SCIENCE CHINA
Information Sciences

• **Supplementary File** •

A zeroth-order Algorithm for distributed optimization with stochastic stripe observations

Yinghui WANG[1], Xianlin ZENG[2*], Wenxiao ZHAO[3] & Yiguang HONG[4]

[1]*School of Automation and Electrical Engineering,*
University of Science and Technology Beijing, Beijing 100083*, China.;*
[2]*Key Laboratory of Intelligent Control and Decision of Complex Systems,*
School of Automation, Beijing Institute of Technology, 100081*, Beijing, China;*
[3]*Key Laboratory of Systems and Control, Academy of Mathematics and Systems Science*
Chinese Academy of Sciences, Beijing, China;
[4]*Tongji University*

Appendix A Assumptions

Consider the following distributed optimization problem over an n-agent network [9]:

$$\begin{aligned} &\min_{x} \quad f(\boldsymbol{x}), \quad f(\boldsymbol{x}) = \sum_{i=1}^{n} f_i(x_i) \\ &s.\ t. \quad x_i = x_j . x \in X \end{aligned} \tag{A1}$$

where $\boldsymbol{x} = \left[x_1^\top, x_2^\top, \ldots, x_n^\top\right]^\top \in \mathcal{R}^{np}$, $x_i \in \mathcal{R}^p$, $f_i : \mathcal{R}^p \to \mathcal{R}$ is a convex function, X is the global constraint.

First, we provide an assumption about the communication topology between agents over the network.

Consider a time-varying multi-agent network. The communication topology between agents over the network is described by a directed graph $\mathcal{G}(k) = \big(\mathcal{N}, \mathcal{E}(k), W(k)\big)$, where $\mathcal{N} = \{1, 2, ...n\}$ is the agent set, $\mathcal{E}(k) \subset \mathcal{N}\times\mathcal{N}$ represents information communication links at time k, and $W(k) = \left[w_{ij}(k)\right]_{ij}$ represents the adjacency matrix at time k. In addition, denote $\mathcal{N}_i(k) = \{j|(i, j) \in \mathcal{E}(k)\}$ as the neighbors of agent i at time k. Each agent interacts with its neighbors in $\mathcal{G}(k) = (\mathcal{N}, \mathcal{E}(k), W(k))$ at time k. The following assumption is on the communication topology $\mathcal{G}(k) = \big(\mathcal{N}, \mathcal{E}(k), W(k)\big)$, which is widely used in distributed time-varying network designs ([8]. [9]).

Assumption 1. The graph $\mathcal{G}(k) = \big(\mathcal{N}, \mathcal{E}(k), W(k)\big)$ satisfies:
(a) There exists a constant η with $0 < \eta < 1$ such that, $\forall k \geqslant 0$ and $\forall i, j$, $w_{ii}(k) \geqslant \eta$; $w_{ij}(k) \geqslant \eta$ if $(j, i) \in \mathcal{E}(k)$.
(b) $W(k)$ is doubly stochastic, i. e. $\sum_{i=1}^{m} w_{ij}(k) = 1$ and $\sum_{j=1}^{m} w_{ij}(k) = 1$.
(c) There is an integer $\kappa \geqslant 1$ such that $\forall k \geqslant 0$ and $\forall (j, i) \in \mathcal{N} \times \mathcal{N}$, $(j, i) \in \mathcal{E}(k) \cup \mathcal{E}(k+1) \cup \cdots \cup \mathcal{E}(k+\kappa-1)$.

The following assumption holds for local functions and constraints of Problem (A1):

Assumption 2. (a) Problem (A1) has solutions.
(b) $f_{L_i}(x)$ and $f_{R_i}(x)$ are convex functions with $f_{L_i}(x) \leqslant f_{R_i}(x)$.
(c) X is a non-empty, compact, convex constraint set in $\mathcal{R}^p$.
(d) The gradients of $f_{L_i}(x)$ and $f_{R_i}(x)$ are locally Lipschitz continuous with constant L.

