CS-E4740 - Federated Learning

L2 - FL Design Principle

Dipl.-Ing. Dr.techn. Alexander Jung
Assoc. Prof. for Machine Learning, Aalto University

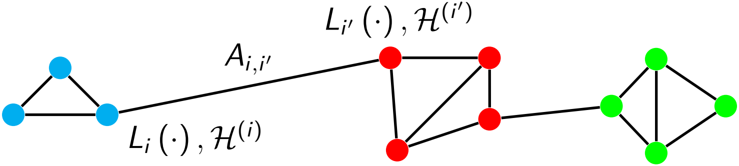
Spring 2026

Calendar **Glossary** **Book** **GitHub**

Table of Contents

Table of Contents

Some federated learning network (FL network)



- devices represented by nodes $i = 1, \ldots, n$
- some i, i' connected by an edge with weight $A_{i,i'} > 0$
- device i learns hypothesis $h^{(i)} \in \mathcal{H}^{(i)}$
- usefulness of $h^{(i)}$ measured by local loss $L_i(\cdot)$

FL via Regularization

consider an FL network where

- each node trains a linear model $h^{(\mathbf{w}^{(i)})}(\mathbf{x}) := \mathbf{x}^T \mathbf{w}^{(i)}$
- each node carries m_i labelled data points
- each data point is characterized by d features

$\Longrightarrow$ node-wise training fails if $m_i \ll d$ (overfitting)

Idea:

use the neighbors $\mathcal{N}^{(i)} := \{i' : \{i, i'\} \in \mathcal{E}\}$ for regularization!

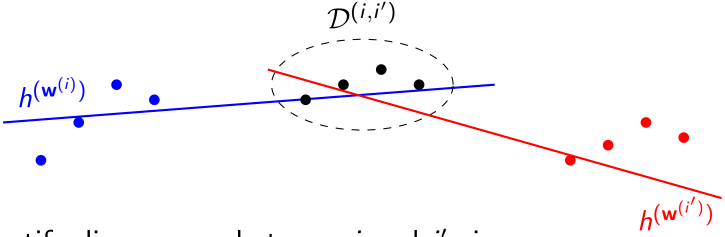
FL via Regularization (ctd.)

regularization can be done either via

- **data augmentation** using information from neighbors,
- **pruning local models** by requiring agreement across edges,
- **adding a penalty term** to the local loss function

Building a Penalty Across Edges

- consider two connected nodes i, i' with datasets $\mathcal{D}^{(i)}, \mathcal{D}^{(i')}$
- assume a non-empty overlap $\mathcal{D}^{(i,i')} = \mathcal{D}^{(i)} \cap \mathcal{D}^{(i')}$



quantify discrepancy between i and i' via

$$\sum_{\mathbf{x} \in \mathcal{D}^{(i,i')}} \left(h^{(\mathbf{w}^{(i)})}(\mathbf{x}) - h^{(\mathbf{w}^{(i')})}(\mathbf{x})\right)^2 = \sum_{\mathbf{x} \in \mathcal{D}^{(i,i')}} \left(\mathbf{x}^T \mathbf{w}^{(i)} - \mathbf{x}^T \mathbf{w}^{(i')}\right)^2$$

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

Discrepancy Measure for Parametric models

use "norm-like" function $\phi(\mathbf{w}^{(i)} - \mathbf{w}^{(i')})$ of difference between model parameter at two nodes i, i'

- $\phi(\mathbf{u}) = \|\mathbf{u}\|_2^2$, or
- $\phi(\mathbf{u}) = \mathbf{u}^T\mathbf{Q}\mathbf{u}$ with positive semi-definite (psd) $\mathbf{Q}$, or
- $\phi(\mathbf{u}) = \|\mathbf{u}\|_1$, or
- $\phi(\mathbf{u}) = \|\mathbf{u}\|$, or
- $\ldots$

