# Adaptive Federated Learning with Auto-tuned Clients via Local Smoothness

Junhyung Lyle Kim (Rice CS), Mohammad Taha Toghani (Rice ECE), CésarA.Uribe (Rice ECE), Anastasios Kyrillidis (Rice CS)

## Federated Learning Overview

- FL is a recent distributed machine learning framework where a global model is trained via multiple collaborative steps by participating clients without sharing data.
- Mathematically, we want to solve

$$\min_{x\in\mathbb{R}^d} \; f(x) = \frac{1}{m}\sum_{i=1}^{m} f_i(x)$$

where $x \in \mathbb{R}^d$ is the shared model parameter, $m$ is the number of clients, and $f_i(x) := \mathbb{E}_{z\sim\mathcal{D}_i}[F_i(x,z)]$ is the individual loss function.

- Offers flexible collaborative learning: the number of clients $m$, their participation rates, and the computing power can all vary and change at any point during the overall training procedure.

$$f_i(x) \leftarrow \mathbb{E}_{z\sim\mathcal{D}_i}[F_i(x,z)]$$

- Therefore, not only $\mathcal{D}_i$ differs for each client $i$, but also the number of samples $z \sim \mathcal{D}_i$, resulting in each client having different $f_i(\cdot)$.

## Challenges in Federated Learning

**Algorithm** FedAvg

**input**: $x_0 \in \mathbb{R}^d$, $\eta > 0$, and $p \in (0,1)$.
**for** each round $t = 0, 1, \ldots, T-1$ **do**
  sample a subset $\mathcal{S}_t$ of clients with size $|\mathcal{S}_t| = p \cdot m$
  **for** each machine in parallel for $i \in \mathcal{S}_t$ **do**
    Set $x_{t,0}^i = x_t$
    **for** local step $k \in [K]$ **do**
      Compute an estimate $g_{t,k-1}^i$ of $\nabla f_i(x_{t,k-1}^i)$
      $x_{t,k}^i = x_{t,k-1}^i - \eta g_{t,k-1}^i$
    **end for**
  **end for**
  $x_{t+1} = \frac{1}{|\mathcal{S}_t|}\sum_{i\in\mathcal{S}_t} x_{t,K}^i = x_t - \frac{1}{|\mathcal{S}_t|}\sum_{i\in\mathcal{S}_t}(x_t - x_{t,K}^i)$
**end for**
**return** $x_T$

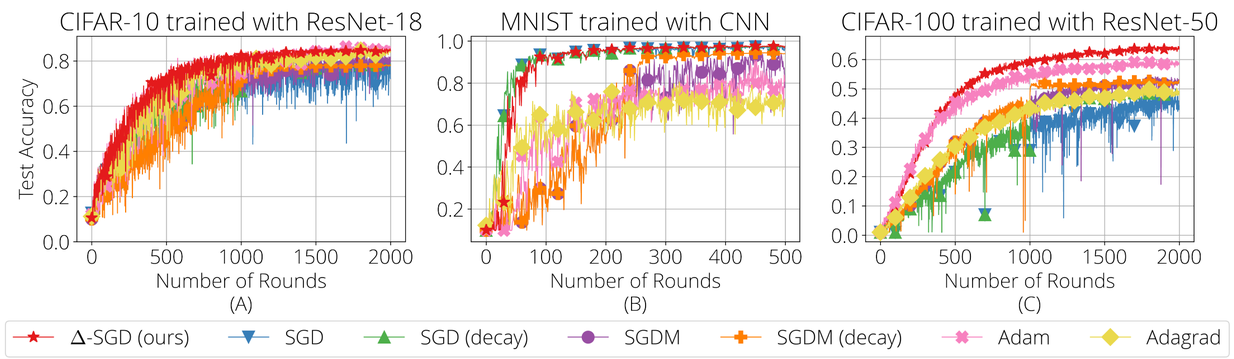
**Server-side:** How do we smartly aggregate the local information coming from each participating client?

- Federated Averaging [McMahan et al., 2017] uses simple averaging
- [Reddi et al., 2021] interpreted averaging as a "pseudo-gradient" step and introduced FedAdam, FedYogi, FedAdagrad, etc.

**Client-side:** How do we make sure each client meaningfully "learns" using local data?

- Federated Averaging [McMahan et al., 2017] uses SGD.
- Does it make sense to use the same $\eta$ for all clients? If not, how should we tune individual step sizes?
- Important to properly tune: many more local updates compared to the aggregation step, as communication is much more expensive.

## Client Optimization Is More Challenging?




- We fine-tune the step size for each client optimizer in task (A), and *intentionally keep it the same* for the other tasks, to highlight the effect of not properly tuning the step size of each client optimizer. Our proposed method, $\Delta$-SGD, exhibits superior performance in all settings, without any additional tuning.
- (A): CIFAR-10 classification task trained on ResNet-18. (B): MNIST classification task trained on shallow CNN. (C): CIFAR-100 classification task trained on ResNet-50.
- All experiments use FedAvg as the server optimizer.

