CS-E4740 - Federated Learning

FL Flavours

Assoc. Prof. Alexander Jung

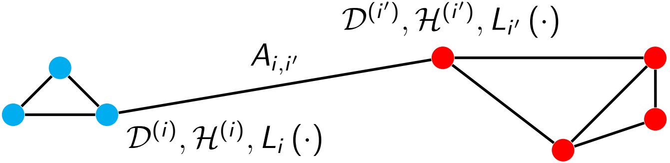
Spring 2025

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 which GTVMin design choices result in

- single-model FL,
- horizontal FL,
- vertical FL,
- clustered FL,
- personalized FL.

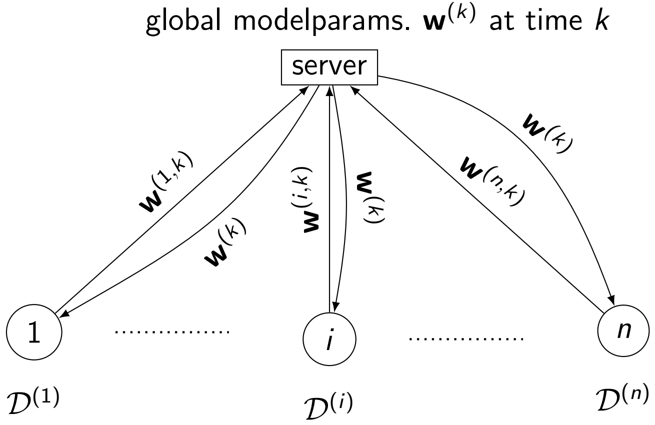
Table of Contents

Basic Workflow

- Server hold global model params. $\mathbf{w}^{(k)} \in \mathbb{R}^d$.
- Clients $i = 1, \ldots, i$ carry local datasets $\mathcal{D}^{(i)}$.
- Use $\mathcal{D}^{(i)}$ to compute update $\mathbf{w}^{(k)} \mapsto \mathbf{w}^{(i,k)}$.[1]
- Sever aggregates $\mathbf{w}^{(i,k)}$ to update $\mathbf{w}^{(k)} \mapsto \mathbf{w}^{(k+1)}$.

[1] How can you use a dataset to update model parameter?

Server-Client Implementation



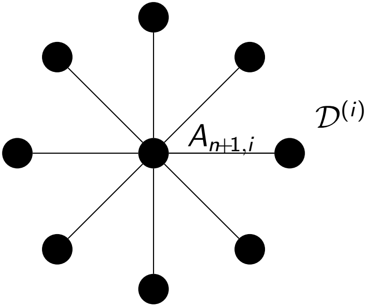
- Each client i computes $\mathbf{w}^{(i,k)}$ using $\mathbf{w}^{(k)}$ and $\mathcal{D}^{(i)}$.
- Server aggregates $\mathbf{w}^{(1,k)}, \ldots, \mathbf{w}^{(n,k)}$ to compute $\mathbf{w}^{(k+1)}$.

Equivalent GTVMin Instance

GTVMin on a star-shaped FL network with $n+1$ nodes,

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

Node $n+1$ has trivial loss func. $L_{n+1}(\cdot) = 0$.



For sufficiently large α, we obtain $\widehat{\mathbf{w}}^{(i)} \approx \widehat{\mathbf{w}}^{(i')}$ for all i, i'.

Table of Contents

Horizontal FL

$$
\begin{array}{cc}
\begin{bmatrix}
x_1^{(1)} & x_2^{(1)} & \cdots & x_d^{(1)} & y^{(1)} \\
x_1^{(2)} & x_2^{(2)} & \cdots & x_d^{(2)} & y^{(2)} \\
x_1^{(3)} & x_2^{(3)} & \cdots & x_d^{(3)} & y^{(3)} \\
x_1^{(4)} & x_2^{(4)} & \cdots & x_d^{(4)} & y^{(4)} \\
x_1^{(5)} & x_2^{(5)} & \cdots & x_d^{(5)} & y^{(5)} \\
\vdots & \vdots & \ddots & \vdots & \vdots \\
x_1^{(m)} & x_2^{(m)} & \cdots & x_d^{(m)} & y^{(m)}
\end{bmatrix}
&
\begin{array}{l}
\mathcal{D}^{(1)} \\
\\
\\
\mathcal{D}^{(i)} \\
\\
\\
\mathcal{D}^{(n)}
\end{array}
\end{array}
$$

Local datasets are (overlapping) subsets of a global dataset.[2]

[2] How could we define similarity between such local datasets?

