Lower Bounds and Optimal Algorithms for Smooth and Strongly Convex Decentralized Optimization over Time-varying Networks

Dmitry Kovalev

Co-authors

Figure 1: Peter Richtarik (KAUST)

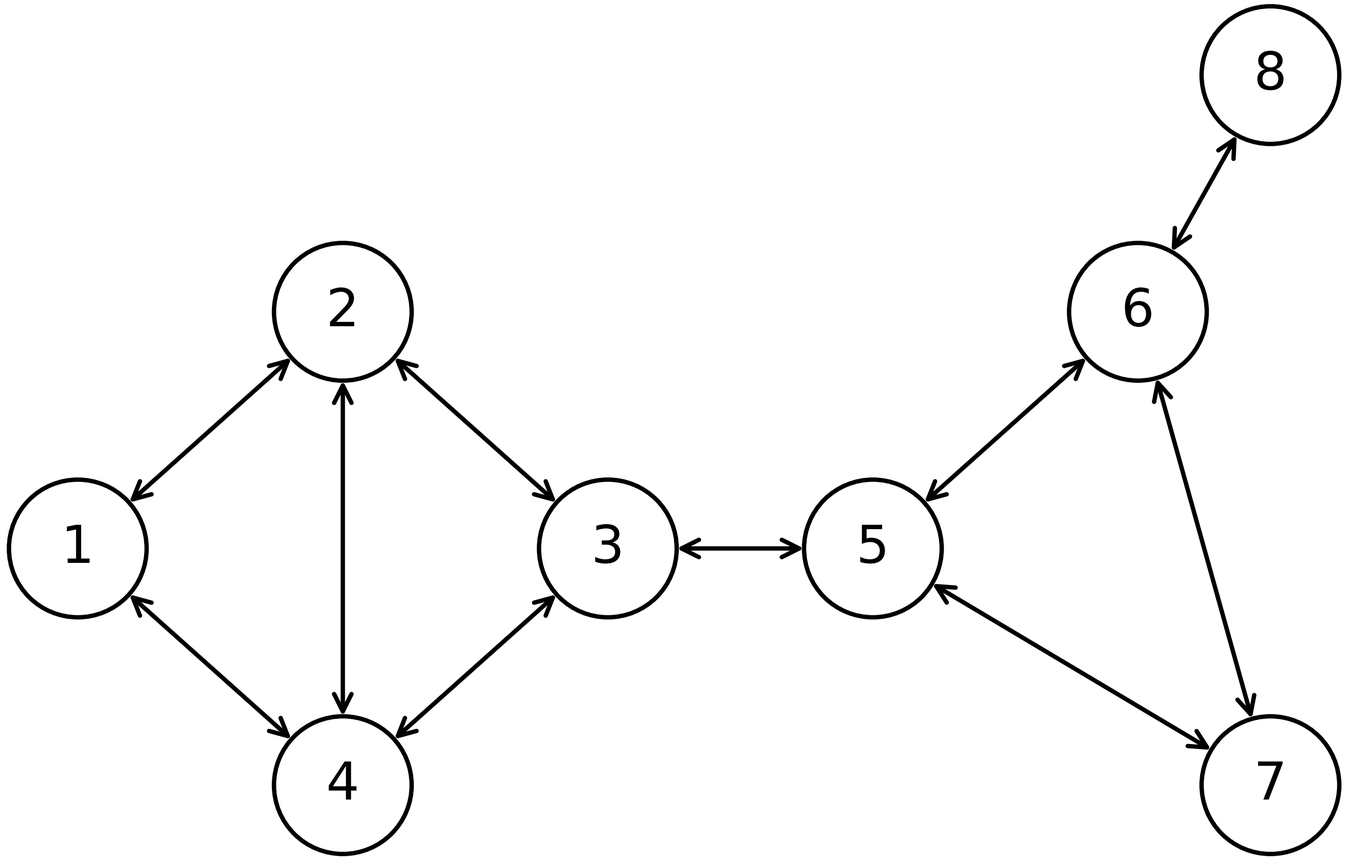
Figure 2: Elnur Gasanov (KAUST)

Figure 3: Alexander Gasnikov (MIPT, HSE)

Introduction

Decentralized Communication Network

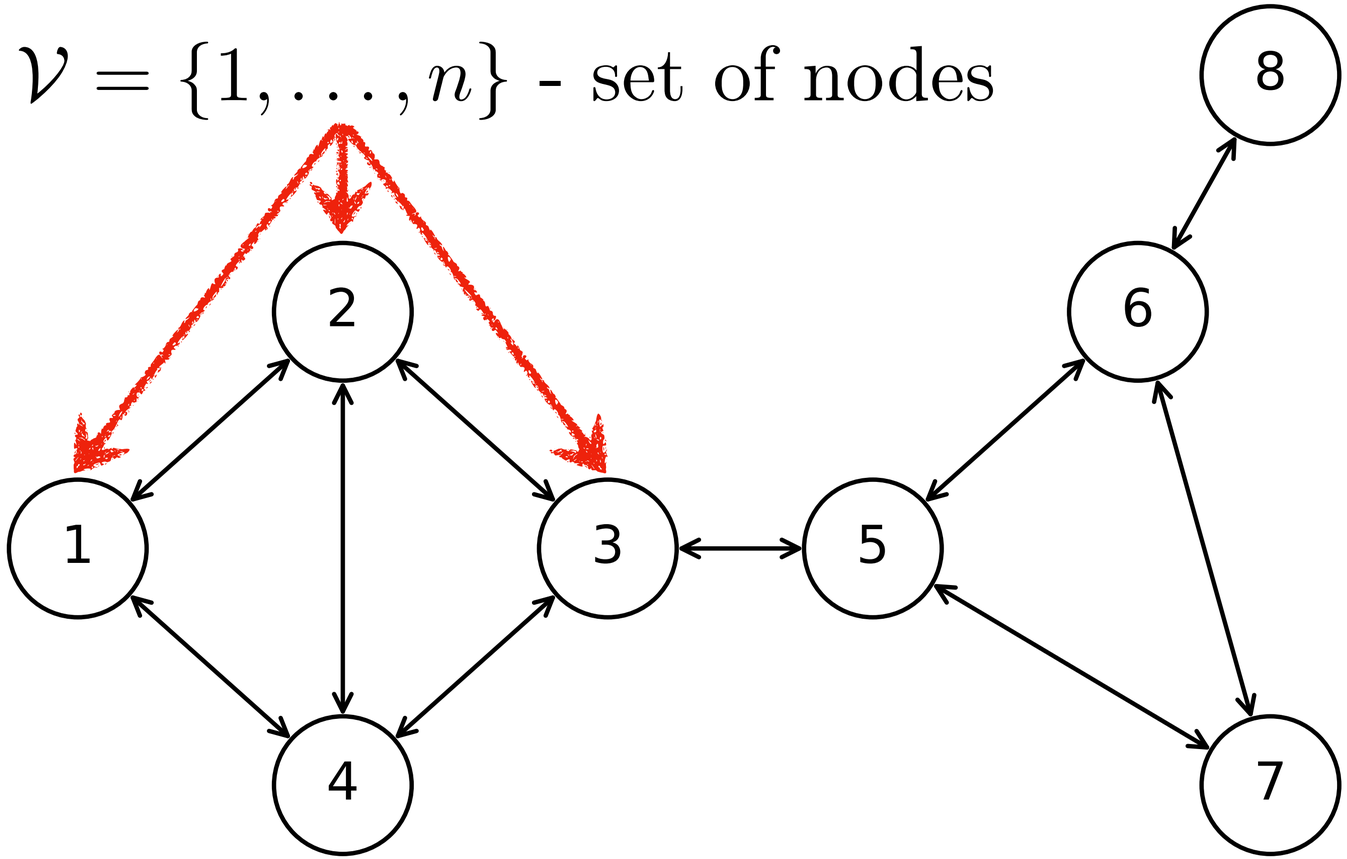
Consider a network $\mathcal{G} = (\mathcal{V}, \mathcal{E})$



