A SURVEY ON DECENTRALIZED OPTIMIZATION

Nhat Trung Nguyen
Moscow Institute of Physics and Technology

December 16, 2024

ABSTRACT

Decentralized optimization is an important direction of research in modern optimization theory. Its applications include large scale machine learning, distributed signal processing, and other distributed control problems. This paper provides a comprehensive survey of existing results and techniques for decentralized optimization, covering smooth and non-smooth problems, saddle point formulations, and optimization with affine constraints. Special attention is given to methods for time-varying networks, which remain challenging due to their dynamic nature. Despite progress in recent works for smooth settings, the non-smooth case and coupled constraints in time-varying networks are still underexplored. This work provides a foundation for further research and highlights open questions in the field.

1 Introduction

In this work, we consider the following minimization problem in a decentralized setting:

$$\min_{x \in \mathcal{X}} f(x) = \frac{1}{n} \sum_{i=1}^{n} f_i(x), \tag{1}$$

where the set $\mathcal{X} \subseteq \mathbb{R}^d$ is convex. We also cover related variants, including saddle point problems and optimization under affine constraints. Each function $f_i(x)$ is held locally at some computational node. The nodes are connected to each other via a decentralized communication network. Since there is no centralized aggregator (server node, master node), each agent locally holds a copy x_i of the decision vector x. Each agent can perform local computations and exchange information with its immediate neighbors in the network. Additionally, the network is allowed to change over time. This type of networks is called *time-varying graphs*.

Problems of the form (1) arise in numerous applications where centralized data aggregation is infeasible due to limitations such as network structure, privacy constraints, or the large volume of data. Notable applications include distributed statistical and machine learning [1, 2], power system control [3, 4], distributed control and signal processing [5, 6].

This work focuses on first-order methods for solving (1) under the assumption that all local objectives $f_i(x)$ are convex. Decentralized optimization using first-order methods involves two main steps: local gradient updates and inter-node communications. We consider the case where communications occur in synchronized rounds. Consequently, the complexity of the method depends on the characteristics of the problem, which include: the number of communication rounds and the number of local oracle calls. These complexity measures depend on various problem characteristics, including the network condition number, the Lipschitz constant L of the gradient in the case of smooth optimization, the strong convexity constant μ in the strongly convex setting, the Lipschitz constant M of the objective in the non-smooth setting, and the desired accuracy ε.

Problem (1) can be reformulated as

$$\min_{\mathbf{x} \in \mathcal{X}^n} F(\mathbf{x}) = \frac{1}{n} \sum_{i=1}^{n} f_i(x_i) \quad \text{s.t.} \quad x_1 = \cdots = x_n. \tag{2}$$

A decentralized optimization algorithm should be designed to minimize the sum of the objective functions while ensuring that the decision vectors across different computational nodes remain approximately consistent.

2 Time-varying graphs

When the network changes over time, the problem is referred to as *decentralized optimization over time-varying graphs*. At each communication round, we use a graph to represent the connection between computing nodes. Denote the network of communications over time by the sequence of graphs $\{\mathcal{G}^k\}_{k=0}^{\infty} = \{(\mathcal{V}, \mathcal{E}^k)\}_{k=0}^{\infty}$, where $\mathcal{V} = \{1, \ldots, n\}$ is the set of nodes and $\mathcal{E}^k$ is the set of available connections at k-th communication round. For each node $i \in \mathcal{V}$, we use the notation $\mathcal{N}_i^k = \{j \in \mathcal{V} | (i,j) \in \mathcal{E}^k\}$ to indicate the set of its neighbors at round k and at that time it can only communicate with nodes in $\mathcal{N}_i^k$.

Assumption 1. Each graph in the time-varying network correspond to a gossip matrix $W^k \in \mathbb{R}^{n \times n}$ that satisfies the following properties.

