Federated Learning with Partial Client Participation

July 26, 2021

Let $\Delta w_t^{(i)} = \frac{1}{\eta}(w_t - w_{t,\tau}^{(i)})$ be the update sent by node i at time t. Let p_i be the probability with which the i-th node sends its update. Let $\mathcal{A}(t)$ be the set of active clients at round t. We study the following setups:

i) FedAvg(Unbiased) or FedAvg(Importance Sampling)

$$G_t = \frac{1}{n} \sum_{i \in \mathcal{A}(t)} \frac{\Delta w_t^{(i)}}{p_i}$$

ii) MIFA

$$\begin{aligned} G_t^{(i)} &= \begin{cases} \Delta w_t^{(i)}, \text{ if } i \in \mathcal{A}(t) \\ G_{t-1}^{(i)} \text{ otherwise} \end{cases} \\ G_t &= \frac{1}{n} \sum_{i \in [n]} G_t^{(i)} \end{aligned}$$

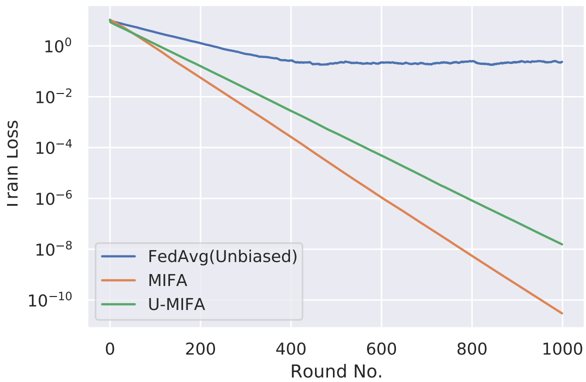
Note that in this case $\mathbb{E}\left[G_t^{(i)}\right] \neq \Delta w_t^{(i)}$

iii) Unbiased MIFA or U-MIFA

$$\begin{aligned} G_t^{(i)} &= \begin{cases} \frac{1}{p_i} \Delta w_t^{(i)} - \left(\frac{1}{p_i} - 1\right) G_{t-1}^{(i)}, \text{ if } i \in \mathcal{A}(t) \\ G_{t-1}^{(i)} \text{ otherwise} \end{cases} \\ G_t &= \frac{1}{n} \sum_{i \in [n]} G_t^{(i)} \end{aligned}$$

Note that in this case $\mathbb{E}\left[G_t^{(i)}\right] = \Delta w_t^{(i)}$.

Simple Quadratic Experiment
10 clients, $p_i = 0.1$, $\eta = 0.02$.



Q) Can we speed things up with momentum?

iv) MIFAm

$$G_t^{(i)} = \begin{cases} \Delta w_t^{(i)}, & \text{if } i \in \mathcal{A}(t) \\ G_{t-1}^{(i)} & \text{otherwise} \end{cases}$$

$$G_t = \frac{1}{n} \sum_{i \in [n]} G_t^{(i)}$$

$$v_t = \beta v_{t-1} + G_t$$

$$w_t = w_{t-1} - \eta v_t$$

v) U-MIFAm

$$G_t^{(i)} = \begin{cases} \frac{1}{p_i} \Delta w_t^{(i)} - \left(\frac{1}{p_i} - 1\right) G_{t-1}^{(i)}, & \text{if } i \in \mathcal{A}(t) \\ G_{t-1}^{(i)} & \text{otherwise} \end{cases}$$

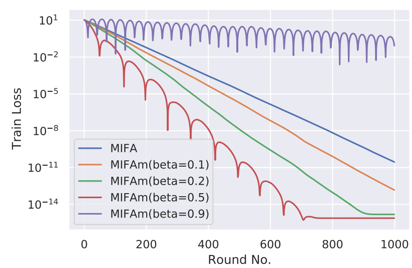
$$G_t = \frac{1}{n} \sum_{i \in [n]} G_t^{(i)}$$

