CS-E4740 Federated Learning

A FL Design Principle

Dipl.-Ing. Dr.techn. Alexander Jung

What are main components of ML and how are they combined?

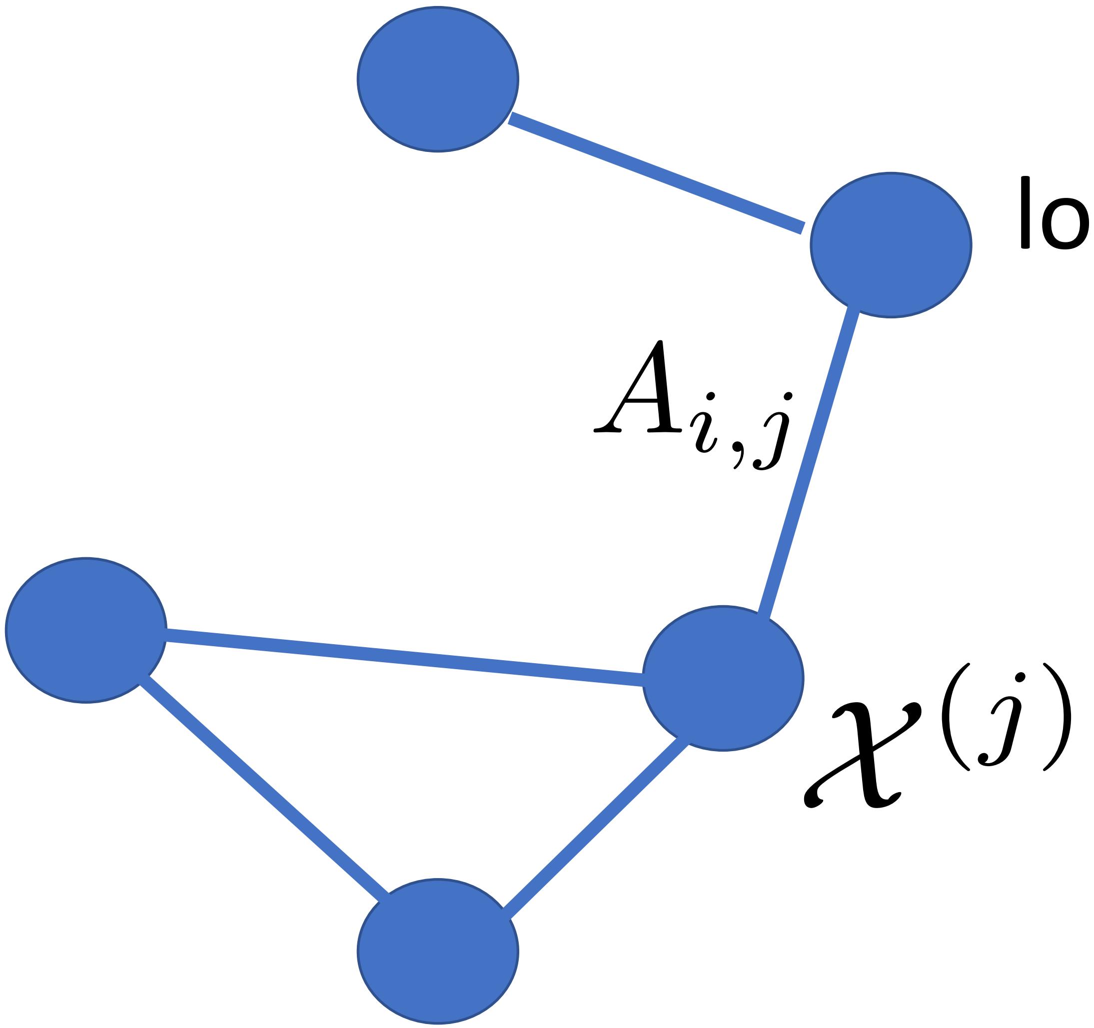
Previous Lecture:
Networked Data and Models

Today:
Loss and Optimization

Weather Stations.

