Training Personalized Models via Total Variation Minimization

Alexander Jung

Assoc. Prof. (tenured), Aalto University

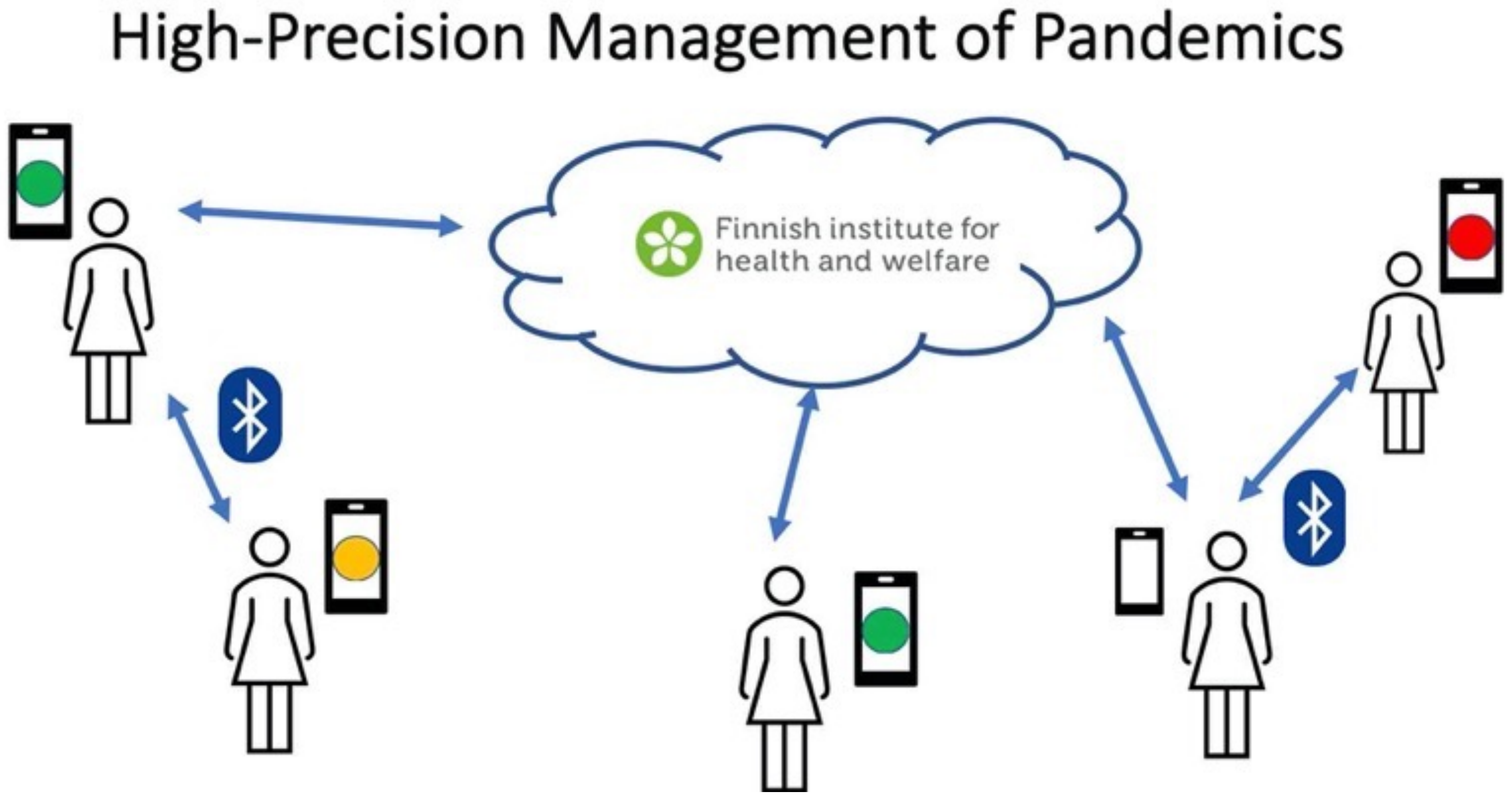
linkedin.com/in/aljung

@alexjung111

About me.

- 2012: Phd in Electrical Engineering, TU Wien
- 2012 – 2015: Post-Doc TUW, ETH Zurich
- 2015 - : Prof. for ML @ Aalto CS A! Aalto University
- 2019- : Instructor at Aalto Executive Education
- 2022 -: Principal AI Scientist at SILO AI
- 2024- : Advisor for ANKERIAS Global chronic pain won't wait, and neither will we.

RA1: Federated Learning.

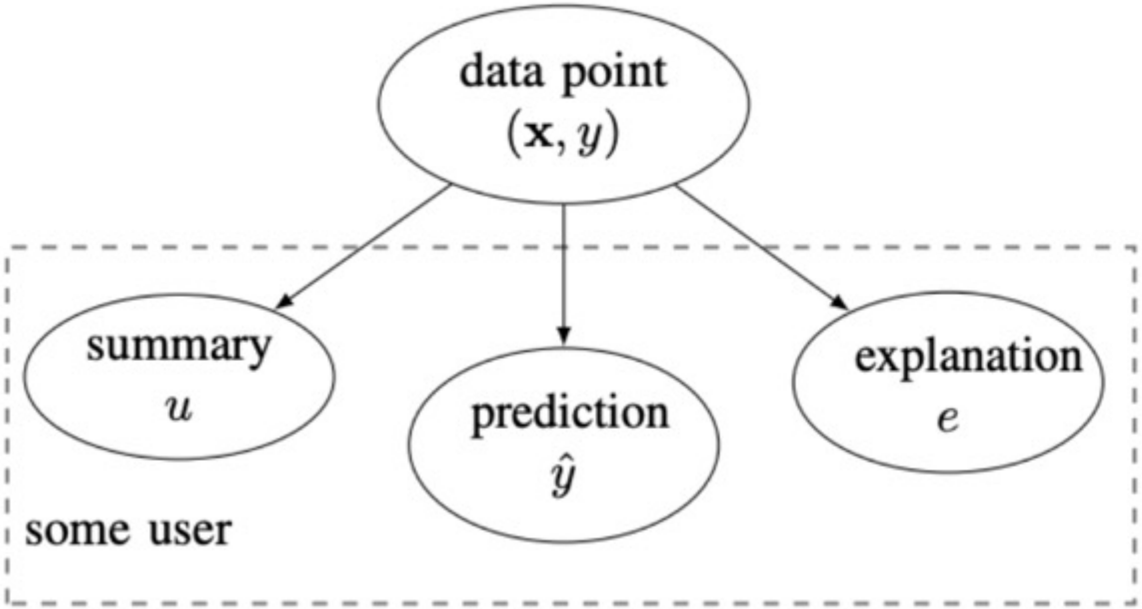


Y. Sarcheshmehpour, M Leinonen and AJ, “Federated Learning From Big Data Over Networks”, IEEE ICASSP, 2021.

AJ, “Networked Exponential Families for Big Data Over Networks,” in IEEE Access, 2020, doi: 10.1109/ACCESS.2020.3033817.

AJ, N. Tran, “Localized Linear Regression in Networked Data,” in IEEE SPL, 2019, doi: 10.1109/LSP.2019.2918933.

RA2: Explainable Machine Learning.



explanation can be:
- relevant example of training set
- subset of features
- counterfactuals
- a free text explanation
- court sentence

AJ, "Explainable Empirical Risk Minimization", arXiv eprint, 2020. weblink

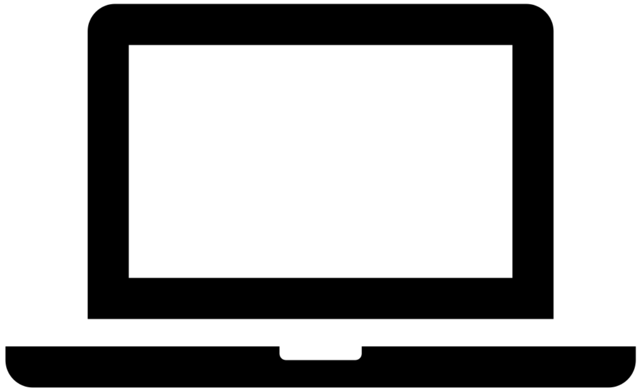
AJ Jung and P. H. J. Nardelli, "An Information-Theoretic Approach to Personalized Explainable Machine Learning," in IEEE SPL, 2020, doi: 10.1109/LSP.2020.2993176.

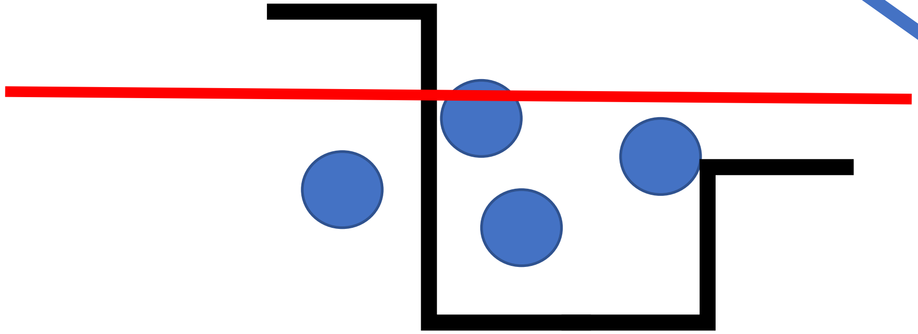
How to train (in a trustworthy fashion) a personalized model by leveraging other’s data?

