

Decentralized Composite Optimization in Stochastic Networks: A Dual Averaging Approach

Zirui Zhou

Joint work with Changxin Liu, Jian Pei, Yong Zhang, Yang Shi

24th Midwest Optimization Meeting, Waterloo, Canada



About Huawei Technologies Canada

About me:

- Name: Zirui Zhou
- Senior Principal Researcher, Huawei Tehnologies Canada, 2020–present
- Assistant Professor, Hong Kong Baptist University, 2018–2020
- Research interest: convex optimization, numerical optimization, solver, ...

About *Big Data & Intelligent Platform Lab* at Vancouver, BC:

- Huawei's mathematical optimization solver: **OptVerse**
 - Linear programming, mixed-integer linear programming, quadratic programming, black-box optimization
- Smart decision cloud platform: Machine Learning + Decision making
 - ML4Opt, ML4Modeling
 - Federated learning, data valuation, digital asset pricing

Multiple research scientist positions available: FTE/Intern/Co-op

Machine Learning for Optimization Solvers

- ML for configuration: black-box optimization + clustering
 - Winner of 2021 NeurIPS ML4CO Competition Configuration Task
- ML for smart branching in B&B: GNN + Attention + Temporal
 - Runner-up of 2021 NeurIPS ML4CO Competition Dual Task
- ML for cut-generation: CutRank + Multiple Instance Learning
 - Average 15% speedup on our internal test set
- ML for basis selection: GNN + Filter Layer + Basis Repair
 - Average 70% speedup on our supply chain LP problems

① Introduction and Motivation

② Dual Averaging for Decentralized Optimization

③ Numerical Experiments

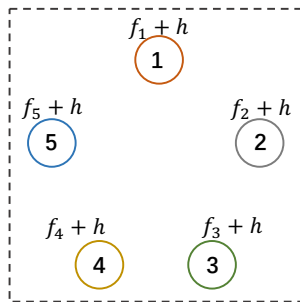
④ Summary

Decentralized convex composite optimization

- Problem statement

$$\min_{\theta} \left\{ F(\theta) := \frac{1}{n} \sum_{i=1}^n f_i(\theta) + h(\theta) \right\}$$

- ▶ Agent-specific convex function: f_i
- ▶ Common convex regularizer: h
- ▶ Decision variable: θ
- ▶ Agent/Machine: $i \in [n]$



Multi-agent system

- Applications included in the framework
 - ▶ Constrained optimization when h is a convex indicator function
 - ▶ Sparse recovery when h is an l_1 -regularizer
 - ▶ ...

Decentralized structure of message-passing

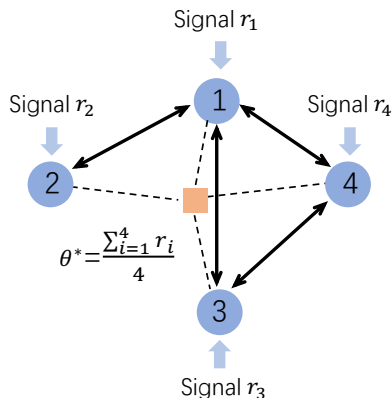
Centralized distributed structure

Decentralized structure

Advantages of the decentralized structure

- Balanced computation with each node
- Robustness to network change
- Preserved privacy

Average consensus (Olfati-Saber and Murray, 2004)



- Average Consensus

$$x_i^{(t)} = \sum_{j \in \mathcal{N}_i \cup \{i\}} p_{ij} x_j^{(t-1)}$$

Lemma (Average Seeking)

If P is doubly stochastic and $x_i^{(0)} = r_i$, then

$$\sum_{i=1}^m x_i^{(t)} = \sum_{i=1}^m r_i, \forall t \geq 0.$$

In addition, if the graph is connected, then

$$x_i^{(\infty)} \rightarrow \theta^* = \frac{1}{m} \sum_{i=1}^m r_i, \forall i \in \mathcal{V}$$

Decentralized projected subgradient method

$$\min_{\theta \in \mathcal{X}} \frac{1}{n} \sum_{i=1}^n F_i(\theta) \Leftrightarrow \min_{x_i \in \mathcal{X}, \forall i \in \mathcal{V}} \frac{1}{n} \sum_{i=1}^n F_i(x_i), \quad x_1 = x_2 = \cdots = x_n$$