## Why is the step size $\eta = 1/L$ popular?

- $L$-smooth functions: $f(y) \le f(x) + \langle \nabla f(x), y-x\rangle + \frac{L}{2}\|y-x\|^2 \;\; \forall x, y$
- Gradient descent: $x_{t+1} = x_t - \eta \nabla f(x_t)$
- Descent lemma:

$$f(x_{t+1}) \le f(x_t) + \langle \nabla f(x_t), x_{t+1} - x_t\rangle + \frac{L}{2}\|x_{t+1} - x_t\|^2$$
$$= f(x_t) - \eta\left(1 - \frac{\eta\cdot L}{2}\right)\|\nabla f(x_t)\|^2$$

- $\eta = 1/L$ maximizes the descent progress.

## Adaptive Step Size via Local Smoothness

- [Malisky & Mishchenko, 2020] proposed the following step size for (centralized) gradient descent:

$$\eta_t = \min\left\{\frac{\|x_t - x_{t-1}\|}{2\|\nabla f(x_t) - \nabla f(x_{t-1})\|}, \sqrt{1 + \frac{\eta_{t-1}}{\eta_{t-2}}}\eta_{t-1}\right\}$$

- The first condition approximates the local smoothness

$$\|\nabla f(x_t) - \nabla f(x_{t-1})\| \le L_t \cdot \|x_t - x_{t-1}\|, \quad \forall t = 1,2,\ldots$$

and the second condition ensures $\eta_t$ to not increase too fast.

- We adapt the above step size to the FL setting:

$$\eta_{t,k}^i = \min\left\{\frac{\|x_{t,k}^i - x_{t,k-1}^i\|}{2\|\tilde{\nabla} f_i(x_{t,k}^i) - \tilde{\nabla} f_i(x_{t,k-1}^i)\|}, \sqrt{1 + \frac{\eta_{t,k}^i}{\eta_{t,k-1}^i}}\eta_{t,k-1}^i\right\}$$

with the stochastic gradients $\tilde{\nabla} f_i(\cdot)$ and the local iterations $k$.

- **Implication:** each client uses its own step size $\eta_t^i$ that is adaptive to the local smoothness of $f_i(\cdot)$