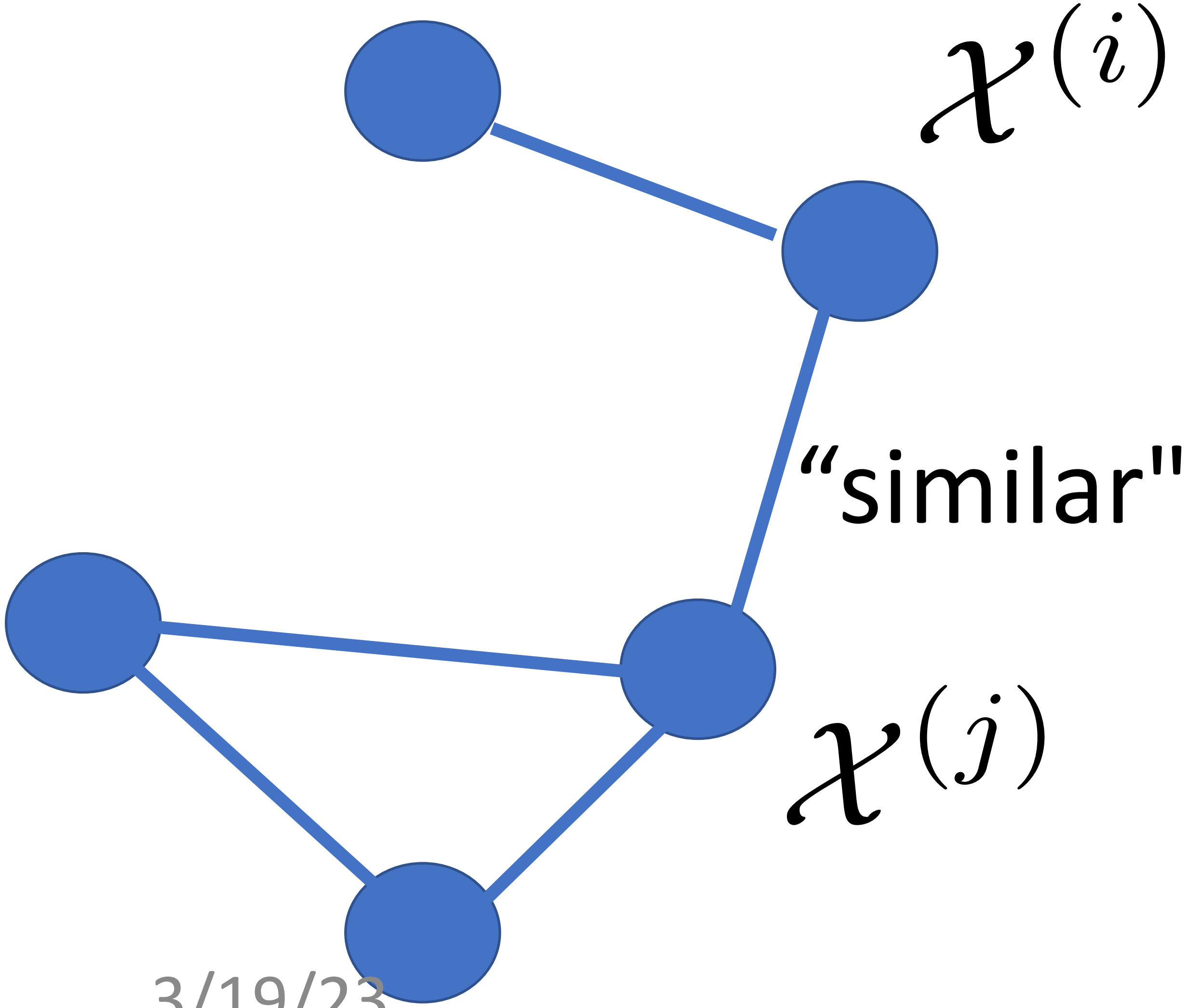
FINNISH METEOROLOGICAL INSTITUTE

The Empirical Graph



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

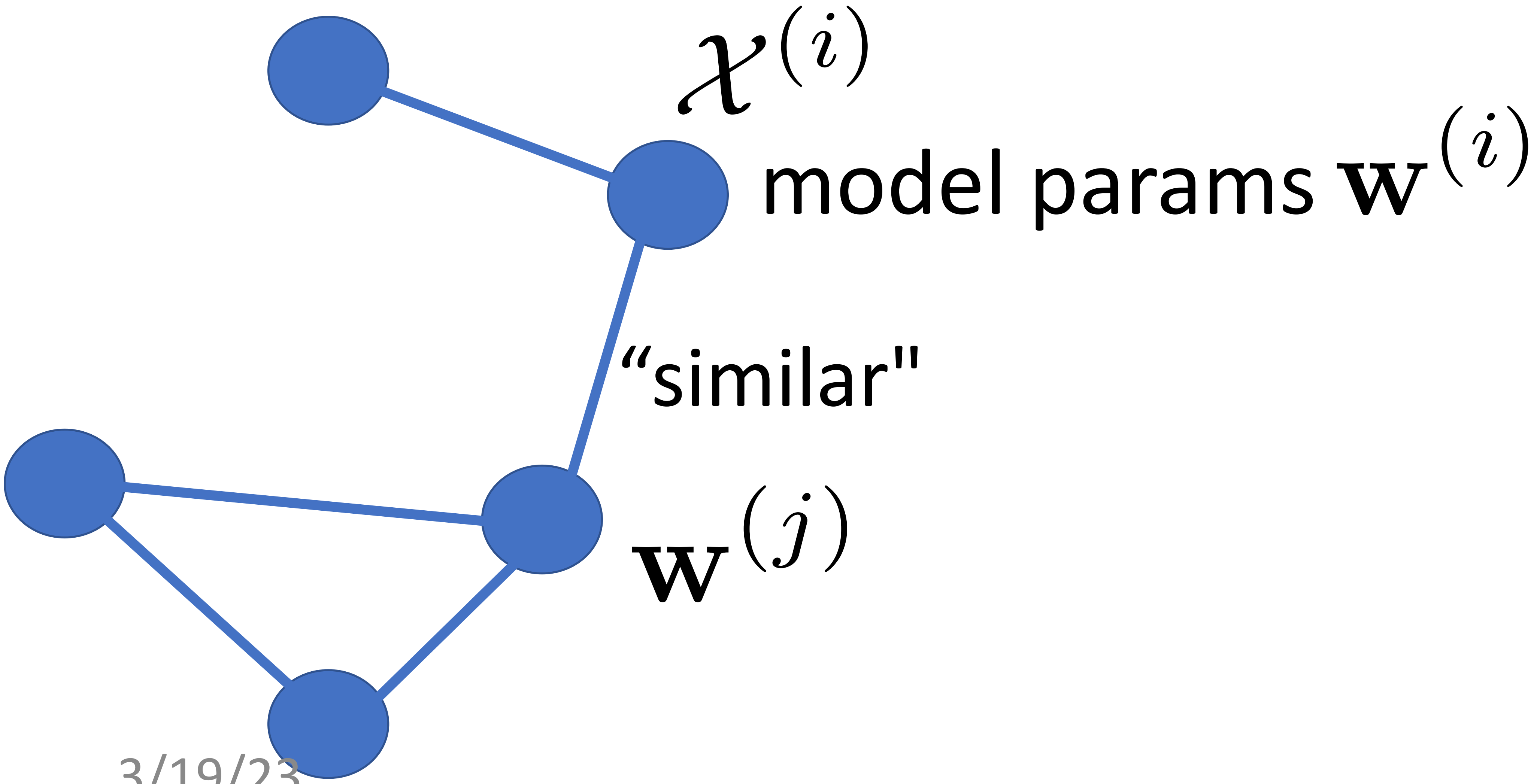
Networked Models.



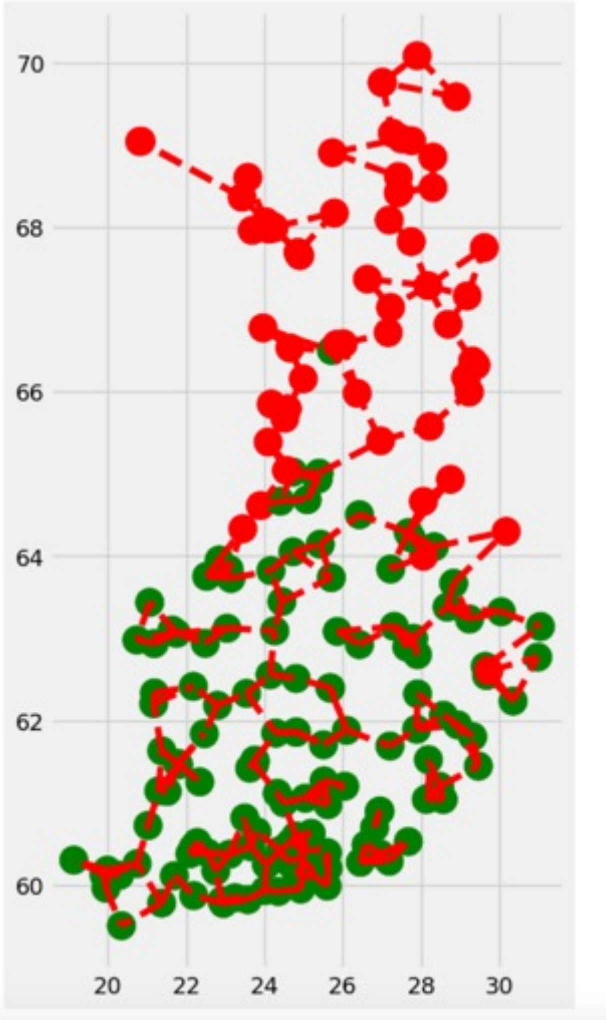
local model for each node

couple models at connected nodes

Networked Parametric Models.



Clustering Assumption



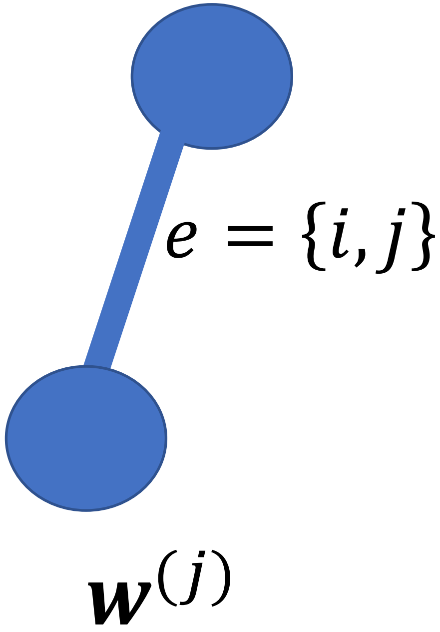
the local datasets form clusters

datasets in same can be approximated as realizations of i.i.d. RVs with prob. dist p(x,y;c)

high edge density inside clusters

Measure Clustering via Variation

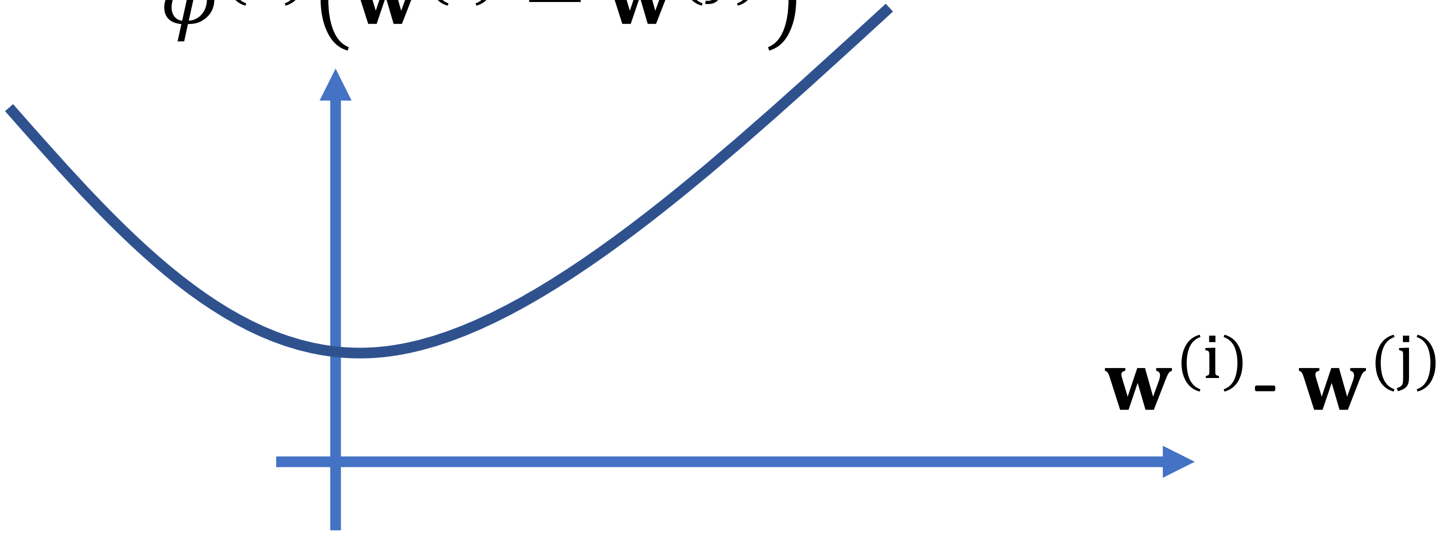
local model params $\boldsymbol{w}^{(i)}$