Examples of Horizontal FL

- **Healthcare:** Local datasets consist of patient records stored at different hospitals.
- **Finance:** Local datasets correspond to account records maintained by individual banks.
- **Condition Monitoring:** Local datasets include sensor recordings collected by car manufacturers.
- **Smart Grids:** Local datasets comprise electricity consumption data gathered by power suppliers.
- **Restaurants:** Local datasets consist of customer reviews for specific restaurants.

Horizontal FL via GTVMin

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

Local loss func. $L_i\left(\mathbf{w}^{(i)}\right) := \frac{1}{|\mathcal{D}^{(i)}|} \sum_{(\mathbf{x},y)\in\mathcal{D}^{(i)}} L\left((\mathbf{x},y), \mathbf{w}^{(i)}\right)$.

Choose edge weights $A_{i,i'}$ based on overlap $\left|\mathcal{D}^{(i)} \cap \mathcal{D}^{(i')}\right|$.

For $\alpha \rightarrow \infty$, we obtain identical local model params. which are copies of global model params. $\widehat{\mathbf{w}} = \mathbf{w}^{(i)}$ for all $i = 1, \ldots, n$.

Table of Contents

Vertical FL (VFL)

$$\begin{array}{c} \mathcal{D}^{(1)} \qquad \mathcal{D}^{(i)} \\ \underbrace{\begin{bmatrix} x_1^{(1)} & x_2^{(1)} & \cdots & x_d^{(1)} & y^{(1)} \\ x_1^{(2)} & x_2^{(2)} & \cdots & x_d^{(2)} & y^{(2)} \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ x_1^{(m)} & x_2^{(m)} & \cdots & x_d^{(m)} & y^{(m)} \end{bmatrix}}_{\mathcal{D}} \end{array}$$

Some VFL Applications

- **Healthcare**: Different healthcare providers store distinct records of the same patient.
- **Finance**: Banks, tax authorities, and other financial institutions hold different data on the same individual.
- **Government**: Social insurance, courts, and tax offices maintain separate databases for the same citizen.
- **Retail**: A single customer has multiple accounts across different loyalty programs (retailers, online marketplaces, etc.).

VFL for Linear Regression

- We want to learn params. $\mathbf{w} \in \mathbb{R}^d$ of a linear model
$$h(\mathbf{x}) = \mathbf{w}^T \mathbf{x} = \sum_{j=1}^{d} w_j x_j.$$
- Training set is distributed over devices $i = 1, \ldots, d$.
- Device i has access to labels and i-th feature. Thus, local dataset consists of $\mathbf{y}$ and $\mathbf{f}^{(i)} = \big(x_i^{(1)}, \ldots, x_i^{(m)}\big)$.
- Linear regression: $\min_{\mathbf{w} \in \mathbb{R}^d} \left\| \mathbf{y} - \sum_{j=1}^{d} \mathbf{f}^{(j)} w_j \right\|_2^2$.

VFL via GTVMin

- Let us rewrite linear regression as

$$\min_{\mathbf{w}\in\mathbb{R}^d,\mathbf{s}} \|\mathbf{y}-\mathbf{s}\|_2^2 \quad \text{s.t.} \quad \mathbf{s} = \sum_{j=1}^{d} \mathbf{f}^{(j)} w_j.$$

- Above problem is equivalent to

$$\min_{\substack{\mathbf{w}\in\mathbb{R}^d \\ \mathbf{s}^{(1)},\ldots,\mathbf{s}^{(d)}}} \sum_{i=1}^{d} \left\|\mathbf{y}-\mathbf{s}^{(i)}\right\|_2^2 \quad \text{s.t.} \quad \mathbf{s}^{(i)} = \sum_{j=1}^{d} \mathbf{f}^{(j)} w_j, \text{ at } i=1,\ldots,d.$$

- We can approximate this by an instance of GTVMin.

