Convergence Results for the Decentralized Gradient Descent and the Gradient-push with Event-triggered Communication

Jimyeong Kim
Department of Mathematics Sungkyunkwan University, KOREA

Ph.D. Dissertation Defense
(Advisors: Woocheol Choi and Ihyeok Seo)

Dec 1, 2023

Outline

- Chapter 1: Introduction
- Chapter 2: Preliminaries
- Chapter 3: Unconstrained Decentralized Gradient Descent
- Chapter 4: Decentralized Projected Gradient Descent
- Chapter 5: Gradient-push algorithm with Event-triggered Communication

Chapter 1: Introduction
Chapter 2: Preliminaries

Decentralized optimization

In this presentation, we consider the following minimization problem:

$$\min_{x \in \Omega} f(x) := \frac{1}{m} \sum_{i=1}^{m} f_i(x).$$

- $f_1, \cdots, f_m$, which are only known to its corresponding agent, are differentiable functions.
- Ω is closed and convex.

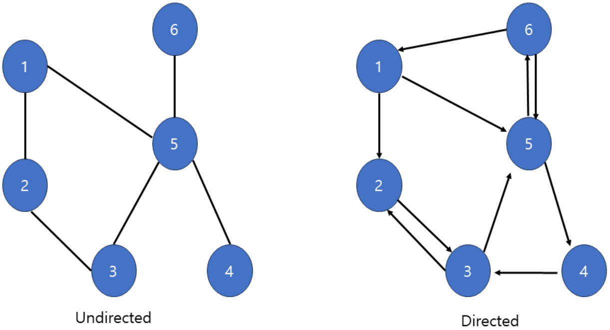
To find a solution x^*, each agent performs local computation using only its own information and that of its neighbors over a network.

GOAL

- (Consensus) $\|x_i(t) - x_j(t)\| \to 0$ as $t \to \infty$
- (Convergence) $\|x_i(t) - x_*\| \to 0$ as $t \to \infty$

Graph

A local agent informs its own information to other agents relying on shared communication networks which are characterized by a graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$.



- Chapter 3, 4: Undirected graph
- Chapter 5: Directed graph

Undirected graph and the mixing matrix

Assumption

- $\mathcal{G}$ is undirected graph, i.e. an edge $\{i, j\} \in \mathcal{E}$ is an unordered pair of distinct nodes i and j.
- $\mathcal{G}$ has no self-loop, i.e. $\{i, i\} \in \mathcal{E}$ for all $i \in \mathcal{V}$.
- $\mathcal{G}$ is connected, i.e. for any $i, j \in \mathcal{V}$, there is a sequence of edges.

We make the mixing matrix $W = \{w_{ij}\}_{1 \leq i,j \leq m}$ with respect to $\mathcal{G}$ satisfying the following assumption

Assumption

- $W = W^T$
- Null$(I - W)$ =Span(1) (doubly stochastic)
- $\beta = \rho(W - (1/m)11^T) < 1$, where $\rho(\cdot)$ denotes the spectral radius of a matrix

The role of the mixing matrix

Note that the mixing matrix W satisfies

$$\lim_{t\to\infty} W^t = W^\infty = \frac{1}{m}1^T 1,$$

where $1 = [1, \cdots, 1]^T$.
Consider the variables $x(t) = [x_1(t), \cdots, x_m(t)]^T \in \mathbb{R}^m$ with $x(0) = [x_1, \cdots, x_m]^T \in \mathbb{R}^m$ satisfying the following dynamic:

$$x(t+1) = Wx(t).$$

Then we have

$$\lim_{t\to\infty} x(t) = \lim_{t\to\infty} W^t x(0) = W^\infty x(0) = \begin{bmatrix} \frac{1}{m}\sum_{i=1}^m x_i \\ \vdots \\ \frac{1}{m}\sum_{i=1}^m x_i \end{bmatrix}$$

L. Xiao, and S. Boyd. "Fast linear iterations for distributed averaging." Systems & Control Letters 53.1 (2004): 65-78.

Chapter 3: Unconstrained Decentralized Gradient Descent

-This chapter is based on the work ‘On the convergence of decentralized gradient descent with diminishing stepsize, revisited’, submitted, with Woocheol Choi

Contributions

- We consider a diminishing stepsize and obtain the exact convergence to the optimal solution.
- We drop a convexity assumption of a local function and present an example highlighting the sharpness of the stepsize condition for obtaining uniform boundedness.