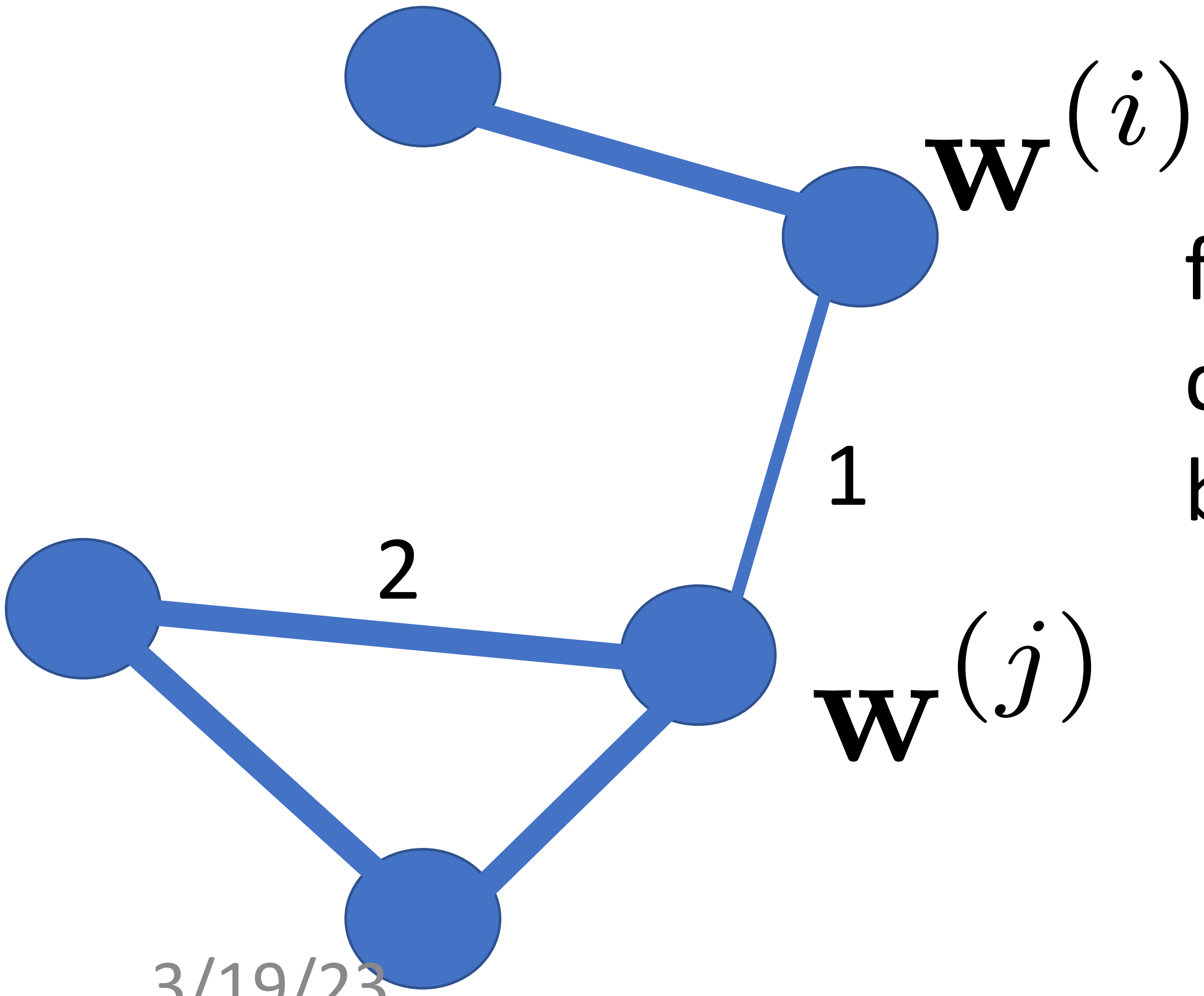
require similar params at ends of edge e

penalty function measures "variation"

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



Generalized Total Variation (GTV)



force model params at well connected nodes to be similar by requiring small GTV

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

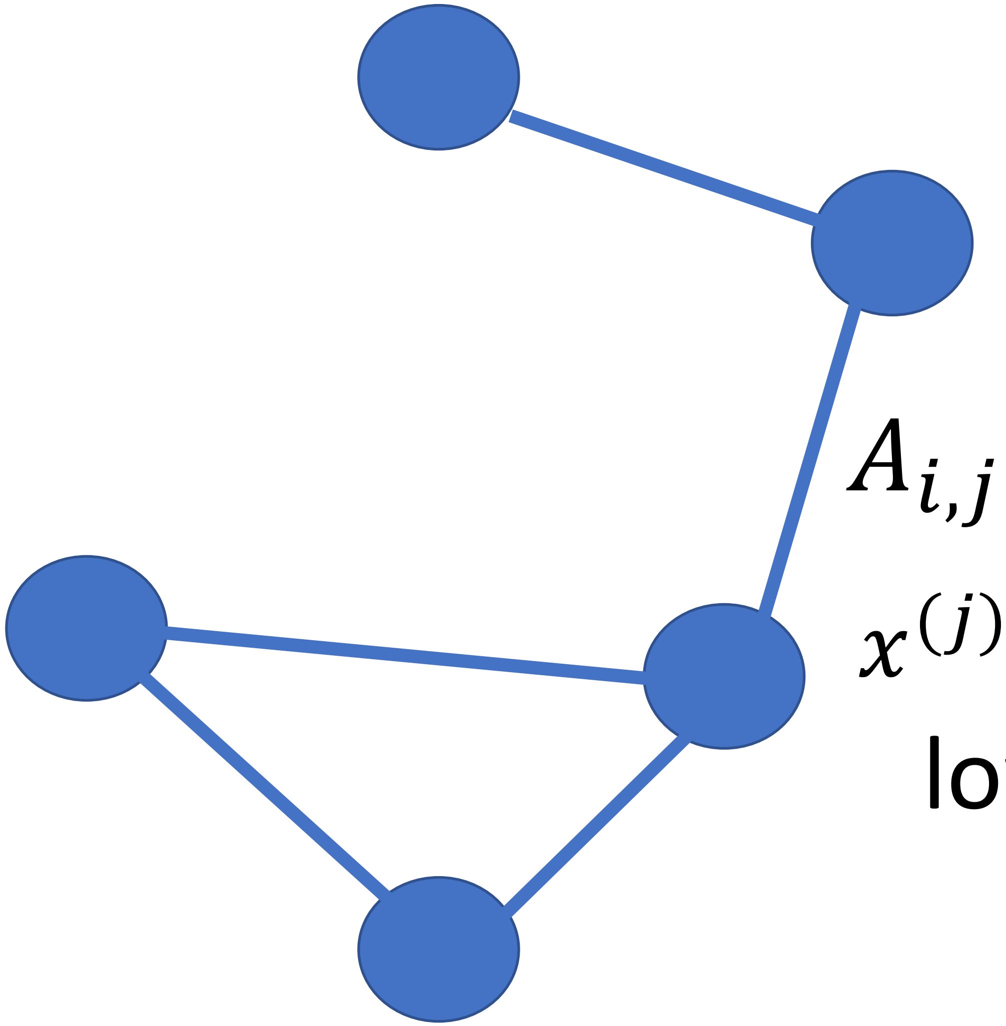
Two Special Cases of GTV.

total variation $\phi(\mathbf{u}) = \|\mathbf{u}\|_2$

graph Laplacian quadratic from is GTV with

$$\phi(\mathbf{u}) = \|\mathbf{u}\|_2^2$$

Smooth Graph Signals.