Then, we assume the following assumption holds for parameter λ_i of Problem (A1):

Assumption 3. There exists a common $\lambda_0 \in (0, 1)$, such that for all agents $i \in \mathcal{N}$,

$$g_i(x) = \lambda_0 f_{L_i}(x) + (1 - \lambda_0) f_{R_i}(x).$$

Assumption 3 gives the characteristics of stripe observations $Y_{L_i}(x)$ and $Y_{R_i}(x)$, which results from inherent errors of measuring devices or methods over the network. $(\hat{x}^*, \lambda_0)$ is the optimal solution to Problem (**??**), and $(\hat{x}^*, \lambda_0)$ is expected exact solution to Problem (A1) with stripe observations $Y_L(x)$ and $Y_R(x)$ under Assumption 3.

Next, we give different probabilistic choices of each agents' initial preferences $\lambda_i(0)$ for Algorithm 1.

Assumption 4. (a) Each agent i has an initial preference $\lambda_i(0)$ with $\lambda_0 = \sum_{i=1}^{n} \lambda_i(0)$.
(b) Each agent i has an initial preference $\lambda_i(0)$. $\lambda_i(0)$s are independent and identically distributed (i.i.d.) random variables with $\mathbb{E}\lambda_i(0) = \lambda_0$.
(c) Each agent i has an initial preference $\lambda_i(0)$. $\lambda_i(0)$s are independent and identically distributed (i.i.d.) random variables with $\mathbb{E}\lambda_i(0) = \lambda_0$, $var\lambda_i(0) = \nu^2$.

Assumption 4(a) is an ideal assumption that although each agent gets incomplete information of $\lambda_i(0)$, they could get the complete information of λ_0 through cooperation. Assumption 4(b) and (c) are general probabilistic assumptions of agents preferences.

Still, the following assumption holds for parameters $\left\{ \triangle_i^q(k)\right\}_{k\geqslant 0}$ of Algorithm 1:

Condition 1. (Parameter selection)

* Corresponding author (email: xianlin.zeng@bit.edu.cn)

(a) Let $\left\{\triangle_i^q(k)\right\}_{k\geqslant 0}$ be a sequence of independent and identically distributed (i. i. d.) random variables, for any fixed (i,q), all $k\geqslant 0$ and (i,q),

$$\left|\triangle_i^q(k)\right|<M_1,\quad \left|\frac{1}{\triangle_i^q(k)}\right|<M_2,\quad \mathbb{E}\left[\frac{1}{\triangle_i^q(k)}\right]=0;$$

(b) $\left\{\triangle_i^q(k)\right\}_{k\geqslant 0}$ and $\left\{\triangle_j^r(k)\right\}_{k\geqslant 0}$ are mutually independent of each other for $i\neq j$ or $q\neq r$.

Remark 1. In order to ensure the almost sure convergence of Algorithm 1, $\{c(k)\}$ and $\{\iota(k)\}$ should satisfy the stochastic approximation assumption [29]

$$\sum_{k=1}^{\infty}\frac{\iota(k)}{c(k)}=\infty,\quad \sum_{k=1}^{\infty}\frac{\iota^2(k)}{c^2(k)}=\infty,$$

and

$$\sum_{k=1}^{\infty}\iota(k)c(k)<\infty.$$

Therefore, $0\leqslant\epsilon<\dfrac{1}{4}$, $\frac{1}{2}-\epsilon>\delta>\epsilon$.

With the stripe observation environment of the distributed problem, $\frac{\iota(k)}{c(k)}$ chosen in this paper differs from that of [17].

Appendix B Proof of Theorem 3

The following lemma, is essential for the proof of Theorem 3.

Lemma 1. With Assumption 3, $\left\|g\left(\boldsymbol{x^*},\boldsymbol{\lambda(0)}\right)-g\left(\boldsymbol{\hat{x}^*},\boldsymbol{\lambda_0}\right)\right\|\leqslant\left|\lambda(0)-\lambda_0\right|\cdot\left\|f_L\left(\hat{x}^*\right)-f_R\left(\hat{x}^*\right)\right\|$.