	Cost	Regularity	Learning rate	Error	Rate
A. Nedic et al. 2009	C	$\|\nabla f_i\|_\infty < \infty$	$\alpha(t) \equiv \alpha$	$f(\tilde{x}_i(t)) - f_*$	$O(\frac{1}{t}) + O(\alpha)$
D. Jakovetic et al. 2014	SC	$\|\nabla f_i\|_\infty < \infty$	$\alpha(t) = \frac{1}{t^{1/3}}$	$\|x_i(t) - x_*\|$	$O(t^{-2/3})$
I. Chen 2012	SC	L-smooth	$\alpha(t) = \frac{1}{\sqrt{k}}$	$\|x_i(t) - x_*\|$	$O(\frac{\log k}{\sqrt{k}})$
K. Yuan et al. 2016	C	L-smooth	$\alpha(t) \equiv \alpha$	$f(x_i(t)) - f^*$	$O(\frac{1}{t}) + O(\alpha)$
K. Yuan et al. 2016	SC	L-smooth	$\alpha(t) \equiv \alpha$	$\|x_i(t) - x_*\|$	$O(e^{-t}) + O(\alpha)$
Choi, K 2022	SC	L-smooth	$\alpha(t) = \frac{a}{(t+w)^p}$	$\|x_i(t) - x_*\|$	$O(t^{-p})$ if $0 < p \leq 1$

W. Choi and J. Kim, 2022. On the convergence of decentralized gradient descent with diminishing stepsize, revisited. arXiv preprint arXiv:2203.09079.

Unconstrained DGD

Decentralized gradient descent (DGD) is written as follows:

$$x_i(t+1) = \sum_{j=1}^{m} w_{ij} x_j(t) - \alpha(t) \nabla f_i(x_i(t)).$$

Yuan-Lin-Yin showed that

$$\|x_i(t) - x_*\| \leq O(e^{-ct}) + O(\alpha)$$

- $\alpha(t) \equiv \alpha < \min\left\{ \frac{1}{\mu+L}, \frac{1+\lambda_m(W)}{L} \right\}$
- The local function f_i is convex and L-smooth and the total cost function f is μ-strongly convex.
- The condition $\frac{1+\lambda_m(W)}{L}$ is needed to obtain boundedness of gradient.

K. Yuan, Q. Ling, W. Yin, On the convergence of decentralized gradient descent. SIAM J. Optim., 26 (2016), 1835–1854.

Gradient descent

Consider the minimization problem

$$\min_{x \in \mathbb{R}^d} f(x),$$

where f is smooth and strongly convex. Let us consider the following gradient descent algorithm:

$$x(t+1) = x(t) - \alpha \nabla f(x(t)).$$

Then it is well-known that

$$\|x(t) - x_*\| \leq O(e^{-ct}).$$

Limitation of the constant stepsize

We rewrite the DGD algorithm as

$$\mathrm{x}(t+1) = W\mathrm{x}(t) - \alpha \nabla F(\mathrm{x}(t)),$$

where

$$\mathrm{x}(t) = \begin{bmatrix} x_1(t) \\ x_2(t) \\ \vdots \\ x_m(t) \end{bmatrix} \text{ and } \nabla F(\mathrm{x}(t)) = \begin{bmatrix} \nabla f_1(x_1(t)) \\ \nabla f_2(x_2(t)) \\ \vdots \\ \nabla f_m(x_m(t)) \end{bmatrix}.$$

Taking the limit, we have

$$\mathrm{x}(\infty) = W\mathrm{x}(\infty) - \alpha \nabla F(\mathrm{x}(\infty)),$$

where $\mathrm{x}(\infty) = \lim_{t\to\infty} \mathrm{x}(t)$.

