

CS-E4740 - Federated Learning

FL Design Principle

Assoc. Prof. Alexander Jung

Spring 2025

YouTube



LinkedIn



GitHub



Table of Contents

Formulating FL as Optimization

Computational Aspects

Statistical Aspects

Interpretations

Table of Contents

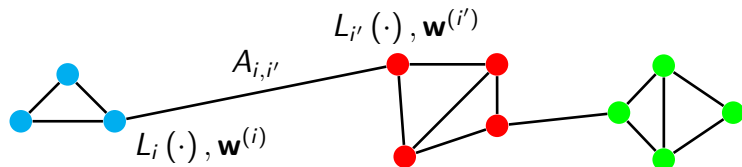
Formulating FL as Optimization

Computational Aspects

Statistical Aspects

Interpretations

An FL Network



- ▶ FL network consists of devices, denoted $i=1, \dots, n$
- ▶ Some i, i' connected by an edge with weight $A_{i,i'} > 0$
- ▶ Device i learns model params. $\mathbf{w}^{(i)} \in \mathbb{R}^d$
- ▶ Usefulness of $\mathbf{w}^{(i)}$ measured by some local loss, e.g.,

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

FL via Regularization

- ▶ Consider a FL network with nodes $i = 1, \dots, n$.
- ▶ Each node carries a linear model $h^{(\mathbf{w}^{(i)})}(\mathbf{x}) := \mathbf{x}^T \mathbf{w}^{(i)}$.
- ▶ Each node carries m_i labelled data points.
- ▶ Node-wise ML fails if $m_i \ll d$ (overfitting).

Idea:

Use the neighbours $\mathcal{N}^{(i)} := \{i' : \{i, i'\} \in \mathcal{E}\}$ to regularize!

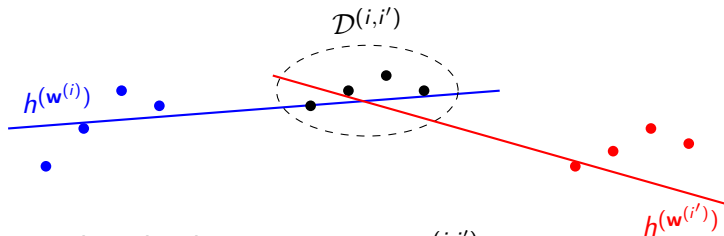
FL via Regularization (ctd.)

As for basic ML, regularization can be done either via

- ▶ **Data augmentation** using data from the neighbours.
- ▶ **Prune local models** by requiring them to agree across edges.
- ▶ **Add a penalty term** to the local loss function.

Building a Penalty Across Edges

- ▶ Consider two nodes i, i' with local datasets $\mathcal{D}^{(i)}, \mathcal{D}^{(i')}$.
- ▶ Assume there is a non-empty overlap $\mathcal{D}^{(i)} \cap \mathcal{D}^{(i')}$.



We penalize the disagreement on $\mathcal{D}(i, i')$:

$$\begin{aligned} \sum_{\mathbf{x} \in \mathcal{D}(i, i')} (h(\mathbf{w}^{(i)})(\mathbf{x}) - h(\mathbf{w}^{(i')})(\mathbf{x}))^2 &= \sum_{\mathbf{x} \in \mathcal{D}(i, i')} (\mathbf{x}^T \mathbf{w}^{(i)} - \mathbf{x}^T \mathbf{w}^{(i')})^2 \\ &= (\mathbf{w}^{(i)} - \mathbf{w}^{(i')})^T \left[\sum_{\mathbf{x} \in \mathcal{D}(i, i')} \mathbf{x}^T \mathbf{x} \right] (\mathbf{w}^{(i)} - \mathbf{w}^{(i')}). \end{aligned}$$

Generalized TV Minimization (GTVMin)

Learn model params. $\widehat{\mathbf{w}}^{(i)}$ by balancing local loss and GTV

$$\min_{\mathbf{w}^{(1)}, \dots, \mathbf{w}^{(n)} \in \mathbb{R}^d} \sum_{i=1}^n \left[L_i(\mathbf{w}^{(i)}) + \alpha \sum_{\{i, i'\} \in \mathcal{E}} A_{i, i'} \phi(\mathbf{w}^{(i)} - \mathbf{w}^{(i')}) \right]$$