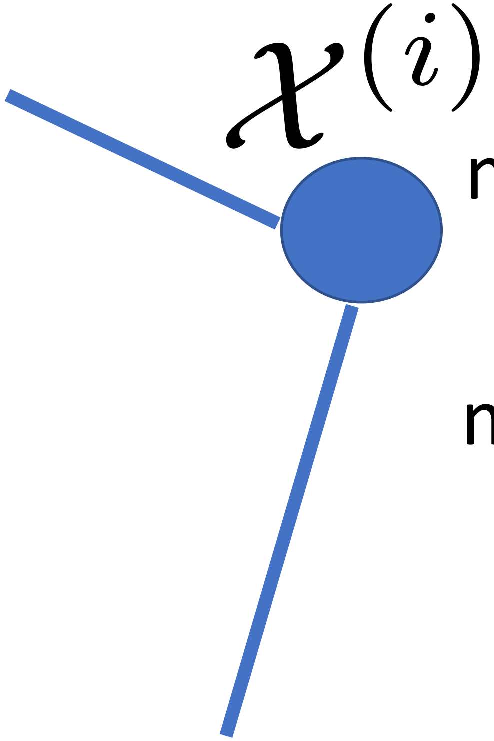
$$x^{(i)} = w^{(i)} + n^{(i)}$$

$$x^{(j)} = w^{(j)} + n^{(j)}$$

low pass constraint on

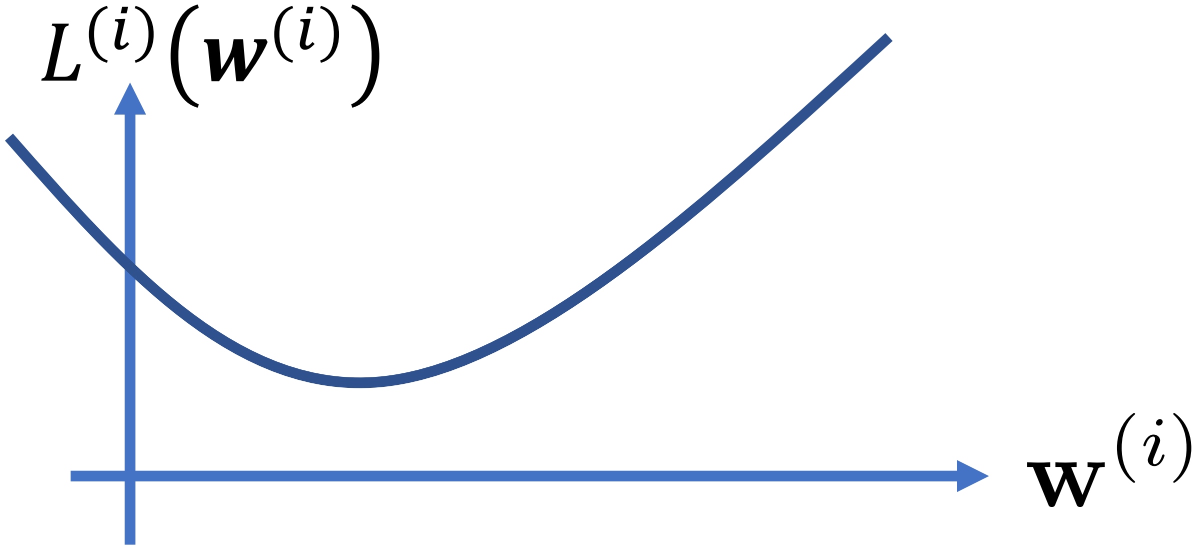
$$\sum_{\{i,j\}} A_{i,j}\left(\mathbf{w}^{(i)} - \mathbf{w}^{(j)}\right)^2$$

Local Loss Function.



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

measure quality of params by local loss function



GTV Minimization

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

increasing λ

average local loss

"clusteredness"

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

Special Case: "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.

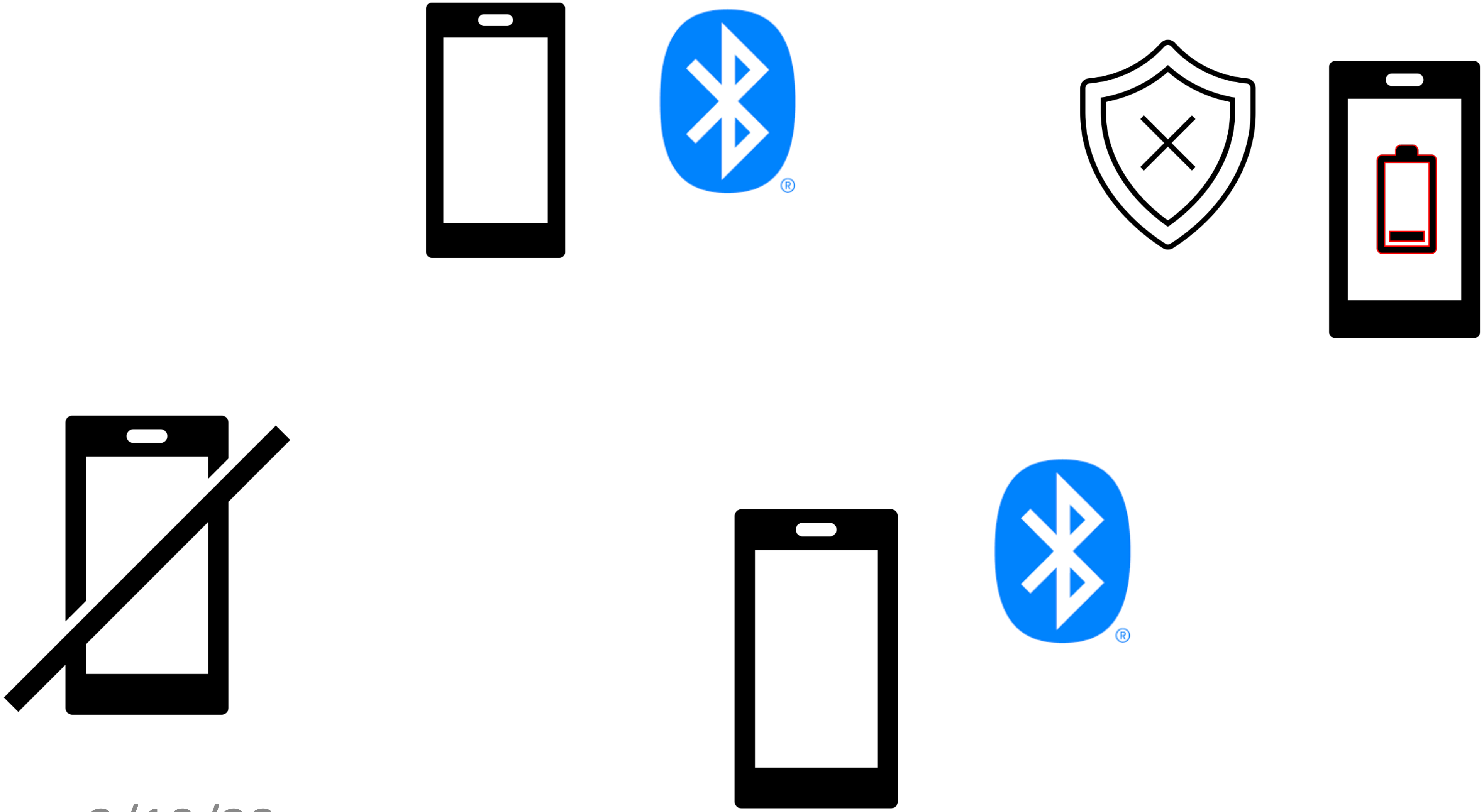
Two Key Questions of ML

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

- computational aspects: how can we efficiently compute solutions (approximately)
- statistical aspects: are the solutions any good?

