PURDUE UNIVERSITY | Elmore Family School of Electrical and Computer Engineering

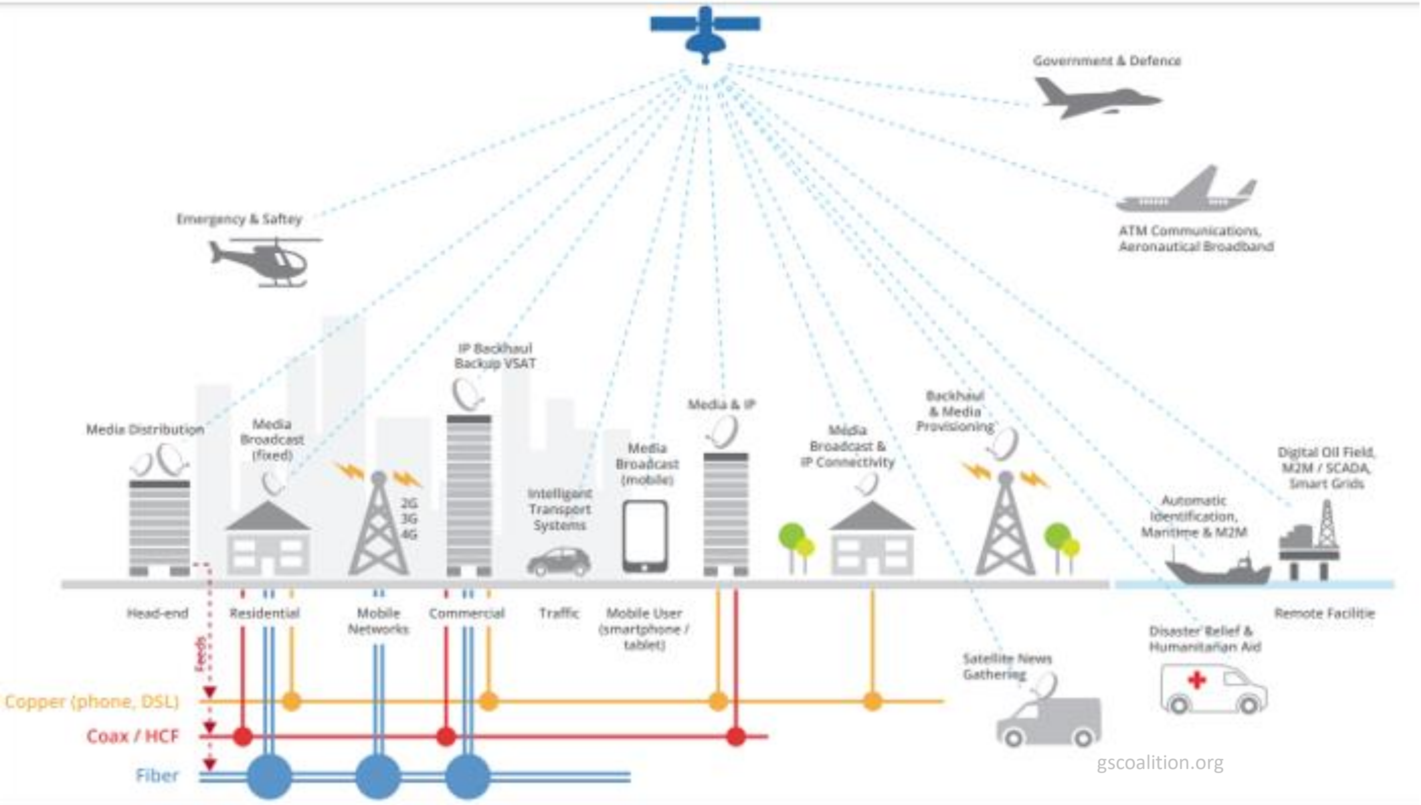
Structured and Resource-Constrained Collaborative Learning

Abolfazl Hashemi

ML Seminar, Purdue University

September 22, 2021

The Era of Collaborative Systems

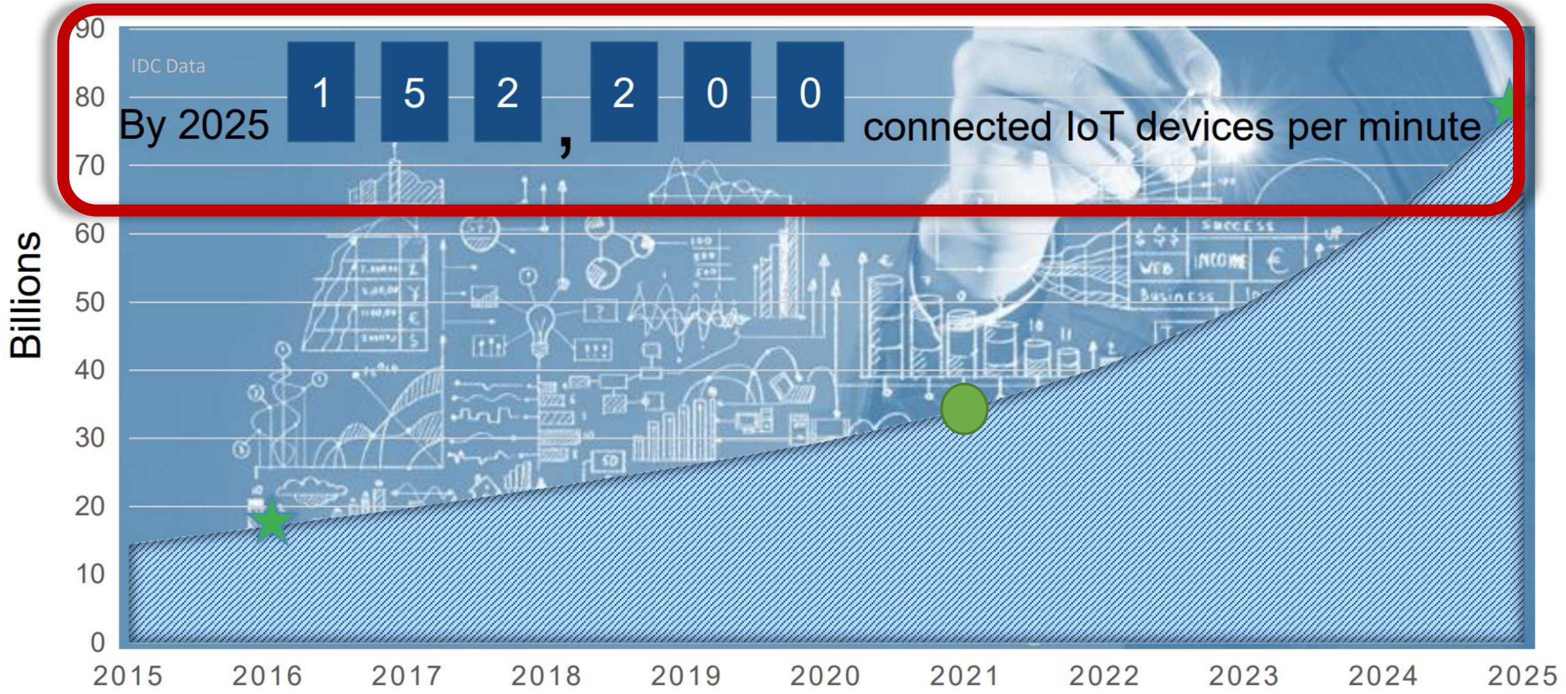


Satellite mesh network

Smart grids

Creating reliable and effective collaborative systems that are highly secure, robust and economically viable

Collaborative Systems: Large-Scale and Heterogenous

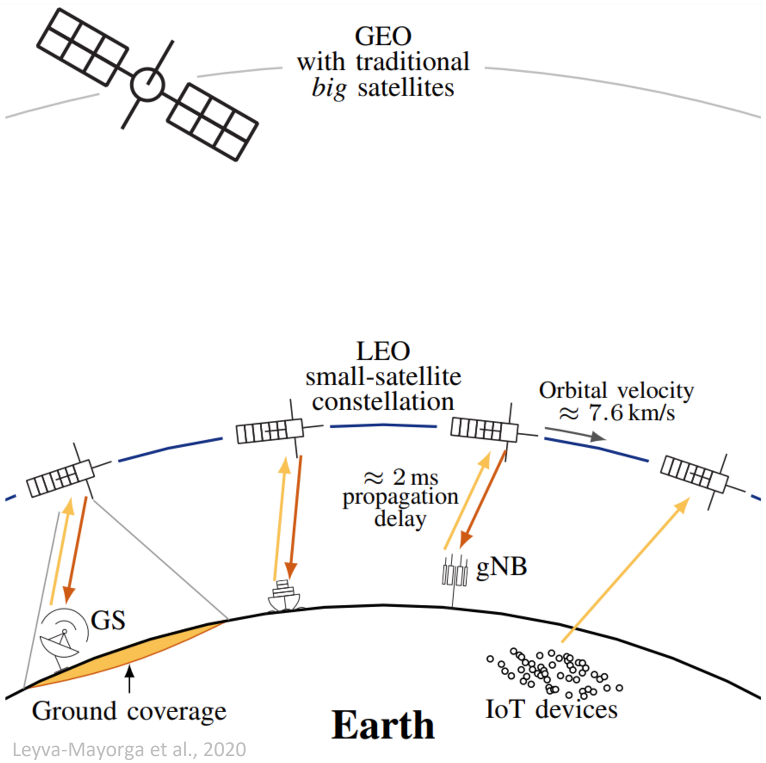


Rapidly increasing network size

High degree of heterogeneity

How can we enable scalable deployment of collaborative learning in presence of heterogeneity?

Collaborative Systems: Embodied Agents



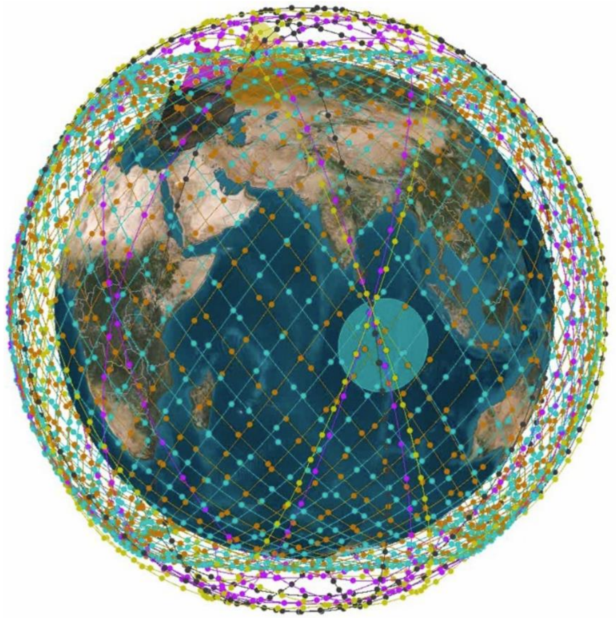
Leyva-Mayorga et al., 2020

Rapidly evolving environments

Limited energy budgets

Costly observation gathering

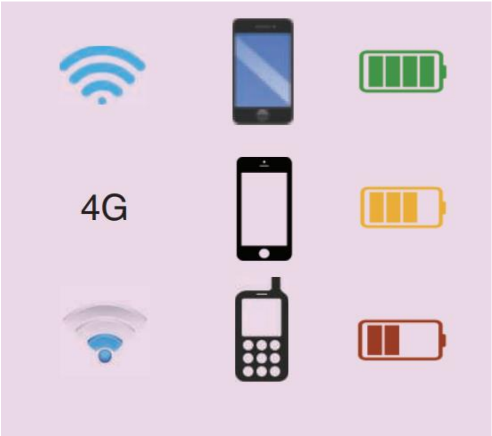
4000+ LEO satellites in SpaceX network

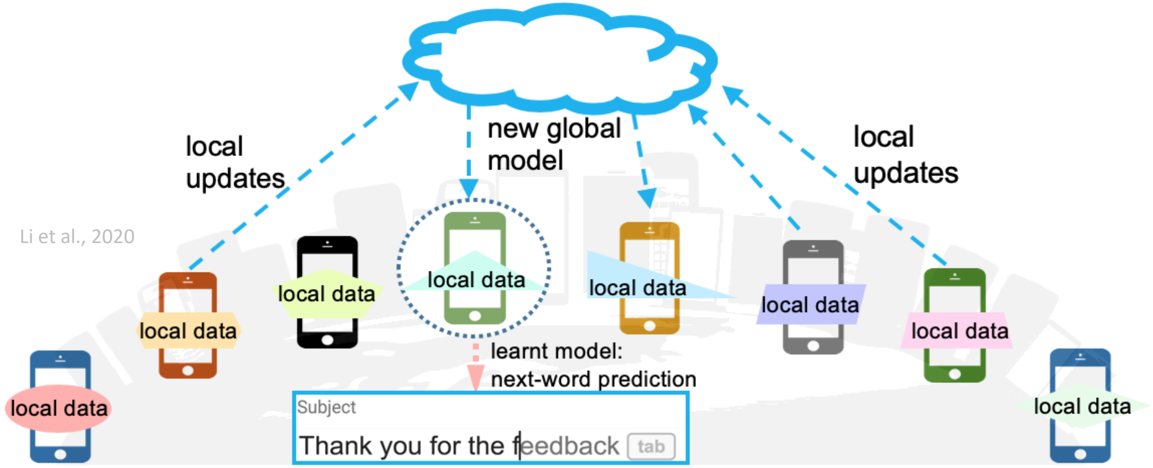


MIT Tech review

How can we design low-cost and energy-efficient collaborative learning systems capable of operating in rapidly evolving environments?

