Network Flows in Federated Learning

Alexander Jung (Aalto University)

linkedin.com/in/aljung

@alexjung111

guiding theme:

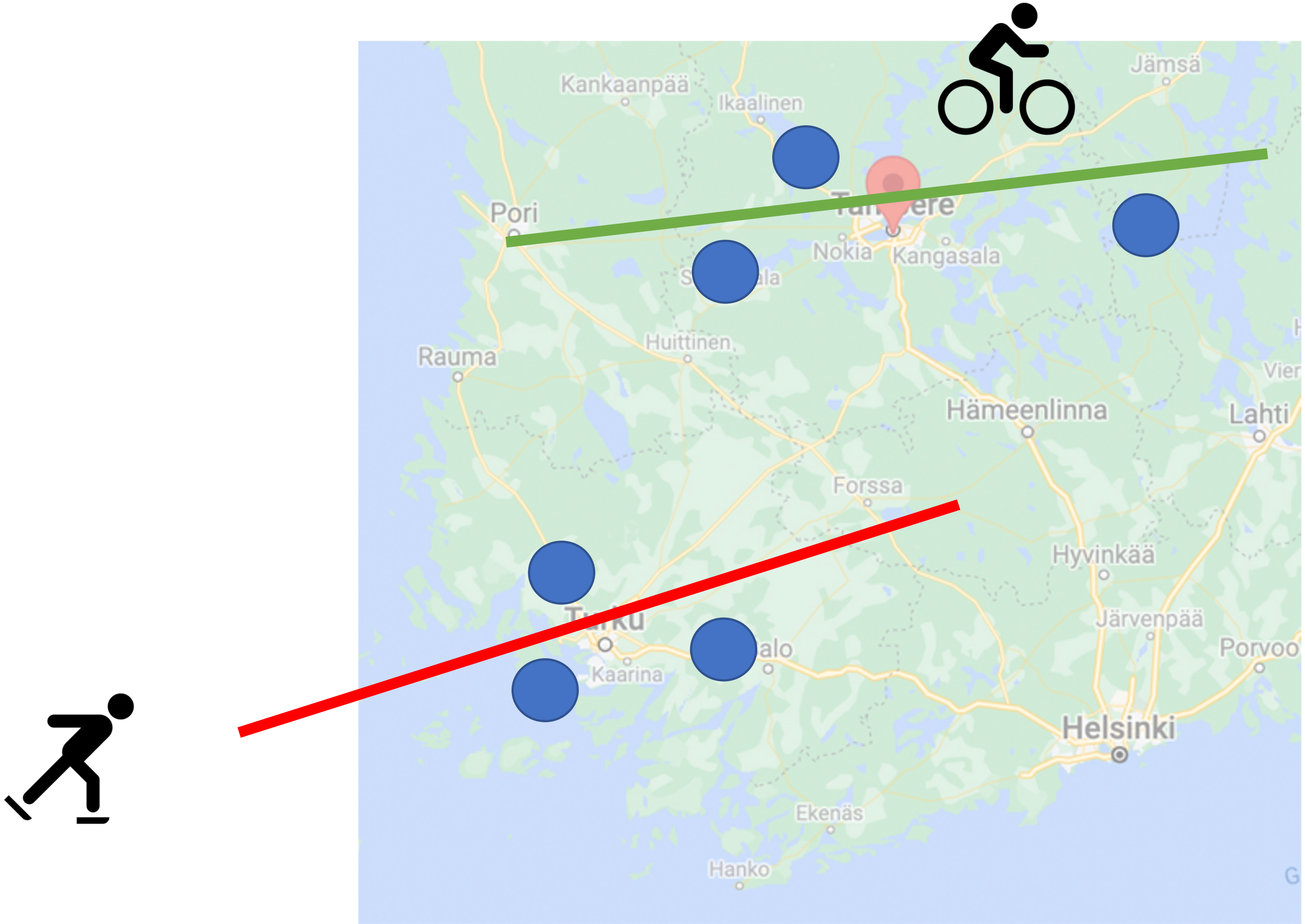
organize data, models and computation for

machine learning as networks.

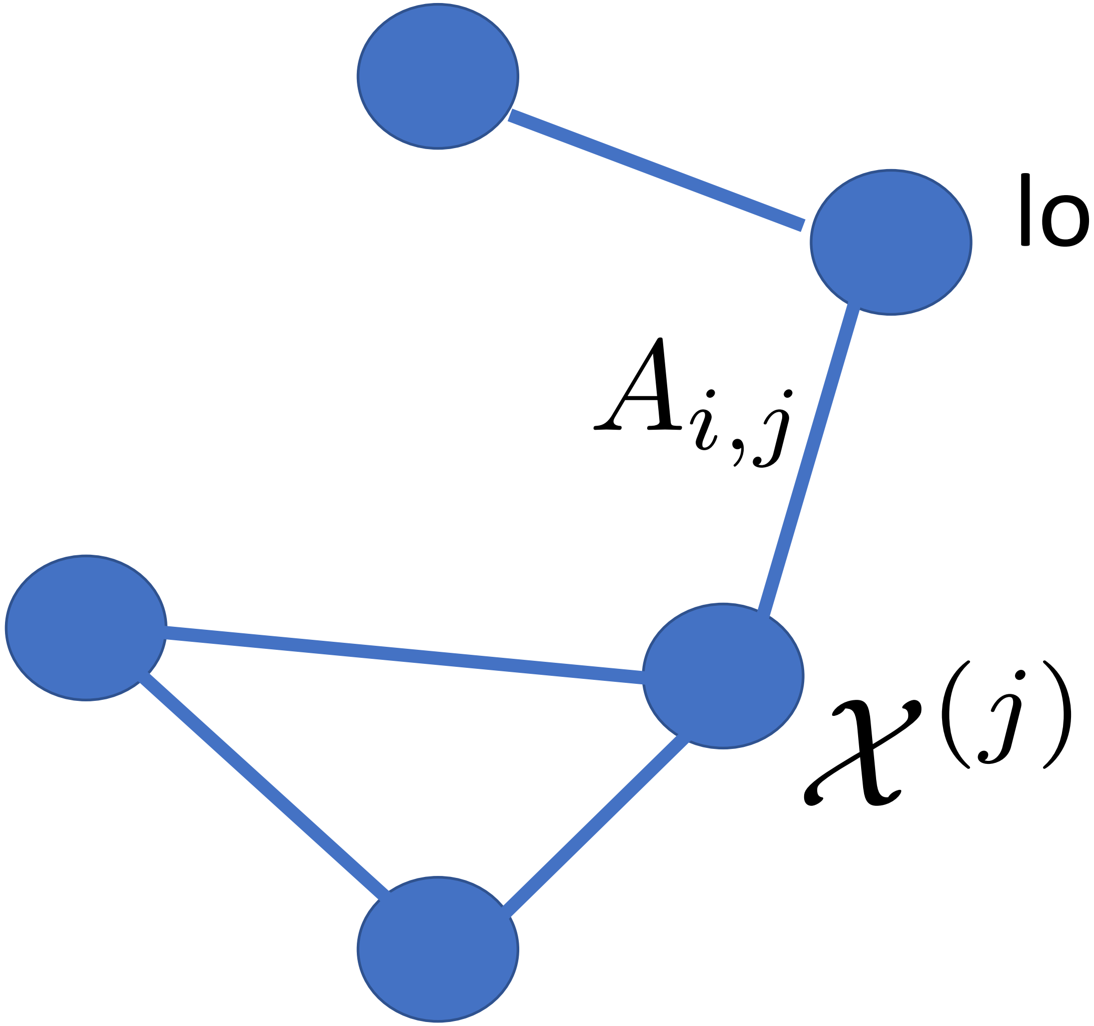
Weather Stations.

