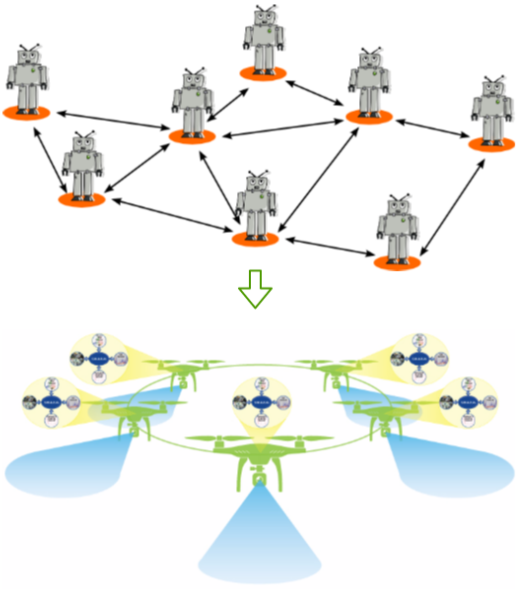
Finite-time Guarantees for Byzantine-Resilient Distributed State Estimation with Noisy Measurements

Lili Su (Northeastern, ECE)
and Shahin Shahrampour (Northeastern, IME)

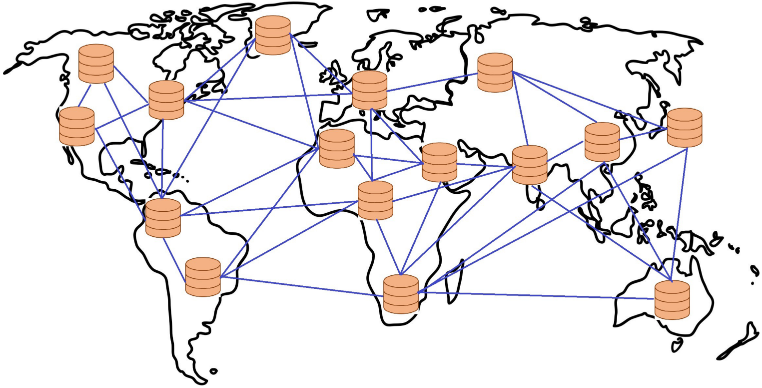
OP21

2021

Fully distributed systems: Multi-Agent Networks



Fully distributed systems: Multi-Agent Networks



A large scale machine learning system

Problem Formulation

State estimation: A static state $\theta^* \in \mathbb{R}^d$ that each of the non-Byzantine agent is interested in learning.

Constraints: an agent can collect *partial* and *noisy* measurements only.

- (Linear observation model) For each agent, its local measurement $y_i(t)$ at time t is generated as

$$y_i(t) := H_i\theta^* + w_i(t),$$

where

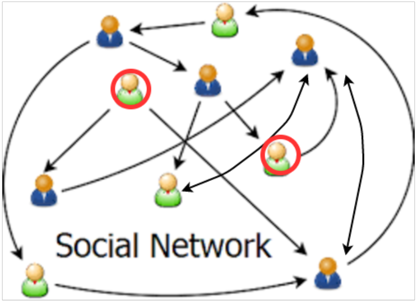
(1) $H_i \in \mathbb{R}^{n_i \times d}$, where $n_i \ll d$, is the local observation matrix

(2) $w_i(t)$'s are the observation noises that are zero mean and bounded. The observation noises across agents are independent.

Applications: IoT, machine learning, wireless networks, sensor networks, and robotic networks

Communication network

- a collection of n agents communicating with each other through a network $G(\mathcal{V}, \mathcal{E})$, where $\mathcal{V} = \{1, \cdots, n\}$ and $\mathcal{E}$ denote the set of agents and communication links, respectively.
- Among the n agents, an *unknown* subset of agents might be compromised and behave adversarially.



An example of an unreliable multi-agent network

Fault/Adversary Model - I

Byzantine Fault Model: There exists a system adversary that can choose up to b out of n agents to compromise and control. Let $\mathcal{A} \subseteq \mathcal{N}$ be the set of compromised agents, referred to as *Byzantine agents*.

