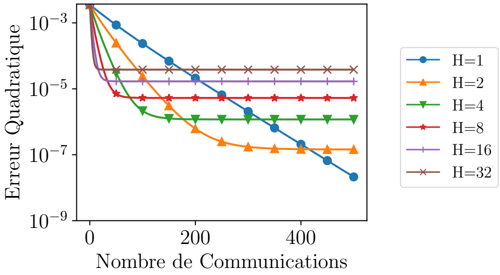
- For $c = 1$ à N in parallel
  - Receive $x^{(t)}$, set $x_c^{(t,0)} = x^{(t)}$
  - For $h = 0$ to $H-1$

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

- Aggregate local models

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

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 $x^\star$

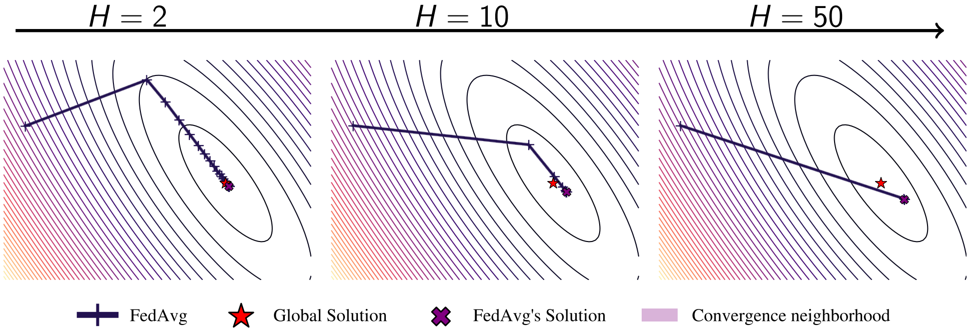
$$\|x^{(T)} - x^\star\|^2 \lesssim (1-\gamma\mu)^{HT}\|x^{(0)} - x^\star\|^2 + \boldsymbol{\chi(\gamma, H, \zeta)} \qquad \text{(for some function } \chi)$$

[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).

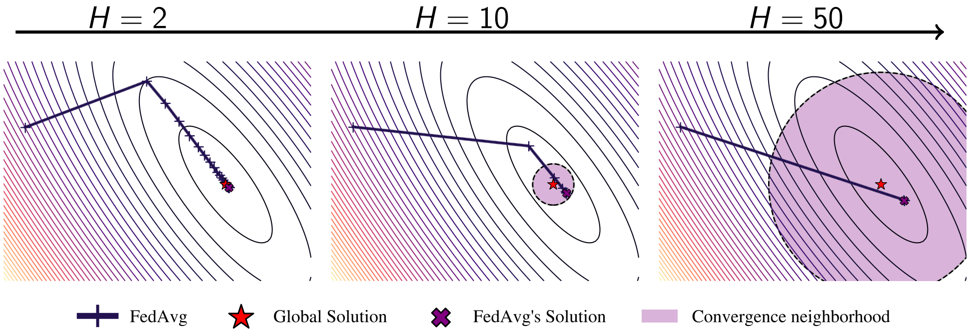
[3] J. Wang et al. "On the Unreasonable Effectiveness of Federated Averaging with Heterogeneous Data". In: *TMLR* 2024 (2024).

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



When the number of local iterations increases, bias incrases

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



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

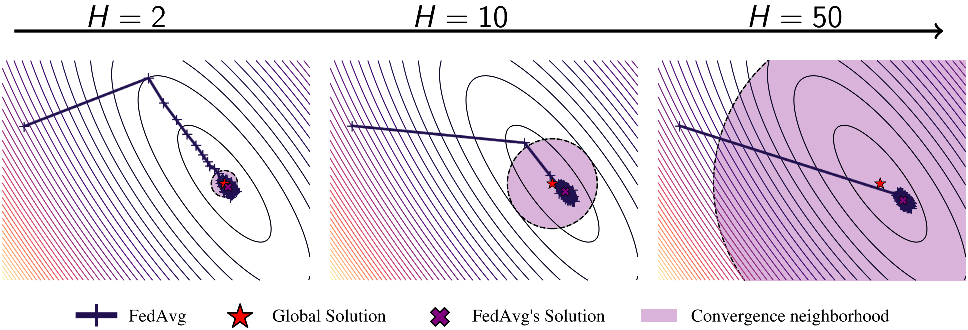
$$\|x_c^{(t+1)} - y_c^{(t+1)}\| \leq (1 - \gamma\mu)^H \|x_c^{(t)} - y_c^{(t)}\|$$

