Lower bounds on decentralized optimization problems under a constant constraint on the change of edges per iteration in a communication network.

Dmitry Metelev, Alexander Rogozin, Alexander Gasnikov

May 19, 2023

Introduction in Decentralized Optimization

We have a graph $G = (V, E)$, where the vertices $v \in V$ are called nodes, and each node v corresponds to a function f_v, which we call the local function of v.

Introduction in Decentralized Optimization

We have a graph $G = (V, E)$, where the vertices $v \in V$ are called nodes, and each node v corresponds to a function f_v, which we call the local function of v.

Intuitively, the graph G is a network of interconnected computers that can communicate only if they are connected by an edge. Each computer stores its function and information in local memory.

Introduction in Decentralized Optimization

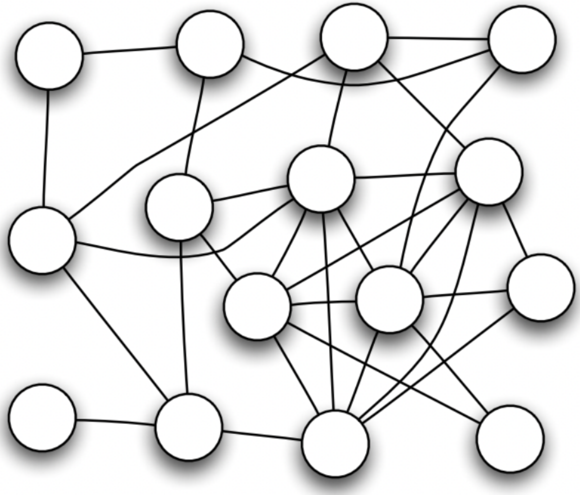
We have a graph $G = (V, E)$, where the vertices $v \in V$ are called nodes, and each node v corresponds to a function f_v, which we call the local function of v.

Intuitively, the graph G is a network of interconnected computers that can communicate only if they are connected by an edge. Each computer stores its function and information in local memory.

The generalized optimization problem is to find such a value

$$\inf_{x \in \mathbb{R}^m} f(x) = \frac{1}{n} \sum_{i=1}^{n} f_i(x).$$

Example of a network



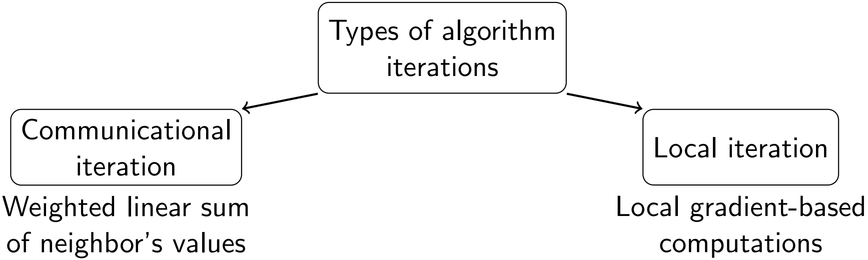
Undirected network

The concept of time-varying networks

At each moment of time, we have a graph representing the network with a fixed set of vertices and corresponding functions. Edges can appear or disappear over time. Consequently, we have a sequence of graphs $\{G_i\}_{i=1}^{\infty}$.

The concept of time-varying networks

At each moment of time, we have a graph representing the network with a fixed set of vertices and corresponding functions. Edges can appear or disappear over time. Consequently, we have a sequence of graphs $\{G_i\}_{i=1}^{\infty}$.

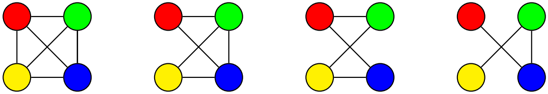


The concept of time-varying networks

At each moment of time, we have a graph representing the network with a fixed set of vertices and corresponding functions. Edges can appear or disappear over time. Consequently, we have a sequence of graphs $\{G_i\}_{i=1}^{\infty}$.

Types of algorithm iterations

- Communicational iteration: Weighted linear sum of neighbor's values
- Local iteration: Local gradient-based computations



Example sequence of graphs on four vertices: $G_1 \to G_2 \to G_3 \to G_4$

Laplacian Matrix of a Graph and Weighted Laplacian

$$L_{ij}(W) = \begin{cases} \sum_{j \sim i} w_{ij} & \text{if } i = j \\ -w_{ij} & \text{if } i \neq j \text{ and } (v_i, v_j) \in E \\ 0 & \text{otherwise} \end{cases}$$

where E is the set of graph edges, and w_{ij} is the weight of the edge (v_i, v_j).