Limitation of the constant stepsize

Assuming that the consensus is achieved, i.e. $\mathrm{x}(\infty) = W\mathrm{x}(\infty)$, it follows that

$$\nabla F(\mathrm{x}(\infty)) = 0,$$

which is equivalent to

$$\nabla f_i(x_i(\infty)) = 0, \text{ for all } i \in \{1, 2, \cdots, m\}.$$

The consensus implies that $x_i(\infty)$ concurrently minimizes f_i for all $i \in \{1, 2, \cdots, m\}$, which is generally not possible since the first optimality condition is

$$\nabla f(x_*) = \sum_{i=1}^{m} \nabla f_i(x_*) = 0.$$

Consensus result

Theorem

Let $\alpha(t)$ be a non-increasing sequence satisfying:

$$\alpha(0) = \frac{a}{w^p} \leq \min\left\{\frac{2}{\mu+L}, \frac{\eta(1-\beta)}{L(\eta+L)}\right\},$$

where $\eta = L\mu/(L+\mu)$. Then for all $t \geq 0$ we have

$$\|\mathrm{x}(t) - \bar{\mathrm{x}}(t)\| \leq \frac{d}{1-\beta}\alpha([t/2]) + \beta^t\|\mathrm{x}(0) - \bar{\mathrm{x}}(0)\| + \frac{\beta^{t/2}d}{1-\beta}\alpha(0),$$

where $\bar{x}(t) = \frac{1}{m}\sum_{i=1}^{m}$, $\mathrm{x}(t) = [x_1(t), \cdots, x_m(t)]$ and $\bar{\mathrm{x}}(t) = [\bar{x}(t), \cdots, \bar{x}(t)]$.

W. Choi and J. Kim, 2022. On the convergence of decentralized gradient descent with diminishing stepsize, revisited. arXiv preprint arXiv:2203.09079.

Convergence results

Theorem (Informal)
Let $\alpha(t) = \frac{a}{(t+w)^p}$. Then, for all $t \geq 0$ we obtain the following estimates:

1. $(0 < p < 1)$
   $\|\bar{x}(t) - x_*\| \leq \frac{Ca}{1-\beta} \cdot ([t/2] + w - 1)^{-p} + C_1(w, a)e^{-t}$.

2. $(p = 1)$
   $\|\bar{x}(t) - x_*\| \leq \left(\frac{w}{t+w}\right)^{\eta a} \|x(0) - x_*\| + \frac{C}{(1-\beta)} \cdot \frac{a}{(t+w+1)} + \frac{C_2(w,a)}{(t+w)^{l+1}} + \frac{\beta^{t/2} C_3(w,a)}{t-1+w}$.

Roughly, the constants $C_1(w, a), C_3(w, a) \approx \frac{q}{w^p}$ and $C_2(w, a) \approx \frac{w^p}{a}$. In addition $\eta a > 1$.

This theorem implies a convergence rate of $O(t^{-p})$.

W. Choi and J. Kim, 2022. On the convergence of decentralized gradient descent with diminishing stepsize, revisited. arXiv preprint arXiv:2203.09079.

Sketch of the proof

By the DGD algorithm, it follows that

$$
\begin{aligned}
&\|\bar{x}(t+1) - x_*\| \\
&= \left\| \bar{x}(t) - x_* - \frac{\alpha(t)}{m} \sum_{i=1}^{m} \nabla f_i(x_i(t)) \right\| \\
&\le \underbrace{\left\| \bar{x}(t) - x_* - \frac{\alpha(t)}{m} \sum_{i=1}^{m} \nabla f_i(\bar{x}(t)) \right\|}_{Gradient\ \ Descent} + \underbrace{\frac{\alpha(t)}{m} \sum_{i=1}^{m} \|\nabla f_i(\bar{x}(t)) - \nabla f_i(x_i(t))\|}_{Smoothness}.
\end{aligned}
$$

Since the total cost function $f(x) = \frac{1}{m}\sum_{i=1}^{m} f_i(x)$ is strongly convex and a local cost function f_i is smooth, it follows that

$$
\begin{aligned}
\|\bar{\mathrm{x}}(t+1) - \mathrm{x}_*\| &\le (1 - \eta\alpha(t))\|\bar{\mathrm{x}}(t) - \mathrm{x}_*\| + \underbrace{L\alpha(t)\|\mathrm{x}(t) - \bar{\mathrm{x}}(t)\|}_{Consensus} \\
&\le (1 - C\alpha(t))\|\bar{\mathrm{x}}(t) - \mathrm{x}_*\| + C'\alpha(t)^2 + C''\beta^t
\end{aligned}
$$

Remarks for the results

- The assumption $\|\nabla F(\mathrm{x}(t))\| \leq d$ can be justified by the uniform boundedness of the sequence. Specifically, there exists a finite value $R > 0$ such that both $\|\bar{\mathrm{x}}(t) - \mathrm{x}_*\| \leq R$ and $\|\mathrm{x}(t) - \bar{\mathrm{x}}(t)\| < R$ hold.