Collaborative Systems: Limited Communication Budget





Unreliable communication

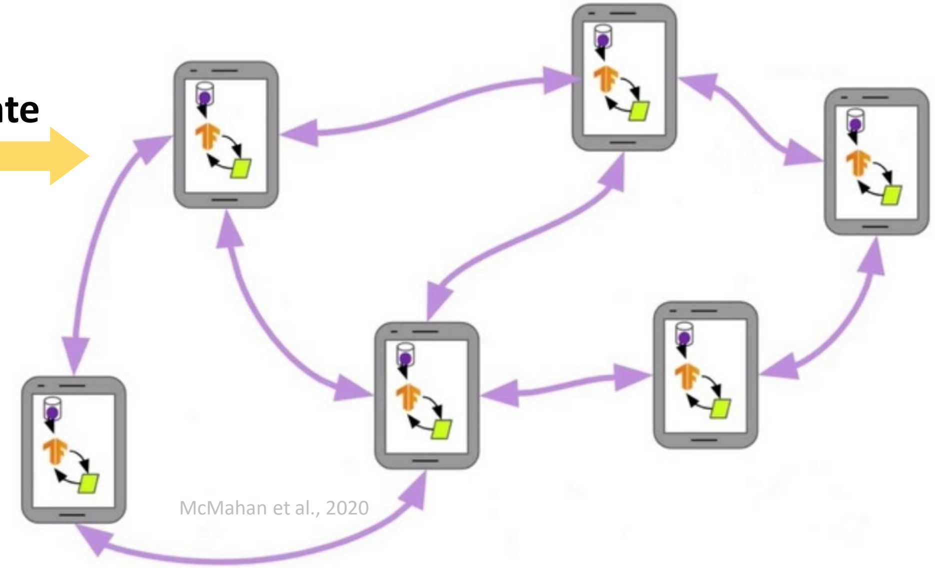
Limited bandwidth

How can we design robust and communication-efficient collaborative learning systems?

Communication-Efficient Federated and Distributed Learning

Local data

Local update



Compressed message-passing

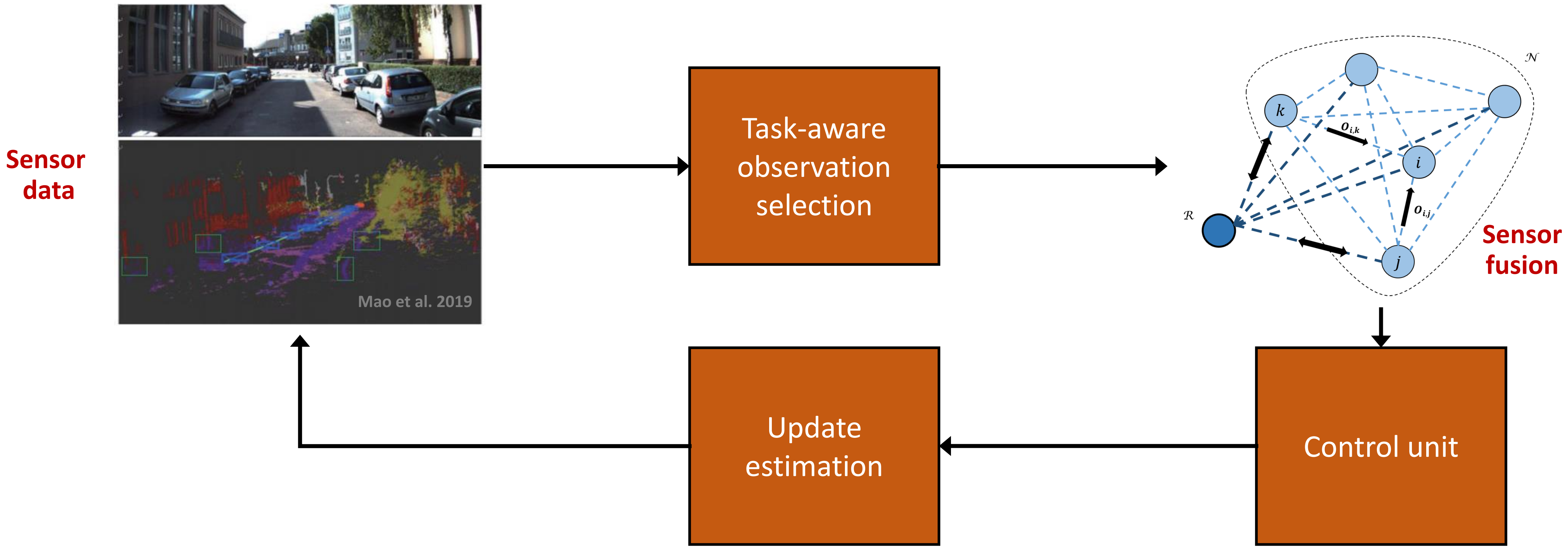
Heterogenous model aggression

Learning

Contributions:

- Model aggregation and communication strategies for distributed learning
- Developing communication-efficient algorithms with provable guarantees

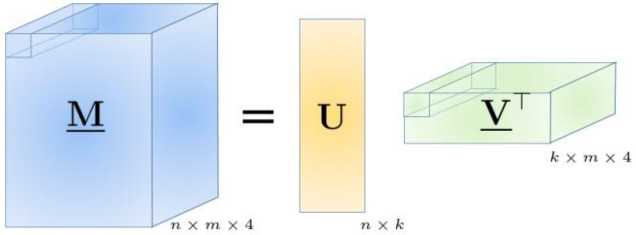
Efficient Observation Selection and Information Gathering



Contributions:

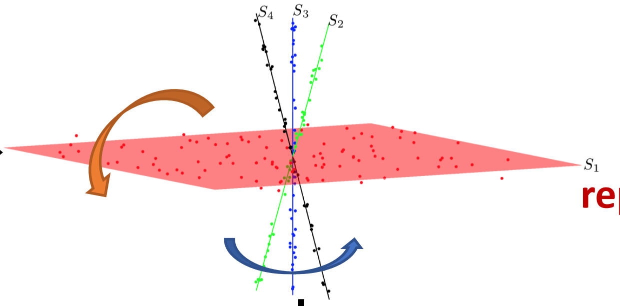
- Task-aware observation selection criteria for sensing networks
- Developing efficient feature selection algorithms with near-optimal utilities

Data-Scarce Parsimonious Representation Learning



Side information

Representation learning



Learned representation

Raw data

Inference

Contributions:

- Representation learning for unsupervised inference from structured data
- Sparse approximation algorithms for function approximation and regression

Communication-Efficient Federated and Distributed Learning

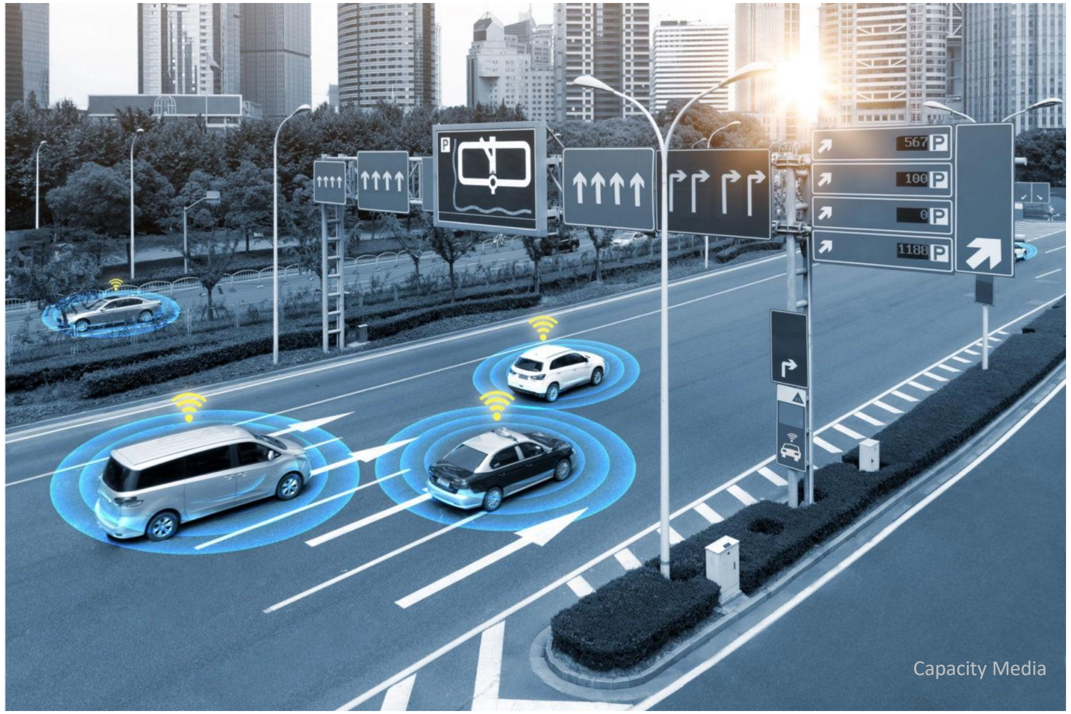
[Das, R., Hashemi, A., Acharya, A., Sanghavi, S., Dhillon, I., Topcu, U., “Faster Non-Convex Federated Learning via Global and Local Momentum,” Submitted, 2021.]

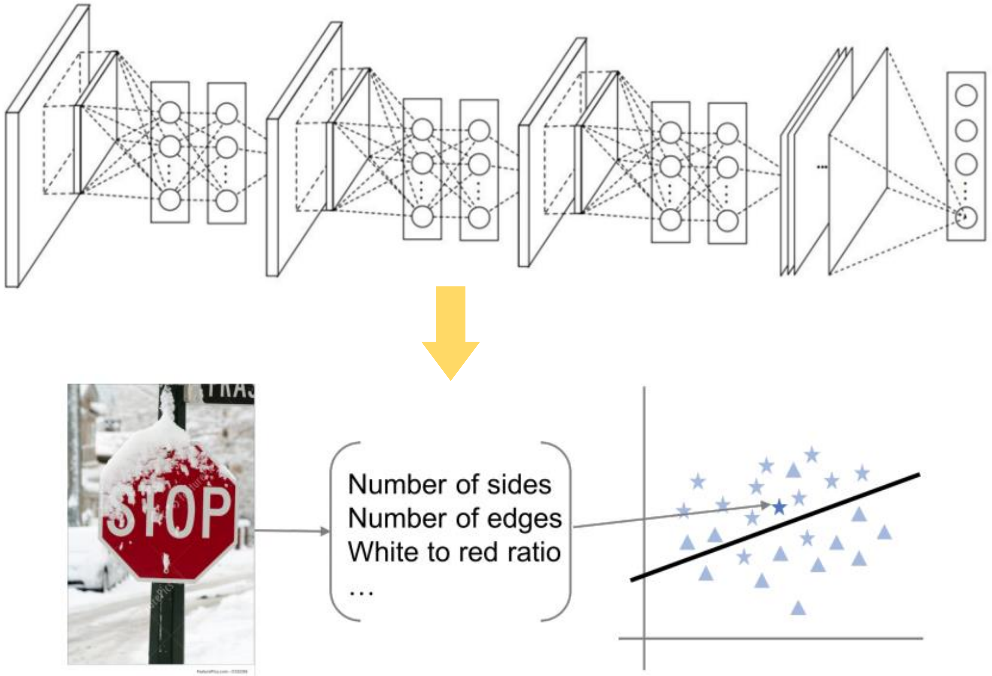
[Chen, Y., Hashemi, A., Vikalo, H., “Communication-Efficient Variance-Reduced Decentralized Stochastic Optimization over Time-Varying Directed Graphs,” Submitted, 2021.]

