The University of Texas at Austin

DECENTRALIZED OPTIMIZATION ON TIME-VARYING DIRECTED GRAPHS UNDER COMMUNICATION CONSTRAINTS

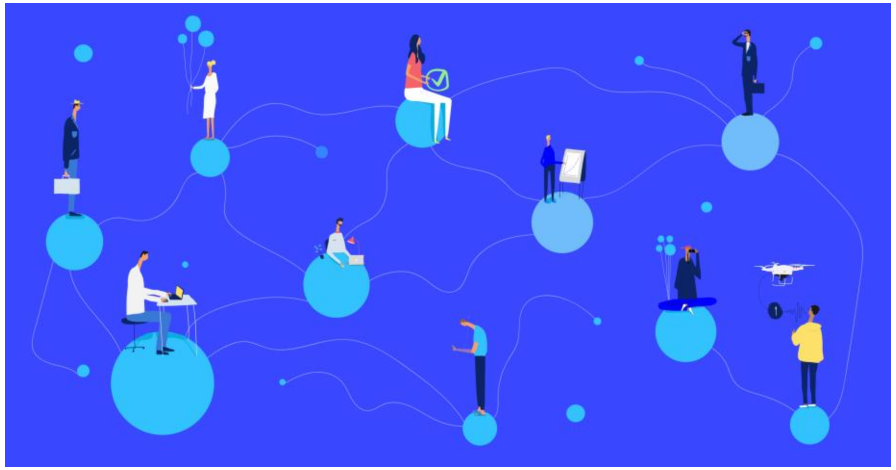
Yiyue Chen, Abolfazl Hashemi, and Haris Vikalo

2021 IEEE International Conference on Acoustics, Speech and Signal Processing

6-11 June 2021 • Toronto, Ontario, Canada

Problem motivation

- Decentralized optimization problems: all clients in the network to collaboratively learn the model via communication



Internet of things (IoT)

Communication network

- Potential issues on privacy, unreliable communication and resource constraint

Problem formulation

Decentralized problems over directed and time-varying networks:

$$\min_{\mathbf{x}\in\mathbb{R}^d}\left[f(\mathbf{x}) := \frac{1}{n}\sum_{i=1}^{n} f_i(\mathbf{x})\right]$$

- The agents collaborate to solve the problem by exchanging information over a network
- The network is modeled by a time-varying directed graph, $\mathcal{G}(t) = (|n|, \mathcal{E}(t))$
- The exchanged information is compressed before communication

Existing work

Algorithm	Directed network?	Time-varying network?	Compression?
Directed decentralized gradient descent [1]	Yes	No	No
Gradient-Push [2]	Yes	Yes	No
Quantized decentralized gradient descent [3]	Yes	No	Yes
This work	Yes	Yes	Yes

[1]: C. Xi, Q.Wu, and U. A. Khan, “On the distributed optimization over directed networks,” Neurocomputing, vol. 267, pp. 508–515, 2017.
[2]: A. Nedi´c and A. Olshevsky, “Distributed optimization over time-varying directed graphs,” IEEE Transactions on Auto-matic Control, vol. 60, no. 3, pp. 601–615, 2014.
[3]: H. Taheri, A. Mokhtari, H. Hassani, and R. Pedarsani, “Quantized decentralized stochastic learning over directed graphs,” in International Conference on Machine Learning (ICML), 2020.