Decentralized Communication Network

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

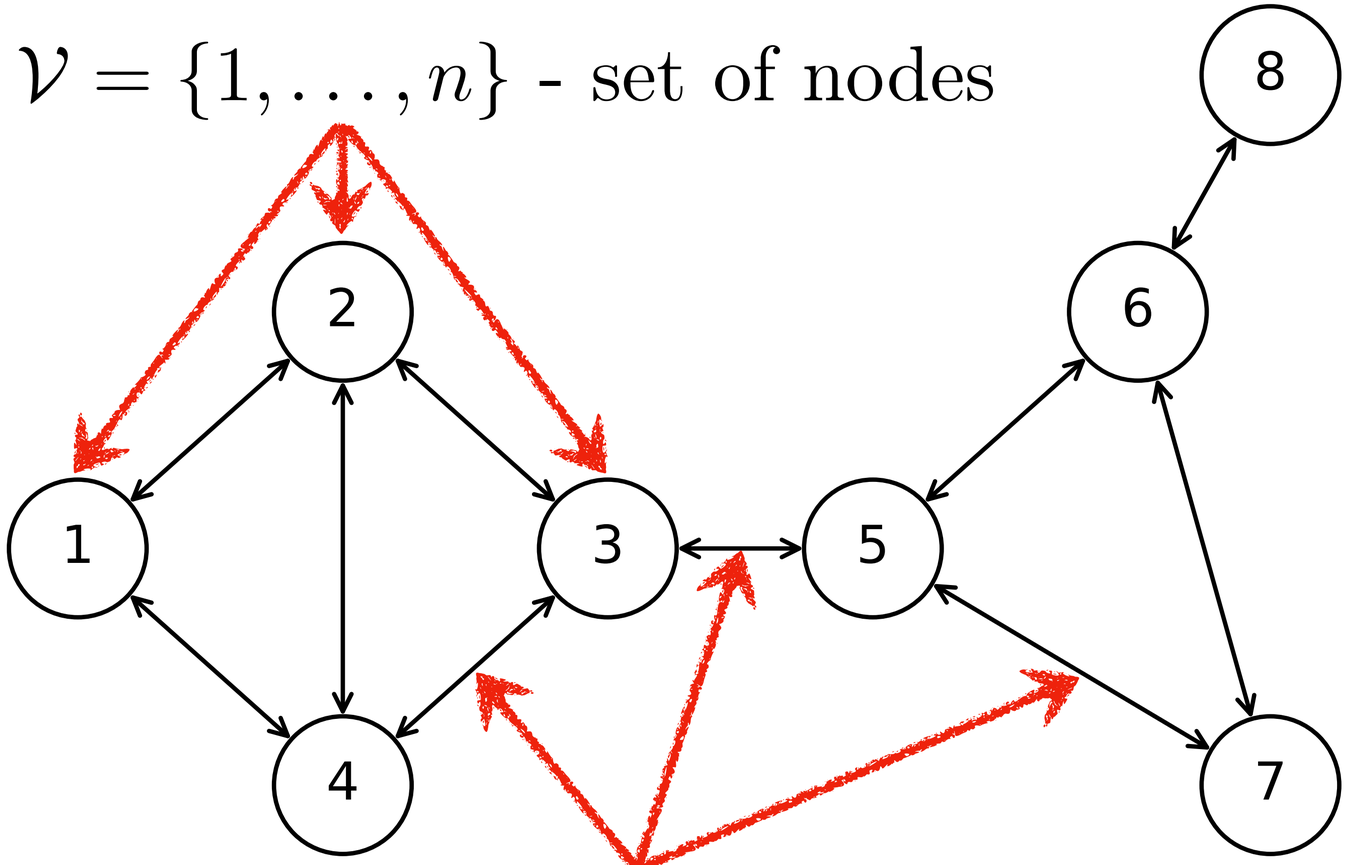
$\mathcal{V} = \{1, \ldots, n\}$ - set of nodes