Discrepancy Measure for Federated GMM

- node i carries Gaussian mixture model (GMM) with model parameters $\mathbf{w}^{(i)}$
- measure discrepancy via Kullback–Leibler divergence (KL divergence)
$$\frac{1}{2}\Big(D^{(\mathrm{KL})}\big(\mathbf{w}^{(i)},\mathbf{w}^{(i')}\big)+D^{(\mathrm{KL})}\big(\mathbf{w}^{(i')},\mathbf{w}^{(i)}\big)\Big)$$
- useful for federated soft clustering

Discrepancy Measure for Federated k-Means

- node i carries local cluster centroids $\mathbf{w}^{(i,1)}, \ldots, \mathbf{w}^{(i,k)}$
- define discrepancy via

$$\sum_{c \in [k]} \min_{c' \in [k]} \left\| \mathbf{w}^{(i,c)} - \mathbf{w}^{(i',c')} \right\|_2^2 + \sum_{c \in [k]} \min_{c' \in [k]} \left\| \mathbf{w}^{(i',c)} - \mathbf{w}^{(i,c')} \right\|_2^2$$

- first term can only vanish if

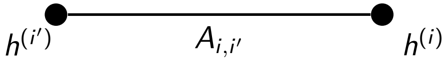
$$\{\mathbf{w}^{(i,c)}\}_{c=1}^{k} \subseteq \{\mathbf{w}^{(i',c)}\}_{c=1}^{k}$$

- second term can only vanish if

$$\{\mathbf{w}^{(i',c)}\}_{c=1}^{k} \subseteq \{\mathbf{w}^{(i,c)}\}_{c=1}^{k}$$

GTV

consider some discrepancy measure $d^{(h^{(i)},h^{(i')})}$ for hypotheses $h^{(i)}, h^{(i')}$ at two connected nodes



we then define the generalized total variation (GTV),

$$\sum_{\{i,i'\}\in\mathcal{E}} A_{i,i'} d^{(h^{(i)},h^{(i')})}$$

by summing the scaled discrepancy measures over all edges

GTVMin

generalized total variation minimization (GTVMin) balances local losses with GTV:

$$\min_{\substack{h^{(1)} \in \mathcal{H}^{(1)} \\ \cdots \\ h^{(n)} \in \mathcal{H}^{(n)}}} \sum_{i=1}^{n} L_i \left(h^{(i)} \right) + \alpha \sum_{\{i,i'\} \in \mathcal{E}} A_{i,i'} d^{(h^{(i)}, h^{(i')})}$$

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

GTVMin - Extreme Cases

GTVMin balances local losses with GTV:

$$\min_{\substack{h^{(1)} \in \mathcal{H}^{(1)} \\ \cdots \\ h^{(n)} \in \mathcal{H}^{(n)}}} \sum_{i=1}^{n} L_i\left(h^{(i)}\right) + \alpha \sum_{\{i,i'\} \in \mathcal{E}} A_{i,i'} d^{(h^{(i)}, h^{(i')})}$$

different regimes depending on regularization strength α

- $\alpha = 0$: GTVMin splits into independent empirical risk minimization (ERM) at each i
- $\alpha \rightarrow \infty$: GTVMin becomes single global ERM with each node i holding a copy of the global learnt hypothesis
- for intermediate α, GTVMin becomes cluster-wise ERM

GTVMin for Parametric models

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

note that this special case of GTVMin requires local models to be parametrized by the same Euclidean space $\mathbb{R}^d$

