Distributed optimization

Yandex School Spring 2025

Demyan Yarmoshik, Alexander Rogozin

Moscow Institute of Physics and Technology

April 16, 2025

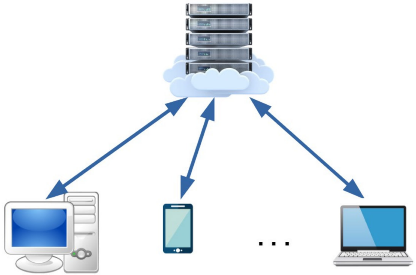
Introduction

Why distributed optimization is needed?

- Large amounts of data.
- Distributed nature of data generation and acquisition.
- Privacy constraints.

Applications:

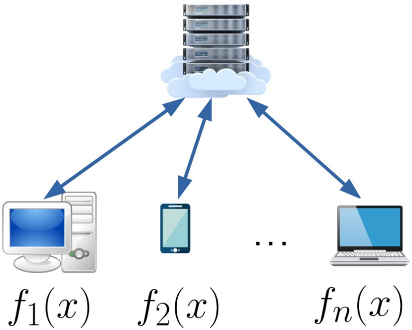
- Distributed
vehicle coordination and control Ren and Beard (2008).
- Power system control Ram et al. (2009).
- Large-scale statistical inference and machine learning Rabbat and Nowak (2004).
- Federated learning Konečný et al. (2016).
- Distributed tracking Granichin and Amelina (2014).
- Formation control Ren (2006).
- Distributed load balancing Amelina et al. (2015).



Centralized optimization

Each node locally holds function f_i and can perform local computations. The agents aim to solve a sum-type problem

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



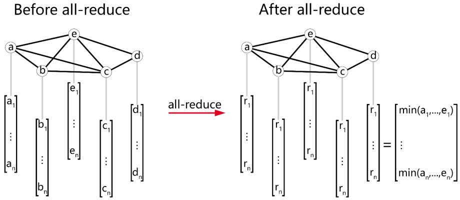
Case of ML.
Model weights: $x \in \mathbb{R}^d$, dataset parts: i, loss functions: f_i.

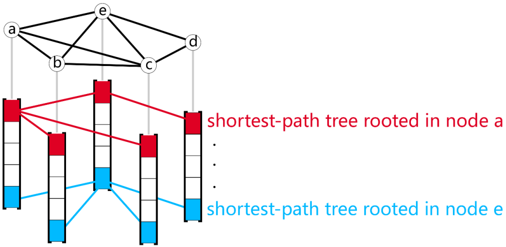
Agents can be represented as
- Computers.
- Nodes of the computing cluster.
- Drones, satellites, unmanned vehicles.
- Smartphones.
- Sensors.

All-Reduce protocol

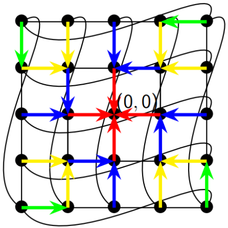
All-Reduce can be used in centralized distributed optimization for better communication efficiency.

The vectors are
split into chunks, each chunk has its own “all-reduce path”.





Different shortest-path trees



A shortest-path tree on torus

Decentralized consensus optimization

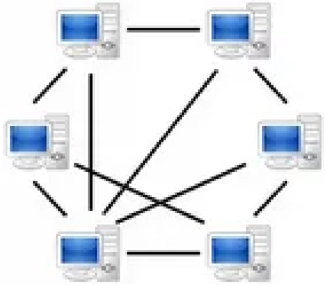
The group of nodes is not coordinated by any centralized server. Each node locally holds f_i and exchanges information only with its immediate neighbors.

$$\min_{x_1,\ldots,x_n \in \mathbb{R}^d} \sum_{i=1}^{n} f_i(x_i)$$
$$\text{s.t. } x_1 = \ldots = x_n.$$

- Agents are not synchronized; each one has its personal optimization trajectory.
- The agents need to maintain approximate consensus.

The optimal point in the decentralized sense should be consensual and optimal, i.e.