- DPSM (Nedic and Ozdaglar, 2009)

$$y_i^{(t)} = \underbrace{\sum_{j \in \mathcal{N}_i \cup \{i\}} p_{ij} x_j^{(t-1)}}_{\text{average consensus}} - \underbrace{a_{t-1} g_i^{(t-1)}}_{\text{gradient search}}$$
$$x_i^{(t)} = \text{proj}_{\mathcal{X}}(y_i^{(t)})$$

- ▶ $g_i^{(t)} \in \partial F_i(x_i^{(t)})$: one of the subgradients of F_i over $x_i^{(t)}$
- ▶ $\{a_t\}_{t \geq 0}$: diminishing stepsize
- ▶ sublinear convergence

A Summary of Research Directions

- **Better rates:** can decentralized algorithms achieve the nice convergence rate of its centralized counterparts?
 - ▶ first-order methods, second-order methods
- **Handle more classes of functions**
 - ▶ smooth, non-smooth, composite
- **Robust to more complicated communication networks**
 - ▶ fixed network, time-varying network, stochastic network

Literature review

	Deterministic network	Stochastic network
Unconstrained	GD+consensus (Linear) ^{1 2} Primal-dual methods (Linear) ^{4 5}	GD+consensus (Linear) ³ Bregman method ($\mathcal{O}(1/t)$) ⁶
Constrained/ Composite	Projected GD+consensus ($\mathcal{O}(1/t)$) ^{7 8} Primal-dual methods (Linear) ¹⁰	DA+consensus ($\mathcal{O}(1/\sqrt{t})$) ⁹

Note: GD=Gradient descent, DA = Dual averaging

¹Nedic et al., DSM for multi-agent optimization, *TAC*, 2009

²Qu and Li, Harness smoothness to accelerate DO, *IEEE TCNS*, 2017

³Xu et al., Convergence of asynchronous DGM over stochastic networks, *TAC*, 2017

⁴Yi et al., Linear convergence for DO without strong convexity, *IEEE CDC*, 2020

⁵Shi et al., On the linear convergence of ADMM in DO, *IEEE TSP*, 2014

⁶Xu et al., A Bregman splitting scheme for DO over networks, *TAC*, 2018

⁷Nedic et al., Constrained consensus and optimization, *TAC*, 2010

⁸Shi et al., A proximal gradient algorithm for DO, *TSP*, 2015

⁹Duchi et al., Dual averaging for distributed optimization, *TAC*, 2011

¹⁰Alghunaim et al., A linearly convergent PGM for DO, *NeurIPS*, 2019

Literature review

	Deterministic network	Stochastic network
Unconstrained	GD+consensus (Linear) ^{1 2}	GD+consensus (Linear) ³
	Primal-dual methods (Linear) ^{4 5}	Bregman method ($\mathcal{O}(1/t)$) ⁶
Constrained/ Composite	Projected GD+consensus ($\mathcal{O}(1/t)$) ^{7 8}	DA+consensus ($\mathcal{O}(1/\sqrt{t})$) ⁹
	Primal-dual methods (Linear) ¹⁰	

Note: GD=Gradient descent, DA = Dual averaging

Can we develop a linearly convergent algorithm for decentralized optimization with composite objective and stochastic network?

¹Nedic et al., DSM for multi-agent optimization, *TAC*, 2009

²Qu and Li, Harness smoothness to accelerate DO, *IEEE TCNS*, 2017

³Xu et al., Convergence of asynchronous DGM over stochastic networks, *TAC*, 2017

⁴Yi et al., Linear convergence for DO without strong convexity, *IEEE CDC*, 2020

⁵Shi et al., On the linear convergence of ADMM in DO, *IEEE TSP*, 2014

⁶Xu et al., A Bregman splitting scheme for DO over networks, *TAC*, 2018

⁷Nedic et al., Constrained consensus and optimization, *TAC*, 2010

⁸Shi et al., A proximal gradient algorithm for DO, *TSP*, 2015

⁹Duchi et al., Dual averaging for distributed optimization, *TAC*, 2011

¹⁰Alghunaim et al., A linearly convergent PGM for DO, *NeurIPS*, 2019

Can we use existing approaches to tackle this problem?