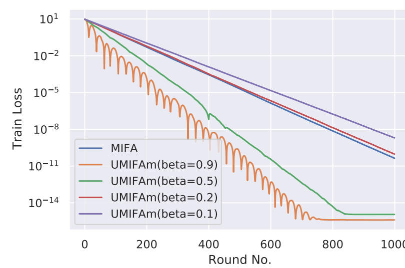
$$v_t = \beta v_{t-1} + G_t$$

$$w_t = w_{t-1} - \eta v_t$$

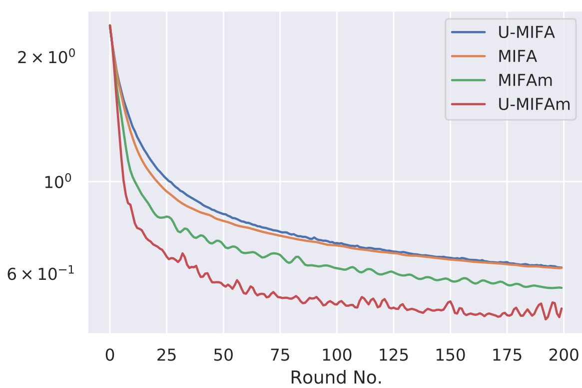
Empirical results suggest that **U-MIFAm** works better in practice.

Experiment on simple quadratic





Experiment on Logistic Regression:



Objective:

$$f_i(w) = \frac{1}{n}\sum_{i=1}^{n} f_i(w) \tag{1}$$

Assumption:Bounded client variance

$$\|\nabla f_i(w_t) - \nabla f(w_t)\|^2 \leq \sigma_G^2$$

i) Non-zero error floor for Partial Client Participation:

$$G_t = \frac{1}{n}\sum_{i \in \mathcal{A}(t)} \frac{\nabla f_i(w_t)}{p}$$

$$\begin{aligned}
f(w_{t+1}) - f(w_t) &\leq \underbrace{-\eta \mathbb{E}\left[\langle \nabla f(w_t), G_t \rangle\right]}_{T_1} + \underbrace{\frac{\eta^2 L}{2}\mathbb{E}\left[\|G_t\|^2\right]}_{T_2} \\
&\leq -\eta \langle \nabla f(w_t), \mathbb{E}\left[G_t\right] \rangle + \frac{\eta^2 L}{2}\mathbb{E}\left[\|G_t\|^2\right] \\
&= -\eta \langle \nabla f(w_t), \nabla f(w_t) \rangle + \frac{\eta^2 L}{2}\mathbb{E}\left[\|G_t\|^2\right] \\
&= -\eta \left\|\nabla f(w_t)\right\|^2 + \frac{\eta^2 L}{2}\left[\left\|\nabla f(w_t)^2\right\|^2 + \frac{1}{n^2}\left(\frac{1}{p} - 1\right)\sum_{i=1}^{n}\left\|\nabla f_i(w_t)\right\|^2\right] \\
&= -\eta \left\|\nabla f(w_t)\right\|^2 + \frac{\eta^2 L}{2}\left[\left\|\nabla f(w_t)^2\right\|^2 + \frac{1}{n^2}\left(\frac{1}{p} - 1\right)\sum_{i=1}^{n}\left\|\nabla f_i(w_t) - \nabla f(w_t) + \nabla f(w_t)\right\|^2\right] \\
&\leq -\eta \left\|\nabla f(w_t)\right\|^2 + \frac{\eta^2 L}{2}\left[\left\|\nabla f(w_t)^2\right\|^2 + \frac{2}{n}\left(\frac{1}{p} - 1\right)\sigma_G^2 + \frac{2}{n}\left(\frac{1}{p} - 1\right)\left\|\nabla f(w_t)\right\|^2\right] \\
&= -\eta \left\|\nabla f(w_t)\right\|^2\left(1 - \frac{\eta L}{2}(1 + c)\right) + \frac{\eta^2 Lc}{2}\sigma_G^2 \\
&\leq -\frac{\eta}{2}\left\|\nabla f(w_t)\right\|^2 + \frac{\eta^2 Lc}{2}\sigma_G^2
\end{aligned}$$

Summing over $t = 0, 1, \ldots, T-1$ we have,

$$\mathbb{E}\left[\frac{1}{T}\sum_{t=0}^{T-1}\left\|\nabla f(w_t)\right\|^2\right] \leq \frac{2(f(w_0) - f^*)}{\eta T} + \frac{\eta^2 Lc}{2}\sigma_G^2$$

ii) Zero error floor for MIFA:

$$G_t^{(i)} = \begin{cases} \nabla f_i(w_t), \text{ if } i \in \mathcal{A}(t) \\ G_{t-1}^{(i)} \text{ otherwise} \end{cases}$$
$$G_t = \frac{1}{n} \sum_{i \in [n]} G_t^{(i)}$$