Plain Old Machine Learning

```python
# We only take the two corresponding features
X = iris.data[:, pair]
y = iris.target

# Train
clf = DecisionTreeClassifier().fit(X, y)
```

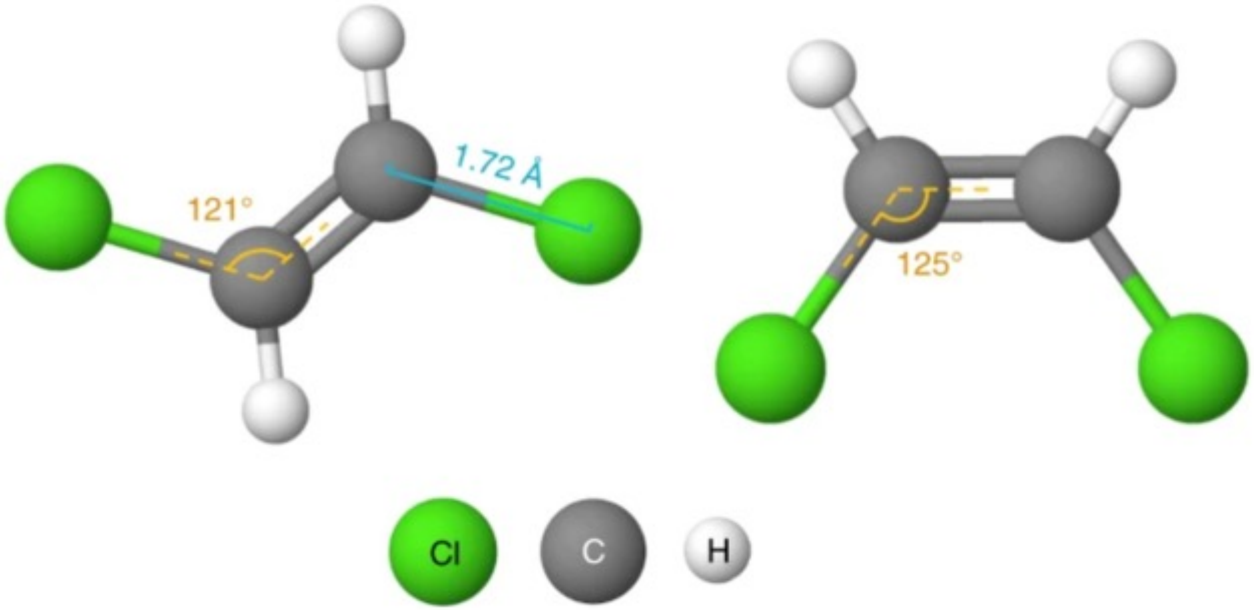




$$\underset{h\in\mathcal{H}}{\operatorname{argmin}}(1/m)\sum_{i=1}^{m} L((\mathbf{x}^{(i)}, y^{(i)}), h)$$

data point = some molecule
features = geometric structure
label = ?

Fig. 1: Comparison between two stereoisomers with the same topology but different geometries.



The two chlorine atoms are on different sides in *trans*-1,2-dichloroethene (left) but the same side in *cis*-1,2-dichloroethene (right).

Fang, X., Liu, L., Lei, J. *et al.* Geometry-enhanced molecular representation learning for property prediction. *Nat Mach Intell* **4**, 127–134 (2022). https://doi.org/10.1038/s42256-021-00438-4

number n of features

“sample-size” m

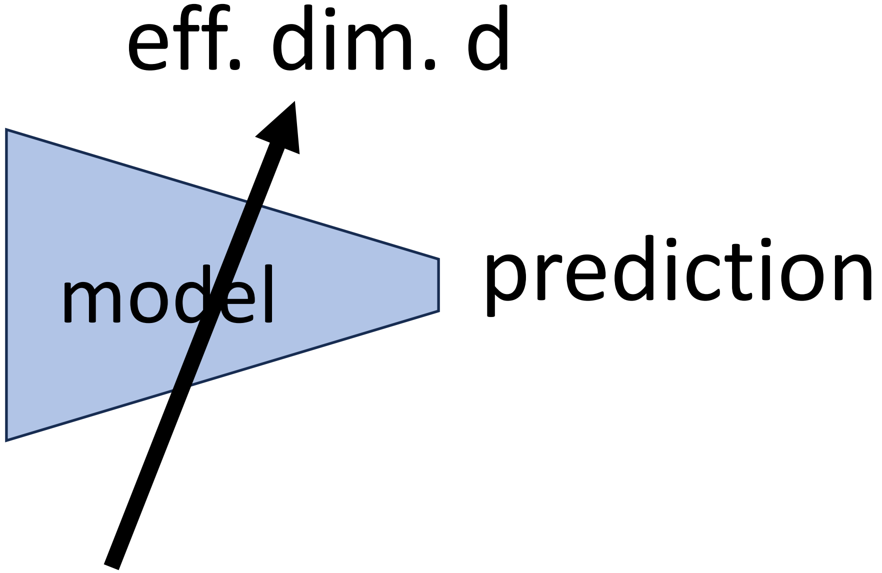
Entry ID	Structure Title	Structure Determination Methodology	Resolution (Å)	Molecular Weight per Deposited Model	Total Number of Polymer Residues per Deposited Model	Number of Non-Hydrogen Atoms per Deposited Model	Experimental Method	Release Date
7GRE	Crystal structure of SARS-CoV-2 main protease in complex with cpd-1	experimental	1.66	68.73	612	5324	X-RAY DIFFRACTION	2024-02-14
7GRF	Crystal structure of SARS-CoV-2 main protease in complex with cpd-2	experimental	1.84	68.76	612	5142	X-RAY DIFFRACTION	2024-02-14
7GRG	Crystal structure of SARS-CoV-2 main protease in complex with cpd-3	experimental	1.54	68.35	612	5313	X-RAY DIFFRACTION	2024-02-14
7GRH	Crystal structure of SARS-CoV-2 main protease in complex with cpd-4	experimental	1.87	69.34	612	5063	X-RAY DIFFRACTION	2024-02-14
7GRI	Crystal structure of SARS-CoV-2 main protease in complex with cpd-5	experimental	1.79	68.29	612	5110	X-RAY DIFFRACTION	2024-02-14
7GRJ	Crystal structure of SARS-CoV-2 main protease in complex with cpd-6	experimental	1.74	68.54	612	5091	X-RAY DIFFRACTION	2024-02-14