Decentralized Communication Network

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

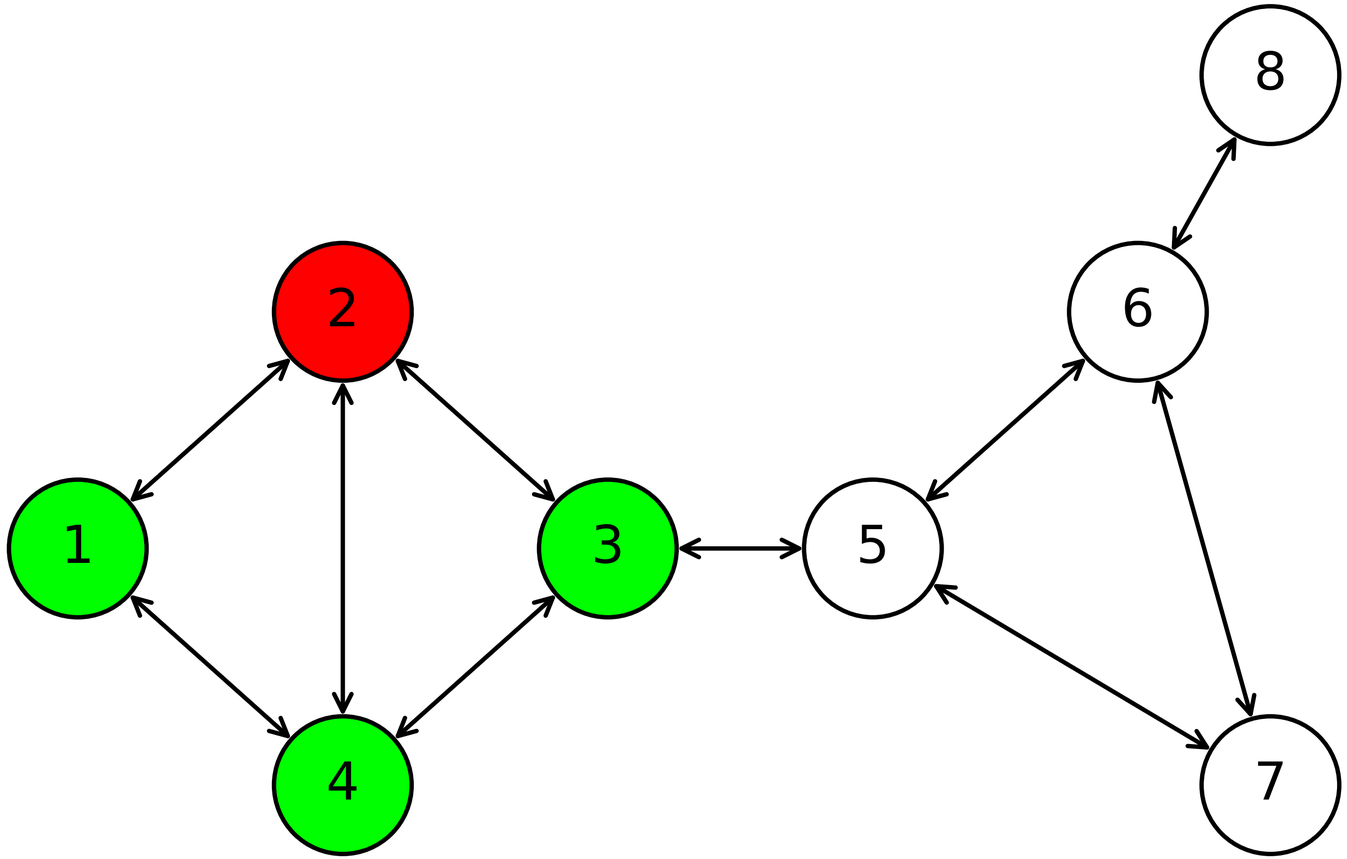
$\mathcal{V} = \{1, \ldots, n\}$ - set of nodes