[Hashemi, A., Acharya, A., Das, R., Vikalo, H., Sanghavi, S., Dhillon, I., “On the Benefits of Multiple Gossip Steps in Communication-Constrained Decentralized Optimization,” Submitted, 2021.]

[Chen, Y., Hashemi, A., Vikalo, H., “Decentralized Optimization on Time-Varying Directed Graphs under Communication Constraints,” International Conference on Acoustic, Speech and Signal Processing (ICASSP), 2021.]

Collaborative Learning in Connected Systems





Limited local data

Collaboration via cloud

Expensive Communication

Systems Heterogeneity

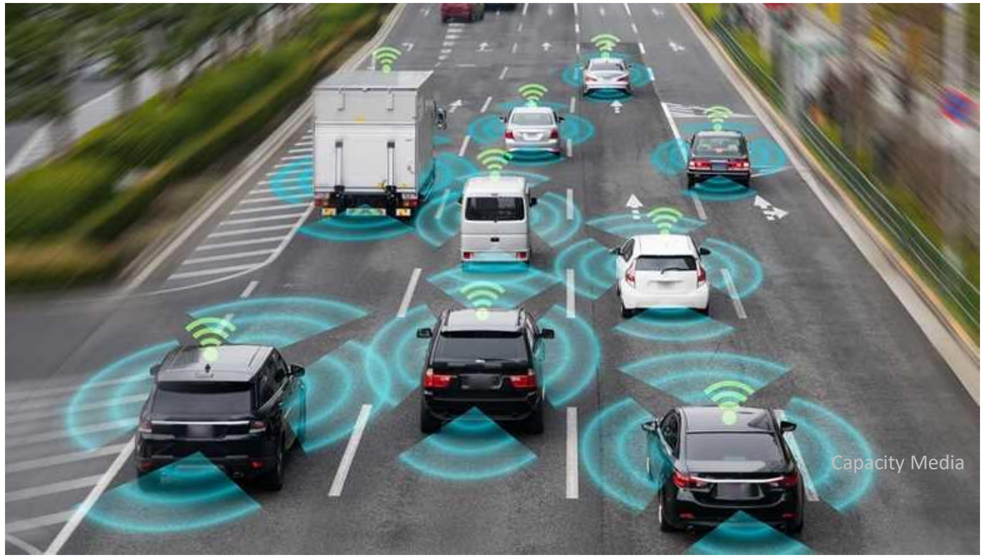
Statistical Heterogeneity

Li et al., 2019

Communication-Efficient Federated Learning

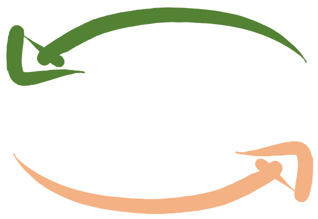
Local optimization

$$\boldsymbol{w}^{(1)}, \boldsymbol{w}^{(2)}, \ldots, \boldsymbol{w}^{(100)}$$



Intermittent high-speed downlink to available vehicles

$$\boldsymbol{w}$$



Intermittent slow uplink from available vehicles

$$\boldsymbol{w}^{(1)}, \boldsymbol{w}^{(2)}, \ldots, \boldsymbol{w}^{(10)}$$

Global aggregation

$$\boldsymbol{w} = \mathrm{agg}(\boldsymbol{w}^{(1)}, \boldsymbol{w}^{(2)}, \ldots, \boldsymbol{w}^{(10)})$$

How to find an effective local update?

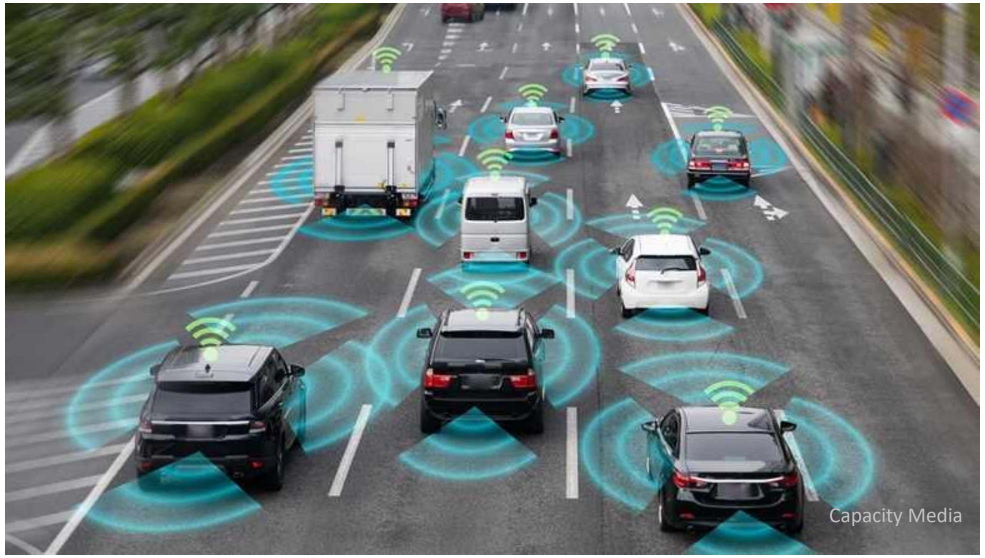
How to send compressed, informative messages?

How to aggregate in the presence of heterogeneity?

Components of the Problem

Task and objective function

Local optimization subroutine



Network structure (availability and heterogeneity)

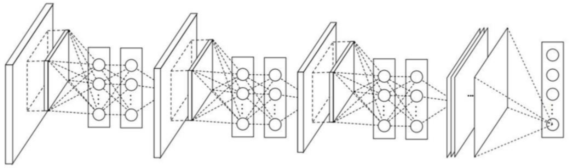
Compressed message-passing

Aggregation scheme

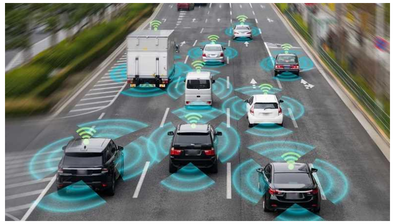
Task and Objective Function

Empirical Risk Minimization (ERM)

Model parameters



$$\min_{\boldsymbol{w}} \quad f(\boldsymbol{w}) = \frac{1}{n}\sum_{i=1}^{n} f_i(\boldsymbol{w}) \quad \text{where} \quad f_i(\boldsymbol{w}) = \frac{1}{n_i}\sum_{j=1}^{n_i} \widehat{f}_{i_j}(\boldsymbol{w})$$



Number of devices

Number data points at device i

Smooth loss functions

$$\|\nabla \widehat{f}_{i_j}(\boldsymbol{x}) - \nabla \widehat{f}_{i_j}(\boldsymbol{y})\| \leq L\|\boldsymbol{x} - \boldsymbol{y}\|$$

Network Structure and Heterogeneity

Periodic message-passing: devices communicate with the server intermittently

K Communication rounds

Training round $t = \underbrace{1, 2, \ldots,}\ E, \underbrace{E+1, E+2, \ldots,}\ 2E, \underbrace{\ldots,}\ KE$

Only local computation

Partial participation: only *r* out of *n* devices available each communication round $(r \ll n)$

$$\|\widetilde{\nabla} f_i(\boldsymbol{w}; \mathcal{B}) - \nabla f_i(\boldsymbol{w})\| \leq \sigma_b$$

Stochastic gradient approximation error

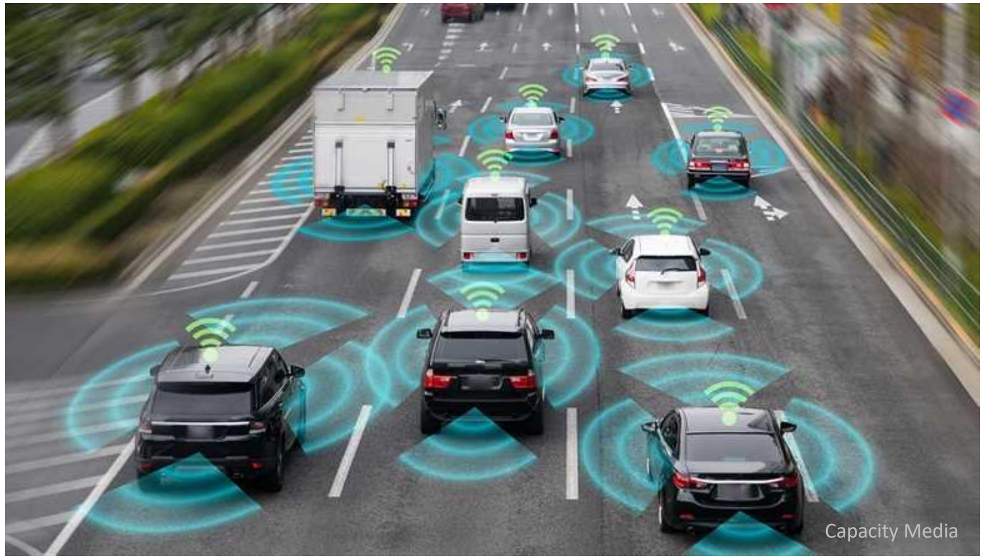
Bounded dissimilarity

$$\|\underbrace{\nabla f_i(\boldsymbol{w})}_{\text{Local functions}} - \underbrace{\nabla f(\boldsymbol{w})}_{\text{Global function}}\|^2 \leq \sigma_r^2$$

Components of the Problem

Task and objective function

Local optimization subroutine



Compressed message-passing

Aggregation scheme

Network structure (availability and heterogeneity)

Stochastic Gradient Descent (SGD)

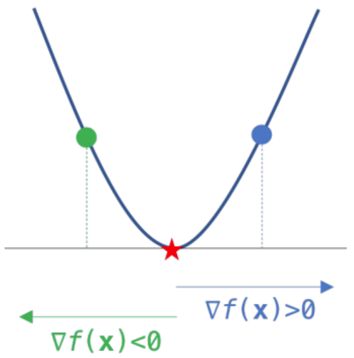
Search for a point where the gradient is small

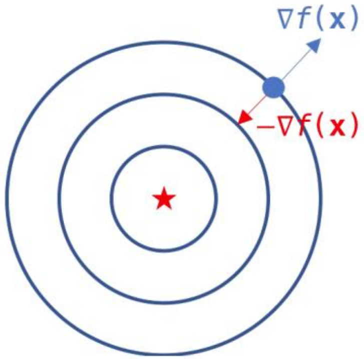
$$\|\nabla f(\boldsymbol{w})\|^2 \leq \epsilon$$

Stochastic first-order update

$$\boldsymbol{w}_{t+1} = \boldsymbol{w}_t - \eta \widetilde{\nabla} f_i(\boldsymbol{w}_t; \mathcal{B}_t)$$

Intuition: Take a step in the direction opposite to the gradient





What do we know about the performance of SGD?