Proof. We get

$$\begin{aligned}&\left\|g\left(\boldsymbol{x^*},\boldsymbol{\lambda(0)}\right)-g\left(\boldsymbol{\hat{x}^*},\boldsymbol{\lambda_0}\right)\right\|\\ =&\sum_{i=1}^{n}\left[\lambda(0)f_{L_i}\left(x^*\right)+(1-\lambda(0))f_{R_i}\left(x^*\right)\right]-\sum_{i=1}^{n}\left[\lambda_0 f_{L_i}\left(\hat{x}^*\right)+(1-\lambda_0)f_{R_i}\left(\hat{x}^*\right)\right]\leqslant\left\|\Gamma_{1n}+\Gamma_{2n}\right\|\end{aligned}\tag{B1}$$

where

$$\begin{aligned}\Gamma_{1n}&=\sum_{i=1}^{n}\left[\lambda(0)\Big(f_{L_i}\left(x^*\right)-f_{L_i}\left(\hat{x}^*\right)\Big)+\left(1-\lambda(0)\right)\Big(f_{R_i}\left(x^*\right)-f_{R_i}\left(\hat{x}^*\right)\Big)\right]\\ \Gamma_{2n}&=\sum_{i=1}^{n}\left[\left(\lambda(0)-\lambda_0\right)\Big(f_{L_i}\left(\hat{x}^*\right)-f_{R_i}\left(\hat{x}^*\right)\Big)\right].\end{aligned}$$

According to Assumption 2(b), we obtain

$$\begin{aligned}f_{L_i}\left(x^*\right)-f_{L_i}\left(\hat{x}^*\right)&\leqslant\Big\langle\nabla f_{L_i}\left(x^*\right),x^*-\hat{x}^*\Big\rangle\\ f_{R_i}\left(x^*\right)-f_{R_i}\left(\hat{x}^*\right)&\leqslant\Big\langle\nabla f_{R_i}\left(x^*\right),x^*-\hat{x}^*\Big\rangle.\end{aligned}\tag{B2}$$

It yields

$$\begin{aligned}\left\|\Gamma_{1n}\right\|=&\left\|\sum_{i=1}^{n}\left[\lambda(0)\Big(f_{L_i}\left(x^*\right)-f_{L_i}\left(\hat{x}^*\right)\Big)+\left(1-\lambda(0)\right)\Big(f_{R_i}\left(x^*\right)-f_{R_i}\left(\hat{x}^*\right)\Big)\right]\right\|\\ \leqslant&\left\|\sum_{i=1}^{n}\left[\lambda(0)f_{L_i}\left(x^*\right)+\left(1-\lambda(0)\right)f_{R_i}\left(x^*\right)\right]\right\|\cdot\left\|f_{L_i}\left(\hat{x}^*\right)-f_{R_i}\left(\hat{x}^*\right)\right\|.\end{aligned}\tag{B3}$$

Since $\left(x^*,\lambda(0)\right)$ is an optimal solution of the distributed problem, we have

$$\sum_{i=1}^{n}\left[\lambda(0)f_{L_i}\left(x^*\right)+\left(1-\lambda(0)\right)f_{R_i}\left(x^*\right)\right]=0.\tag{B4}$$

Therefore,

$$\left\|\Gamma_{1n}\right\|=0.\tag{B5}$$

Combining (B5) with (B1), yields

$$\left\|g\left(\boldsymbol{x^*},\boldsymbol{\lambda(0)}\right)-g\left(\boldsymbol{\hat{x}^*},\boldsymbol{\lambda_0}\right)\right\|\leqslant\left|\lambda(0)-\lambda_0\right|\cdot\left\|f_L\left(\hat{x}^*\right)-f_R\left(\hat{x}^*\right)\right\|.\tag{B6}$$

Lemma 1 gives an upper bound of $\left\|g\left(\boldsymbol{x^*},\boldsymbol{\lambda(0)}\right)-g\left(\boldsymbol{\hat{x}^*},\boldsymbol{\lambda_0}\right)\right\|$.

Next, we provide the proof of Theorem 3.

Proof.