FINNISH METEOROLOGICAL INSTITUTE

Personalized Weather Forecast



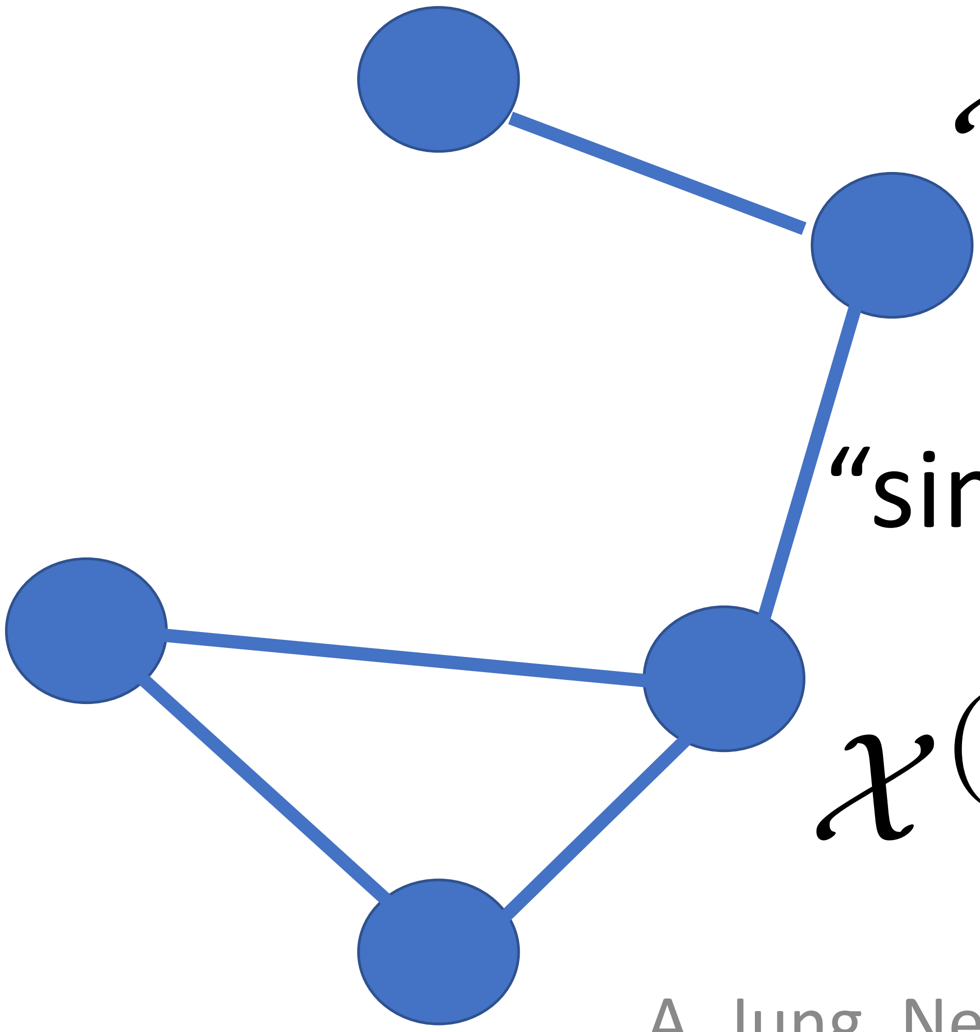
The Empirical Graph



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

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

Networked Models.



local model for each node

couple models at
connected nodes

TV Minimization

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

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

Rewrite GTVMin

$$\widehat{\mathbf{w}} \in \arg\min_{\mathbf{w} \in \mathcal{W}} \; f(\mathbf{w}) + g(\mathbf{D}\mathbf{w})$$

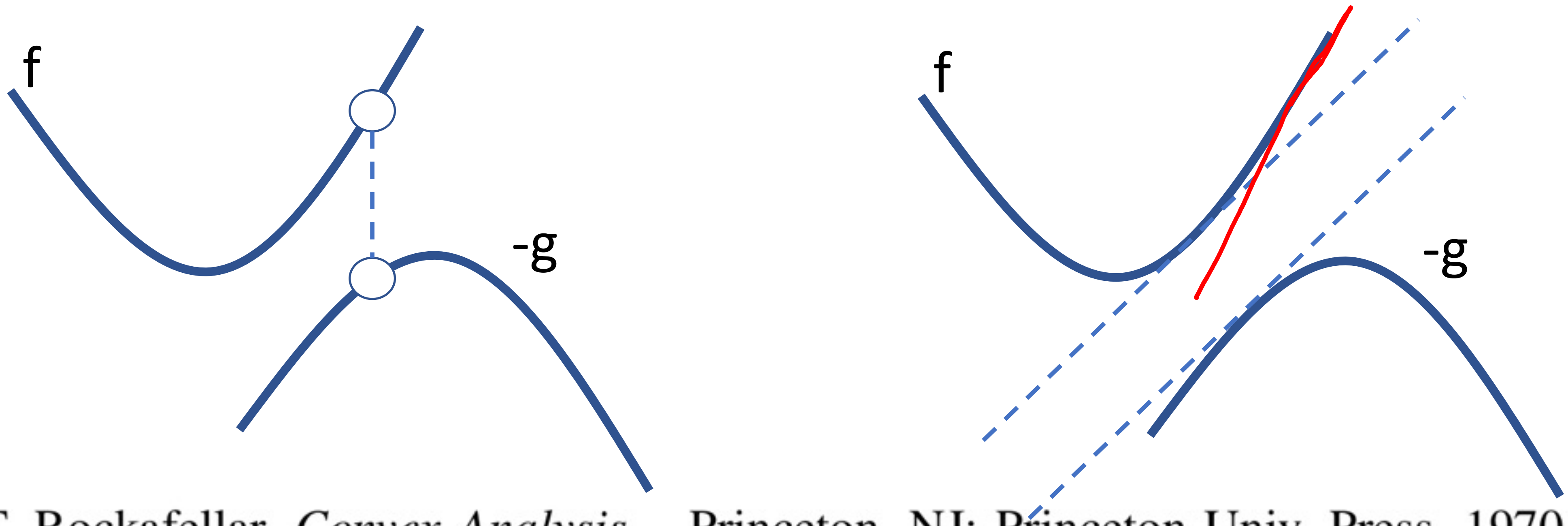
$$\text{with } f(\mathbf{w}) := \sum_{i \in \mathcal{V}} L_i\left(\mathbf{w}^{(i)}\right) \;, \text{ and } g(\mathbf{u}) := \lambda \sum_{e \in \mathcal{E}} A_e \phi\left(\mathbf{u}^{(e)}\right).$$

with incidence matrix/operator

$$\mathbf{D} : \mathcal{W} \rightarrow \mathcal{U} : \mathbf{w} \mapsto \mathbf{u} \text{ with } \mathbf{u}^{(e)} = \mathbf{w}^{(e_+)} - \mathbf{w}^{(e_-)}.$$

Fenchel's Duality.