- ▶ Penalty function $\phi(\mathbf{u})$ is a design choice.
- ▶ Previous slide used $\phi(\mathbf{u}) = \mathbf{u}^T \mathbf{Q} \mathbf{u}$ with $\mathbf{Q} := \sum_{\mathbf{x} \in \mathcal{D}(i, i')} \mathbf{x} \mathbf{x}^T$
- ▶ Our focus is on the choice $\phi(\mathbf{u}) := \|\mathbf{u}\|_2^2$.
- ▶ Another popular choice is $\phi(\mathbf{u}) := \|\mathbf{u}\|$.¹

¹Y. SarcheshmehPour, et.al, "Clustered Federated Learning via Generalized Total Variation Minimization," in IEEE Trans. Sig. Proc, 2023, doi: 10.1109/TSP.2023.3322848.

Model-Agnostic GTVMin

Replacing $\phi(\mathbf{w}^{(i)} - \mathbf{w}^{(i')})$ with the disagreement measure

$$D(h^{(i)}, h^{(i')}) := \sum_{\mathbf{x} \in \mathcal{D}^{(i, i')}} (h^{(i)}(\mathbf{x}) - h^{(i')}(\mathbf{x}))^2$$

yields a model-agnostic generalization of GTVMin

$$\min_{\substack{h^{(i)} \in \mathcal{H}^{(i)} \\ i \in \mathcal{V}}} \sum_{i \in \mathcal{V}} \left[L_i(h^{(i)}) + \alpha \sum_{\{i, i'\} \in \mathcal{E}} A_{i, i'} D(h^{(i)}, h^{(i')}) \right].$$

This allows for VERY heterogeneous FL networks, e.g., $\mathcal{H}^{(1)} = \text{lin.model}$, $\mathcal{H}^{(2)} = \text{LLM}$, $\mathcal{H}^{(3)} = \text{decision tree}$.

Table of Contents

Formulating FL as Optimization

Computational Aspects

Statistical Aspects

Interpretations

Computational Aspects

$$\min_{\mathbf{w}^{(1)}, \dots, \mathbf{w}^{(n)} \in \mathbb{R}^d} \sum_{i=1}^n \left[L_i(\mathbf{w}^{(i)}) + \alpha \sum_{\{i, i'\} \in \mathcal{E}} A_{i, i'} \phi(\mathbf{w}^{(i)} - \mathbf{w}^{(i')}) \right]$$

- ▶ How can we solve it efficiently over an FL network?
- ▶ How much compute/comm. is needed at least?
- ▶ What is the effect of different choices for the edges \mathcal{E} , loss funcs. $L_i(\cdot)$, and GTV penalty ϕ ?

Computational Aspects - Smooth GTVmin

Consider a GTVMin instance

$$\min_{\mathbf{w}^{(1)}, \dots, \mathbf{w}^{(n)} \in \mathbb{R}^d} \sum_{i=1}^n \left[L_i(\mathbf{w}^{(i)}) + \alpha \sum_{\{i, i'\} \in \mathcal{E}} A_{i, i'} \left\| \mathbf{w}^{(i)} - \mathbf{w}^{(i')} \right\|_2^2 \right]$$

with a smooth (differentiable) $L_i(\cdot)$.

If we use (distributed) gradient descent to solve GTVMin,

- ▶ how many iterations should we run?
- ▶ what is a good choice for the learning rate?
- ▶ how to communicate gradients over comm. links?

Characterizing GTVMin Solutions

- ▶ Consider GTVMin solution $\hat{\mathbf{w}}^{(i)} \in \mathbb{R}^d$, for $i = 1, \dots, n$.
- ▶ We stack them into a long vector

$$\hat{\mathbf{w}} := (\hat{\mathbf{w}}^{(1)}, \dots, \hat{\mathbf{w}}^{(n)})^T \in \mathbb{R}^{dn}.$$

- ▶ We characterize the solutions as a fixed-point of some \mathcal{F} ,

$$\hat{\mathbf{w}} \text{ solves GTVMin} \Leftrightarrow \hat{\mathbf{w}} = \mathcal{F}\hat{\mathbf{w}}$$

- ▶ The operator \mathcal{F} is not unique (design choice!).