GTVMin for Federated Linear regression

consider FL network where each node trains a linear model using local dataset $\left(\mathbf{x}^{(1)}, y^{(1)}\right), \ldots, \left(\mathbf{x}^{(m_i)}, y^{(m_i)}\right)$

federated linear regression via GTVMin

$$\min_{\substack{\mathbf{w}^{(1)} \in \mathbb{R}^d \\ \ldots \\ \mathbf{w}^{(n)} \in \mathbb{R}^d}} \sum_{i=1}^{n} \frac{1}{m_i} \left\| \mathbf{y}^{(i)} - \mathbf{X}^{(i)} \mathbf{w}^{(i)} \right\|_2^2 + \alpha \sum_{\{i,i'\} \in \mathcal{E}} A_{i,i'} \left\| \mathbf{w}^{(i)} - \mathbf{w}^{(i')} \right\|_2^2$$

here, we used the local feature matrix $\mathbf{X}^{(i)} := \left(\mathbf{x}^{(1)}, \ldots, \mathbf{x}^{(n)}\right)^T$ and the local label vector $\mathbf{y}^{(i)} := \left(y^{(1)}, \ldots, y^{(n)}\right)^T$

Table of Contents

Interpretations

we next discuss some interpretations of GTVMin

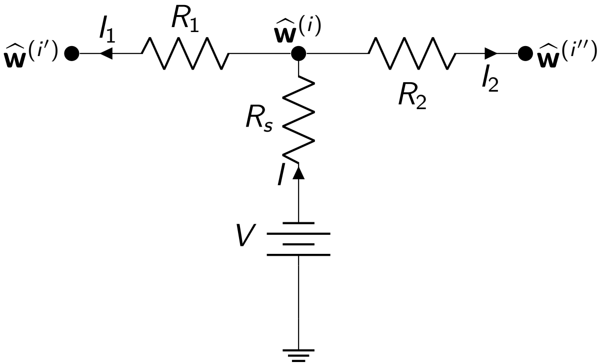
$$\min_{\mathbf{w}^{(1)},\ldots,\mathbf{w}^{(n)} \in \mathbb{R}^d} \sum_{i=1}^{n} L_i\left(\mathbf{w}^{(i)}\right) + \alpha \sum_{\{i,i'\}\in\mathcal{E}} A_{i,i'} \left\|\mathbf{w}^{(i)} - \mathbf{w}^{(i')}\right\|_2^2$$

with smooth and convex loss functions $L_i\left(\mathbf{w}^{(i)}\right)$

we assume that there exists a solution $\widehat{\mathbf{w}}^{(1)},\ldots,\widehat{\mathbf{w}}^{(n)}$ (do we really need this assumption?)

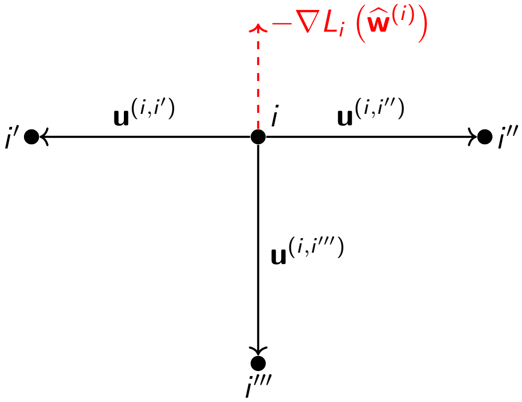
Electronic Circuit

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



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

Vector-Valued Flows

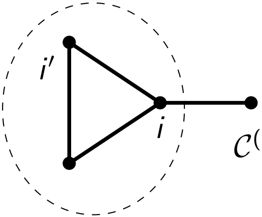


Vector-valued flow $\mathbf{u}^{(i,i')} := \nabla\phi\big(\mathbf{u}\big)\Big|_{\mathbf{u}=\widehat{\mathbf{w}}^{(i)}-\widehat{\mathbf{w}}^{(i')}}$.

AJ, "On the Duality Between Network Flows and Network Lasso," in IEEE Signal Processing Letters, 2020.

