TEXAS
The University of Texas at Austin

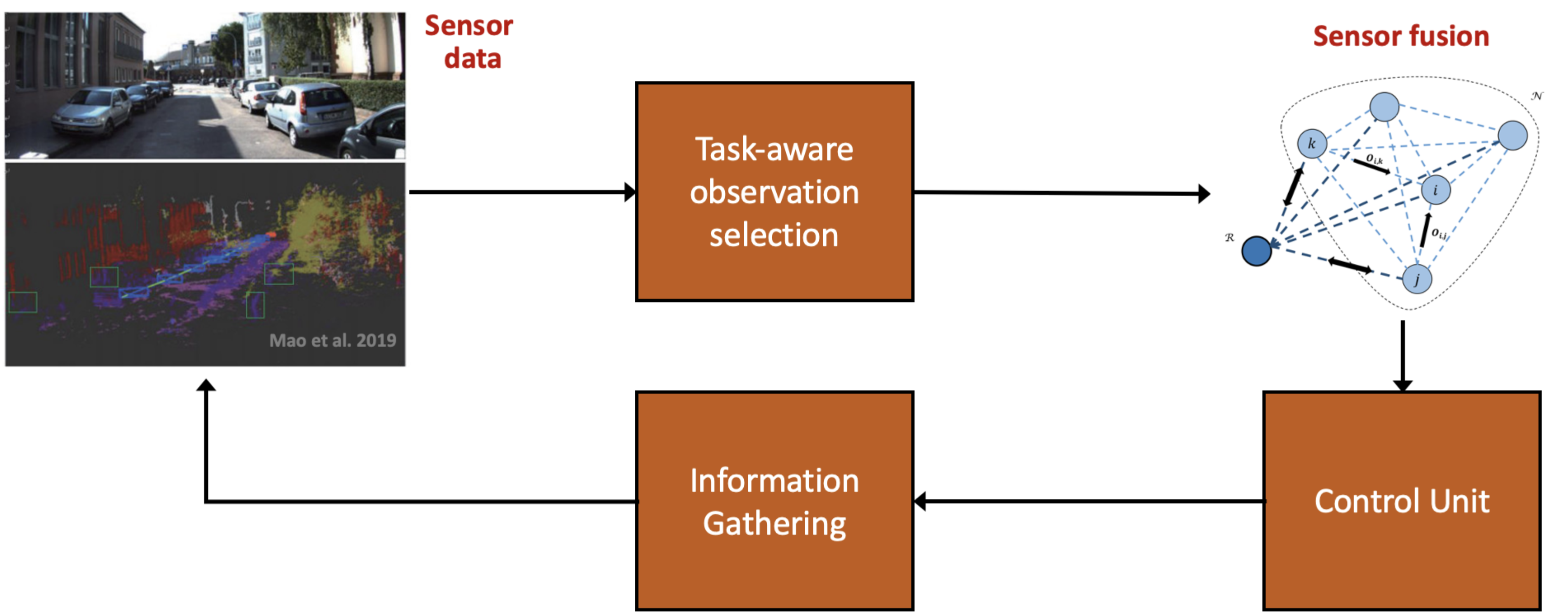
Sensing and Learning in Distributed Systems Operating under Resource Constraints

Haris Vikalo

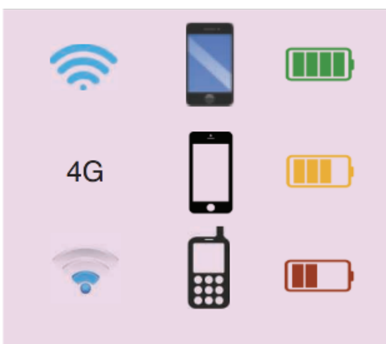
Texas A&M, April 9, 2021

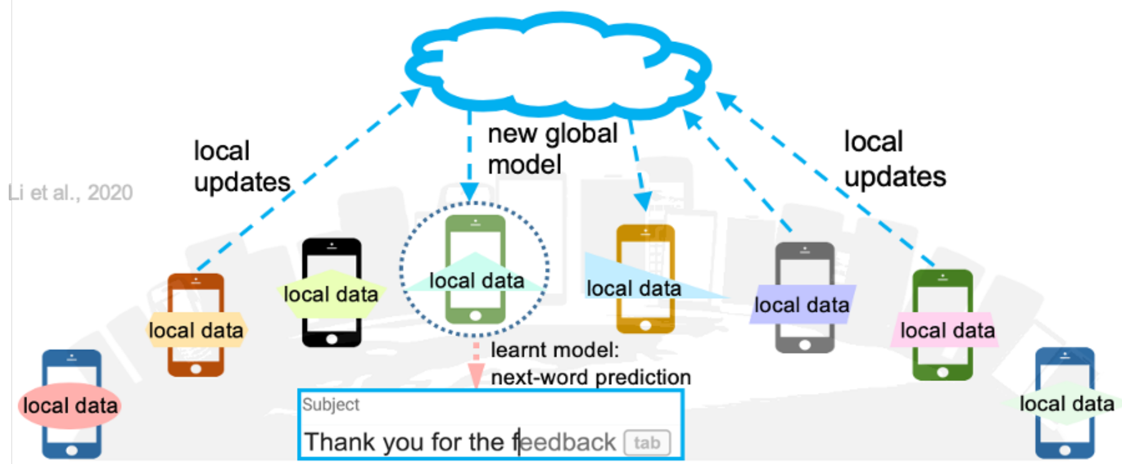
Observation Selection and Information Gathering

- Sensor networks often operate under restrictions on communication bandwidth and computational capabilities



- Federated learning systems: ameliorating privacy concerns yet still communication-intensive





Organization of the Talk

Information Gathering

- Linear models
    - Weak submodularity of the MSE objective
    - Greedier than greedy: Randomized greedy selection
- Beyond linear models: Observation selection for quadratic models
    - Exploiting Van Trees' bound

Privacy preserving ML: Federated Learning

- Client selection as the remote estimation problem
- Exploring communication-accuracy tradeoff

Roadmap

Information Gathering

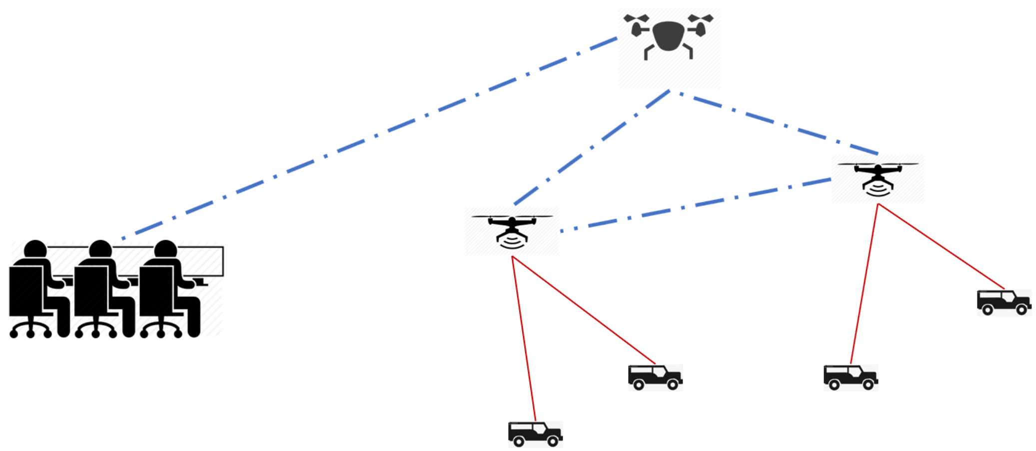
- Linear models
  - Weak submodularity of the MSE objective
  - Greedier than greedy: Randomized greedy selection
- Beyond linear models: Observation selection for quadratic models
  - Exploiting Van Trees' bound

Privacy preserving ML: Federated Learning

- Client selection as the remote estimation problem
- Exploring communication-accuracy tradeoff

An Illustrative Example

- An example of a large-scale sensor network: A swarm of UAVs
  - UAVs gathering measurements of targets' positions
  - location estimation and tracking in a remote control unit



- **The goal**: Computationally efficient selection of informative measurements for accurate (in terms of MSE) target tracking