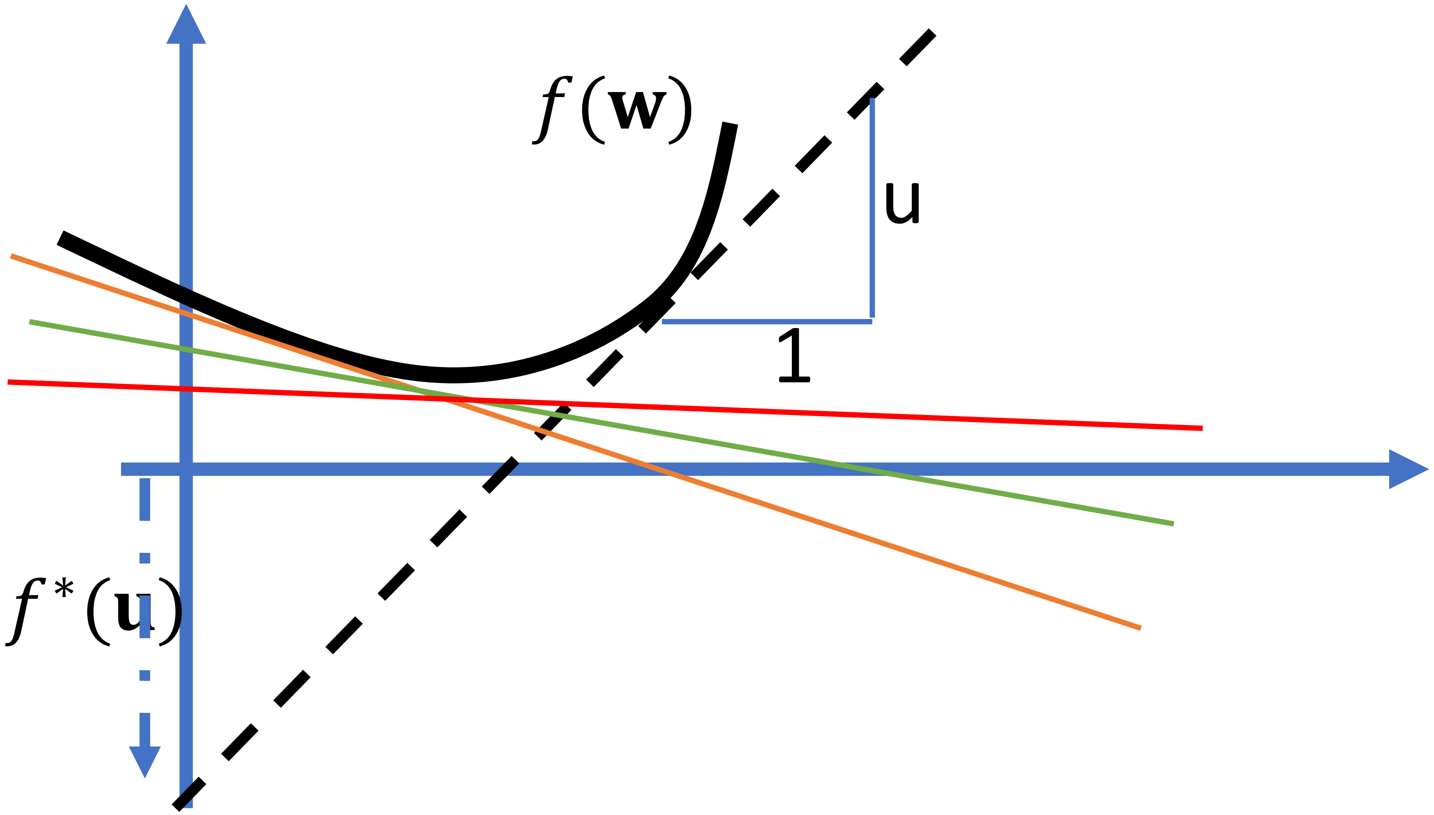
$$\min_{\mathbf{w}\in\mathcal{W}} f(\mathbf{w}) + g(\mathbf{D}\mathbf{w}) = \max_{\mathbf{u}\in\mathcal{U}} -g^*(\mathbf{u}) - f^*(-\mathbf{D}^T\mathbf{u}).$$



R. T. Rockafellar, *Convex Analysis*. Princeton, NJ: Princeton Univ. Press, 1970.
https://en.wikipedia.org/wiki/Fenchel%27s_duality_theorem

Convex Conjugate.

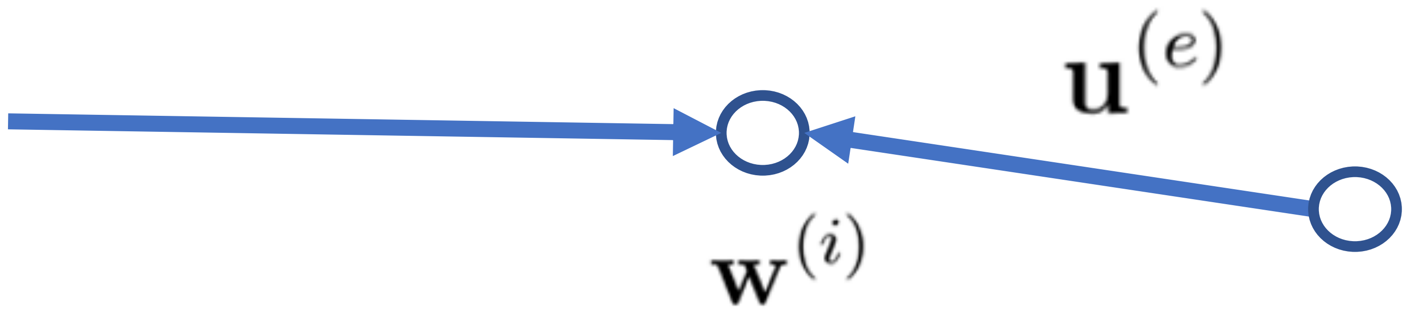
$$f^*(\mathbf{w}) := \sup_{\mathbf{z} \in \mathbb{R}^{n|\mathcal{V}|}} \mathbf{w}^T\mathbf{z} - f(\mathbf{z}) \qquad g^*(\mathbf{u}) := \sup_{\mathbf{z} \in \mathbb{R}^{n|\mathcal{E}|}} \mathbf{u}^T\mathbf{z} - g(\mathbf{z})$$



The Dual of GTVMin.

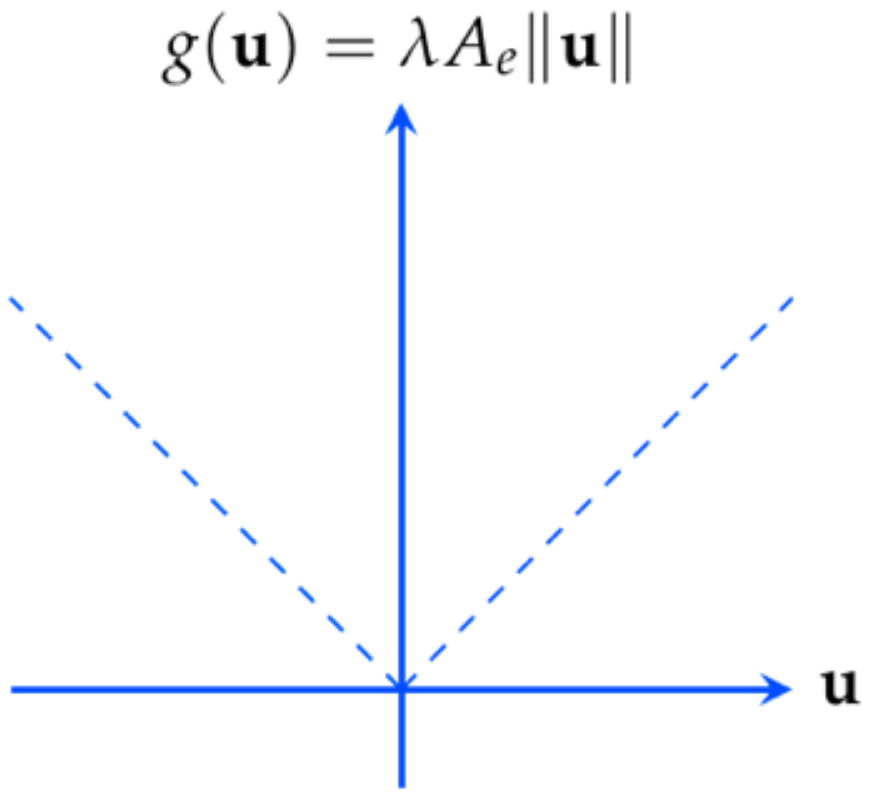
$$\max_{\mathbf{u}\in\mathcal{U}} -\sum_{i\in\mathcal{V}} L_i^*\left(\mathbf{w}^{(i)}\right) - \lambda \sum_{e\in\mathcal{E}} A_e \phi^*\left(\mathbf{u}^{(e)}/(\lambda A_e)\right)$$