Convex and Smooth GTVMin

Consider GTVMin with a smooth and convex $L_i(\cdot)$,

$$\min_{\mathbf{w}^{(1)}, \dots, \mathbf{w}^{(n)} \in \mathbb{R}^d} \sum_{i=1}^n \left[L_i(\mathbf{w}^{(i)}) + \alpha \sum_{\{i, i'\} \in \mathcal{E}} A_{i, i'} \left\| \mathbf{w}^{(i)} - \mathbf{w}^{(i')} \right\|_2^2 \right] \quad (1)$$

$$\hat{\mathbf{w}} \text{ solves (1)} \Leftrightarrow \hat{\mathbf{w}} = \mathcal{F} \hat{\mathbf{w}}$$

with \mathcal{F} mapping $\mathbf{u} = (\mathbf{u}^{(1)}, \dots, \mathbf{u}^{(n)})^T$ to $\mathbf{v} = (\mathbf{v}^{(1)}, \dots, \mathbf{v}^{(n)})^T$,

$$\mathbf{v}^{(i)} = \mathbf{u}^{(i)} - \eta \left[\nabla L_i(\mathbf{u}^{(i)}) + 2\alpha \sum_{i' \in \mathcal{N}^{(i)}} A_{i, i'} (\mathbf{u}^{(i)} - \mathbf{u}^{(i')}) \right].$$

Different choices for “step-size” $\eta > 0$ yield different \mathcal{F} .

Table of Contents

Formulating FL as Optimization

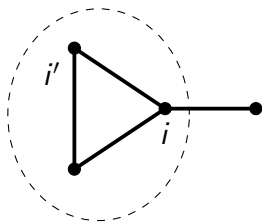
Computational Aspects

Statistical Aspects

Interpretations

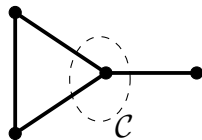
Statistical Aspects

- ▶ GTVMin solution yields model params. $\hat{\mathbf{w}}^{(i)}$, $i = 1, \dots, n$
- ▶ How useful are these model params. ?
- ▶ The local loss $L_i(\hat{\mathbf{w}}^{(i)})$ can be misleading (why?)
- ▶ Better to use aggregate $\sum_{i \in \mathcal{C}^{(i)}} L_i(\hat{\mathbf{w}}^{(i)})$, with cluster

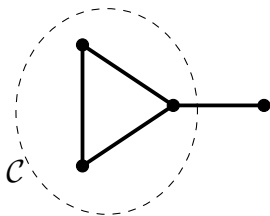


$$\mathcal{C}^{(i)} := \{i' : \hat{\mathbf{w}}^{(i)} \approx \hat{\mathbf{w}}^{(i')}\}.$$

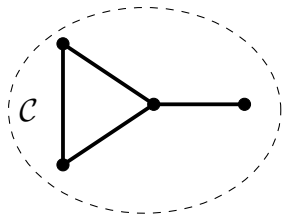
Clustering of GTVMin



small α



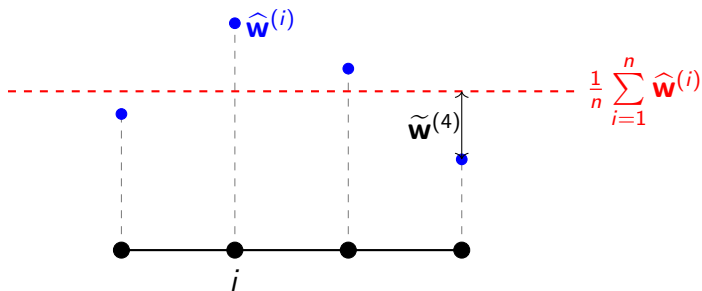
moderate α



large α

Analysis of Clustering - Assumptions

- ▶ Consider a connected FL network \mathcal{G} with $\lambda_2 > 0$.
- ▶ Assume loss funcs. satisfy $\min_{\mathbf{v} \in \mathbb{R}^d} \sum_{i=1}^n L_i(\mathbf{v}) \leq \varepsilon$
- ▶ Use GTVMin to learn local params. $\widehat{\mathbf{w}}^{(i)}$.
- ▶ Define the variation $\widetilde{\mathbf{w}}^{(i)} := \widehat{\mathbf{w}}^{(i)} - \frac{1}{n} \sum_{i=1}^n \widehat{\mathbf{w}}^{(i)}$.