System Model

- A (linearized) dynamical model:

$$\mathbf{x}_{k+1} = \mathbf{A}_k \mathbf{x}_k + \mathbf{w}_k$$
$$\mathbf{y}_k = \mathbf{H}_k \mathbf{x}_k + \mathbf{v}_k$$

- State and measurement noises: $\mathbf{w}_k = \mathcal{N}(0, \mathbf{Q}_k)$, $\mathbf{v}_k = \mathcal{N}(0, \mathbf{R}_k)$
- At each step k, select a subset S_k of size K from n measurements
- Control unit: track the state vector via (extended) Kalman filter based on the communicated measurements:

(predicted error covariance) $\mathbf{P}_{k|k-1} = \mathbf{A}_k \mathbf{P}_{k-1|k-1} \mathbf{A}_k^\top + \mathbf{Q}_k$

(filtered error covariance) $\mathbf{P}_{k|k,S_k} = \left(\mathbf{P}_{k|k-1}^{-1} + \mathbf{H}_{k,S_k}^\top \mathbf{R}_{k,S_k}^{-1} \mathbf{H}_{k,S_k} \right)^{-1}$

Selection of Observations

- Mean-square error of the state estimate at k: $\mathrm{MSE}_{S_k} = \mathrm{Tr}(\mathbf{P}_{k|k,S_k})$
- Select a subset S of size K to achieve the lowest estimation MSE

$$\begin{aligned} \underset{S}{\text{minimize}} \quad & \mathrm{Tr}\left(\mathbf{F}_S^{-1}\right) \\ \text{subject to} \quad & S \subset [n], \quad |S| = K \end{aligned}$$

  - $\mathbf{F}_S = \mathbf{P}_{k|k,S}^{-1}$: The Fisher information matrix
- Challenges:
  - An NP-hard, combinatorial problem [Natarajan'95]; due to high computational complexity, resort to approximate methods
  - Massive amounts of sensory data $\rightarrow$ need accelerated schemes

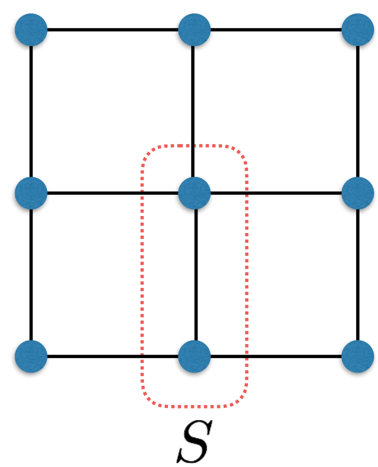
Related Work and Our Contribution

- Existing approaches
  - Using a surrogate objective function (e.g., $\log\det(\mathbf{P}_{k|k,S_k})$) [Joshi'09, Shamaiah'10, Mirzasoleyman'15, Tzoumas'16]
    - submodular (and thus efficient algorithms come with performance guarantees) but not explicitly related to MSE, the desired objective
  - Greedy schemes for MSE formulation [Singh'17, Chamon'17]
    - iteratively selecting sensors, one at each iteration
    - $\mathcal{O}(nKm^2)$ complexity $\rightarrow$ not suitable for large-scale networks
- Our work: A **randomized greedy** algorithm for the **MSE** objective
  - demonstrating, exploiting weak submodularity of the MSE
  - $\mathcal{O}(nm^2)$ complexity $\rightarrow$ $\mathcal{O}(K)$ gain in speed
  - theoretical bound on worst-case MSE, near-optimal performance

Background: Set Functions

- Set function: a function that assigns a value to each subset of the ground set X (e.g., the set of all sensors in a network)

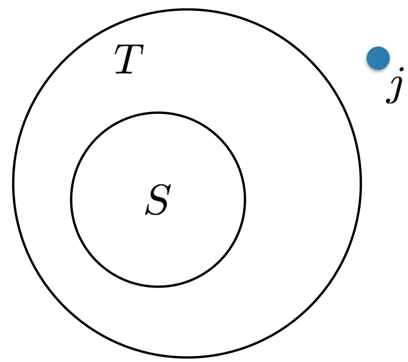
  Example: The value of a cut $f(S)$ for all $S \subseteq V$ in an undirected graph $G = (V, E)$.



- Monotonicity: $f(S) \leq f(T)$ for all $S \subseteq T \subseteq X$

Background Cont'd: (Weak) Submodularity

- Marginal gain: $f_j(S) = f(S \cup \{j\}) - f(S)$, i.e., the gain obtained by adding j to S



- Submodularity: $f_j(T) \leq f_j(S)$ for all $S \subseteq T \subset X$ and $j \in X \backslash T$
  - diminishing returns property
- Weak Submodularity: $f_j(T) \leq \mathcal{C} \times f_j(S)$ where $\mathcal{C} > 1$ is the max (over all combinations of (S, T, j)) element-wise curvature of f

Weak Submodularity of Sensor Selection

- Define $f(S) = \mathrm{Tr}\left(\mathbf{P}_{k|k-1} - \mathbf{F}_S^{-1}\right)$ (inverse additive of MSE)
  - a maximization task equivalent to MMSE:

$$\begin{aligned} \max_{S} \quad & f(S) \\ \text{s.t.} \quad & S \subset [n], \quad |S| = K. \end{aligned}$$

- Useful observations:
  - $f(S)$ is monotone (higher values as we keep selecting more sensors)
  - An efficient formula for marginal gain using matrix inversion lemma:

$$f_j(S) = \frac{\mathbf{h}_{k,j}^\top \mathbf{F}_S^{-2} \mathbf{h}_{k,j}}{\sigma_j^2 + \mathbf{h}_{k,j}^\top \mathbf{F}_S^{-1} \mathbf{h}_{k,j}}$$

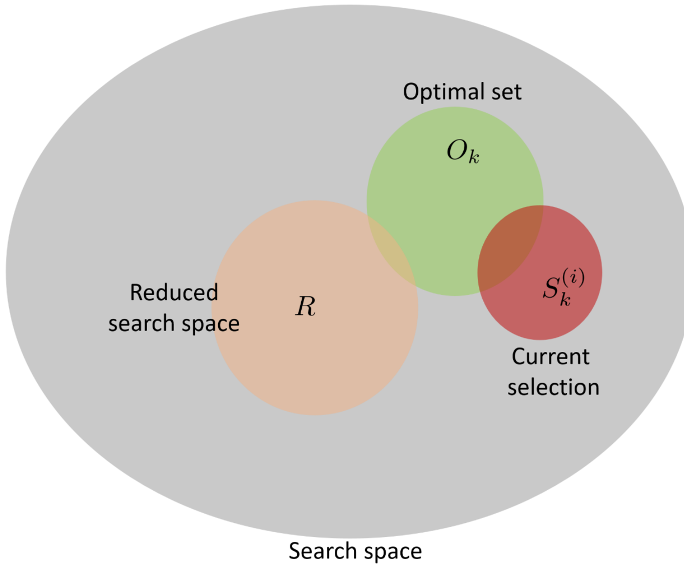
  where $\mathbf{R}_k = \mathrm{diag}(\sigma_1^2, \ldots, \sigma_n^2)$ (independent measurements)

Weak Submodularity of Sensor Selection Cont'd

- While not submodular, under certain conditions $f(S)$ has bounded maximum element-wise curvature [Hashemi et al., 2021]
  - deterministic bound on $\mathcal{C}$ under a constraint on $\lambda_{\mathbf{max}}(\mathbf{H}_k^T\mathbf{H}_k)$
  - probabilistic bounds if $\mathbf{h}_{k,j}$ are i.i.d. with bounded variance
- Informally, these results imply that for a well-conditioned $P_{k|k-1}$, the curvature of $f(S)$ is small (i.e., $f(S)$ is weak submodular)
- We still need fast algorithms for solving large scale sensor selection problems...

