ADOM: Accelerated Decentralized Optimization Method for Time-Varying Networks

Dmitry Kovalev, **Egor Shulgin**, Peter Richtárik, Alexander Rogozin, Alexander Gasnikov

ICML
International Conference On Machine Learning

July, 2021

Egor Shulgin, shulgin-egor.github.io

Co-authors

Dmitry Kovalev
PhD student
(KAUST)

Peter Richtárik
Professor
(KAUST)

Alexander Rogozin
PhD student
(MIPT, HSE)

Alexander Gasnikov
Professor
(MIPT, HSE)

King Abdullah University
of Science and Technology

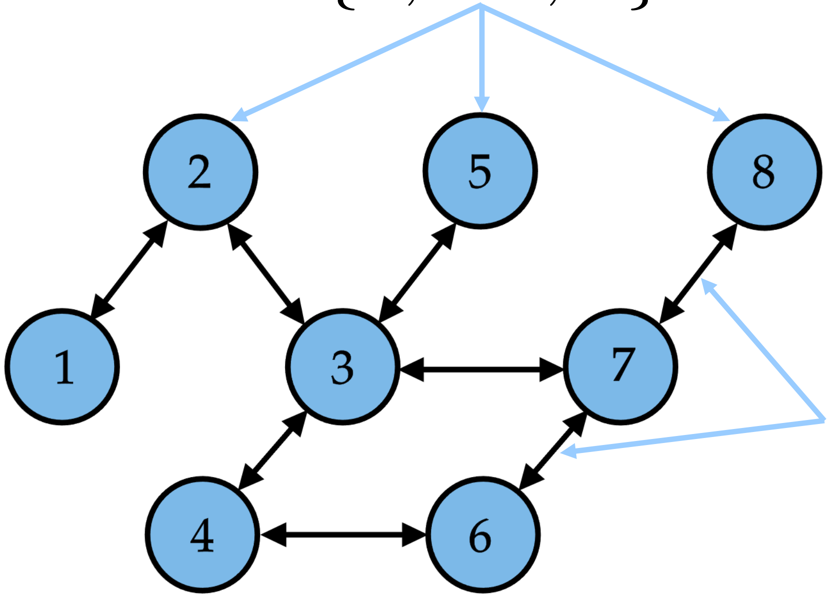
MIPT
MOSCOW INSTITUTE OF PHYSICS AND TECHNOLOGY

HIGHER SCHOOL OF ECONOMICS
NATIONAL RESEARCH UNIVERSITY

Decentralized Setting

Consider an undirected network $\mathcal{G} = (\mathcal{V}, \mathcal{E})$

$\mathcal{V} = \{1, \ldots, n\}$ – set of computing nodes