- **Decentralized primal-dual algorithms (Jakovetic et al. 2015)**

- ▶ Step 1: Reformulation via lifting

$$\begin{aligned} \min_{x_1 \cdots, x_n \in \mathbb{R}^m} \quad & \frac{1}{n} \sum_{i=1}^n (f_i(x_i) + h(x_i)) \\ \text{s.t.} \quad & \overbrace{\left(\underbrace{\mathcal{L}}_{\substack{\text{a priori known}}} \otimes I \right) ([x_1^T, \cdots, x_n^T]^T)}^{\text{consensus constraint}} = 0 \end{aligned}$$

- ▶ Step 2: Solve the linearly constrained optimization problem via primal-dual algorithms (with coordinate change)
- When the network is random, i.e., $\mathcal{L}$ unknown, it does NOT fit in

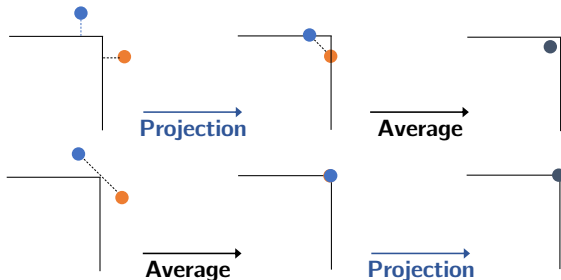
Can we use existing approaches to tackle this problem?

- **Consensus-based decentralized gradient method (DGM) (Nedic et al. 2010)**

$$y_i^{(t)} = \sum_{j \in \mathcal{N}_i \cup \{i\}} p_{ij} x_j^{(t-1)} - a_{t-1} g_i^{(t-1)}, \quad x_i^{(t)} = \arg \min_{x \in \mathcal{X}} \|x - y_i^{(t)}\|^2$$

Observation

- In DGM, **averaging** is tightly coupled with **projection** and they do NOT necessarily commute



- ① Introduction and Motivation
- ② Dual Averaging for Decentralized Optimization
- ③ Numerical Experiments
- ④ Summary

Centralized dual averaging

- Dual averaging for nonsmooth convex functions (Nesterov, 2009)

$$z^{(t)} = z^{(t-1)} + g^{(t-1)}, g^{(t-1)} \in \partial F(\theta^{(t-1)}), \quad \theta^{(t)} = \arg \min_{\theta \in \mathcal{X}} \left\{ a_t \langle z^{(t)}, \theta \rangle + \underbrace{d(\theta)}_{\frac{1}{2} \|\theta\|^2} \right\}$$

- Dual averaging for smooth and (μ -strongly) convex functions (Lu et al., 2018)

$$\theta^{(t)} = \arg \min_{\theta \in \mathcal{X}} \left\{ \sum_{\tau=0}^{t-1} a_{\tau+1} \left(\langle \nabla f(\theta^{(\tau)}), \theta \rangle + \frac{\mu}{2} \|\theta - \theta^{(\tau)}\|^2 \right) + d(\theta) \right\}$$

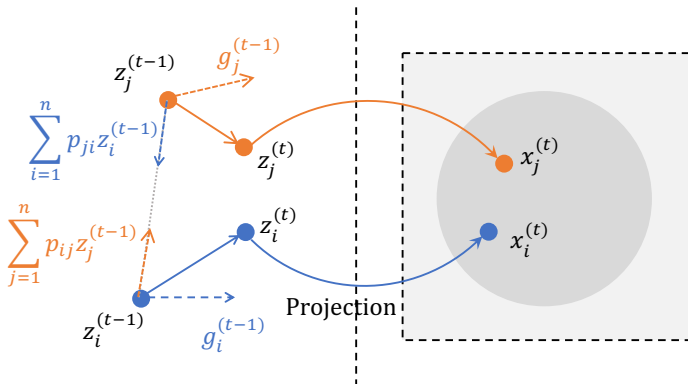
- Achieves linear convergence when $\mu > 0$ (Lu et al., 2018)

Decentralized dual averaging

- Decentralized dual averaging (DDA) (Duchi et al., 2011)

$$z_i^{(t)} = \sum_{j \in \mathcal{N}_i \cup \{i\}} p_{ij} z_j^{(t-1)} + g_i^{(t-1)}, g_i^{(t-1)} \in \partial F_i(x_i^{(t-1)})$$

$$x_i^{(t)} = \arg \min_{\theta \in \mathcal{X}} \{a_t \langle z_i^{(t)}, \theta \rangle + d(\theta)\}$$



