CS-E4740 - Federated Learning

FL Algorithms

Assoc. Prof. Alexander Jung

Spring 2025

Playlist

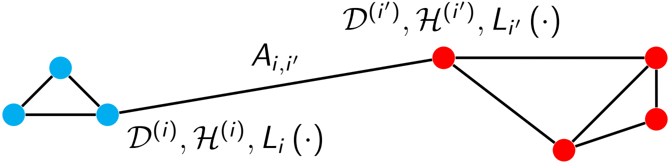
Glossary

Course Site

Outline

Table of Contents

FL Network as Mathematical Model for FL



- An FL network consists of devices $i = 1, \ldots, n$.
- Some i, i' connected by edge with the weight $A_{i,i'} > 0$.
- Device i **generates data** $\mathcal{D}^{(i)}$ and **trains model** $\mathcal{H}^{(i)}$.
- Data $\mathcal{D}^{(i)}$ used to construct loss func. $L_i(\cdot)$.

GTV Minimization (for Parametric Models)

We train local models in a collaborative fashion by solving

$$\min_{\mathbf{w}^{(1)},\ldots,\mathbf{w}^{(n)}} \sum_{i=1}^{n} L_i\left(\mathbf{w}^{(i)}\right) + \alpha \sum_{\{i,i'\}\in\mathcal{E}} A_{i,i'} \left\|\mathbf{w}^{(i)} - \mathbf{w}^{(i')}\right\|_2^2 \quad \text{(GTVMin)}.$$

- Solution consists of learnt modelparams. $\widehat{\mathbf{w}}^{(i)}$.
- Tuning parameter $\alpha \geq 0$ controls clustering of $\widehat{\mathbf{w}}^{(i)}$.
- For $\alpha = 0$, GTVMin reduces to separate ERM for each i.
- Increasing α makes $\widehat{\mathbf{w}}^{(i)}$ more similar across nodes i.

Learning Goals

After completing this module, you know how

- FL alg. can be obtained from **gradient descent**,
- to implement GD as **message passing**,
- to generalize GD to handle **non-parametric models**,
- to implement **asynchronous FL algos.**

Table of Contents

Gradient Step for GTVMin

Starting from initial local params. $\mathbf{w}^{(i,0)}$, repeat grad. steps

$$\mathbf{w}^{(i,k+1)} = \mathbf{w}^{(i,k)} - \eta_{k,i}\left[\nabla L_i\left(\mathbf{w}^{(i,k)}\right) + 2\alpha \sum_{i' \in \mathcal{N}^{(i)}} A_{i,i'}\left(\mathbf{w}^{(i,k)} - \mathbf{w}^{(i',k)}\right)\right]$$

- The learn. rate $\eta_{k,i}$ determines extent of update.
- $\nabla L_i\left(\mathbf{w}^{(i,k)}\right)$ **steers** update towards **min. local loss**.
- $\left(\mathbf{w}^{(i,k)} - \mathbf{w}^{(i',k)}\right)$ **steers** update to **agree with neigh.**
- $\alpha A_{i,i''}$ balances those two steering effects.

Synchronous Operation

The gradient step

$$\mathbf{w}^{(i,k+1)} = \mathbf{w}^{(i,k)} - \eta_{k,i} \left[\nabla L_i \left(\mathbf{w}^{(i,k)} \right) + 2\alpha \sum_{i' \in \mathcal{N}^{(i)}} A_{i,i'} \left(\mathbf{w}^{(i,k)} - \mathbf{w}^{(i',k)} \right) \right]$$

has to be carried out by all nodes $i = 1, \ldots, n$.

When these (local) gradient steps are completed, each node shares its new model params. with its neighbours.

After sharing the model params., start new iteration $k := k+1$.

Message Passing Implementation

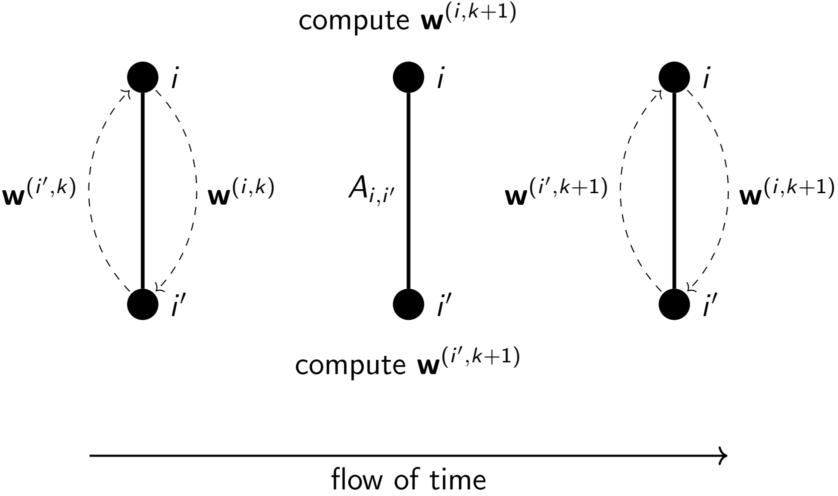


Table of Contents

Federated Gradient Descent (FedGD)