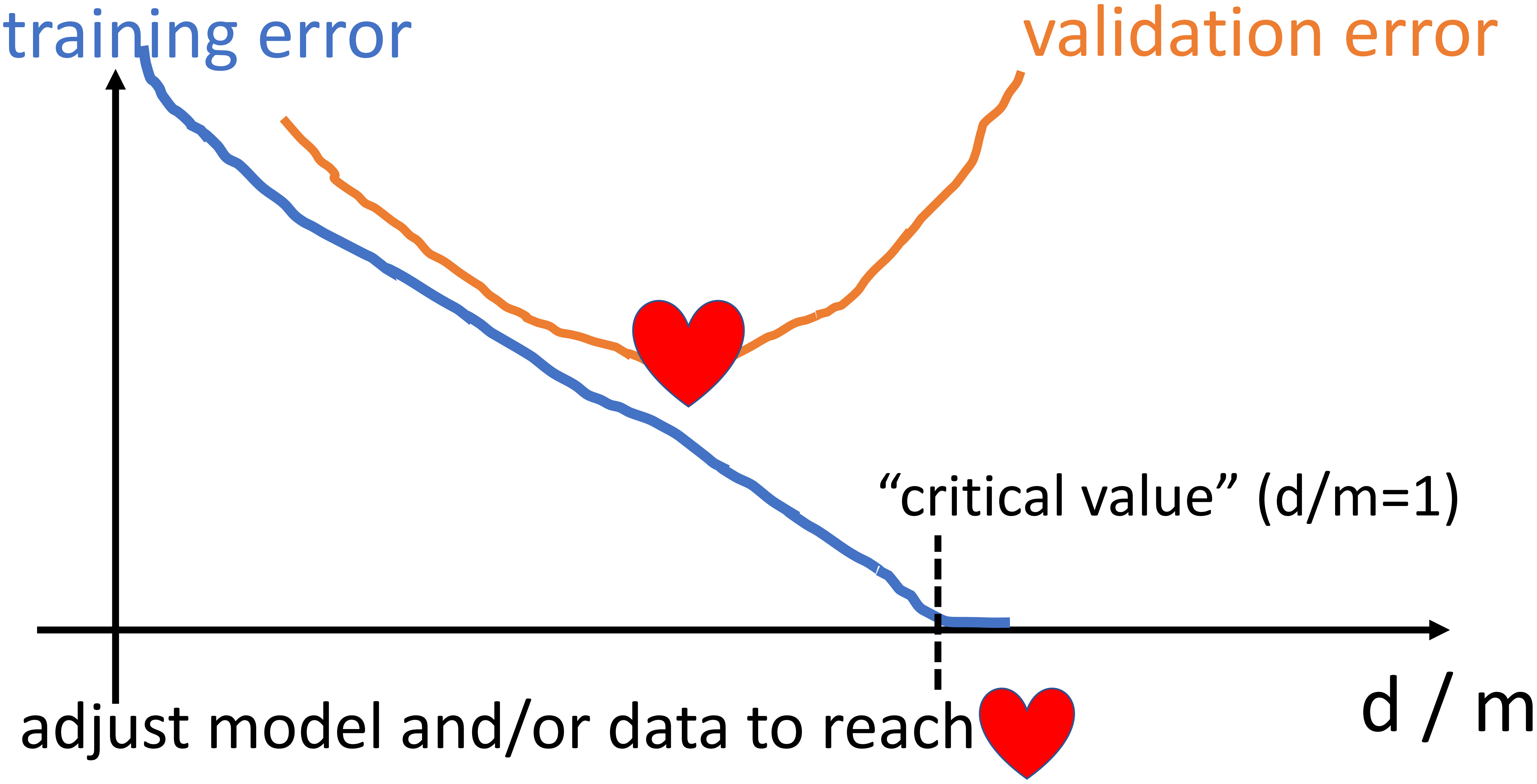
RCSB PDB PROTEIN DATA BANK

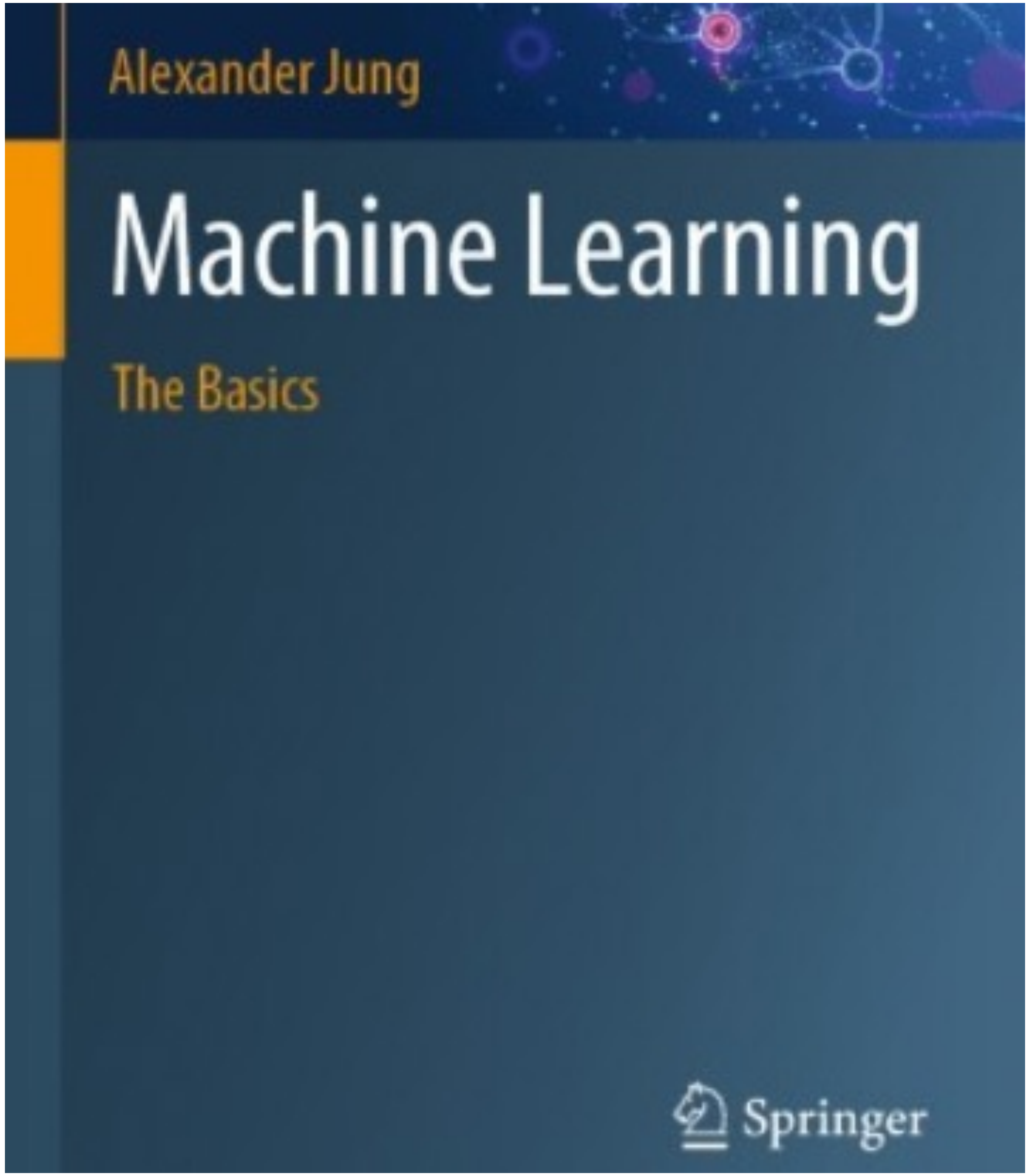
215,908 Structures from the PDB

1,068,577 Computed Structure Models (CSM)

PDB-101 wwPDB EMDataResource NAKB




training error
validation error
“critical value” (d/m=1)
d / m
adjust model and/or data to reach


Alexander Jung
Machine Learning
The Basics
Springer


Alexander Jung
Maschinelles Lernen
Die Grundlagen
Springer

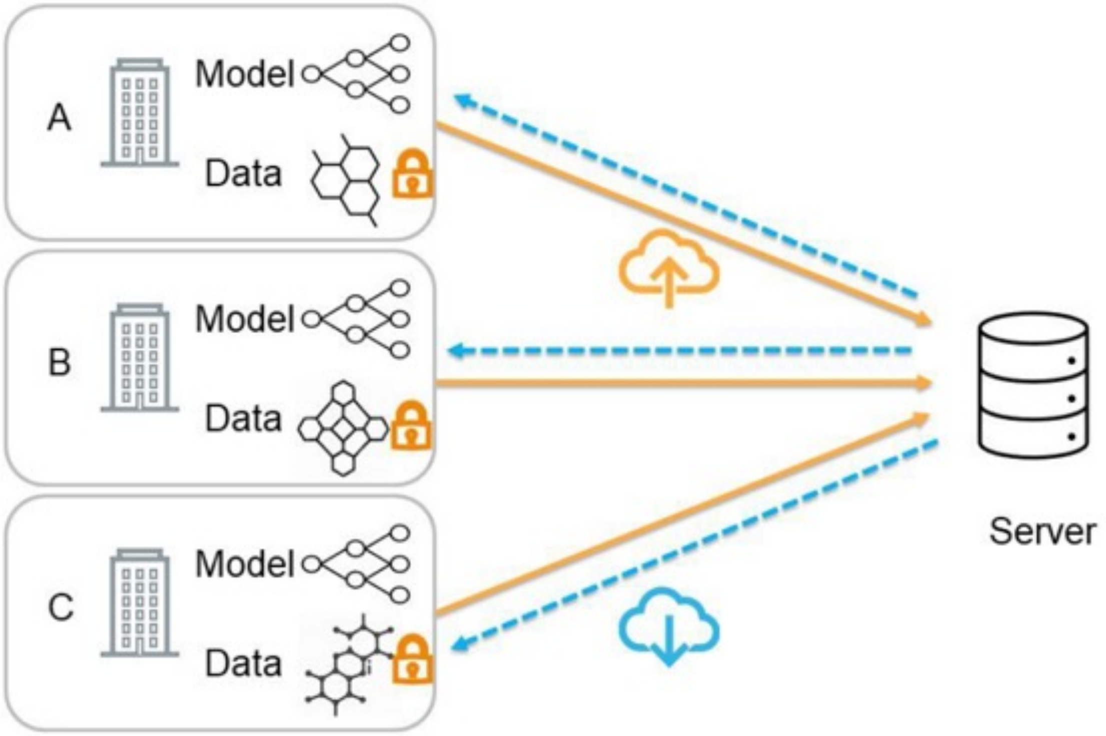


Figure 1: We illustrate heterogeneous federated molecular learning where three institutions focus on different types of molecules. The server has no access to training data.

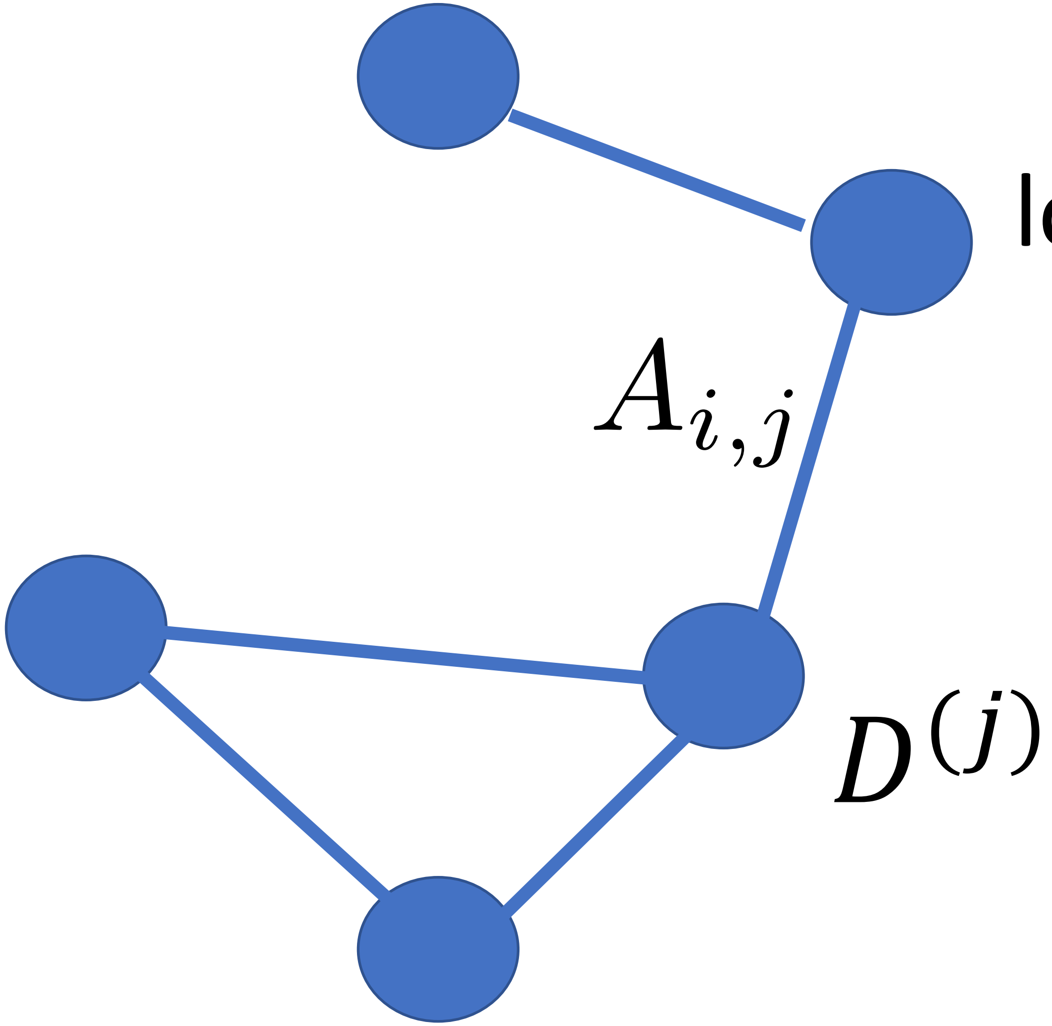
Zhu W, Luo J, White AD. Federated learning of molecular properties with graph neural networks in a heterogeneous setting. Patterns (N Y). 2022 Jun 2;3(6):100521. doi: 10.1016/j.patter.2022.100521. PMID: 35755872; PMCID: PMC9214329.

Machine Learning:
choose right model to ensure $d/m < 1$

Federated Learning:
pool right data to ensure $d/m < 1$

FL Design Principle

Networked Data.



local dataset $D^{(i)}$

edge weights $A_{i,j}$ quantify "similarities"

Testing Similarity

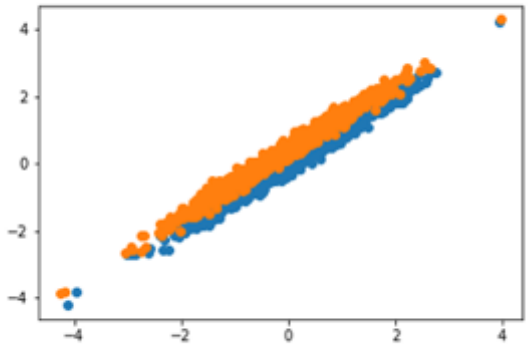


Figure 1: The figure shows the random samples from the distribution of X (Blue points) and X' (Orange points) with $d = 2$, $\bar{d} = 1$, $G(z) = (z, z)$, $u = 0.3$ and $\sigma = 0.1$. We can see that the shape and the location of the two scatter plots are quite similar, yet the ℓ_p distance is quite large due to the support mismatching.

Kyoungjae Lee. Kisung You. Lizhen Lin. "Bayesian Optimal Two-Sample Tests for High-Dimensional Gaussian Populations." Bayesian Anal. Advance Publication 1 - 25, 2023. https://doi.org/10.1214/23-BA1373

Measuring Similarity

1. map local dataset i to a vector z_i

2. measure similarity between i,i' via z_i and $z_{i'}$

3. how to map dataset to a vector?