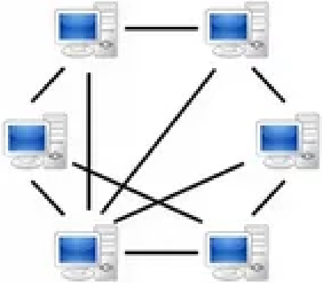
$$x_1 = \ldots = x_n = x^* = \operatorname*{argmin}_{x \in \mathbb{R}^d} \frac{1}{n} \sum_{i=1}^{n} f_i(x).$$



Coupled constraints optimization

Each node locally holds f_i and constraint matrix A_i.

$$\begin{aligned} \min_{x_1 \in \mathbb{R}^{d_1}, \dots, x_n \in \mathbb{R}^{d_n}} \quad & \sum_{i=1}^{n} f_i(x_i) \\ \text{s.t.} \quad & \sum_{i=1}^{n} (A_i x_i - b_i) = 0. \end{aligned}$$



- Agent local vectors are tied by distributed affine constraints.
- Consensus optimization is a special case of coupled constraints.

Federated learning is used in analysis of locally held user data.

- Local user devices (smartphones, laptops, etc.) hold user data.
- We need to train a model over all the dataset.
- Personal data should not leave the user's device.

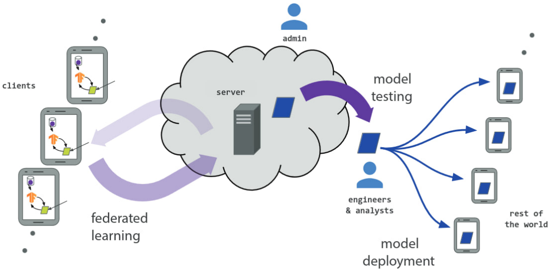


Рис.: Federated learning Kairouz et al. (2021)

Energy system control

Energy system consists of several areas that are controlled by sensors.

- Centralized data aggregation is unavailable.
- Each sensor only captures a part of the network.
- Areas corresponding to sensors have common points.
- Intersections of network parts is represented as coupled constraints.

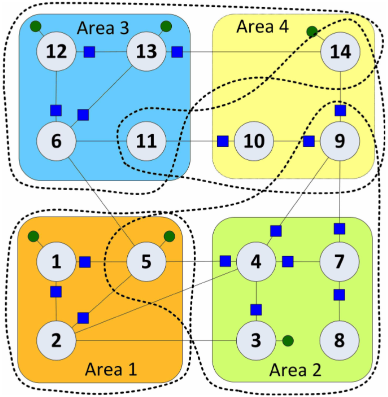


Рис.: Areas corresponding to one measurement device are circled by dotted lines Kekatos et al. (2020)

Distributed vehicle coordination

Consider a group of autonomous vehicles that act collectively.

- Each of the devices can communicate to others.
- The connection is wireless and is established only if the agents are within some radius.
- The central server is not present.
- The aim is to collectively solve a problem.

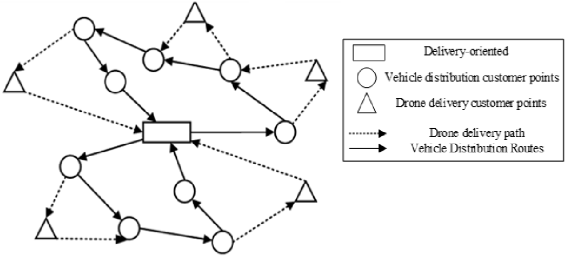


Рис.: Distributed vehicle coordination Li et al. (2022)

Question 1: how to solve decentralized optimization?

Assumption (1)

Consider undirected graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$. Gossip matrix $W \in \mathbb{R}^{m \times m}$ has the following properties.

- *(Decentralized property) If $(i, j) \notin \mathcal{E}$, then $[W]_{ij} = 0$.*
- *(Symmetry and positive semi-definiteness) $W = W^\top$ and $W \succeq 0$.*
- *(Kernel property) $Wx = 0$ if and only if $x_1 = \ldots = x_n$.*
- *(Contraction property) There exists $\chi > 1$ such that*

$$\left\| Wx - x \right\|_2 \leq (1 - \chi^{-1}) \left\| x \right\|_2 \text{ for all } x \in \operatorname{Im} W, \text{ i.e. } x_1 + \ldots + x_n = 0.$$

Example of gossip matrix: $W = \frac{L(\mathcal{G})}{\lambda_{\max}(L(\mathcal{G}))}$, where $L(\mathcal{G}) = D(\mathcal{G}) - A(\mathcal{G})$ denotes the graph Laplacian.