- To obtain the uniform boundedness, the condition $\alpha(t) \leq \frac{\eta(1-\beta)}{L(\eta+L)}$ is needed ($\eta = (\mu L)/(\mu + L)$).

- [K. Yuan et al (2016)] showed the uniform boundedness result when $\alpha(t) \leq \frac{1+\lambda_m(W)}{L}$, which is less restrictive than $\alpha(t) \leq \frac{\eta(1-\beta)}{L(\eta+L)}$.

- On the other hand, the assumptions on [Choi, K. (2022)] allow each local cost function to be nonconvex.

Sharpness of the condition for uniform boundedness

Consider the following functions:

$$f_1(x) = \frac{L}{2}x^2 \text{ and } f_2(x) = -\left(\frac{L}{2} - \mu\right)x^2, \; x \in \mathbb{R},$$

where L and μ are positive values satisfying $L > 2\mu > 0$. Then the total cost function $f = f_1 + f_2$ is strongly convex. We take a value $\gamma \in (0, 1/2]$ and set a doubly stochastic matrix W by

$$W = \begin{bmatrix} 1-\gamma & \gamma \\ \gamma & 1-\gamma \end{bmatrix}$$

Lemma

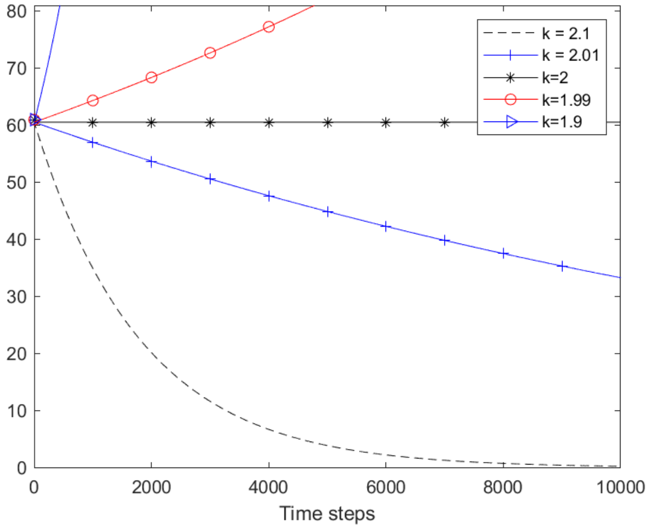
If $\alpha > \frac{2\mu\gamma}{L(L-2\mu)}$, then the sequence $\{(x_1(t), x_2(t))\}$ generated by DGD with any initial data $(x_1(0), x_2(0)) \in (\mathbb{R} \setminus \{0\})^2$ diverges.

Since $\frac{2\mu\gamma}{L(L-2\mu)} / \frac{\eta(1-\beta)}{L(\eta+L)} = \frac{L-2\mu}{L+2\mu}$, it follows that

$$\lim_{L\to\infty} \frac{L-2\mu}{L+2\mu} = \lim_{\mu\to 0} \frac{L-2\mu}{L+2\mu} = 1$$

Sharpness of the condition for uniform boundedness

- Set $f_1(x) = a_1 x^2$ and $f_2(x) = a_2 x^2$. Then we have $\frac{2\mu\gamma}{L(L-2\mu)} = \frac{\gamma(a_1+a_2)}{2a_1a_2}$
- We test the following stepsize $\frac{\gamma(a_1+a_2)}{ka_1a_2}$



Chapter 4: Constrained Decentralized Gradient Descent

-This chapter is based on the work ‘On the convergence analysis of the decentralized projected gradient descent’, submitted to SIAM J. Optim under major revision, with Woocheol Choi.

Contributions

- We obtain an $O(\sqrt{\alpha})$ error.
- We obtain an exact convergence result for a diminishing stepsize.