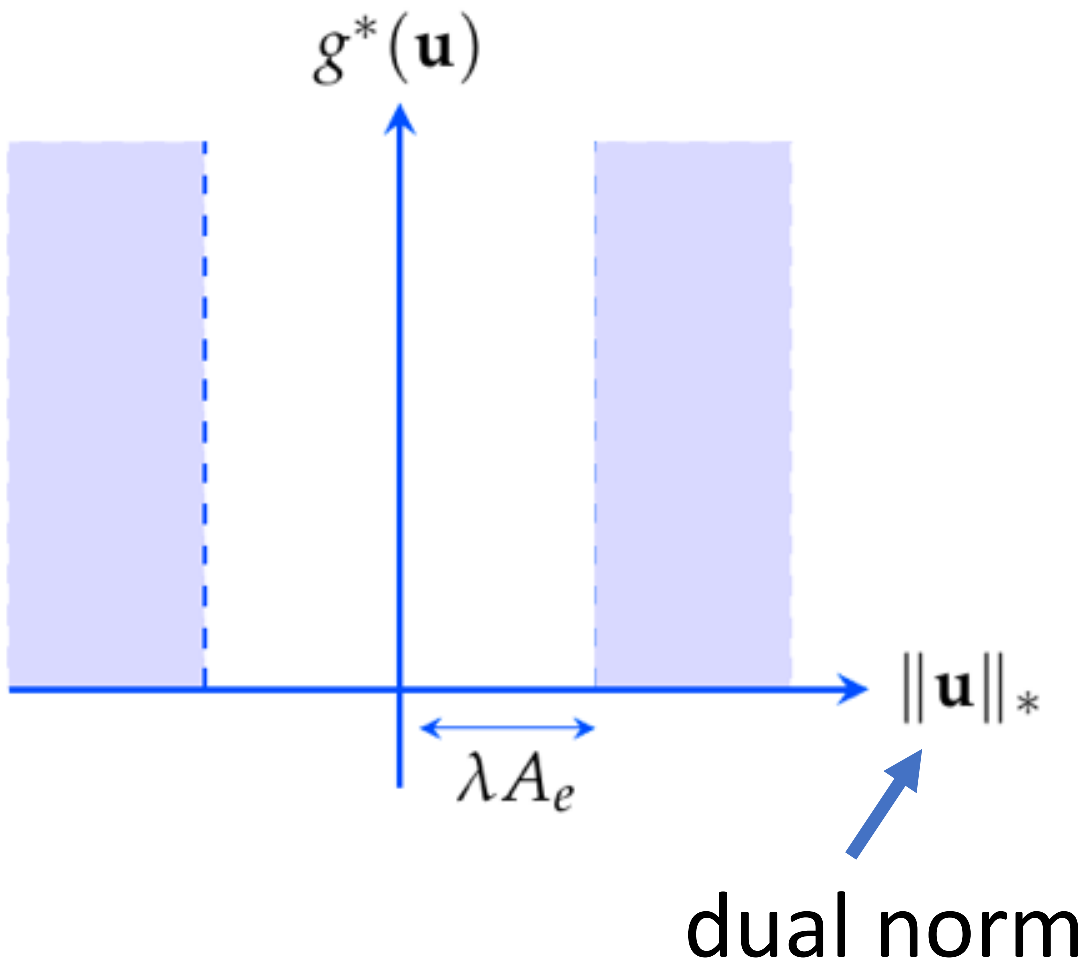
$$\text{subject to } -\mathbf{w}^{(i)} = \sum_{e\in\mathcal{E}} \sum_{i=e_{+}} \mathbf{u}^{(e)} - \sum_{i=e_{-}} \mathbf{u}^{(e)} \text{ for all nodes } i \in \mathcal{V}.$$