1. W^k is positive semi-definite,
2. $W_{i,j}^k = 0$ if $i \neq j$ and $(i,j) \notin \mathcal{E}^k$,
3. $\ker W^k = \operatorname{span}(\mathbf{1})$, where $\mathbf{1} = (1, \ldots, 1) \in \mathbb{R}^n$.

The number $\chi(W) = \frac{\lambda_{\max}(W)}{\lambda_{\min}^+(W)}$ is called the *condition number* of gossip matrix W, where $\lambda_{\max}(W)$ and $\lambda_{\min}^+(W)$ denote the largest and smallest positive eigenvalue of W. For a time-varying network $\{\mathcal{G}^k\}_{k=1}^{\infty}$, the condition number is given by $\chi = \sup_{k \in \mathbb{N} \cup \{0\}} \frac{\lambda_{\max}(W^k)}{\lambda_{\min}^+(W^k)}$.

An example of gossip matrix is the *Laplacian matrix* associated with the graph:

$$W_{ij}^k = \begin{cases} -1, & \text{if } (i,j) \in \mathcal{E}^k, \\ \deg(i), & \text{if } i = j, \\ 0, & \text{other wise.} \end{cases}$$

Multi-step communication procedures are widely utilized in distributed optimization over static graphs. In the time-invariant scenario, a technique known as *Chebyshev acceleration* is employed to reduce the communication complexity to $\sqrt{\chi}$. Specifically, given a gossip matrix W with condition number χ, one can approximate W using a Chebyshev polynomial $P_K(W)$ of degree $K = \lceil \sqrt{\chi} \rceil$. The condition number of $P_K(W)$ becomes $\mathcal{O}(1)$ due to the inherent structure of the polynomial (see [7]). However, such acceleration of communication steps cannot be achieved in the time-varying setting. In this case, it is not possible to reduce χ to $\sqrt{\chi}$, as the lower complexity bounds impose stricter limitations (see [8]).

3 Optimization problems

3.1 Notations and Assumptions

Definition 1 (Convexity, Strong Convexity, and Smoothness)**.** Consider a function $h : \mathbb{R}^d \to \mathbb{R}$.

1. h is **convex** if for any $x_1, x_2 \in \mathbb{R}^d$, it holds:
$$h(x_2) \geq h(x_1) + \langle \nabla h(x_1), x_2 - x_1 \rangle.$$
2. h is μ**-strongly convex** if for any $x_1, x_2 \in \mathbb{R}^d$, it holds:
$$h(x_2) \geq h(x_1) + \langle \nabla h(x_1), x_2 - x_1 \rangle + \frac{\mu}{2} \|x_2 - x_1\|^2.$$
3. h is L**-smooth** if for any $x_1, x_2 \in \mathbb{R}^d$, it holds:
$$h(x_2) \leq h(x_1) + \langle \nabla h(x_1), x_2 - x_1 \rangle + \frac{L}{2} \|x_2 - x_1\|^2.$$

The following assumptions on objective functions are standard in convex optimization.

Assumption 2. For every $i = 1, \ldots, n$, the function f_i is convex.

Assumption 3. For every $i = 1, \ldots, n$, the function f_i is μ_i-strongly convex.

Assumption 4. For every $i = 1, \ldots, n$, the function f_i is L_i-smooth.

Assumption 5. For every $i = 1, \ldots, n$, the norm of the subgradients of f_i is bounded by M_i, i.e.,

$$\|\partial f_i(x)\| \leq M_i.$$

Assumption 6. We assume that the feasible region $\mathcal{X}$ is convex and bounded, and denote by R the radius of a ball containing $\mathcal{X}$, i.e.

$$\forall x \in \mathcal{X} : \ \|x\| \leq R. \tag{3}$$

We define following local constants for the problem:

$$L = \max_{i=1,\ldots,m} L_i, \quad M = \max_{i=1,\ldots,m} M_i, \quad \mu = \min_{i=1,\ldots,m} \mu_i.$$