Theorem (Convergence of SGD)

$$T = \mathcal{O}\left(\frac{\sigma_b^2}{\epsilon^2} + \frac{1}{\epsilon}\right)$$

Bottou 2018

Theorem (Lower bounds)

$$T = \mathcal{O}\left(\frac{1}{\epsilon^{1.5}}\right)$$

Arjevani et al. 2019

Local Momentum-Based Variance Reduction

A recursive stochastic first-order update

$$\boldsymbol{v}_\tau^{(i)} = \widetilde{\nabla} f_i(\boldsymbol{w}_\tau^{(i)}; \mathcal{B}_\tau^{(i)}) + (\boldsymbol{v}_{\tau-1}^{(i)} - \widetilde{\nabla} f_i(\boldsymbol{w}_{\tau-1}^{(i)}; \mathcal{B}_\tau^{(i)}))$$

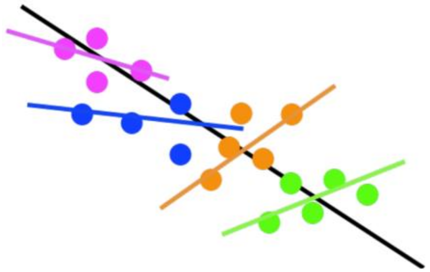
$$\boldsymbol{w}_{\tau+1}^{(i)} = \boldsymbol{w}_t^{(i)} - \eta \boldsymbol{v}_\tau^{(i)}$$

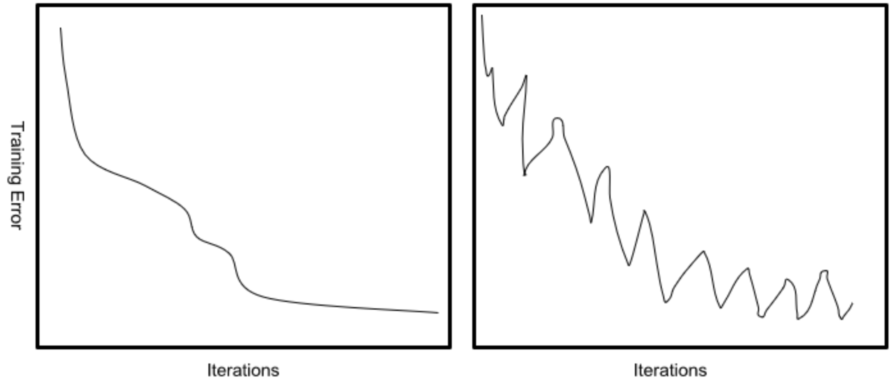
Lemma: Reduces the variance of stochastic gradient

$$\sum_{\tau=0}^{E-1} \mathbb{E}[\|\boldsymbol{v}_\tau^{(i)} - \nabla f_i(\boldsymbol{w}_\tau^{(i)})\|^2] \leq O(E^2\eta^2)$$

Hashemi et al., 2021

$$\|\underbrace{\nabla f_i(\boldsymbol{w})}_{\text{Local functions}} - \underbrace{\nabla f(\boldsymbol{w})}_{\text{Global function}}\|^2 \leq \sigma_r^2$$





Proposed SGD

How about heterogeneity?

Global Momentum-Based Variance Reduction

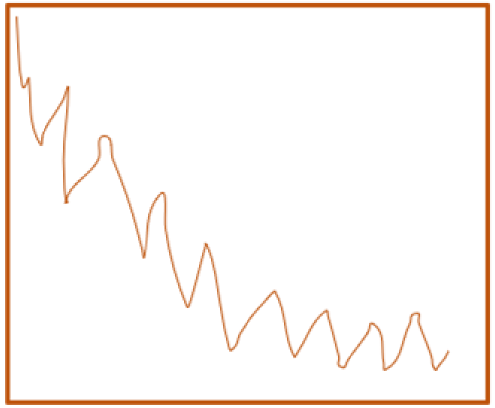
Simple aggregation: $\boldsymbol{w} \leftarrow \frac{1}{r}\sum_{i\in\mathcal{S}} \boldsymbol{w}_E^{(i)}$

$$\boldsymbol{w} \leftarrow \boldsymbol{w} - \frac{\eta}{r}\sum_{i\in\mathcal{S}_k} \frac{\boldsymbol{w}-\boldsymbol{w}_E^{(i)}}{\eta}$$

A generalized stochastic gradient

Using momentum-based variance reduction for model parameters

Similar issue, but now because of heterogeneity



Theorem (optimal rate)

To get $\mathbb{E}\|\nabla f(\boldsymbol{w}_K)\|^2 \leq \epsilon$ we need

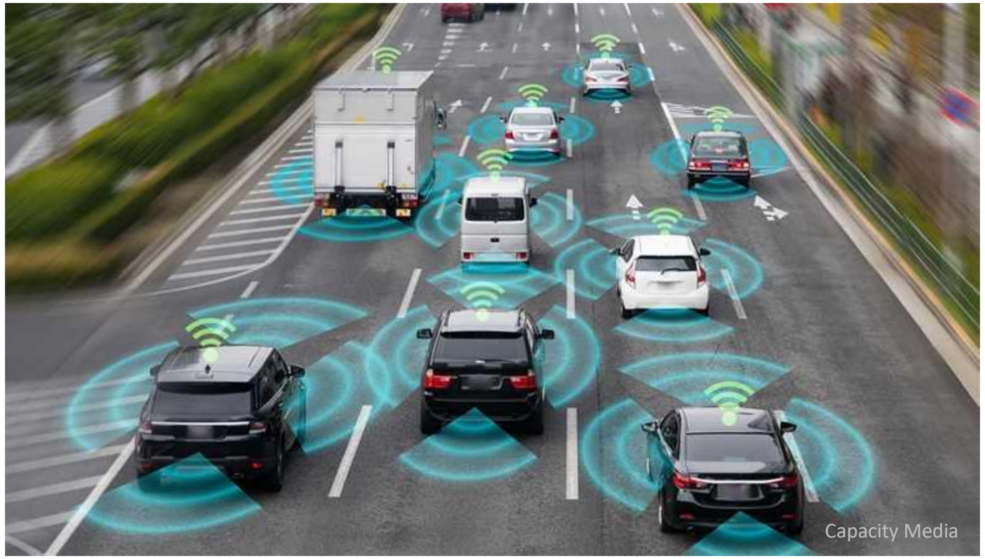
$$K = \mathcal{O}\left(\frac{1}{\epsilon^{1.5}}\right)$$

Hashemi et al., 2021

Components of the Problem

Task and objective function

Local optimization subroutine



Compressed message-passing

Aggregation scheme

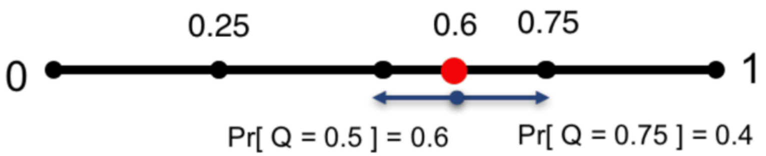
Network structure (availability and heterogeneity)

Quantized Uplink Communication

s-level Stochastic quantization

[Alistarh et al., 2017]

$$Q_D(v_i) = \|\boldsymbol{v}\|.\mathrm{sgn}(v_i).\xi_i(\boldsymbol{v}, s)$$



Unbiased with small variance

Uplink messages

Each learner sends

$$Q_D(\underbrace{\boldsymbol{w}_k - \boldsymbol{w}^{(i)}_{k,E}}_{\text{previous global model – current local model}})$$

Local and Global Momentum-Based Variance Reduction

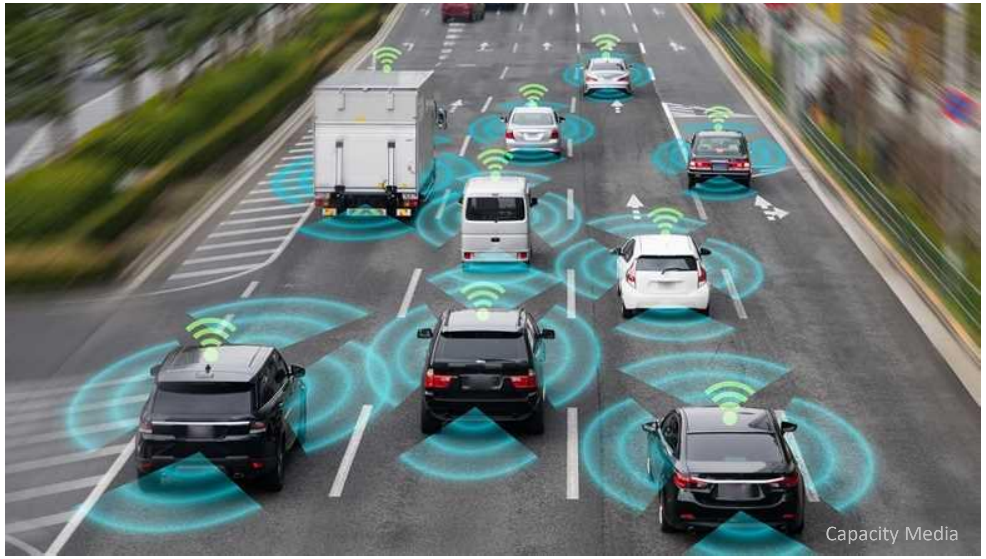
Local momentum-based variance reduction

Global momentum-based variance reduction

$$\boldsymbol{v}_{\tau}^{(i)} = \widetilde{\nabla} f_i(\boldsymbol{w}_{\tau}^{(i)}; \mathcal{B}_{\tau}^{(i)}) + \left(\boldsymbol{v}_{\tau-1}^{(i)} - \widetilde{\nabla} f_i(\boldsymbol{w}_{\tau-1}^{(i)}; \mathcal{B}_{\tau}^{(i)})\right)$$

$$\boldsymbol{w}_{\tau+1}^{(i)} = \boldsymbol{w}_{t}^{(i)} - \eta \boldsymbol{v}_{\tau}^{(i)}$$

$$\boldsymbol{w}_{k+1} = \mathrm{agg}(\{Q_D(\boldsymbol{w}_k - \boldsymbol{w}_{k,E}^{(i)})\})$$



$$\boldsymbol{w}_{k+1}$$

Intermittent slow uplink from available vehicles

$$Q_D(\boldsymbol{w}_k - \boldsymbol{w}_{k,E}^{(i)})$$

Collaborative Learning of Multiclass Classifiers

10 classes, 50,000 images

n = 50 collaborative learners

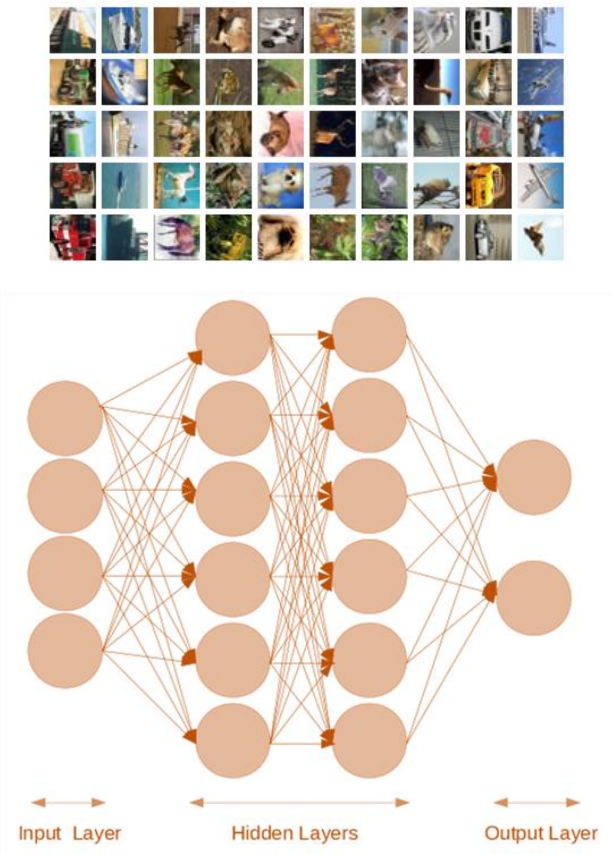
Communication protocol:
***50%* dropout rate** *(r=25)*

Communication **every 10 rounds** (Intermittency)

Heterogenous case:
2% of data available **locally**, from at most **two classes**