The Gradient…

…maps a dataset to ….

$$\underbrace{(1/m)\sum_{r=1}^{m}\left(y^{(r)} - \mathbf{w}^{T}\mathbf{x}^{(r)}\right)^{2}}_{:=f(\mathbf{w})}.$$

…a vector.

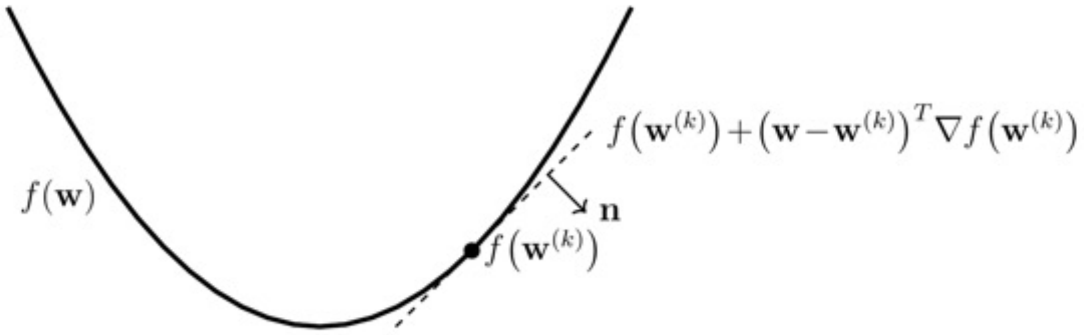
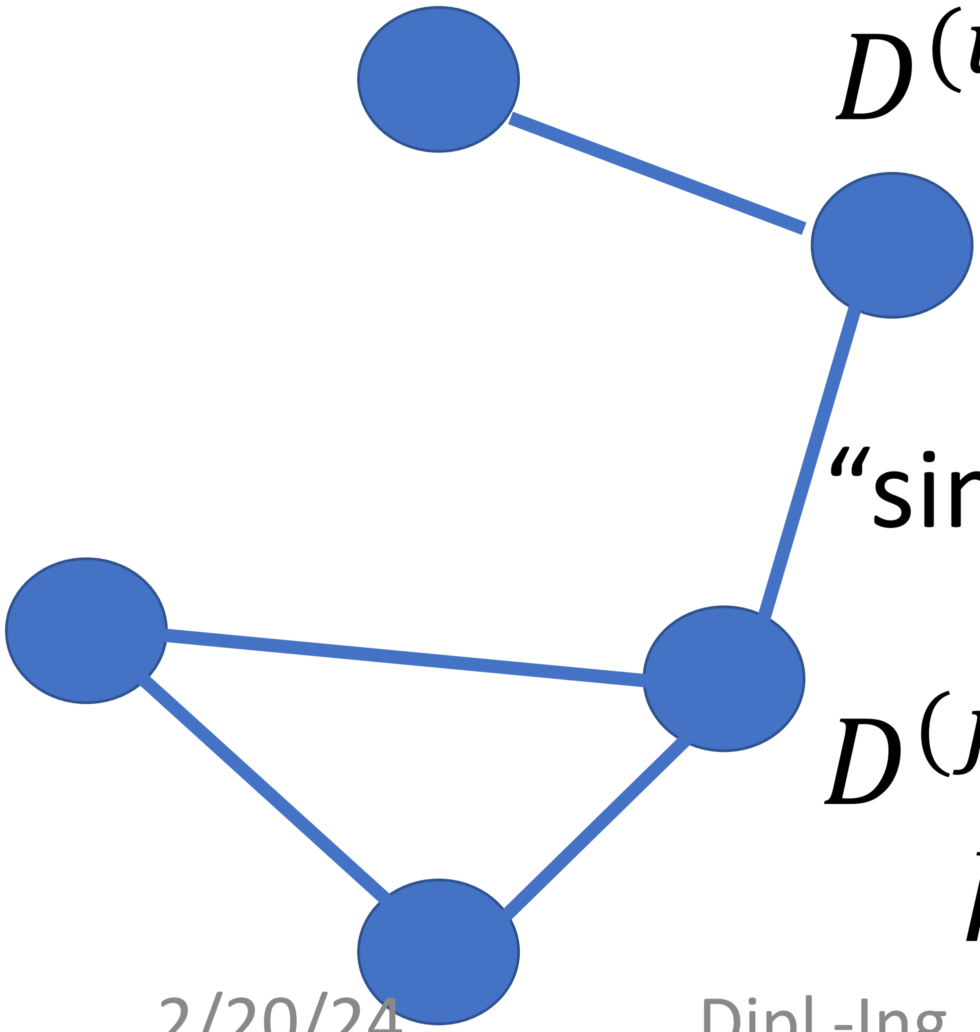


Figure 4.1: We can approximate a differentiable function $f(\mathbf{w})$ locally around a point $\mathbf{w}^{(k)} \in \mathbb{R}^{d}$ using the linear function $f\big(\mathbf{w}^{(k)}\big)+\big(\mathbf{w}-\mathbf{w}^{(k)}\big)^{T}\nabla f\big(\mathbf{w}^{(k)}\big)$. Geometrically, we approximate the graph of $f(\mathbf{w})$ by a hyperplane with normal vector $\mathbf{n} = (\nabla f\big(\mathbf{w}^{(k)}\big), -1)^{T} \in \mathbb{R}^{d+1}$ of this approximating hyperplane is determined by the gradient $\nabla f\big(\mathbf{w}^{(k)}\big)$ [5].

Networked Models.



$D^{(i)}$ $h^{(i)} \epsilon \mathcal{H}^{(i)}$

local model for each node

“similar"