Computational Aspects

A FL Setting



Requirements

- run in ad-hoc nets of low-cost devices
- robustness against node/link failures
- robustness against “stragglers”

Another FL Setting...

https://www.google.com/about/datacenters/

https://en.wikipedia.org/wiki/Optical_fiber

GTV Min. for Local Lin.Reg.

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

using stacked parameters $\mathbf{w} = \left(\mathbf{w}^{(1)}, \cdots, \mathbf{w}^{(n)}\right)^T$,

$$\min_{\mathbf{w}} \mathbf{w}^T \mathbf{Q} \mathbf{w} + \mathbf{w}^T \mathbf{q}$$

with psd matrix $\mathbf{Q}$ and vector $\mathbf{q}$ that depend on local datasets, GTVMin parameter λ and empirical graph

GTV Min. for Local Lin.Reg.

$$\min_{\mathbf{w}} \mathbf{w}^T \, \mathbf{Q} \, \mathbf{w} + \mathbf{w}^T \mathbf{q}$$

can be solved using gradient methods

$$\boldsymbol{w}^{(\boldsymbol{k}+\mathbf{1})} = \boldsymbol{w}^{(\boldsymbol{k})} - \boldsymbol{\alpha}_{\boldsymbol{k}}\left(\mathbf{2}\mathbf{Q}\boldsymbol{w}^{(\boldsymbol{k})} + \mathbf{q}\right)$$

Statistical Aspects

GTV Min. for Local Lin.Reg.

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

using stacked parameters $\mathbf{w} = \left(\mathbf{w}^{(1)}, \cdots, \mathbf{w}^{(n)} \right)^T$,

$$\sum_{\{i,j\}} A_{i,j} \left\| \mathbf{w}^{(i)} - \mathbf{w}^{(j)} \right\|^2 = \mathbf{w}^T (\mathbf{L} \otimes \mathbf{I})\, \mathbf{w}$$

with the graph Laplacian $\mathbf{L}$

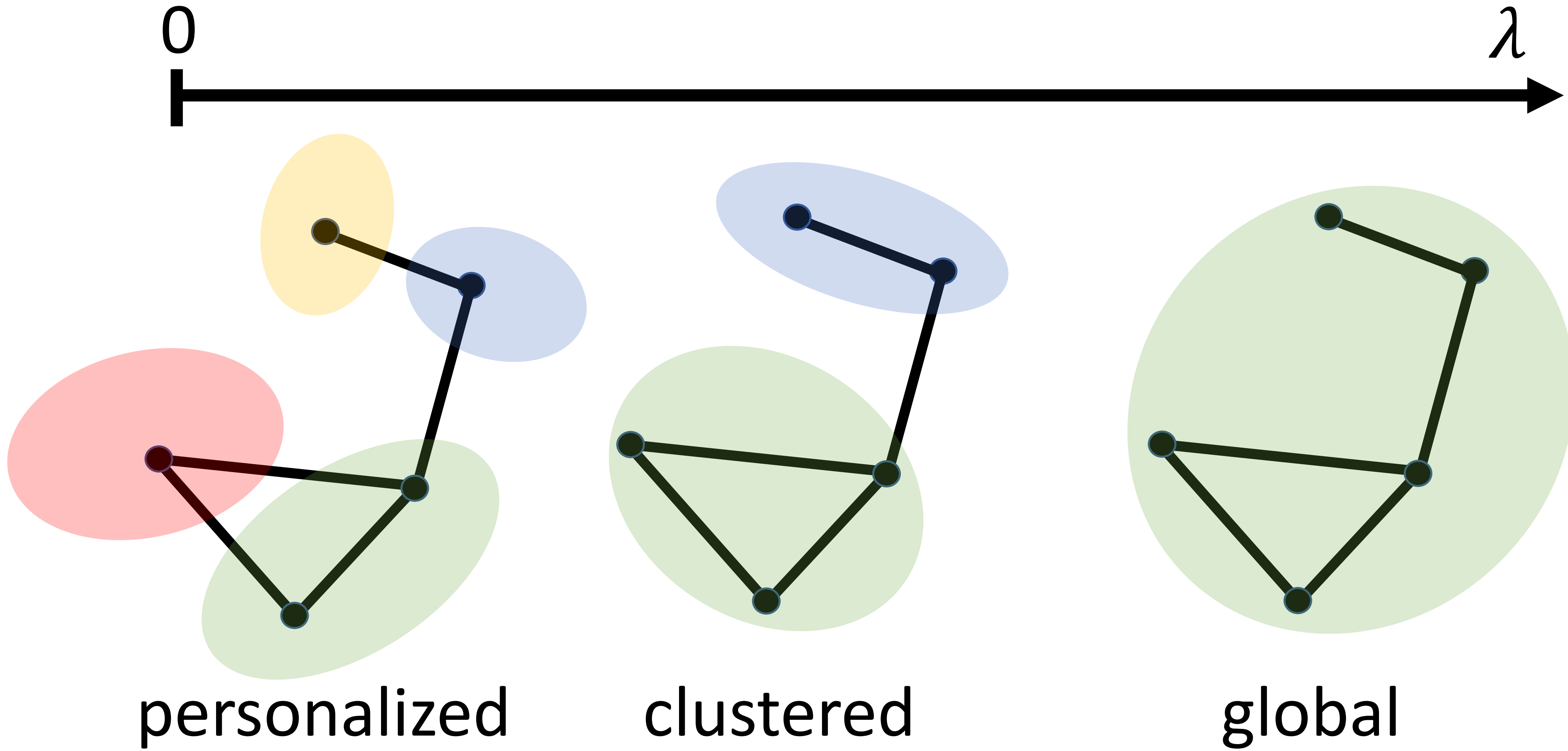
Spectral Clustering

for large λ, GTVMin is to minimize

$$\sum_{\{i,j\}} A_{i,j} \left\| \mathbf{w}^{(i)} - \mathbf{w}^{(j)} \right\|^2 = \mathbf{w}^T (\mathbf{L} \otimes \mathbf{I})\, \mathbf{w}$$

$\Rightarrow$ local model parameters should be composed of eigvecs. of L corresponding to smallest eig.vals