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

Convex Conjugate of Norm

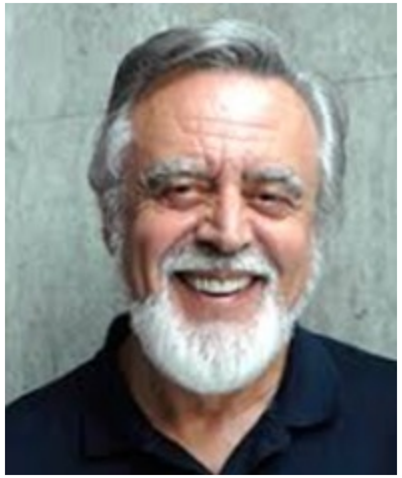


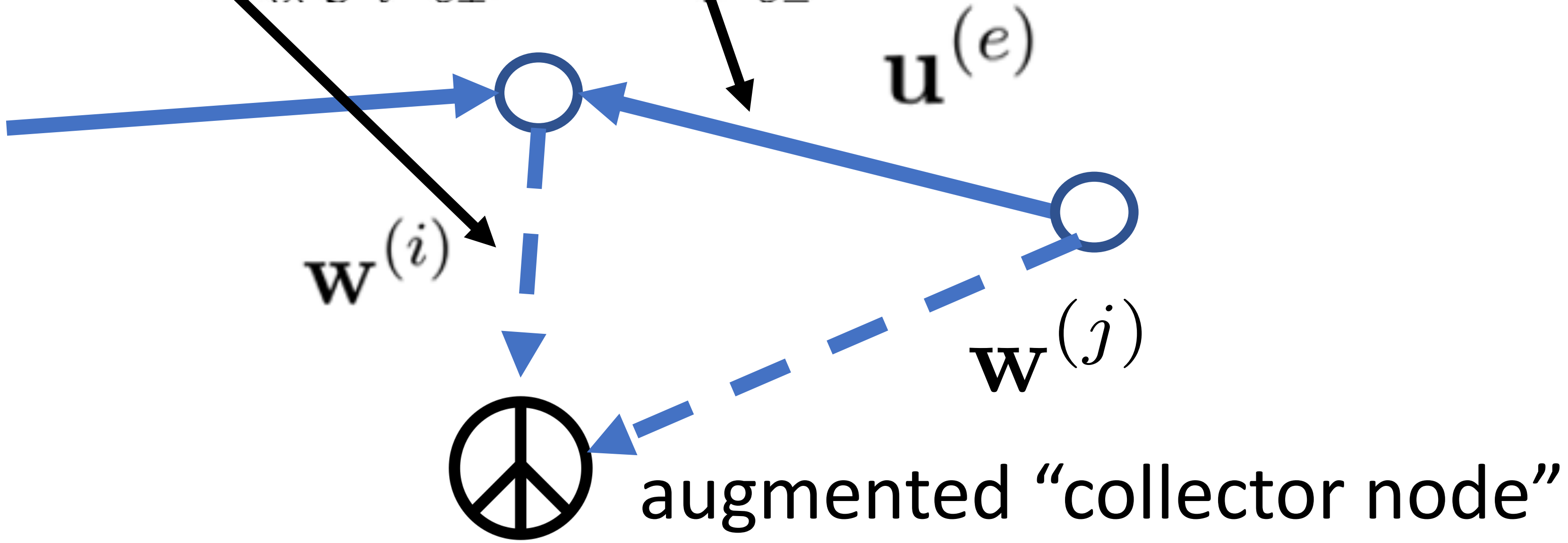


Non-Linear Min-Cost-Flow

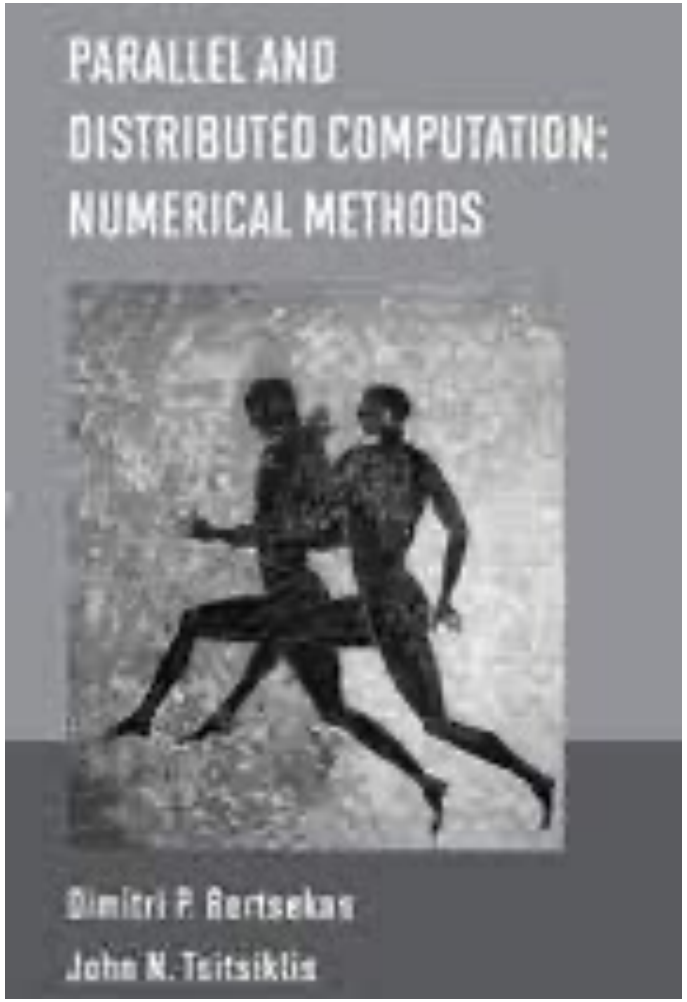
$$\max_{\mathbf{u} \in \mathcal{U}} - \sum_{i \in \mathcal{V}} L_i^{*}\left(\mathbf{w}^{(i)}\right) - \lambda \sum_{e \in \mathcal{E}} A_e \phi^{*}\left(\mathbf{u}^{(e)}/(\lambda A_e)\right)$$

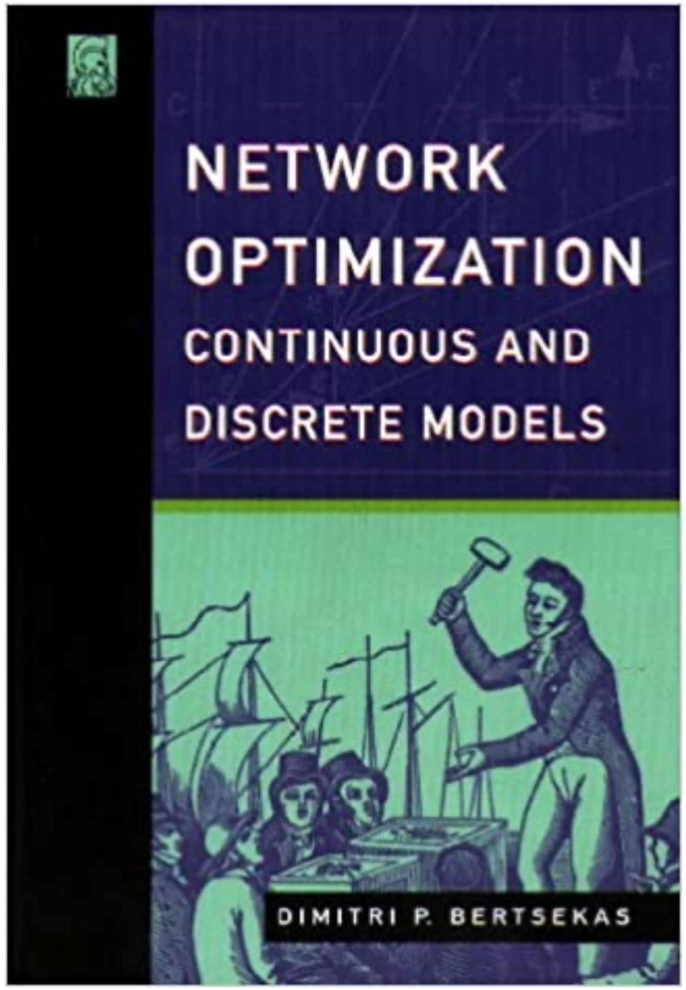
$$\text{subject to} \quad -\mathbf{w}^{(i)} = \sum_{e \in \mathcal{E}} \sum_{i=e_{+}} \mathbf{u}^{(e)} - \sum_{i=e_{-}} \mathbf{u}^{(e)} \text{ for all nodes } i \in \mathcal{V}.$$





Non-Linear Min-Cost-Flow

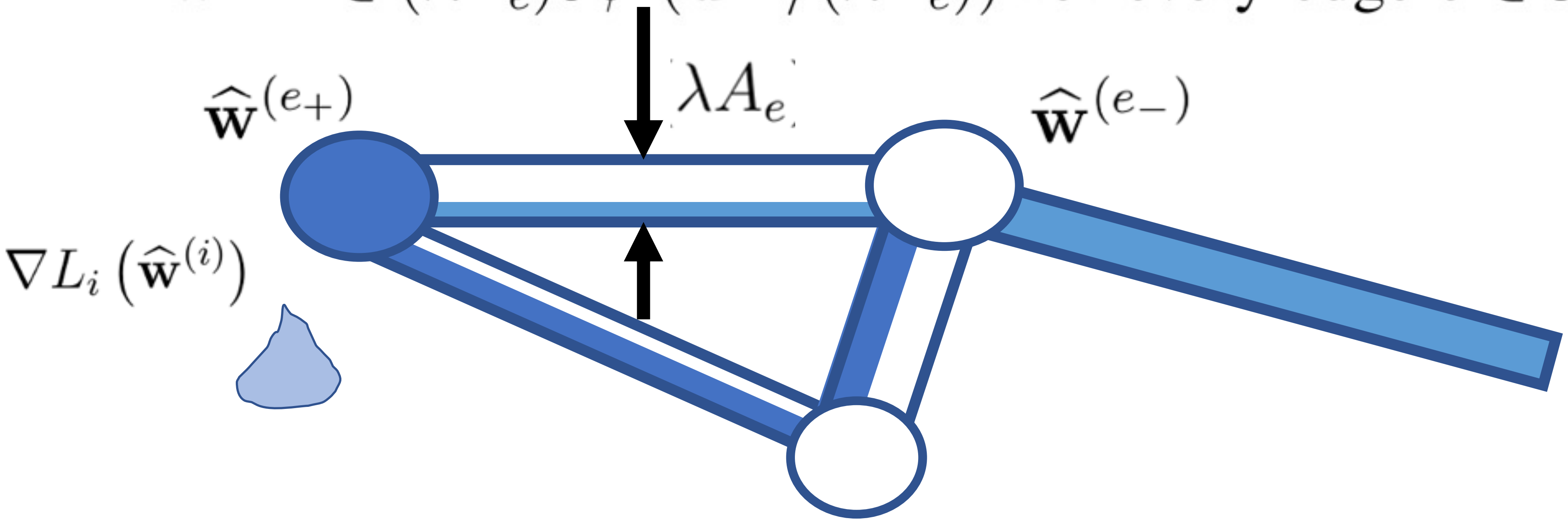




Primal and Dual Optimality.