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

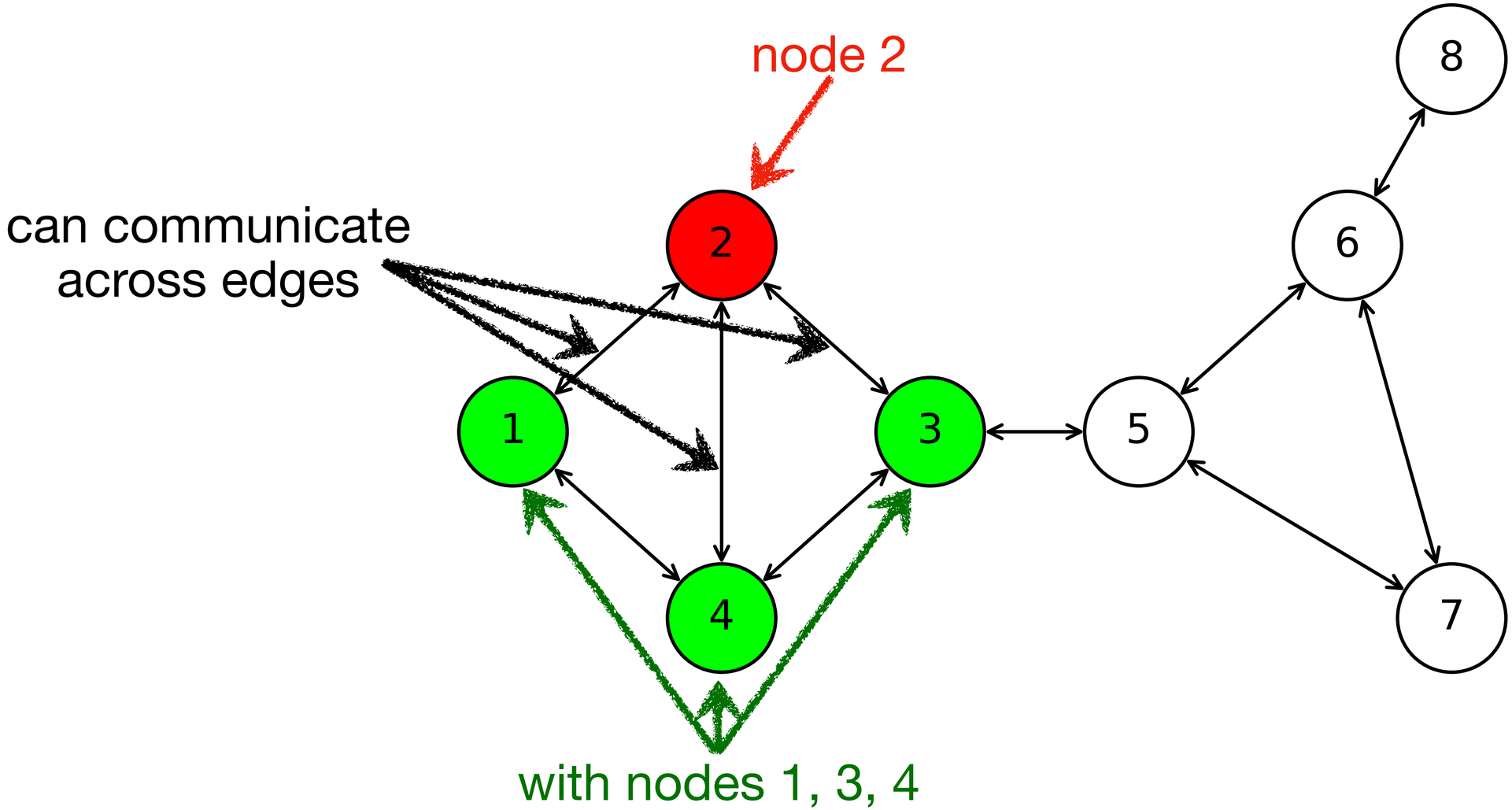
Decentralized Communication

Is done only across edges of the network



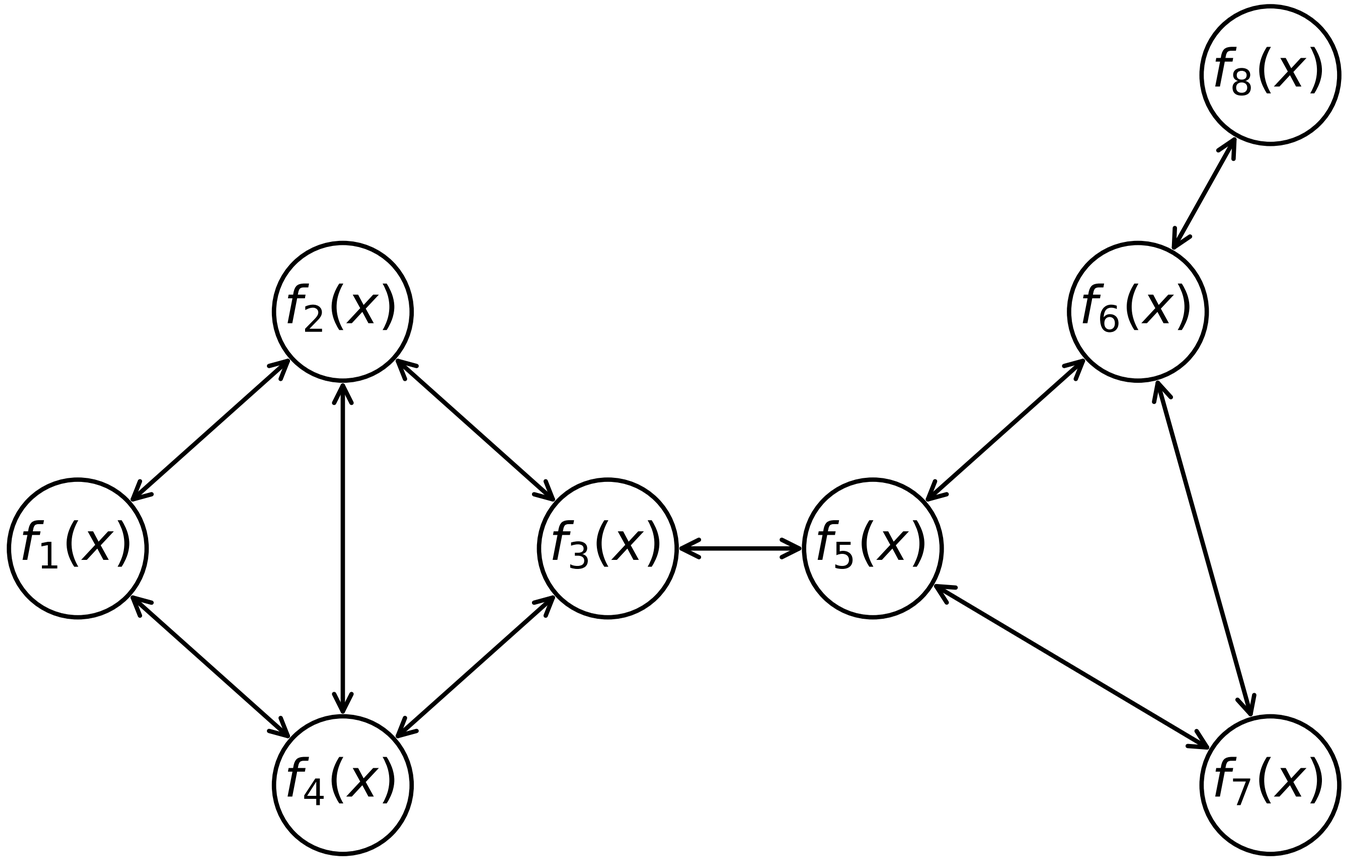
Decentralized Communication

Is done only across edges of the network



Decentralized Optimization Problem

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

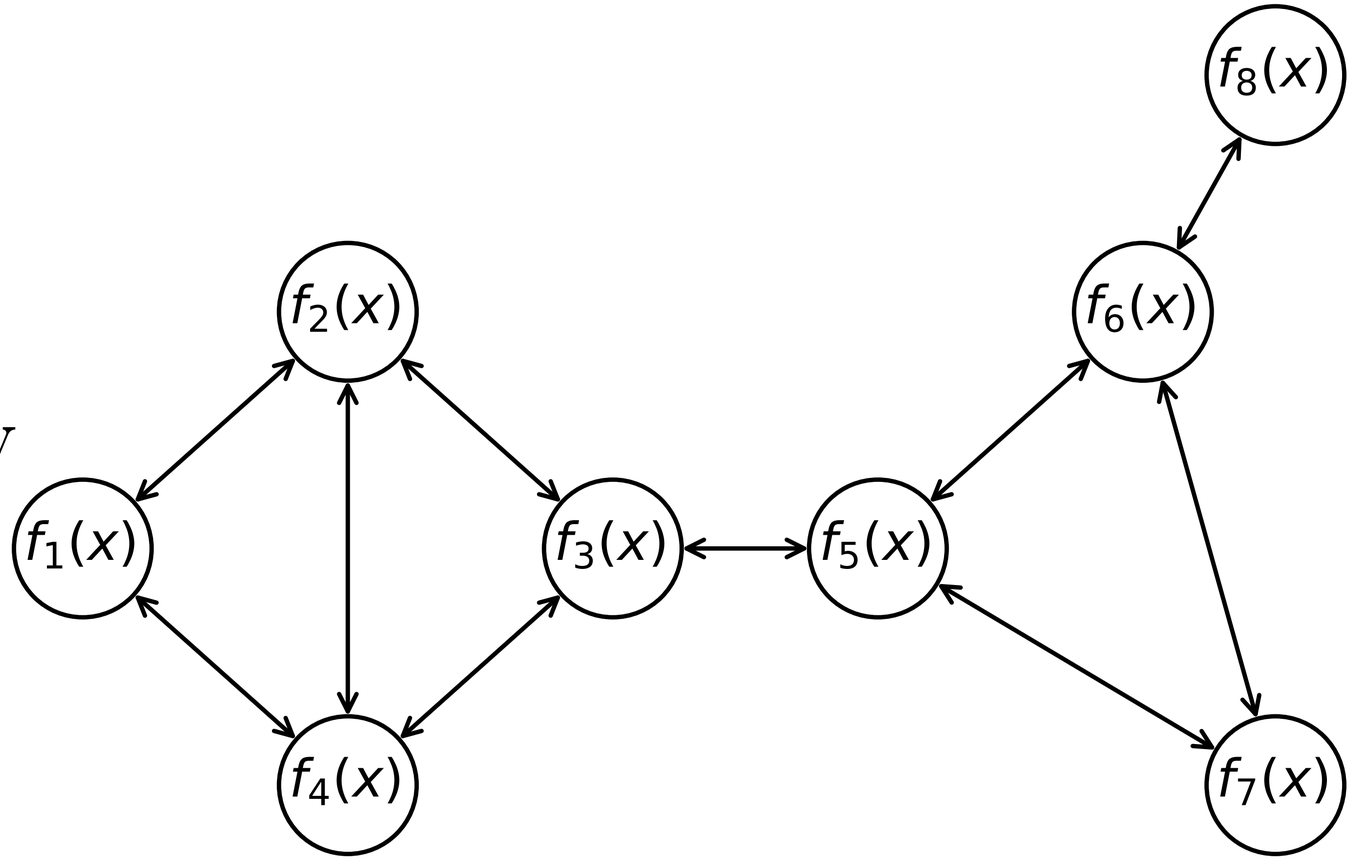


Decentralized Optimization Problem

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

$$f_i(x) : \mathbb{R}^d \to \mathbb{R}$$

- is stored on node i only
- L-smooth
- μ-strongly convex



Gossip Matrix

$\mathbf{W} \in \mathbb{R}^{n \times n} -$ gossip matrix

Gossip Matrix

$$\mathbf{W} \in \mathbb{R}^{n \times n} - \text{ gossip matrix}$$

Property 1

$$\mathbf{W}_{i,j} \neq 0 \text{ if and only if } (i,j) \in \mathcal{E} \text{ or } i = j$$

