CS-E4740 Federated Learning

"FL Design Principle"

Dipl.-Ing. Dr.techn. Alexander Jung

How Are You ? – Feedback Samples

- *Lack of feedback. The material is very interesting, but I'm not sure if I'm grasping it at an adequate depth.*

- *The quizzes and exercises help to figure out the depth required, but would need feedback.*

- *…would be nice to know the total points for quizzes even before the deadline. Would it be possible to only show total points, but not points per tasks? This would help to prevent gaming*

- *it would be better to give some explanations of the correct answers for quizzes*

- *In the second quiz (regression), the review of the first attempt was never made available. I blindly used my second attempt, "just in case", after which I still couldn't see the review...*

- *My suggestion is that number of points from quiz can be seen right after the quiz. Then it would motivate more to study the subjects more when trying to get the answers right with 2. try.*

- *Lecture notes are fine, and quizzes have some questions that feel a bit 50/50 in terms how to interpret them and some questions are terribly easy having some random number as answer of two alternative*

- *I need feedback for quizzes.*

- *It would be helpful to get some sort of feedback from the quizzes.*

- *The course is quite theoretical, I really did enjoy the approach in the machine learning and deep learning with python courses, where there were hands-on exercises in addition to the lectures.*

- *I learn best by coding the solutions. Coding exercises would be nice I wish there are also coding exercises (or at least some demos) on FL, which will be hugely beneficial for the project.*

- *Some hands-on exercises in addition to the labs could be nice. I suppose that is the purpose of the project, but I probably can't fit a whole project in my schedule.*

- *"I think having the lectures on campus in a lecture hall worked much better than the online-only lectures that we have currently had, as there was more substantial two-way communication with students."*

- *Maybe there should be less "free" points given out, e.g. even doing just two (50p) exercises in the lecture notes can get you a grade 5 to my understanding.*

Consider a dataset of 10 data points, indexed by $i = 1, 2, \ldots, 10$. The i-th data point is characterized by 20 numeric features that are stacked into the feature vector $\mathbf{x}^{(i)} = \left(x_1^{(i)}, \ldots, x_{20}^{(i)}\right)^T$. The feature vectors $\mathbf{x}^{(1)}, \ldots, \mathbf{x}^{(10)} \in \mathbb{R}^{20}$ can be reasonably well modeled as i.i.d realizations of a multivariate normal distribution with an invertible covariance matrix. We learn a linear hypothesis map by linear regression, i.e., minimizing the average squared error loss. Choose the correct answer(s) below.

☑ a.   There is always a linear map that only uses the first two features and still achieves minimum training error. ✗

☐ b.   The training error of the learnt hypothesis will be zero (up to numerical errors) with probability one.

☑ c.   There might be several different hypothesis maps which achieve the minimum training error. ✓

What are the main components of ML and how are they combined?
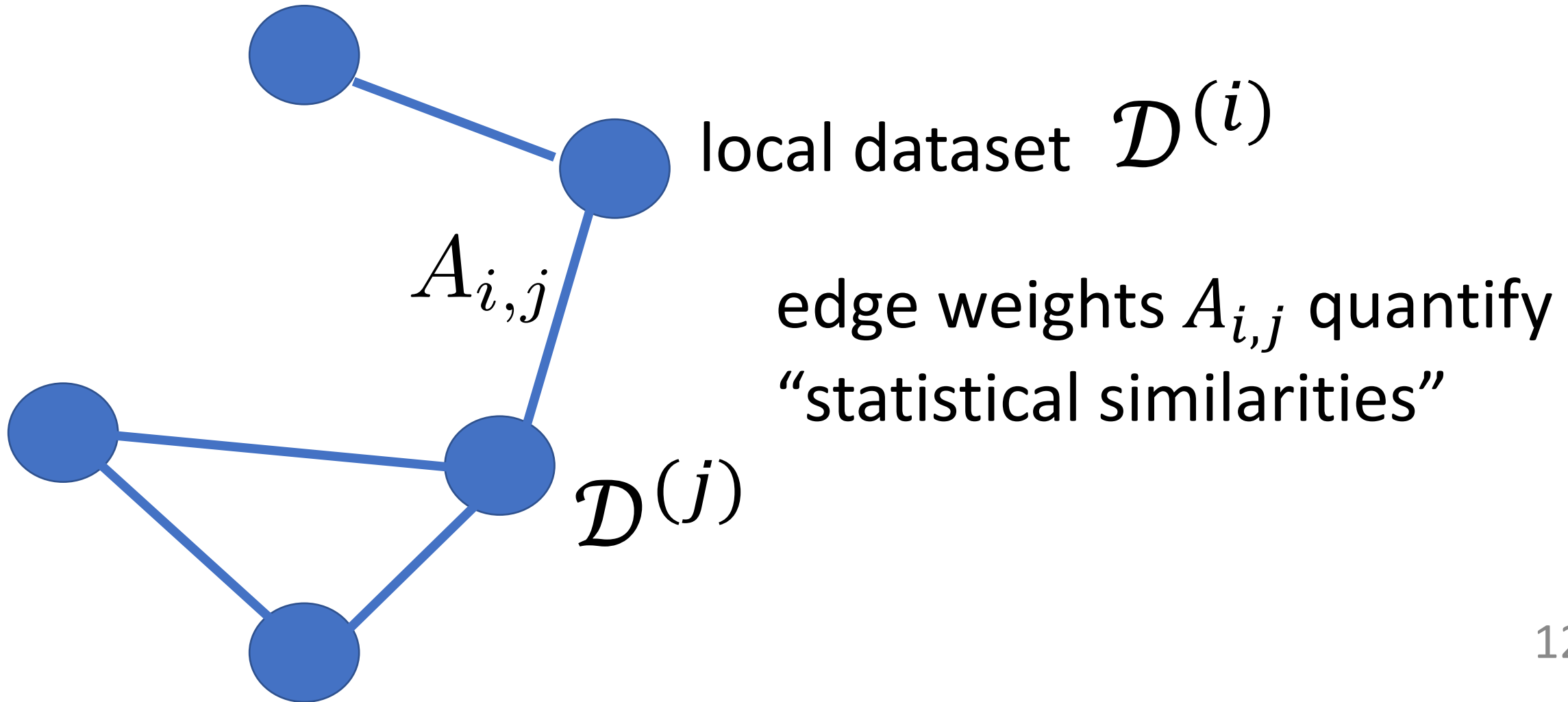
Previous Lecture:
Networked Data and Models
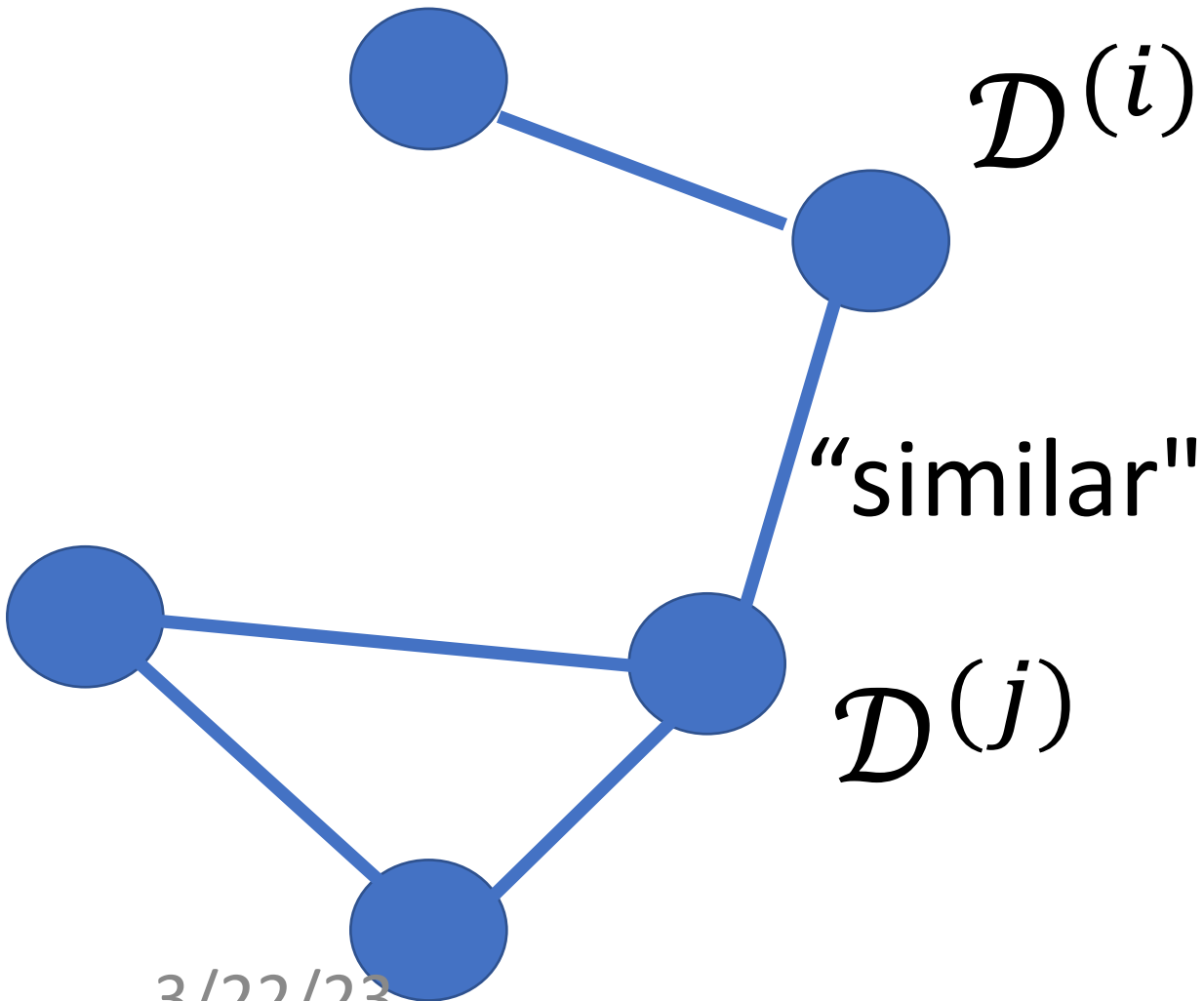
Today:
Loss and Optimization

