ROYAL INSTITUTE
OF TECHNOLOGY

Exponential Convergence for Distributed Optimization Under the Restricted Secant Inequality Condition

Xinlei Yi

Joint work with Shengjun Zhang, Tao Yang, Tianyou Chai, and Karl H. Johansson

July 12, 2020

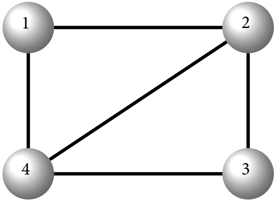
School of Electrical Engineering and Computer Science
KTH Royal Institute of Technology
Stockholm, Sweden

Distributed optimization

A network of agents cooperatively solve a global optimization problem, where

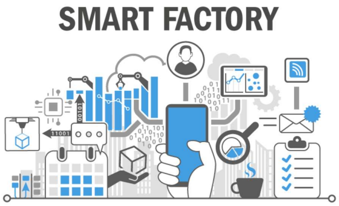
$$\min_{x \in \mathbb{R}^p} f(x) = \sum_{i=1}^{n} f_i(x).$$

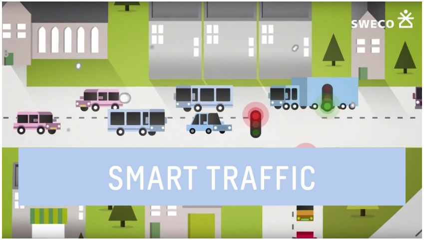
- each agent i has a local private objective $f_i(x)$
- all agents collaborate together to find the solution to minimize $f(x)$
- local information exchange via the underlying communication network

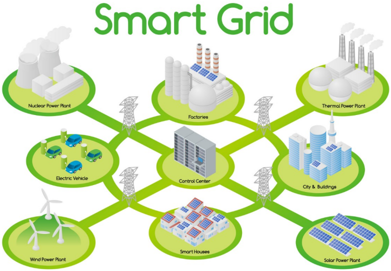


- An important component of many machine learning techniques with data parallelism, e.g., deep learning and federated learning

Applications







(Images are from the Internet.)

Motivation

Existing algorithms

Continuous- and discrete-time distributed algorithms

Existing result

A standard assumption for proving exponential/linear convergence of existing distributed algorithms is

strong convexity of the cost functions

Question

Could strong convexity be relaxed?
For example, quadratic functions may be not strongly convex.

Answer in our paper

Yes, it can be relaxed by the restricted secant inequality condition.

Restricted secant inequality condition (1/2)

Let $f(x):\ \mathbb{R}^p \mapsto \mathbb{R}$ be a differentiable function, $f^* = \min_{x\in\mathbb{R}^p} f(x)$, $\mathbb{X}^* = \arg\min_{x\in\mathbb{R}^p} f(x)$, and x_p is the projection of x onto the set $\mathbb{X}^*$.

- Strong Convexity (SC): For all x and y,

$$f(y) \geq f(x) + \langle \nabla f(x), y - x \rangle + \frac{\mu}{2}\|y - x\|^2$$

- Essential Strong Convexity (ESC): For all x and y with $x_p = y_p$,

$$f(y) \geq f(x) + \langle \nabla f(x), y - x \rangle + \frac{\mu}{2}\|y - x\|^2$$

- Weak Strong Convexity (WSC): For all x,

$$f^* \geq f(x) + \langle \nabla f(x), x_p - x \rangle + \frac{\mu}{2}\|x_p - x\|^2$$

- Restricted Secant Inequality (RSI): For all x

$$\langle \nabla f(x), x - x_p \rangle \geq \frac{\mu}{2}\|x_p - x\|^2$$

Restricted secant inequality condition (2/2)

A function f satisfies the RSI condition with constant $\mu > 0$ if

$$\langle \nabla f(x), x - x_p \rangle \geq \frac{\mu}{2} \|x_p - x\|^2$$

Remark 1

- SC $\Rightarrow$ ESC $\Rightarrow$ WSC $\Rightarrow$ RSI
- Every stationary point is a minimizer
- It does not imply that $\mathbb{X}^*$ is a singleton
- It does not imply convexity of f
- It is difficult to verify this condition

One special case

Let $f(x) = g(Ax)$, where $g : \mathbb{R}^p \to \mathbb{R}$ is a strongly convex function and $A \in \mathbb{R}^{p \times p}$ is a matrix, then f satisfies the RSI condition.

