Streaming Federated Learning with Markovian Data

NeurIPS 2025

Tan Khiem Huynh [1] Malcolm Egan [1] Giovanni Neglia [2] Jean-Marie Gorce [1]

[1]Inria Lyon, France, [2]Inria Sophia Antipolis, France

Problems in the classical Federated Learning setting

- M clients collaborate to minimize the following loss function:

Definition (Objective function)

$$F(w) = \frac{1}{M} \sum_{m=1}^{M} F_m(w)$$

- Currently, we solve the above problem by assuming:
  - **ERM:** *data is sampled uniformly from a fix, pre-collected local dataset* $\mathcal{D}_m$.
  - **Stochastic Optimization:** *data is sampled I.I.D. from the local distribution* π_m *(usually unknown)*.

Problems in the classical Federated Learning setting

- What happens, if
  - Having a large pre-collected dataset is costly?
    - In IoT: sensors/edge devices with limited memory continually collect data to train their models.

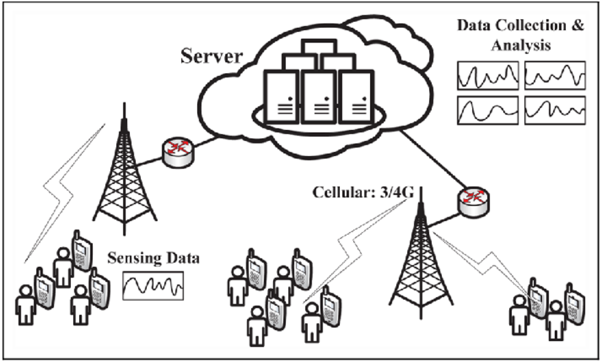


Figure: Crowdsensing

Problems in the classical Federated Learning setting

- What happens, if
  - Obtaining i.i.d. samples from the local distributions is hard / not possible ?
    - In Time-Series Analysis: temporal dependence between data samples.

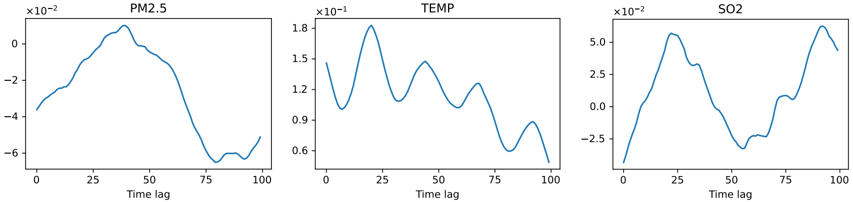


Figure: Auto-correlation of some air-quality measurements from the Beijing Multi-site Air-Quality dataset

Problems in the classical Federated Learning setting

- What happens, if
  - Having a large pre-collected dataset is costly?
    - In IoT: sensors/edge devices with limited memory continually collect data to train their models.
  - Obtaining i.i.d. samples from the local distributions is hard / not possible ?
    - In Time-Series Analysis: temporal dependence between data samples.

Question

What is the performance of FL algorithms when I.I.D or pre-collected data is not available ?

Non-I.I.D. Data Stream

- Observations in many real-world physical and biological systems are often modeled by **Markov processes**:

$$\mathbb{P}(x_i \mid x_{i-1}, \ldots, x_1) = \mathbb{P}(x_i \mid x_{i-1})$$

- Assumption: Each client has access to samples drawn from a Markov chain that **converges to the corresponding local distribution** π_{m}.
  - Speed of convergence: measured by *mixing time*.

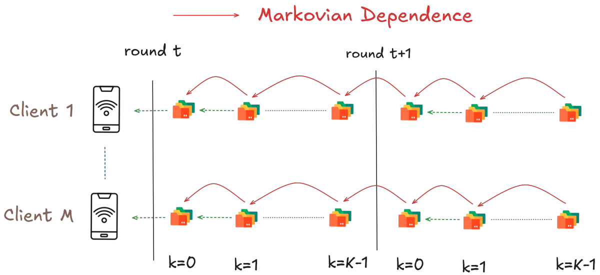


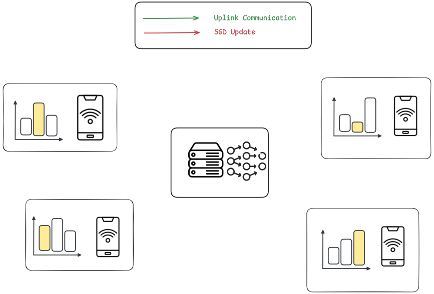
Figure: Streaming FL with Markovian Data

Main contributions

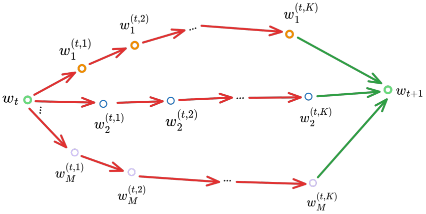
- We prove that:
  - FL with Markovian Data requires *as little communication as in the i.i.d. setting*.
  - *Collaboration in FL remains beneficial* under this regime.
- This has been shown in Federated Reinforcement Learning under more restricted setting:
  - *Linear Least-Square* loss.
  - *Stationary and fast-mixing* Markov chains.
- We focus on more general setting:
  - Non-convex loss.
  - No assumption on the speed of convergence for the underlying Markov chains.