DDA versus DPGM

- DDA

$$z_i^{(t)} = \sum_{j \in \mathcal{N}_i \cup \{i\}} p_{ij} z_j^{(t-1)} + g_i^{(t-1)}$$

$$x_i^{(t)} = \arg \min_{\theta \in \mathcal{X}} \{a_t \langle z_i^{(t)}, \theta \rangle + d(\theta)\}$$

- DPGM

$$y_i^{(t)} = \sum_{j \in \mathcal{N}_i \cup \{i\}} p_{ij} x_j^{(t-1)} - a_{t-1} g_i^{(t-1)}$$

$$x_i^{(t)} = \text{proj}_{\mathcal{X}}(y_i^{(t)})$$

Comparison

- DDA equally weights the (sub)gradients obtained so far
- **In DDA, consensus-building is decoupled from projection**
 - ▶ DDA has the advantage of handling stochastic networks and nonsmoothness *simultaneously*

DDA versus DPGM

- DDA

$$z_i^{(t)} = \sum_{j \in \mathcal{N}_i \cup \{i\}} p_{ij} z_j^{(t-1)} + g_i^{(t-1)}$$
$$x_i^{(t)} = \arg \min_{\theta \in \mathcal{X}} \{a_t \langle z_i^{(t)}, \theta \rangle + d(\theta)\}$$

- DPGM

$$y_i^{(t)} = \sum_{j \in \mathcal{N}_i \cup \{i\}} p_{ij} x_j^{(t-1)} - a_{t-1} g_i^{(t-1)}$$
$$x_i^{(t)} = \text{proj}_{\mathcal{X}}(y_i^{(t)})$$

Comparison

- DDA equally weights the (sub)gradients obtained so far
- **In DDA, consensus-building is decoupled from projection**
 - ▶ DDA has the advantage of handling stochastic networks and nonsmoothness *simultaneously*

DDA and all later extensions considered nonsmooth problems, and have an $\mathcal{O}(1/\sqrt{t})$ **rate of convergence**

Speeding up DDA

- Decentralized composite optimization

$$\min_{\theta} \left\{ \underbrace{\frac{1}{n} \sum_{i=1}^n f_i(\theta)}_{f: \text{smooth}} + \underbrace{h(\theta)}_{\text{nonsmooth}} \right\}$$

- Reformulation of centralized dual averaging for composite functions

$$\begin{aligned} \theta^{(t)} &= \arg \min_{\theta} \left\{ \sum_{\tau=0}^{t-1} a_{\tau+1} \left(\left\langle \nabla f(\theta^{(\tau)}), x \right\rangle + \frac{\mu}{2} \|\theta - \theta^{(\tau)}\|_2^2 + h(\theta) \right) + d(\theta) \right\} \\ &= \arg \min_{\theta} \left\{ \underbrace{\left\langle \sum_{\tau=0}^{t-1} a_{\tau+1} (\nabla f(\theta^{(\tau)}) - \mu \theta^{(\tau)}), \theta \right\rangle}_{z^{(t)}} + \underbrace{\sum_{\tau=0}^{t-1} a_{\tau+1} \left(\frac{\mu}{2} \|\theta\|_2^2 + h(\theta) \right)}_{A_t} + d(\theta) \right\} \\ &= \arg \min_{\theta} \left\{ \left\langle z^{(t)}, \theta \right\rangle + \underbrace{\frac{\mu A_t}{2} (\|\theta\|^2 + h(\theta))}_{\text{common knowledge to all agents}} + d(\theta) \right\} \end{aligned}$$

- If $z^{(t)}$ can be accurately estimated by each agent, decentralized optimization may be achieved

Speeding Up DDA (cont'd)

- How to estimate $z^{(t)} = \sum_{\tau=0}^{t-1} a_{\tau+1} (\nabla f(\theta^{(\tau)}) - \mu \theta^{(\tau)})$?
- The scheme in conventional DDA (Duchi et al., 2011)

$$z_i^{(t)} = \sum_{j \in \mathcal{N}_i^{(t-1)} \cup \{i\}} p_{ij}^{(t-1)} z_j^{(t-1)} + a_t \left(\nabla f_i(x_i^{(t-1)}) - \mu x_i^{(t-1)} \right)$$

- Then, each agent uses $z_i^{(t)}$ to run a local dual averaging step

$$x_i^{(t)} = \arg \min_{x \in \mathbb{R}^m} \left\{ \langle z_i^{(t)}, x \rangle + A_t \left(\frac{\mu}{2} \|x\|^2 + h(x) \right) + d(x) \right\}$$