Gossip Matrix

$\mathbf{W} \in \mathbb{R}^{n \times n}$ — gossip matrix

Property 1

$$\mathbf{W}_{i,j} \neq 0 \text{ if and only if } (i,j) \in \mathcal{E} \text{ or } i = j$$

$$(y_1, \ldots, y_n)^\top = \mathbf{W} \cdot (x_1, \ldots, x_n)^\top$$

$$\Downarrow$$

$$y_i \in \text{span}(\{x_j : j \text{ is a neighbor of } i\})$$

Gossip Matrix

$\mathbf{W} \in \mathbb{R}^{n \times n}$ — gossip matrix

Properties 2 and 3

$$\ker \mathbf{W} \supset \{(x_1, \ldots, x_n) \in \mathbb{R}^n : x_1 = \ldots = x_n\}$$

$$\text{range } \mathbf{W} \subset \{(x_1, \ldots, x_n) \in \mathbb{R}^n : \textstyle\sum_{i=1}^{n} x_i = 0\}$$

technical assumptions

Gossip Matrix

$$\mathbf{W} \in \mathbb{R}^{n \times n} - \text{ gossip matrix}$$

Property 4

there exists $\chi \geq 1$, such that

$$\|\mathbf{W}x - x\|^2 \leq \left(1 - \chi^{-1}\right) \|x\|^2 \text{ for all } x \in \{(x_1, \ldots, x_n) \in \mathbb{R}^n : \sum_{i=1}^n x_i = 0\}$$

Gossip Matrix

$$\mathbf{W} \in \mathbb{R}^{n \times n} - \text{ gossip matrix}$$

Property 4

there exists $\chi \geq 1$, such that

$\|\mathbf{W}x - x\|^2 \leq \left(1 - \chi^{-1}\right) \|x\|^2$ for all $x \in \{(x_1, \ldots, x_n) \in \mathbb{R}^n : \sum_{i=1}^n x_i = 0\}$

contraction property

Gossip Matrix

$$\mathbf{W} \in \mathbb{R}^{n \times n} - \text{ gossip matrix}$$

Property 4

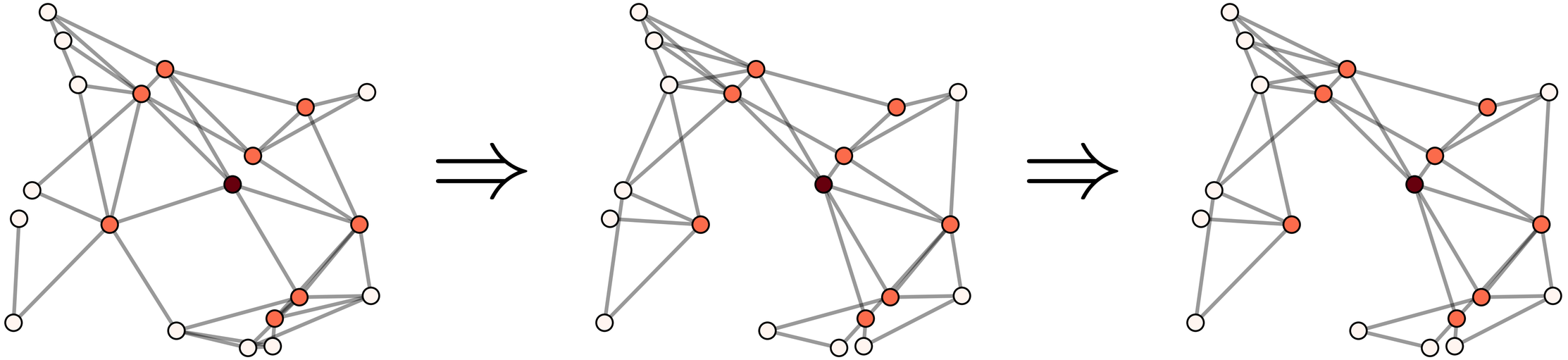
there exists $\chi \geq 1$, such that

$\|\mathbf{W}x - x\|^2 \leq \left(1 - \chi^{-1}\right) \|x\|^2$ for all $x \in \{(x_1, \ldots, x_n) \in \mathbb{R}^n : \sum_{i=1}^n x_i = 0\}$

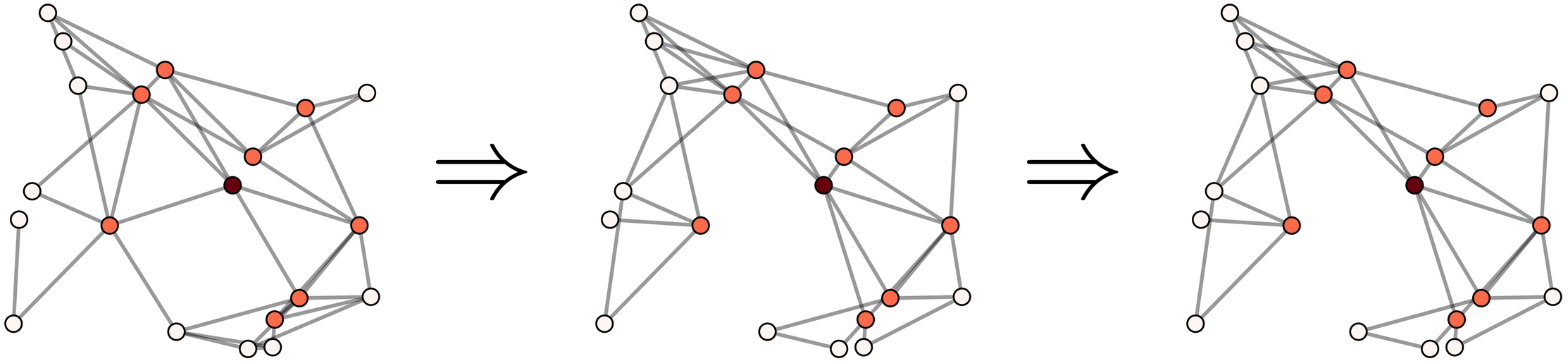
contraction property

χ – condition number of network

Time-varying networks



Time-varying networks



$$\mathcal{G} = (\mathcal{V}, \mathcal{E}) \to \mathcal{G}^k = (\mathcal{V}, \mathcal{E}^k)$$
$$\mathbf{W} \to \mathbf{W}(k)$$

set of edges and gossip matrix change in time

New Algorithm: ADOM+

Problem Reformulation 1

Reformulation via Lifting

$$\min_{x \in \mathcal{L}} F(x)$$

Problem Reformulation 1

Reformulation via Lifting

$$\min_{x \in \mathcal{L}} F(x)$$

block-separable function

$$F(x) = \sum_{i \in \mathcal{V}} f_i(x_i), \text{ where } x = (x_1, \ldots, x_n) \in \left(\mathbb{R}^d\right)^{\mathcal{V}}$$