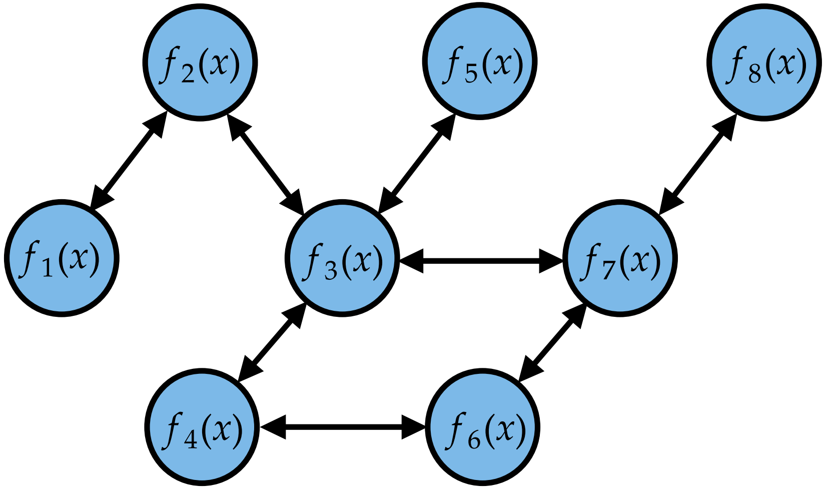


$\mathcal{E} \subset \mathcal{V} \times \mathcal{V}$ – set of edges

Decentralized Optimization

$$\min_{x \in \mathbb{R}^d} \sum_{i \in \mathcal{V}} f_i(x)$$

$f_i(x) : \mathbb{R}^d \to \mathbb{R}$ is stored on node i only

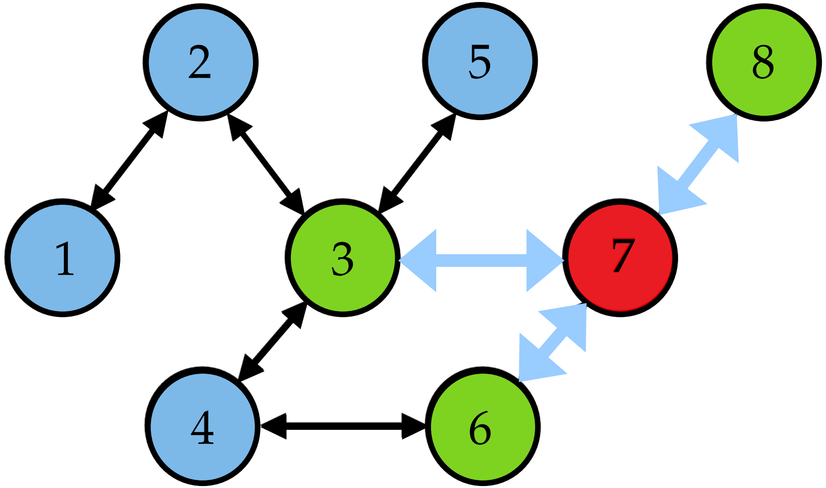


Each $f_i(x)$ is:

- L-smooth
- μ-strongly convex

Decentralized Communication

Is done only across edges $e \in \mathcal{E}$



Decentralized Communication via Gossip

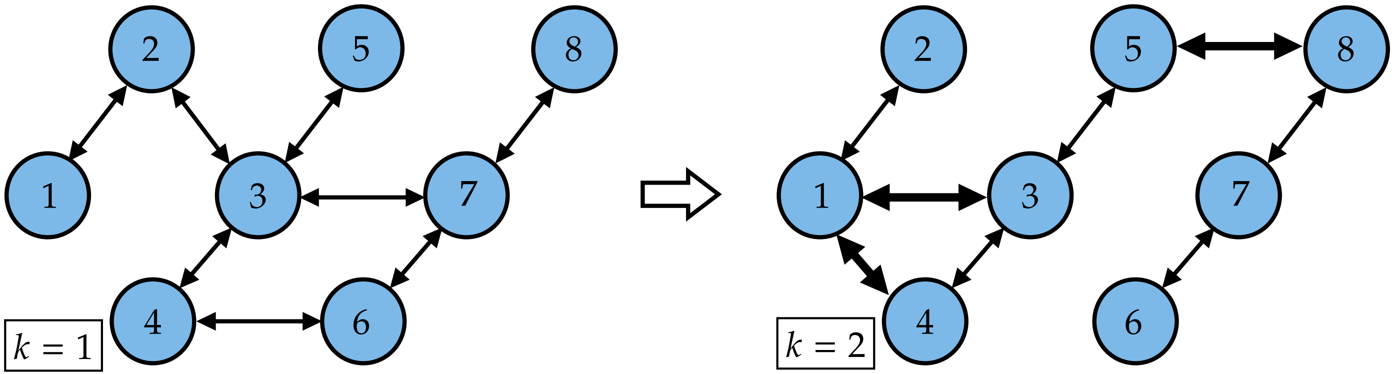
Gossip matrix $\mathbf{W} \in \mathbb{R}^{n \times n}$:

- $\mathbf{W}$ is symmetric positive semidefinite
- $\mathbf{W}_{ij} \neq 0$ iff $i = j$ or $(i, j) \in \mathcal{E}$
- $\ker \mathbf{W} = \{x \in \mathbb{R}^n : x_1 = \ldots = x_n\}$