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

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



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

Kirchhoff's Current Law

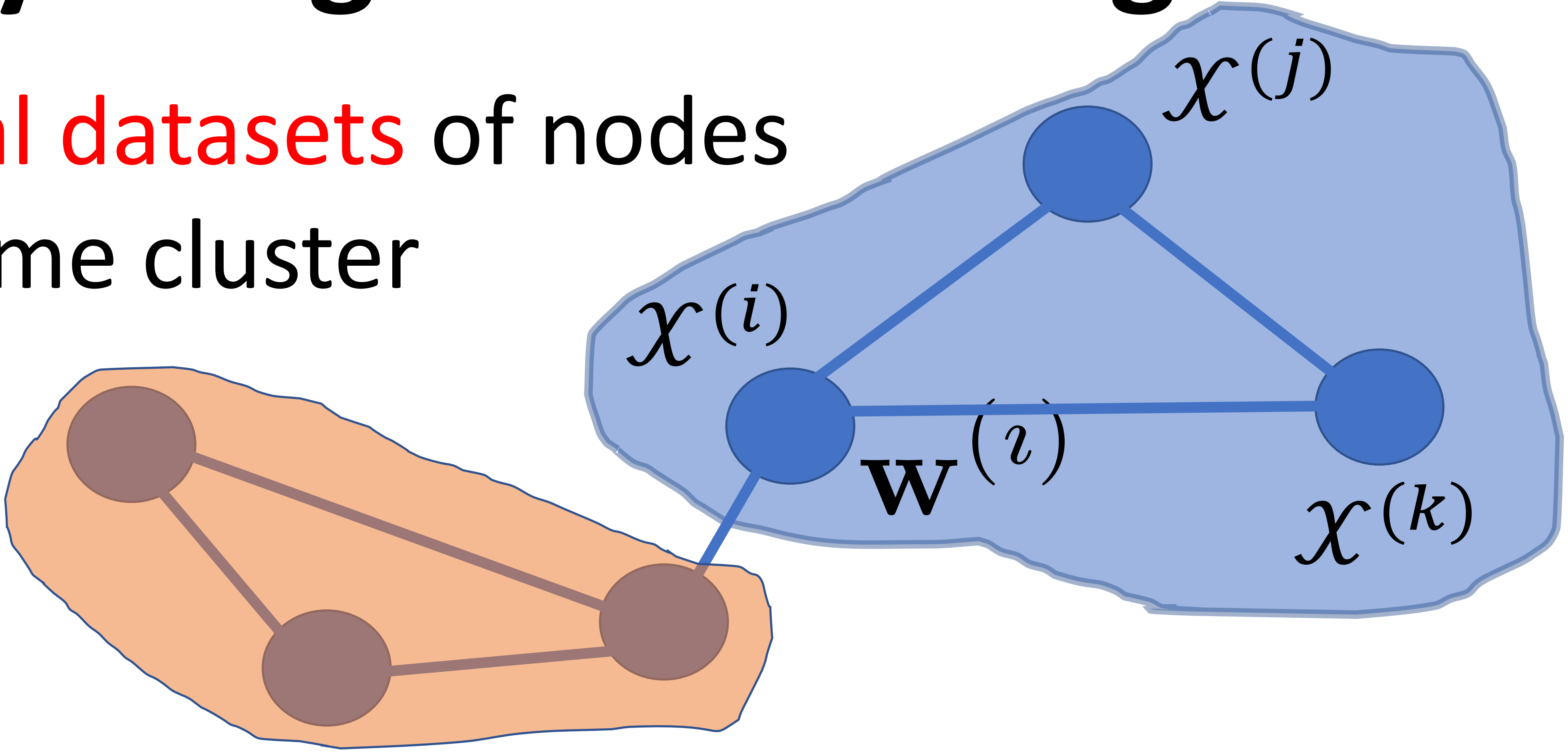
$$\sum_{e \in \mathcal{E}} \sum_{i=e_+} \widehat{\mathbf{u}}^{(e)} - \sum_{i=e_-} \widehat{\mathbf{u}}^{(e)} = -\nabla L_i \left(\widehat{\mathbf{w}}^{(i)}\right) \text{ for all nodes } i \in \mathcal{V}$$

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

Generalized Ohm Law

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.

Primal-Dual Optimality Conditions.

(assuming convexity of loss functions and GTV penalty)

$$\mathbf{M}^{-1}\begin{pmatrix}\partial f & \mathbf{D}^{T} \\ -\mathbf{D} & \partial g^{*}\end{pmatrix}\begin{pmatrix}\widehat{\mathbf{w}} \\ \widehat{\mathbf{u}}\end{pmatrix} \ni \mathbf{0} \text{ with } \mathbf{M} := \begin{pmatrix}\mathbf{T}^{-1} & -\mathbf{D}^{T} \\ -\mathbf{D} & \mathbf{\Sigma}^{-1}\end{pmatrix}$$

$$\Updownarrow$$

$$\begin{pmatrix}\widehat{\mathbf{w}} \\ \widehat{\mathbf{u}}\end{pmatrix} = \left(\mathbf{I} + \mathbf{M}^{-1}\begin{pmatrix}\partial f & \mathbf{D}^{T} \\ -\mathbf{D} & \partial g^{*}\end{pmatrix}\right)^{-1}\begin{pmatrix}\widehat{\mathbf{w}} \\ \widehat{\mathbf{u}}\end{pmatrix}$$

this is again a fixed-point problem !

Proximal Point Algorithm.

primal and dual variables $\widehat{\mathbf{w}}, \widehat{\mathbf{u}}$ optimal if and only if

$$\mathbf{M}^{-1}\begin{pmatrix} \partial f & \mathbf{D}^T \\ -\mathbf{D} & \partial g^* \end{pmatrix}\begin{pmatrix} \widehat{\mathbf{w}} \\ \widehat{\mathbf{u}} \end{pmatrix} \ni \mathbf{0} \text{ with } \mathbf{M} := \begin{pmatrix} \mathbf{T}^{-1} & -\mathbf{D}^T \\ -\mathbf{D} & \mathbf{\Sigma}^{-1} \end{pmatrix}$$

solve iteratively by proximal point algorithm

$$\begin{pmatrix} \widehat{\mathbf{w}}^{(k+1)} \\ \widehat{\mathbf{u}}^{(k+1)} \end{pmatrix} = \left(\mathbf{I} + \mathbf{M}^{-1}\begin{pmatrix} \partial f & \mathbf{D}^T \\ -\mathbf{D} & \partial g^* \end{pmatrix}\right)^{-1}\begin{pmatrix} \widehat{\mathbf{w}}^{(k)} \\ \widehat{\mathbf{u}}^{(k)} \end{pmatrix}$$