Locally Weighted Learning

GTVMin delivers model parameters $\widehat{\mathbf{w}}^{(i)}$ that form clusters



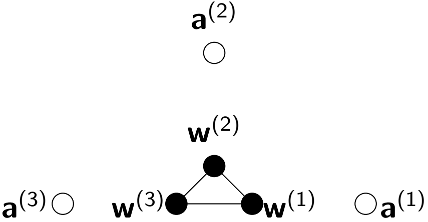
$$\mathcal{C}^{(i)} := \left\{ i' \in \mathcal{V} : \widehat{\mathbf{w}}^{(i)} \approx \widehat{\mathbf{w}}^{(i')} \right\}.$$

for node i, GTVMin is the same as locally weighted learning

$$\min_{\mathbf{w}^{(i)} \in \mathbb{R}^d} \sum_{i'=1}^{n} L_{i'}\left(\mathbf{w}^{(i)}\right) \rho_{i'} \text{ with } \rho_{i'} = \begin{cases} 1 & \text{if } i' \in \mathcal{C}^{(i)} \\ 0 & \text{otherwise.} \end{cases}$$

C. G. Atkeson, S. A. Schaal and Andrew W, Moore, Locally Weighted Learning, AI Review,Volume 11, Pages 11-73 (Kluwer Publishers) 1997.

Generalized Convex Clustering



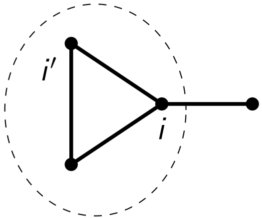
$$\min_{\mathbf{w}^{(1)},\ldots,\mathbf{w}^{(n)}} \sum_{i=1}^{n} \left\| \mathbf{w}^{(i)} - \mathbf{a}^{(i)} \right\|_2^2 + \alpha \sum_{i,i' \in \mathcal{V}} \left\| \mathbf{w}^{(i)} - \mathbf{w}^{(i')} \right\|_2 .$$

D. Sun, et.al, Convex Clustering: Model, Theoretical Guarantee and Efficient Algorithm, JMLR, 2021.

Table of Contents

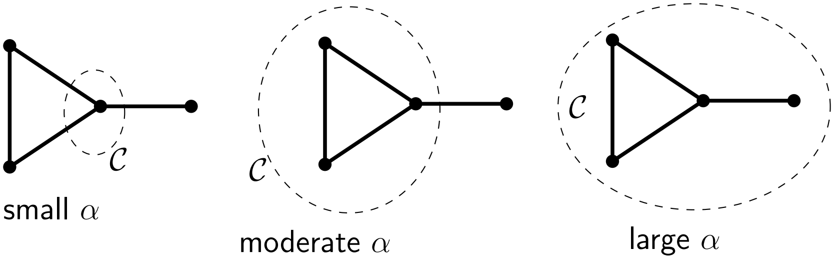
Statistical Aspects

- GTVMin solution yields model parameters $\widehat{\mathbf{w}}^{(i)}$, $i = 1, \ldots, n$
- how useful are these?
- loss value $L_i\left(\widehat{\mathbf{w}}^{(i)}\right)$ can be misleading (why?)
- better to use aggregate loss $\sum_{i \in \mathcal{C}^{(i)}} L_i\left(\widehat{\mathbf{w}}^{(i)}\right)$, with cluster



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

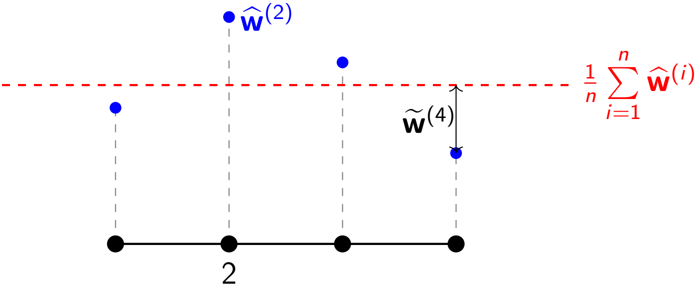
Clustering of GTVMin



Y. SarcheshmehPour, Y. Tian, L. Zhang and A. Jung, "Clustered Federated Learning via Generalized Total Variation Minimization," in IEEE Transactions on Signal Processing, 2023.