VFL via GTVMin (ctd.)

Consider linear regression problem

$$\min_{\substack{\mathbf{w}\in\mathbb{R}^d \\ \mathbf{s}^{(1)},\ldots,\mathbf{s}^{(d)}}} \sum_{i=1}^{d} \left\|\mathbf{y}-\mathbf{s}^{(i)}\right\|_2^2 \text{ s.t. } \mathbf{s}^{(i)} = \sum_{j=1}^{d} \mathbf{f}^{(j)} w_j, \text{ at } i=1,\ldots,d. \text{ (A)}$$

- Consider $\mathbf{s}^{(i)}$ as auxiliary local model params.
- For given $\left(w_1,\ldots,w_d\right)^T$, the constraint $\mathbf{s}^{(i)} = \sum_{j=1}^{d} \mathbf{f}^{(j)} w_j$ holds approximately by solutions of

$$\min_{\mathbf{s}^{(1)},\ldots,\mathbf{s}^{(d)}} \sum_{i=1}^{d} \left\|\mathbf{s}^{(i)} - \mathbf{f}^{(i)} w_i\right\|_2^2 + \underbrace{\alpha}_{\rightarrow\infty} \sum_{\{i,i'\}\in\mathcal{E}} \left\|\mathbf{s}^{(i)} - \mathbf{s}^{(i')}\right\|_2^2 . \text{(B)}$$

- Edges $\mathcal{E}$ must form a connected FL network (e.g., a star).

VFL via GTVMin (ctd.)

- Combining (A) with (B) yields

$$\min_{\substack{w_1,\ldots,w_d \\ \mathbf{s}^{(1)},\ldots,\mathbf{s}^{(d)}}} \sum_{i=1}^{d} \left\|\mathbf{s}^{(i)} - \mathbf{y}\right\|_2^2 + \beta \left\|\mathbf{s}^{(i)} - \mathbf{f}^{(i)} w_i\right\|_2^2 + \alpha \sum_{\{i,i'\} \in \mathcal{E}} \left\|\mathbf{s}^{(i)} - \mathbf{s}^{(i')}\right\|_2^2.$$

- GTVMin with local model params. $w_i, \mathbf{s}^{(i)}$, $i = 1, \ldots, d$.

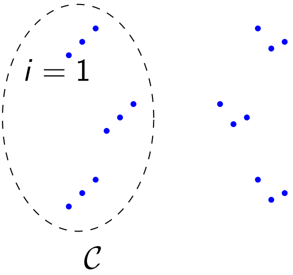
- Edges $\mathcal{E}$ must form a connected FL network (e.g., a star).

- Need sufficiently large $\beta > 0$ to ensure solutions satisfy

$$\mathbf{s}^{(i)} = \sum_{j=1}^{d} \mathbf{f}^{(j)} w_j, \text{ for each } i = 1, \ldots, d.$$

Table of Contents

Clustered FL (CFL)



A simple prob. model for FL applications is as follows.

- Local dataset $\mathcal{D}^{(i)}$ drawn i.i.d. from prob. dist. $p^{(i)}(\mathbf{x}, y)$.
- Nodes i with similar $p^{(i)}$ form a cluster $\mathcal{C} \subseteq \mathcal{V}$.
- CFL aims at learning cluster-wise model params.