Problem Reformulation 1

Reformulation via Lifting

$$\min_{x \in \mathcal{L}} F(x)$$

block-separable function

$$F(x) = \sum_{i \in \mathcal{V}} f_i(x_i), \text{ where } x = (x_1, \ldots, x_n) \in \left(\mathbb{R}^d\right)^{\mathcal{V}}$$

$$\mathcal{L} = \left\{(x_1, \ldots, x_n) \in \left(\mathbb{R}^d\right)^{\mathcal{V}} : x_1 = \ldots = x_n\right\}$$

consensus space

Problem Reformulation 2

Saddle Point Reformulation

$$\min_{x \in \mathcal{L}} F(x)$$

Problem Reformulation 2

Saddle Point Reformulation

$$\min_{x \in \mathcal{L}} F(x)$$

$$\min_{x,w \in (\mathbb{R}^d)^\nu} F(x) - \frac{\nu}{2}\|x\|^2 + \frac{\nu}{2}\|w\|^2$$
$$\text{s.t. } x = w,\ x \in \mathcal{L}$$

Problem Reformulation 2

Saddle Point Reformulation

$$\min_{x \in \mathcal{L}} F(x)$$

$$\min_{x,w \in (\mathbb{R}^d)^{\nu}} F(x) - \frac{\nu}{2}\|x\|^2 + \frac{\nu}{2}\|w\|^2$$

$$\text{s.t. } x = w,\ x \in \mathcal{L}$$

$$\nu \in (0, \mu)$$

Problem Reformulation 2

Saddle Point Reformulation

$$\min_{x \in \mathcal{L}} F(x)$$

$$\min_{x,w \in (\mathbb{R}^d)^{\mathcal{V}}} F(x) - \frac{\nu}{2}\|x\|^2 + \frac{\nu}{2}\|w\|^2$$
$$\text{s.t. } x = w,\ x \in \mathcal{L}$$

$\nu \in (0, \mu)$

$$\min_{x,w \in (\mathbb{R}^d)^{\mathcal{V}}} \max_{y \in (\mathbb{R}^d)^{\mathcal{V}}} \max_{z \in \mathcal{L}^\perp} F(x) - \frac{\nu}{2}\|x\|^2 + \frac{\nu}{2}\|w\|^2 + \langle y, w - x\rangle + \langle z, w\rangle$$

Problem Reformulation 2

Saddle Point Reformulation

$$\min_{x,w \in (\mathbb{R}^d)^{\mathcal{V}}} \max_{y \in (\mathbb{R}^d)^{\mathcal{V}}} \max_{z \in \mathcal{L}^{\perp}} F(x) - \frac{\nu}{2}\|x\|^2 + \frac{\nu}{2}\|w\|^2 + \langle y, w - x \rangle + \langle z, w \rangle$$

Problem Reformulation 2

Saddle Point Reformulation

$$\min_{x \in (\mathbb{R}^d)^{\mathcal{V}}} \max_{y \in (\mathbb{R}^d)^{\mathcal{V}}} \max_{z \in \mathcal{L}^\perp} F(x) - \frac{\nu}{2}\|x\|^2 - \langle y, x \rangle - \frac{1}{2\nu}\|y + z\|^2$$

$$\min_{x,w \in (\mathbb{R}^d)^{\mathcal{V}}} \max_{y \in (\mathbb{R}^d)^{\mathcal{V}}} \max_{z \in \mathcal{L}^\perp} F(x) - \frac{\nu}{2}\|x\|^2 + \frac{\nu}{2}\|w\|^2 + \langle y, w - x \rangle + \langle z, w \rangle$$

Problem Reformulation 3

Monotone Inclusion Reformulation

$$\min_{x \in (\mathbb{R}^d)^{\mathcal{V}}} \max_{y \in (\mathbb{R}^d)^{\mathcal{V}}} \max_{z \in \mathcal{L}^{\perp}} F(x) - \frac{\nu}{2}\|x\|^2 - \langle y, x \rangle - \frac{1}{2\nu}\|y + z\|^2$$

Problem Reformulation 3

Monotone Inclusion Reformulation

$$\min_{x \in (\mathbb{R}^d)^{\mathcal{V}}} \max_{y \in (\mathbb{R}^d)^{\mathcal{V}}} \max_{z \in \mathcal{L}^\perp} F(x) - \frac{\nu}{2}\|x\|^2 - \langle y, x \rangle - \frac{1}{2\nu}\|y + z\|^2$$

optimality conditions

$$(x^*, y^*, z^*) \in \left(\mathbb{R}^d\right)^{\mathcal{V}} \times \left(\mathbb{R}^d\right)^{\mathcal{V}} \times \mathcal{L}^\perp$$

$$0 = \nabla F\left(x^*\right) - \nu x^* - y^*$$

$$0 = \nu^{-1}\left(y^* + z^*\right) + x^*$$

$$\mathcal{L} \ni y^* + z^*$$

Problem Reformulation 3

Monotone Inclusion Reformulation

$$
\begin{aligned}
&(x^*, y^*, z^*) \in \left(\mathbb{R}^d\right)^{\mathcal{V}} \times \left(\mathbb{R}^d\right)^{\mathcal{V}} \times \mathcal{L}^{\perp} \\
&0 = \nabla F\left(x^*\right) - \nu x^* - y^* \\
&0 = \nu^{-1}\left(y^* + z^*\right) + x^* \\
&\mathcal{L} \ni y^* + z^*
\end{aligned}
$$

$$
\begin{aligned}
&(x^*, y^*, z^*) \in \left(\mathbb{R}^d\right)^{\mathcal{V}} \times \left(\mathbb{R}^d\right)^{\mathcal{V}} \times \mathcal{L}^{\perp} \\
&A(x^*, y^*, z^*) + B(x^*, y^*, z^*) = 0
\end{aligned}
$$