Weather Stations

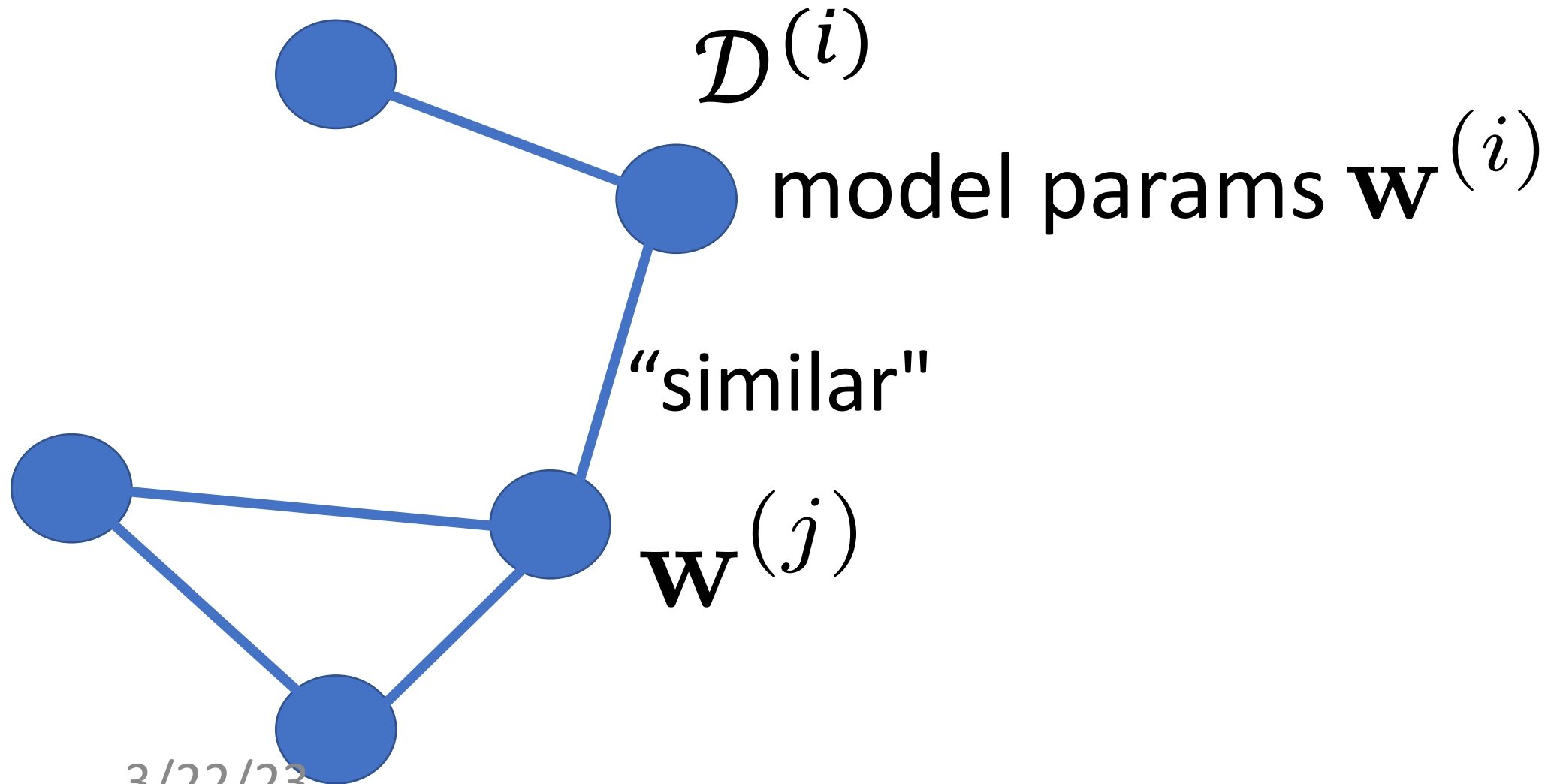FINNISH METEOROLOGICAL
INSTITUTE

The Empirical Graph

local dataset $\mathcal{D}^{(i)}$

$A_{i,j}$

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

$\mathcal{D}^{(j)}$

Networked Models.