- It is only guaranteed that $\|z_i^{(t)} - \frac{1}{n} \sum_{j=1}^n z_j^{(t)}\|$ is bounded (**cannot achieve exact optimization if $\{a_t\}_{t \geq 0}$ is constant or increasing**)

For fast convergence, a more accurate estimate is necessary to validate the use of constant or geometrically increasing $\{a_t\}_{t \geq 0}$

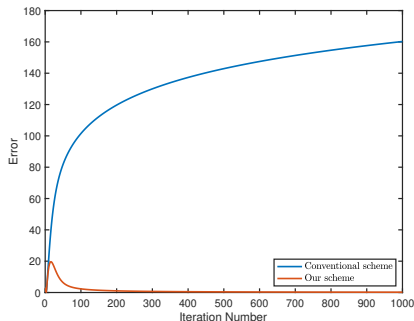
Speeding up DDA (cont'd)

- Tailored dynamic average consensus for DDA

$$z_i^{(t)} = \sum_{j \in \mathcal{N}_i^{(t-1)} \cup \{i\}} p_{ij}^{(t-1)} \left(z_j^{(t-1)} + a_t \left(\nabla f_j(x_j^{(t-1)}) - \mu x_j^{(t-1)} \right) \right)$$

$s_i^{(t-1)}$

$$s_i^{(t)} = \sum_{j \in \mathcal{N}_i^{(t-1)} \cup \{i\}} p_{ij}^{(t-1)} s_j^{(t-1)} + \left(\nabla f_i(x_i^{(t)}) - \mu x_i^{(t)} \right) - \left(\nabla f_i(x_i^{(t-1)}) - \mu x_i^{(t-1)} \right)$$



- ▶ A system of 5 agents, $a_t = 1$

- ▶ $\nabla f_i(x_i^{(t)}) - \mu x_i^{(t)} = i - \frac{5}{t}$

- ▶ Error:

$$\left\| z_i^{(t)} - \sum_{\tau=0}^{t-1} \frac{\sum_{i=1}^5 \left(\nabla f_i(x_i^{(\tau)}) - \mu x_i^{(\tau)} \right)}{n} \right\|$$

DDA algorithm

- **Initialization:** $a_0 = a > 0$, $A_0 = 0$, $x_i^{(0)} = x^{(0)}$, $s_i^{(0)} = \nabla f_i(x^{(0)}) - \mu x^{(0)}$ and $z_i^{(0)} = 0$
- **Parameter update:** $a_t = \frac{a_{t-1}}{1-a\mu}$ and $A_t = A_{t-1} + a_t$
- **Consensus step:**

$$z_i^{(t)} = \sum_{j=1}^n p_{ij}^{(t-1)} \left(z_j^{(t-1)} + a_t s_j^{(t-1)} \right)$$

$$s_i^{(t)} = \sum_{j=1}^n p_{ij}^{(t-1)} s_j^{(t-1)} + \left(\nabla f_i(x_i^{(t)}) - \mu x_i^{(t)} \right) - \left(\nabla f_i(x_i^{(t-1)}) - \mu x_i^{(t-1)} \right)$$

- **Local dual averaging:**

$$x_i^{(t)} = \arg \min_{x \in \mathbb{R}^m} \left\{ \langle z_i^{(t)}, x \rangle + A_t \left(\frac{\mu}{2} \|x\|^2 + h(x) \right) + d(x) \right\}$$

- Set $t = t + 1$ and go to **Parameter update**

Assumptions

- Assumptions for objective functions

- ▶ f_i is (strongly) convex with modulus $\mu \geq 0$

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

- ▶ $\nabla f_i(x)$ is Lipschitz continuous with constant $L > 0$

$$\|\nabla f_i(x) - \nabla f_i(y)\| \leq L \|x - y\|$$

- Assumptions for the mixing matrix

- ▶ $P^{(t)}$ is independent of the random events that occur up to time $t - 1$
- ▶ There exists a constant $\beta \in (0, 1)$ such that

$$\sqrt{\rho \left(\mathbb{E}_t \left[P^{(t)T} P^{(t)} \right] - \frac{\mathbf{1}\mathbf{1}^T}{n} \right)} \leq \beta$$

where $\rho(\cdot)$ denotes the spectral radius.