- For consensus optimization $x_1 = \ldots = x_n$, we write $\boldsymbol{W}\boldsymbol{x} = 0$, where $\boldsymbol{W} = W \otimes \mathrm{I}_d$, $\boldsymbol{x} = \mathrm{col}(x_1, \ldots, x_n)$.
- For coupled constraints $\sum_{i=1}^n (A_i x_i - b_i) = 0$, we write $\boldsymbol{A}\boldsymbol{x} + \boldsymbol{W}\boldsymbol{y} - \boldsymbol{b} = 0$, where $\boldsymbol{A} = \mathrm{diag}(A_1, \ldots, A_n)$, $\boldsymbol{W} = W \otimes \mathrm{I}_d$, $\boldsymbol{b} = \mathrm{col}(b_1, \ldots, b_n)$.

Question 1: how to solve decentralized optimization?

We come to affinely constrained optimization: $\min_{u \in \mathbb{R}^p} \; G(u)$ s.t. $\boldsymbol{B} u = c$.

- Consensus optimization: $u = \boldsymbol{x}, \; p = nd, \; G(u) = F(\boldsymbol{x}), \; \boldsymbol{B} = \boldsymbol{W}, \; c = 0$.
- Coupled constraints optimization:

$u = (\boldsymbol{x}, \; \boldsymbol{y}), \; p = 2(d_1 + \ldots + d_n), \; G(u) = F(\boldsymbol{x}), \; \boldsymbol{B} = [\boldsymbol{A} \; \boldsymbol{W}], \; c = \boldsymbol{b}$.

Algorithm APAPC

Parameters: $u^0 \in \mathbb{R}^d \; \eta, \theta, \alpha > 0, \; \tau \in (0, 1)$

Set $u_f^0 = u^0, \; z^0 = 0 \in \mathbb{R}^d$

for $k = 0, 1, 2, \ldots$ **do**

$$u_g^k \overset{\text{def}}{=} \tau u^k + (1 - \tau) u_f^k$$

$$u^{k+\frac{1}{2}} \overset{\text{def}}{=} (1 + \eta\alpha)^{-1}(u^k - \eta(\nabla G(u_g^k) - \alpha u_g^k + z^k))$$

$$z^{k+1} \overset{\text{def}}{=} z^k + \theta \boldsymbol{B}^\top (\boldsymbol{B} u^{k+\frac{1}{2}} - c)$$

$$u^{k+1} \overset{\text{def}}{=} (1 + \eta\alpha)^{-1}(u^k - \eta(\nabla G(u_g^k) - \alpha u_g^k + z^{k+1}))$$

$$u_f^{k+1} \overset{\text{def}}{=} u_g^k + \tfrac{2\tau}{2-\tau}(u^{k+1} - u^k)$$

end for

Question 2: how efficient is the method?

Assumption (2)

Function $G(u)$ is μ-strongly convex and L-smooth, i.e. for any $u, v \in \mathbb{R}^p$ it holds

$$\frac{\mu}{2}\|u - v\|_2^2 \leq G(v) - G(u) + \langle \nabla G(u), v - u \rangle \leq \frac{L}{2}\|u - v\|_2^2 .$$

We introduce

$$\kappa_{\boldsymbol{B}} = \frac{\lambda_{\max}(\boldsymbol{B}^\top \boldsymbol{B})}{\lambda_{\min^+}(\boldsymbol{B}^\top \boldsymbol{B})}.$$

Theorem

Let Assumption 2 hold. There exists a set of parameters for Algorithm 1 such that to yield x^N satisfying $\left\|x^N - x^*\right\|_2^2 \leq \varepsilon$ it requires $N = O(\kappa_{\boldsymbol{B}}\sqrt{L/\mu}\log(1/\varepsilon))$ iterations.

Question 2: how efficient is the method?

Prob.	Compl.	Complexity	Lower bound
Consensus optim.	Grad.	$\sqrt{\frac{L}{\mu}}\log\left(\frac{1}{\varepsilon}\right)$	$\sqrt{\frac{L}{\mu}}\log\left(\frac{1}{\varepsilon}\right)$
	Comm.	$\sqrt{\kappa_W}\sqrt{\frac{L}{\mu}}\log\left(\frac{1}{\varepsilon}\right)$	$\sqrt{\kappa_W}\sqrt{\frac{L}{\mu}}\log\left(\frac{1}{\varepsilon}\right)$
	Paper	S. et al., 2017	S. et al., 2017
Coupled constr.	Grad.	$\sqrt{\frac{L}{\mu}}\log\left(\frac{1}{\varepsilon}\right)$	$\sqrt{\frac{L}{\mu}}\log\left(\frac{1}{\varepsilon}\right)$
	Mat.	$\sqrt{\kappa_A}\sqrt{\frac{L}{\mu}}\log\left(\frac{1}{\varepsilon}\right)$	$\sqrt{\kappa_A}\sqrt{\frac{L}{\mu}}\log\left(\frac{1}{\varepsilon}\right)$
	Comm.	$\sqrt{\kappa_W}\sqrt{\kappa_A}\sqrt{\frac{L}{\mu}}\log\left(\frac{1}{\varepsilon}\right)$	$\sqrt{\kappa_W}\sqrt{\kappa_A}\sqrt{\frac{L}{\mu}}\log\left(\frac{1}{\varepsilon}\right)$
	Paper	Y. et al., 2024	Y. et al., 2024