3.2 Lower Bounds and Optimal Algorithms for Smooth and Strongly Convex Optimization

For optimization over networks lower bounds are known as well as corresponding optimal algorithms. For static graphs, lower bounds were derived in [7] and in the same paper the optimal dual (i.e. using a dual oracle) algorithm was proposed. Optimal decentralized methods with primal oracle were proposed in [9]. Considering time-varying graphs (with arbitrary changes at each iteration), lower complexity bounds were proposed in [8]. Optimal primal algorithm was proposed in the same paper [8], and optimal dual method first appeared in [10]. After that, lower bounds for slowly time-varying graphs with different velocities of network changes were studied in [11]. In [12], it was shown that it is sufficient to change only two edges at each iteration to make communication complexity equal to arbitrary time-varying graph. The overview of optimal convergence rates for decentralized optimization is presented in Table 1 (notation $\mathcal{O}(\cdot)$ is omitted).

	static	time-var.	slowly time-var.
comm.	$\sqrt{\chi}\sqrt{\frac{L}{\mu}}\log\frac{1}{\varepsilon}$	$\chi\sqrt{\frac{L}{\mu}}\log\frac{1}{\varepsilon}$	$\chi\sqrt{\frac{L}{\mu}}\log\frac{1}{\varepsilon}$
oracle	$\sqrt{\frac{L}{\mu}}\log\frac{1}{\varepsilon}$	$\sqrt{\frac{L}{\mu}}\log\frac{1}{\varepsilon}$	$\sqrt{\frac{L}{\mu}}\log\frac{1}{\varepsilon}$
paper	[7], [9]	[8]	[12], [8]

Table 1: Convergence rates for smooth and strongly convex optimization

3.3 Lower Bounds and Optimal Algorithms for Non-smooth Optimization

In the context of decentralized non-smooth optimization over networks, both lower complexity bounds and optimal algorithms have been studied. For static graphs, [13] established the lower bounds and introduced the optimal primal algorithm. Later, [14] introduced the *decentralized communication sliding* method, which also achieves optimal complexity. In the case of time-varying graphs, [15] obtained the associated lower bounds and developed an optimal primal algorithm.

3.4 Problems with affine constraints

We introduce the decentralized optimization problem with affine constraints:

$$\min_{x \in \mathcal{K}} f(x) = \frac{1}{n}\sum_{i=1}^{n} f_i(x) \quad \text{s.t.} \quad Ax = b, \ A \in \mathbb{R}^{m \times d}, \ \mathcal{K} \subseteq \mathbb{R}^d. \tag{4}$$

In the time-static case a range of algorithms was proposed in [16, 17, 18, 19, 20]. The algorithms are based either on dual decomposition or on primal approaches to affinely constrained minimization. For time-varying architectures, [21] established the lower bounds and introduced an optimal dual algorithm. However, the design of optimal primal algorithms for time-varying architectures remains an open problem. Extending these results to non-smooth objectives also represents an important open research direction.

Setting	Complexity type	Static networks	Time-varying networks
Strongly convex	Decentralized communication	$\sqrt{\chi}\frac{M}{\sqrt{\mu\varepsilon}}$	$\chi\frac{M}{\sqrt{\mu\varepsilon}}$
	Subgradient computation	$\frac{M^2}{\mu\varepsilon}$	$\frac{M^2}{\mu\varepsilon}$
Convex	Decentralized communication	$\sqrt{\chi}\frac{MR}{\varepsilon}$	$\chi\frac{MR}{\varepsilon}$
	Subgradient computation	$\frac{M^2R^2}{\varepsilon^2}$	$\frac{M^2R^2}{\varepsilon^2}$
Paper		[13], [14]	[15]

Table 2: Convergence rates for non-smooth optimization

We consider an additional area of interest, which is the decentralized optimization with coupled constraints

$$\min_{x_1\in\mathbb{R}^{d_1},\ldots,x_n\in\mathbb{R}^{d_n}} f(x)=\sum_{i=1}^{n} f_i(x) \quad \text{s.t.} \quad \sum_{i=1}^{n}(A_ix_i-b_i)=0, \tag{5}$$

where $A_i\in\mathbb{R}^{m\times d_i}$ and $b_i\in\mathbb{R}^m$. We assume that the functions f_i, matrices A_i, and vectors b_i are stored locally by the nodes of a computational network. Some progress has been made in [22], but generalizing these results to time-varying networks and/or nonsmooth objectives remains an open challenge.