Problem Reformulation 3

Monotone Inclusion Reformulation

monotone operators

$$A(x,y,z) = \begin{bmatrix} \nabla F(x) - \nu x \\ \nu^{-1}(y+z) \\ \mathbf{P}\nu^{-1}(y+z) \end{bmatrix} \qquad B(x,y,z) = \begin{bmatrix} -y \\ x \\ 0 \end{bmatrix}$$

$$(x^*, y^*, z^*) \in \left(\mathbb{R}^d\right)^{\mathcal{V}} \times \left(\mathbb{R}^d\right)^{\mathcal{V}} \times \mathcal{L}^{\perp}$$

$$0 = \nabla F\left(x^*\right) - \nu x^* - y^*$$

$$0 = \nu^{-1}\left(y^* + z^*\right) + x^*$$

$$\mathcal{L} \ni y^* + z^*$$

$$(x^*, y^*, z^*) \in \left(\mathbb{R}^d\right)^{\mathcal{V}} \times \left(\mathbb{R}^d\right)^{\mathcal{V}} \times \mathcal{L}^{\perp}$$

$$A(x^*, y^*, z^*) + B(x^*, y^*, z^*) = 0$$

Problem Reformulation 3

Monotone Inclusion Reformulation

monotone operators

$$A(x,y,z) = \begin{bmatrix} \nabla F(x) - \nu x \\ \nu^{-1}(y+z) \\ \mathbf{P}\nu^{-1}(y+z) \end{bmatrix} \qquad B(x,y,z) = \begin{bmatrix} -y \\ x \\ 0 \end{bmatrix}$$

$\mathbf{P}$ – orthogonal projection matrix onto $\mathcal{L}^{\perp}$

$$(x^*, y^*, z^*) \in \left(\mathbb{R}^d\right)^{\mathcal{V}} \times \left(\mathbb{R}^d\right)^{\mathcal{V}} \times \mathcal{L}^{\perp}$$

$$A(x^*, y^*, z^*) + B(x^*, y^*, z^*) = 0$$

Algorithm Design

Forward-Backward Algorithm

$$\left(x^{+}, y^{+}, z^{+}\right) = J_{\omega B}[(x, y, z) - \omega A(x, y, z)]$$

Algorithm Design

ω – stepsize

Forward-Backward Algorithm

$$(x^+, y^+, z^+) = J_{\omega B}[(x, y, z) - \omega A(x, y, z)]$$

Algorithm Design

Forward-Backward Algorithm

$$(x^+, y^+, z^+) = J_{\omega B}[(x, y, z) - \omega A(x, y, z)]$$

ω – stepsize

$J_{\omega B} = (I + \omega B)^{-1}$ – resolvent

identity mapping

Algorithm Design

Forward-Backward Algorithm

$$(x^+, y^+, z^+) = J_{\omega B}[(x, y, z) - \omega A(x, y, z)]$$

+

ω – stepsize

$J_{\omega B} = (I + \omega B)^{-1}$ – resolvent

identity mapping

Algorithm Design

Forward-Backward Algorithm

$$(x^+, y^+, z^+) = J_{\omega B}[(x, y, z) - \omega A(x, y, z)]$$

ω – stepsize

$J_{\omega B} = (I + \omega B)^{-1}$ – resolvent

identity mapping

+

Nesterov Acceleration

Algorithm Design

Forward-Backward Algorithm

$$(x^+, y^+, z^+) = J_{\omega B}[(x, y, z) - \omega A(x, y, z)]$$

$$A(x, y, z) = \begin{bmatrix} \nabla F(x) - \nu x \\ \nu^{-1}(y + z) \\ \mathbf{P}\nu^{-1}(y + z) \end{bmatrix}$$

+

Nesterov Acceleration

Algorithm Design

Forward-Backward Algorithm

$$(x^+, y^+, z^+) = J_{\omega B}[(x, y, z) - \omega A(x, y, z)]$$

+

Nesterov Acceleration

$$A(x, y, z) = \begin{bmatrix} \nabla F(x) - \nu x \\ \nu^{-1}(y + z) \\ \mathbf{P}\nu^{-1}(y + z) \end{bmatrix}$$

$$\mathbf{P} = \left(\mathbf{I}_n - \frac{1}{n}\mathbf{1}_n\mathbf{1}_n^\top\right) \otimes \mathbf{I}_d$$

Algorithm Design

Forward-Backward Algorithm

$$(x^+, y^+, z^+) = J_{\omega B}[(x, y, z) - \omega A(x, y, z)]$$

+

Nesterov Acceleration

$$A(x, y, z) = \begin{bmatrix} \nabla F(x) - \nu x \\ \nu^{-1}(y + z) \\ \mathbf{P}\nu^{-1}(y + z) \end{bmatrix}$$

$$\mathbf{P} = \left(\mathbf{I}_n - \frac{1}{n}\mathbf{1}_n\mathbf{1}_n^\top\right) \otimes \mathbf{I}_d$$

requires full averaging over the network

Algorithm Design

Forward-Backward Algorithm

$$(x^+, y^+, z^+) = J_{\omega B}[(x, y, z) - \omega A(x, y, z)]$$

+

Nesterov Acceleration

$$A(x, y, z) = \begin{bmatrix} \nabla F(x) - \nu x \\ \nu^{-1}(y + z) \\ \mathbf{P}\nu^{-1}(y + z) \end{bmatrix}$$

replace $\mathbf{P}\nu^{-1}(y + z)$ with $(\mathbf{W}(k) \otimes \mathbf{I}_d)\,\nu^{-1}(y + z)$

Biased Gradient vs Projected Gradient

$$(\mathbf{W}(k) \otimes \mathbf{I}_d)\, \nu^{-1}(y+z) \quad \text{vs} \quad \mathbf{P}\nu^{-1}(y+z)$$

Biased Gradient vs Projected Gradient

$$(\mathbf{W}(k) \otimes \mathbf{I}_d)\, \nu^{-1}(y+z) \quad \text{vs} \quad \mathbf{P}\nu^{-1}(y+z)$$

Recall property 1

$$\mathbf{W}_{i,j} \neq 0 \text{ if and only if } (i,j) \in \mathcal{E} \text{ or } i = j$$

Biased Gradient vs Projected Gradient

$$(\mathbf{W}(k) \otimes \mathbf{I}_d)\, \nu^{-1}(y + z) \quad \text{vs} \quad \mathbf{P}\nu^{-1}(y + z)$$

Recall property 1

$$\mathbf{W}_{i,j} \neq 0 \text{ if and only if } (i, j) \in \mathcal{E} \text{ or } i = j$$

$(\mathbf{W}(k) \otimes \mathbf{I}_d)\, \nu^{-1}(y + z)$ can be computed with decentralized communication

Biased Gradient vs Projected Gradient

$$(\mathbf{W}(k) \otimes \mathbf{I}_d)\, \nu^{-1}(y+z) \quad \text{vs} \quad \mathbf{P}\nu^{-1}(y+z)$$

Recall property 3

$$\text{range } \mathbf{W} \subset \{(x_1, \ldots, x_n) \in \mathbb{R}^n : \textstyle\sum_{i=1}^n x_i = 0\}$$

