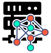
Convergence and Linear Speed-Up in Stochastic Federated Learning

Paul Mangold

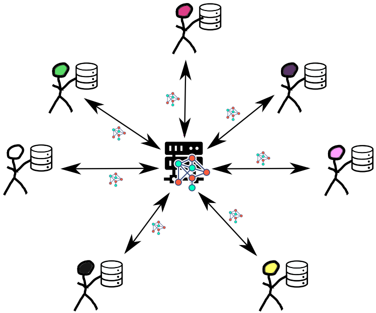
Workshop Fondation Mathématiques de l'IA

March 25th, 2025

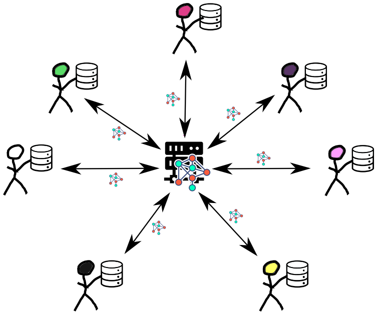
Optimisation fédérée



Optimisation fédérée



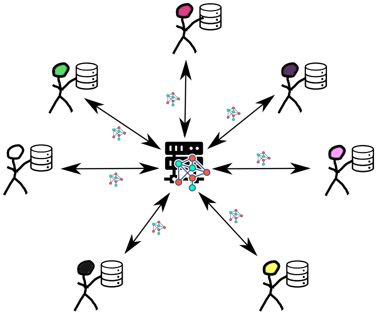
Optimisation fédérée



Optimisation collaborative

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

Optimisation fédérée



Optimisation collaborative

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

Difficultés centrales : hétérogénéité des données et des moyens de calcul

$+$ communication lente et difficile à établir

Optimisation fédérée

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

Federated Averaging[1] (FedAvg)

À chaque itération globale :

- Pour $c = 1$ à N en parallèle
  - Recevoir $x^{(t)}$, initialiser $x_c^{(t,0)} = x^{(t)}$
  - Pour $h = 0$ à $H-1$

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

- Agrégation des modèles

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

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

Optimisation fédérée

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

Federated Averaging[1] (FedAvg)

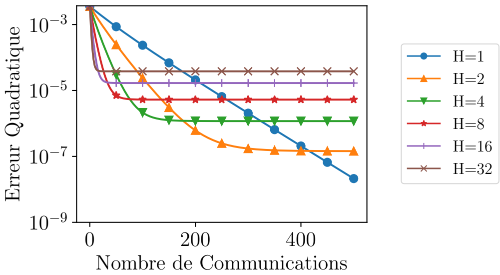
À chaque itération globale :

- Pour $c = 1$ à N en parallèle
  - Recevoir $x^{(t)}$, initialiser $x_c^{(t,0)} = x^{(t)}$
  - Pour $h = 0$ à $H - 1$

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

- Agrégation des modèles

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



Plus d'itérations locales

✓ convergence plus rapide

✗ biais plus grand

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

FedAvg avec gradients stochastiques converge ![1]

Si f_c trois fois dérivable, μ-fortement convexe, ∇f_c est L-Lipschitz, et $\gamma \leq 1/L$

- FedAvg converge en Wasserstein vers une distribution $\pi^{(\gamma,H)}$

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

FedAvg avec gradients stochastiques converge ![1]

Si f_c trois fois dérivable, μ-fortement convexe, ∇f_c est L-Lipschitz, et $\gamma \leq 1/L$

- FedAvg converge en Wasserstein vers une distribution $\pi^{(\gamma,H)}$
  - et si $x^{(t)} \sim \psi_{x^{(t)}}$,

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

  - où $\mathcal{W}_2$ est la distance de Wasserstein d'ordre 2

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

FedAvg avec gradients stochastiques converge ![1]

Si f_c trois fois dérivable, μ-fortement convexe, ∇f_c est L-Lipschitz, et $\gamma \leq 1/L$

- FedAvg converge en Wasserstein vers une distribution $\pi^{(\gamma,H)}$
- Biais de FedAvg (pour γ, H petits)

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

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

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

FedAvg avec gradients stochastiques converge ![1]

Biais d'hétérogénéité

Disparaît quand $\nabla^2 f_c(x^\star) = \nabla^2 f(x^\star)$
ou quand $\nabla f_c(x^\star) = \nabla f(x^\star)$

exe,

une

Biais de stochasticité

A est un opérateur linéaire
$C(x^\star)$ est la covariance de ∇f en $x^\star$

- Biais de FedAvg (pour γ, H petits)