Homogenous case:
2% of data available **locally**, i.i.d. among the devices

CIFAR-10



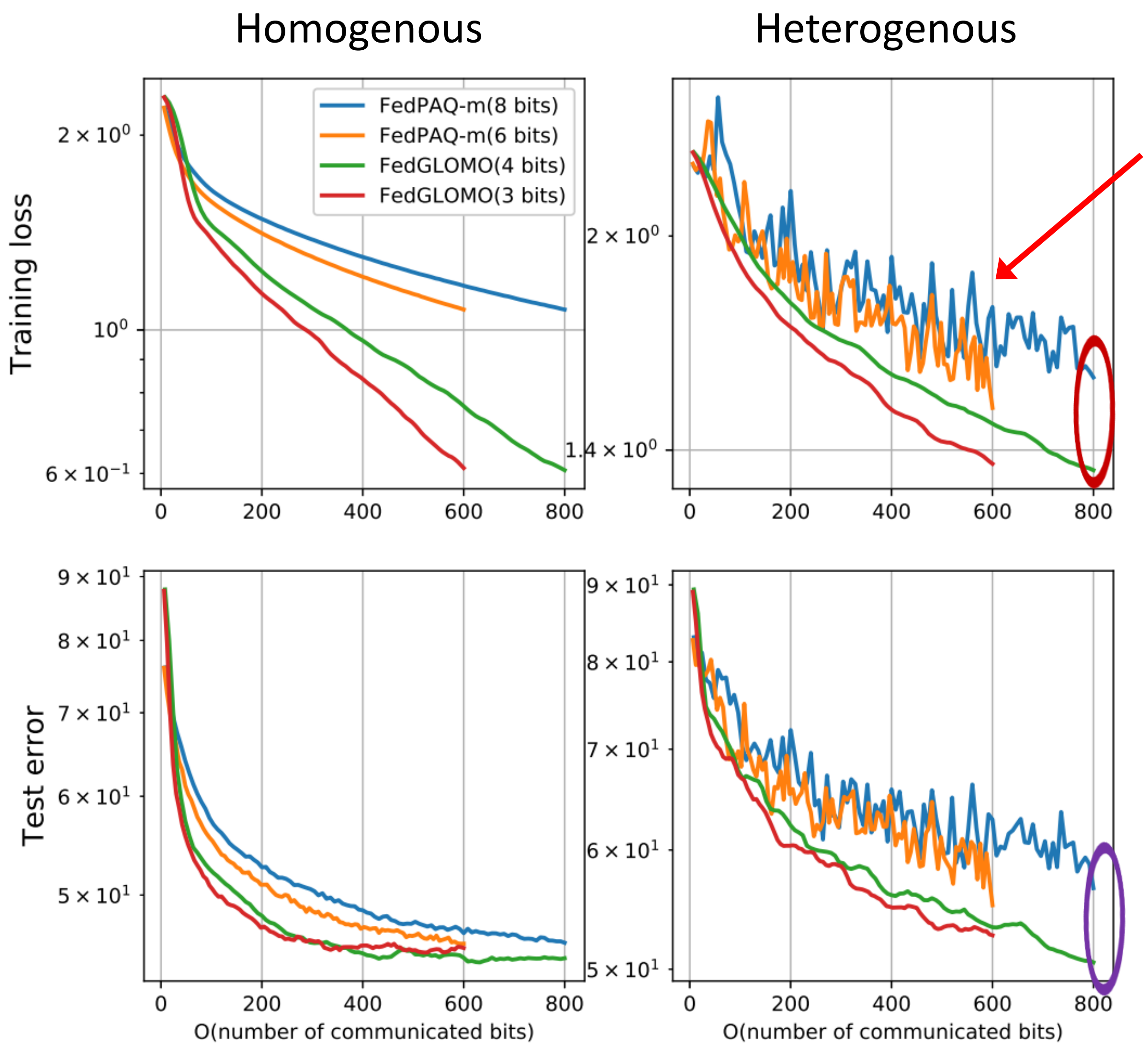
600 neurons

Efficacy of Quantization and Momentum Mechanisms

Baseline: FedPaQ

Proposed: FedGLOMO

An order of magnitude savings in communication resources

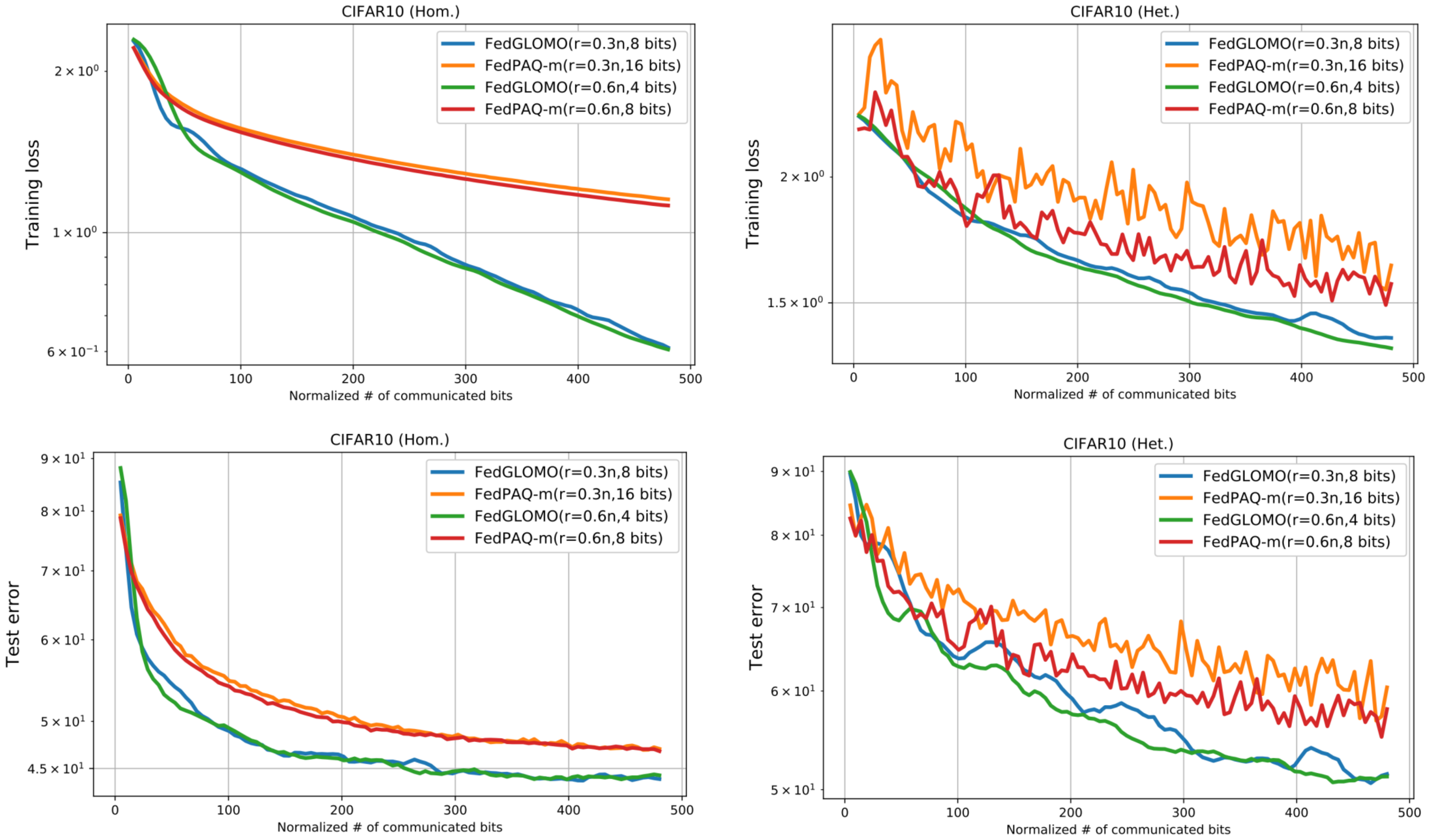


Heterogeneity

Faster training

Better generalization

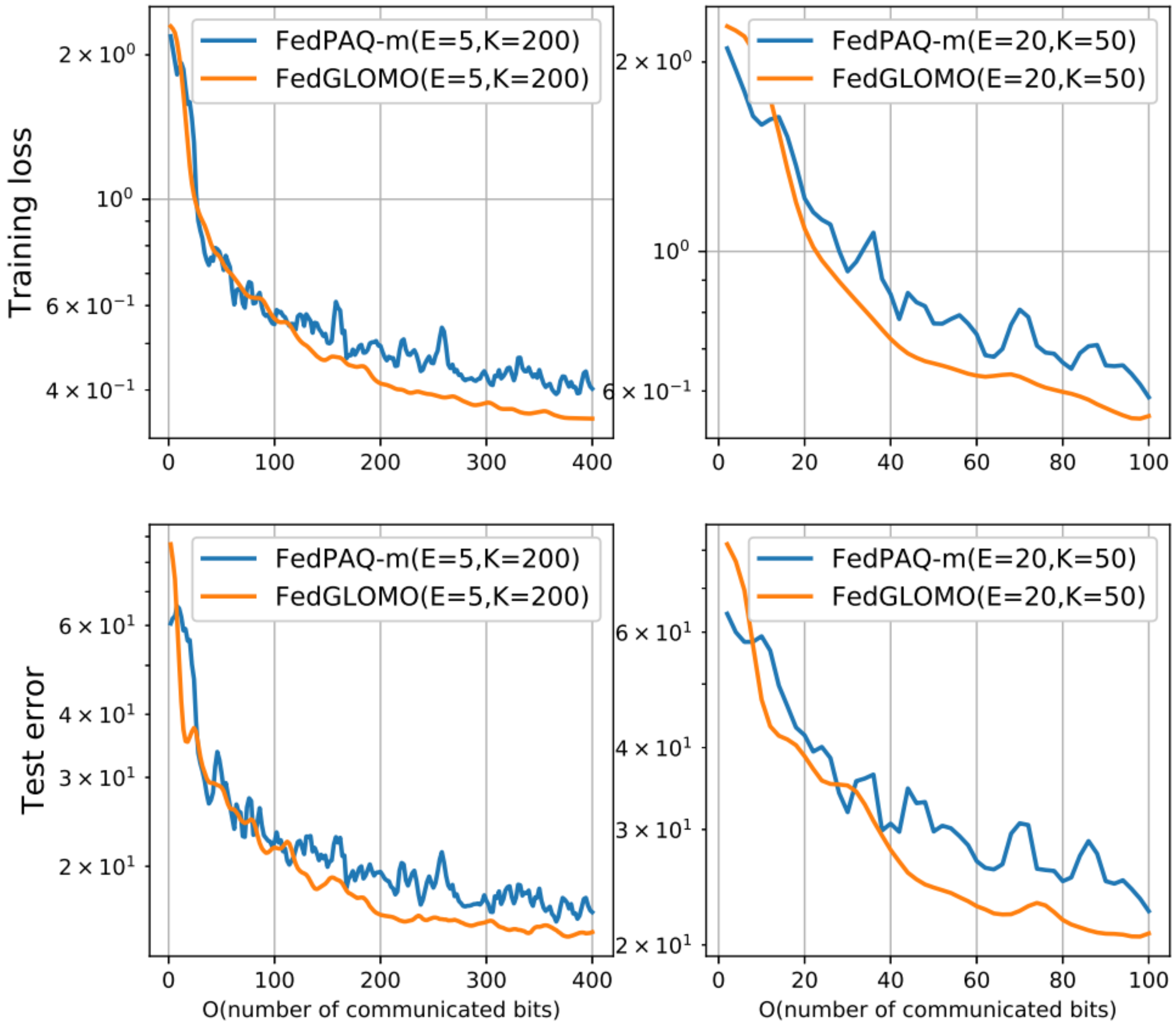
Robustness to Unreliable Communication



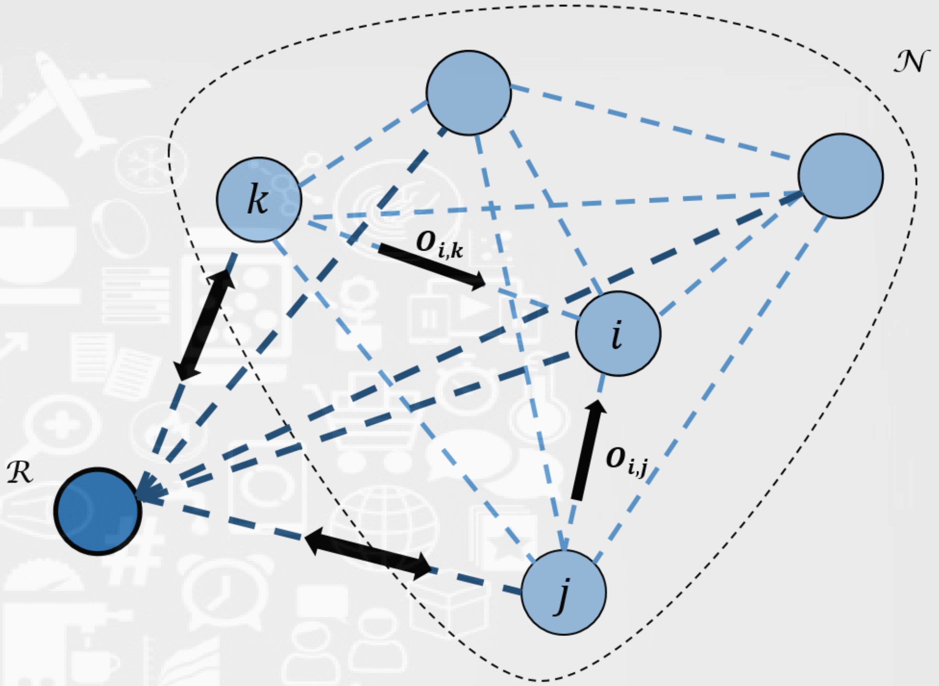
Resiliency to device dropout (smaller *r*)