Algorithm description (1/2)

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

Each f_i is smooth, f satisfies the RSI condition, and $\mathcal{G}$ is connected

- It is equivalent to the following constrained optimization problem:

$$\min_{\boldsymbol{x} \in \mathbb{R}^{np}} \quad \tilde{f}(\boldsymbol{x}) = \sum_{i=1}^{n} f_i(x_i)$$

$$\text{subject to} \quad \boldsymbol{L}^{1/2}\boldsymbol{x} = \mathbf{0}_{np}, \ (\Leftrightarrow x_i = x_j, \ \forall i, j \in [n])$$

  where $\boldsymbol{x} = (x_1, \dots, x_n)$ and $\boldsymbol{L} = L \otimes \mathbf{I}_p$.
- The associated augmented Lagrangian function is

$$\mathcal{A}(\boldsymbol{x}, \boldsymbol{u}) = \tilde{f}(\boldsymbol{x}) + \frac{\alpha}{2}\|\boldsymbol{L}^{1/2}\boldsymbol{x}\|^2 + \beta \boldsymbol{u}^\top \boldsymbol{L}^{1/2}\boldsymbol{x},$$

  where $\boldsymbol{u}$ is the dual variable, $\alpha > 0$ and $\beta > 0$ are parameters to be designed later.

Algorithm description (2/2)

$$\mathcal{A}(\boldsymbol{x}, \boldsymbol{u}) = \tilde{f}(\boldsymbol{x}) + \frac{\alpha}{2}\boldsymbol{x}^\top \boldsymbol{L}\boldsymbol{x} + \beta \boldsymbol{u}^\top \boldsymbol{L}^{1/2}\boldsymbol{x}.$$

- A continuous-time distributed primal-dual algorithm is

$$\dot{\boldsymbol{x}}(t) = -\frac{\partial \mathcal{A}(\boldsymbol{x}(t), \boldsymbol{u}(t))}{\boldsymbol{x}} = -\alpha \boldsymbol{L}\boldsymbol{x}(t) - \beta \boldsymbol{L}^{1/2}\boldsymbol{u}(t) - \nabla \tilde{f}(\boldsymbol{x}(t)),$$
$$\dot{\boldsymbol{u}}(t) = \frac{\partial \mathcal{A}(\boldsymbol{x}(t), \boldsymbol{u}(t))}{\boldsymbol{u}} = \beta \boldsymbol{L}^{1/2}\boldsymbol{x}(t),\ \forall \boldsymbol{x}(0),\ \boldsymbol{u}(0) \in \mathbb{R}^{np}.$$

- Denote $\boldsymbol{v}(t) = \boldsymbol{L}^{1/2}\boldsymbol{u}(t)$. Then,

$$\dot{\boldsymbol{x}}(t) = -\alpha \boldsymbol{L}\boldsymbol{x}(t) - \beta \boldsymbol{v}(t) - \nabla \tilde{f}(\boldsymbol{x}(t)),$$
$$\dot{\boldsymbol{v}}(t) = \beta \boldsymbol{L}\boldsymbol{x}(t),\ \forall \boldsymbol{x}(0) \in \mathbb{R}^{np},\ \sum_{j=1}^{n} v_j(0) = \boldsymbol{0}_p.$$

Algorithm extension

- Special initialization:

$$\dot{x}_i(t) = -\alpha \sum_{j=1}^{n} L_{ij} x_j(t) - \beta v_i(t) - \nabla f_i(x_i(t)),$$

$$\dot{v}_i(t) = \beta \sum_{j=1}^{n} L_{ij} x_j(t), \ \forall x_i(0) \in \mathbb{R}^p, \ \sum_{j=1}^{n} v_j(0) = \mathbf{0}_p.$$

$(v_i(0) = \mathbf{0}_p, \ \forall i \in [n]$, or $v_i(0) = \sum_{j=1}^{n} L_{ij} x_j(0), \ \forall i \in [n])$

- Arbitrary initialization:

$$\dot{x}_i(t) = -\alpha \sum_{j=1}^{n} L_{ij} x_j(t) - \beta \sum_{j=1}^{n} L_{ij} v_j(t) - \nabla f_i(x_i(t)),$$

$$\dot{v}_i(t) = \beta \sum_{j=1}^{n} L_{ij} x_j(t), \ \forall x_i(0), v_i(0) \in \mathbb{R}^p.$$

(Additional communication of $v_j(t)$)

Convergence analysis

Theorem 1

If each f_i is smooth, f satisfies the RSI condition with constant $\mu > 0$, and $\mathcal{G}$ is connected, then $\sum_{i=1}^{n} \|x_i(t) - \mathcal{P}_{\mathbb{X}^*}(\bar{x}(t))\|$ exponentially converges to 0, where $\bar{x}(t) = \frac{1}{n}\sum_{i=1}^{n} x_i(t)$.