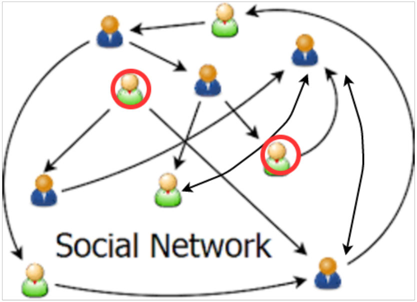
"The Byzantine Generals Problem", LAMPORT, SHOSTAK, and PEASE

- The adversary has complete knowledge of the network
  - the local program that each good agent is supposed to run;
  - the current status of the system;
  - running history of the system.

Fault/Adversary Model - II

The Byzantine agents are capable of

- colluding with each other;
- deviate from their pre-specified local programs to *arbitrarily* misrepresent information to the good agents;
- can mislead each of the good agents in a unique fashion, i.e., letting $m_{ij}(t) \in \mathbb{R}^d$ be the message sent from agent $i \in \mathcal{A}$ to agent $j \in \mathcal{V} \setminus \mathcal{A}$ at time t, it is possible that $m_{ij}(t) \neq m_{ij'}(t)$ for $j \neq j' \in \mathcal{V} \setminus \mathcal{A}$.



An example of an unreliable multi-agent network

Problem Formulation

State estimation: A static state $\theta^* \in \mathbb{R}^d$ that each of the non-Byzantine agent is interested in learning.

Constraints: an agent can collect *partial* and *noisy* measurements only.

- (Linear observation model) For each agent, its local measurement $y_i(t)$ at time t is generated as

$$y_i(t) := H_i \theta^* + w_i(t),$$

where

(1) $H_i \in \mathbb{R}^{n_i \times d}$, where $n_i \ll d$, is the local observation matrix

(2) $w_i(t)$'s are the observation noises that are zero mean and bounded. The observation noises across agents are independent.

*The local observation of a Byzantine agent is well-defined.

Reaching agreement in the presence of Byzantine faults is far from trivial.

Example: For binary consensus, even in complete graphs, no distributed algorithms can tolerate more than 1/3 of the agents to be Byzantine. [Lamport, Shostak, and Pease, 82]

Reaching agreement in the presence of Byzantine faults is far from trivial.

Example: For binary consensus, even in complete graphs, no distributed algorithms can tolerate more than 1/3 of the agents to be Byzantine. [Lamport, Shostak, and Pease, 82]

The reached agreement could be biased and the amount of bias is out of the control of the good agents.

Related Work

- **Adversary-resilient State Estimation**

  There is a rich line of work on the adversary-resilient state estimation problem wherein the existence of a fusion center is assumed. [Kosut-Jia-Thomas-Tong '11] [Kim and Poor '11] [Sou-Sandberg-Johansson '13] . . .
- **Adversary-resilient Distributed State Estimation**

  [Sundaram-Hadjicostis '11] [Chen-Kar-Moura '18 a, b,c,d,e] [Mitra-Sundaram '18] [Mitra-Ghawash-Sundaram-Abbas '21]. . .

Related Work

- **Adversary-resilient State Estimation**

  There is a rich line of work on the adversary-resilient state estimation problem wherein the existence of a fusion center is assumed. [Kosut-Jia-Thomas-Tong '11] [Kim and Poor '11] [Sou-Sandberg-Johansson '13] ...

- **Adversary-resilient Distributed State Estimation**

  [Sundaram-Hadjicostis '11] [Chen-Kar-Moura '18 a, b,c,d,e] [Mitra-Sundaram '18] [Mitra-Ghawash-Sundaram-Abbas '21]...

Our focus:

Noisy measurements, partially observable local matrix, and finite-time guarantees.

A Distributed Optimization Prospective: Asymptotic local function

For each agent $i \in \mathcal{V}$, define its *asymptotic* local function $f_i : \mathbb{R}^d \to \mathbb{R}$ as

$$f_i(x) := \frac{1}{2}\mathbb{E}\left[\|H_i x - y_i\|_2^2\right],$$

where the expectation of $f_i(x)$ is taken over the randomness of w_i.

1* f_i is well-defined for each agent regardless of whether it is a good agent or a Byzantine agent

2* Since the distribution of w_i is unknown to agent i, at any finite t, function f_i is not accessible to agent i.