Randomized Greedy Sensor Selection

- **The main idea:** Perform greedy search over only a subset of the search space



- Construct R by sampling uniformly at random (no replacement)
- A condition for accuracy: intersection of R with O_k
  - $|R| = \frac{n}{K} \log(\frac{1}{\epsilon}) \rightarrow$ intersection with high probability
  - $0 < \epsilon < 1$: controlling size of the search space

Outline of the Algorithm

- Initialize: $S_k^{(0)} = \emptyset$, $\mathbf{F}^{-1}_{S_k^{(\mathbf{0})}} = \mathbf{P}_{k|k-1}$ (initial Fisher information)
- In each iteration:
  - select a subset R of size $\frac{n}{K}\log(\frac{1}{\epsilon})$ uniformly at random and without replacement from the set of all sensors
  - identify sensor $i_s \in R$ with the largest marginal gain
  - update the selected subset:

$$S_k^{(i+1)} = S_k^{(i)} \cup \{i_s\}$$

Theoretical Performance Guarantees

- On expectation, not too far from the optimal solution

$$\mathbb{E}[f(S_k)] \geq \underbrace{(1 - e^{-\frac{1}{c}} - \frac{\epsilon^\beta}{c})}_{\alpha} f(O_k),$$

where $c = \max\{1, \mathcal{C}\}$, $e^{-K} \leq \epsilon \leq 1$, and $\beta \geq 1$ is a function of $|R|$.

- Bound on expected MSE:

$$\mathbb{E}\left[\text{MSE}_{S_k}\right] \leq \alpha \text{MSE}_{O_k} + (1 - \alpha)\text{Tr}(\mathbf{P}_{k|k-1}).$$

- Running time of the algorithm is $\mathcal{O}(nm^2 \log(\frac{1}{\epsilon}))$
    - $\mathcal{O}(K)$ gain in speed compared to greedy

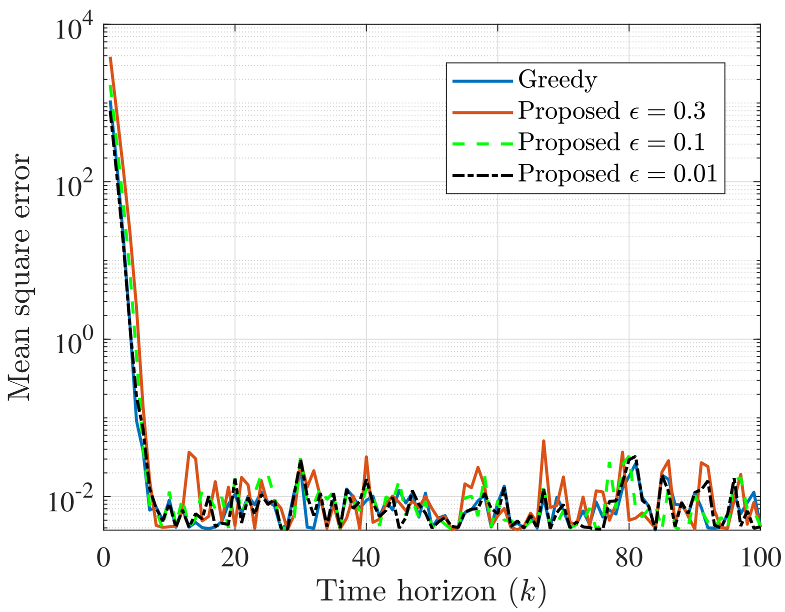
Simulation Results

A comparison with the classic greedy algorithm and the SDP relaxation

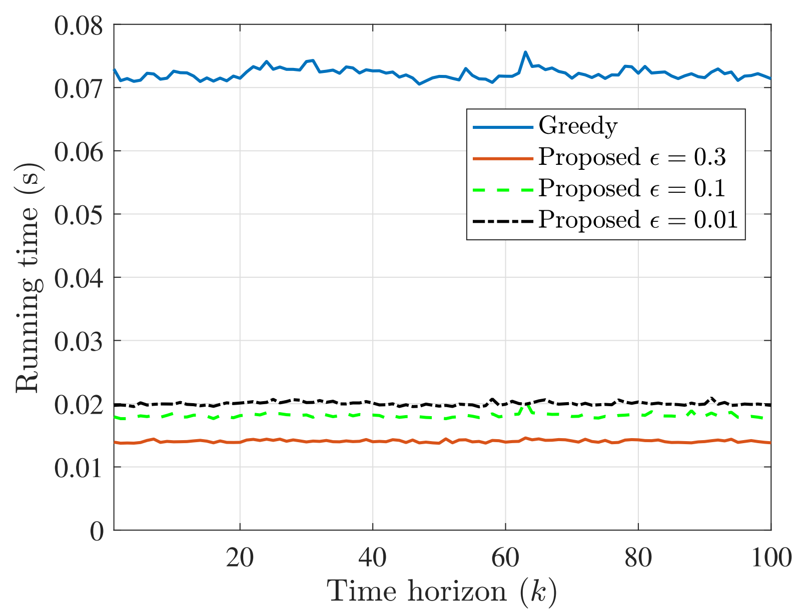
- The settings: State estimation in linear/linearized systems with Kalman filter / EKF
- Investigated accuracy/runtime tradeoff, scalability (network size) and the impact of search randomization

Results (1): Accelerated Multi-Target Tracking

- Tracking the state vector over a period of 100 time steps
- There are $m = 20$ targets; we select $K = 100$ out of $n = 600$ measurements



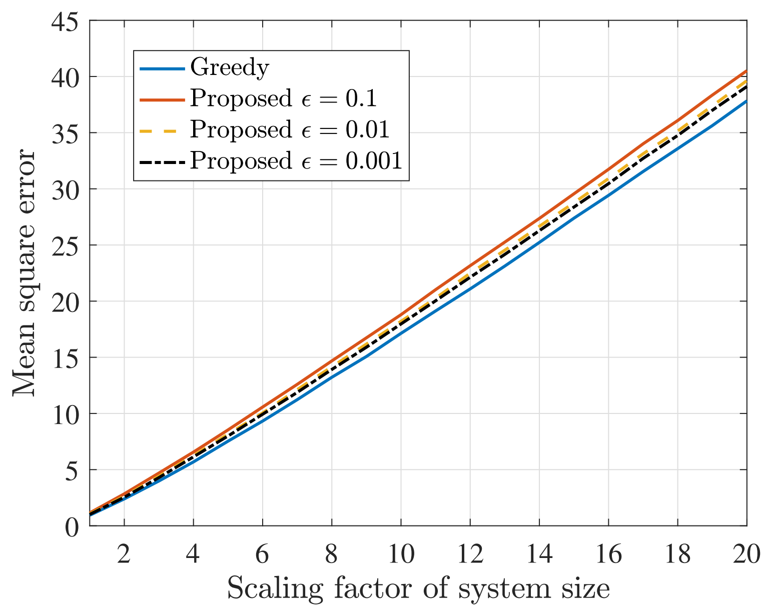
(a) MSE comparison



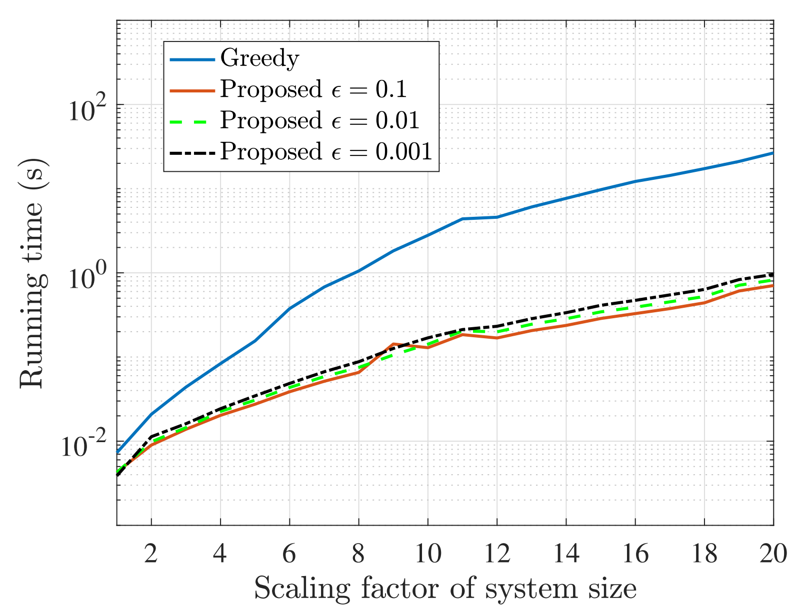
(b) Running time comparison

Results (2): A Scalability Study

- Start with a linear dynamical system with $m = 20$, $n = 200$, $K = 25$
- Scaling it up to 20X