- Contribution: exponential convergence without strong convexity, even without convexity
- Potential drawback: the constant μ is used to choose the parameter α

Algorithm description

- Recall continuous-time distributed primal-dual algorithm

$$\dot{x}_i(t) = -\alpha \sum_{j=1}^{n} L_{ij} x_j(t) - \beta v_i(t) - \nabla f_i(x_i(t)),$$

$$\dot{v}_i(t) = \beta \sum_{j=1}^{n} L_{ij} x_j(t), \ \forall x_i(0) \in \mathbb{R}^p, \ \sum_{j=1}^{n} v_j(0) = \mathbf{0}_p.$$

$\dot{y}(t) \approx \frac{y(t+h)-y(t)}{h}$ (Euler's approximation method) $\Rightarrow$

- Discrete-time distributed primal-dual algorithm

$$x_i(k+1) = x_i(k) - h(\alpha \sum_{j=1}^{n} L_{ij} x_j(k) + \beta v_i(k) + \nabla f_i(x_i(k))),$$

$$v_i(k+1) = v_i(k) + h\beta \sum_{j=1}^{n} L_{ij} x_j(k), \ \forall x_i(0) \in \mathbb{R}^p, \ \sum_{j=1}^{n} v_j(0) = \mathbf{0}_p,$$

where $h > 0$ is a fixed stepsize.

Algorithm comparison

- Discrete-time distributed primal-dual algorithm

$$x_i(k+1) = x_i(k) - h(\alpha \sum_{j=1}^{n} L_{ij} x_j(k) + \beta v_i(k) + \nabla f_i(x_i(k))),$$

$$v_i(k+1) = v_i(k) + h\beta \sum_{j=1}^{n} L_{ij} x_j(k), \ \forall x_i(0) \in \mathbb{R}^p, \ \sum_{j=1}^{n} v_j(0) = \mathbf{0}_p.$$

- Distributed gradient tracking algorithm

$$x_i(k+1) = \sum_{j=1}^{n} W_{ij} x_j(k) - h s_i(k), \ \forall x_i(0) \in \mathbb{R}^p, \ s_i(0) = \nabla f_i(x_i(0)),$$

$$s_i(k+1) = \sum_{j=1}^{n} W_{ij} s_j(k) + \nabla f_i(x_i(k+1)) - \nabla f_i(x_i(k)).$$

(Additional communication of $s_j(k)$, and strong convexity is needed to show linear convergence.)

Convergence analysis

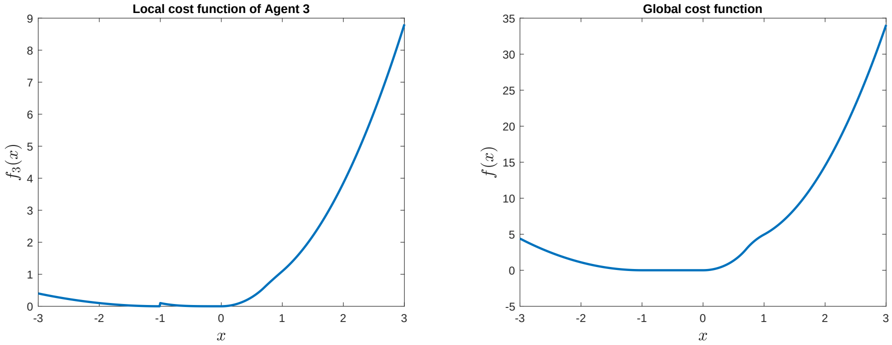
Theorem 2

If each f_i is smooth, f satisfies the RSI condition with constant $\mu > 0$, and $\mathcal{G}$ is connected, then $\sum_{i=1}^{n} \|x_i(k) - \mathcal{P}_{\mathbb{X}^*}(\bar{x}(k))\|$ linearly converges to 0, where $\bar{x}(k) = \frac{1}{n}\sum_{i=1}^{n} x_i(k)$.

- Contribution: linear convergence without strong convexity, even without convexity
- Potential drawback: the constant μ is used to choose the parameter α

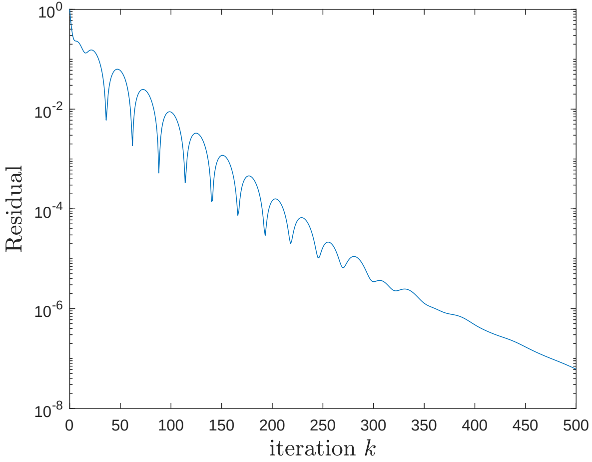
Simulation: settings

- Each $f_i : \mathbb{R} \mapsto \mathbb{R}$ is non-convex but differentiable and smooth
- $f(x) = \sum_{i=1}^{n} f_i(x)$ satisfies the RSI condition
- The optimal set $\mathbb{X}^* = [-1, 0]$
- A ring graph with $n = 10$ agents