4 Saddle Point Problems

This paper studies min-max optimization problems of type

$$\min_{x\in\mathcal{X}}\max_{y\in\mathcal{Y}} f(x,y):=\frac{1}{n}\sum_{i=1}^{n} f_i(x,y), \tag{6}$$

where functions $f_i(x,y)$ are convex in x and concave in y and the sets $\mathcal{X}\subseteq\mathbb{R}^{n_x}$ and $\mathcal{Y}\subseteq\mathbb{R}^{n_y}$ are closed convex sets.

Saddle point optimization is widely used in real-world applications that involve competing objectives. In machine learning, it powers methods like adversarial training and Generative Adversarial Networks (GANs), balancing tasks such as minimizing loss and maximizing robustness. In distributed systems, it helps with resource allocation, load balancing, and network control.

4.1 Notations and Assumptions

Smoothness and strong convexity.

Additionally, we introduce the set $\mathcal{Z}=\mathcal{X}\times\mathcal{Y}\subseteq\mathbb{R}^{n_z}$, $z=(x,y)$, $n_z=n_x+n_y$, and the operator F:

$$F_i(z)=F_i(x,y)=\begin{pmatrix}\nabla_x f_i(x,y)\\ -\nabla_y f_i(x,y)\end{pmatrix},\quad F(z)=\frac{1}{n}\sum_{m=1}^{n}F_i(z).$$

Assumption 7. Let functions $f(x,y)$ and $f_m(x,y)$ satisfy following properties:

1. Function $f(x,y)$ is L-smooth, i.e. for all $z_1,z_2\in\mathcal{Z}$ it holds
$$\|F(z_1)-F(z_2)\|\le L\|z_1-z_2\|.$$
2. For all i, $f_i(x,y)$ is $L_{\max}$-smooth, i.e. for all $z_1,z_2\in\mathcal{Z}$ it holds
$$\|F_i(z_1)-F_i(z_2)\|\le L_{\max}\|z_1-z_2\|.$$
3. Function $f(x,y)$ is strongly-convex-strongly-concave with constant μ, i.e. for all $z_1,z_2\in\mathcal{Z}$ it holds
$$\langle F(z_1)-F(z_2),z_1-z_2\rangle\ge\mu\|z_1-z_2\|^2.$$

4.2 Lower Bounds and Optimal Algorithms for Smooth and Strongly Convex-Strongly Concave Saddle Point Problems

The results for decentralized saddle point problems are analogical to optimization. Lower bounds for min-max optimization over static graphs were given in [23]. The same paper [23] proposed algorithms optimal up to a logarithmic term. The case of (arbitrary) time-varying graphs was studied in [24] along with methods optimal up to a logarithmic factor. Optimal algorithms for sum-type variational inequalities (a generalization of saddle point problem) were proposed in [25]. Lower bounds for saddle point problems over slowly time-varying graphs (only two edge changes per iteration) were provided in [26]. The corresponding results are presented in Table 3. It is worth noting that the complexity are same as for optimization (Table 1), except for replacing $\sqrt{L/\mu}$ by L/μ.

	static	time-var.	slowly time-var.
comm.	$\sqrt{\chi}\frac{L}{\mu}\log\frac{1}{\varepsilon}$	$\chi\frac{L}{\mu}\log\frac{1}{\varepsilon}$	$\chi\frac{L}{\mu}\log\frac{1}{\varepsilon}$
oracle	$\frac{L}{\mu}\log\frac{1}{\varepsilon}$	$\frac{L}{\mu}\log\frac{1}{\varepsilon}$	$\frac{L}{\mu}\log\frac{1}{\varepsilon}$
paper	[23], [25]	[24], [25]	[26], [25]