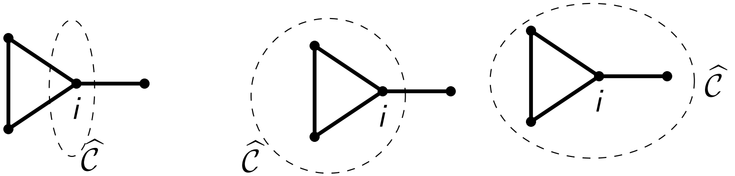
$$\mathbf{w}^{(i)} = \overline{\mathbf{w}}^{(\mathcal{C})} \text{ for all } i \in \mathcal{C}.$$

Some CFL Applications

- **Healthcare.** Personalizing models for hospitals or wearable devices based on patient population similarities.
- **Smart Homes.** Grouping devices by household behavior patterns for energy consumption prediction or automation.
- **Industrial IoT.** Adapting predictive maintenance models to clusters of machines with similar operational profiles.
- **Retail.** Tailoring recommendation engines to different store locations or customer demographics.
- **Mobility.** Building location-specific traffic prediction models for ride-sharing or delivery services.

CFL via GTVMin

Local model params. delivered by GTVMin tend to be clustered over well-connected subsets of nodes.[3]



small α moderate α large α

How to ensure that clusters of GTVMin are correct ($\widehat{\mathcal{C}} \approx \mathcal{C}$)?

[3] Y. SarcheshmehPour, Y. Tian, L. Zhang and A. Jung, "Clustered Federated Learning via Generalized Total Variation Minimization," in IEEE Transactions on Signal Processing, vol. 71, pp. 4240-4256, 2023.

Designing FL Network for CFL

- Edge weights $A_{i,i'}$ are design choice for FL methods.
- More edges $\Rightarrow$ means more computation.
- Need sufficiently many edges within a cluster $\mathcal{C}$.
- Avoid boundary edges that leave a cluster $\mathcal{C}$.

Graph Learning Methods

Data-driven (using $\mathcal{D}^{(1)}, \ldots, \mathcal{D}^{(n)}$) constructions of edges:

- Use statistical tests[1] for $p^{(i)} \stackrel{?}{=} p^{(i')}$.
- Choose $A_{i,i'}$ via (est.) KL-divergence[2] $D^{(\mathrm{KL})}\big(p^{(i)}, p^{(i')}\big)$.
- Compare gradients[3] $\nabla L_i(\mathbf{w}), \nabla L_{i'}(\mathbf{w})$.
- Compare vector representation (embedding)[4] $\mathbf{z}^{(i)}, \mathbf{z}^{(i')}$.

---

[1] Schrab et.al., MMD Aggregated Two-Sample Test, JMLR, 2023

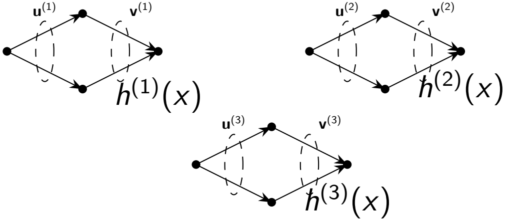
[2] Y. Bu et.al., "Estimation of KL Divergence: Optimal Minimax Rate," in IEEE Transactions on Information Theory, 2018

[3] Werner et.al.,Provably Personalized and Robust Federated Learning, TMLR, 2023.

[4] Petukhova et.al, Text Clustering with LLM Embeddings, 2024.

Table of Contents

Personalized FL



- GTVMin for ANNs $h^{(i)}$ with hidden layer.
- Local model params $\mathbf{w}^{(i)} = \left(\left(\mathbf{u}^{(i)}\right)^T, \left(\mathbf{v}^{(i)}\right)^T\right)^T$.
- Use GTV penalty $\phi = \left\|\mathbf{u}^{(i)} - \mathbf{u}^{(i')}\right\|_2^2$.

Table of Contents

Wrap Up

We discussed how FL flavours are obtained by specific design choices for FL networks and GTVMin.

- Global-Model FL equivalent to GTVMin over star graph.
- HFL: Nodes access same features of datapoints.
- VFL: Nodes access different sets of features.
- CFL: Construct edges by statistical similarities between local datasets.
- PersFL: GTV penalty ϕ uses only parts of a model (e.g., input layers).

What's Next?

The next (and final) module discusses key requirements for trustworthy FL, including robustness, privacy-protection and explainability.

We can ensure these requirements by specific design choices for FL networks and GTVMin.

Further Resources

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