A. Chambolle, T. Pock. An introduction to continuous optimization for imaging. Acta Numerica, 2016.

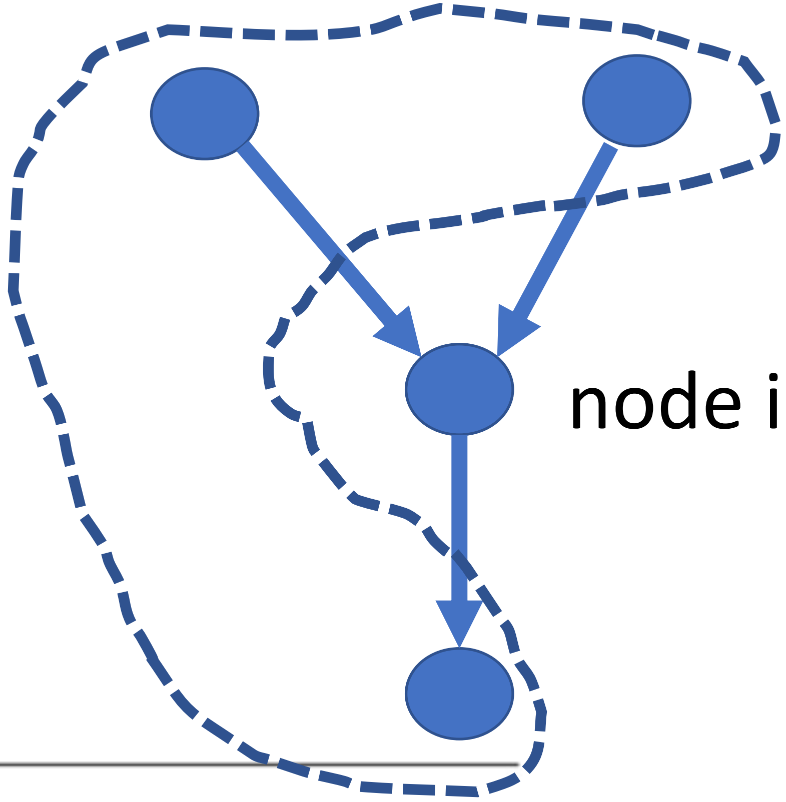
After Some Manipulations.

Algorithm 1 Primal-Dual Method for Networked FL

Input: empirical graph $\mathcal{G}=(\mathcal{V},\mathcal{E},\mathbf{A})$; training set $\{\mathbf{X}^{(i)}\}_{i\in\mathcal{M}}$; regularization parameter λ; loss $\mathcal{L}$; GTV penalty ϕ

Initialize: $k:=0;\widehat{\mathbf{w}}_0:=\mathbf{0};\widehat{\mathbf{u}}_0:=\mathbf{0}$; $\sigma_e=1/2$ and $\tau_i=1/|\mathcal{N}_i|$

while stopping criterion is not satisfied **do**
    for all nodes $i\in\mathcal{V}$ **do**
        $\widehat{\mathbf{w}}^{(i)}_{k+1}:=\widehat{\mathbf{w}}^{(i)}_{k}-\tau_i\sum_{e\in\mathcal{E}}D_{e,i}\widehat{\mathbf{u}}^{(e)}_{k}$
    end for
    for nodes in the training set $i\in\mathcal{M}$ **do**
        $\widehat{\mathbf{w}}^{(i)}_{k+1}:=\mathcal{PU}^{(i)}\{\widehat{\mathbf{w}}^{(i)}_{k+1}\}$
    end for
    for all edges $e\in\mathcal{E}$ **do**
        $\widehat{\mathbf{u}}^{(e)}_{k+1}:=\widehat{\mathbf{u}}^{(e)}_{k}+\sigma_e\big(2\big(\widehat{\mathbf{w}}^{(e_+)}_{k+1}-\widehat{\mathbf{w}}^{(e_-)}_{k+1}\big)-\big(\widehat{\mathbf{w}}^{(e_+)}_{k}-\widehat{\mathbf{w}}^{(e_-)}_{k}\big)\big)$
        $\widehat{\mathbf{u}}^{(e)}_{k+1}:=\mathcal{DU}^{(e)}\{\widehat{\mathbf{u}}^{(e)}_{k+1}\}$
    end for
    $k:=k+1$
end while

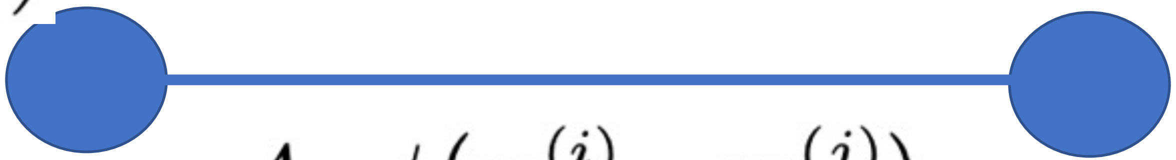


Algorithm 1 is Attractive for NFL…

- decentralized implementation (mess. pass.)
- robust against various imperfections
  - approximate primal/dual updates
  - node/link failures
- privacy friendly; no raw data exchanged

Local Computations in Algorithm 1.

$$L^{(i)}\left(\mathcal{X}^{(i)}, \mathbf{w}^{(i)}\right)$$

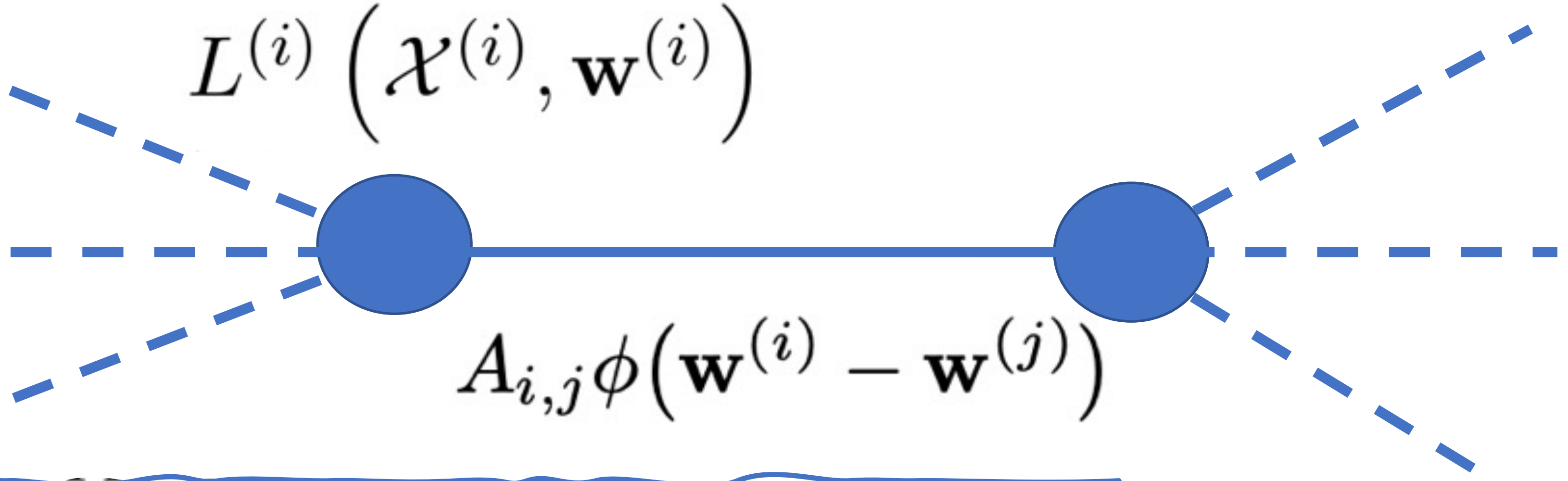


$$A_{i,j}\phi\left(\mathbf{w}^{(i)} - \mathbf{w}^{(j)}\right)$$

node-wise primal update: $$\mathcal{PU}^{(i)}\{\mathbf{v}\} := \operatorname*{argmin}_{\mathbf{z}\in\mathbb{R}^n} L^{(i)}\left(\mathbf{z}\right)+(1/2\tau_i)\|\mathbf{v}-\mathbf{z}\|^2.$$

edge-wise dual update: $$\mathcal{DU}^{(e)}\{\mathbf{v}\} := \operatorname*{argmin}_{\mathbf{z}\in\mathbb{R}^n} \lambda A_e\phi^*\left(\mathbf{z}/(\lambda A_e)\right)+(1/2\sigma_e)\|\mathbf{v}-\mathbf{z}\|^2.$$

Spreading Local Results.