Main contributions

- Two standard baselines:
  - Minibatch SGD: clients compute gradient at the same points $\rightarrow$ server perform SGD steps on the average gradients.
  - Local SGD: clients perform SGD on their local model $\rightarrow$ server averages the local models.
- Heterogeneity slows down Local SGD, as previously shown in the I.I.D. setting [Woodworth et al., 2020].



(a) Heterogeneity in FL



(b) Local SGD update scheme: clients' local model tend to converge toward their corresponding local optimum.

Main contributions

- Heterogeneity slows down Local SGD, as previously shown in the I.I.D. setting [Woodworth et al., 2020].
- We then propose a momentum-based variant of Local SGD that helps mitigating the heterogeneity effect.

Algorithm 1 Local SGD-M

Input: initial model w_0 and gradient estimate v_0, local learning rate η, global learning rate γ and momentum β
for $t = 0$ **to** $T-1$ **do**
  for every client $m \in [M]$ in parallel **do**
    Initialize local model $w_t^{(m,0)} = w_t$
    for $k = 0$ **to** $K-1$ **do**
      $v_t^{(m,k)} = \beta \nabla f_m \left(w_t^{(m,k)}; x_t^{(m,k)} \right) + (1-\beta) v_t$
      $w_t^{(m,k+1)} = w_t^{(m,k)} - \eta v_t^{(m,k)}$
    end for
    Communicate $w_t^{m,K}$
  end for
  Aggregate: $v_{t+1} = \frac{1}{\eta M K} \sum_{m=1}^{M} \left(w_t - w_t^{(m,K)} \right)$
  Server update: $w_{t+1} = w_t - \gamma v_{t+1}$
end for
Output: $\hat{w}_T$ sampled uniformly from $w_0, \ldots, w_{T-1}$.

Figure: Local SGD with Momentum

- Each local update: a convex combination of the current stochastic gradient and **the average of all local updates performed in the previous rounds.**

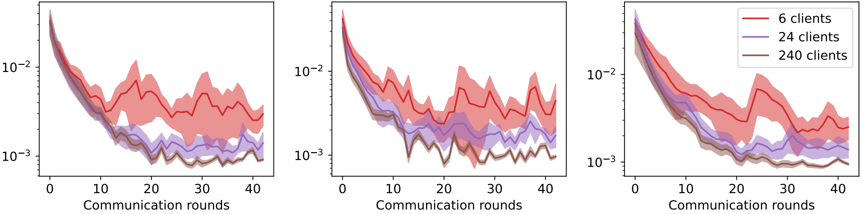
Results

	Communication complexity	Sample complexity per client
Minibatch SGD	$\frac{L\Delta_0}{\epsilon}$	$\frac{\tau\sigma^2 L\Delta_0}{M\epsilon^2}$
Local SGD	$\frac{L\Delta_0}{\epsilon}\max\left\{\delta^2, \frac{\sigma^2+\phi^2}{\epsilon}\right\}$	$\frac{\tau\sigma^2 L\Delta_0}{M\epsilon^2}\max\left\{\delta^2, \frac{\sigma^2+\phi^2}{\epsilon}\right\}$
Local SGD-M	$\frac{L\Delta_0}{\beta\epsilon}$	$\frac{\tau\sigma^2 L\Delta_0}{M\epsilon^2}$
Lower bound I.I.D. [Patel et al., 2022]	$\frac{L\Delta_0}{\epsilon}$	$\frac{\sigma L\Delta_0}{M\epsilon^{3/2}}$

Table: Communication & sample complexity of FL algorirthms with Markovian Data
τ: maximum mixing time of clients' Markov chains, δ^2, ϕ^2: heterogeneity constants, σ^2: noise in stochastic gradient estimates.

Experimental results

- Beijing Multi-Site Air-Quality dataset: hourly measurements of different air-quality indicators during 4 years.
  - Prediction target: seasonality-adjusted PM2.5 concentration.
- Collaboration in FL is still beneficial with Markovian data!
  - The sample complexity per client scales inversely with the number of clients.

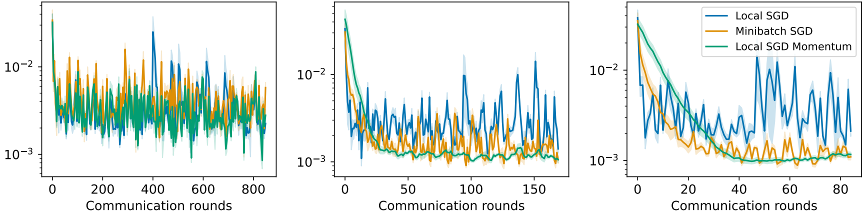


(a) Local SGD **(b)** Minibatch SGD **(c)** Local SGD-M

Figure: Grad. norm trajectory with different number of clients

Experimental results

- Beijing Multi-Site Air-Quality dataset: hourly measurements of different air-quality indicators during 4 years.
  - Prediction target: seasonality-adjusted PM2.5 concentration.
- Local SGD suffers from heterogeneity, while Minibatch SGD & Local SGD with Momentum do not.



(a) $K = 10$ **(b)** $K = 50$ **(c)** $K = 100$

Figure: Grad. norm trajectory for with different number of local steps K.

Thank you, see you in San Diego!

Poster session: Thu 4 Dec 4:30 p.m. - 7:30 p.m. PST