## Experimental Results: $\Delta$-SGD exhibits superior performance in all settings without any additional tuning

| *Non-iidness* | *Optimizer* | *Dataset / Model* | | | | |
|---|---|---|---|---|---|---|
| Dir($\alpha \cdot \mathbf{p}$) | | MNIST CNN | FMNIST CNN | CIFAR-10 ResNet-18 | CIFAR-100 ResNet-18 | CIFAR-100 ResNet-50 |
| $\alpha = 1$ | SGD | **98.3**$_{\downarrow(0.2)}$ | 86.5$_{\downarrow(0.8)}$ | 87.7$_{\downarrow(2.1)}$ | 57.7$_{\downarrow(4.2)}$ | 53.0$_{\downarrow(12.8)}$ |
| | SGD (↓) | 97.8$_{\downarrow(0.7)}$ | 86.3$_{\downarrow(1.0)}$ | 87.8$_{\downarrow(2.0)}$ | **61.9**$_{\downarrow(0.0)}$ | 60.9$_{\downarrow(4.9)}$ |
| | SGDM | **98.5**$_{\downarrow(0.0)}$ | 85.2$_{\downarrow(2.1)}$ | 88.7$_{\downarrow(1.1)}$ | 58.8$_{\downarrow(3.1)}$ | 60.5$_{\downarrow(5.3)}$ |
| | SGDM (↓) | **98.4**$_{\downarrow(0.1)}$ | **87.2**$_{\downarrow(0.1)}$ | **89.3**$_{\downarrow(0.5)}$ | **61.4**$_{\downarrow(0.5)}$ | 63.3$_{\downarrow(2.5)}$ |
| | Adam | 94.7$_{\downarrow(3.8)}$ | 71.8$_{\downarrow(15.5)}$ | **89.4**$_{\downarrow(0.4)}$ | 55.6$_{\downarrow(6.3)}$ | 61.4$_{\downarrow(4.4)}$ |
| | Adagrad | 64.3$_{\downarrow(34.2)}$ | 45.5$_{\downarrow(41.8)}$ | 86.6$_{\downarrow(3.2)}$ | 53.5$_{\downarrow(8.4)}$ | 51.9$_{\downarrow(13.9)}$ |
| | SPS | 10.1$_{\downarrow(88.4)}$ | 85.9$_{\downarrow(1.4)}$ | 82.7$_{\downarrow(7.1)}$ | 1.0$_{\downarrow(60.9)}$ | 50.0$_{\downarrow(15.8)}$ |
| | $\Delta$-SGD | **98.4**$_{\downarrow(0.1)}$ | **87.3**$_{\downarrow(0.0)}$ | **89.8**$_{\downarrow(0.0)}$ | **61.5**$_{\downarrow(0.4)}$ | **65.8**$_{\downarrow(0.0)}$ |
| $\alpha = 0.1$ | SGD | **98.1**$_{\downarrow(0.0)}$ | 83.6$_{\downarrow(2.8)}$ | 72.1$_{\downarrow(12.9)}$ | 54.4$_{\downarrow(6.7)}$ | 44.2$_{\downarrow(19.9)}$ |
| | SGD (↓) | **98.0**$_{\downarrow(0.1)}$ | 84.7$_{\downarrow(1.7)}$ | 78.4$_{\downarrow(6.6)}$ | 59.3$_{\downarrow(1.8)}$ | 48.7$_{\downarrow(15.4)}$ |
| | SGDM | **97.6**$_{\downarrow(0.5)}$ | 83.6$_{\downarrow(2.8)}$ | 79.6$_{\downarrow(5.4)}$ | 58.8$_{\downarrow(2.3)}$ | 52.3$_{\downarrow(11.8)}$ |
| | SGDM (↓) | **98.0**$_{\downarrow(0.1)}$ | **86.1**$_{\downarrow(0.3)}$ | 77.9$_{\downarrow(7.1)}$ | 60.4$_{\downarrow(0.7)}$ | 52.8$_{\downarrow(11.3)}$ |
| | Adam | 96.4$_{\downarrow(1.7)}$ | 80.4$_{\downarrow(6.0)}$ | **85.0**$_{\downarrow(0.0)}$ | 55.4$_{\downarrow(5.7)}$ | 58.2$_{\downarrow(5.9)}$ |
| | Adagrad | 89.9$_{\downarrow(8.2)}$ | 46.3$_{\downarrow(40.1)}$ | 84.1$_{\downarrow(0.9)}$ | 49.6$_{\downarrow(11.5)}$ | 48.0$_{\downarrow(16.1)}$ |
| | SPS | 96.0$_{\downarrow(2.1)}$ | 85.0$_{\downarrow(1.4)}$ | 70.3$_{\downarrow(14.7)}$ | 42.2$_{\downarrow(18.9)}$ | 42.2$_{\downarrow(21.9)}$ |
| | $\Delta$-SGD | **98.1**$_{\downarrow(0.0)}$ | **86.4**$_{\downarrow(0.0)}$ | **84.5**$_{\downarrow(0.5)}$ | **61.1**$_{\downarrow(0.0)}$ | **64.1**$_{\downarrow(0.0)}$ |
| $\alpha = 0.01$ | SGD | 96.8$_{\downarrow(0.7)}$ | 79.0$_{\downarrow(1.2)}$ | 22.6$_{\downarrow(11.3)}$ | 30.5$_{\downarrow(1.3)}$ | 24.3$_{\downarrow(7.1)}$ |
| | SGD (↓) | **97.2**$_{\downarrow(0.3)}$ | 79.3$_{\downarrow(0.9)}$ | **33.9**$_{\downarrow(0.0)}$ | 30.3$_{\downarrow(1.5)}$ | 24.6$_{\downarrow(6.8)}$ |
| | SGDM | 77.9$_{\downarrow(19.6)}$ | 75.7$_{\downarrow(4.5)}$ | 28.4$_{\downarrow(5.5)}$ | 24.8$_{\downarrow(7.0)}$ | 22.0$_{\downarrow(9.4)}$ |
| | SGDM (↓) | 94.0$_{\downarrow(3.5)}$ | 79.5$_{\downarrow(0.7)}$ | 29.0$_{\downarrow(4.9)}$ | 20.9$_{\downarrow(10.9)}$ | 14.7$_{\downarrow(16.7)}$ |
| | Adam | 80.8$_{\downarrow(16.7)}$ | 60.6$_{\downarrow(19.6)}$ | 22.1$_{\downarrow(11.8)}$ | 18.2$_{\downarrow(13.6)}$ | 22.6$_{\downarrow(8.8)}$ |
| | Adagrad | 72.4$_{\downarrow(25.1)}$ | 45.9$_{\downarrow(34.3)}$ | 12.5$_{\downarrow(21.4)}$ | 25.8$_{\downarrow(6.0)}$ | 22.2$_{\downarrow(9.2)}$ |
| | SPS | 69.7$_{\downarrow(27.8)}$ | 44.0$_{\downarrow(36.2)}$ | 21.5$_{\downarrow(12.4)}$ | 22.0$_{\downarrow(9.8)}$ | 17.4$_{\downarrow(14.0)}$ |
| | $\Delta$-SGD | **97.5**$_{\downarrow(0.0)}$ | **80.2**$_{\downarrow(0.0)}$ | 31.6$_{\downarrow(2.3)}$ | **31.8**$_{\downarrow(0.0)}$ | **31.4**$_{\downarrow(0.0)}$ |

- We examine the performance difference when we change the dataset, model architecture, or level of heterogeneity induced by the Dirichlet concentration parameter $\alpha$.
- Step size for each client optimizer is fine-tuned for CIFAR-10 classification task trained with ResNet-18, with $\alpha = 0.1$.