Analysis of GTVMin - Assumptions

- consider a connected FL network with $\lambda_2 > 0$
- assume loss functions satisfy $\min_{\mathbf{v} \in \mathbb{R}^d} \sum_{i=1}^{n} L_i(\mathbf{v}) \leq \varepsilon$
- use GTVMin to learn model parameters $\widehat{\mathbf{w}}^{(i)}$
- define variation $\widetilde{\mathbf{w}}^{(i)} := \widehat{\mathbf{w}}^{(i)} - \frac{1}{n} \sum_{i=1}^{n} \widehat{\mathbf{w}}^{(i)}$



Analysis of GTVMin - Upper Bound

the variation $\widetilde{\mathbf{w}}^{(i)}$ is upper bounded as

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

this bound involves the

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

$\boxed{\text{Large } \alpha\lambda_2 \text{ enforces similar model parameters } \widehat{\mathbf{w}}^{(i)}.}$

Table of Contents

Computational Aspects

$$\min_{\substack{h^{(1)} \in \mathcal{H}^{(1)} \\ \cdots \\ h^{(n)} \in \mathcal{H}^{(n)}}} \sum_{i=1}^{n} L_i\left(h^{(i)}\right) + \alpha \sum_{\{i,i'\} \in \mathcal{E}} A_{i,i'} d^{(h^{(i)},h^{(i')})}$$

- how can we solve it efficiently over an FL network?
- how much compute/communication is needed at least?
- what is the effect of edges $\mathcal{E}$, loss functions $L_i\left(\cdot\right)$, and discrepancy measure $d^{(\cdot,\cdot)}$?

Fixed-Point Characterization

consider hypotheses $\widehat{h}^{(i)}$, for $i=1,\ldots,n$, that solve

$$\min_{h^{(1)},\ldots,h^{(n)}} \sum_{i=1}^{n} L_i\left(h^{(i)}\right) + \alpha \sum_{\{i,i'\}\in\mathcal{E}} A_{i,i'} d^{(h^{(i)},h^{(i')})}$$

trivially, for each node i, solution $\widehat{h}^{(i)}$ must satisfy

$$\widehat{h}^{(i)} \in \underbrace{\operatorname*{arg\,min}_{h\in\mathcal{H}^{(i)}} L_i\left(h\right) + \alpha \sum_{i'\in\mathcal{N}^{(i)}} A_{i,i'} d^{(h,\widehat{h}^{(i')})}}_{\mathcal{F}^{(i)}\left(\widehat{h}^{(1)},\ldots,\widehat{h}^{(n)}\right)}$$

(when is this necessary condition also sufficient?)

Fixed-Point Characterization (ctd.)

necessary condition for GTVMin solution:

$$\widehat{h}^{(i)} \in \underbrace{\operatorname*{arg\,min}_{h \in \mathcal{H}^{(i)}} L_i\left(h\right) + \alpha \sum_{i' \in \mathcal{N}^{(i)}} A_{i,i'} d^{(h,\widehat{h}^{(i')})}}_{\mathcal{F}^{(i)}\left(\widehat{h}^{(1)},\ldots,\widehat{h}^{(n)}\right)}$$

- when is this necessary condition also sufficient?
- nothing but regularization of $L_i\left(h\right)$ using penalty term

$$\alpha \sum_{i' \in \mathcal{N}^{(i)}} A_{i,i'} d^{(h,\widehat{h}^{(i')})}$$

Fixed-Point Characterization for Parametric Models

consider GTVMin with a smooth and convex $L_i\left(\cdot\right)$,

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

$$\boxed{\widehat{\mathbf{w}} \text{ solves } (1) \Leftrightarrow \widehat{\mathbf{w}} = \mathcal{F}^{(\eta)}\widehat{\mathbf{w}}}$$

$\mathcal{F}^{(\eta)}$ maps $\mathbf{u}=\left(\mathbf{u}^{(1)},\ldots,\mathbf{u}^{(n)}\right)^T$ to $\mathbf{v}=\left(\mathbf{v}^{(1)},\ldots,\mathbf{v}^{(n)}\right)^T$,

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

different learning rate $\eta > 0$ yields different $\mathcal{F}^{(\eta)}$

Fixed-Point Iterations

Q: how to compute a fixed point $\widehat{\mathbf{w}}$ of $\mathcal{F}$?