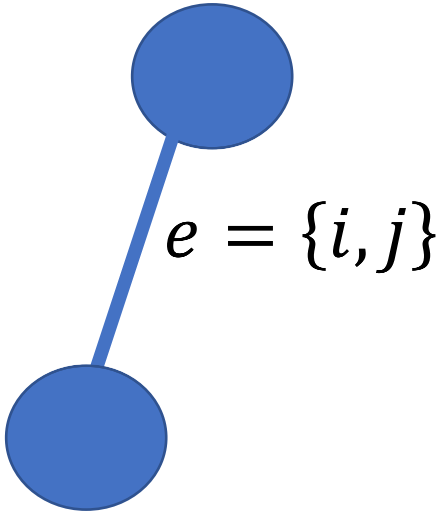
couple models at connected nodes

$D^{(j)}$

$h^{(j)} \epsilon \mathcal{H}^{(j)}$

Measuring Variation over Edge

model params $\boldsymbol{w}^{(i)}$

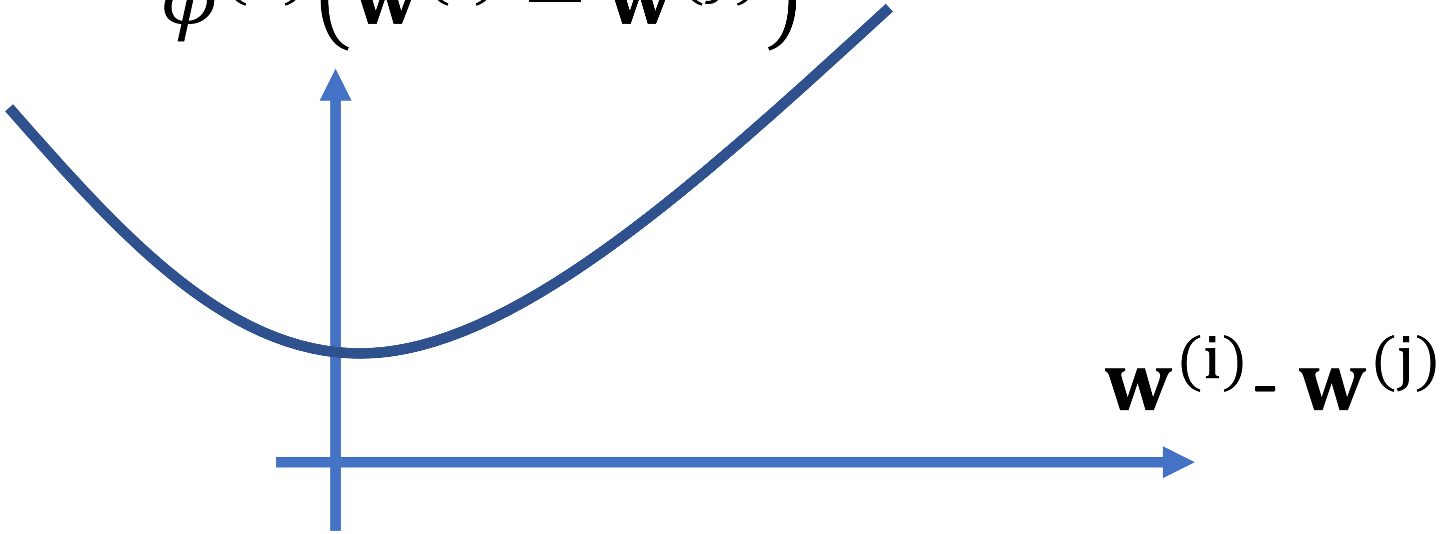


$\boldsymbol{w}^{(j)}$

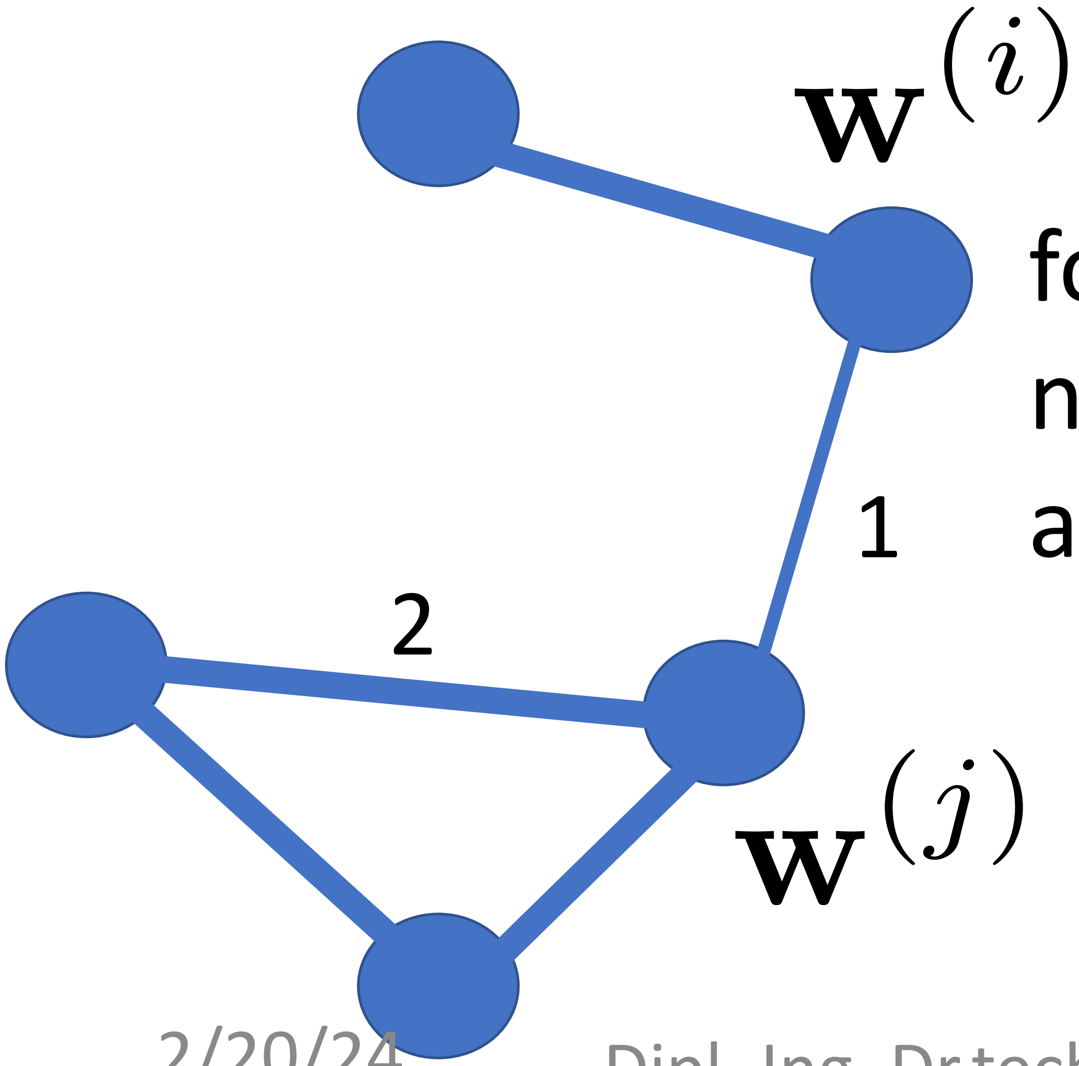
require similar params at ends of edge e

penalty function measures “tension”

$\phi^{(e)}\left(\mathbf{w}^{(i)} - \mathbf{w}^{(j)}\right)$



Generalized Total Variation (GTV)



force params of well connected nodes to be similar by requiring a small GTV

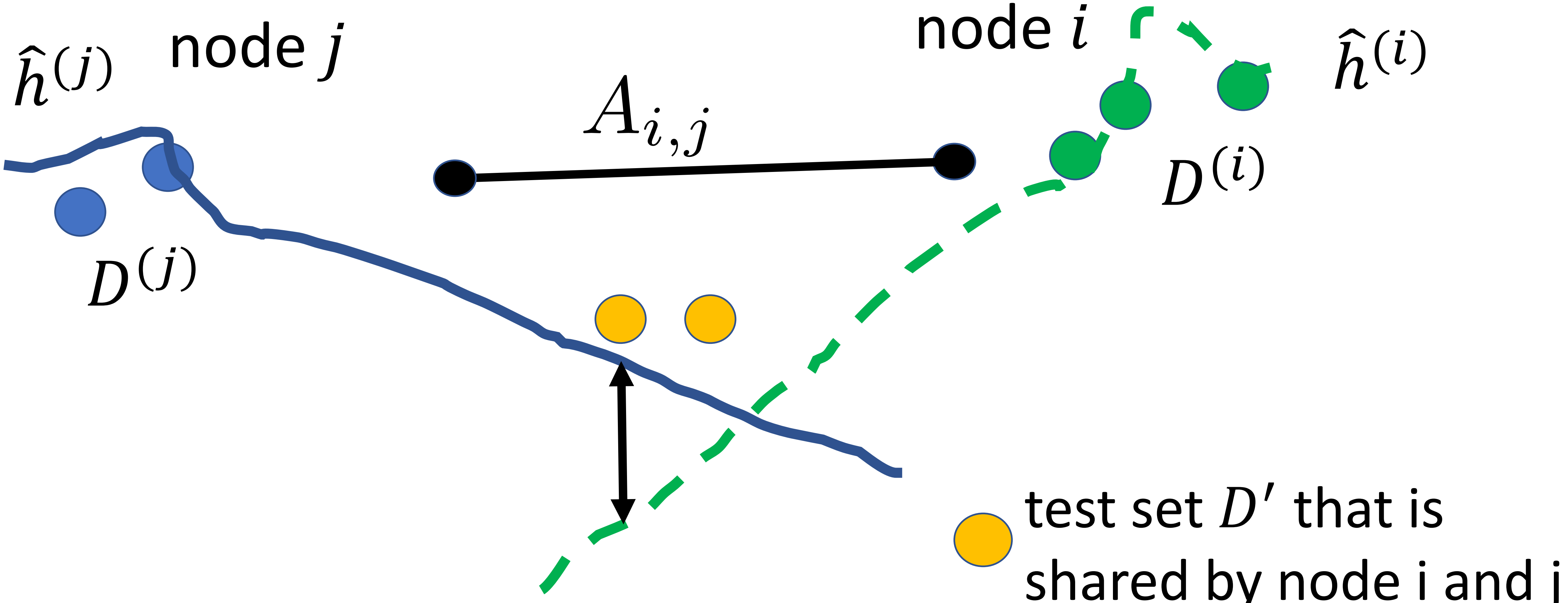
$$\sum_{\{i,j\}} A_{i,j}\phi\big(\mathbf{w}^{(i)} - \mathbf{w}^{(j)}\big)$$

Design Choice: Penalty Function

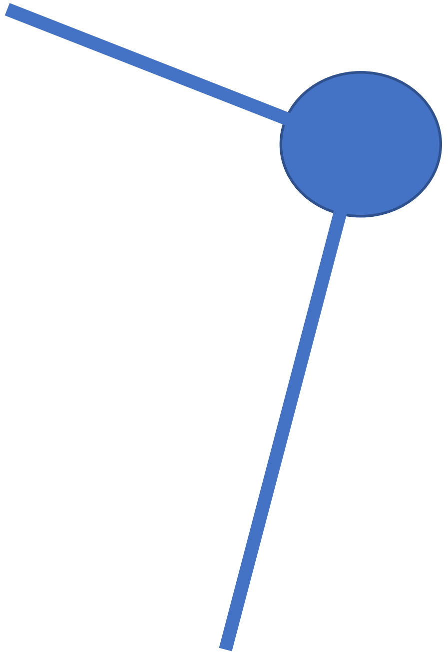
MOCHA: $\phi = \left\| \mathbf{w}^{(i)} - \mathbf{w}^{(j)} \right\|^2$

Lasso: $\phi = \left\| \mathbf{w}^{(i)} - \mathbf{w}^{(j)} \right\|$

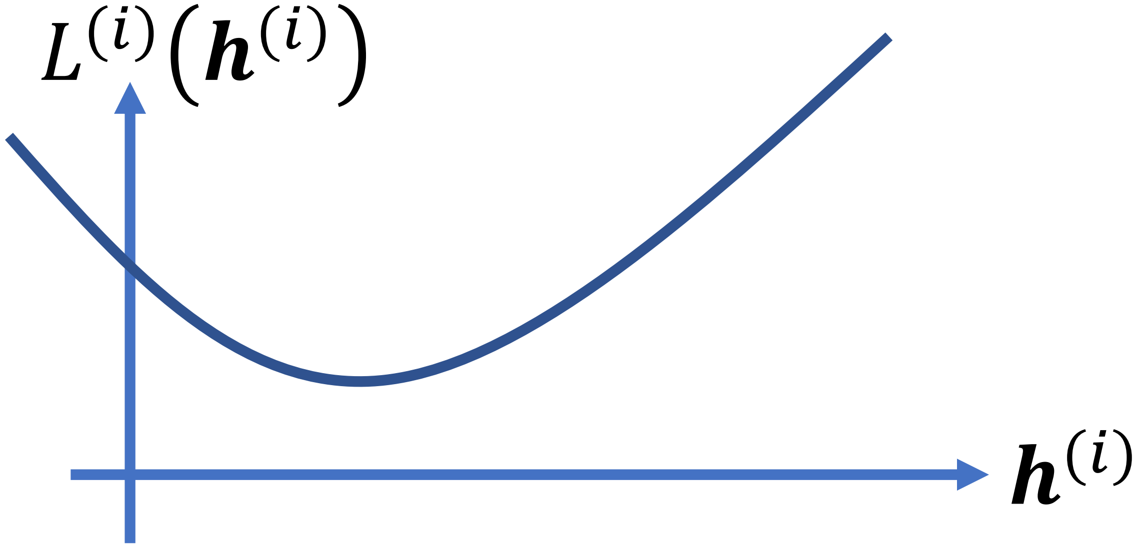
Variation of Non-Param. Models



Local Loss Functions.



measure quality of hypothesis by local loss function



GTV Minimization

$$\min_{\boldsymbol{h}^{(i)} \epsilon \mathcal{H}^{(i)}} \sum_{i} L^{(i)}\left(\boldsymbol{h}^{(i)}\right) + \lambda \sum_{\{i,j\}} A_{i,j} \phi\left(\boldsymbol{h}^{(i)}; \boldsymbol{h}^{(j)}\right)$$

local loss/
fit to local data

increasing λ

“clusteredness”

Some Special Cases of GTVMin

Network Lasso

$$\min_{\mathbf{w}} \sum_{i} L^{(i)}\left(w^{(i)}\right) + \lambda \sum_{\{i,j\}} A_{i,j} \left\| w^{(i)} - w^{(j)} \right\|$$

Network Lasso: Clustering and Optimization in Large Graphs

by D Hallac · 2015 · Cited by 206 — **Network Lasso**: **Clustering and Optimization in Large Graphs** ... Keywords: Convex **Optimization**, ADMM, **Network Lasso**. Go to: ... 2013 [**Google Scholar**]. 2.

Abstract · INTRODUCTION · CONVEX PROBLEM... · EXPERIMENTS

“MOCHA”

$$\min_{w} \sum_{i} L^{(i)}\left(w^{(i)}\right) + \lambda \sum_{\{i,j\}} A_{i,j} \left\| w^{(i)} - w^{(j)} \right\|^2$$

https://papers.nips.cc › paper › 7029-federated-m... ▾ PDF

Federated Multi-Task Learning - NIPS Proceedings

by V Smith · 2017 · Cited by 501 — 3.2 MOCHA: A Framework for **Federated Multi-Task Learning**. In the **federated** setting, the aim is to train statistical models directly on the edge, and thus we solve (1) while assuming that the data {X1,..., Xm} is distributed across m nodes or devices.