Each node $i = 1, \ldots, n$ initializes

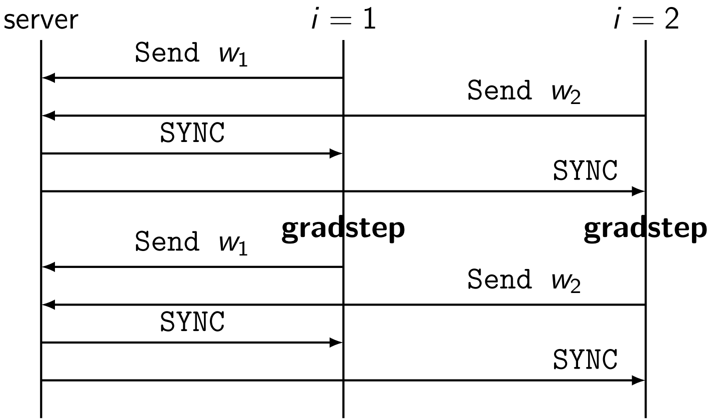
- local model params. $\widehat{\mathbf{w}}_0^{(i)} := \mathbf{0}$, and
- iteration counter $k := 0$.

Repeat the following steps at each node i:

- Send $\widehat{\mathbf{w}}_k^{(i)}$ to all neighbours $\mathcal{N}^{(i)}$.
- Do a gradient step.
- Increment iteration counter $k := k + 1$.

CAUTION: Nodes must execute steps synchronously!

Implementing FedGD with a Sync-Server



Python demo

Federated Stochastic Gradient Descent (FedSGD)

- Consider FL network with node i carrying local dataset

$$\mathcal{D}^{(i)} = \left\{ \left(\mathbf{x}^{(i,1)}, y^{(i,1)}\right), \ldots, \left(\mathbf{x}^{(i,m_i)}, y^{(i,m_i)}\right)\right\}.$$

- Node i uses local loss function

$$L_i\left(\mathbf{w}^{(i)}\right) := (1/m_i) \sum_{r=1}^{m_i} \left(y^{(i,r)} - \left(\mathbf{w}^{(i)}\right)^T \mathbf{x}^{(i,r)} \right)^2.$$

- FedGD requires to compute gradient,

$$\nabla L_i\left(\mathbf{w}^{(i)}\right) = (-2/m_i) \sum_{r=1}^{m_i} \mathbf{x}^{(i,r)} \left(y^{(i,r)} - \left(\mathbf{w}^{(i)}\right)^T \mathbf{x}^{(i,r)} \right).$$

Stochastic Gradient Approximation

For some applications, the computation of

$$\sum_{r=1}^{m_i} \mathbf{x}^{(i,r)} \left(y^{(i,r)} - \left(\mathbf{w}^{(i)}\right)^T \mathbf{x}^{(i,r)} \right)$$

is intractable, e.g., too many data points or too slow access.

$\Rightarrow$ Use instead a sum over random subset $\mathcal{B} \subseteq \{1, \ldots, m_i\}$,

$$\sum_{r \in \mathcal{B}} \mathbf{x}^{(i,r)} \left(y^{(i,r)} - \left(\mathbf{w}^{(i)}\right)^T \mathbf{x}^{(i,r)} \right).$$

We refer to $\mathcal{B}$ as batch with batch size $|\mathcal{B}|$.

Federated Averaging (FedAvg)

- Some FL applications use common model at all nodes,

$$\mathbf{w}^{(i)} = \mathbf{w}^{(i')} \quad \forall\, i, i' \in \mathcal{V}.$$

- GTVMin becomes constrained optimization problem:

$$\min_{\mathbf{w}^{(1)},\ldots,\mathbf{w}^{(n)}} \sum_{i \in \mathcal{V}} L_i\left(\mathbf{w}^{(i)}\right) \text{s.t.} \quad \mathbf{w}^{(i)} = \mathbf{w}^{(i')} \quad \forall\, i, i' \in \mathcal{V}.$$

- For diff. $L_i\left(\mathbf{w}^{(i)}\right)$ we can apply projected GD.