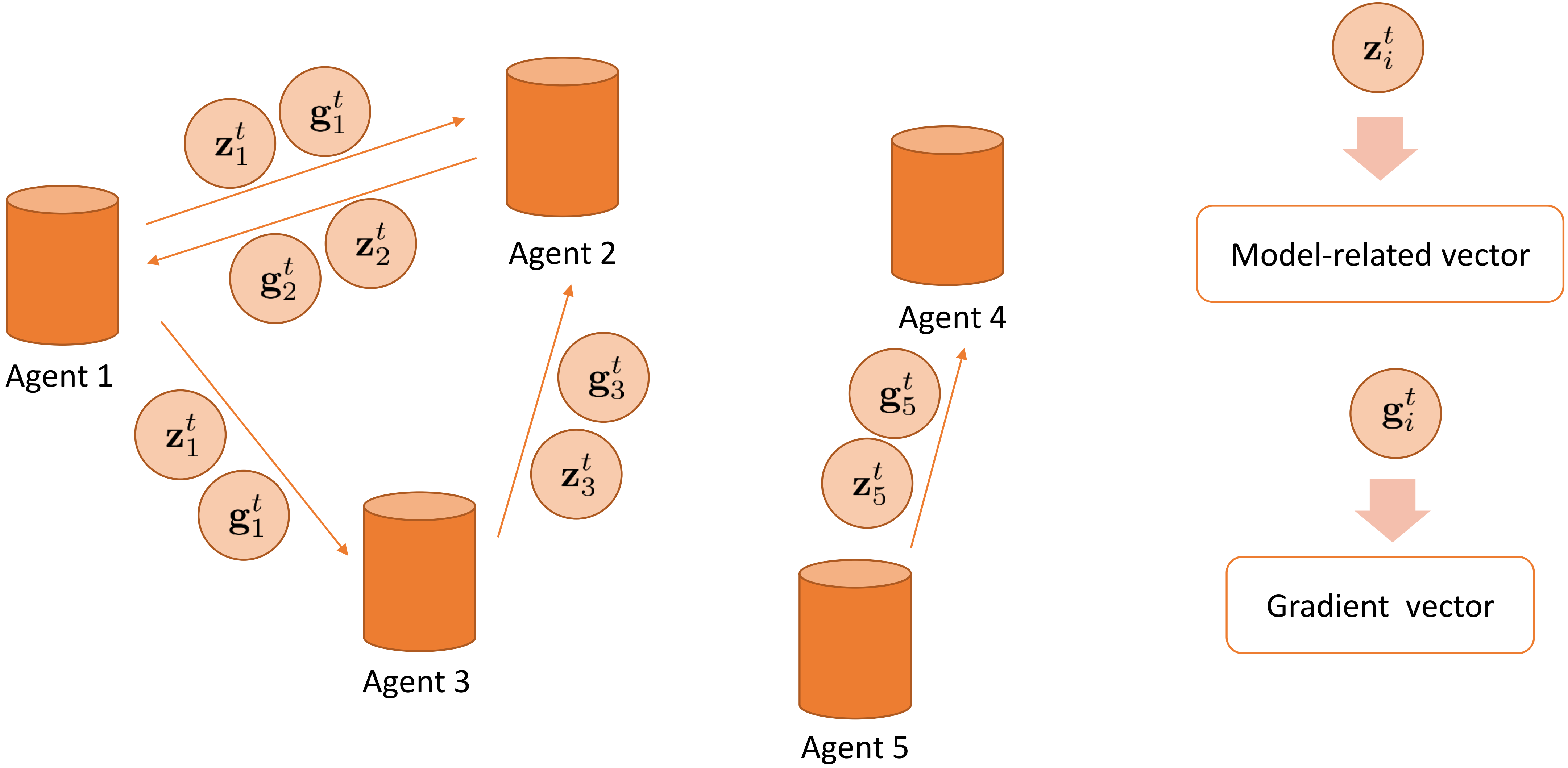
Challenges

- Communication imbalance in directed time-varying networks
- Bias induced by the compression operator

Compression operator – uniformly select k out of d entries from a d-dimensional message:

$$Q : \mathbb{R}^d \to \mathbb{R}^d$$

Algorithm



$\mathbf{z}_1^t$
$\mathbf{g}_1^t$
$\mathbf{z}_2^t$
$\mathbf{g}_2^t$
Agent 1
Agent 2
$\mathbf{z}_1^t$
$\mathbf{g}_1^t$
$\mathbf{g}_3^t$
$\mathbf{z}_3^t$
Agent 3
Agent 4
$\mathbf{g}_5^t$
$\mathbf{z}_5^t$
Agent 5
$\mathbf{z}_i^t$
Model-related vector
$\mathbf{g}_i^t$
Gradient vector