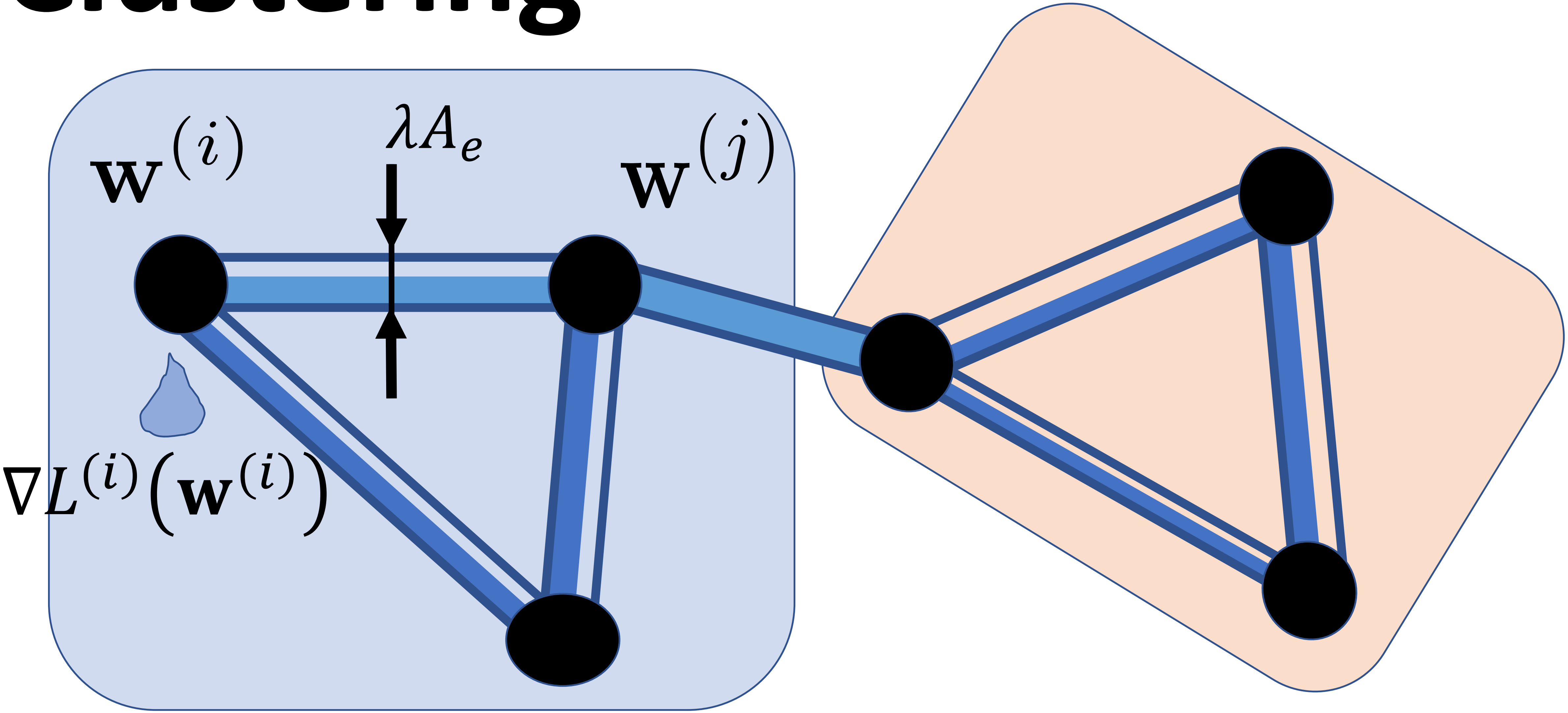
for all nodes $i \in \mathcal{V}$ **do**
$\quad \widehat{\mathbf{w}}_{k+1}^{(i)} := \widehat{\mathbf{w}}_{k}^{(i)} - \tau_i \sum_{e \in \mathcal{E}} D_{e,i} \widehat{\mathbf{u}}_{k}^{(e)}$
end for

for all edges $e \in \mathcal{E}$ **do**
$\quad \widehat{\mathbf{u}}_{k+1}^{(e)} := \widehat{\mathbf{u}}_{k}^{(e)} + \sigma_e\Big(2\big(\widehat{\mathbf{w}}_{k+1}^{(e_+)} - \widehat{\mathbf{w}}_{k+1}^{(e_-)}\big) - \big(\widehat{\mathbf{w}}_{k}^{(e_+)} - \widehat{\mathbf{w}}_{k}^{(e_-)}\big)\Big)$

Are GTVMin Solutions Any Good?

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

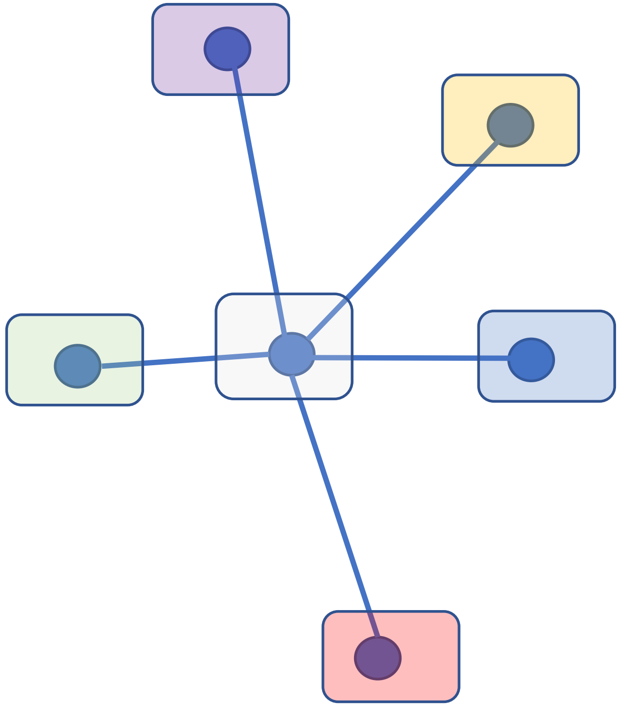
Clustering



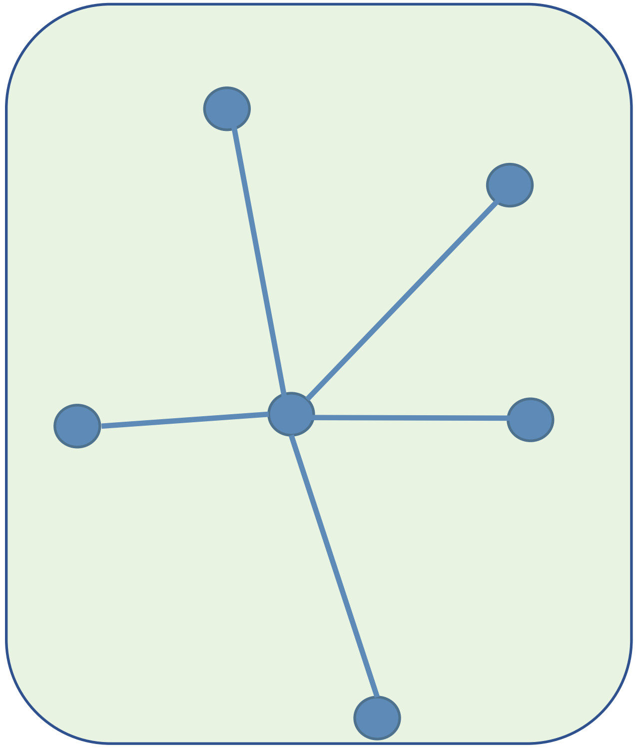
parameter vectors can only change over saturated links !

Person. vs. Globalization

small λ, edges easily saturated



large λ, edges hard to saturate



References

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

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

Y. SarcheshmehPour, Y. Tian, L. Zhang, AJ, “Networked Federated Learning”, <i>arXiv e-prints</i>, 2021.