- Projection step amounts to averaging $(1/n) \sum_{i=1}^{n} \mathbf{w}^{(i)}$.

(Almost) FedAvg

Init. counter (clock) $k := 0$ and model params $\widehat{\mathbf{w}} := \mathbf{0}$.

1. **Broadcast.** Server sends $\widehat{\mathbf{w}}$ to all nodes $i \in \mathcal{V}$.
2. **Local Gradient Step.** Each node computes
$$\mathbf{w}^{(i,k)} = \widehat{\mathbf{w}} - \eta_{k,i} \nabla L_i \left(\widehat{\mathbf{w}} \right).$$
3. **Collect.** Nodes send $\mathbf{w}^{(i,k)}$ back to server.
4. **Aggregate.** Server computes $\widehat{\mathbf{w}} := (1/n) \sum_{i=1}^{n} \mathbf{w}^{(i,k)}$.
5. **Clock Tick.** Server increments $k := k + 1$. Go to step 1.

FedAvg

We obtain FedAvg via the following modifications:[1]

- Use (stochastic) approximations of gradients.
- Instead of single GD step, compute several GD steps.
- Only a subset of nodes compute local updates during each iteration.

[1] B. McMahan et.al., Communication-Efficient Learning of Deep Networks from Decentralized Data, PMLR, 2017

FedProx

FedProx replaces GD steps in FedAvg with[2]

$$\mathbf{w}^{(i)} := \underset{\mathbf{v} \in \mathbb{R}^d}{\operatorname{argmin}} \left[L_i\left(\mathbf{v}\right) + (1/\eta) \left\| \mathbf{v} - \widehat{\mathbf{w}}^{(\mathrm{global})} \right\|_2^2 \right].$$

Empirical studies found FedProx to result in more robust FL systems compared to FedAvg.

FedProx seems to handle well varying computational power of devices.

[2] T. Li, et.al, Federated Optimization in Heterogeneous Networks, Proc. of Machine Learning and Systems 2, 2020.

Federated Relaxation (FedRelax)

- Consider GTVMin objective function

$$f\left(\mathbf{w}^{(1)},\ldots,\mathbf{w}^{(n)}\right)=\sum_{i=1}^{n} L_i\left(\mathbf{w}^{(i)}\right)+\alpha\sum_{\{i,i'\}\in\mathcal{E}} A_{i,i'}\left\|\mathbf{w}^{(i)}-\mathbf{w}^{(i')}\right\|_2^2.$$

- Complicated due to coupling terms $A_{i,i'}\left\|\mathbf{w}^{(i)}-\mathbf{w}^{(i')}\right\|_2^2$.
- Without coupling, GTVMin would be much easier.
- Optimize $f(\cdot)$ w.r.t. $\mathbf{w}^{(i)}$, holding $\left\{\mathbf{w}^{(i')}\right\}_{i'\in\mathcal{V}\setminus\{i\}}$ fixed![3]

[3]Similar idea is used in the Jacobi method for solving linear equations.

FedRelax for Parametric Models

- **Init.** Set counter $k := 0$, local model params. $\widehat{\mathbf{w}}_0^{(i)} := \mathbf{0}$.
- **Repeat until stopping criterion:**
  - Each node i shares $\widehat{\mathbf{w}}_k^{(i)}$ with neighbours $\mathcal{N}^{(i)}$.
  - **Local Update.** Each node i computes

    $$\widehat{\mathbf{w}}_{k+1}^{(i)} := \operatorname*{argmin}_{\mathbf{w}^{(i)} \in \mathbb{R}^d} L_i\left(\mathbf{w}^{(i)}\right) + \alpha \sum_{i' \in \mathcal{N}^{(i)}} A_{i,i'} \left\| \mathbf{w}^{(i)} - \widehat{\mathbf{w}}_k^{(i')} \right\|_2^2 .$$

  - **Clock Tick.** $k := k + 1$.

Note the similarity of local update with ridge regression!

FedRelax for Non-Parametric Models

- **Init.** $k := 0$, construct test-set $\mathcal{D}^{\{i,i'\}}$ for each $\{i, i'\} \in \mathcal{E}$
- **Repeat until stopping criterion:**
  - Each i shares $\widehat{h}_k^{(i)}(\mathbf{x})$ for each $\mathbf{x} \in \mathcal{D}^{\{i,i'\}}$ and $i' \in \mathcal{N}^{(i)}$.
  - **Local Update.** Each node i computes
  $$\widehat{h}_{k+1}^{(i)} \in \underset{h^{(i)} \in \mathcal{H}^{(i)}}{\operatorname{argmin}} L_i\left(h^{(i)}\right) + \alpha \sum_{i' \in \mathcal{N}^{(i)}} A_{i,i'} D\big(h^{(i)}, \widehat{h}_k^{(i')}\big).$$
  - **Clock Tick.** $k := k + 1$.