Robustness to Unreliable Communication

Resiliency to device intermittent availability (larger *E*)



Information Management in Resource-Constrained Sensing Networks



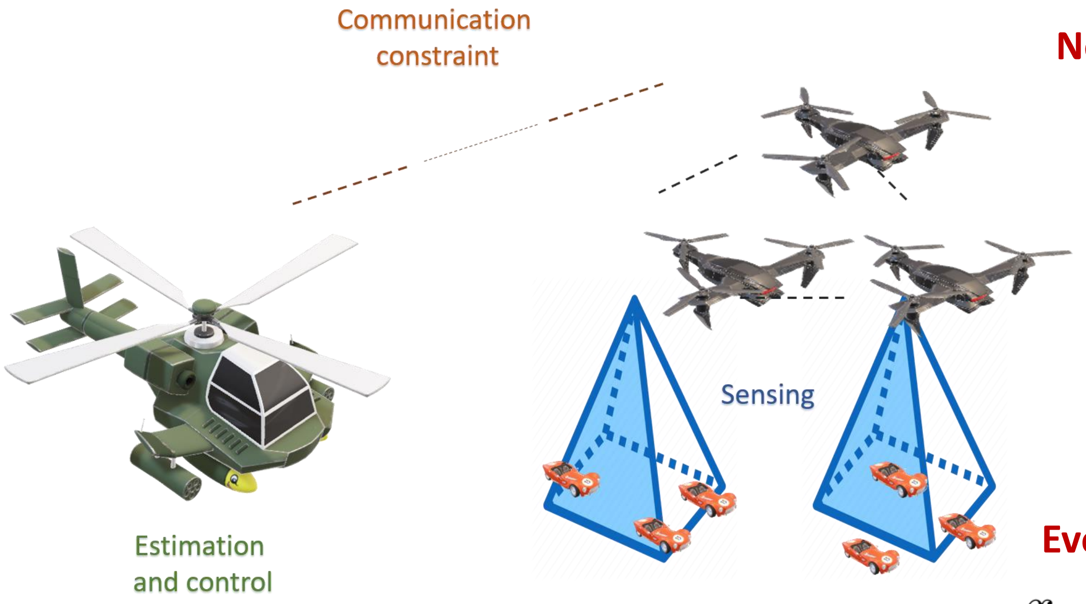
[Hashemi, A., Vikalo, H., de Veciana, G., "Progressive Stochastic Greedy Sparse Reconstruction and Support Selection," Submitted, 2021.]

[Hashemi, A., Ghasemi, M., Vikalo, H., Topcu, U., "Randomized greedy sensor selection: Leveraging weak submodularity," IEEE Transactions on Automatic Control, Jan. 2021.]

[Hashemi, A., Vikalo, H., de Veciana, G., "On the Performance-Complexity Tradeoff in Stochastic Greedy Weak Submodular Optimization," International Conference on Acoustic, Speech and Signal Processing (ICASSP), 2021.]

[Hashemi, A., Ghasemi, M., Vikalo, H., Topcu, U., "Submodular Observation Selection and Information Gathering for Quadratic Models," International Conference on Machine Learning (ICML), Long Beach, CA, June 2019.]

Observation Selection for Sensing Networks



Noisy and time-varying observation model

$$y_t = h(x_t) + \nu_t$$

Evolving unknown states

$$x_t = g(x_{t-1}, u_t) + v_t$$

Observation selection and sharing for low-MSE estimation

$$\mathrm{MSE}_t(S) = \|\hat{x}_t(S) - x_t\|^2$$

<u>Questions</u>

- What should be be the selection criteria?
- How can we perform the selection efficiently and with guaranteed performance?

Observation Selection Criteria

Scalar functions of the predicted error covariance matrix $f\,(P_t(S))$

Constrained combinatorial optimization

$$\hat{S} = \arg\max_{|S| \leq k} f(S)$$

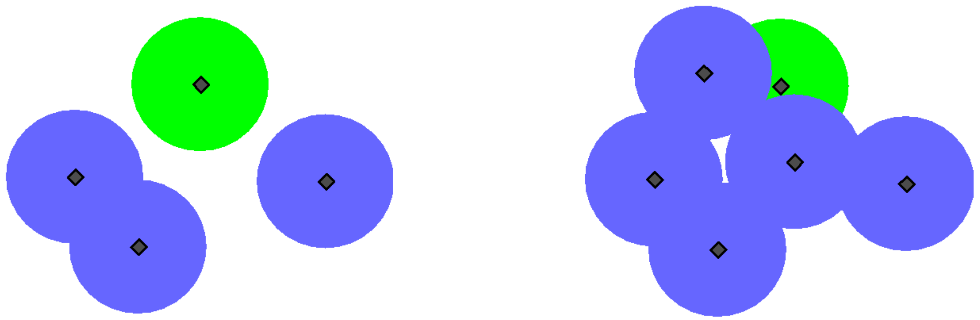
NP-hard

Krause, 2011

f(S) has nice properties:

- Monotonicity
- Weak-submodularity

Hashemi et al., 2018



$$f(A \cup \{d\}) - f(A) \;\geq\; f(B \cup \{d\}) - f(B)$$

Approximate greedy solution

Weak-submodularity constant $0 < \alpha \leq 1$

$$f(\hat{S}) \geq \left(1 - e^{-\alpha}\right) f(S^\star)$$

Optimal approximation guarantee

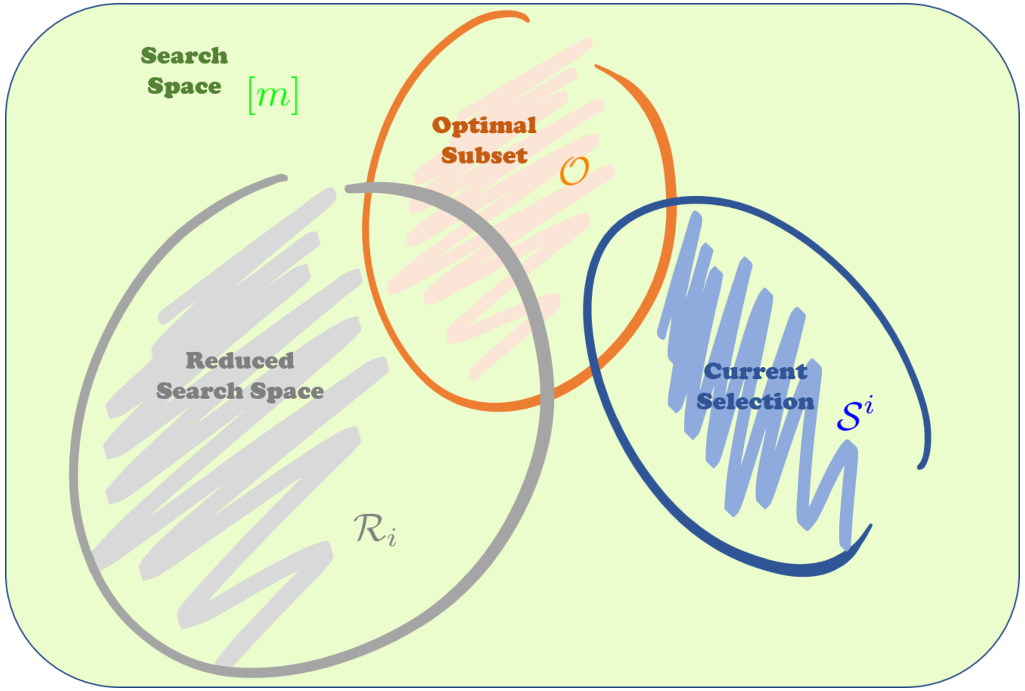
Krause, 2011

Observation Selection in Large-Scale Networks

Greedy selection → Tight approximation guarantee

Greedy selection → Prohibitive computational cost

Reduce the space of greedy by **random sampling**



How to construct R_i

Theorem

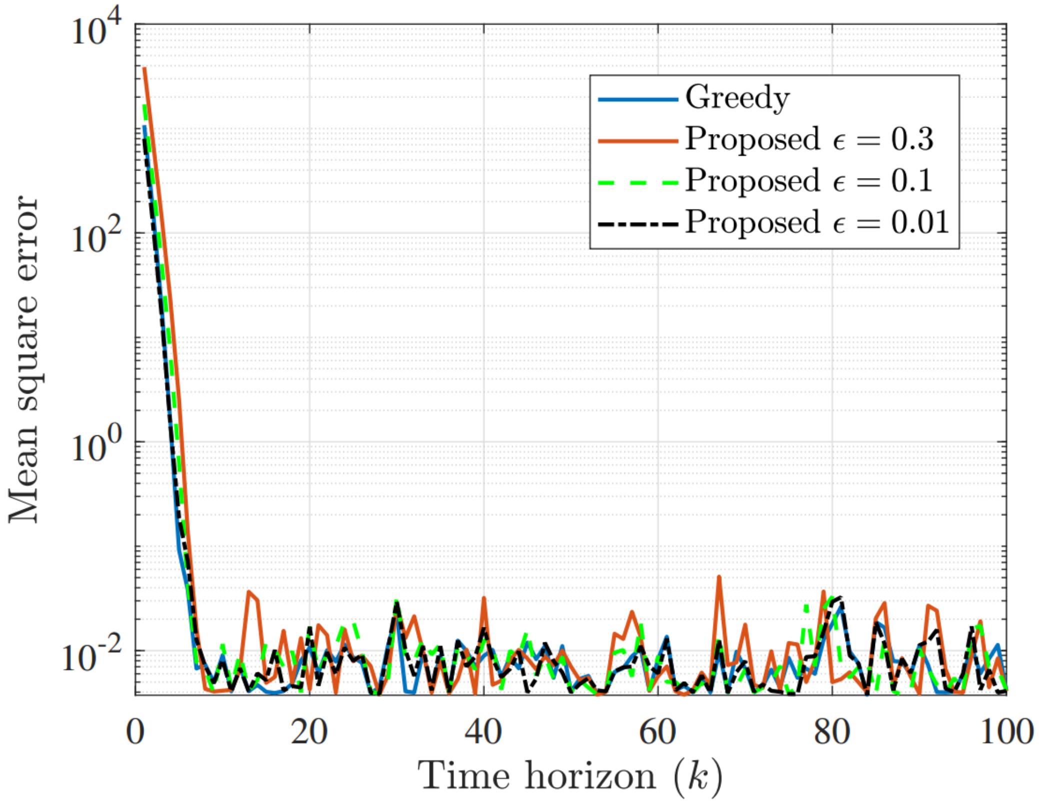
An increasing schedule is required to ensure the intersection is nonempty