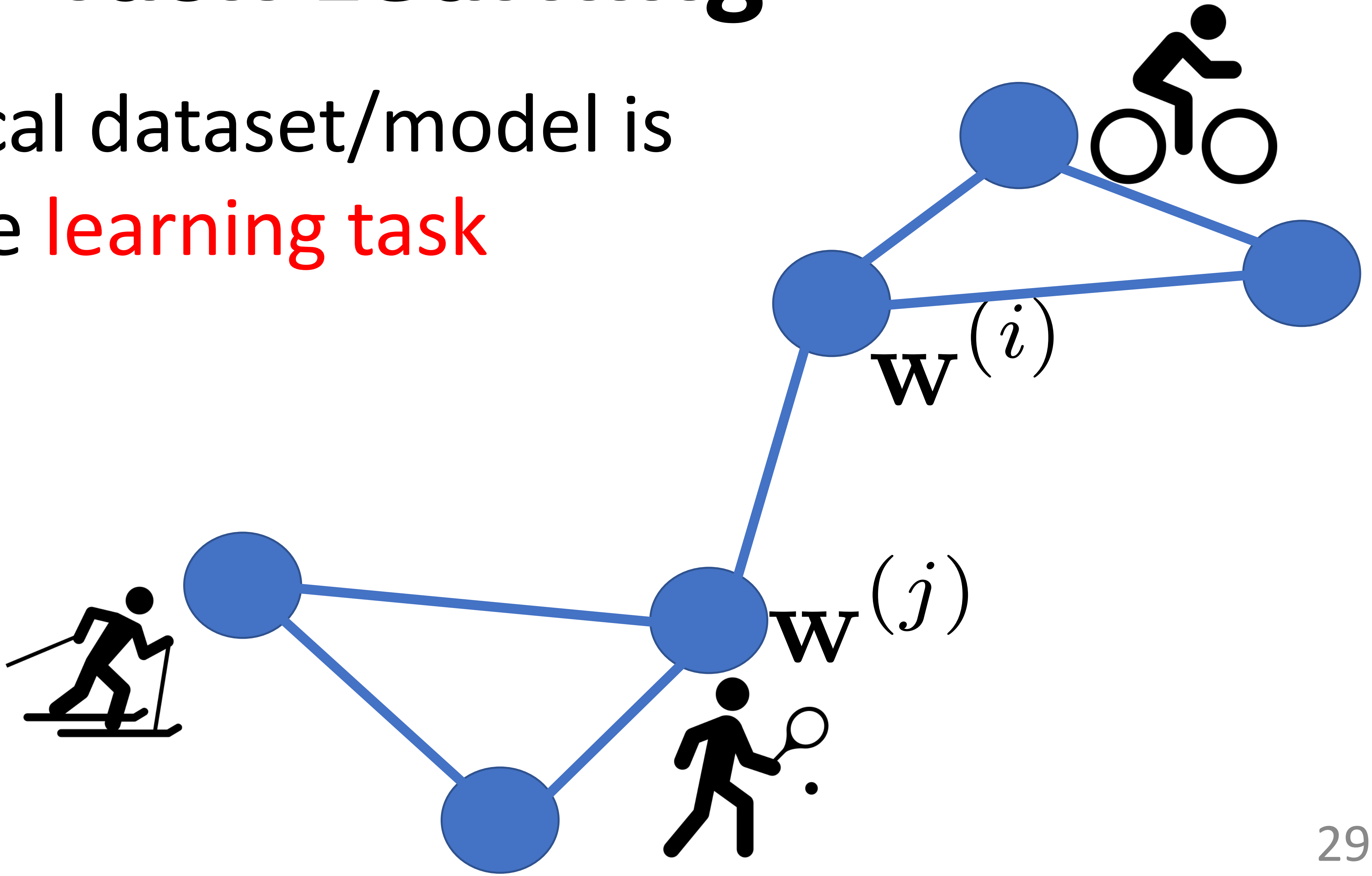
Clustering of GTVMin Solutions



Interpretations

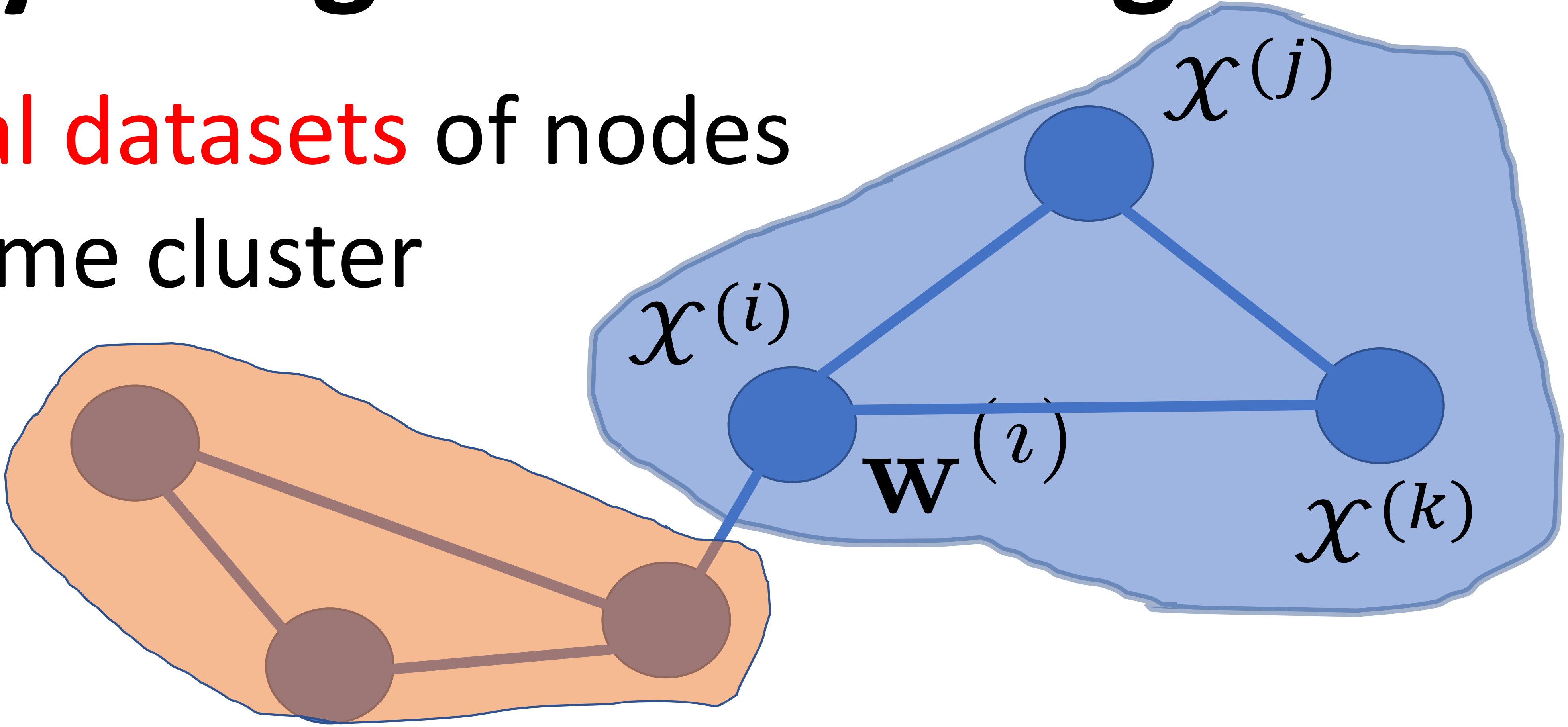
Multi-Task Learning

each local dataset/model is separate learning task



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.

Generalized Convex Clustering

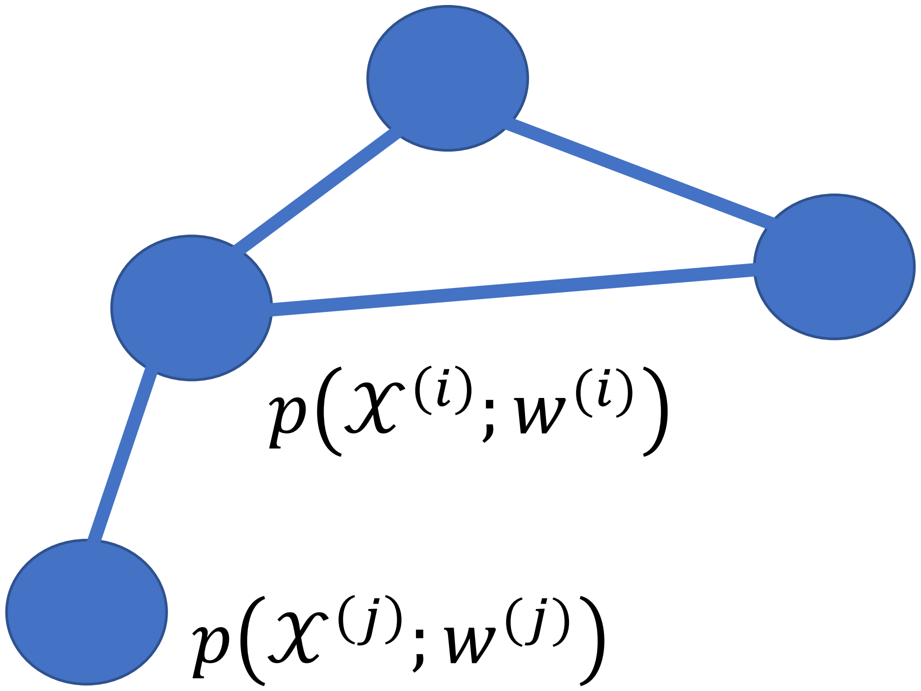
$$\min_{\mathbf{w}} \sum_{i} \left\| w^{(i)} - a^{(i)} \right\|^2 + \lambda \sum_{\{i,j\}} A_{i,j} \left\| w^{(i)} - w^{(j)} \right\|_p$$