Here, we use the discrepancy measure

$$D\big(h^{(i)}, h^{(i')}\big) := (1/m') \sum_{\mathbf{x} \in \mathcal{D}^{\{i,i'\}}} \left[h^{(i)}(\mathbf{x}) - h^{(i')}(\mathbf{x})\right]^2.$$

Table of Contents

FL as Fixed-Point Iterations

- Consider FL net. with parametric local models.
- FL algorithms presented so far can be written as
$$\mathbf{w}^{(i,k+1)} = \mathcal{F}^{(i)}\big(\mathbf{w}^{(1,k)}, \ldots, \mathbf{w}^{(n,k)}\big).$$
- This is a synchronous fixed-point iteration with
$$\mathcal{F}^{(i)} : \mathbb{R}^{nd} \rightarrow \mathbb{R}^{d} \text{ , for each node } i = 1, \ldots, n.$$
- **FL algorithm is determined by its fixed-point operators** $\mathcal{F}^{(1)}, \ldots, \mathcal{F}^{(n)}$, encoding local update rules.[4]

[4] Local updates can be time-varying, i.e., using $\mathcal{F}^{(i,k)}$ varying with k.

Challenges of Synchronous FL

Implementing synchronous fixed-point iteration is challenging.

- Devices might have **limited computational resources**.
- Evaluating the local loss may require **data collection**.
- **Message passing is unreliable** over wireless links.
- Devices may **spontaneously join or drop out**.

Modelling Asynchronous Federated Learning

Consider an FL algorithm with fixed-point operators $\mathcal{F}^{(i)}$.

We obtain an asynchronous variant by the update

$$\mathbf{w}^{(i,k+1)} = \begin{cases} \mathcal{F}^{(i)}\Big(\mathbf{w}^{(1,s_{i,1}^{(k)})}, \ldots, \mathbf{w}^{(n,s_{i,i'}^{(k)})}\Big) & \text{for } k \in T^{(i)} \\ \mathbf{w}^{(i,k)} & \text{otherwise.} \end{cases}$$

- The iteration index k enumerates update events.
- Node i runs local update only during events $k \in T^{(i)}$.
- **Delay** $k - s_{i,i'}^{(k)}$ from i' to i during update event $k \in T^{(i)}$.

Totally Asynchronous Algorithms

Consider asynchronous FL algorithm

$$\mathbf{w}^{(i,k+1)} = \begin{cases} \mathcal{F}^{(i)}\left(\mathbf{w}^{(1,s_{i,1}^{(k)})}, \ldots, \mathbf{w}^{(n,s_{i,i'}^{(k)})}\right) & \text{for } k \in T^{(i)} \\ \mathbf{w}^{(i,k)} & \text{otherwise.} \end{cases}$$

A totally asynchronous algorithm "works" under the following minimal assumptions:[5]

- The set $T^{(i)}$ is infinite for each $i = 1, \ldots, n$.
- The delayed update times $s_{i,i'}^{(k)}$ are unbounded,

$$\lim_{\substack{k\to\infty \\ k\in T^{(i)}}} s_{i,i'}^{(k)} = \infty.$$

[5] see Ch. 6 of D. Bertsekas, J. Tsitsiklis, "Parallel and Distributed Computation: Numerical Methods," 2015.

Partially Asynchronous Algorithms

A partially asynchronous FL algorithm

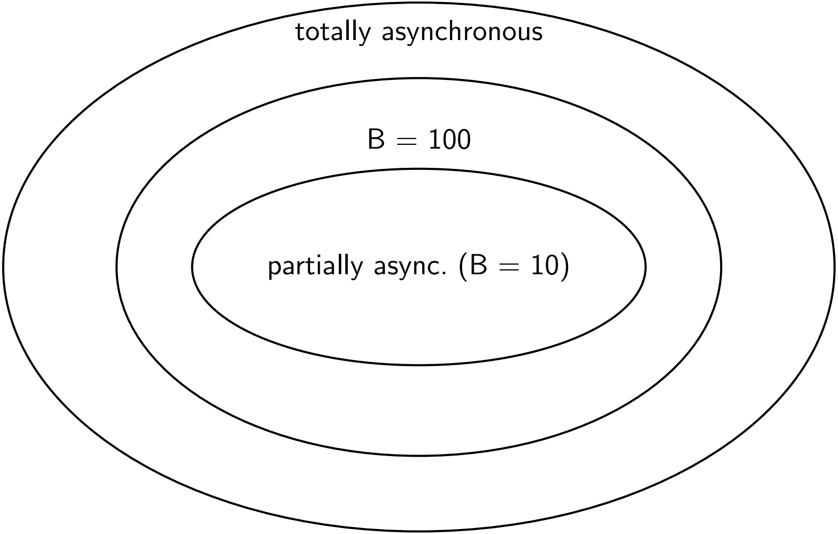
$$\mathbf{w}^{(i,k+1)} = \begin{cases} \mathcal{F}^{(i)}\left(\mathbf{w}^{(1,s_{i,i'}^{(k)})}, \ldots, \mathbf{w}^{(n,s_{i,i'}^{(k)})}\right) & \text{for } k \in T^{(i)} \\ \mathbf{w}^{(i,k)} & \text{otherwise.} \end{cases}$$