Bounding T_1
We have,

$$\begin{aligned} T_1 &= -\eta \langle \nabla f(w_t), G_t \rangle \\ &= -\frac{\eta}{2} \|\nabla f(w_t)\|^2 - \frac{\eta}{2} \|G_t\|^2 + \frac{\eta}{2} \|\nabla f(w_t) - G_t\|^2 \end{aligned}$$

$$\begin{aligned} \|\nabla f(w_t) - G_t\|^2 &= \left\| \frac{1}{n} \sum_{i=1}^n \nabla f_i(w_t) - \frac{1}{n} \sum_{i=1}^n f_i(w_{t-\tau(t,i)}) \right\|^2 \\ &\leq \frac{1}{n} \sum_{i=1}^n \left\| f_i(w_t) - f_i(w_{t-\tau(t,i)}) \right\|^2 \\ &\leq \frac{L^2}{n} \sum_{i=1}^n \left\| w_t - w_{t-\tau(t,i)} \right\|^2 \\ &\leq \frac{L^2 \eta^2}{n} \sum_{i=1}^n \tau(t,i) \sum_{j=t-\tau(t,i)}^{t} \|G_j\|^2 \end{aligned}$$

We have therefore,

$$\begin{aligned} f(w_{t+1}) - f(w_t) &\leq -\frac{\eta}{2} \|\nabla f(w_t)\|^2 - \frac{\eta}{2} \mathbb{E}\left[\|G_t\|^2\right] + \frac{\eta^3 L^2}{2n} \sum_{i=1}^n \tau(t,i) \sum_{j=t-\tau(t,i)}^{t} \mathbb{E}\left[\|G_j\|^2\right] + \frac{\eta^2 L}{2} \mathbb{E}\left[\|G_t\|^2\right] \\ &= -\frac{\eta}{2} \|\nabla f(w_t)\|^2 + \frac{\eta^3 L^2}{2n} \sum_{i=1}^n \tau(t,i) \sum_{j=t-\tau(t,i)}^{t} \mathbb{E}\left[\|G_j\|^2\right] - \frac{\eta}{2}(1 - \eta L) \mathbb{E}\left[\|G_t\|^2\right] \end{aligned}$$

Summing over $t = 1, 2, \ldots T-1$ we have,

$$\begin{aligned}
f(w_T) - f(w_1) &\leq -\frac{\eta}{2} \sum_{t=1}^{T-1} \|\nabla f(w_t)\|^2 + \frac{\eta^3 L^2}{2n} \sum_{t=1}^{T-1} \sum_{i=1}^{n} \tau(t,i) \sum_{j=t-\tau(t,i)}^{t} \mathbb{E}\left[\|G_j\|^2\right] - \frac{\eta}{2}(1-\eta L) \sum_{t=1}^{T-1} \mathbb{E}\left[\|G_t\|^2\right] \\
&\leq -\frac{\eta}{2} \sum_{t=1}^{T-1} \|\nabla f(w_t)\|^2 + \frac{\eta^3 L^2}{2n} \sum_{t=1}^{T-1} \sum_{i=1}^{n} \tau_{\max} \sum_{j=t-\tau_{\max}}^{t} \mathbb{E}\left[\|G_j\|^2\right] - \frac{\eta}{2}(1-\eta L) \sum_{t=1}^{T-1} \mathbb{E}\left[\|G_t\|^2\right] \\
&\leq -\frac{\eta}{2} \sum_{t=1}^{T-1} \|\nabla f(w_t)\|^2 + \frac{\eta^3 L^2 \tau_{\max}^2}{2} \sum_{t=1}^{T-1} \mathbb{E}\left[\|G_j\|^2\right] - \frac{\eta}{2}(1-\eta L) \sum_{t=1}^{T-1} \mathbb{E}\left[\|G_t\|^2\right] \\
&\leq -\frac{\eta}{2} \sum_{t=1}^{T-1} \|\nabla f(w_t)\|^2 - \frac{\eta}{2}\left(1 - \eta L - \eta^2 L^2 \tau_{\max}^2\right) \sum_{t=1}^{T-1} \mathbb{E}\left[\|G_t\|^2\right] \\
&\leq -\frac{\eta}{2} \sum_{t=1}^{T-1} \|\nabla f(w_t)\|^2
\end{aligned}$$