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

(Probabilistic) Graphical Model

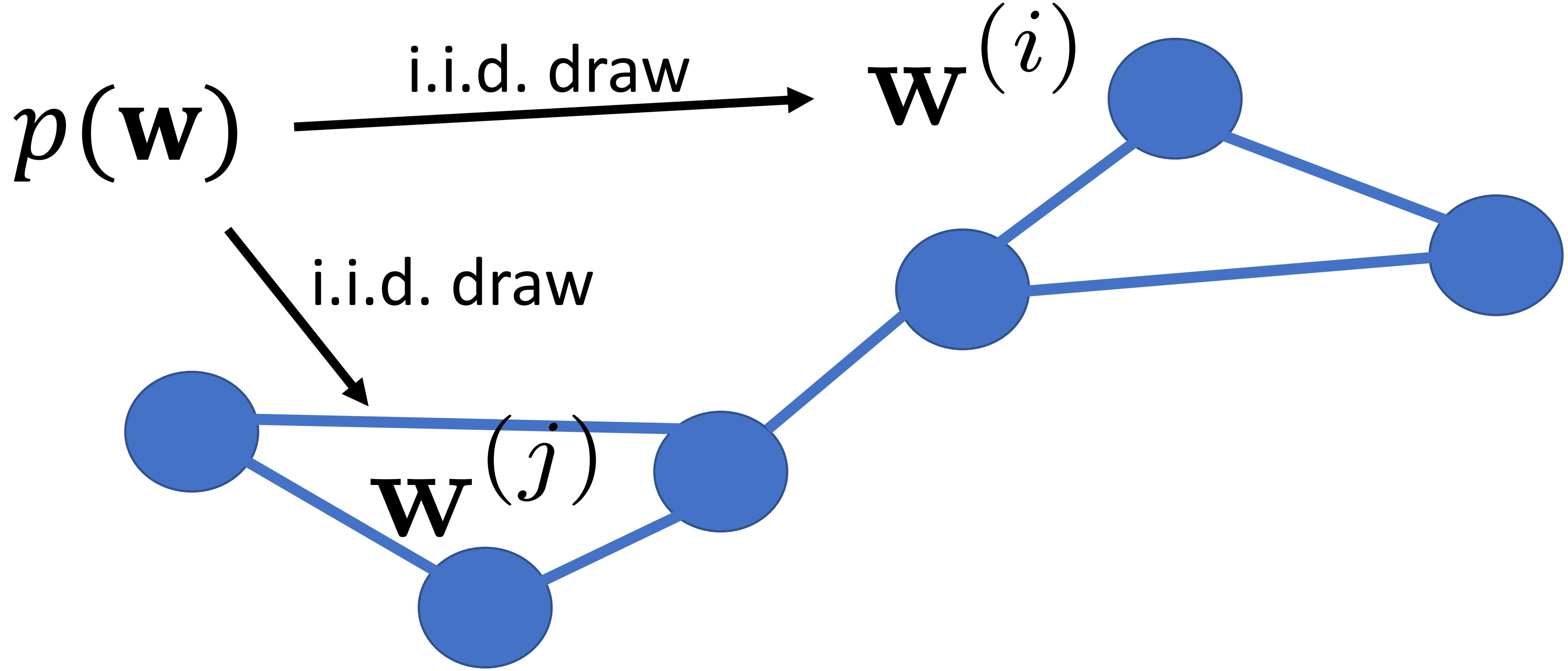
separate prob. space for each local dataset

traditionally, PGMs use a common prob. space for all local datasets



AJ, "**Networked Exponential Families for Big Data Over Networks,**" in *IEEE Access*, vol. 8, pp. 202897-202909, 2020, doi: 10.1109/ACCESS.2020.3033817.

Approx. Hierarch. Bayes' Model

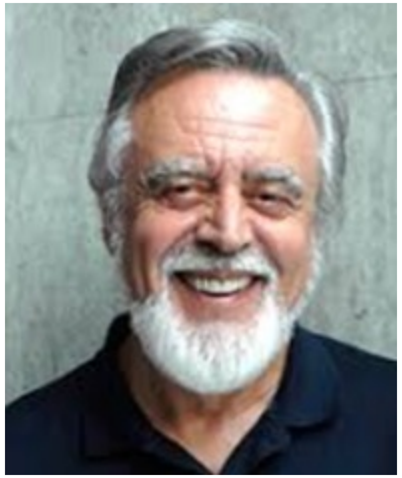


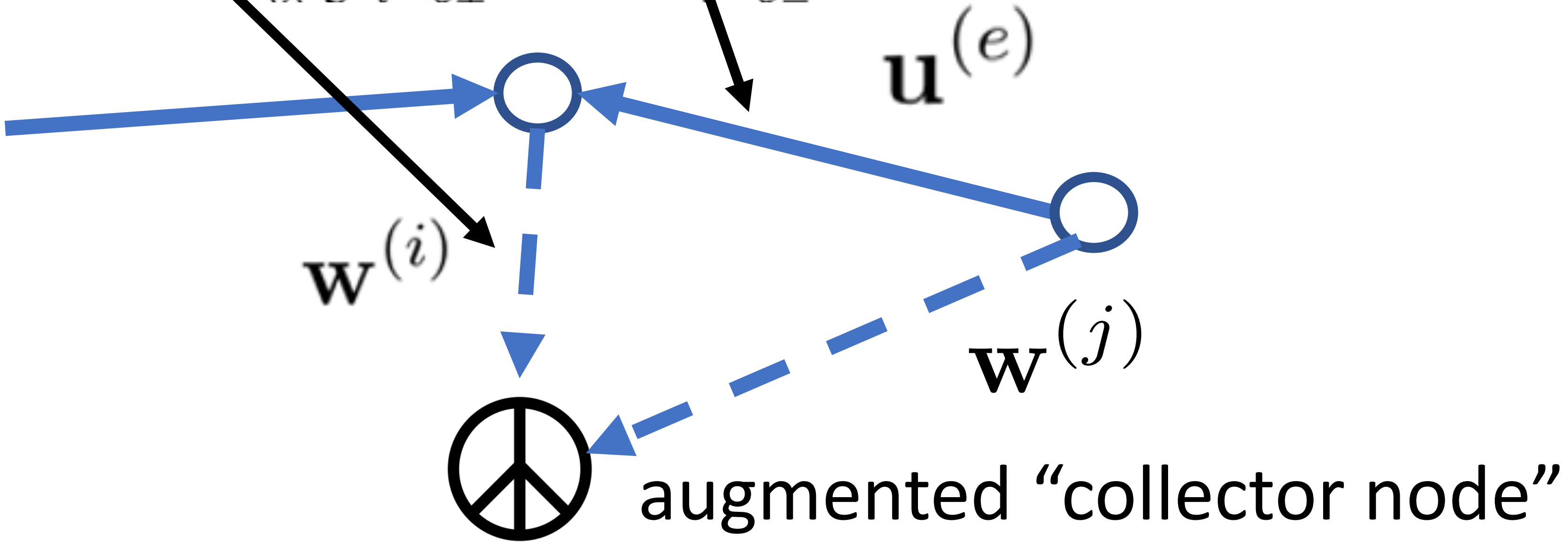
Lyu, B., Hanzely, F., and Kolar, M., "Personalized Federated Learning with Multiple Known Clusters", *arXiv e-prints*, 2022. doi:10.48550/arXiv.2204.13619.