$\mathcal{D}^{(i)}$

"similar"

$\mathcal{D}^{(j)}$

local model for each node

couple models at connected nodes

Local Parametric Models

$\mathcal{D}^{(i)}$

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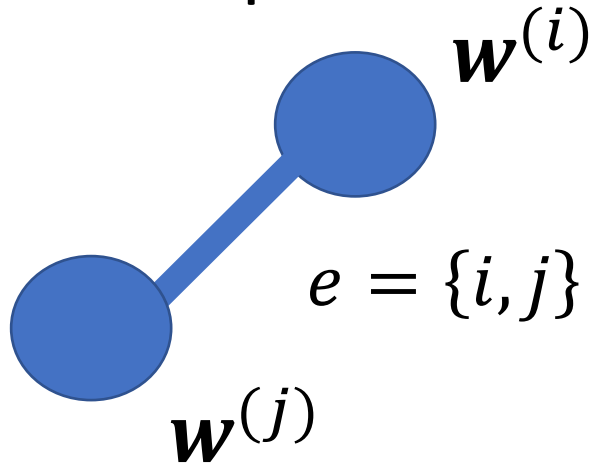
"similar"

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

Clustering Assumption
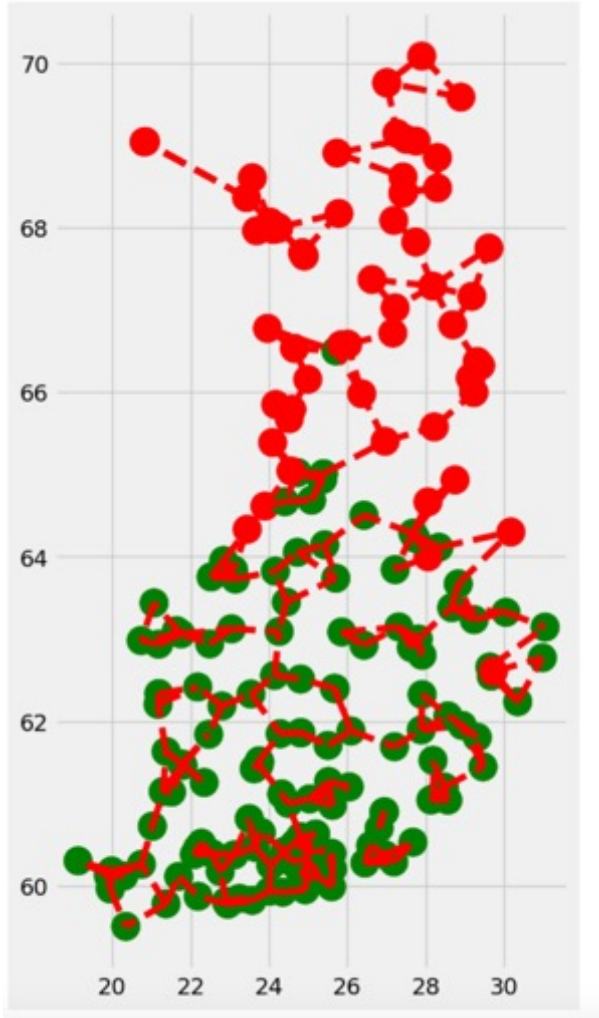
the local datasets form clusters

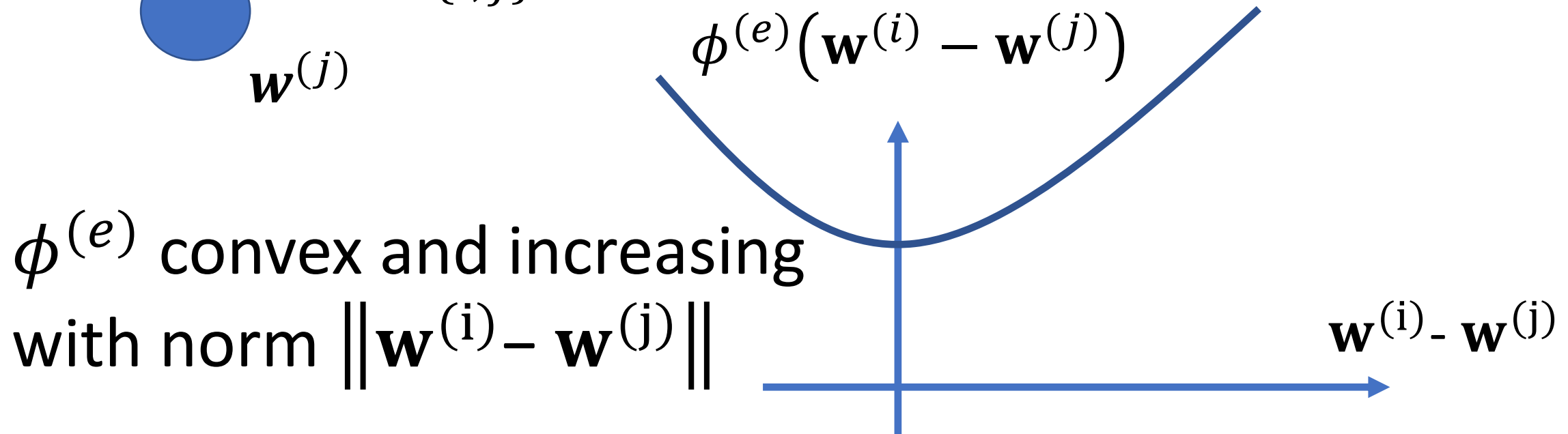datasets in same can be approximated as realizations of i.i.d. RVs with prob. dist p(x,y;c)

more edges inside clusters
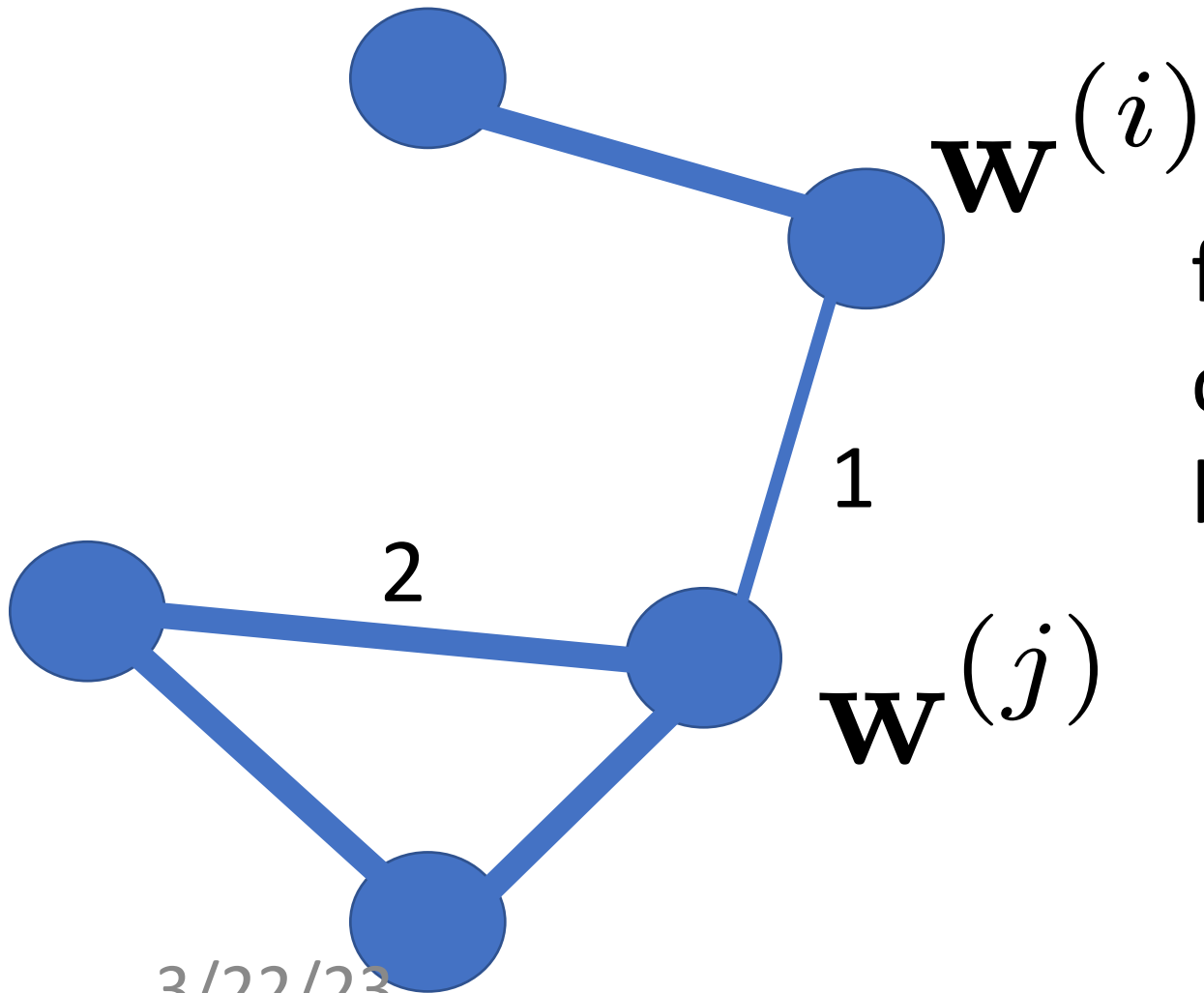
Measure Clustering via Variation

local model params

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

$e = \{i, j\}$

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

require similar params at ends of edge e

penalty function measures "variation"