A Distributed Optimization Prospective: Finite-time local function

The agent has access to the *finite-time* or *empirical* local function $f_{i,t}$ defined as

$$f_{i,t}(x) := \frac{1}{2t} \sum_{s=1}^{t} \|H_i x - y_i(s)\|_2^2 ,$$

whose gradient at x is

$$\nabla f_{i,t}(x) = \frac{1}{t} \sum_{s=1}^{t} H_i^\top (H_i x - y_i(s))$$

$$= H_i^\top H_i(x - \theta^*) - H_i^\top \frac{1}{t} \sum_{s=1}^{t} w_i(s).$$

A First Thought?

Question: Combine the local gradient descent with multi-dimensional Byzantine resilient consensus?

- The computation complexity of the relevant consensus component is prohibitively high
  - which typically relies on using Tverberg points
- assured convergence rate scales poorly in d

Proposed Algorithm

High-level idea:

Each good agent iteratively aggregates the received messages by, for each coordinate, discarding the largest b and the smallest b values, and averaging the remaining.

- *Local gradient descent:* Agent i first computes the noisy local gradient $\nabla f_{i,t}(x_i(t-1))$, and performs local gradient descent to obtain $z_i(t)$, i.e.,

$$z_i(t) = x_i(t-1) - \nabla f_{i,t}(x_i(t-1)).$$

Proposed Algorithm (continued)

- *Information exchange:* It exchanges $z_i(t)$ with other agents in its local neighborhood. Recall that $m_{ij}(t) \in \mathbb{R}^d$ is the message sent from agent i to agent j at time t. It relates to $z_i(t)$ as follows:

$$m_{ij}(t) = \begin{cases} z_i(t) & \text{if } i \in (\mathcal{V}/\mathcal{A}); \\ \star & \text{if } i \in \mathcal{A}, \end{cases}$$

  where $\star$ denotes an arbitrary value.
- *Robust aggregation:* For each coordinate $k = 1, \ldots, d$, the agent computes the trimmed mean to obtain $x_i(t)$.

Main results: Complete graphs

for ease of illustration: Applicable to computer networks and wireless networks with message forwarding

Lemma

For each iteration t, each good agent $i \in \mathcal{V}/\mathcal{A}$, and each k, there exist coefficients $\left(\beta_{ij}^k(t),\ j \in \mathcal{V}/\mathcal{A}\right)$ such that

- $x_i^k(t) = \sum_{j\in\mathcal{V}/\mathcal{A}} \beta_{ij}^k(t) \left\langle z_j(t), e_k \right\rangle$;
- $0 \le \beta_{ij}^k(t) \le \frac{1}{\phi - b}$ *for all* $j \in \mathcal{V}/\mathcal{A}$ *and* $\sum_{j\in\mathcal{V}/\mathcal{A}} \beta_{ij}^k(t) = 1$,

where $\phi = |\mathcal{V}/\mathcal{A}|$.

Main results: Complete graphs

for ease of illustration: Applicable to computer networks and wireless networks with message forwarding

Lemma

For each iteration t, each good agent $i \in \mathcal{V}/\mathcal{A}$, and each k, there exist coefficients $\left(\beta_{ij}^k(t),\ j \in \mathcal{V}/\mathcal{A}\right)$ such that

- $x_i^k(t) = \sum_{j\in\mathcal{V}/\mathcal{A}} \beta_{ij}^k(t) \langle z_j(t), e_k \rangle$;
- $0 \le \beta_{ij}^k(t) \le \frac{1}{\phi - b}$ *for all* $j \in \mathcal{V}/\mathcal{A}$ *and* $\sum_{j\in\mathcal{V}/\mathcal{A}} \beta_{ij}^k(t) = 1$,

where $\phi = |\mathcal{V}/\mathcal{A}|$.