(c) MSE comparison



(d) Running time comparison

Roadmap

Information Gathering

- Linear models
  - Weak submodularity of the MSE objective
  - Greedier than greedy: Randomized greedy selection
- Beyond linear models: Observation selection for quadratic models
  - Exploiting Van Trees' bound

Privacy preserving ML: Federated Learning

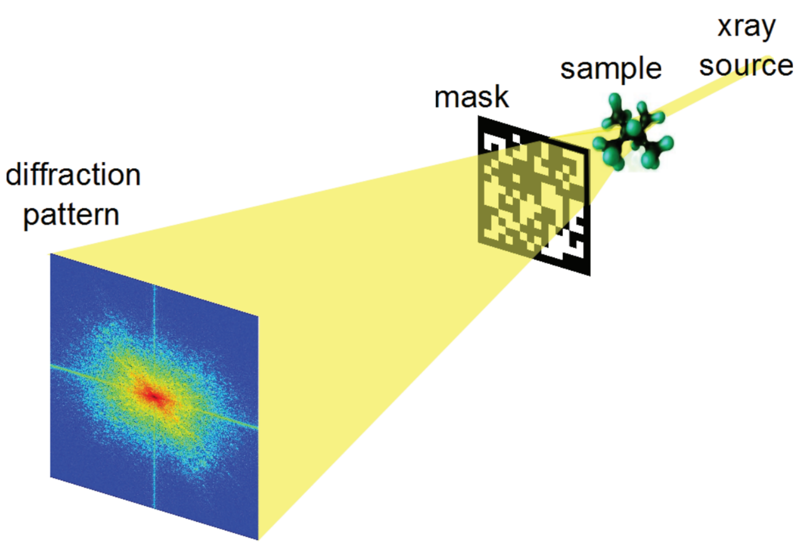
- Client selection as the remote estimation problem
- Exploring communication-accuracy tradeoff

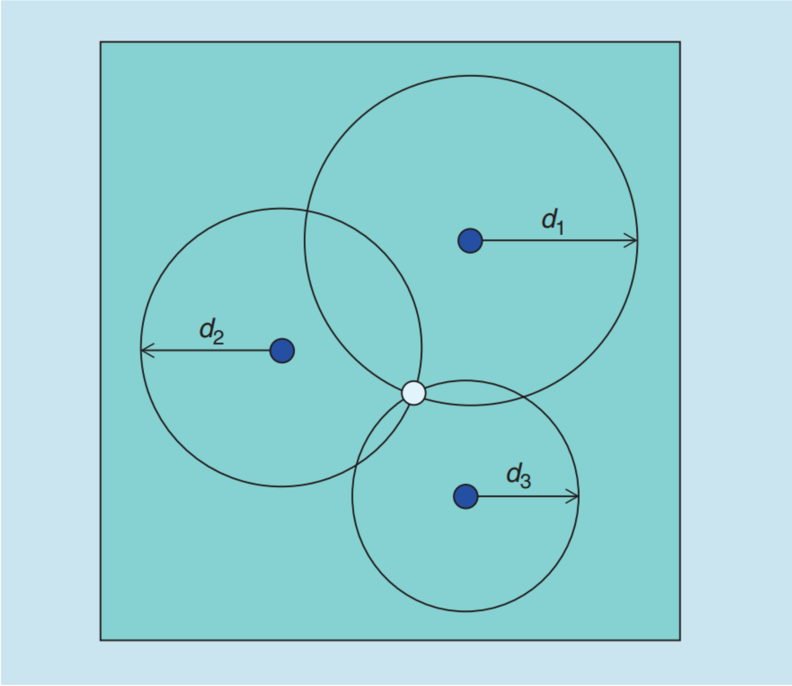
Beyond Linear Observation Models

- Measurement models are often non-linear
  - phase retrieval, object tracking and localization in robotics and autonomous systems
- Existing methods for information gathering selection typically rely on Monte Carlo methods or linearization of the utility function
  - determining informativeness of an observation in terms of metrics of interest becomes challenging
  - greedy algorithms no longer come with performance guarantees

Quadratic relation between observations and unknown parameters

$$\mathbf{y}_i = \underbrace{\frac{1}{2}\mathbf{x}^\top \mathbf{Z}_i \mathbf{x} + \mathbf{h}_i^\top \mathbf{x}}_{g_i(\mathbf{x})} + \mathbf{v}_i\ , \quad i \in \{1, 2, \ldots, n\}$$





(a) Phase retrieval: $y_i = \frac{1}{2}\mathbf{x}^*(\mathbf{z}_i\mathbf{z}_i^*)\mathbf{x} + v_i$ (b) Localization: $\mathbf{y}_i = \frac{1}{2}\|\mathbf{h}_i - \mathbf{x}\|_2^2 + \mathbf{v}_i$

(Figures from [Candes'15] and [Gezici'05])

Prior Work on Selection in Quadratic Models

- Challenge: Unknown optimal estimator and error covariance matrix
- Locally-optimal selection [Flaherty'06, Krause'08]: Linearize around a guess $\mathbf{x}_0$

$$\hat{y}_i := y_i - g_i(\mathbf{x}_0) \approx \nabla g_i(\mathbf{x}_0)^\top \mathbf{x} + v_i,$$

  and find an approximate covariance matrix:

$$\hat{\mathbf{P}}_{\mathcal{S}} = \left(\boldsymbol{\Sigma}_x^{-1} + \sum_{i \in \mathcal{S}} \frac{1}{\sigma_i^2} \nabla g_i(\mathbf{x}_0) \nabla g_i(\mathbf{x}_0)^\top \right)^{-1}$$

- The observation selection becomes

$$\begin{aligned} \underset{\mathcal{S}}{\text{minimize}} \quad & \mathrm{Tr}\left(\hat{\mathbf{P}}_{\mathcal{S}}\right) \\ \text{s.t.} \quad & \mathcal{S} \subset [n], \;\; |\mathcal{S}| = K \end{aligned}$$

Proposed Approach: VTB for Quadratic Models

TEXAS
The University of Texas at Austin

Main Idea

Exploiting Van Trees' bound (VTB) on the error covariance matrix of potentially biased estimators

- A closed-form expression for VTB of quadratic models

Theorem

For any weakly biased estimator $\hat{\mathbf{x}}_{\mathcal{S}}$ with error covariance $\mathbf{P}_{\mathcal{S}}$ it holds that

$$\mathbf{P}_{\mathcal{S}} \succeq \left(\sum_{i \in \mathcal{S}} \frac{1}{\sigma_i^2} \left(\mathbf{Z}_i \boldsymbol{\Sigma}_x \mathbf{Z}_i^\top + \mathbf{h}_i \mathbf{h}_i^\top \right) + \mathbf{I}_x \right)^{-1} = \mathbf{B}_{\mathcal{S}}$$

- Proposed method: Find $\mathcal{S}$ by greedily maximizing Tr(.) scalarization of $\mathbf{B}_{\mathcal{S}}$: $f^A(\mathcal{S}) := \mathrm{Tr}(\mathbf{I}_x^{-1} - \mathbf{B}_{\mathcal{S}})$

Characterizing $f^A(\mathcal{S})$

Theorem

$f^A(\mathcal{S})$ is a monotone, weak submodular set function (i.e., bounded α_{f^A}).