Theoretical results

- Auxiliary variable

$$y^{(t)} = \arg \min_{x \in \mathbb{R}^m} \left\{ \left\langle \frac{1}{n} \sum_{i=1}^n z_i^{(t)}, x \right\rangle + A_t \left(\frac{\mu}{2} \|x\|^2 + h(x) \right) + d(x) \right\}$$

Theorem 1

Suppose that ^a

$$\frac{1}{a} > \left\{ \frac{\beta(2L+3\mu)}{(1-\beta)^2} + \mu, \quad 2L - \mu + \frac{4L-2\mu}{(1-a\mu)(1-\rho(\mathbf{M}))^2} \right\}.$$

where

$$\mathbf{M} = \begin{bmatrix} \beta & \beta \\ \frac{a(L+\mu)}{1-a\mu} \left(\beta + \frac{1}{1-a\mu} \right) & \frac{\beta+a\beta(L+\mu)}{1-a\mu} \end{bmatrix}.$$

Then

$$\mathbb{E}[F(\tilde{y}^{(t)})] - F(x^*) \leq \frac{C}{A_t}, \quad \mathbb{E}[\|\tilde{x}_i^{(t)} - \tilde{y}^{(t)}\|^2] \leq \frac{D}{A_t}$$

where $\tilde{x}_i^{(t)} = A_t^{-1} \sum_{\tau=1}^t a_\tau x_i^{(\tau)}$ and $\tilde{y}^{(t)} = A_t^{-1} \sum_{\tau=1}^t a_\tau y^{(\tau)}$.

^a $a = \Theta((1-\beta)^2/\kappa)$ where $\kappa = L/\mu$.

Consequences of Theorem 1

Corollary 1

Suppose the premise of Theorem 1 holds. If $\mu > 0$, then

$$\mathbb{E}[\|\tilde{x}_i^{(t)} - x^*\|^2] \leq \frac{2}{a} \left(\frac{2C}{\mu} + D \right) (1 - a\mu)^t$$

Corollary 2

Suppose the premise of Theorem 1 holds. If $\mu = 0$, then

$$\mathbb{E}[F(\tilde{y}^{(t)})] - F(x^*) \leq \frac{C}{at}, \quad \mathbb{E}[\|\tilde{x}_i^{(t)} - \tilde{y}_i^{(t)}\|^2] \leq \frac{D}{at}.$$

In addition, if $h \equiv 0$, $d(x) = \|x\|^2/2$, and

$$\frac{1}{a} > 2L \cdot \max \left\{ \frac{\beta}{(1 - \beta)^2}, 1 + \frac{6}{(1 - \nu)^2} \right\},$$

then we further have

$$\mathbb{E}[F(\tilde{x}_i^{(t)})] - F(x^*) \leq \frac{E}{t}.$$

① Introduction and Motivation

② Dual Averaging for Decentralized Optimization

③ Numerical Experiments

④ Summary

Numerical experiments I

- Decentralized logistic regression (Spambase Data Set in UCI Machine Learning Repository)

$$\min_{\theta \in \mathbb{R}^n} \frac{1}{n} \sum_{i=1}^n f_i(\theta) + \phi \|\theta\|_1$$

where

$$f_i(\theta) = \frac{1}{m_i} \sum_{j=1}^{m_i} \ln \left(1 + \exp \left(-y_j^i (M_j^{iT} \theta) \right) \right) + \frac{\chi}{2} \|\theta\|^2$$

- Parameters

n	30
m_i	100
M_j^i	features $M_j^i \in \mathbb{R}^{58}$
y_j^i	labels $y_j^i \in \{-1, 1\}$
χ	0.02
ϕ	0.001
θ^*	ground truth by centralized Proximal gradient descent
Fixed graph	Erdos-Renyi graphs with connectivity ratios 0.2, 0.4
Weight matrix P	Metropolis-Hastings rule