Biased Gradient vs Projected Gradient

$$(\mathbf{W}(k) \otimes \mathbf{I}_d)\, \nu^{-1}(y+z) \quad \text{vs} \quad \mathbf{P}\nu^{-1}(y+z)$$

Recall property 3

$$\text{range } \mathbf{W} \subset \{(x_1, \ldots, x_n) \in \mathbb{R}^n : \textstyle\sum_{i=1}^n x_i = 0\}$$

$(\mathbf{W}(k) \otimes \mathbf{I}_d)\, \nu^{-1}(y+z)$ belongs to $\mathcal{L}^{\perp}$

no need for projection

Biased Gradient vs Projected Gradient

$$(\mathbf{W}(k) \otimes \mathbf{I}_d)\, \nu^{-1}(y+z) \quad \text{vs} \quad \mathbf{P}\nu^{-1}(y+z)$$

Recall property 4 (contraction property)

$$\|\mathbf{W}x - x\|^2 \leq \left(1 - \chi^{-1}\right) \|x\|^2 \text{ for all } x \in \{(x_1, \ldots, x_n) \in \mathbb{R}^n : \textstyle\sum_{i=1}^n x_i = 0\}$$

Biased Gradient vs Projected Gradient

$$(\mathbf{W}(k) \otimes \mathbf{I}_d)\, \nu^{-1}(y+z) \quad \text{vs} \quad \mathbf{P}\nu^{-1}(y+z)$$

Recall property 4 (contraction property)

$$\|\mathbf{W}x - x\|^2 \leq \left(1-\chi^{-1}\right)\|x\|^2 \text{ for all } x \in \{(x_1,\ldots,x_n) \in \mathbb{R}^n : \textstyle\sum_{i=1}^n x_i = 0\}$$

$$\Downarrow$$

$$\|\,(\mathbf{W}(k) \otimes \mathbf{I}_d)\, \nu^{-1}(y+z) - \mathbf{P}\nu^{-1}(y+z)\|^2 \leq (1-\chi^{-1})\|\mathbf{P}\nu^{-1}(y+z)\|^2$$

Biased Gradient vs Projected Gradient

$$(\mathbf{W}(k) \otimes \mathbf{I}_d)\, \nu^{-1}(y+z) \quad \text{vs} \quad \mathbf{P}\nu^{-1}(y+z)$$

Recall property 4 (contraction property)

$$\|\mathbf{W}x - x\|^2 \leq \left(1-\chi^{-1}\right)\|x\|^2 \text{ for all } x \in \{(x_1,\ldots,x_n) \in \mathbb{R}^n : \textstyle\sum_{i=1}^n x_i = 0\}$$

$$\Downarrow$$

$$\|(\mathbf{W}(k) \otimes \mathbf{I}_d)\, \nu^{-1}(y+z) - \mathbf{P}\nu^{-1}(y+z)\|^2 \leq (1-\chi^{-1})\|\mathbf{P}\nu^{-1}(y+z)\|^2$$

Biased Gradient vs Projected Gradient

$$(\mathbf{W}(k) \otimes \mathbf{I}_d)\, \nu^{-1}(y+z) \quad \text{vs} \quad \mathbf{P}\nu^{-1}(y+z)$$

Recall property 4 (contraction property)

$$\|\mathbf{W}x - x\|^2 \leq \left(1-\chi^{-1}\right)\|x\|^2 \text{ for all } x \in \{(x_1,\ldots,x_n) \in \mathbb{R}^n : \textstyle\sum_{i=1}^n x_i = 0\}$$

$$\Downarrow$$

$$\|(\mathbf{W}(k) \otimes \mathbf{I}_d)\, \nu^{-1}(y+z) - \mathbf{P}\nu^{-1}(y+z)\|^2 \leq (1-\chi^{-1})\|\mathbf{P}\nu^{-1}(y+z)\|^2$$

apply *Error-Feedback* mechanism for biased contraction operator

Algorithm Design

Forward-Backward Algorithm

$$\left(x^{+}, y^{+}, z^{+}\right) = J_{\omega B}[(x, y, z) - \omega A(x, y, z)]$$

Nesterov Acceleration

Algorithm Design

Forward-Backward Algorithm

$$\left(x^{+}, y^{+}, z^{+}\right) = J_{\omega B}[(x, y, z) - \omega A(x, y, z)]$$

Nesterov Acceleration

Error-Feedback

Result: ADOM+

Algorithm 1 ADOM+

input: $x^0, y^0, m^0 \in (\mathbb{R}^d)^{\mathcal{V}}$, $z^0 \in \mathcal{L}^{\perp}$
$x_f^0 = x^0$, $y_f^0 = y^0$, $z_f^0 = z^0$
for $k = 0, 1, 2, \ldots$ **do**
  $x_g^k = \tau_1 x^k + (1 - \tau_1) x_f^k$
  $x^{k+1} = x^k + \eta\alpha(x_g^k - x^{k+1}) - \eta\left[\nabla F(x_g^k) - \nu x_g^k - y^{k+1}\right]$
  $x_f^{k+1} = x_g^k + \tau_2(x^{k+1} - x^k)$
  $y_g^k = \sigma_1 y^k + (1 - \sigma_1) y_f^k$
  $y^{k+1} = y^k + \theta\beta(\nabla F(x_g^k) - \nu x_g^k - y^{k+1}) - \theta\left[\nu^{-1}(y_g^k + z_g^k) + x^{k+1}\right]$
  $y_f^{k+1} = y_g^k + \sigma_2(y^{k+1} - y^k)$
  $z_g^k = \sigma_1 z^k + (1 - \sigma_1) z_f^k$
  $z^{k+1} = z^k + \gamma\delta(z_g^k - z^k) - (\mathbf{W}(k) \otimes \mathbf{I}_d)\left[\gamma\nu^{-1}(y_g^k + z_g^k) + m^k\right]$
  $m^{k+1} = \gamma\nu^{-1}(y_g^k + z_g^k) + m^k - (\mathbf{W}(k) \otimes \mathbf{I}_d)\left[\gamma\nu^{-1}(y_g^k + z_g^k) + m^k\right]$
  $z_f^{k+1} = z_g^k - \zeta(\mathbf{W}(k) \otimes \mathbf{I}_d)(y_g^k + z_g^k)$
end for

Iteration Complexity of ADOM+

$$\mathcal{O}\left(\chi\sqrt{L/\mu}\log\frac{1}{\epsilon}\right)$$

Iteration Complexity of ADOM+

$$\mathcal{O}\left(\chi\sqrt{L/\mu}\log\frac{1}{\epsilon}\right)$$

- This is an optimal communication complexity (we prove lower complexity bounds)
- We can reach optimal gradient computation complexity using multi-consensus procedure
- There is a dual-based optimal algorithm ADOM, which is based on similar ideas for a slightly different reformulation