Laplacian Matrix of a Graph and Weighted Laplacian

$$L_{ij}(W) = \begin{cases} \sum_{j\sim i} w_{ij} & \text{if } i = j \\ -w_{ij} & \text{if } i \neq j \text{ and } (v_i, v_j) \in E \\ 0 & \text{otherwise} \end{cases}$$

where E is the set of graph edges, and w_{ij} is the weight of the edge (v_i, v_j).

Example:

Consider a graph with 3 vertices and edges $(v_1, v_2), (v_2, v_3)$ with weights $w_{12} = 2$ and $w_{23} = 1$.

The weighted Laplacian matrix for this graph looks as follows:

$$L(W) = \begin{bmatrix} 2 & -2 & 0 \\ -2 & 3 & -1 \\ 0 & -1 & 1 \end{bmatrix}$$

Gossip Matrices

- For each round of decentralized communication $k \in \{0, 1, 2, \ldots\}$, consider a matrix $W_k \in R^{n \times n}$ that satisfies the following conditions:
  1. $[W_k]_{i,j} = 0$, if $i \neq j$ and $(i, j) \notin E_k$,
  2. $\ker W_k \supseteq \{(x_1, \ldots, x_n) \in R^n : x_1 = \cdots = x_n\}$,
  3. $\operatorname{Im} W_k \subseteq \left\{(x_1, \ldots, x_n) \in R^n : \sum_{i=1}^n x_i = 0\right\}$,
  4. There exists $\chi \geq 1$, such that $||Wx - x||^2 \leq (1 - \chi^{-1})||x||^2$ for all $x \in \{(x_1, \ldots, x_n) \in R^n : \sum_{i=1}^n x_i = 0\}$.

Gossip Matrices

- For each round of decentralized communication $k \in \{0, 1, 2, \dots\}$, consider a matrix $W_k \in R^{n \times n}$ that satisfies the following conditions:
  1. $[W_k]_{i,j} = 0$, if $i \neq j$ and $(i, j) \notin E_k$,
  2. $\ker W_k \supseteq \{(x_1, \dots, x_n) \in R^n : x_1 = \dots = x_n\}$,
  3. $\operatorname{Im} W_k \subseteq \{(x_1, \dots, x_n) \in R^n : \sum_{i=1}^n x_i = 0\}$,
  4. There exists $\chi \geq 1$, such that $||Wx - x||^2 \leq (1 - \chi^{-1})||x||^2$ for all $x \in \{(x_1, \dots, x_n) \in R^n : \sum_{i=1}^n x_i = 0\}$.
- An example of a gossip matrix: $W = \lambda_{\max}^{-1} \cdot L(G))$, where $L(G)$ is the Laplacian of an undirected connected graph G. Then $\chi = \frac{\lambda_{\max}(L(G_k))}{\lambda_{\min}^{+}(L(G_k))}$.

Gossip Matrices

- For each round of decentralized communication $k \in \{0, 1, 2, \dots\}$, consider a matrix $W_k \in R^{n \times n}$ that satisfies the following conditions:
  1. $[W_k]_{i,j} = 0$, if $i \neq j$ and $(i, j) \notin E_k$,
  2. $\ker W_k \supseteq \{(x_1, \dots, x_n) \in R^n : x_1 = \dots = x_n\}$,
  3. $\operatorname{Im} W_k \subseteq \{(x_1, \dots, x_n) \in R^n : \sum_{i=1}^n x_i = 0\}$,
  4. There exists $\chi \geq 1$, such that $||Wx - x||^2 \leq (1 - \chi^{-1})||x||^2$ for all $x \in \{(x_1, \dots, x_n) \in R^n : \sum_{i=1}^n x_i = 0\}$.
- An example of a gossip matrix: $W = \lambda_{\max}^{-1} \cdot L(G))$, where $L(G)$ is the Laplacian of an undirected connected graph G. Then $\chi = \frac{\lambda_{\max}(L(G_k))}{\lambda_{\min}^{+}(L(G_k))}$.
- For a time-varying network, χ can be chosen as an upper bound for the numbers χ_k, defined as $\chi = \sup\limits_k \chi_k$.