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

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

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

FedAvg avec gradients stochastiques converge ![1]

Si f_c trois fois dérivable, μ-fortement convexe, ∇f_c est L-Lipschitz, et $\gamma \leq 1/L$

- FedAvg converge en Wasserstein vers une distribution $\pi^{(\gamma,H)}$
- Biais de FedAvg (pour γ, H petits)
- Variance de FedAvg (pour γ, H petits)

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

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

FedAvg avec gradients stochastiques converge ![1]

Si f_c trois fois dérivable, μ-fortement convexe, ∇f_c est L-Lipschitz, et $\gamma \leq 1/L$

- FedAvg conver… istribution $\pi^{(\gamma,H)}$
- Biais de FedAv…
- Variance de FedAvg (pour γ, H petits)

Linear speed-up !

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

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

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

Scaffold

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

À chaque itération globale :

- Pour $c = 1$ à N en parallèle
  - Recevoir $x^{(t)}$, initialiser $x_c^{(t,0)} = x^{(t)}$
  - Pour $h = 0$ à $H-1$

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

- Agrégation des modèles

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

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

Scaffold

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

À chaque itération globale :

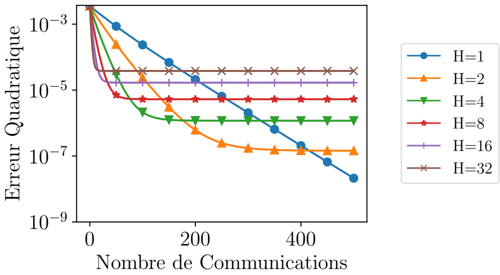
- Pour $c = 1$ à N en parallèle
  - Recevoir $x^{(t)}$, initialiser $x_c^{(t,0)} = x^{(t)}$
  - Pour $h = 0$ à $H-1$

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

- Agrégation des modèles

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

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



$\longrightarrow$ No more heterogeneity bias!

Scaffold also converges ![1]

Si f_c trois fois dérivable, μ-fortement convexe, ∇f_c est L-Lipschitz, et $\gamma \leq 1/L$

- Scaffold converges if $\gamma HL \leq 1$ en Wasserstein vers une distribution $\pi^{(\gamma,H)}$

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

Scaffold also converges ![1]

Si f_c trois fois dérivable, μ-fortement convexe, ∇f_c est L-Lipschitz, et $\gamma \leq 1/L$

- Scaffold converges if $\gamma H L \leq 1$ en Wasserstein vers une distribution $\pi^{(\gamma,H)}$
  - et si $x^{(t)} \sim \psi_{x^{(t)}}$,

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

  - où $\mathcal{W}_2$ est la distance de Wasserstein d'ordre 2

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

Scaffold also converges ![1]

Si f_c trois fois dérivable, μ-fortement convexe, ∇f_c est L-Lipschitz, et $\gamma \leq 1/L$

- Scaffold converges if $\gamma H L \leq 1$ en Wasserstein vers une distribution $\pi^{(\gamma,H)}$
- Biais de Scaffold (pour γ, H petits)

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

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

Scaffold also c

Biais de stochasticité

A est un opérateur linéaire
$C(x^\star)$ est la covariance de ∇f en $x^\star$

Si f_c trois fois dérivable, μ-fortement convexe, ∇f_c est L-Lipschitz, et $\gamma \leq 1/L$

- Scaffold converges if $\gamma HL \leq 1$ en Wasserstein vers une distribution $\pi^{(\gamma,H)}$
- Biais de Scaffold (pour γ, H petits)

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

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

Scaffold also converges ![1]

Si f_c trois fois dérivable, μ-fortement convexe, ∇f_c est L-Lipschitz, et $\gamma \leq 1/L$

- Scaffold converges if $\gamma HL \leq 1$ en Wasserstein vers une distribution $\pi^{(\gamma,H)}$
- Biais de Scaffold (pour γ, H petits)
- Variance de FedAvg (pour γ, H petits)

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

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

Scaffold also converges ![1]

Si f_c trois fois dérivable, μ-fortement convexe, ∇f_c est L-Lipschitz, et $\gamma \leq 1/L$

Linear speed-up !

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

- Scaffold conver… …ein vers une distribution $\pi^{(\gamma,H)}$
- Biais de Scaffo…
- Variance de FedAvg (pour γ, H petits)

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

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

New Convergence Rate for Scaffold

Linear Speed-Up!

Numerical Illustrations

Conclusion