Algorithm

$$\mathbf{z}_i^t = \begin{cases} \mathbf{x}_i^t, & i \in \{1, ..., n\} \\ \mathbf{y}_{i-n}^t, & i \in \{n+1, ..., 2n\} \end{cases}$$

$\mathbf{x}_i^t$

Vector of parameters

$\mathbf{y}_i^t$

Auxiliary vector to record the network imbalance bias

Elementwise update:

$$z_{im}^{t+1} = \sum_{j=1}^{2n} [\bar{M}_m^t]_{ij} [Q(\mathbf{z}_j^t)]_m + \mathbf{1}_{\{t \bmod \mathcal{B}=\mathcal{B}-1\}} \epsilon [F]_{ij} z_{jm}^{\mathcal{B}\lfloor t/\mathcal{B} \rfloor} - \mathbf{1}_{\{t \bmod \mathcal{B}=\mathcal{B}-1\}} \alpha_{\lfloor t/\mathcal{B} \rfloor} g_{im}^{\mathcal{B}\lfloor t/\mathcal{B} \rfloor}$$

Reweighting mixing matrix to cancel out compression bias

Stored message facilitating consensus convergence in jointly-connected networks

Display local gradient descent in jointly-connected network

Assumptions

The mixing matrices, stepsizes, and the local objectives satisfy:

(i) The product of mixing matrices, $M_m((k+1)\mathcal{B}-1 : k\mathcal{B})$, has a non-zero spectral gap.

Network joint connectivity

(ii) For a fixed $\epsilon \in (0,1)$, the set of all possible mixing matrices $\{\bar{M}_m^t\}$ is a finite set.

Mixing matrix weight policy

(iii) The sequence of stepsizes, $\{\alpha_t\}$, is non-negative and satisfies $\sum_{t=0}^{\infty} \alpha_t = \infty$, $\sum_{t=0}^{\infty} \alpha_t^2 < \infty$.

Step size

(iv) Each entry of the gradient is bounded ($|g_{im}^t| < D$).

Bounded gradient

Convergence result

Suppose the previous assumptions hold. Let $\mathbf{x}^*$ be the unique optimal solution and $f^* = f(\mathbf{x}^*)$.

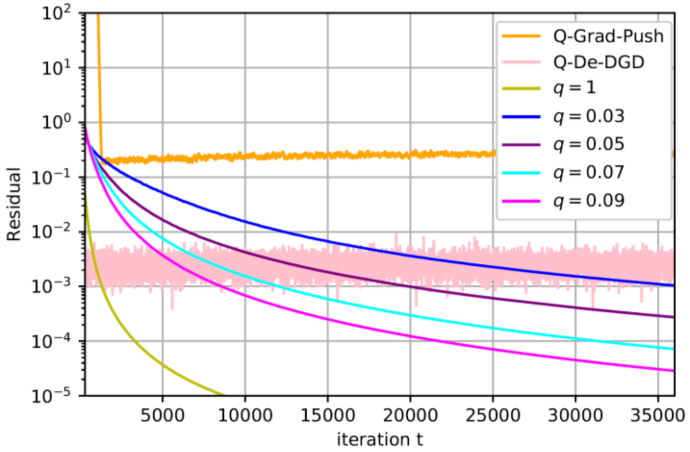
$$2\sum_{k=0}^{\infty}\alpha_k(f(\bar{\mathbf{z}}^{k\mathcal{B}}) - f^*) \leq n\|\bar{\mathbf{z}}^0 - \mathbf{x}^*\| + nD'^2\sum_{k=0}^{\infty}\alpha_k^2 + \frac{4D'}{n}\sum_{i=1}^{n}\sum_{k=0}^{\infty}\alpha_k\|\mathbf{z}_i^{k\mathcal{B}} - \bar{\mathbf{z}}^{k\mathcal{B}}\|,$$

where $D' = \sqrt{d}D$ and $\bar{\mathbf{z}}^t = \frac{1}{n}\sum_{i=1}^{n}\mathbf{x}_i^t + \frac{1}{n}\sum_{i=1}^{n}\mathbf{y}_i^t$.

For the stepsize $\alpha_t = \mathcal{O}(1/\sqrt{t})$, the algorithm attains the convergence rate $\mathcal{O}(\frac{\ln T}{\sqrt{T}})$.