Heterogeneous Federated Regression

$$\min_{w} \sum_{i} L^{(i)}\left(h^{(i)}\right) + \lambda \sum_{\{i,j\}} A_{i,j} \sum_{D'} \left(h^{(i)}(x) - h^{(j)}(x)\right)^{2}$$

Computer Science > Machine Learning

[Submitted on 8 Feb 2023]

Towards Model-Agnostic Federated Learning over Networks

A. Jung

We present a model-agnostic federated learning method for decentralized data with an intrinsic network structur
between the (statistics of) local datasets and, in turn, their associated local models. Our method is an instance of
regularization term that is constructed from the network structure of data. In particular, we require well-connecte
predictions on a common test set. In principle our method can be applied to any collection of local models. The c

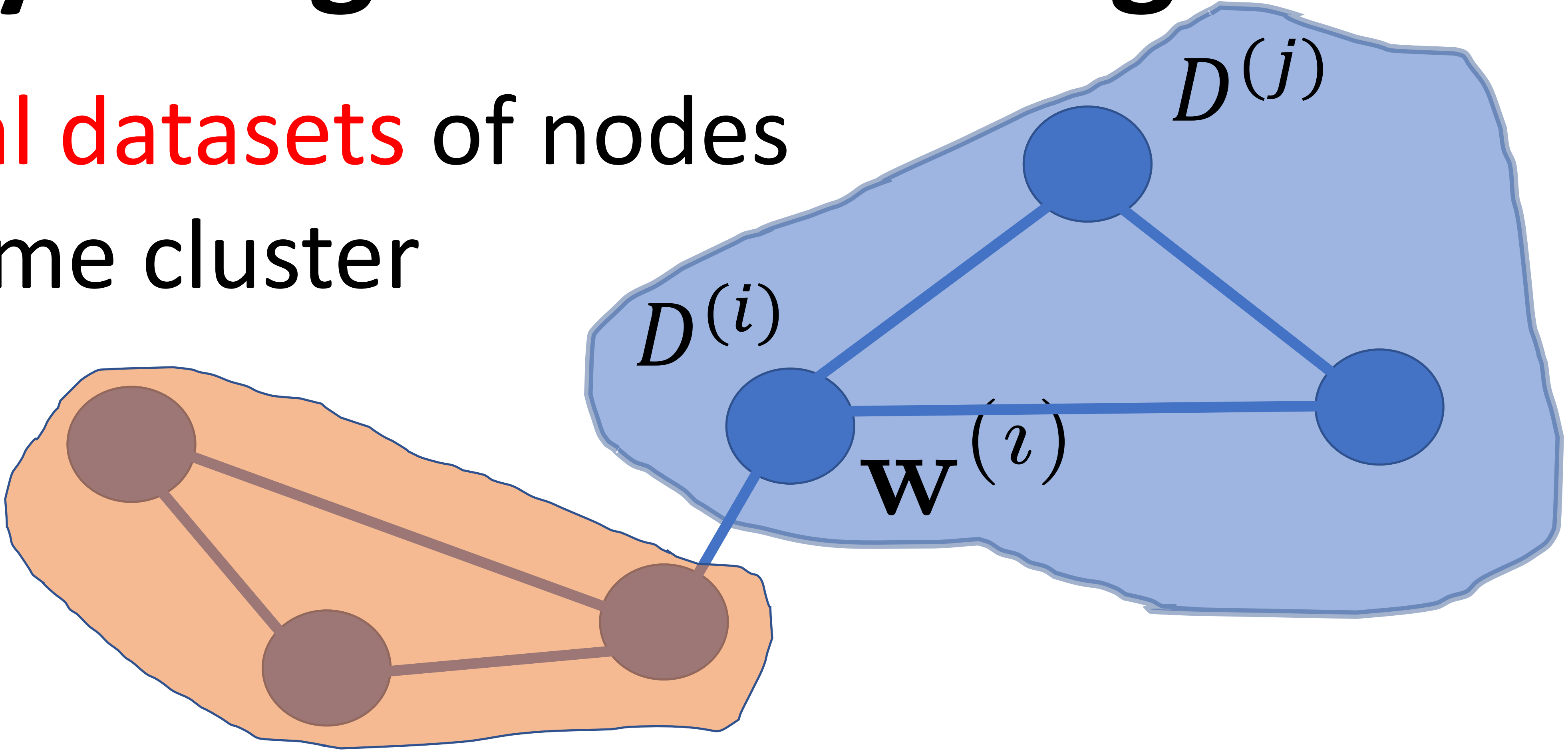
Convex Clustering

$$\min_{\mathbf{w}} \sum_i \left\| w^{(i)} - a^{(i)} \right\|^2 + \lambda \sum_{\{i,j\}} A_{i,j} \left\| w^{(i)} - w^{(j)} \right\|_p$$

D. Sun, K.-C. Toh, Y. Yuan;
Convex Clustering: Model, Theoretical Guarantee and Efficient Algorithm, JMLR, 22(9):1–32, 2021

Locally Weighted Learning

pool local datasets of nodes in the same cluster

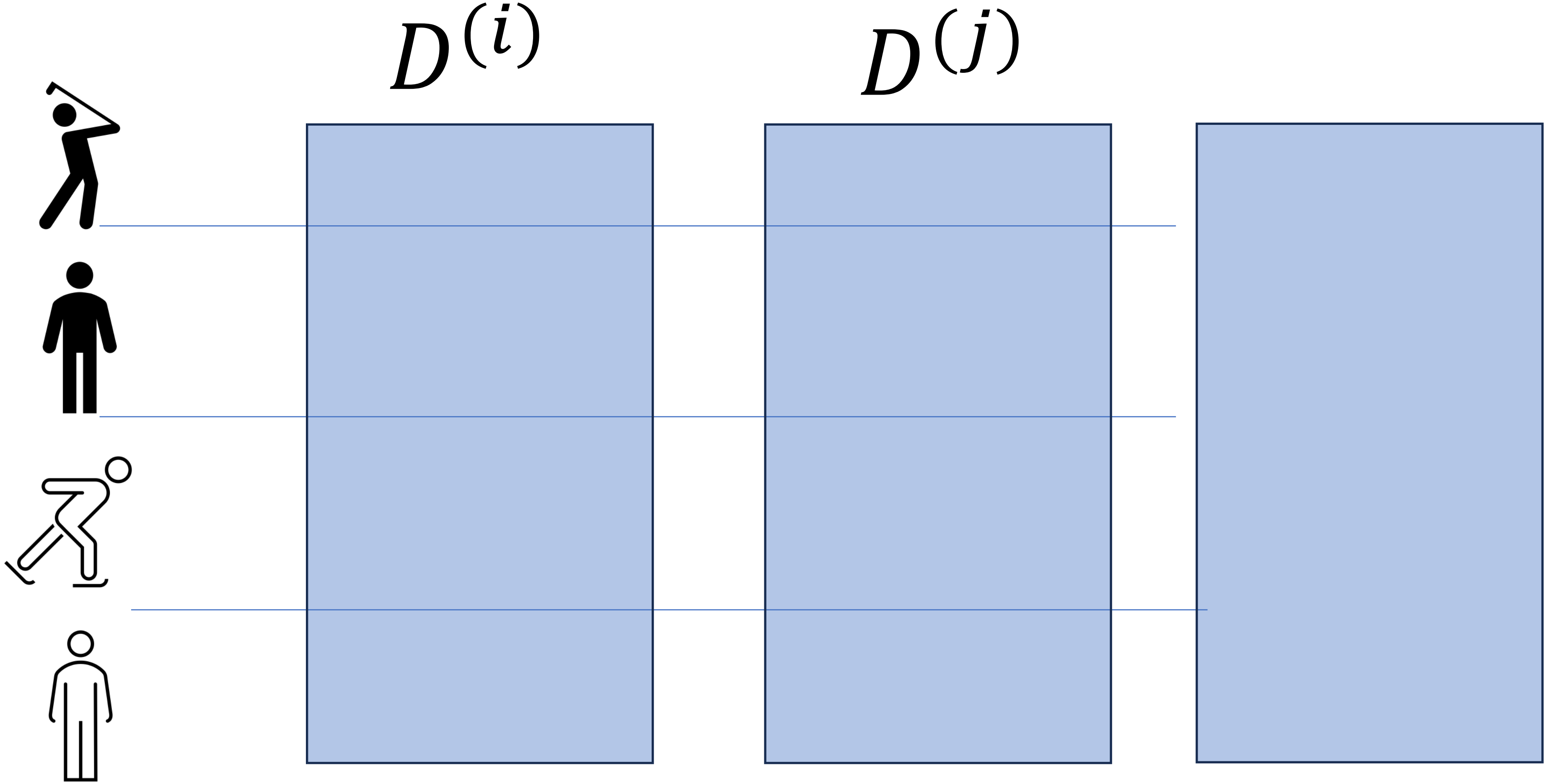


William S. Cleveland, Susan J. Devlin, Eric Grosse,
"Regression by local fitting: Methods, properties, and computational algorithms,"
Journal of Econometrics, Volume 37, Issue 1, 1988.

Vertical FL

Bundesministerium
Finanzen

jö



7 Key Requirements

R1. Human agency and oversight

R2. Technical robustness and safety

R3. Privacy and data governance

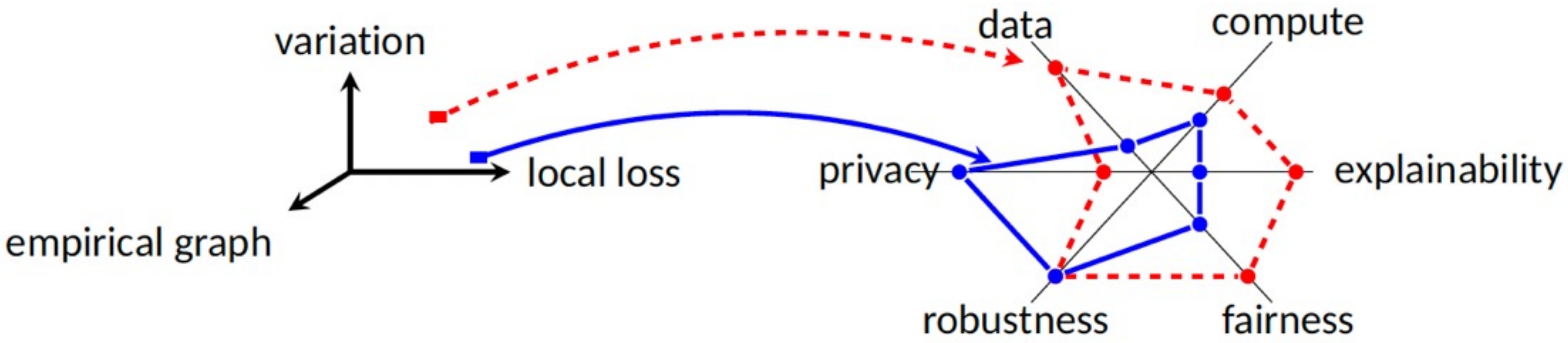
R4. Transparency

R5. Diversity, non-discrimination and fairness

R6. Societal and environmental wellbeing

R7. Accountability

https://digital-strategy.ec.europa.eu/en/library/ethics-guidelines-trustworthy-ai


variation
local loss
empirical graph
data
compute
privacy
explainability
robustness
fairness