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

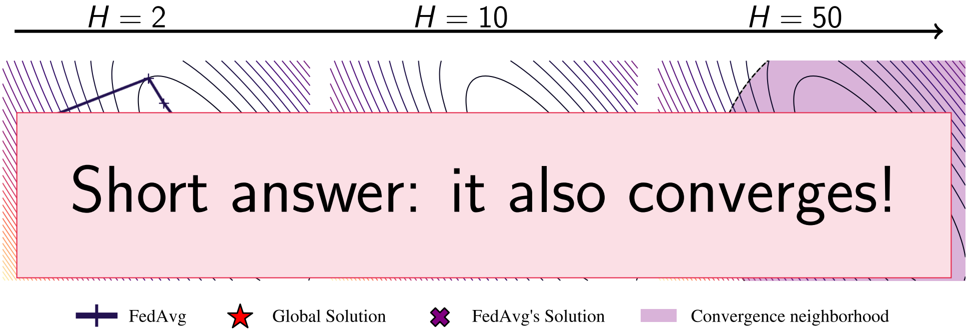
[1] G. Malinovskiy et al. "From local SGD to local fixed-point methods for federated learning". In: **ICML**. 2020.

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 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.

...ic gr... [1]

...oth, μ-st...

Heterogeneity bias

vanishes when $\nabla^2 f_c(x^\star) = \nabla^2 f(x^\star)$
or when $\nabla f_c(x^\star) = \nabla f(x^\star)$
or when $H = 1$ (one local update)

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 $x^\star$

- ...stribution $\pi^{(\gamma, H)}$...
- 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

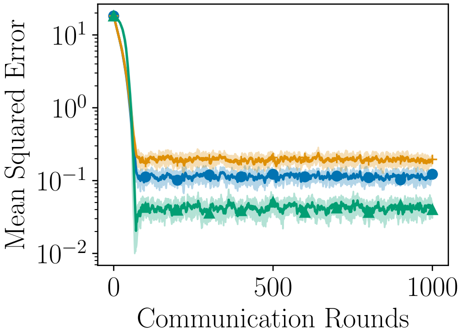
We can co

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

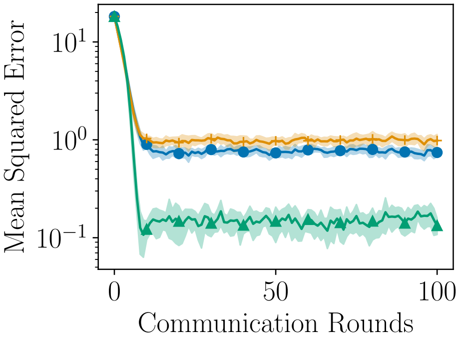
$\rightarrow$ bias is effectively reduced!!

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

Numerical Illustration: FedAvg



(a) $H = 10$



(b) $H = 100$

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

Conclusion

- FedAvg converges (even with stochastic gradients)
- This allows to derive new analyses for these problems, with exact first-order expression for bias
- FedAvg (and friends) are still biased!
  - there is still a lot to do in federated optimization
  - ... especially when gradients are stochastic,
    ... which is the most interesting setting!
- Similar results hold for Scaffold (see next slide)

PS: Extension to Scaffold

Little teaser :)

Similar results hold for Scaffold, and we prove that:

- Scaffold's iterates converge
- Scaffold eliminates heterogeneity bias *but not stochasticity bias*
- new convergence rate for Scaffold:

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

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