Refined Analysis of Constant Step Size Federated Averaging and Federated Richardson-Romberg Extrapolation

Paul Mangold [1], Alain Durmus [1], Aymeric Dieuleveut [1], Sergey Samsonov [2], Eric Moulines [1,3]
[1]CMAP, Ecole polytechnique, France [2]HSE University, Russia [3]MBZUAI, UAE

Federated Learning

Take N functions $f_c(\theta) = \mathbb{E}[F_c^{Z_c}(\theta)]$, with $Z_c \sim \xi_c$

Goal: solve collaboratively

$$\theta^\star \in \arg\min_{\theta\in\mathbb{R}^d} f(\theta) = \frac{1}{N}\sum_{c=1}^{N} f_c(\theta)$$

(a) f_c is μ-strongly convex with $\mu > 0$, F_c^z if convex
(b) F_c^z is L-smooth and $\nabla^2 f_c(\theta) \preccurlyeq L\mathrm{Id}$
(c) the third derivative of f_c is uniformly bounded

Like SGD [1], FedAvg [2] is a Markov Chain

FedAvg's iterates are a Markov chain with kernel κ:

$$\kappa(\theta_t, \mathsf{B}) \triangleq \mathbb{E}[\mathbf{1}_{\mathsf{B}}(\theta_{t+1})|\theta_t] \text{ for } \mathsf{B} \in \mathcal{B}(\mathbb{R}^d)$$

Theorem: FedAvg's global iterates, with step γ and H local updates, converge to $\pi^{(\gamma,H)}$ with rate

$$\mathbf{W}_2^2(\rho\kappa^t, \pi^{(\gamma,H)}) \leq (1-\gamma\mu)^{Ht}\mathbf{W}_2^2(\rho, \pi^{(\gamma,H)})$$

Its expectation and covariance in the limit are

Limit point $\bar\theta_{\text{sto}}^{(\gamma,H)} \triangleq \int \vartheta \pi^{(\gamma,H)}(\mathrm{d}\vartheta)$

Covariance $\bar\Sigma_{\text{sto}}^{(\gamma,H)} \triangleq \int \{\vartheta - \theta^\star\}^{\otimes 2} \pi^{(\gamma,H)}(\mathrm{d}\vartheta)$

FedAvg as a Markov chain: it converges to a stationary distribution

We give an exact expression of its bias and a new method to reduce bias

FedAvg's bias : $$\bar\theta_{\text{sto}}^{(\gamma,H)} - \theta^\star = \underbrace{\frac{\gamma(H-1)}{2N}\sum_{c=1}^{N} \nabla^2 f(\theta^\star)^{-1}(\nabla^2 f_c(\theta^\star) - \nabla^2 f(\theta^\star))\nabla f_c(\theta^\star)}_{\text{Heterogeneity Bias}}$$
$$\underbrace{-\frac{\gamma}{2N}\nabla^2 f(\theta^\star)^{-1}\nabla^3 f(\theta^\star)\mathbf{A}\mathcal{C}(\theta^\star)}_{\text{Stochastic Bias}} + O(\gamma^2H^2 + \gamma^{3/2}H)$$

FedAvg's Linear Speed-Up

The variance of FedAvg at stationarity is:

$$\bar\Sigma_{\text{sto}}^{(\gamma,H)} = \frac{\gamma}{N}\mathbf{A}\mathcal{C}(\theta^\star) + O(\gamma^2H^2 + \gamma^{3/2}H)$$

where $$\mathbf{A} \triangleq (\mathrm{Id}\otimes\nabla^2 f(\theta^\star) + \nabla^2 f(\theta^\star)\otimes\mathrm{Id})^{-1}$$

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

New method: Federated Richardson-Romberg

New algorithm: run FedAvg twice with step sizes γ and 2γ and combine iterates

$$\vartheta_t^{(\gamma,H)} \triangleq 2\theta_t^{(\gamma,H)} - \theta_t^{(2\gamma,H)}$$

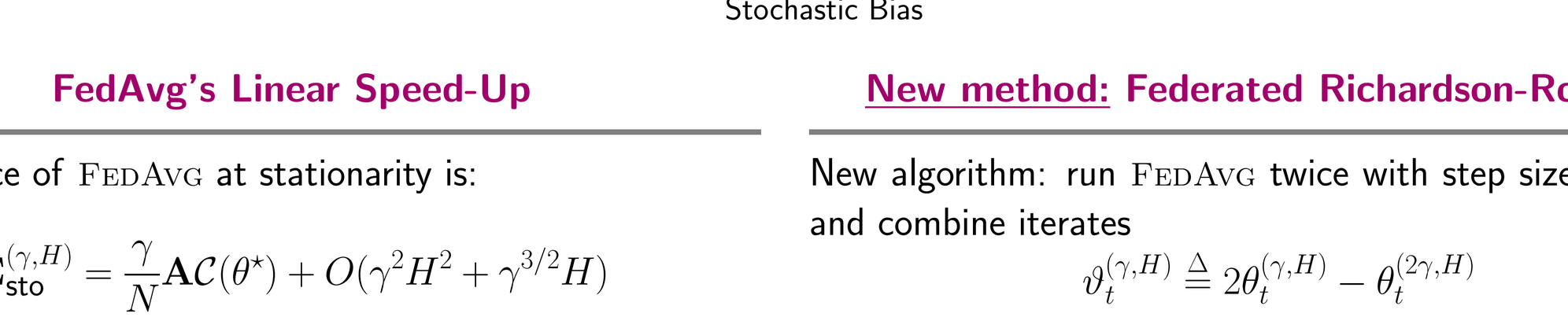
It converges to $\bar\vartheta_{\text{sto}}^{(\gamma,H)} = 2\bar\theta_{\text{sto}}^{(\gamma,H)} - \bar\theta_{\text{sto}}^{(2\gamma,H)}$ with bias

$$\bar\vartheta_{\text{sto}}^{(\gamma,H)} - \theta^\star = O(\gamma^2H^2 + \gamma^{3/2}H)$$

$\rightarrow$ reduces **both bias** without control variates

Numerical Illustration

Logistic regression: (a-b) homogeneous data with large gradient variance, (c-d) heterogeneous data with small variance
Blue = FedAvg [2], Green = FedAvg with Richardson-Romberg [ours], Orange = Scaffold [3]



(a) $H = 10$ (b) $H = 100$ (c) $H = 10$ (d) $H = 100$

References

[1] A. Dieuleveut, A. Durmus, and F. Bach. "Bridging the gap between constant step size stochastic gradient descent and Markov chains". In: *The Annals of Statistics* 48.3 (2020), pp. 1348–1382.
[2] B. McMahan et al. "Communication-efficient learning of deep networks from decentralized data". In: *Artificial intelligence and statistics*. PMLR. 2017, pp. 1273–1282.
[3] S. P. Karimireddy et al. "Scaffold: Stochastic controlled averaging for federated learning". In: *International conference on machine learning*. PMLR. 2020, pp. 5132–5143.