R2. Technical robustness and safety

Some Assumptions

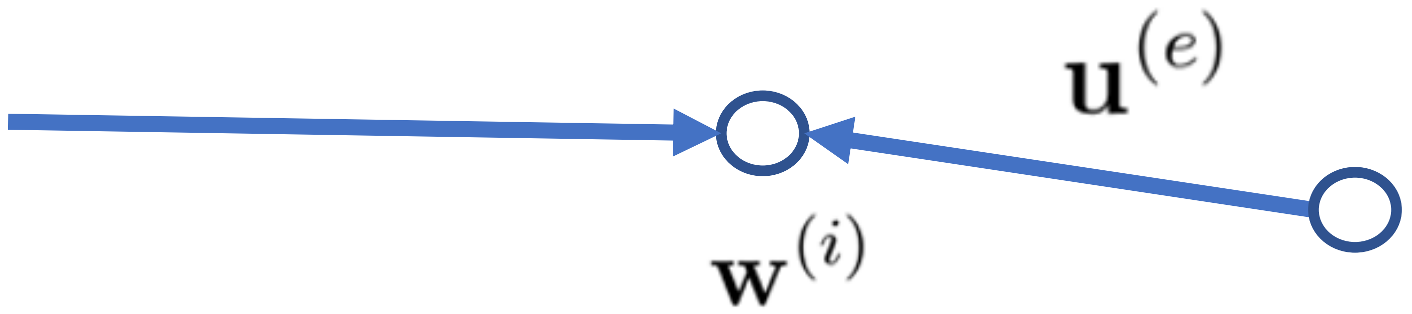
$$\min_{\mathbf{w}} \sum_{i} L^{(i)}\left(w^{(i)}\right) + \lambda \sum_{\{i,j\}} A_{i,j} \left\| w^{(i)} - w^{(j)} \right\|$$

- parametrized local models
- use some norm as penalty
- local functions are convex and diffable.

The Dual of GTVMin

$$\max_{\mathbf{u}\in\mathcal{U}} - \sum_{i\in\mathcal{V}} L_i^*\left(\mathbf{w}^{(i)}\right) - \lambda \sum_{e\in\mathcal{E}} A_e \phi^*\left(\mathbf{u}^{(e)}/(\lambda A_e)\right)$$

$$\text{subject to} \quad -\mathbf{w}^{(i)} = \sum_{e\in\mathcal{E}} \sum_{i=e_{+}} \mathbf{u}^{(e)} - \sum_{i=e_{-}} \mathbf{u}^{(e)} \text{ for all nodes } i \in \mathcal{V}.$$

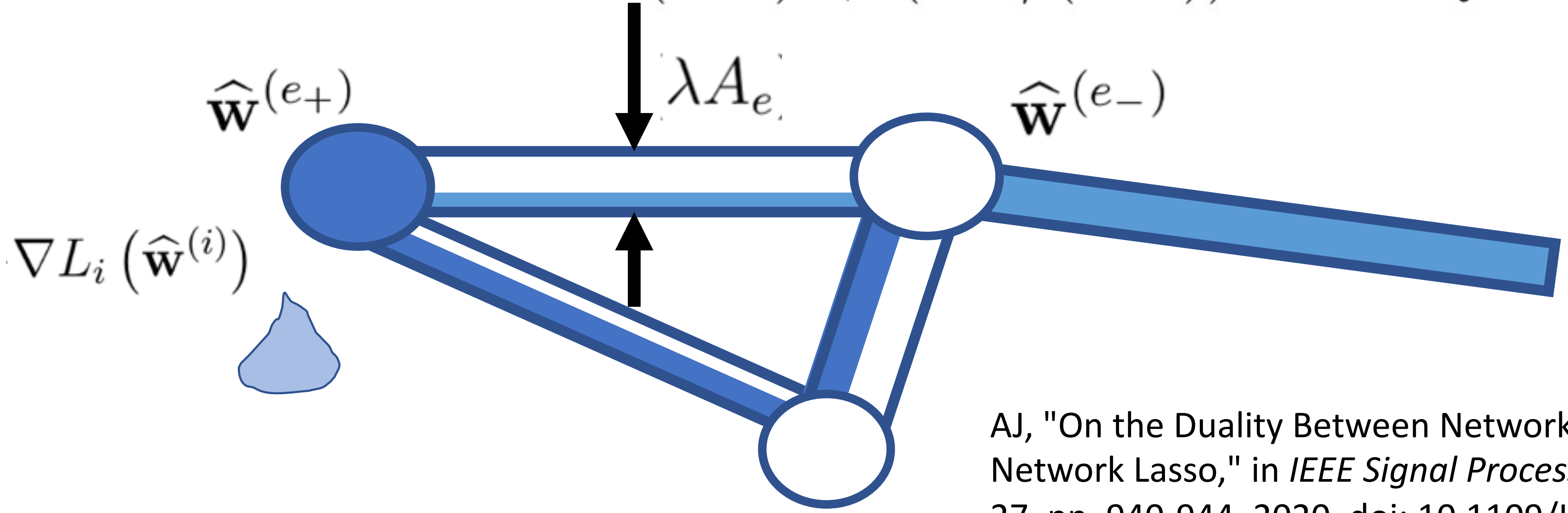


dual variables $\mathbf{u}^{(e)}$ for each (oriented) edge $e = (j, i)$

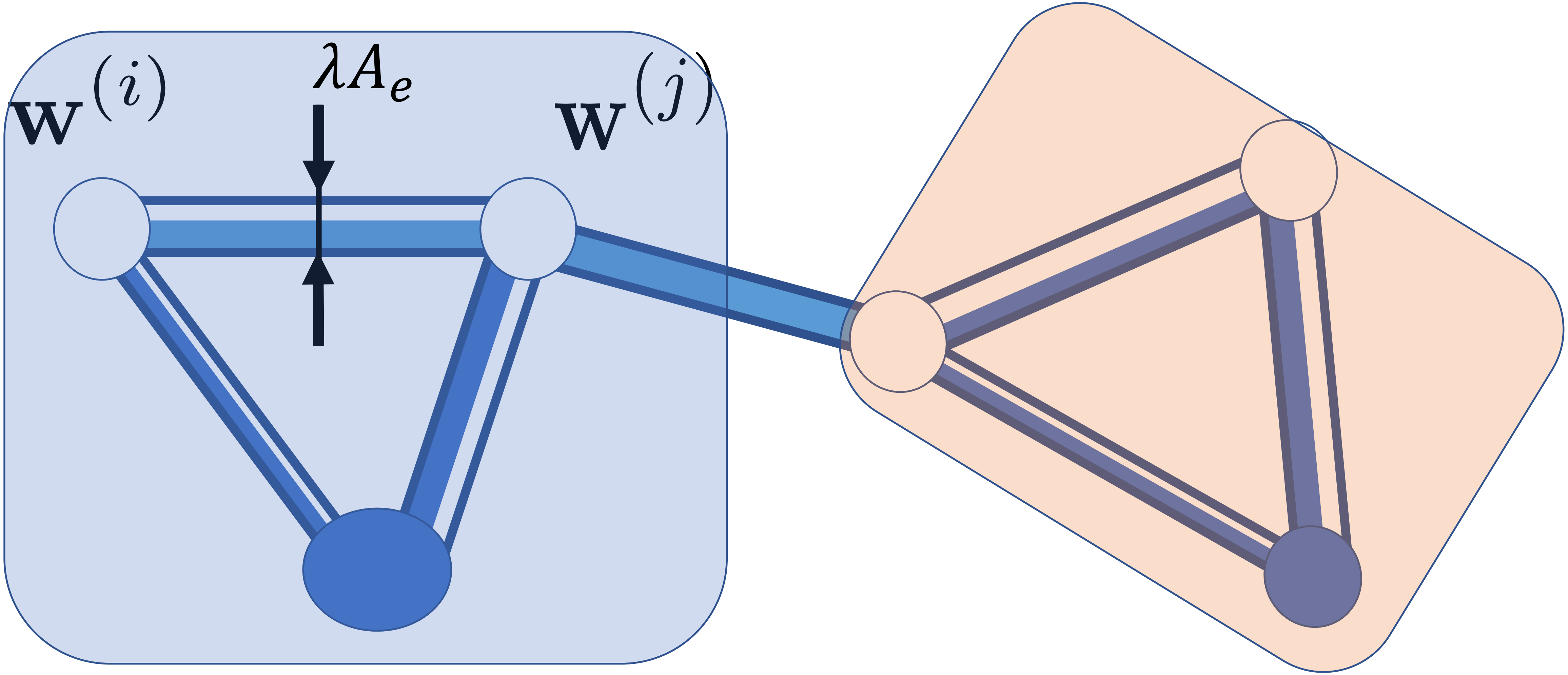
Primal and Dual Optimality.

$$\sum_{e \in \mathcal{E}} \sum_{i=e_+} \widehat{\mathbf{u}}^{(e)} - \sum_{i=e_-} \widehat{\mathbf{u}}^{(e)} = -\nabla L_i\left(\widehat{\mathbf{w}}^{(i)}\right) \text{ for all nodes } i \in \mathcal{V}$$

$$\widehat{\mathbf{w}}^{(e_+)} - \widehat{\mathbf{w}}^{(e_-)} \in (\lambda A_e)\partial\phi^*(\widehat{\mathbf{u}}^{(e)}/(\lambda A_e)) \text{ for every edge } e \in \mathcal{E}.$$



AJ, "On the Duality Between Network Flows and Network Lasso," in *IEEE Signal Processing Letters*, vol. 27, pp. 940-944, 2020, doi: 10.1109/LSP.2020.2998400.