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

$\phi^{(e)}$ convex and increasing
with norm $\|\mathbf{w}^{(i)} - \mathbf{w}^{(j)}\|$

$\mathbf{w}^{(i)} - \mathbf{w}^{(j)}$

Generalized Total Variation (GTV)

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

$$\sum_{\{i,j\}} A_{i,j} \phi\big(\mathbf{w}^{(i)} - \mathbf{w}^{(j)}\big)$$
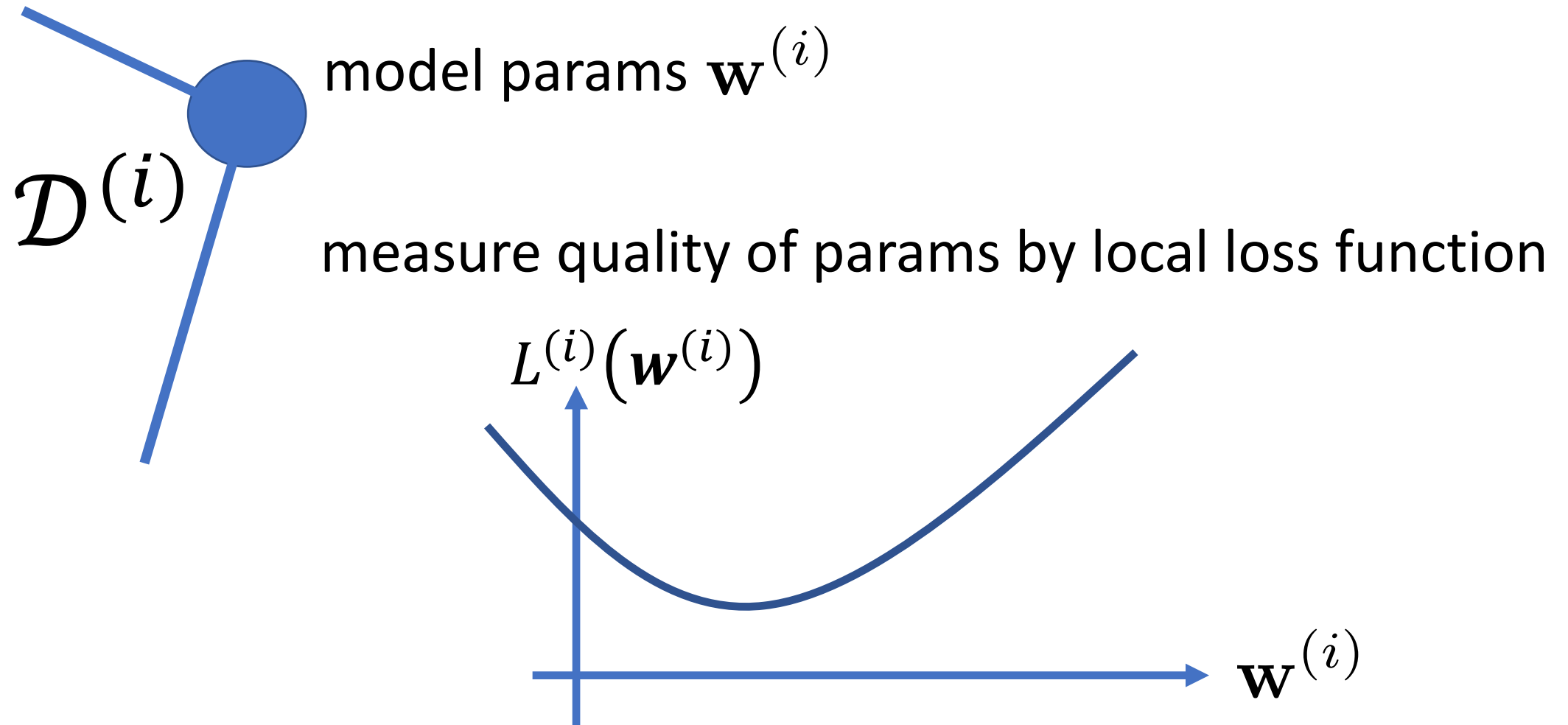
Two Special Cases of GTV

total variation $\phi(\mathbf{u}) = \|\mathbf{u}\|_2$
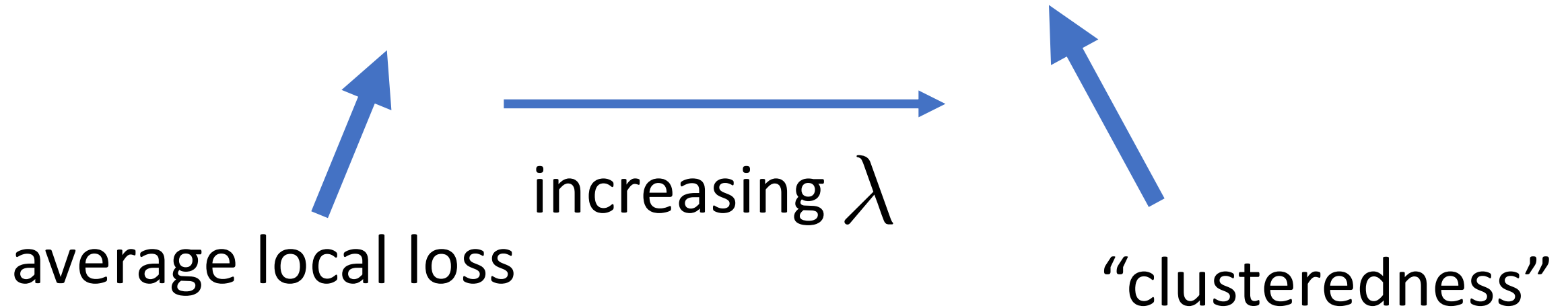
graph Laplacian quadratic from is GTV with

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

Local Loss Function

$\mathcal{D}^{(i)}$

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

measure quality of params by local loss function

$L^{(i)}\big(\boldsymbol{w}^{(i)}\big)$

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

GTV Minimization

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

average local loss

increasing λ

"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.
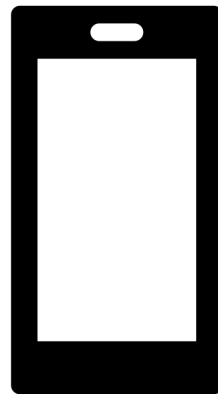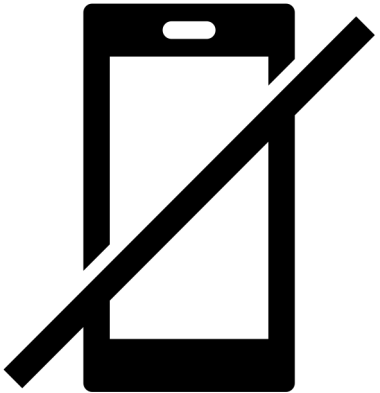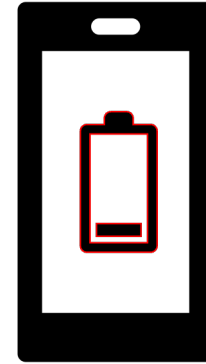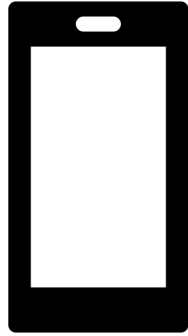
Two Key Questions of ML

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

- computational aspects: how to compute (approximate) solutions efficiently ?