(a) Theorem 3 is a direct conclusion of Lemma 1 and the fact that $\lambda_0 = \lambda(0)$ with Assumption 4(a).
(b) According to the law of large numbers [23],

$$\lim_{n\to\infty} P\big(\|\lambda(0) - \lambda_0\| < \epsilon\big) = 1.$$

It yields

$$\lim_{n\to\infty} P\bigg(\Big\|g\big(\boldsymbol{x}^{\boldsymbol{*}}, \boldsymbol{\lambda(0)}\big) - g\big(\hat{\boldsymbol{x}}^{\boldsymbol{*}}, \boldsymbol{\lambda_0}\big)\Big\| < \epsilon c_0\big(\hat{x}^*\big)\bigg) = 1.$$

(c) The following equality holds according to the central limit theorem [23]:

$$\lim_{n\to\infty} P\bigg(\frac{\sqrt{n}\big(\lambda(0) - \lambda_0\big)}{\nu} \leqslant x\bigg) = \phi(x)$$

where $\phi(x)$ is the cumulative distribution function of Gaussian distribution $\mathcal{N}(0,1)$. With Lemma 1, we get

$$\begin{aligned}&\lim_{n\to\infty} P\bigg(\Big\|g\big(\boldsymbol{x}^{\boldsymbol{*}}, \boldsymbol{\lambda(0)}\big) - g\big(\hat{\boldsymbol{x}}^{\boldsymbol{*}}, \boldsymbol{\lambda_0}\big)\Big\| \leqslant a c_0\big(\hat{x}^*\big)\bigg) \geqslant \lim_{n\to\infty} P\bigg(\big|\lambda(0) - \lambda_0\big| \leqslant a\bigg)\\ =&P\bigg(\frac{\sqrt{n}\big|\lambda(0) - \lambda_0\big|}{\nu} \leqslant \frac{\sqrt{n}a}{\nu}\bigg) = \phi\bigg(\frac{\sqrt{n}a}{\nu}\bigg) - \phi\bigg(-\frac{\sqrt{n}a}{\nu}\bigg) = 2\phi\bigg(\frac{\sqrt{n}a}{\nu}\bigg) - 1\end{aligned}$$

and

$$\begin{aligned}&\lim_{n\to\infty} P\bigg(\Big\|g\big(\boldsymbol{x}^{\boldsymbol{*}}, \boldsymbol{\lambda(0)}\big) - g\big(\hat{\boldsymbol{x}}^{\boldsymbol{*}}, \boldsymbol{\lambda_0}\big)\Big\| \geqslant b c_0\big(\hat{x}^*\big)\bigg) \leqslant \lim_{n\to\infty} P\bigg(\big|\lambda(0) - \lambda_0\big| \geqslant b\bigg)\\ =&P\bigg(\frac{\sqrt{n}\big|\lambda(0) - \lambda_0\big|}{\nu} \geqslant \frac{\sqrt{n}b}{\nu}\bigg) = 1 - \phi\bigg(\frac{\sqrt{n}b}{\nu}\bigg) + \phi\bigg(-\frac{\sqrt{n}b}{\nu}\bigg) = 2 - 2\phi\bigg(\frac{\sqrt{n}b}{\nu}\bigg).\end{aligned}$$