Observations

- The update of x_i uses the information provided by the *good* agents only;
- each of the good agent has limited impact on x_i;

Remaining analysis is still non-trivial because
$\left(\beta_{ij}^k(t),\ j \in \mathcal{V}/\mathcal{A}\right) \neq \left(\beta_{ij}^{k'}(t),\ j \in \mathcal{V}/\mathcal{A}\right)$ for $k \neq k'$

Main results: Complete graphs

Assumption 1

For all $k = 1, \cdots, d$, we have that

$$\frac{1}{\phi - b} \sum_{j \in \mathcal{V}/\mathcal{A}} \left\| \left(\mathbf{I} - H_j^\top H_j \right) e_k \right\|_1 < 1.$$

- Note that it can well be the case that $\left\| \left(\mathbf{I} - H_j^\top H_j \right) e_k \right\|_1 \geq 1$ for some good agents.
- None of the agents are required to satisfy $\left\| \left(\mathbf{I} - H_j^\top H_j \right) e_k \right\|_1 < 1$ simultaneously for all $k = 1, \cdots, d$.

Main theorem

Let $\rho \triangleq \max_{k:1\le k\le d} \frac{\sum_{j\in\mathcal{V}/\mathcal{A}} \left\| \left(\mathbf{I}-H_j^\top H_j\right) e_k \right\|_1}{\phi - b}$, and
$C_0 \triangleq \max_{i\in\mathcal{V}/\mathcal{A}} \|H_i\|_2$.

Theorem

Suppose Assumption 1 holds and the graph is complete. Then

$$\max_{i\in\mathcal{V}/\mathcal{A}} \|x_i(t) - \theta^*\|_\infty \overset{a.s.}{\to} 0.$$

Moreover, with probability at least
$1 - \phi \exp\left(\frac{-\epsilon^2(1-\rho)^2 t}{8C^2}\right)$*, it holds that*

$$\max_{i\in\mathcal{V}/\mathcal{A}} \|x_i(t) - \theta^*\|_\infty \le \rho^t \max_{i\in\mathcal{V}/\mathcal{A}} \|x_i(0) - \theta^*\|_\infty$$
$$+ C_0 \left(\sum_{i\in\mathcal{V}/\mathcal{A}} \sqrt{\mathrm{trace}(\Sigma_j)}\right) \sum_{m=1}^{t-1} \frac{\rho^m}{\sqrt{t-m}} + \phi\epsilon.$$

Main results: Incomplete graphs

Theorem

Under the assumption that ensures Byzantine consensus with scalar inputs, if an assumption analogous to Assumption 1 holds, then any given $\delta \in (0, 1)$*, any* $\epsilon > 0$*, and*

$$t \geq \Omega\left(n^2/\epsilon^2\right)\left(\log \frac{1}{\delta} + \log n\right),$$

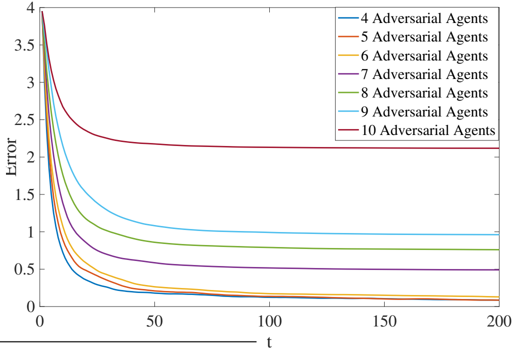
with probability at least $1 - \delta$*, it holds that*

$$\max_{i \in \mathcal{V}/\mathcal{A}} \|x_i(t) - \theta^*\|_\infty \leq \tilde{\rho}^t \max_{i \in \mathcal{V}/\mathcal{A}} \|x_i(0) - \theta^*\|_\infty + \tilde{C}_0 n \sum_{m=1}^{t-1} \frac{\tilde{\rho}^m}{\sqrt{t-m}} + \epsilon,$$

where $\tilde{\rho} \in (0, 1)$*.*

Numerical Examples: Energy Efficiency Data Set

- Regression dataset on UCI Machine Learning Repository[1]: In this dataset, the vector $\theta^* \in \mathbb{R}^8$, including eight features.
- We consider a network of $|\mathcal{V} \setminus \mathcal{A}| = 160$ agents. Each agent i observes only one feature corrupted by a Gaussian noise $\mathcal{N}(0, 0.25)$. Also, each agent i is connected to 40 agents $i-20, i-19, \ldots, i+19, i+20$.



[1] https://archive.ics.uci.edu/ml/datasets/Energy+efficiency