- 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.

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

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

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

Statistical Aspects

GTV Min. for Local Lin.Reg.

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

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

$$\Sigma_{\{i,j\}} A_{i,j} \|\mathbf{w}^{(i)} - \mathbf{w}^{(j)}\|^2 = \mathbf{w}^T (\mathbf{L} \otimes \mathbf{I}) \, \mathbf{w}$$
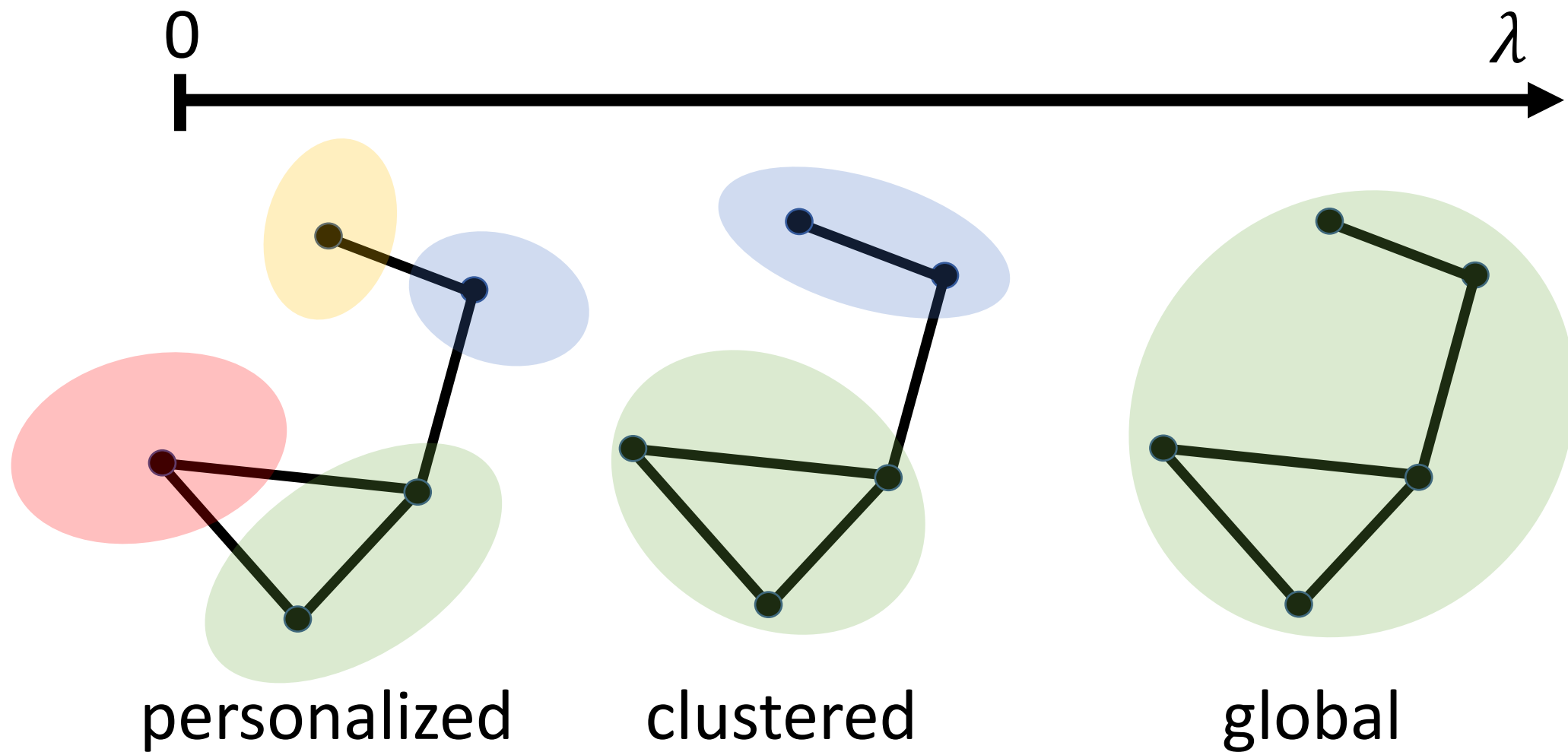
with the graph Laplacian $\mathbf{L}$