References

1 Anit K S, Dusan J, Dragana B, et al. Distributed zeroth order optimization over random networks: A Kiefer-Wolfowitz stochastic approximation approache[J]. arXiv preprint arXiv:1803.07836, 2018.
2 Nedić Liu Q, Le X, Li K. A distributed optimization algorithm based on multiagent network for economic dispatch with region partitioning[J]. IEEE transactions on cybernetics, 2019, 51(5): 2466-2475.
3 Yi P, Hong Y, Liu F. Initialization-free distributed algorithms for optimal resource allocation with feasibility constraints and application to economic dispatch of power systems[J]. Automatica, 2016, 74: 259-269.
4 Li Z, Liu X, Ren W, et al. Distributed tracking control for linear multiagent systems with a leader of bounded unknown input[J]. IEEE Transactions on Automatic Control, 2012, 58(2): 518-523.
5 Lei J, Chen H F, Fang H T. Asymptotic properties of primal-dual algorithm for distributed stochastic optimization over random networks with imperfect communications[J]. SIAM Journal on Control and Optimization, 2018, 56(3): 2159-2188.
6 Wang Y, Lin P, Hong Y. Distributed regression estimation with incomplete data in multi-agent networks[J]. Science China Information Sciences, 2018, 61(9): 1-14.
7 Liu S, Qiu Z, Xie L. Convergence rate analysis of distributed optimization with projected subgradient algorithm[J]. Automatica, 2017, 83: 162-169.
8 Nedic A, Ozdaglar A. Distributed subgradient methods for multi-agent optimization[J]. IEEE Transactions on Automatic Control, 2009, 54(1): 48-61.
9 Sundhar Ram S, Nedi04 A, Veeravalli V V. Distributed stochastic subgradient projection algorithms for convex optimization[J]. Journal of optimization theory and applications, 2010, 147(3): 516-545.
10 Yang T, Yi X, Wu J, et al. A survey of distributed optimization[J]. Annual Reviews in Control, 2019, 47: 278-305.
11 Liang S, Yin G. Exponential convergence of distributed primalCdual convex optimization algorithm without strong convexity[J]. Automatica, 2019, 105: 298-306.
12 Liang S, Yin G. Dual averaging push for distributed convex optimization over time-varying directed graph[J]. IEEE Transactions on Automatic Control, 2019, 65(4): 1785-1791.
13 Tsianos K I, Lawlor S, Rabbat M G. Push-sum distributed dual averaging for convex optimization[C]//2012 ieee 51st ieee conference on decision and control (cdc). IEEE, 2012: 5453-5458.
14 Yuan D, Ho D W C, Xu S. Zeroth-order method for distributed optimization with approximate projections[J]. IEEE transactions on neural networks and learning systems, 2015, 27(2): 284-294.
15 Yuan D, Ho D W C. Randomized gradient-free method for multiagent optimization over time-varying networks[J]. IEEE Transactions on Neural Networks and Learning Systems, 2014, 26(6): 1342-1347.
16 Marinopoulos A G, Alexiadis M C, Dokopoulos P S. Energy losses in a distribution line with distributed generation based on stochastic power flow[J]. Electric Power Systems Research, 2011, 81(10): 1986-1994.
17 Wang Y, Zhao W, Hong Y, et al. Distributed subgradient-free stochastic optimization algorithm for nonsmooth convex functions over time-varying networks[J]. SIAM Journal on Control and Optimization, 2019, 57(4): 2821-2842.
18 Wang Y, Tu Z, Qin H. Distributed stochastic mirror descent algorithm for resource allocation problem[J]. Control Theory and Technology, 2020, 18(4): 339-347.
19 Chang Y, Yan L, Wu T, et al. Remote sensing image stripe noise removal: From image decomposition perspective[J]. IEEE Transactions on Geoscience and Remote Sensing, 2016, 54(12): 7018-7031.
20 Chen Y, Huang T Z, Zhao X L, et al. Stripe noise removal of remote sensing images by total variation regularization and group sparsity constraint[J]. Remote sensing, 2017, 9(6): 559.
21 Chen C, Wang F, Zhou B, et al. An interval optimization based day-ahead scheduling scheme for renewable energy management in smart distribution systems[J]. Energy Conversion and Management, 2015, 106: 584-596.
22 Li Y Z, Wu Q H, Jiang L, et al. Optimal power system dispatch with wind power integrated using nonlinear interval optimization and evidential reasoning approach[J]. IEEE Transactions on Power Systems, 2015, 31(3): 2246-2254.

23 Durrett R. Probability: theory and examples[M]. Cambridge university press, 2019.
24 Bhurjee A K, Panda G. Efficient solution of interval optimization problem[J]. Mathematical Methods of Operations Research, 2012, 76(3): 273-288.
25 Nedi04 A, Olshevsky A. Stochastic gradient-push for strongly convex functions on time-varying directed graphs[J]. IEEE Transactions on Automatic Control, 2016, 61(12): 3936-3947.
26 Polyak B T. Introduction to optimization. optimization software[J]. Inc., Publications Division, New York, 1987, 1-32.
27 Maeda T. On characterization of fuzzy vectors and its applications to fuzzy mathematical programming problems[J]. Fuzzy Sets and Systems, 2008, 159(24): 3333-3346.
28 DeGroot M H. Reaching a consensus[J]. Journal of the American Statistical Association, 1974, 69(345): 118-121.
29 Robbins H, Monro S. A stochastic approximation method[J]. The annals of mathematical statistics, 1951: 400-407.
30 Clarke F H, Ledyaev Y S, Stern R J, et al. Nonsmooth analysis and control theory[M]. Springer Science & Business Media, 2008.
31 Carmona R, Lacker D. A probabilistic weak formulation of mean field games and applications[J]. The Annals of Applied Probability, 2015, 25(3): 1189-1231.