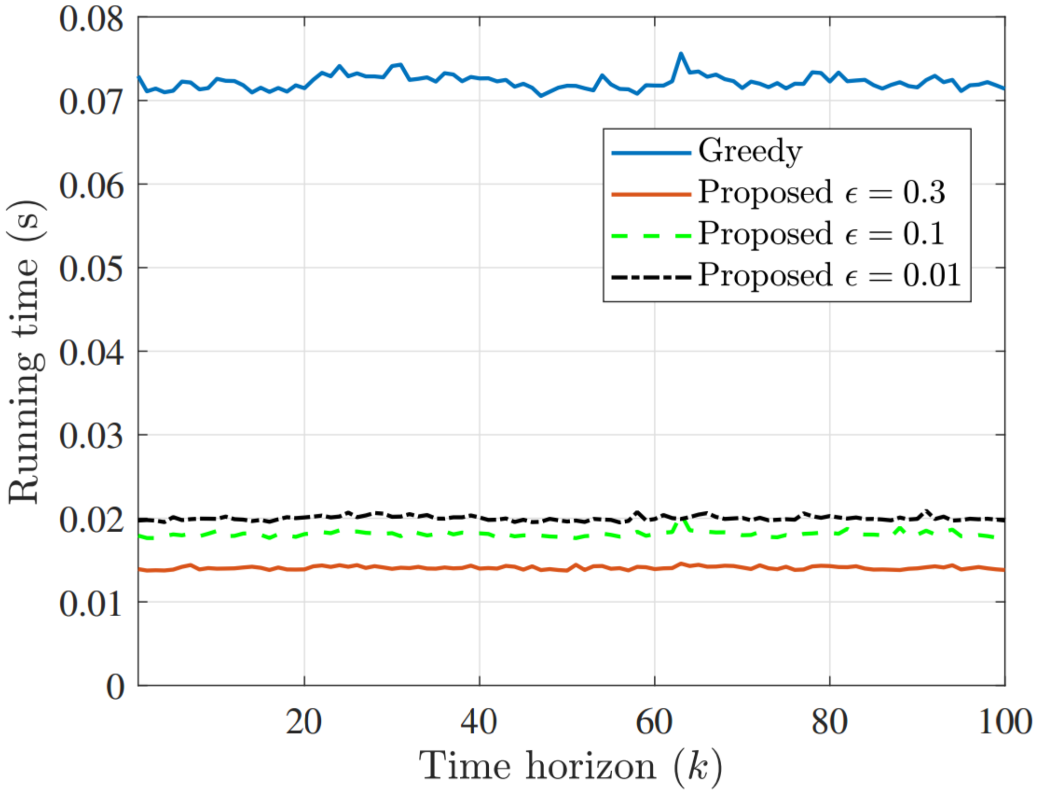
Theorem

Near-optimal expected approximation guarantee

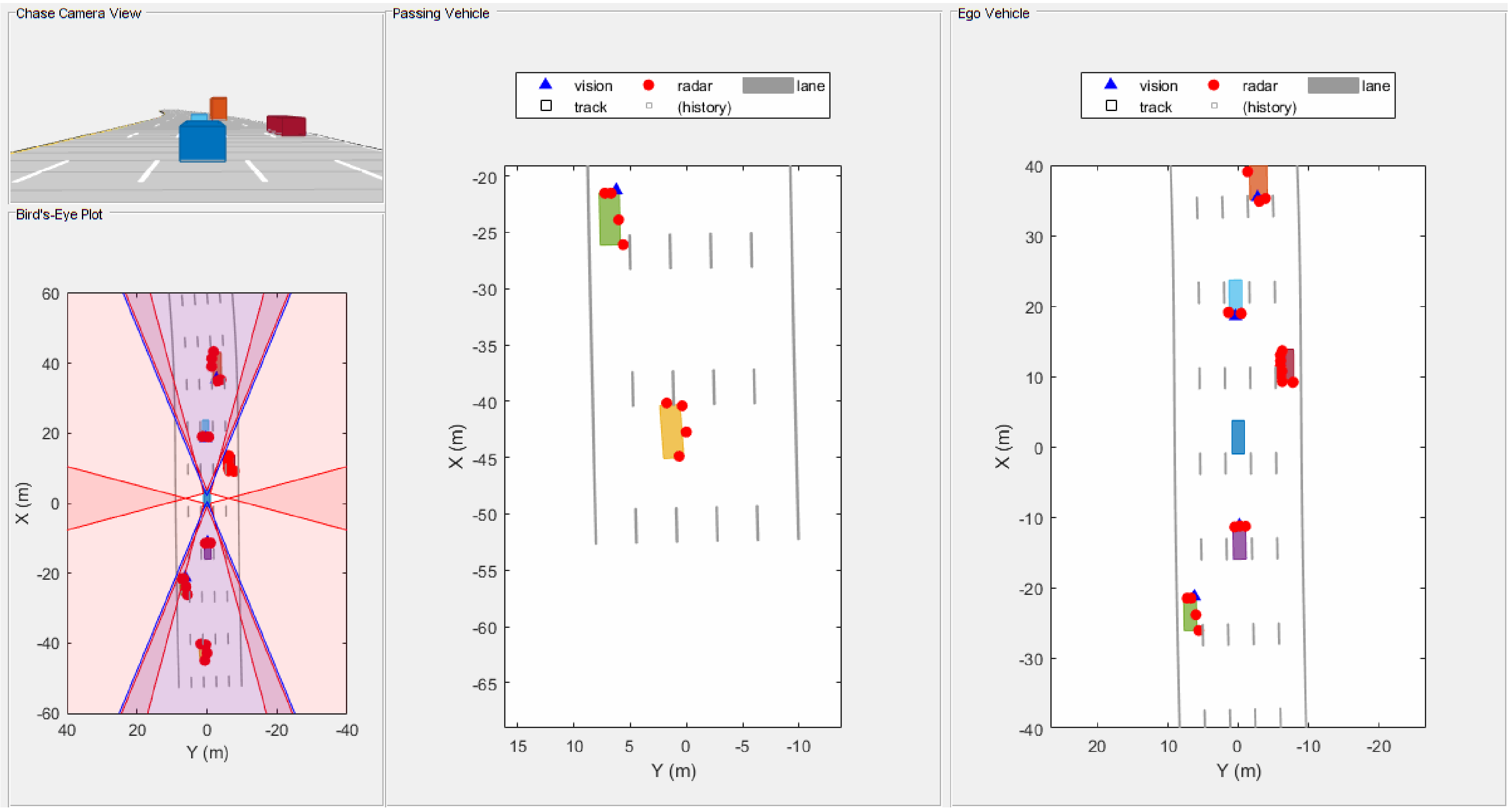
$$\mathbb{E}[f(\hat{S})] \geq \left(1 - e^{-\alpha} - \alpha\epsilon\right) f(S^{\star})$$

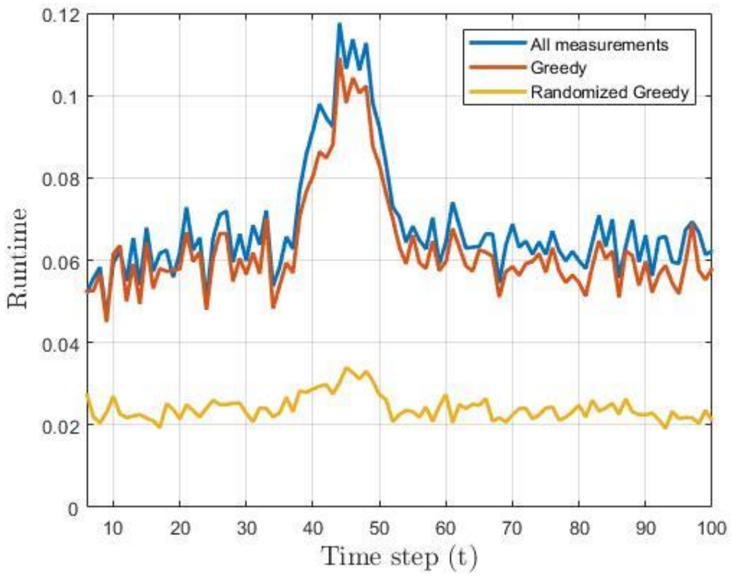
UAV-Based Target Tracking

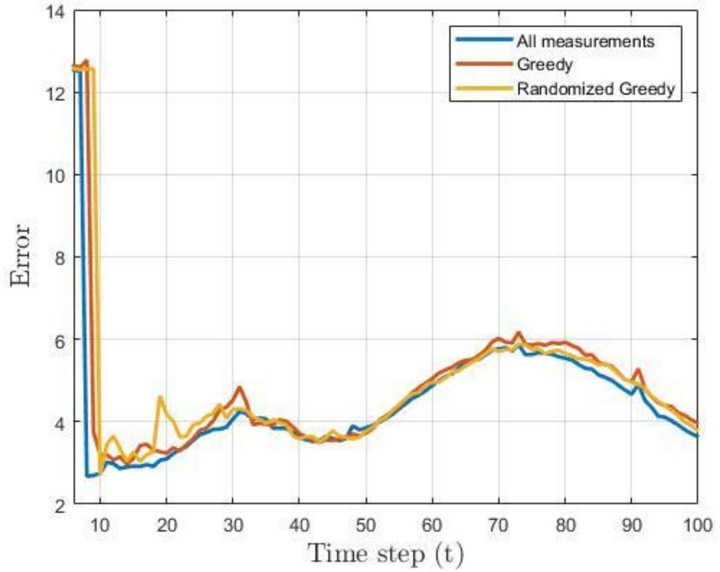




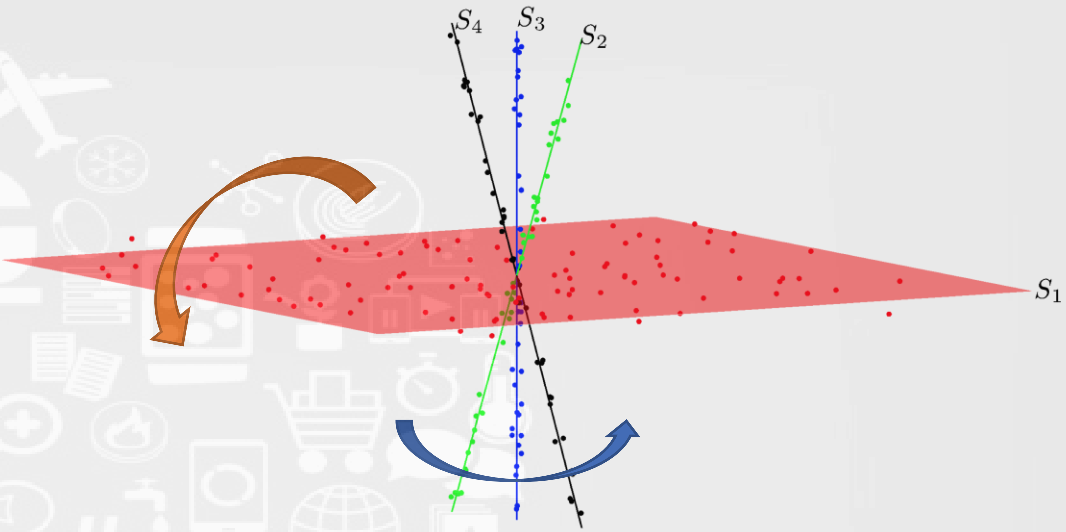
Application in Autonomous Driving







Data-Scarce Parsimonious Representation Learning



[Hashemi, A., Schaeffer, H., Shi, B., Tran, G., Ward, R., "Generalization Bounds for Sparse Random Features Expansions", 2021.]

[Hashemi, A., Zhu, B., Vikalo, H., "Sparse Tensor Decomposition for Haplotype Assembly of Diploids and Polyploids," BMC Genomics, Mar. 2018.]

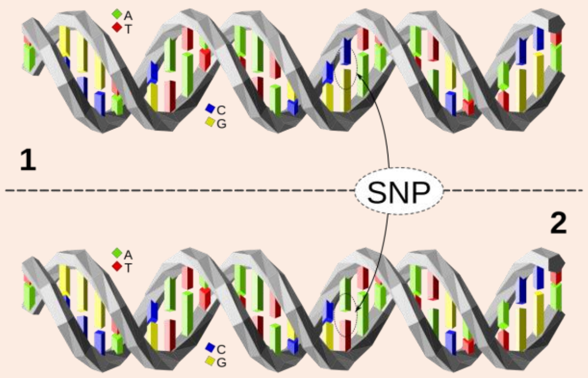
[Hashemi, A. and Vikalo, H., "Evolutionary Self-Expressive Models for Subspace Clustering," IEEE Journal of Selected Topics in Signal Processing, Dec. 2018.]

[Hashemi, A. and Vikalo, H., "Accelerated Orthogonal Least-Squares for Large-Scale Sparse Reconstruction," Digital Signal Processing, Nov. 2018.]