Table 3: Convergence rates for smooth and strongly convex-strongly concave saddle point problems

4.3 Non-smooth Decentralized Saddle Point Problems

Non-smooth min-max problems can be solved with a subgradient method [27]. Other algorithms have been proposed to solve non-smooth decentralized saddle point problems, for example, [28] applied the *mirror-prox sliding* method for solving the composite problem obtained from penalization for the consensus affine constraint. However, they are not optimal in communications (even lower bounds have not been established) and therefore non-smooth decentralized saddle point problems are an open direction of research for both static and time-varying graphs. Extending saddle point problems to include affine and coupled constraints represents a valuable direction for further study.

5 Conclusions

Decentralized optimization finds applications in various domains, including distributed control and sensing, large-scale machine learning, and distributed statistical inference. One of the key challenges in this field is developing methods that can handle changes in network structure. There is still a lack of comprehensive studies addressing the non-smooth case, leaving an important gap in the current research landscape. This paper reviews existing results and techniques and highlights several open questions to guide future research.

References

[1] Pedro A Forero, Alfonso Cano, and Georgios B Giannakis. Consensus-based distributed linear support vector machines. In *Proceedings of the 9th ACM/IEEE international conference on information processing in sensor networks*, pages 35–46, 2010.

[2] Angelia Nedić, Alex Olshevsky, and César A Uribe. Fast convergence rates for distributed non-bayesian learning. *IEEE Transactions on Automatic Control*, 62(11):5538–5553, 2017.

[3] Sundhar Srinivasan Ram, Venugopal V Veeravalli, and Angelia Nedic. Distributed non-autonomous power control through distributed convex optimization. In *IEEE INFOCOM 2009*, pages 3001–3005. IEEE, 2009.

[4] Lingwen Gan, Ufuk Topcu, and Steven H Low. Optimal decentralized protocol for electric vehicle charging. *IEEE Transactions on Power Systems*, 28(2):940–951, 2012.

[5] Alex Olshevsky. Efficient information aggregation strategies for distributed control and signal processing. *arXiv preprint arXiv:1009.6036*, 2010.

[6] Juan Andrés Bazerque and Georgios B Giannakis. Distributed spectrum sensing for cognitive radio networks by exploiting sparsity. *IEEE Transactions on Signal Processing*, 58(3):1847–1862, 2009.

[7] 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.

[8] Dmitry Kovalev, Elnur Gasanov, Alexander Gasnikov, and Peter Richtarik. Lower bounds and optimal algorithms for smooth and strongly convex decentralized optimization over time-varying networks. *Advances in Neural Information Processing Systems*, 34, 2021.

[9] Dmitry Kovalev, Adil Salim, and Peter Richtárik. Optimal and practical algorithms for smooth and strongly convex decentralized optimization. *Advances in Neural Information Processing Systems*, 33:18342–18352, 2020.

[10] Dmitry Kovalev, Egor Shulgin, Peter Richtárik, Alexander V Rogozin, and Alexander Gasnikov. Adom: accelerated decentralized optimization method for time-varying networks. In *International Conference on Machine Learning*, pages 5784–5793. PMLR, 2021.

[11] Dmitry Metelev, Alexander Rogozin, Dmitry Kovalev, and Alexander Gasnikov. Is consensus acceleration possible in decentralized optimization over slowly time-varying networks? In *International Conference on Machine Learning*, pages 24532–24554. PMLR, 2023.

[12] Dmitry Metelev, Aleksandr Beznosikov, Alexander Rogozin, Alexander Gasnikov, and Anton Proskurnikov. Decentralized optimization over slowly time-varying graphs: Algorithms and lower bounds. *arXiv preprint arXiv:2307.12562*, 2023.