Gossip Matrices

- For each round of decentralized communication $k \in \{0, 1, 2, \dots\}$, consider a matrix $W_k \in R^{n \times n}$ that satisfies the following conditions:
  1. $[W_k]_{i,j} = 0$, if $i \neq j$ and $(i, j) \notin E_k$,
  2. $\ker W_k \supseteq \{(x_1, \dots, x_n) \in R^n : x_1 = \dots = x_n\}$,
  3. $\operatorname{Im} W_k \subseteq \left\{(x_1, \dots, x_n) \in R^n : \sum_{i=1}^n x_i = 0\right\}$,
  4. There exists $\chi \geq 1$, such that $||Wx - x||^2 \leq (1 - \chi^{-1})||x||^2$ for all $x \in \{(x_1, \dots, x_n) \in R^n : \sum_{i=1}^n x_i = 0\}$.
- An example of a gossip matrix: $W = \lambda_{\max}^{-1} \cdot L(G))$, where $L(G)$ is the Laplacian of an undirected connected graph G. Then $\chi = \frac{\lambda_{\max}(L(G_k))}{\lambda_{\min}^{+}(L(G_k))}$.
- For a time-varying network, χ can be chosen as an upper bound for the numbers χ_k, defined as $\chi = \sup\limits_k \chi_k$.
- We will call χ the condition number of the time-varying network.

Consensus Procedure

Consensus from the point of view of each node:

$$x_i^{k+1} = w_{ii}^k x_i^k + \sum_{(i,j)\in E_k} w_{ij}^k x_j^k$$

Consensus Procedure

Consensus from the point of view of each node:

$$x_i^{k+1} = w_{ii}^k x_i^k + \sum_{(i,j)\in E_k} w_{ij}^k x_j^k$$

Consensus in matrix form:

$$X^{k+1} = W^k X^k$$

Known Communication Complexity Lower Bounds

Table 1: Known lower bounds on communication complexity for decentralized optimization. Here $\alpha > 0$ is a scalar, and $d \in \mathbb{N}$ is a constant. Complexity depends on the maximum number of edge changes allowed per iteration.

Number of changes	Lower Bound	Citation
no changes	$\Omega\left(\chi^{1/2}\sqrt{\frac{L}{\mu}}\log\frac{1}{\epsilon}\right)$	Scaman et al. [2017]
$O(n)$	$\Omega\left(\chi\sqrt{\frac{L}{\mu}}\log\frac{1}{\epsilon}\right)$	Kovalev et al. [2021]
$O(n^{\alpha})$	$\Omega\left(\chi\sqrt{\frac{L}{\mu}}\log\frac{1}{\epsilon}\right)$	Metelev et al. [2023]
$O(\log n)$	$\Omega\left(\frac{\chi}{\log\chi}\sqrt{\frac{L}{\mu}}\log\frac{1}{\epsilon}\right)$	Metelev et al. [2023]
$12(d-1)$	$\Omega\left(\chi^{d/(d+1)}\sqrt{\frac{L}{\mu}}\log\frac{1}{\epsilon}\right)$	Metelev et al. [2023]

Main Result

Table 2: Lower bound on communication complexity for decentralized optimization with two edge changes per iteration.

Number of changes	Lower Bound
2	$\Omega\left(\chi\sqrt{\frac{L}{\mu}}\log\frac{1}{\epsilon}\right)$

Such a result was obtained due to a properly chosen counterexample matrix and the following lemma:

Lemma

For every graph $G = (V, E)$ there exists a weighted Laplacian $L(G)$ with $\chi(L(G)) \leq 2Dn$, where $n = |V|$ and D is the diameter of G.

Example of a network

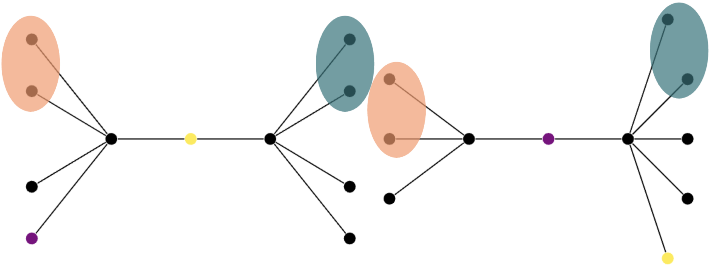


Figure 2: Counter-example graph for the last result

References

Kevin Scaman, Francis Bach, Sébastien Bubeck, Yin Tat Lee, and Laurent Massoulié. Optimal algorithms for smooth and strongly convex distributed optimization in networks. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, pages 3027–3036. JMLR. org, 2017.

Dmitry Kovalev, Egor Shulgin, Peter Richtárik, Alexander Rogozin, and Alexander Gasnikov. Adom: Accelerated decentralized optimization method for time-varying networks. *arXiv preprint arXiv:2102.09234*, 2021.

Dmitriy Metelev, Alexander Rogozin, Dmitry Kovalev, and Alexander Gasnikov. Is consensus acceleration possible in decentralized optimization over slowly time-varying networks?, 2023.