pooling over cluster results in sufficiently large training sets

optimize robustness of GTVmin by network design

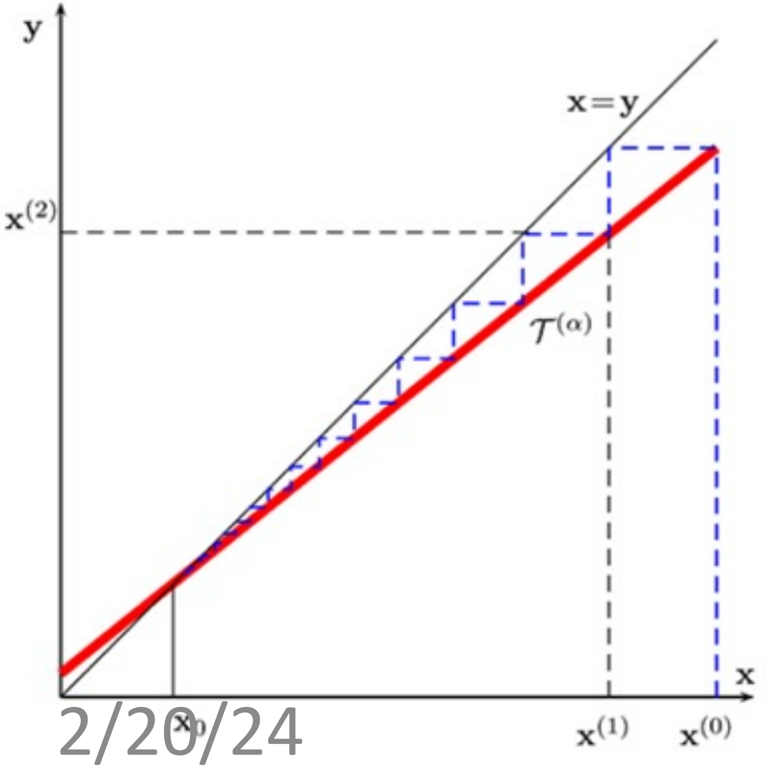
GTV Minimization

$$\min_{\boldsymbol{h}^{(i)} \epsilon \mathcal{H}^{(i)}} \sum_{i} L^{(i)}\left(\boldsymbol{h}^{(i)}\right) + \lambda \sum_{\{i,j\}} A_{i,j} \phi\left(\boldsymbol{h}^{(i)}; \boldsymbol{h}^{(j)}\right)$$

how to efficiently compute (approximate) solutions ?

Iterative Algorithms

$$w^{(k+1)} = \mathcal{T}^{(k)}\left(w^{(k)}\right)$$



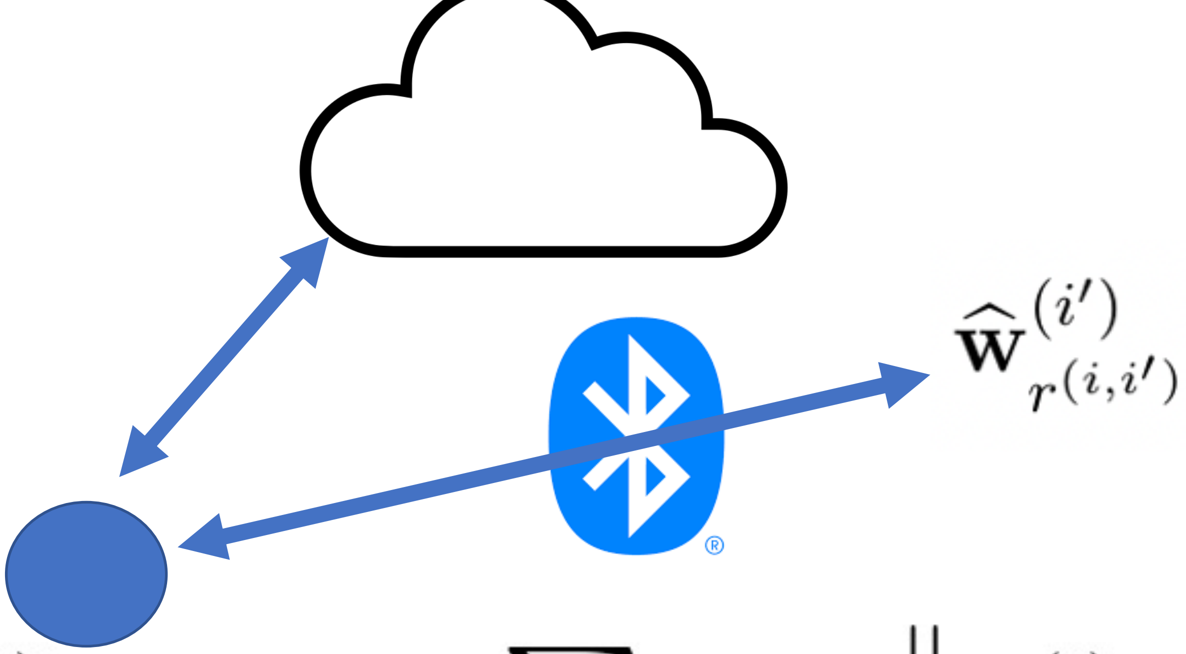
AJ, “A Fixed-Point of View on Gradient Methods for Big Data”, Front. Appl. Math. Stat., 2017.

Some Iterative Algos.

$$w^{(k+1)} = \mathcal{T}^{(k)}\left(w^{(k)}\right)$$

- gradient descent (FedSGD)
- primal-dual methods (ADMM et.al.)
- block-coordinate optimization (FedRelax)

FedRelax



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

delay $\left|r - r^{(i,i')}\right|$ due to stragglers, link failures,…

$$\widehat{\mathbf{w}}^{(i)}_{r+1} := \underbrace{\operatorname*{argmin}_{\mathbf{w}^{(i)} \in \mathbb{R}^d} L_i\left(\mathbf{w}^{(i)}\right) + (\lambda/2) \sum_{i' \in \mathcal{N}^{(i)}} A_{i,i'} \left\| \mathbf{w}^{(i)} - \widehat{\mathbf{w}}^{(i')}_{r(i,i')} \right\|_2^2}_{\mathcal{T}^{(i)}}$$

if $\mathcal{T}^{(i)}$ is a contraction under max-norm then FedRelax convergences for any max. delay

see Sec. 6.3 of D. Bertsekas, J. Tsitsiklis "Parallel and Distributed Computation: Numerical Methods", Athena, 2014

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

how to ensure update is a contraction ?

-> use a “nice” loss function (strongly convex)

Bauschke, H.H., Moffat, S.M. & Wang, X. Firmly Nonexpansive Mappings and Maximally Monotone Operators: Correspondence and Duality. *Set-Valued Anal* **20**, 131–153 (2012). https://doi.org/10.1007/s11228-011-0187-7

R3. Privacy and data governance

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

updates might leak sensitive information

diff. privacy can be ensured by perturbing updates or loss function itself

Differentially Private Empirical Risk Minimization. K. Chaudhuri, C. Monteleoni, and A. Sarwate. J. Mach. Learn. Res. (2011).

R3. Privacy and data governance

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

updates might leak sensitive information

ensure diff. privacy by perturbing updates or loss function

Differentially Private Empirical Risk Minimization. K. Chaudhuri, C. Monteleoni, and A. Sarwate. J. Mach. Learn. Res. (2011).

R4. Transparency

"the data, system and AI business models should be transparent. ..Moreover, AI systems and their decisions should be explained in a manner adapted to the stakeholder concerned..."

Explainable ERM (EERM)

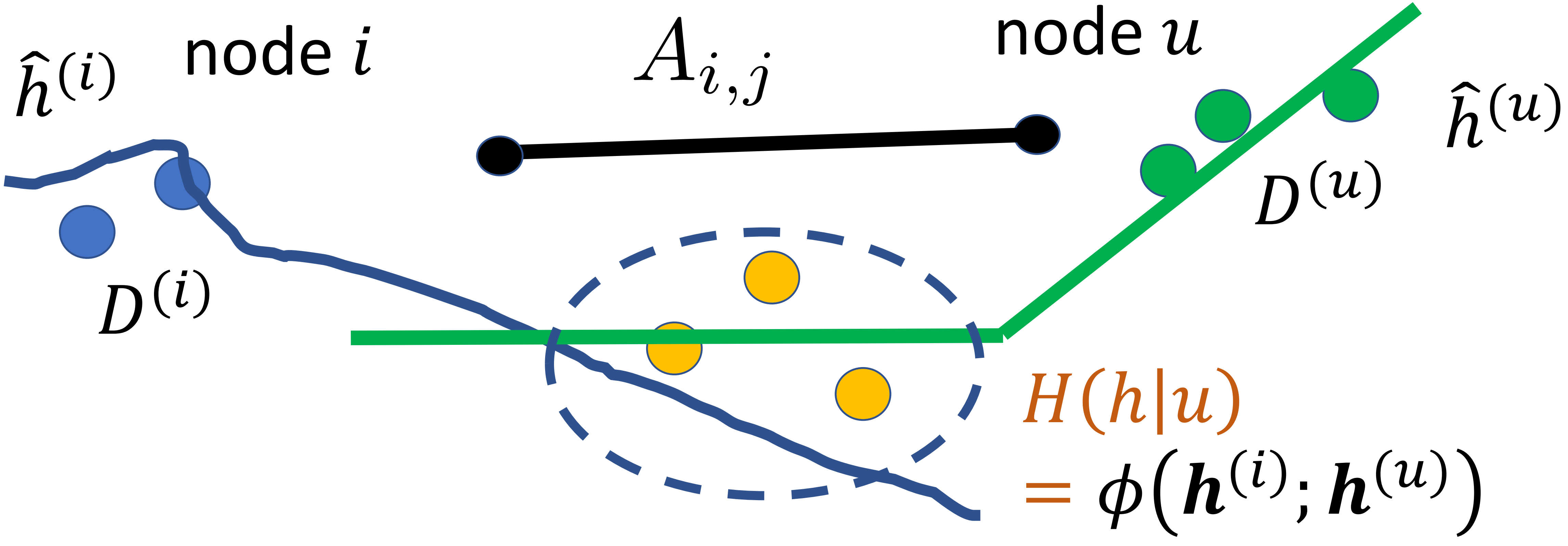
$$\min_{h \epsilon \mathcal{H}} \frac{1}{m} \sum_{i=1}^{m} L\left(\left(x^{(i)}, y^{(i)}\right), h\right) + \lambda H(h|u)$$

- $H(h|u)$ measures (lack of) subjective explainability
- enforce similar $h(x)$ for data points with similar user signal u

Zhang, L., Karakasidis, G., Odnoblyudova, A., Dogruel, L., and AJ"Explainable Empirical Risk Minimization, 2020. doi:10.48550/arXiv.2009.01492.

AJ and P. H. J. Nardelli, "An Information-Theoretic Approach to Personalized Explainable Machine Learning," in *IEEE Signal Processing Letters*, vol. 27, pp. 825-829, 2020, doi: 10.1109/LSP.2020.2993176.

User Nodes for Explainability



data points with identical user signal u

Wrap Up

- GTVmin as flexible design principle for FL
- design choices: local models, network structure, variation measure
- guided by key requirements for trustworthy AI

Happy to collaborate on …

- fundamental limits for personalized FL
- fundamental trade-offs between explainabillity and accuracy

Thank you for your attention!