[13] Kevin Scaman, Francis Bach, Sébastien Bubeck, Laurent Massoulié, and Yin Tat Lee. Optimal algorithms for non-smooth distributed optimization in networks. *Advances in Neural Information Processing Systems*, 31, 2018.

[14] Guanghui Lan, Soomin Lee, and Yi Zhou. Communication-efficient algorithms for decentralized and stochastic optimization. *Mathematical Programming*, 180(1):237–284, 2020.

[15] Dmitry Kovalev, Ekaterina Borodich, Alexander Gasnikov, and Dmitrii Feoktistov. Lower bounds and optimal algorithms for non-smooth convex decentralized optimization over time-varying networks. *arXiv preprint arXiv:2405.18031*, 2024.

[16] Ion Necoara and Valentin Nedelcu. Distributed dual gradient methods and error bound conditions. *arXiv preprint arXiv:1401.4398*, 2014.

[17] Ion Necoara and Valentin Nedelcu. On linear convergence of a distributed dual gradient algorithm for linearly constrained separable convex problems. *Automatica*, 55:209–216, 2015.

[18] Ion Necoara, Valentin Nedelcu, and Ioan Dumitrache. Parallel and distributed optimization methods for estimation and control in networks. *Journal of Process Control*, 21(5):756–766, 2011.

[19] Alexander Rogozin, Demyan Yarmoshik, Ksenia Kopylova, and Alexander Gasnikov. Decentralized strongly-convex optimization with affine constraints: Primal and dual approaches. In *International Conference on Optimization and Applications*, pages 93–105. Springer, 2022.

[20] Demyan Yarmoshik, Alexander Rogozin, Oleg O Khamisov, Pavel Dvurechensky, and Alexander Gasnikov. Decentralized convex optimization under affine constraints for power systems control. In *International Conference on Mathematical Optimization Theory and Operations Research*, pages 62–75. Springer, 2022.

[21] Demyan Yarmoshik, Alexander Rogozin, and Alexander Gasnikov. Decentralized optimization with affine constraints over time-varying networks. *Computational Management Science*, 21(1):10, 2024.

[22] Demyan Yarmoshik, Alexander Rogozin, Nikita Kiselev, Daniil Dorin, Alexander Gasnikov, and Dmitry Kovalev. Decentralized optimization with coupled constraints. *arXiv preprint arXiv:2407.02020*, 2024.

[23] Aleksandr Beznosikov, Valentin Samokhin, and Alexander Gasnikov. Distributed saddle-point problems: Lower bounds, optimal algorithms and federated gans. *arXiv preprint arXiv:2010.13112*, 2021.

[24] Aleksandr Beznosikov, Alexander Rogozin, Dmitry Kovalev, and Alexander Gasnikov. Near-optimal decentralized algorithms for saddle point problems over time-varying networks. In *International Conference on Optimization and Applications*, pages 246–257. Springer, 2021.

[25] Dmitry Kovalev, Aleksandr Beznosikov, Abdurakhmon Sadiev, Michael Persiianov, Peter Richtárik, and Alexander Gasnikov. Optimal algorithms for decentralized stochastic variational inequalities. *Advances in Neural Information Processing Systems*, 35:31073–31088, 2022.

[26] Nhat Trung Nguyen, Alexander Rogozin, Dmitry Metelev, and Alexander Gasnikov. Min-max optimization over slowly time-varying graphs. In *Doklady Mathematics*, volume 108, pages S300–S309. Springer, 2023.

[27] Angelia Nedić and Asuman Ozdaglar. Subgradient methods for saddle-point problems. *Journal of optimization theory and applications*, 142:205–228, 2009.

[28] Ilya Kuruzov, Alexander Rogozin, Demyan Yarmoshik, and Alexander Gasnikov. The mirror-prox sliding method for non-smooth decentralized saddle-point problems. *arXiv preprint arXiv:2210.06086*, 2022.