Spectral Clustering

for large λ, GTVMin is to minimize

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

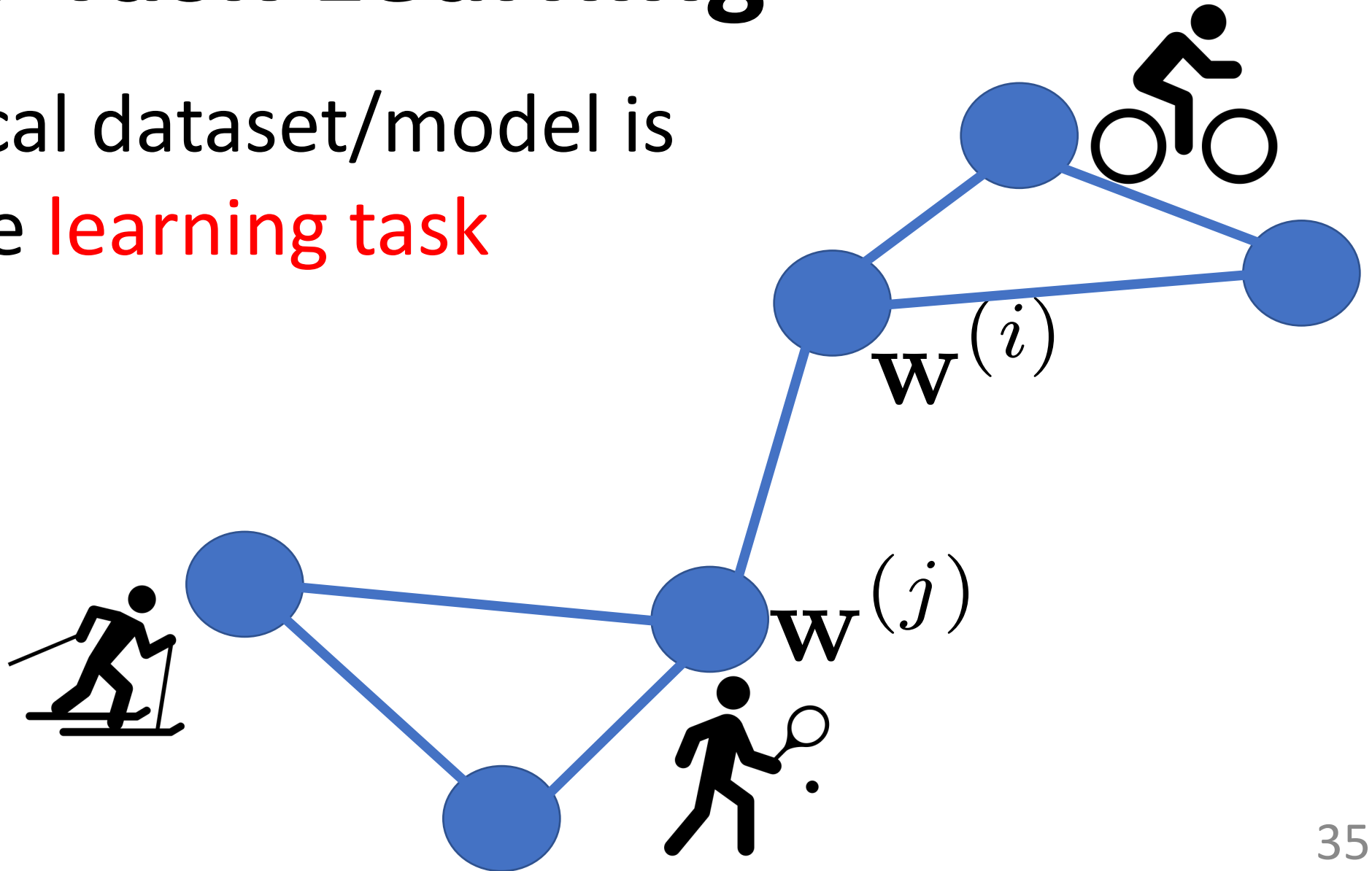$\Rightarrow$ local model parameters composed of eigvecs. of L corresponding to smallest eig.vals
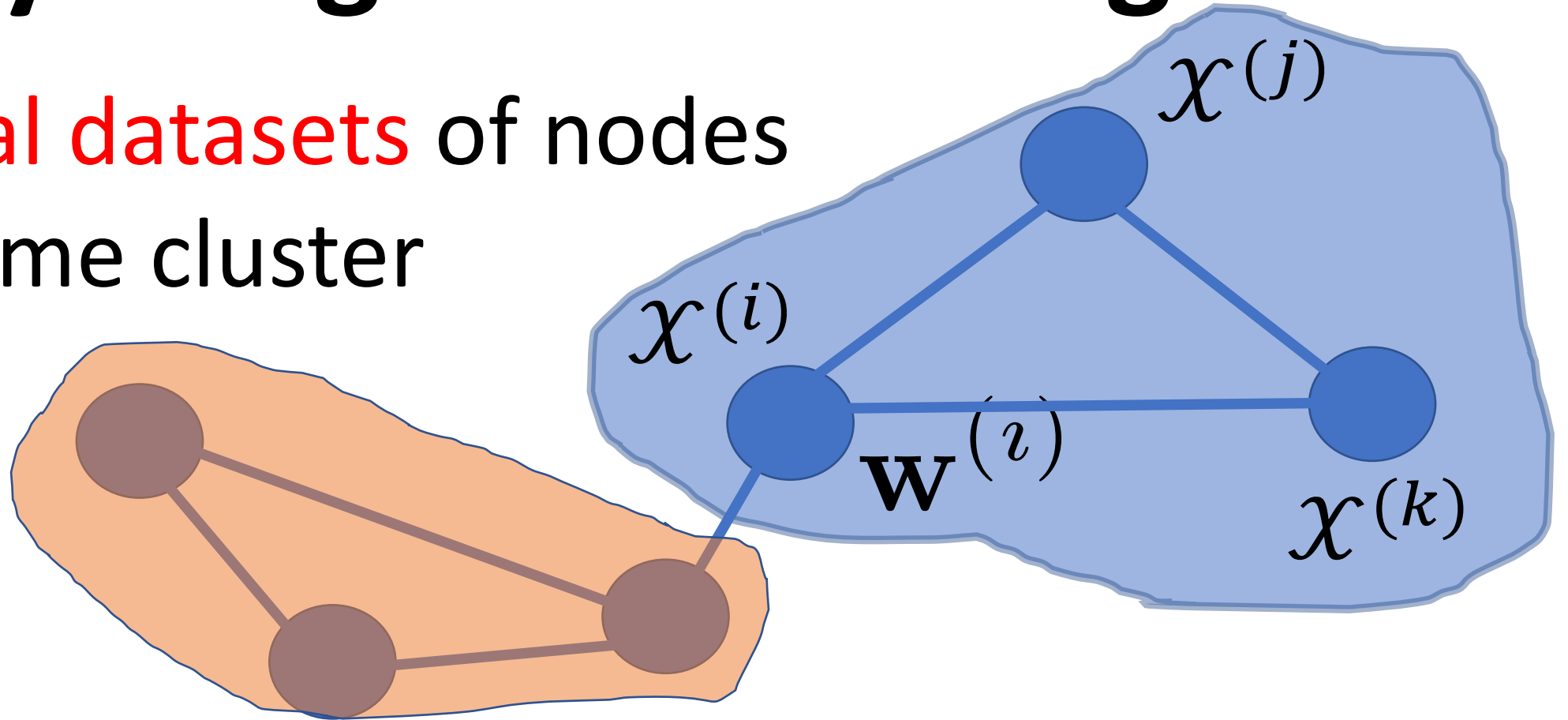
Clustering of GTVMin Solutions