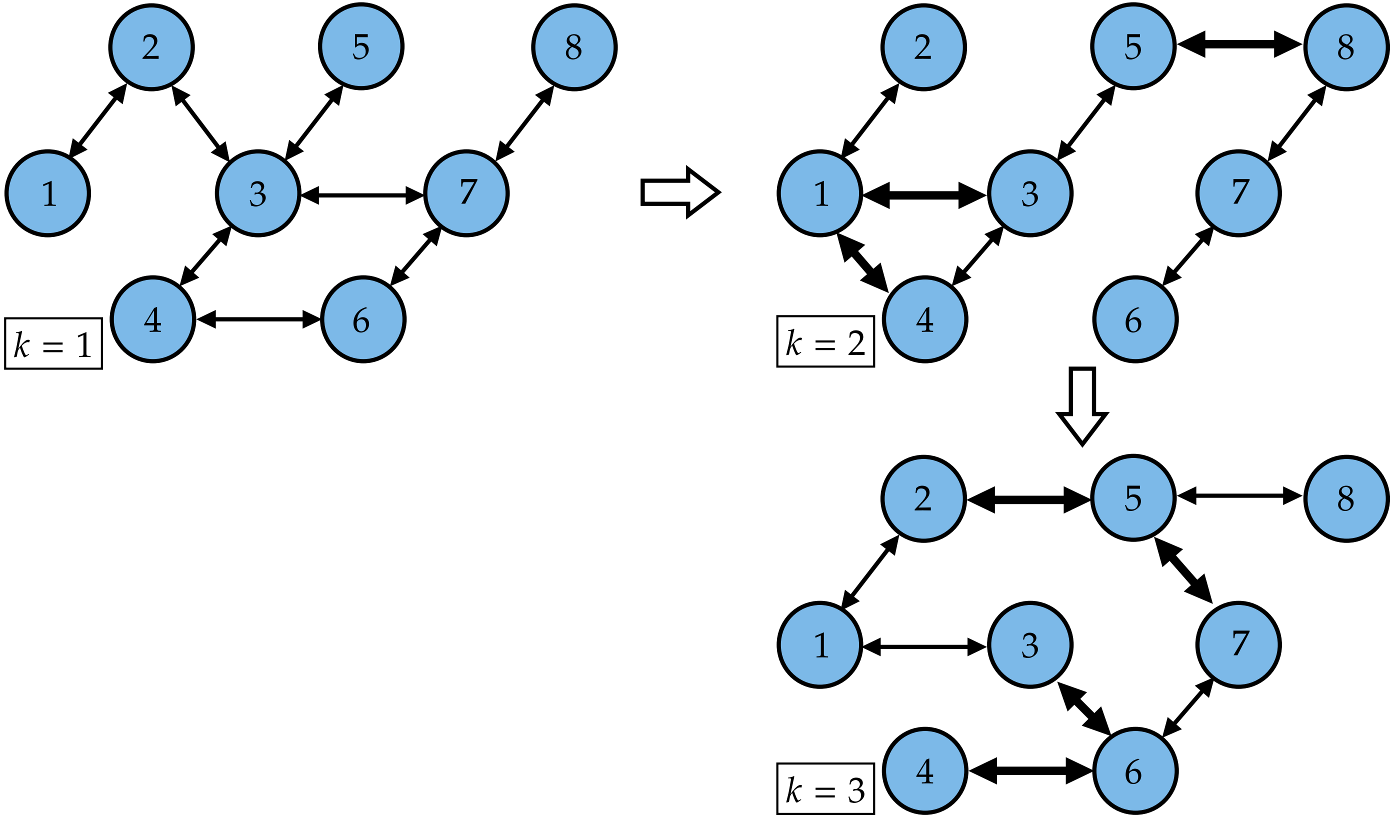
Communication can be represented as multiplication of vector by $\mathbf{W}$

$$[\mathbf{W}x]_i \in \operatorname{span}\left(\{x_j : j \text{ is neighbor of } i\}\right)$$