Analysis of Clustering - Upper Bound

The variation $\tilde{\mathbf{w}}^{(i)}$ is upper bounded as

$$\sum_{i=1}^n \|\tilde{\mathbf{w}}^{(i)}\|_2^2 \leq \frac{\varepsilon}{\alpha \lambda_2}.$$

This bound involves the

- ▶ connectivity of FL network (via λ_2),
- ▶ the properties of local loss functions (via ε), and
- ▶ the GTVMin parameter α .

A large $\alpha \lambda_2$ results in nearly identical local params. $\tilde{\mathbf{w}}^{(i)} \approx \mathbf{0}$.

Table of Contents

Formulating FL as Optimization

Computational Aspects

Statistical Aspects

Interpretations

Interpretations

We next discuss some interpretations of GTVMin

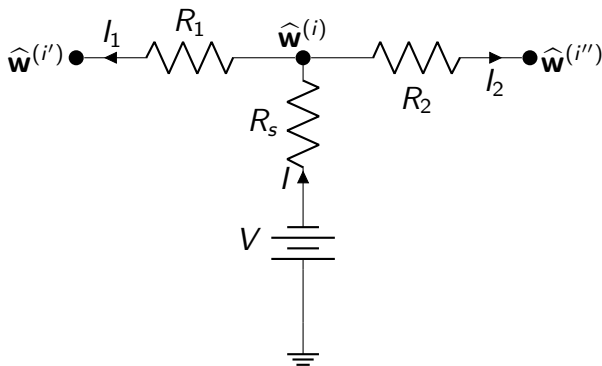
$$\min_{\mathbf{w}^{(1)}, \dots, \mathbf{w}^{(n)} \in \mathbb{R}^d} \sum_{i=1}^n \left[L_i(\mathbf{w}^{(i)}) + \alpha \sum_{\{i, i'\} \in \mathcal{E}} A_{i, i'} \left\| \mathbf{w}^{(i)} - \mathbf{w}^{(i')} \right\|_2^2 \right]$$

for some FL network with weighted undirected graph \mathcal{G} and smooth and convex loss func. $L_i(\mathbf{w}^{(i)})$.

We assume that there exists a solution $\widehat{\mathbf{w}}^{(1)}, \dots, \widehat{\mathbf{w}}^{(n)}$.

Electronic Circuit

Consider a node i with neighbours $\mathcal{N}^{(i)} = \{i', i''\}$.

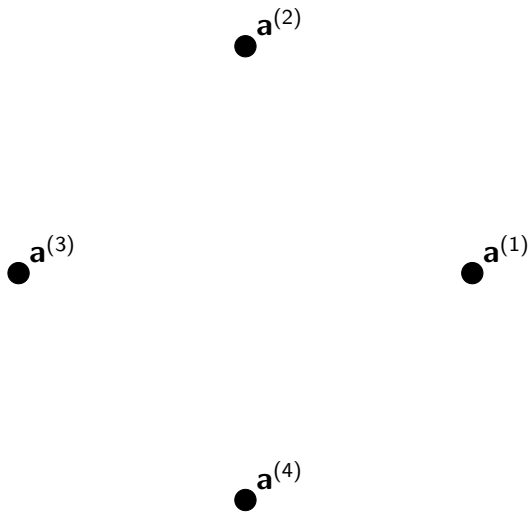


$$\underbrace{-\nabla L_i}_{\textcolor{red}{I}}(\hat{\mathbf{w}}^{(i)}) = \left[\underbrace{A_{i,i'}(\hat{\mathbf{w}}^{(i)} - \hat{\mathbf{w}}^{(i')})}_{\textcolor{red}{I_1}} + \underbrace{(A_{i,i''}\hat{\mathbf{w}}^{(i)} - \hat{\mathbf{w}}^{(i'')})}_{\textcolor{red}{I_2}} \right]$$

Vector-Valued Flows

Locally Weighted Learning

Generalized Convex Clustering



What's Next?

The next module applies optimization methods to solve GTVMin.

We can implement these methods as message passing over the edges of an FL network.

Further Resources

- ▶ **YouTube:** [@alexjung111](#)
- ▶ **LinkedIn:** [Alexander Jung](#)
- ▶ **GitHub:** [alexjungaalto](#)