with asynchronous measure $B \in \mathbb{N}$ works as long as[6]

- $\{k, \ldots, k+B-1\} \cap T^{(i)} \neq \emptyset$ for each $k \in \mathbb{N}$, and
- bounded delay $k - s_{i,i'}^{(k)} \leq B$ for all $k \in T^{(i)}$.

[6] see Ch. 7 of D. Bertsekas, J. Tsitsiklis, "Parallel and Distributed Computation: Numerical Methods," 2015.

Asynchronous Algorithms



When Does it Work ?

An asynchronous FL algorithm is fully specified by:

- The fixed-point operators $\mathcal{F}^{(i)}$ for $i = 1, \ldots, n$.
- The update events $T^{(i)}$, for $i = 1, \ldots, n$.
- The delays $k - s_{i,i'}^{(k)}$ for $i, i' \in \mathcal{V}$, $k \in T^{(i)}$.

What are sufficient conditions on those components such that the resulting algorithm is totally (partially) asynchronous?[7]

[7] H. R. Feyzmahdavian and M. Johansson, "On the convergence rates of asynchronous iterations," Proc. IEEE CDC, 2014

Pseudo-Contractions w.r.t. Block-Maximum Norm

- Stacked local model params. $\mathbf{w} = \mathrm{stack}\{\mathbf{w}^{(i)}\}_{i=1}^{n}$.
- Define the block-maximum norm

$$\|\mathbf{w}\|_{\infty} := \max_{i=1,\ldots,n} \left\|\mathbf{w}^{(i)}\right\|_{i}.$$

FL algo. $\mathcal{F} = \left(\mathcal{F}^{(1)}, \ldots, \mathcal{F}^{(n)}\right)$ is pseudo-contraction if[8]

$$\|\mathcal{F}\mathbf{w} - \widehat{\mathbf{w}}\|_{\infty} \leq \kappa \|\mathbf{w} - \widehat{\mathbf{w}}\|_{\infty} \tag{1}$$

with fixed point $\widehat{\mathbf{w}} = \mathcal{F}\widehat{\mathbf{w}}$ and some $\kappa \in [0, 1)$.

[8] H. R. Feyzmahdavian and M. Johansson, "Asynchronous Iterations in Optimization: New Sequence Results and Sharper Algorithmic Guarantees," JMLR, 2023

A Convergence Result

Consider FL algo. which is a pseudo-contr. with $\kappa < 1$. Then,

- it converges in totally asynchronous setting,[9] and
- for partially asynchronous setting[10]

$$\left\| \mathbf{w}^{(k)} - \widehat{\mathbf{w}} \right\|_{\infty} \leq \kappa^{\frac{k}{2B+1}} \left\| \mathbf{w}^{(0)} - \widehat{\mathbf{w}} \right\|_{\infty}.$$

Thus, we should use small factor κ, and small async. measure B. How can we ensure this?

[9] Thm. 23 in H. R. Feyzmahdavian and M. Johansson, JMLR, 2023.
[10] Thm. 24 in H. R. Feyzmahdavian and M. Johansson, JMLR, 2023.

Wrap Up

- FL alg. as fixed-point iterations $\mathbf{w}^{(k+1)} = \mathcal{F}\mathbf{w}^{(k)}$.
- Fixed point of $\mathcal{F}$ is a solution of GTVMin.
- Convergence depends on contraction properties of $\mathcal{F}$.
- Tolerant against asynchronous implementation.

What's Next?

The next module studies main flavours of FL.

These flavours are characterized by specific design choices arising in FL networks and GTVMin.

Further Resources

- **YouTube:** @alexjung111
- **LinkedIn:** Alexander Jung
- **GitHub:** alexjungaalto