We get,

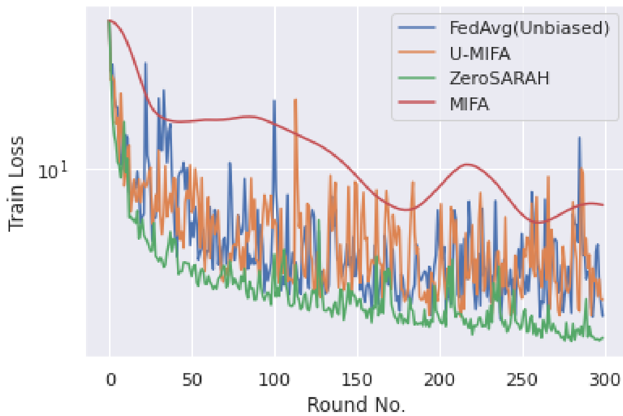
$$\mathbb{E}\left[\frac{1}{T-1} \sum_{t=1}^{T-1} \|\nabla f(w_t)\|^2\right] \leq \frac{2(f(w_1) - f^*)}{\eta T - 1}$$

Equivalence between variance reduction methods in SGD and partial client participation

Methods:- SAG,SAGA, SVRG, SARAH

SAG,SAGA methods require you to pay a memory cost $(n \times d)$

SVRG, SARAH require you to compute updates from all clients occasionally.



General analysis for MIFA and U-MIFA:

$$\begin{aligned}
f(w_{t+1}) - f(w_t) &\leq -\eta \mathbb{E}\left[\langle \nabla f(w_t), G_t \rangle\right] + \frac{\eta^2 L}{2} \mathbb{E}\left[\|G_t\|^2\right] \\
&= -\frac{\eta}{2}\|\nabla f(w_t)\|^2 - \frac{\eta}{2}\mathbb{E}\left[\|G_t\|^2\right] + \frac{\eta}{2}\mathbb{E}\left[\|\nabla f(w_t) - G_t\|^2\right] + \frac{\eta^2 L}{2}\mathbb{E}\left[\|G_t\|^2\right] \\
&= -\frac{\eta}{2}\|\nabla f(w_t)\|^2 + \frac{\eta}{2}\mathbb{E}\left[\|\nabla f(w_t) - G_t\|^2\right] - \frac{\eta}{2}(1 - \eta L)\,\mathbb{E}\left[\|G_t\|^2\right] \\
&\leq -\frac{\eta}{2}\|\nabla f(w_t)\|^2 + \frac{\eta}{2}\mathbb{E}\left[\|\nabla f(w_t) - G_t\|^2\right]
\end{aligned}$$

The second step uses $\langle x, y \rangle = -\frac{1}{2}\|x\|^2 - \frac{1}{2}\|y\|^2 + \frac{1}{2}\|x - y\|^2$. The fourth step uses $\eta L \leq 1$.

Analyzing $\mathbb{E}\left[\|\nabla f(w_t) - G_t\|^2\right]$ for MIFA:

$$\begin{aligned}
\mathbb{E}\left[\|\nabla f(w_t) - G_t\|^2\right] &= \mathbb{E}\left[\left\|\frac{1}{n}\sum_{i=1}^{n}\nabla f_i(w_t) - \frac{1}{n}\sum_{i=1}^{n} G_t^{(i)}\right\|^2\right] \\
&\leq \mathbb{E}\left[\frac{1}{n}\sum_{i=1}^{n}\left\|\nabla f_i(w_t) - G_t^{(i)}\right\|^2\right] \\
&= \frac{1-p}{n}\sum_{i=1}^{n}\left\|\nabla f_i(w_t) - G_{t-1}^{(i)}\right\|^2
\end{aligned}$$

Analyzing $\mathbb{E}\left[\|\nabla f(w_t) - G_t\|^2\right]$ for U-MIFA:

$$\begin{aligned}
\mathbb{E}\left[\|\nabla f(w_t) - G_t\|^2\right] &= \mathbb{E}\left[\left\|\frac{1}{n}\sum_{i=1}^{n}\nabla f_i(w_t) - \frac{1}{n}\sum_{i=1}^{n} G_t^{(i)}\right\|^2\right] \\
&= \mathbb{E}\left[\frac{1}{n^2}\sum_{i=1}^{n}\left\|\nabla f_i(w_t) - G_t^{(i)}\right\|^2\right] \\
&= \frac{1-p}{n^2 p}\sum_{i=1}^{n}\left\|\nabla f_i(w_t) - G_{t-1}^{(i)}\right\|^2 \\
&= \frac{1}{np}\left(\frac{1-p}{n}\sum_{i=1}^{n}\left\|\nabla f_i(w_t) - G_{t-1}^{(i)}\right\|^2\right)
\end{aligned}$$