Interpretations

Multi-Task Learning

each local dataset/model is
separate learning task

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

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

Locally Weighted Learning

pool local datasets of nodes

in the same cluster

$\mathcal{X}^{(j)}$

$\mathcal{X}^{(i)}$

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

$\mathcal{X}^{(k)}$

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

$$\min_{\mathbf{w}} \sum_i \left\| w^{(i)} - a^{(i)} \right\|^2 + \lambda \sum_{\{i,j\}} A_{i,j} \left\| w^{(i)} - w^{(j)} \right\|_p$$
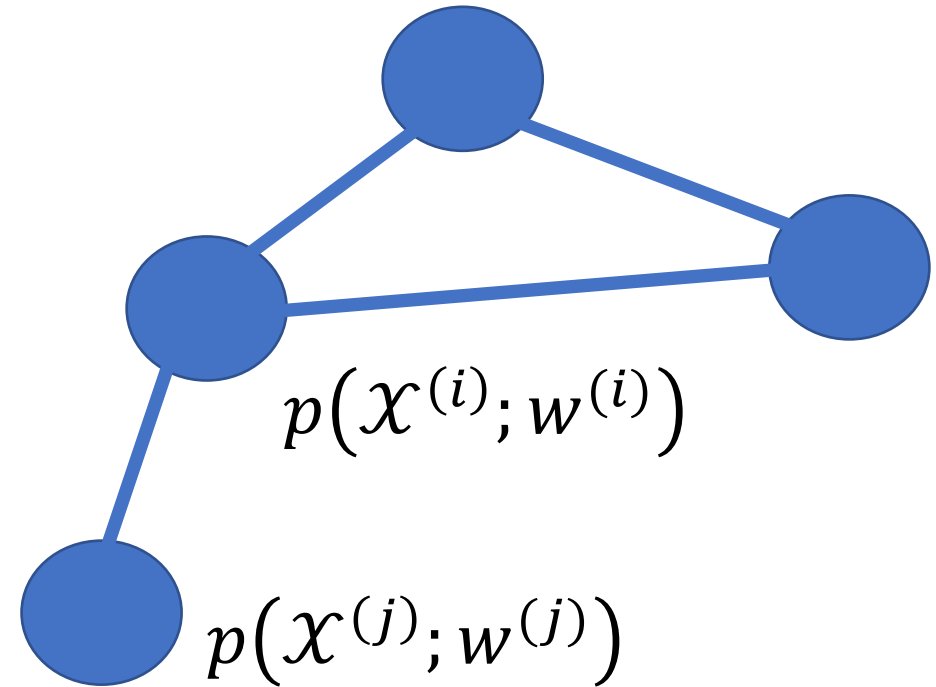
D. Sun, K.-C. Toh, Y. Yuan;
Convex Clustering: Model, Theoretical Guarantee and Efficient Algorithm, JMLR, 22(9):1–32, 2021
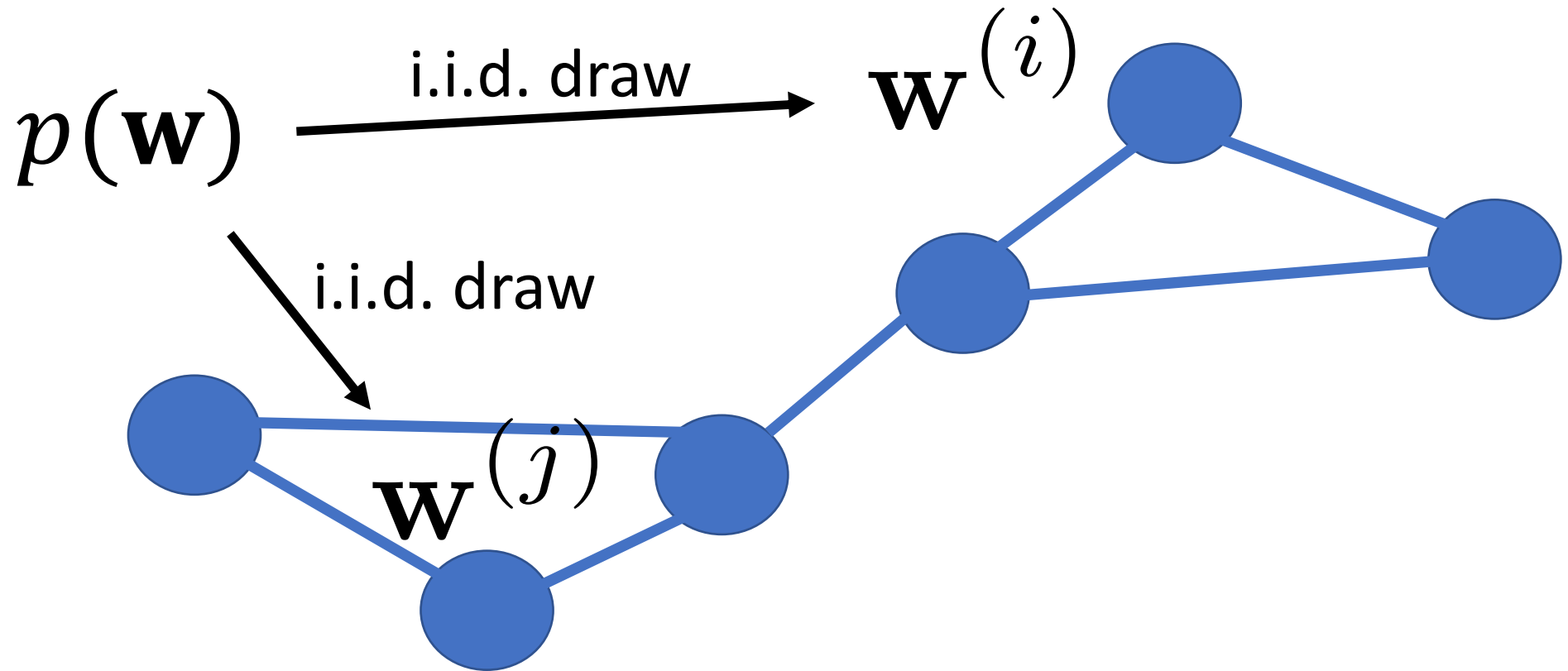
(Probabilistic) Graphical Model

separate prob. space for each local dataset

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

$$p(\mathcal{X}^{(i)}; w^{(i)})$$

$$p(\mathcal{X}^{(j)}; w^{(j)})$$

Approx. Hierarch. Bayes' Model

$p(\mathbf{w})$ $\xrightarrow{\text{i.i.d. draw}}$ $\mathbf{w}^{(i)}$
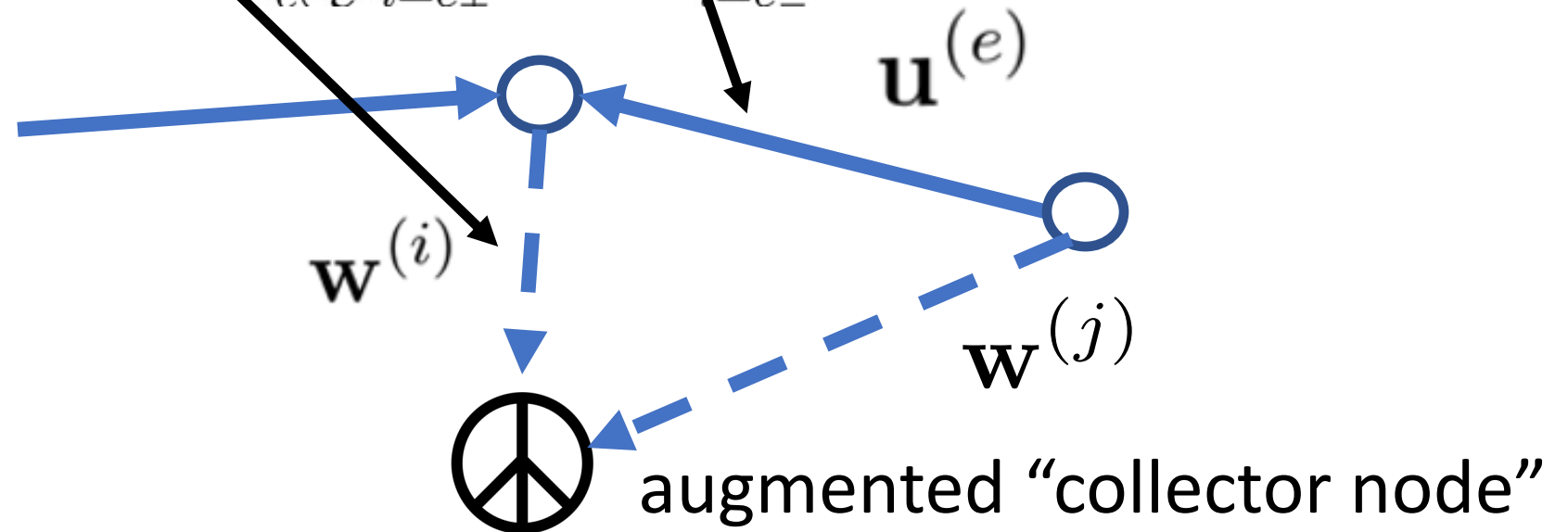
i.i.d. draw