Structured Function Approximation

Low-rank structure

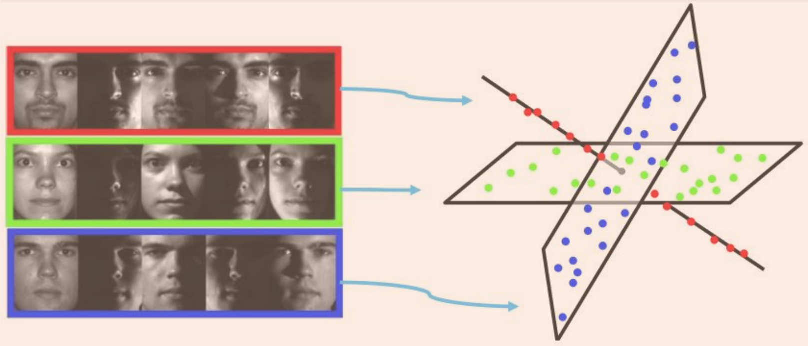
Genome sequencing



Wikipedia

Sparsity

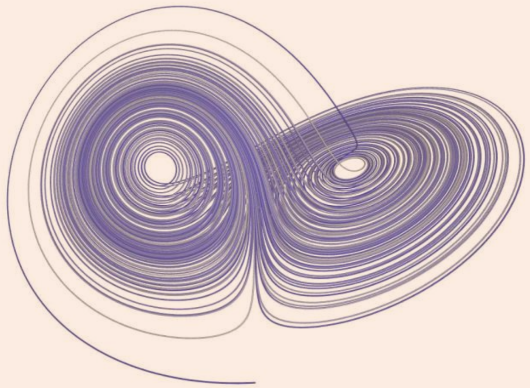
Clustering



You et al., 2018

Low-order structure

Dynamical systems



$$\dot{x}_i = x_{i+1}x_{i-1} - x_{i-1}x_{i-2} - x_i + 8, \quad i = 1, \ldots d$$

Wikipedia

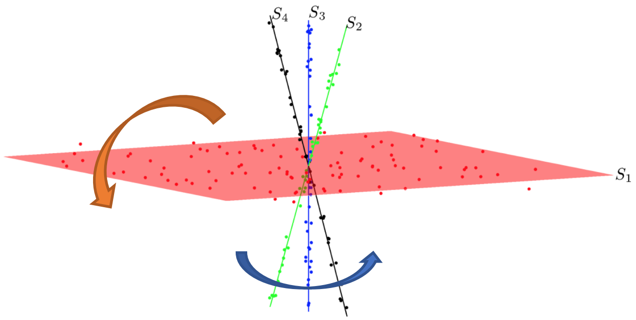
<u>**Goal**</u>

- Learning structured unknown functions from limited measurements

$$\min_{\theta} \sum_{i=1}^{m} \mathrm{dist}(y_i, f(x_i;\theta)) \quad s.t. \quad f(x;\theta) \in \mathcal{F}$$

Parsimonious Representation Learning

Elhamifar et al., 2013



Data X_t

Desired structures

Structure-aware representation learning

Multi-variate, non-convex problems

Alternating minimization

Subspace structure

Evolutionary structure

$$\min_{\mathbf{U},\alpha} \quad \|\mathbf{X}_t - \mathbf{X}_t(\alpha\mathbf{U} + (1-\alpha)\mathbf{C}_{t-1})\|_F^2$$

$$\text{s.t.} \quad \text{diag}(\mathbf{U}) = \mathbf{0}, \quad \|\mathbf{U}\|_0 \leq k, \quad 0 \leq \alpha \leq 1$$

Subspace structure

Theorem

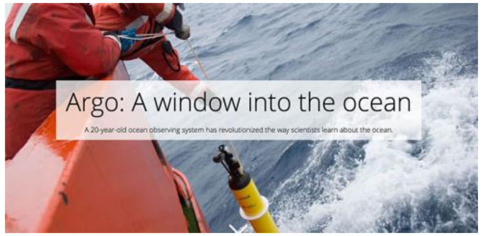
Bound on amount of required data for a target accuracy

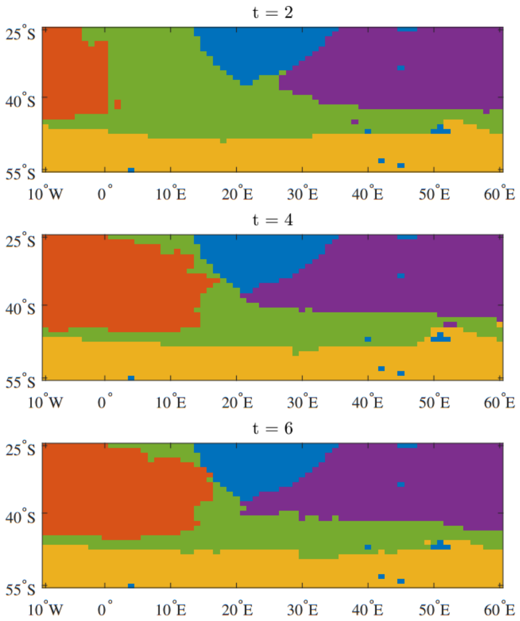
Empirical Applications in Unsupervised Learning

Real-time motion segmentation

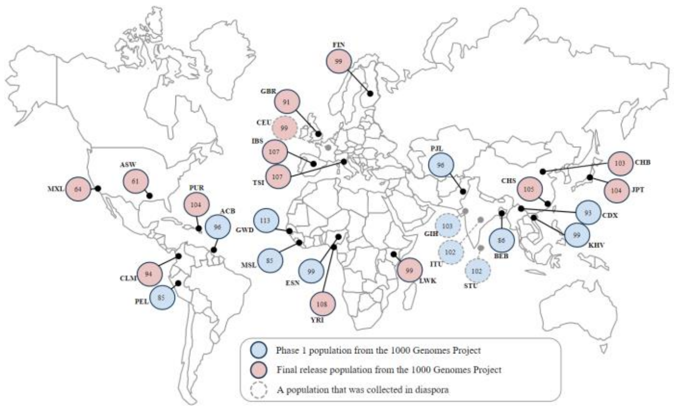
Method	Error (%)	Runtime (s)
Proposed	**5.60**	**1.69**
Baseline	10.76	46.16

Tracking water masses near South Africa





Study of genetic variation



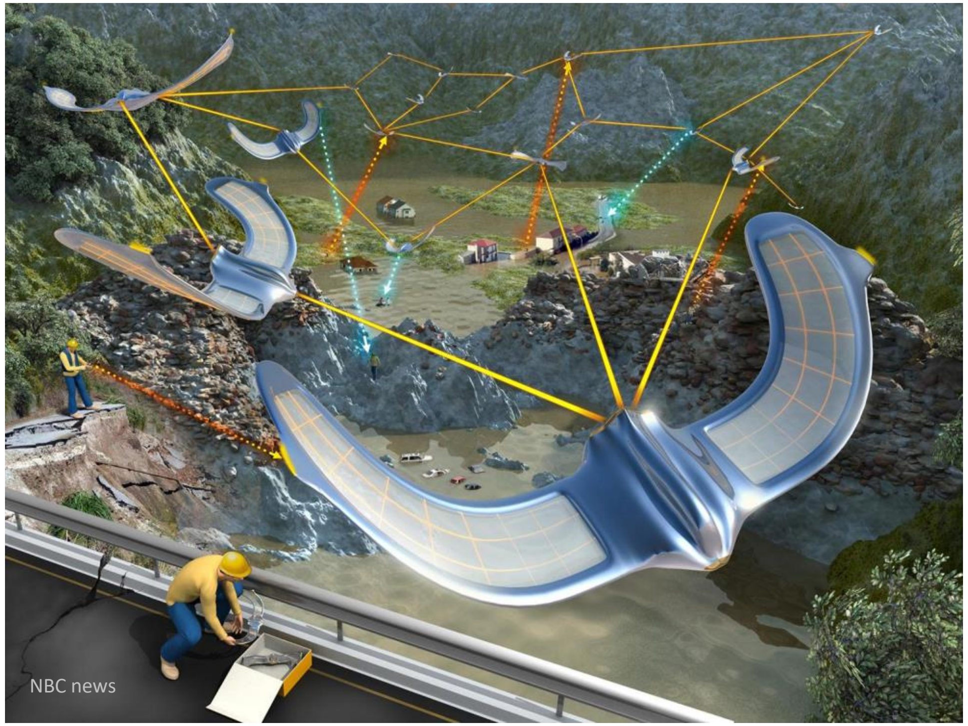
Method	Error (%)	Runtime (s)
Proposed	**0.042**	**4.60**
Baseline	0.152	11.42

Ongoing and Future Work

Collaborative Learning in Dynamic Environments

Adaptive representation learning of dynamic data

Resource-constrained collaborative learning under uncertainty and dynamic heterogeneity



[Ghasemi, M., **Hashemi, A.,** Vikalo, H., Topcu, U., "No-Regret Learning with High-Probability in Adversarial Markov Decision Processes," Conference on Uncertainty in Artificial Intelligence (UAI), 2021]

[Ghasemi, M., **Hashemi, A.,** Topcu, U., Vikalo, H., "Online Learning with Implicit Exploration in Episodic Markov Decision Processes," American Control Conference (ACC), 2021]

Robustness and Security

Collaboration against unexpected contingencies and adversaries

Integrating robust hypothesis testing into information acquisition and representation learning

Exploring the trade-off between privacy and robustness

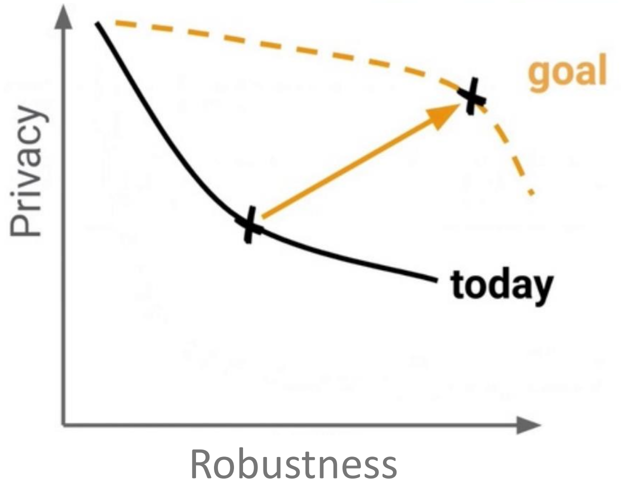
WIRED

Hackers Remotely Kill a Jeep on the Highway—With Me in It

ANDY GREENBERG SECURITY 07.21.15 6:00 AM

Update: Chrysler recalls 1.4M vehicles after Jeep hack

COMPUTERWORLD



[Acharya, A., Hashemi, A., Jain, P., Sanghavi, S., Dhillon, I., Topcu, U., “Robust Training in High Dimensions via Block Coordinate Geometric Median Descent,” Preprint, 2021]

[Das, R., Hashemi, A., Sanghavi, S., Dhillon, I., “DP-NormFedAvg: Normalizing Client Updates for Privacy-Preserving Federated Learning,” Preprint, 2021]

Structured and Resource-Constrained Collaborative Learning

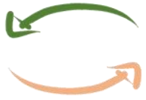
Abolfazl Hashemi

Local momentum-based variance reduction

$$\boldsymbol{v}_\tau^{(i)} = \widetilde{\nabla} f_i(\boldsymbol{w}_\tau^{(i)}; \mathcal{B}_\tau^{(i)}) + (\boldsymbol{v}_{\tau-1}^{(i)} - \widetilde{\nabla} f_i(\boldsymbol{w}_{\tau-1}^{(i)}; \mathcal{B}_\tau^{(i)}))$$

$$\boldsymbol{w}_{\tau+1}^{(i)} = \boldsymbol{w}_t^{(i)} - \eta \boldsymbol{v}_\tau^{(i)}$$

$\boldsymbol{w}_{k+1}$



Intermittent slow uplink from available vehicles

$Q_D(\boldsymbol{w}_k - \boldsymbol{w}_{k,E}^{(i)})$

Global momentum-based variance reduction

$$\boldsymbol{w}_{k+1} = \mathrm{agg}(\{Q_D(\boldsymbol{w}_k - \boldsymbol{w}_{k,E}^{(i)})\})$$