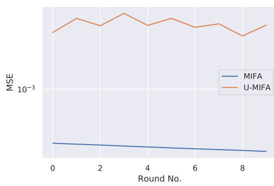


Figure 1: $p = 0.01$

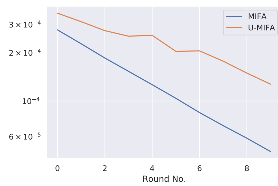


Figure 2: $p = 0.1$

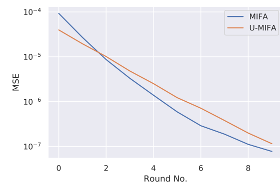


Figure 3: $p = 0.5$

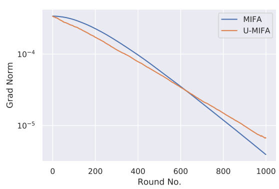


Figure 4: $p = 0.01$

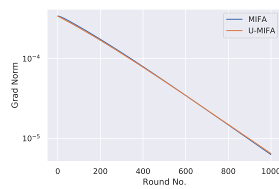


Figure 5: $p = 0.1$

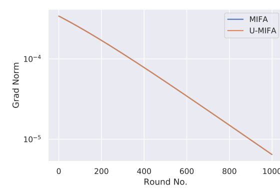
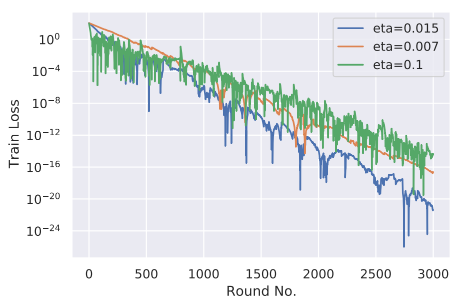


Figure 6: $p = 0.5$

Learning Rate:

MIFA paper suggests using $\eta \leq \frac{1}{Ln}$. Reddi paper suggest we can use something like $\frac{1}{L\eta^{2/3}}$



i) Convex and $L-$smooth objective

Criteria - $\min_{t \in T} f(x_t) - f(x^*) \leq \epsilon$

Assumptions

i) $\|\nabla f_i(x) - \nabla f(x)\|^2 \leq \sigma_G^2$
ii) sample b clients at each round without replacement

Method	η bound	Convergence	Complexity
FedGD	$\mathcal{O}(\frac{1}{L})$	$\frac{1}{T}(\frac{1}{2\eta}\|x_0 - x^*\|^2 + f(x_0) - f(x^*))$	$\mathcal{O}(\frac{n}{\epsilon})$
FedPGD	$\mathcal{O}(\frac{1}{L})$	$\frac{1}{T}(\frac{1}{2\eta}\|x_0 - x^*\|^2 + 2(f(x_0) - f(x^*)) + \frac{\eta L}{n-1}(\frac{n}{b} - 1)\sigma_G^2$	$\mathcal{O}(\frac{1}{\epsilon^2})$ w.d
SAGA/U-MIFA	$\mathcal{O}(\frac{1}{L})$	$\frac{1}{T}(\frac{8}{\eta}\|x_0 - x^*\|^2 + 8n(f(x_0) - f(x^*))$	$\mathcal{O}(\frac{n}{\epsilon})$??

i) Non-Convex and $L-$smooth objective

Criteria - $\min_{t \in T} \|\nabla f(x_t)\|^2 \leq \epsilon$

Assumptions

i) $\|\nabla f_i(x) - \nabla f(x)\|^2 \leq \sigma^2$
ii) sample b clients at each round without replacement

Method	η bound	Convergence	Complexity
FedGD	$\mathcal{O}(\frac{1}{L})$	$\frac{2(f(x_0)-f(x^*)}{\eta T}$	$\mathcal{O}(\frac{nL}{\epsilon})$
FedPGD	$\mathcal{O}(\frac{1}{L})$	$\frac{2(f(x_0)-f(x^*)}{\eta T} + \frac{\eta L}{n-1}(\frac{n}{b} - 1)\sigma^2$	$\mathcal{O}(\frac{L}{\epsilon^2})$
SAGA/U-MIFA	$\mathcal{O}(\frac{1}{Ln^{2/3}})$	$\frac{\mathcal{O}(1)(f(x_0)-f(x^*))}{\eta T}$	$\mathcal{O}(\frac{n^{2/3}L}{\epsilon})$
ZeroSarah ??	$\mathcal{O}(\frac{1}{Ln^{1/2}})$	$\frac{\mathcal{O}(1)(f(x_0)-f(x^*))+\Delta}{\eta T}$	$\mathcal{O}(\frac{n^{1/2}L}{\epsilon})$