Time-Varying Graphs

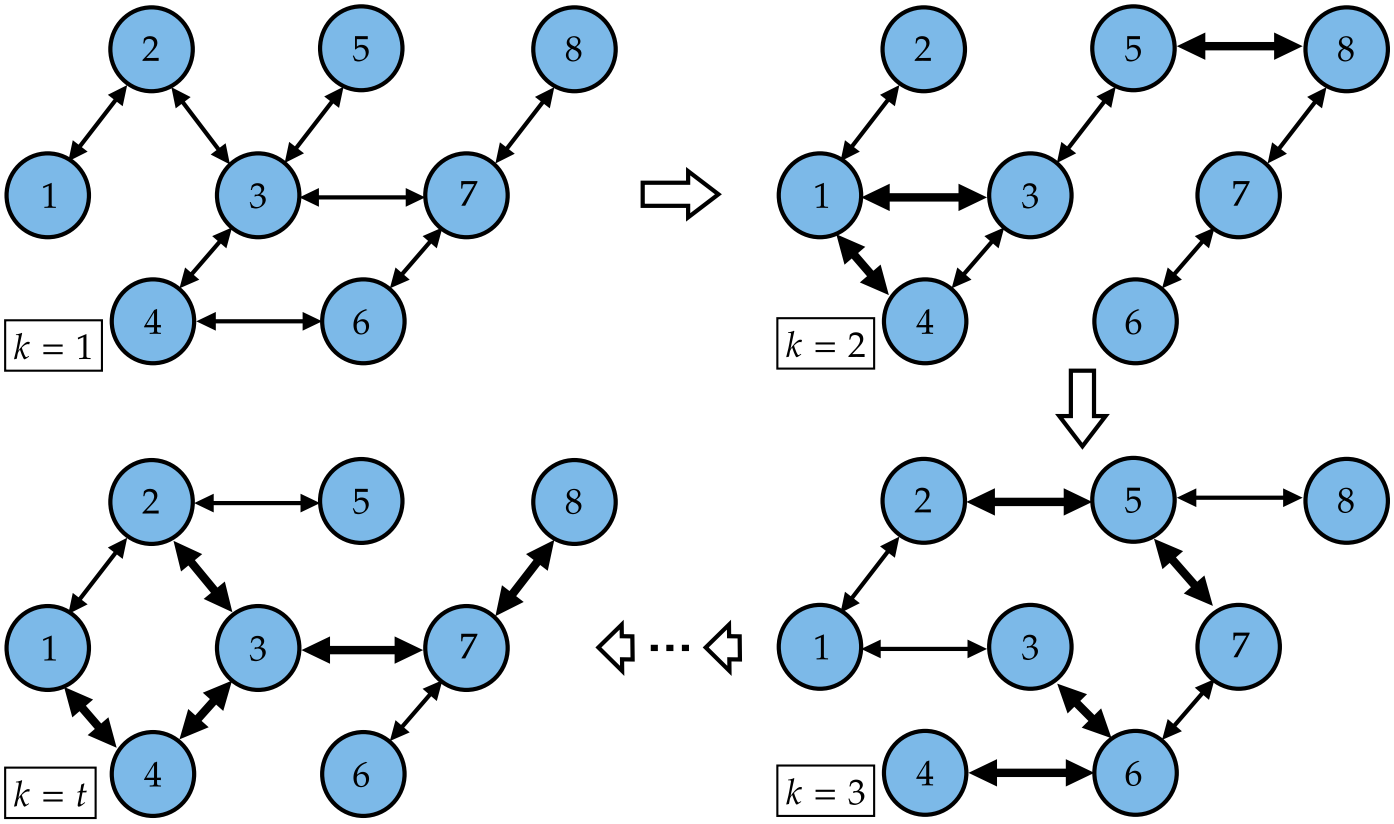


Time-Varying Graphs



Time-Varying Graphs

Time-varying network is modeled as a sequence of graphs $\{\mathcal{G}_k\}_{k=1}^{\infty}$ with gossip matrices $\mathbf{W}(k)$



Problem Reformulation

Original problem

$$\min_{x \in \mathbb{R}^d} \sum_{i \in \mathcal{V}} f_i(x)$$

Lifted problem (Primal)

$$\min_{\substack{x=(x_1,\ldots,x_n)\in(\mathbb{R}^d)^{\mathcal{V}} \\ x_1=\cdots=x_n}} F(x)$$

$$F(x) := \sum_{i \in \mathcal{V}} f_i(x_i)$$

Dual formulation:

$$\min_{\substack{z=(z_1,\ldots,z_n)\in(\mathbb{R}^d)^{\mathcal{V}} \\ \sum_{i=1}^n z_i=0}} F^*(z)$$

Projected Nesterov Gradient Descent

$$z_g^k = \tau z^k + (1-\tau) z_f^k$$
$$z^{k+1} = z^k + \eta\alpha(z_g^k - z^k) - \eta \mathbf{P}\nabla F^*(z_g^k)$$
$$z_f^{k+1} = z_g^k - \theta \mathbf{P}\nabla F^*(z_g^k)$$

Converges with rate: $\mathcal{O}\left(\kappa^{1/2} \log 1/\epsilon\right) \quad \kappa = L/\mu$

Can not be implemented in decentralized fashion

Key Idea

Decentralized communication can be seen as the application of a contractive compression operator

$$\|\sigma \mathbf{W}(k) z - z\|^2 \leq \left(1 - \sigma \lambda_{\min}^{+}\right) \|z\|^2$$

$$\lambda_{\min}^{+} = \inf_{k} \lambda_{\min}^{+}(\hat{\mathbf{W}}(k))$$

Error-Feeback Mechanism

Contractive compressor: $\|\mathcal{C}(z) - z\|^2 \leq (1-\delta)\|z\|^2$

Gradient Descent with Contractive (biased) compression operators may not converge.

$$v^k = m^k - \gamma g^k \quad \text{// vector to compress}$$

$$z^{k+1} = z^k + \mathcal{C}(v^k) \quad \text{// gradient step}$$

$$m^{k+1} = v^k - \mathcal{C}(v^k) \quad \text{// update error}$$

Comparison to Existing Work

ADOM achieves the new state-of-the-art rate for decentralized optimization over time-varying graphs.