A: start with initial guess $\widehat{\mathbf{w}}^{(0)}$ and iterate

$$\widehat{\mathbf{w}}^{(t)} = \mathcal{F}\widehat{\mathbf{w}}^{(t-1)}, \text{ for } t = 1, 2, \ldots .$$

if $\mathcal{F}$ is **firmly non-expansive operator**, $\lim_{t \to \infty} \widehat{\mathbf{w}}^{(t)} = \widehat{\mathbf{w}}$

if $\mathcal{F}$ is a **contractive operator** with constant $\kappa < 1$,

$$\left\| \widehat{\mathbf{w}}^{(t)} - \widehat{\mathbf{w}} \right\|_2 \leq \kappa^t \left\| \widehat{\mathbf{w}}^{(0)} - \widehat{\mathbf{w}} \right\|_2$$

H. Bauschke, P. Combettes, "Convex Analysis and Monotone Operator Theory in Hilbert Spaces," Springer, 2017.

GD as Fixed-point iteration

gradient descent (GD) for smooth and convex $f(\mathbf{w})$,

$$\mathbf{w}^{(t)} = \mathbf{w}^{(t-1)} - \eta \nabla f\left(\mathbf{w}^{(t-1)}\right)$$

is a fixed-point iteration with $\mathcal{F}^{(\eta)} : \mathbf{w} \mapsto \mathbf{w} - \eta \nabla f(\mathbf{w})$

- in general, $\mathcal{F}^{(\eta)}$ is neither a firmly non-expansive operator nor a contractive operator
- convergence can be ensured if η is sufficiently small
- e.g., use learning rate $\eta_t = 1/t$ for smooth $f\left(\mathbf{w}\right)$

FL Algorithm as Fixed-point iteration

turn GTVMin optimality condition into FL algorithm

$$\widehat{h}^{(i,t+1)} = \mathcal{F}^{(i)}\big(\widehat{h}^{(1,t)}, \ldots, \widehat{h}^{(n,t)}\big) \text{ for } t = 0, 1, \ldots$$

- the node-wise update operator $\mathcal{F}^{(i)}$ depends on
  - local loss function $L_i(\cdot)$
  - neighbors $\mathcal{N}^{(i)}$ and edge weights
  - discrepancy measure $d^{(\cdot,\cdot)}$
- design choices ensure that iteration solves GTVMin
- update is an instance of regularized empirical risk minimization (RERM)

Online FL Algorithms

a more general form of FL algorithms is

$$\widehat{h}^{(i,t+1)} = \mathcal{F}^{(i,t)}\big(\widehat{h}^{(1,t)}, \ldots, \widehat{h}^{(n,t)}\big)$$

with time-varying update operators $\mathcal{F}^{(i,t)}$

- allows for data points arriving continuously
- relevant for online learning or reinforcement learning (RL)
- $\mathcal{F}^{(i,t)}$ depends on data arriving at node i and time t

Table of Contents

The FL Workflow of this Course

this lecture revolved around the following workflow

1. formulate FL application as GTVMin
2. GTVMin solutions are trained local models
3. find a fixed-point characterization of GTVMin solutions
4. solve GTVMin via fixed-point iteration

Two Research Question

two core questions:

- where do fixed-point iteration converge to ?
- how to efficiently compute fixed-point iteration ?

What's Next?

the next lecture discusses the design and study of fixed-point iterations for solving GTVMin

Further Resources

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