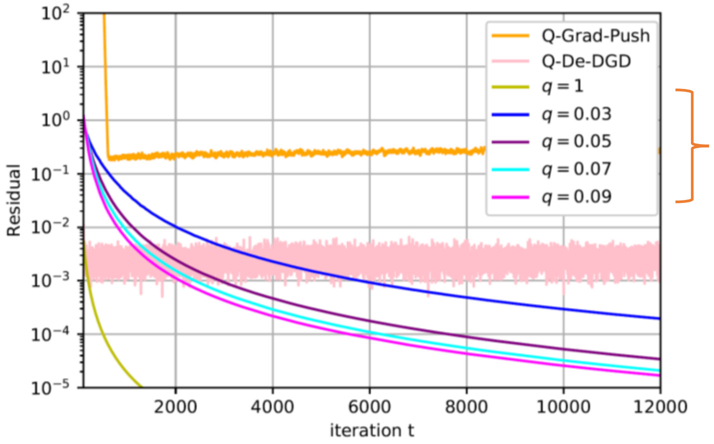
Simulation – linear regression

Decentralized linear regression with 10 agents



Algorithm with different sparsification levels

Joint connectivity, B=3



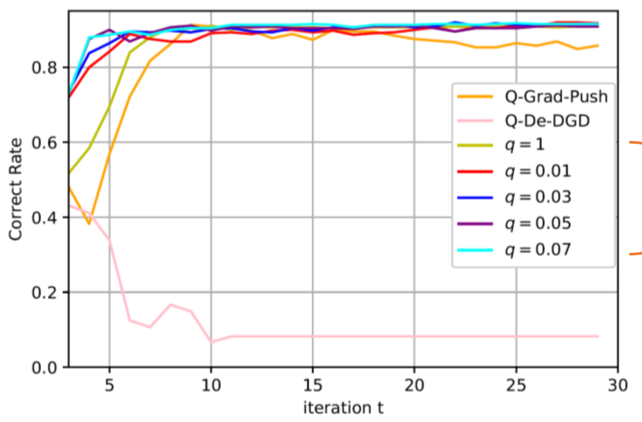
Algorithm with different sparsification levels

Strong connectivity, B=1

Algorithm converges faster with stronger connectivity and smaller sparsification level.

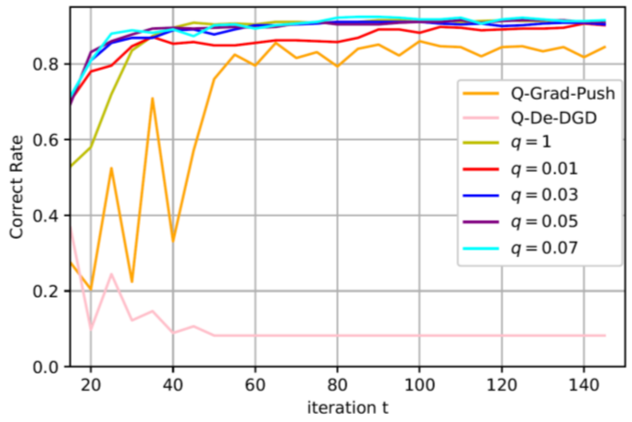
Simulation – logistic regression

Decentralized logistic regression with 10 agents



Algorithm with different sparsification levels

Strong connectivity, B=1



Algorithm with different sparsification levels

Joint connectivity, B=5

Algorithm reaches higher accuracy faster with stronger connectivity and smaller sparsification level.

Conclusion and future work

- Proposed a communication-sparsifying algorithm for decentralized convex optimization over directed time-varying graphs.
- Proved the convergence rate of the proposed algorithm.
- Justified the performance of the proposed algorithm.

Future work

- Apply stochastic variance-reduced gradient method to the decentralized algorithm to reach faster convergence rate.
- Extend the algorithm to non-convex optimization problems.

Thank you!

DECENTRALIZED OPTIMIZATION ON TIME-VARYING DIRECTED GRAPHS UNDER COMMUNICATION CONSTRAINTS