Non-Linear Min-Cost-Flow

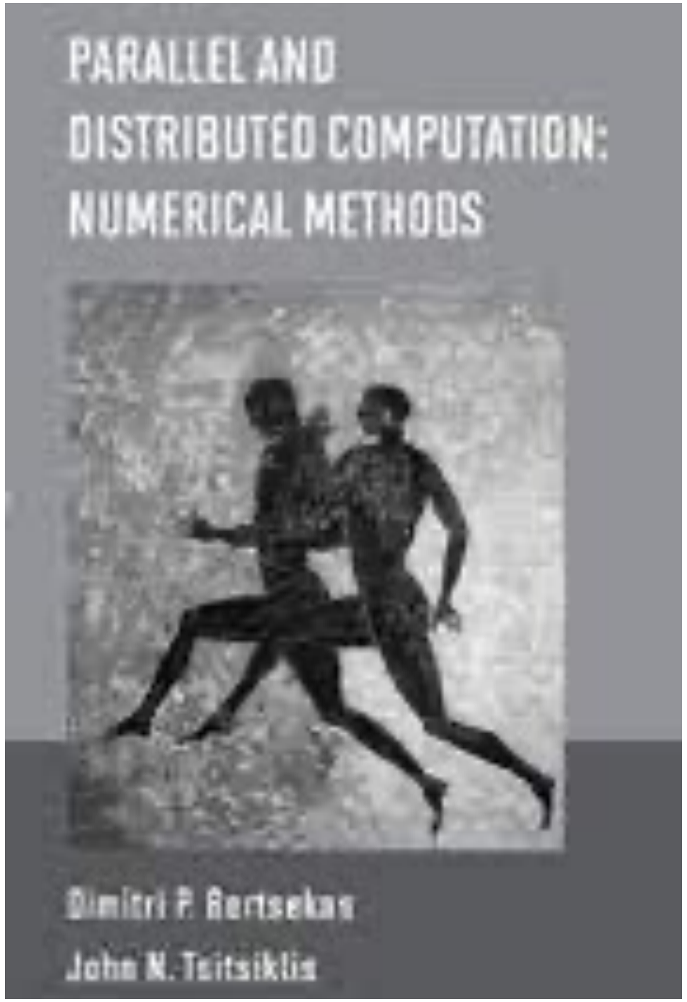
$$\max_{\mathbf{u}\in\mathcal{U}} -\sum_{i\in\mathcal{V}} L_i^*\left(\mathbf{w}^{(i)}\right) - \lambda\sum_{e\in\mathcal{E}} A_e\phi^*\left(\mathbf{u}^{(e)}/(\lambda A_e)\right)$$

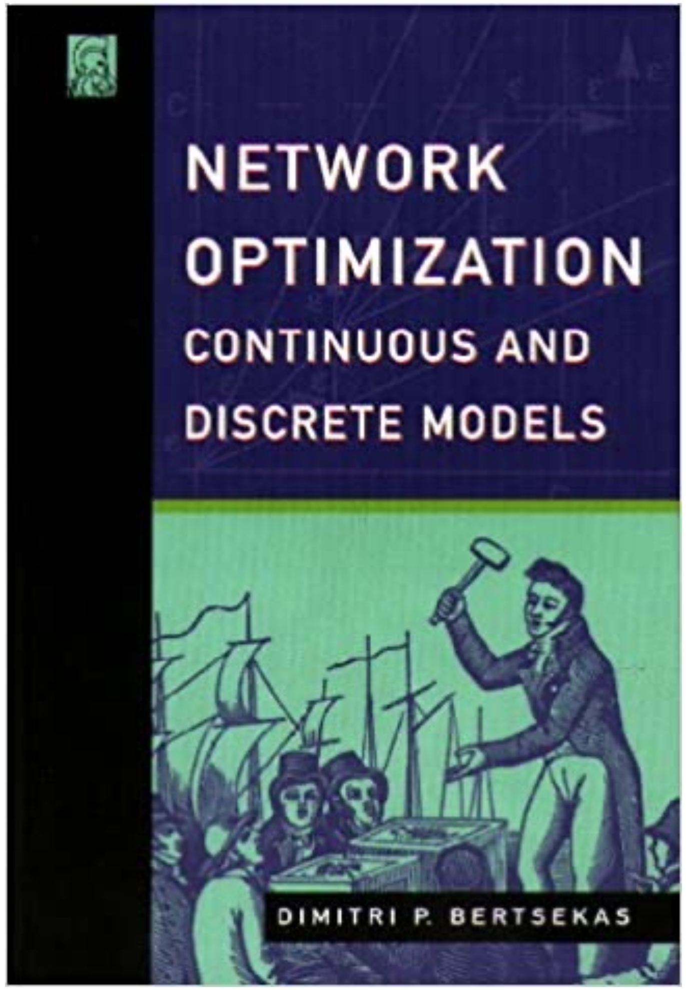
$$\text{subject to} \quad -\mathbf{w}^{(i)} = \sum_{e\in\mathcal{E}}\sum_{i=e_+}\mathbf{u}^{(e)} - \sum_{i=e_-}\mathbf{u}^{(e)} \quad \text{for all nodes } i\in\mathcal{V}.$$





Non-Linear Min-Cost-Flow





Electrical Network.
(“AI is new Electricity!”)

Kirchhoff’s Current Law

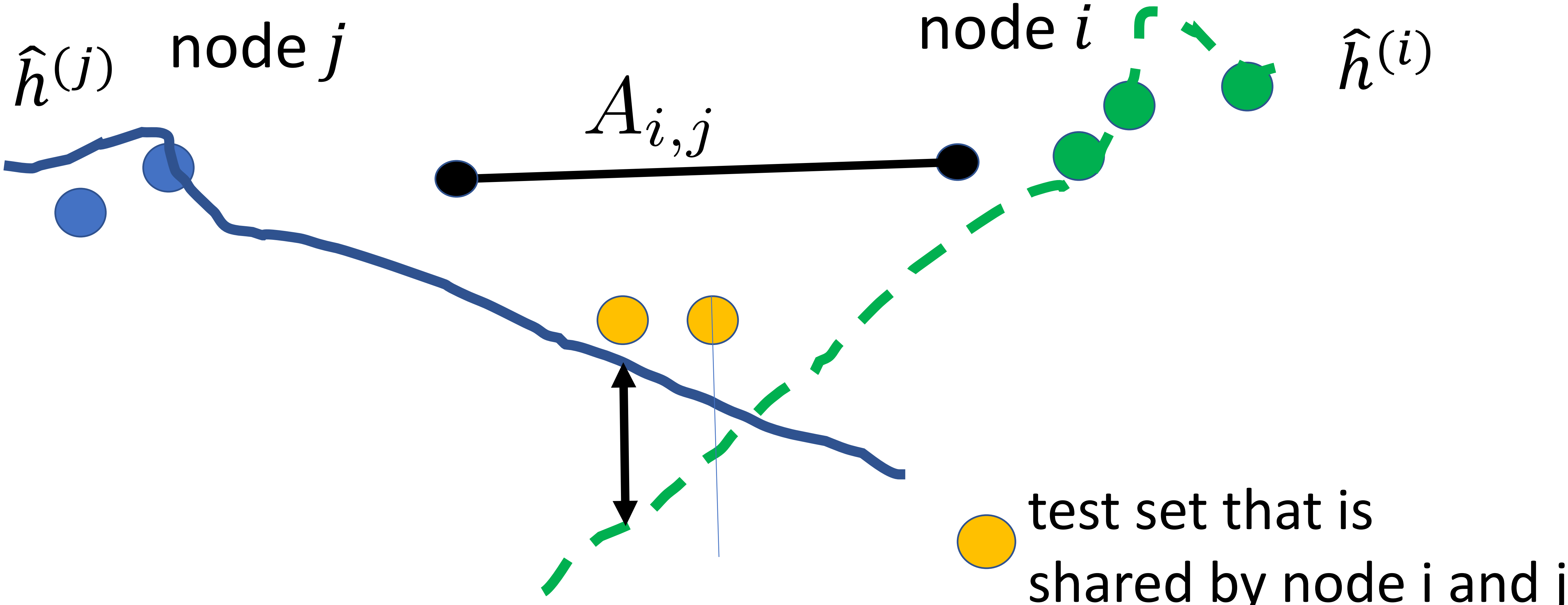
$$\sum_{e\in\mathcal{E}}\sum_{i=e_+}\widehat{\mathbf{u}}^{(e)} - \sum_{i=e_-}\widehat{\mathbf{u}}^{(e)} = -\nabla L_i\left(\widehat{\mathbf{w}}^{(i)}\right) \text{ for all nodes } i \in \mathcal{V}$$

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

Generalized Ohm Law

GTVMin
for
Non-Param. Models

Variation of Non-Param. Models



Wrap Up.

- couple local model training via regularization
- regularizer obtained via GTV (over empirical graph)
- FL algorithms = optimization methods for GTV min
- GTVmin pools local datasets into clusters
- cluster structure depends on emp.graph and local data!

Thank you for your attention!