Compared Algorithms and Network Configurations

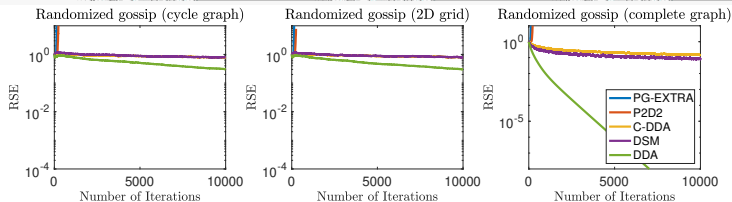
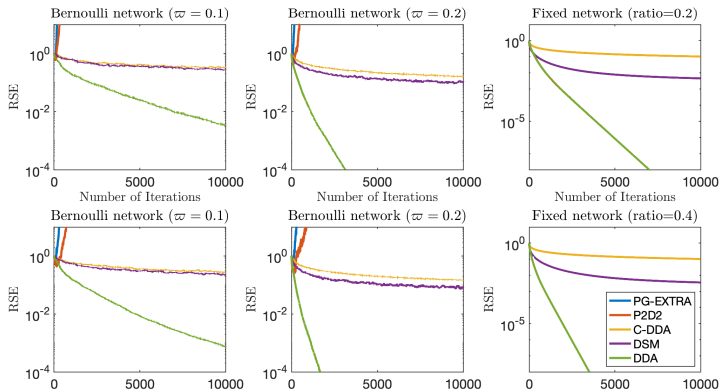
- Compared algorithms and their parameters

PG-EXTRA (Shi et al. 2015)	$a = 0.1, 0.2, 0.0001$
P2D2 (Alghunaim et al. 2019)	$a = 0.1, 0.2, 0.0001$
DSM (Lobel & Ozdaglar 2010)	$a_t = 1/\sqrt{t+1}$
C-DDA (Duchi et al. 2010)	$a_t = 1/\sqrt{t+1}$
DDA (this work)	$a_t = 0.1/(0.998)^t, 0.2/(0.996)^t, 0.1/(0.998)^t$
$d(x)$ for DA-type algorithms	$\ x\ ^2/2$
Relative square error (RSE)	$\frac{\sum_{i=1}^n \ x_i^{(t)} - \theta^*\ ^2}{\sum_{i=1}^n \ x_i^{(0)} - \theta^*\ ^2}$

- Network configurations

Bernoulli protocol	each link is activated with probability $\varpi = 0.1, 0.2$
Randomized gossip	a single link (i, j) is sampled at each t with probability $\frac{1}{n(\mathcal{N}_i +1)}$

Comparison Results



Numerical experiments II

- Decentralized LASSO (Synthetic Data Set)

$$\min_{\theta \in \mathbb{R}^m} \frac{1}{n} \sum_{i=1}^n \frac{1}{2} \|b_i - C_i \theta\|^2, \quad \text{s.t. } \|\theta\|_1 \leq R,$$

- Parameters

n	30
(C_i, b_i)	$C_i \in \mathbb{R}^{60 \times 50}$, $b_i \in \mathbb{R}^{60}$ randomly generated
$x^\sharp$	$x^\sharp \in \mathbb{R}^{50}$ randomly generated
R	$R = 1.1 \ x^\sharp\ _1$
L	1
μ	0.5
Graph	Erdos-Renyi graph with connection probability 0.3
Weight matrix P	Metropolis-Hastings rule

Compared algorithms and results

- Compared algorithms and their parameters

PG-EXTRA (Shi et al. 2015)	$a = 1$
P2D2 (Alghunaim et al. 2019)	$a = 1$
DSM (Lobel & Ozdaglar 2010)	$a_t = 1/\sqrt{t+1}$
C-DDA (Duchi et al. 2010)	$a_t = 1/\sqrt{t+1}$
DDA (this work)	$a_t = 1/(0.5)^t$
$d(x)$ for DA-type algorithms	$\ x\ ^2/2$
Stochastic communication	each link is activated with probability 0.6
Relative square error (RSE)	$\frac{\sum_{i=1}^n \ x_i^{(t)} - \theta^*\ ^2}{\sum_{i=1}^n \ x_i^{(0)} - \theta^*\ ^2}$

Table 1: Mean and standard deviation of the number of iterations to achieve an accuracy of 10^{-10} for 100 random instances of the decentralized LASSO problem.

Algorithms	DDA (Bernoulli network)	C-DDA	DDA	PG-EXTRA	P2D2
No. of Iterations	318.85(± 86.70)	N/A	125.54(± 41.67)	157.30(± 41.86)	337.88(± 88.43)

① Introduction and Motivation

② Dual Averaging for Decentralized Optimization

③ Numerical Experiments

④ Summary

Summary of this work

- Proposed a dual averaging method for decentralized optimization with composite objective and stochastic communication network
- Proved linear convergence for the proposed method

Thank You!