- interpretation of bound on α_{f^A} as an SNR condition

Greedy maximization performance:

$$f^A(\mathcal{S}) \geq (1 - e^{-\frac{1}{\alpha_{f^A}}}) f(\mathcal{O})$$

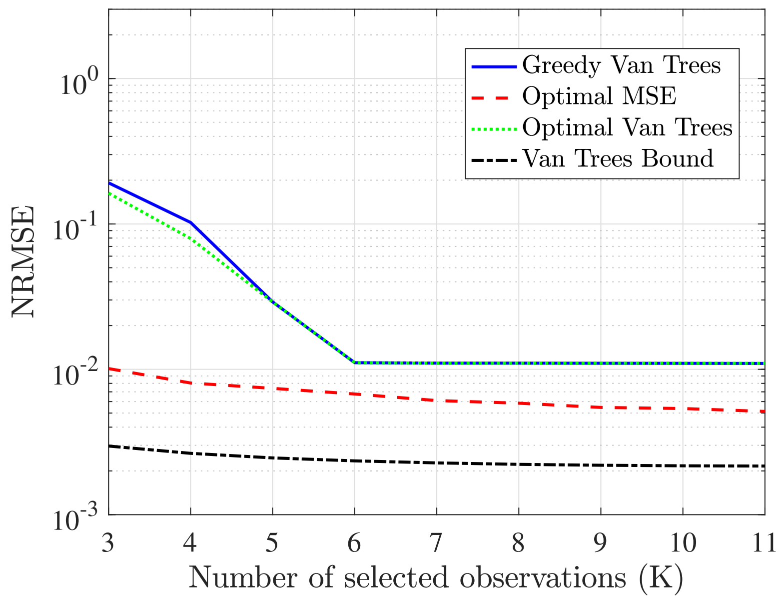
Remark: Obtained submodularity characterization for other criteria:

- $f^T(\mathcal{S}) = \mathrm{Tr}(\mathbf{B}_{\mathcal{S}}^{-1}) - \mathrm{Tr}(\mathbf{I}_x)$ is monotone modular
- $f^D(\mathcal{S}) = \log\det(\mathbf{B}_{\mathcal{S}}^{-1}) - \log\det(\mathbf{I}_x)$ is monotone submodular
- $f^E(\mathcal{S}) = \lambda_{\mathbf{min}}(\mathbf{B}_{\mathcal{S}}^{-1}) - \lambda_{\mathbf{min}}(\mathbf{I}_x)$ is monotone and weak submodular

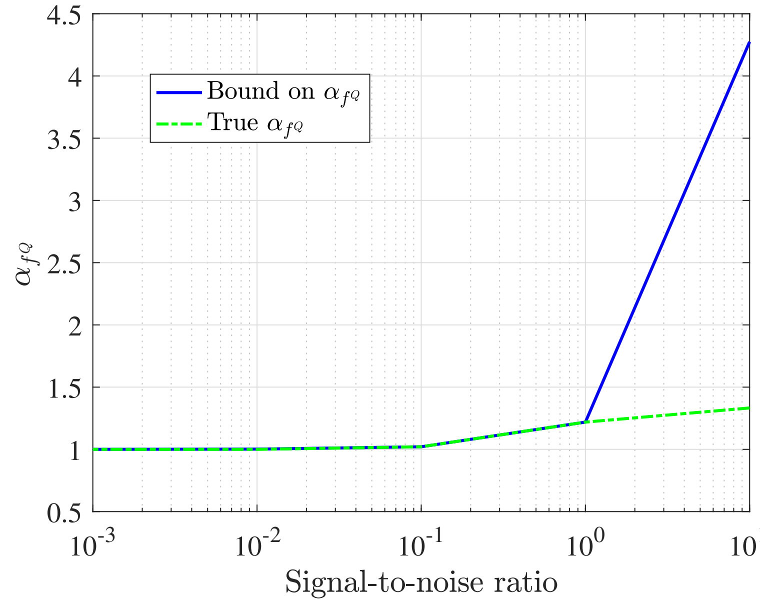
Simulation Results

- Demonstration of the tightness of the Van Trees bound
- A comparison of the VTB based observation selection vs. selection based on linearization of quadratic models
  - applications to phase retrieval, multi-target tracking

- The phase retrieval problem with $n = 12$ observations



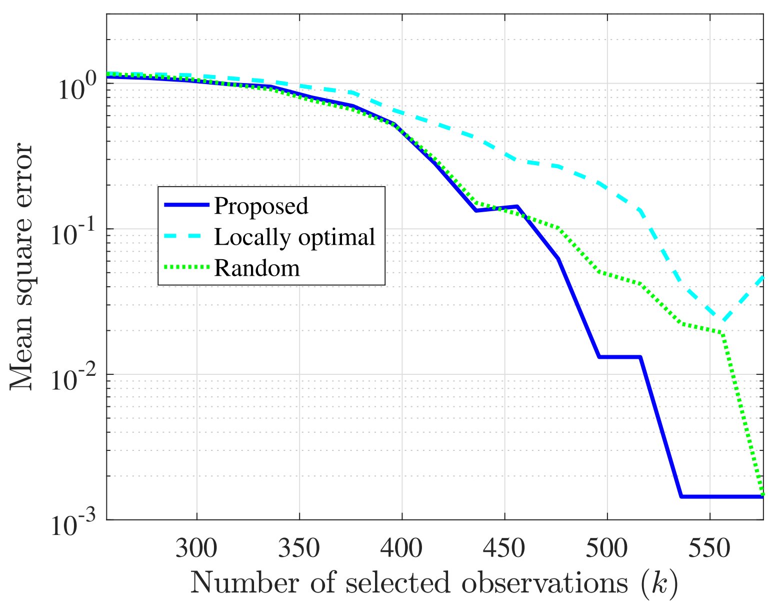
(a) Tightness of VTB



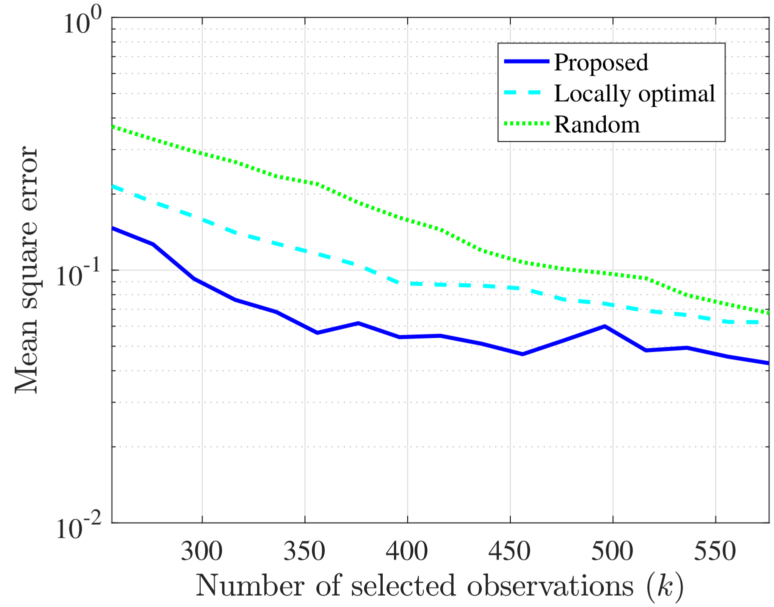
(b) Bound on α_{fA}

- Asymptotic tightness of VTB
- Tightness of weak submodularity bound in low SNR regime

- The phase retrieval problem with $n = 1280$ observations
  - Wirtinger flow [Candes'15] as the estimator