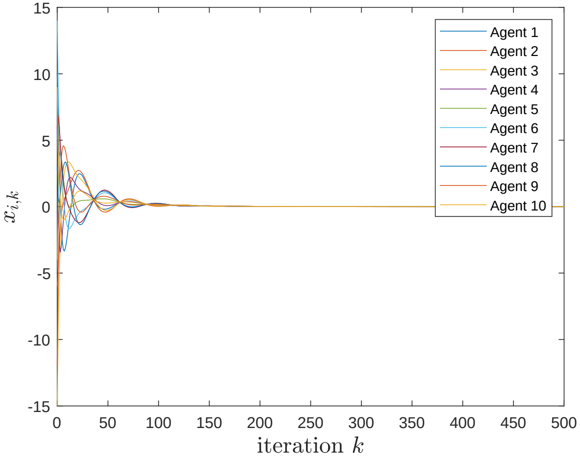
Simulation: evolutions of residual

Residual: $\sum_{i=1}^{n} \|x_i(k) - \mathcal{P}_{\mathbb{X}^*}(\bar{x}(k))\| / \sum_{i=1}^{n} \|x_i(0) - \mathcal{P}_{\mathbb{X}^*}(\bar{x}(0))\|$

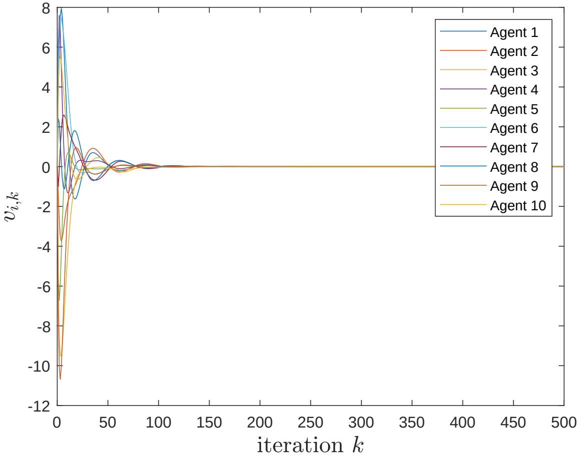


(Linear convergence is established.)

Simulation: evolutions of local primal variables



Simulation: evolutions of local dual variables



Conclusions

- Problem: distributed optimization
- Method: continuous- and discrete-time primal-dual algorithms
- Results: exponential/linear convergence under the RSI condition, which is weaker than strong convexity
- Extensions:
  - Overcoming the potential drawback (the constant μ is used)
  - Relaxing the RSI condition by the Polyak-Łojasiewicz condition
  - Considering communication efficiency: compression and quantization
  - Studying the scenarios where gradient information is unavaiable

References

[1] A. Nedić, "Convergence rate of distributed averaging dynamics and optimization in networks," Foundations and Trends in Systems and Control, vol. 2, no. 1, pp. 1–100, 2015.

[2] I. Necoara, Y. Nesterov, and F. Glineur, "Linear convergence of first order methods for non-strongly convex optimization," Mathematical Programming, vol. 175, no. 1–2, pp. 69–107, 2019.

[3] H. Karimi, J. Nutini, and M. Schmidt, "Linear convergence of gradient and proximal-gradient methods under the Polyak-Łojasiewicz condition," in Joint European Conference on Machine Learning and Knowledge Discovery in Databases, 2016, pp. 795–811.

[4] J. Lu and C. Y. Tang, "Zero-gradient-sum algorithms for distributed convex optimization: The continuous-time case," IEEE Transactions on Automatic Control, vol. 57, no. 9, pp. 2348–2354, 2012.

[5] S. S. Kia, J. Cortés, and S. Martínez, "Distributed convex optimization via continuous-time coordination algorithms with discrete-time communication," Automatica, vol. 55, pp. 254–264, 2015.

[6] S. Liang, L. Y. Wang, and G. Yin, "Exponential convergence of distributed primal-dual convex optimization algorithm without strong convexity," Automatica, vol. 105, pp. 298–306, 2019.

[7] Y. Tang, J. Zhang, and N. Li, "Distributed zero-order algorithms for nonconvex multi-agent optimization," arXiv preprint arXiv:1908.11444, 2019.

[8] X. Yi, S. Zhang, T. Yang, T. Chai, and K. H. Johansson, "Exponential convergence for distributed smooth optimization under the restricted secant inequality condition," arXiv preprint arXiv:1909.03282, 2019.

[9] X. Yi, S. Zhang, T. Yang, T. Chai, and K. H. Johansson, "Linear convergence of first- and zeroth-order primal-dual algorithms for distributed nonconvex optimization," arXiv preprint arXiv:1912.12110, 2019.

Thanks for your time!