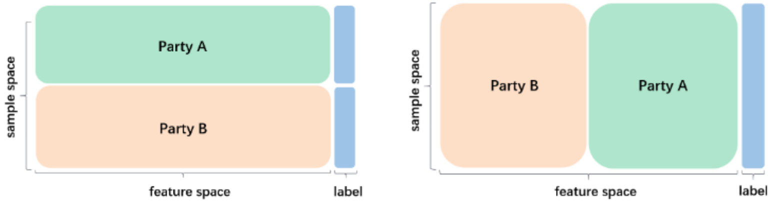
Таблица: Convergence rates for decentralized smooth optimization.

Project: vertical federated learning

Let F be the matrix of features, split vertically between compute nodes into submatrices F_i, so that each node possesses its own subset of features for all data samples. Let $l \in \mathbb{R}^m$ denote the vector of labels, and let $x_i \in \mathbb{R}^{d_i}$ be the vector of model parameters owned by the i-th node. VFL problem formulates as

$$\min_{\substack{z \in \mathbb{R}^m \\ x_1 \in \mathbb{R}^{d_1}, \dots, x_n \in \mathbb{R}^{d_n}}} \ell(z, l) + \sum_{i=1}^{n} r_i(x_i) \quad \text{s.t.} \quad \sum_{i=1}^{n} F_i x_i = z, \tag{1}$$

where ℓ is a loss function, and r_i are regularizers.



(a) Horizontal Federated Learning (b) Vertical Federated Learning

References I

Amelina, N., Fradkov, A., Jiang, Y., and Vergados, D. J. (2015). Approximate consensus in stochastic networks with application to load balancing. *IEEE Transactions on Information Theory*, 61(4):1739–1752.

Granichin, O. and Amelina, N. (2014). Simultaneous perturbation stochastic approximation for tracking under unknown but bounded disturbances. *IEEE Transactions on Automatic Control*, 60(6):1653–1658.

Kairouz, P., McMahan, H. B., Avent, B., Bellet, A., Bennis, M., Bhagoji, A. N., Bonawitz, K., Charles, Z., Cormode, G., Cummings, R., et al. (2021). Advances and open problems in federated learning. *Foundations and Trends® in Machine Learning*, 14(1–2):1–210.

Kekatos, V., Wang, G., Zhu, H., and Giannakis, G. B. (2020). Psse redux.

Konečný, J., McMahan, H. B., Yu, F. X., Richtárik, P., Suresh, A. T., and Bacon, D. (2016). Federated learning: Strategies for improving communication efficiency. *arXiv preprint arXiv:1610.05492*.

Li, J., Liu, H., Lai, K. K., and Ram, B. (2022). Vehicle and uav collaborative delivery path optimization model. *Mathematics*, 10(20):3744.

References II

Rabbat, M. and Nowak, R. (2004). Distributed optimization in sensor networks. In *Proceedings of the 3rd international symposium on Information processing in sensor networks*, pages 20–27.

Ram, S. S., Veeravalli, V. V., and Nedic, A. (2009). Distributed non-autonomous power control through distributed convex optimization. In *IEEE INFOCOM 2009*, pages 3001–3005. IEEE.

Ren, W. (2006). Consensus based formation control strategies for multi-vehicle systems. In *2006 American Control Conference*, pages 6–pp. IEEE.

Ren, W. and Beard, R. W. (2008). *Distributed consensus in multi-vehicle cooperative control*, volume 27. Springer.

Scaman, K., Bach, F., Bubeck, S., Lee, Y. T., and Massoulié, L. (2017). Optimal algorithms for smooth and strongly convex distributed optimization in networks. In *International Conference on Machine Learning*, pages 3027–3036.

Yarmoshik, D., Rogozin, A., Kiselev, N., Dorin, D., Gasnikov, A., and Kovalev, D. (2024). Decentralized optimization with coupled constraints. *arXiv preprint arXiv:2407.02020*.