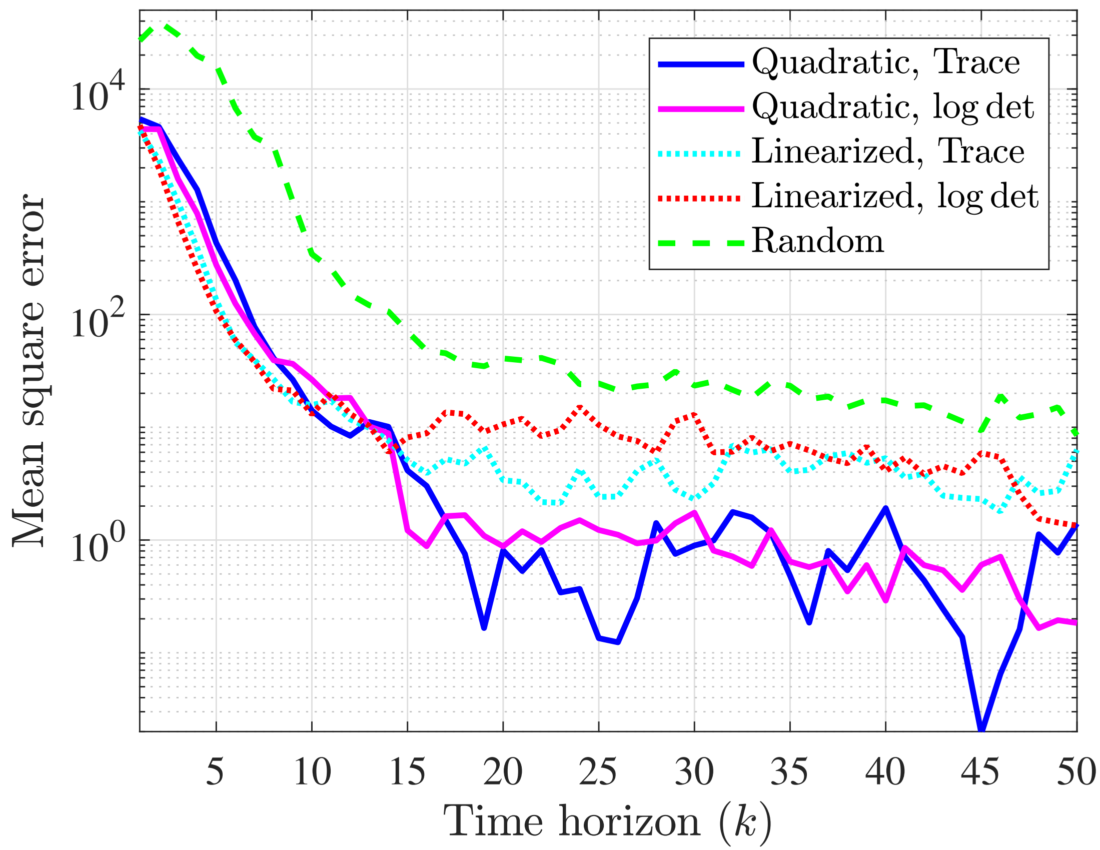
(a) Gaussian observations



(b) DFT observations

Results (3): Multi-Target Tracking

- The setting: 10 UAVs, 10 targets
- Selecting 10% of radar observations



Conclusions and Extensions

- Established weak submodularity of the MSE for linear models
- Exploited weak submodularity to establish performance guarantees of a randomized greedy algorithm for observation selection
- Utilized VTB as a surrogate to MSE for quadratic models and showed its weak submodularity
- Future work: Beyond quadratic models

Roadmap

Information Gathering

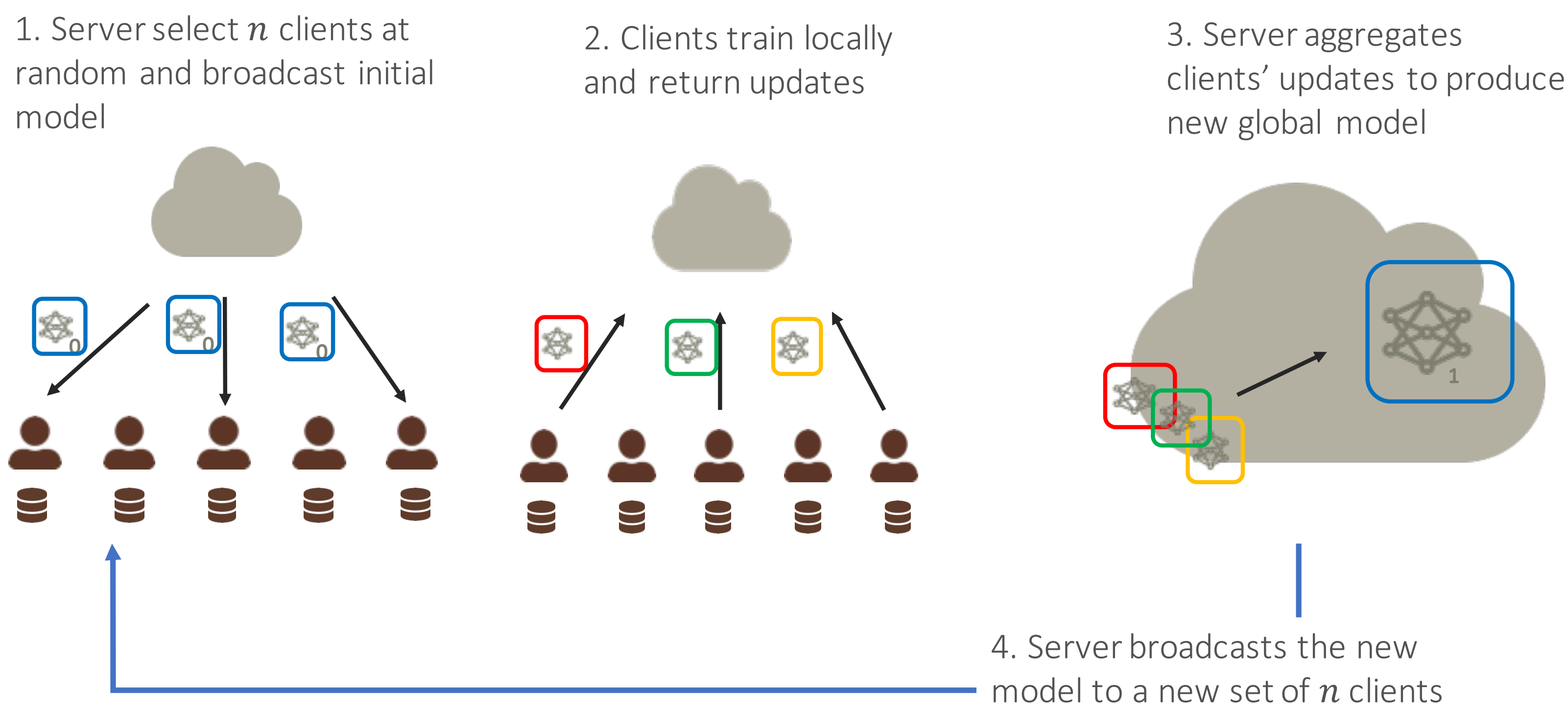
- Linear models
  - Weak submodularity of the MSE objective
  - Greedier than greedy: Randomized greedy selection
- Beyond linear models: Observation selection for quadratic models
  - Exploiting Van Trees' bound

Privacy preserving ML: Federated Learning

- Client selection as the remote estimation problem
- Exploring communication-accuracy tradeoff

Federated Learning

Private and efficient framework for learning a *global model* in settings where data is distributed across many clients.



One global round of FL

Federated Learning Cont'd: Challenges

- A large number of clients, potentially in millions
- Memory and bandwidth-intensive ML models; e.g., VGG-16 has 138M parameters, 500MB
- Highly dynamic systems: new users may join, new data may be generated by old users
  - may require a large number of global FL rounds

Prior Work on Improving Communications Efficiency

TEXAS The University of Texas at Austin

- Reducing individual users' communication
  - compression, sparsification, subsampling, low-rank approximation of weights' matrices [Konecny et al., 2016; Alistarh et al., 2017; Konecny et al. 2018; Horvath et al., 2019; Cho et al. 2020]
- Client subsampling [Hsieh et al. 2017; Chen et al., 2018; Singh et al., 2019; Cho et al., 2020]
  - introduces bias and/or increases variance of model estimation in each round, causing model variations and slowing down the convergence
  - relies on hyperparameters which have to be determined (e.g., k in "top-k" selection methods)

Proposed Approach to Client Selection

- A framework for selecting clients with the most informative updates, estimating aggregate update of the clients not selected
    - a computationally efficient FL algorithm that reduces communication
    - a reduced bias and variance gradient estimator
- Extensive experimental verification of the developed methodology in realistic federated learning settings