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

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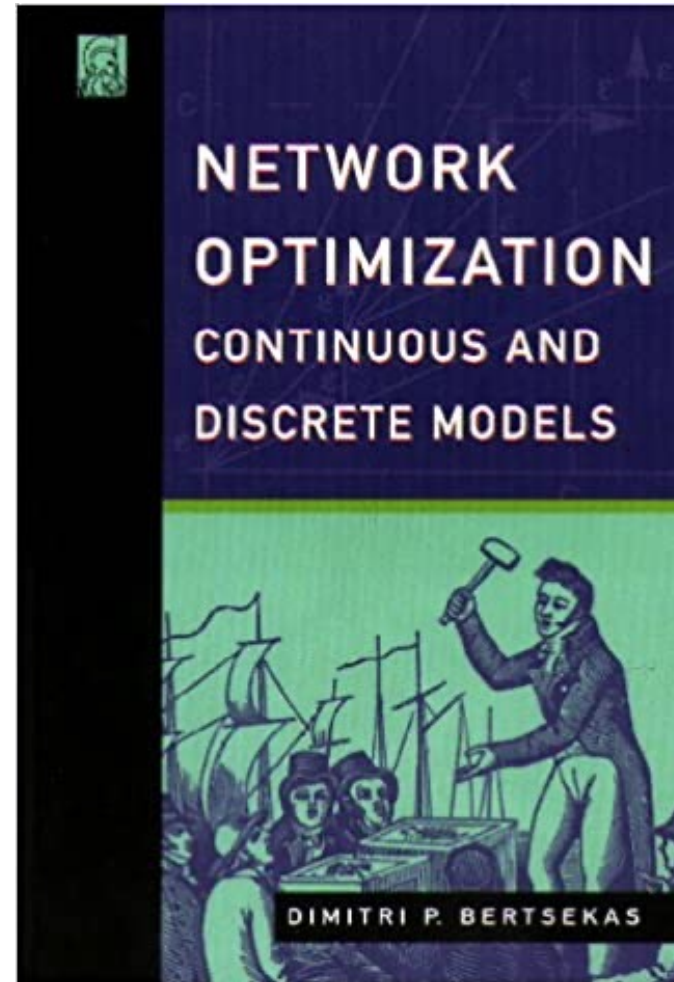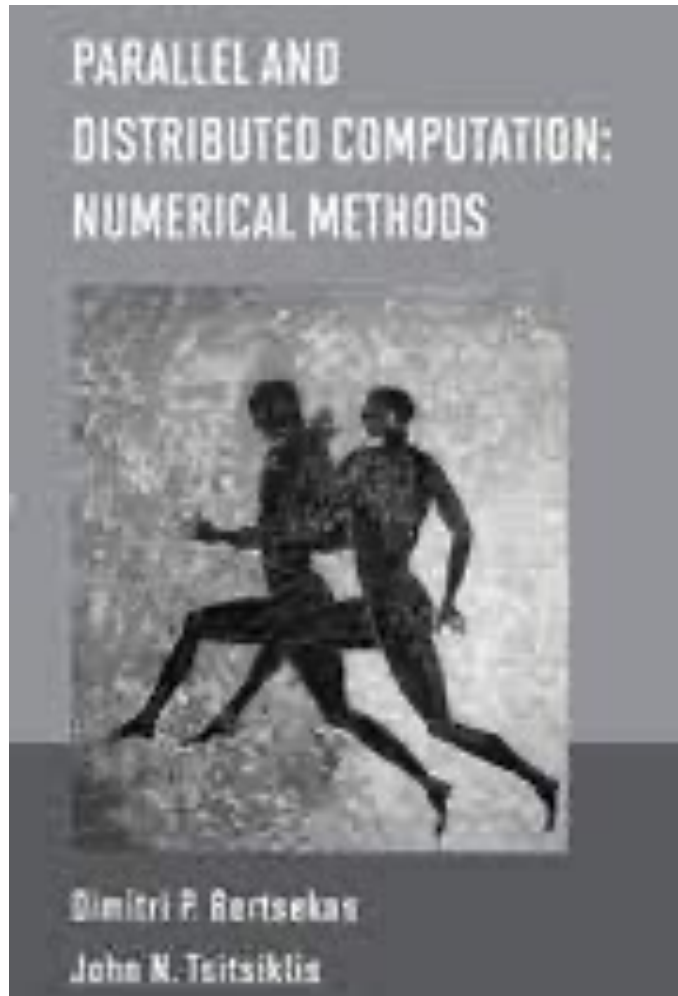
Non-Linear Min-Cost-Flow

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

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

$\mathbf{u}^{(e)}$

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

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

augmented "collector node"

Non-Linear Min-Cost-Flow

Electrical Network. ("AI is new Electricity!")

Kirchhoff's Current Law

$$\sum_{e \in \mathcal{E}} \sum_{i=e_+} \widehat{\mathbf{u}}^{(e)} - \sum_{i=e_-} \widehat{\mathbf{u}}^{(e)} = -\nabla L_i\left(\widehat{\mathbf{w}}^{(i)}\right) \text{ for all nodes } i \in \mathcal{V}$$
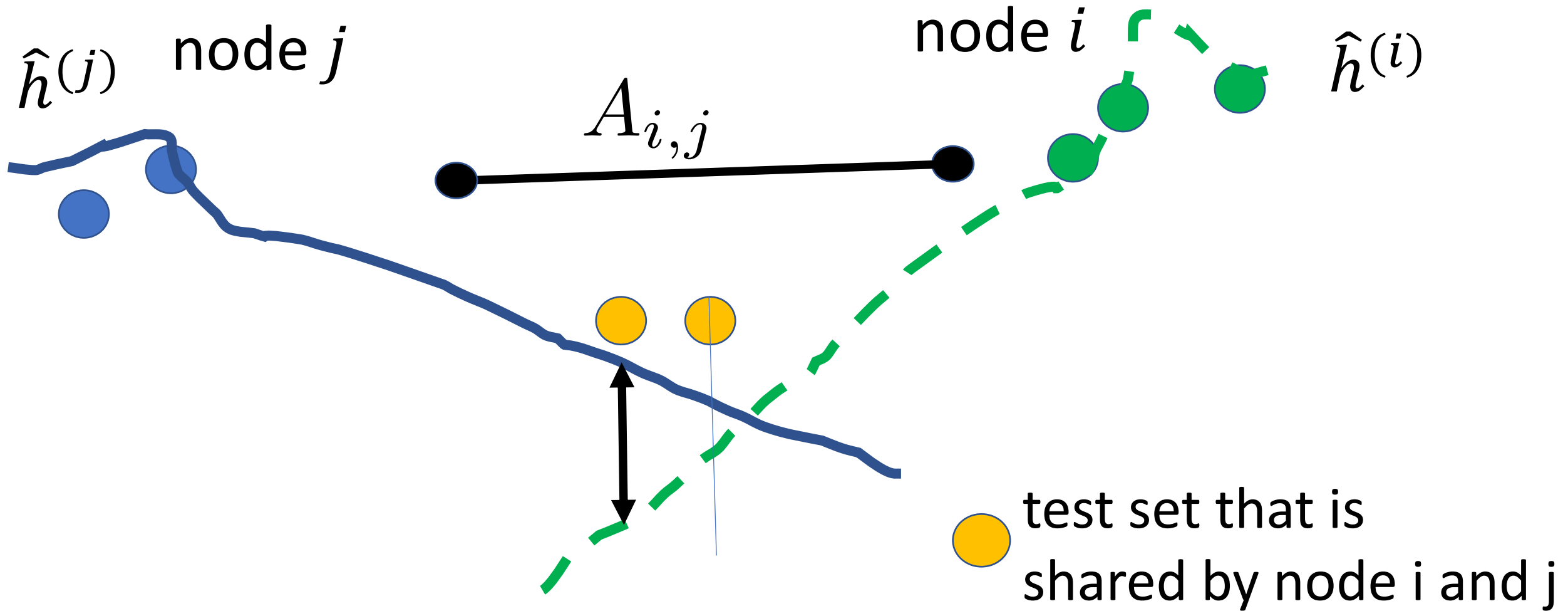
$$\widehat{\mathbf{w}}^{(e_+)} - \widehat{\mathbf{w}}^{(e_-)} \in (\lambda A_e)\partial\phi^*(\widehat{\mathbf{u}}^{(e)}/(\lambda A_e)) \text{ for every edge } e \in \mathcal{E}.$$

Generalized Ohm Law

GTVMin for Non-Param. Models

Variation of Non-Param. Models

$\hat{h}^{(j)}$ node j node i $\hat{h}^{(i)}$

$A_{i,j}$

test set that is shared by node i and j

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!