Algorithm	Communication complexity
DIGing *Nedic et al. (2017)*	$\mathcal{O}\left(n^{1/2}\chi^2\kappa^{3/2}\log\frac{1}{\epsilon}\right)$
PANDA *Maros & Jaldén (2018)*	$\mathcal{O}\left(\chi^2\kappa^{3/2}\log\frac{1}{\epsilon}\right)$
Acc-DNGD *Qu & Li (2019)*	$\mathcal{O}\left(\chi^{3/2}\kappa^{5/7}\log\frac{1}{\epsilon}\right)$
APM *Li et al. (2018)*	$\mathcal{O}\left(\chi\kappa^{1/2}\log^2\frac{1}{\epsilon}\right)$
Mudag *Ye et al. (2020)*	$\mathcal{O}\left(\chi\kappa^{1/2}\log(\kappa)\log\frac{1}{\epsilon}\right)$
ADOM **Our Work**	$\mathcal{O}\left(\chi\kappa^{1/2}\log\frac{1}{\epsilon}\right)$

Our method combines error-feedback with Nesterov acceleration

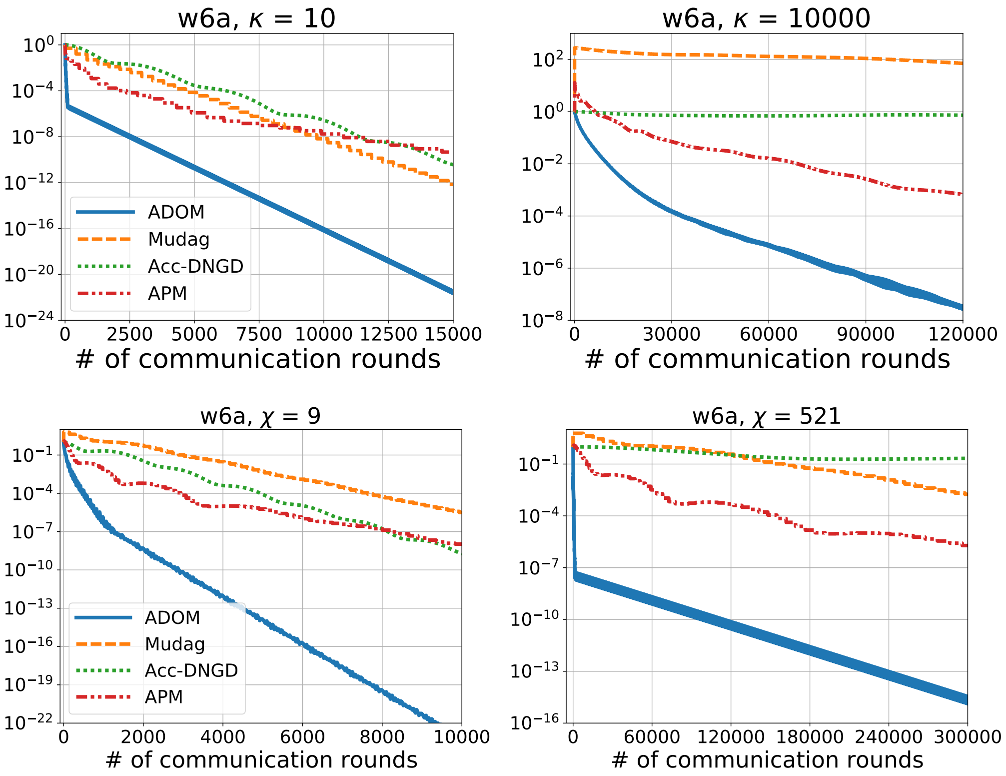
$$\kappa = L/\mu$$

$$\chi = \sup_k \frac{\lambda_{\max}(\hat{\mathbf{W}}(k))}{\lambda_{\min}^{+}(\hat{\mathbf{W}}(k))}$$

Obtained communication complexity is optimal. See preprint arXiv:2106.04469 by Kovalev et al.

Experimental Results

ADOM converges linearly and <u>outperforms</u> all known algorithms for every set of parameters.



Regularized Logistic Regression Problem

$$f_i(x) = \frac{1}{m}\sum_{j=1}^{m} \log\left(1 + \exp\left(-b_{ij}a_{ij}^\top x\right)\right) + \frac{r}{2}\|x\|^2$$

with LibSVM dataset *w6a* $(n = 17188, d = 300)$

Time-varying network simulated as a sequence of geometric random graphs with Laplacians $\mathbf{W}(k)$.

More results (including real networks!) in the paper