Preliminaries

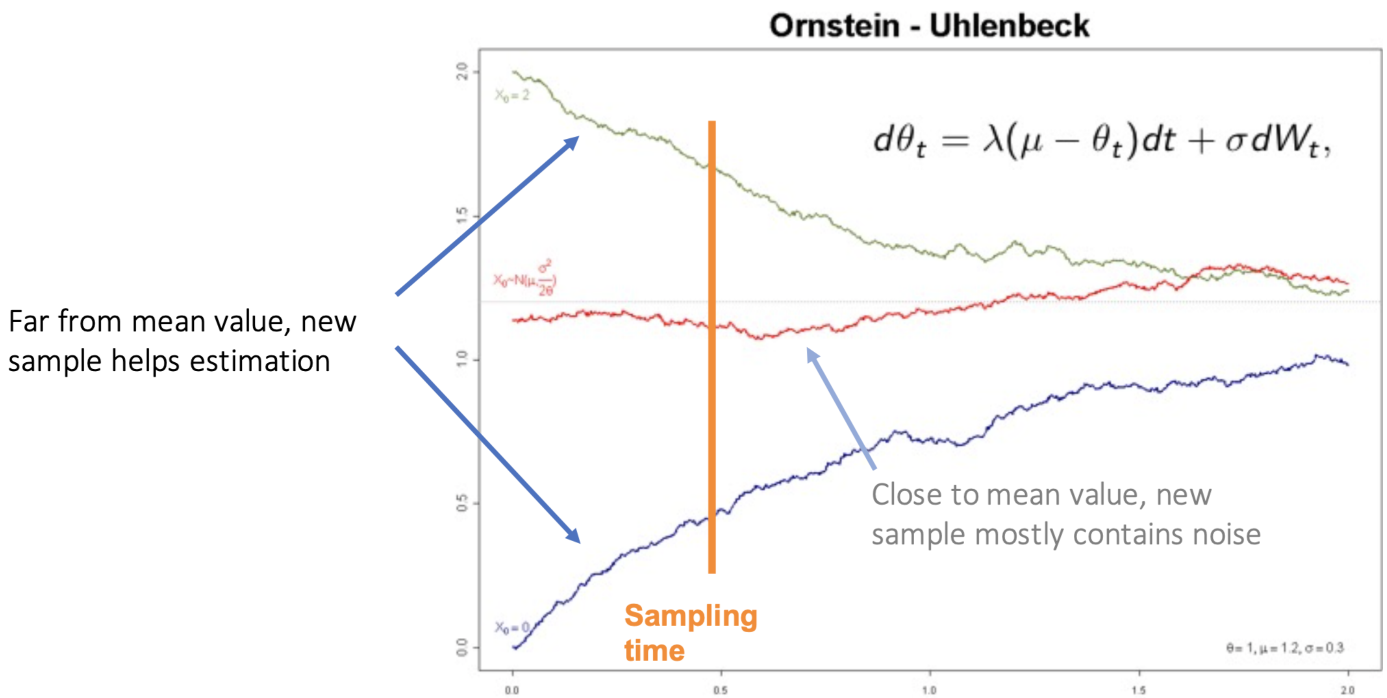
- SGD can be thought as a discretization of an OU process [Blanc et al., 2019; Wang et al., 2017; Li et al., 2018; Mandt et al., 2016]
- Ornstein–Uhlenbeck process: A stationary (Gauss-Markov) process θ_t which, over time, drifts towards its mean function
  - letting W_t denote the standard Wiener process,

$$d\theta_t = \lambda(\mu - \theta_t)dt + \sigma dW_t$$

- Basic idea: rely on the proximity of a sample path to the mean to assess informativeness of an update

Client Selection in FL

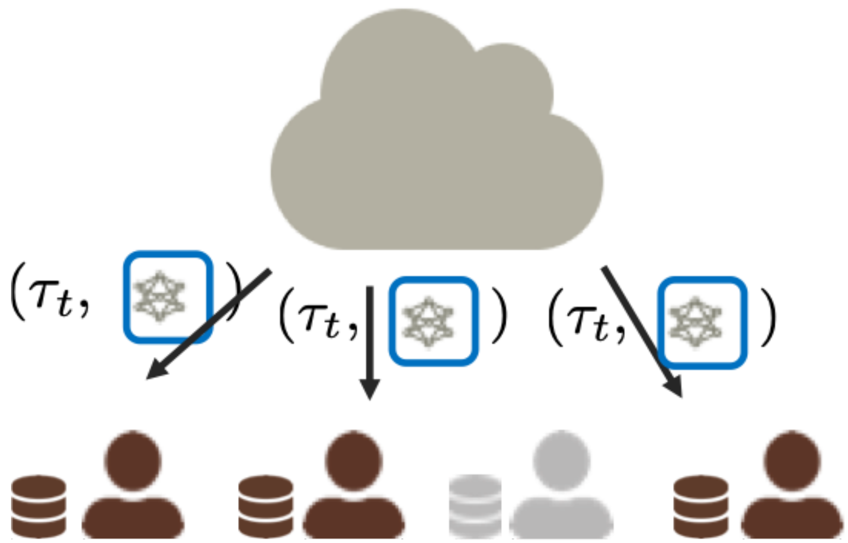
Revised model update strategy: collect only the updates with magnitude that exceed a threshold τ is the optimal sampling strategy



Estimate/predict the update of the clients that did not communicate

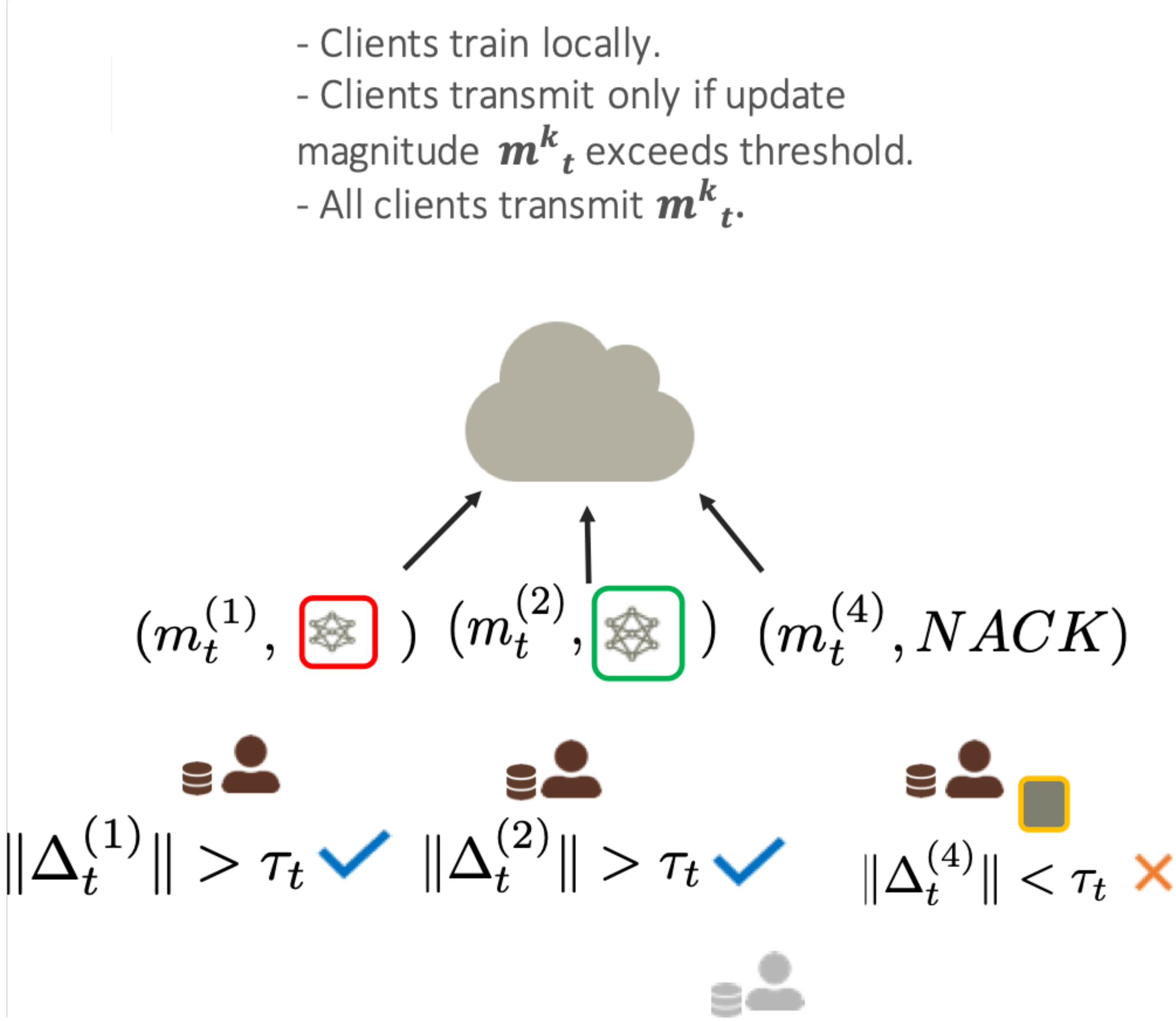
Step 1:

At time t, server select n clients at random, broadcast initial model and threshold τ_t

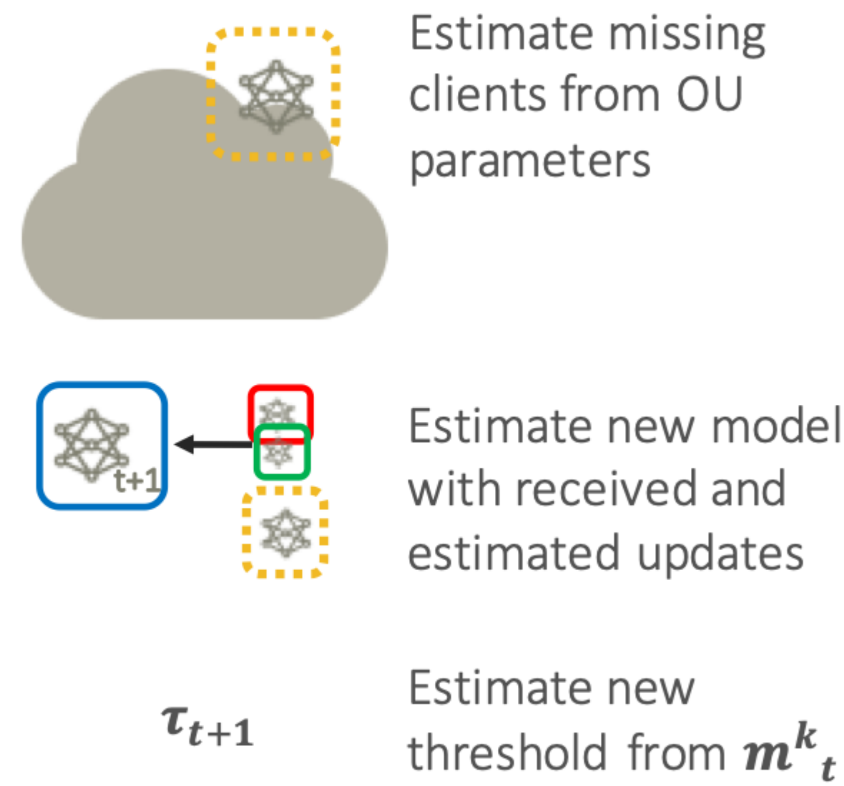


Communication-Efficient FL Cont'd

Step 2:



Step 3:

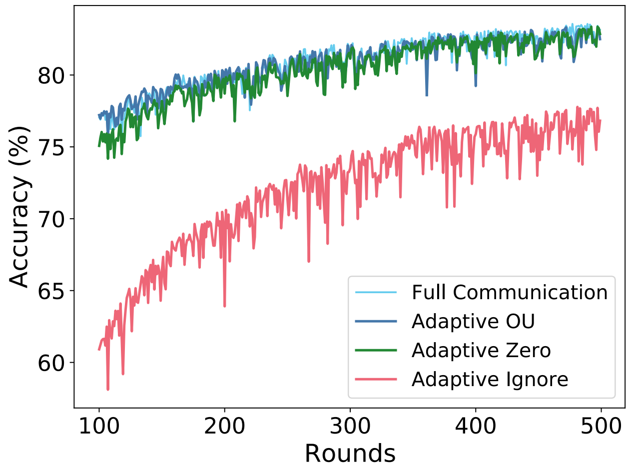

Estimate missing clients from OU parameters
t+1
Estimate new model with received and estimated updates
τ_{t+1}
Estimate new threshold from $\boldsymbol{m}^k{}_t$

Simulation and Analytical Results

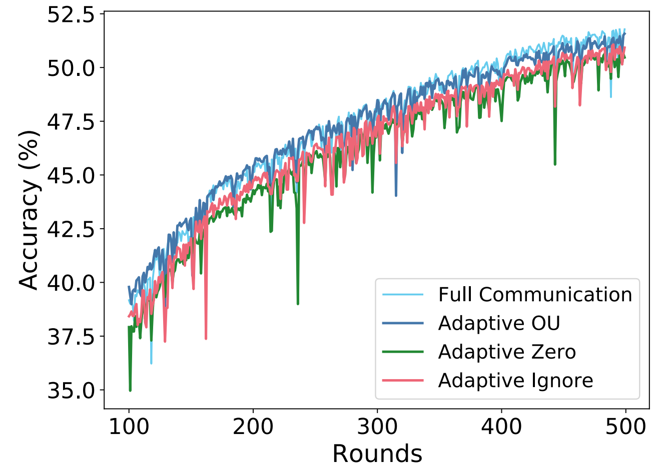
- We consider two client selection strategies:
  (a) adaptive threshold adjusted according to the gradient magnitude
  (b) random selection of a pre-fixed number of participating clients
- Compare the following model estimation strategies:
  - Our proposed OU process based estimation (`OU` strategy)
  - The strategy in [Li et al., 2019], where missing updates are replaced by the previous global model (`Ignore` strategy)
  - Dismiss missing, average transmitted updates (`Zero` strategy) [Hsieh et al., 2017]

Results: Accuracy on Different Data Sets

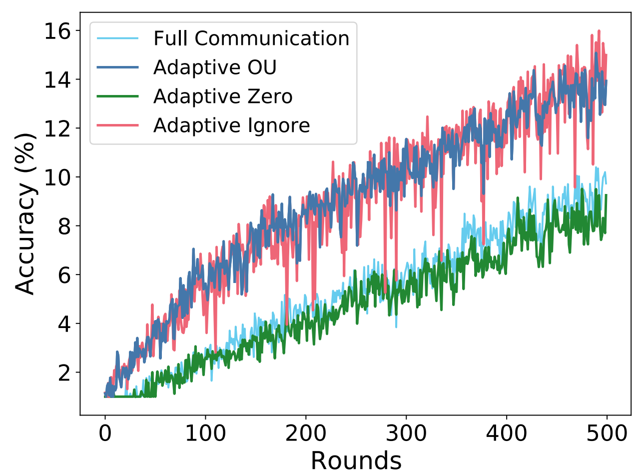
Test accuracy of EMNIST CNN.



Test accuracy of Shakespeare RNN.



Test accuracy of CIFAR100 Resnet.



Accuracy on Different Data Sets: Comments

- Dismissing clients' updates slows down the convergence in all experiments
  - in fact, `Zero` strategy yields biased model estimate
- `Ignore` strategy exhibits a significantly more unstable convergence than other techniques
  - both `OU` and `Zero` achieve lower variance
- OU estimation incorporates missing updates without rendering the convergence slow nor unstable
- Analytical result: Formally showed that the variance of `OU` strategy is lower than the variance of `Ignore` strategy

Accuracy vs. Communication Cost: Comments

- Thresholding-based sampling combined with `OU` estimation cuts communication **up to 50%** while achieving accuracy comparable to the baseline (i.e., to the full communication scheme)
- The `Zero` estimation strategy has communication savings similar to `OU` but with slower convergence rates and inferior final accuracy
- The `Ignore` strategy achieves the highest communication savings due to threshold inflation caused by ignoring clients, which then lead to even fewer clients in the following rounds
    - ultimately, the `Ignore` strategy is not capable of matching the accuracy of the `OU` method on Shakespeare and CIFAR100 datasets

Conclusions and Extensions

- Proposed a new way of selecting clients in a FL system
    - an efficient algorithm, guaranteed convergence, variance reduced w.r.t. alternative technique
- Future work: Client selection/sampling as a fairness mechanism

Acknowledgements

Joint work with Abolfazl Hashemi, Monica Ribero, Mahsa Ghasemi, and Ufuk Topcu