	Cost	Regularity	Learning rate	Error	Rate
S.S. Ram et al. 2010	C	$\|\nabla f_i\|_\infty < \infty$	$\sum \alpha(t) = \infty$ $\sum \alpha(t)^2 < \infty$	$\|x_i(t) - x_*\|$	$o(1)$
I.-A. Chen et al. 2012	C	$\|\nabla f_i\|_\infty < \infty$	$\alpha(t) = ct^{-\alpha}$	$f(x_i(t)) - f^*$	$O(t^{-p})$ if $p \in (0, 1/2)$ $O(\frac{\log t}{\sqrt{t}})$ if $p = 1/2$ $O(t^{p-1})$ if $p \in (1/2, 1)$
S. Liu et al. 2017	SC	$\|\nabla f_i\|_\infty < \infty$	$\alpha(t) = ct^{-1}$	$\|x_i(t) - x_*\|$	$O(1/\sqrt{t})$
C. Liu et al 2020	SC	L-smooth	$\alpha(t) \equiv \alpha$	$\|x_i(t) - x_*\|$	$O(e^{-t}) + O(\alpha) + O(1)$
Choi, K 2023	SC	L-smooth	$\alpha(t) \equiv \alpha$	$\|x_i(t) - x_*\|$	$O(e^{-t}) + O(\sqrt{\alpha})$
Choi, K 2023	SC	L-smooth	$\alpha(t) = ct^{-p}$	$\|x_i(t) - x_*\|$	$O(t^{-p/2})$ if $p \in (0, 1]$

W. Choi and J. Kim, 2023. On the convergence analysis of the decentralized projected gradient descent, arXiv:2303.08412.

Contributions

	Cost	Regularity	Learning rate	Error	Rate
S.S. Ram et al. 2010	C	$\|\nabla f_i\|_\infty < \infty$	$\sum \alpha(t) = \infty$ $\sum \alpha(t)^2 < \infty$	$\|x_i(t) - x_*\|$	$o(1)$
I.-A. Chen et al. 2012	C	$\|\nabla f_i\|_\infty < \infty$	$\alpha(t) = ct^{-\alpha}$	$f(x_i(t)) - f^*$	$O(t^{-p})$ if $p \in (0, 1/2)$ $O(\frac{\log t}{\sqrt{t}})$ if $p = 1/2$ $O(t^{p-1})$ if $p \in (1/2, 1)$
S. Liu et al. 2017	SC	$\|\nabla f_i\|_\infty < \infty$	$\alpha(t) = ct^{-1}$	$\|x_i(t) - x_*\|$	$O(1/\sqrt{t})$
C. Liu et al 2020	SC	L-smooth	$\alpha(t) \equiv \alpha$	$\|x_i(t) - x_*\|$	$O(e^{-t}) + O(\alpha) + O(1)$
Choi, K 2023	SC	L-smooth	$\alpha(t) \equiv \alpha$	$\|x_i(t) - x_*\|$	$O(e^{-t}) + O(\sqrt{\alpha})$
Choi, K 2023	SC	L-smooth	$\alpha(t) = ct^{-p}$	$\|x_i(t) - x_*\|$	$O(t^{-p/2})$ if $p \in (0, 1]$

For the unconstrained case, i.e. $\Omega = \mathbb{R}^d$, Yuan-Lin-Yin [K. Yuan et al (2016)] showed that a convergence error of $O(\alpha)$.

Question: Can we improve the convergence error from $O(\sqrt{\alpha})$ to $O(\alpha)$?

W. Choi and J. Kim, 2023. On the convergence analysis of the decentralized projected gradient descent, arXiv:2303.08412.

Contributions

- We obtain an $O(\sqrt{\alpha})$ error.
- We obtain an exact convergence result for a diminishing stepsize.
- We present specific examples that achieve $O(\alpha)$ errors

	Cost	Regularity	Learning rate	Error	Rate
S.S. Ram et al. 2010	C	$\|\nabla f_i\|_\infty < \infty$	$\sum \alpha(t) = \infty$ $\sum \alpha(t)^2 < \infty$	$\|x_i(t) - x_*\|$	$o(1)$
I.-A. Chen et al. 2012	C	$\|\nabla f_i\|_\infty < \infty$	$\alpha(t) = ct^{-\alpha}$	$f(x_i(t)) - f^*$	$O(t^{-p})$ if $p \in (0, 1/2)$ $O(\frac{\log t}{\sqrt{t}})$ if $p = 1/2$ $O(t^{p-1})$ if $p \in (1/2, 1)$
S. Liu et al. 2017	SC	$\|\nabla f_i\|_\infty < \infty$	$\alpha(t) = ct^{-1}$	$\|x_i(t) - x_*\|$	$O(1/\sqrt{t})$
C. Liu et al 2020	SC	L-smooth	$\alpha(t) \equiv \alpha$	$\|x_i(t) - x_*\|$	$O(e^{-t}) + O(\alpha) + O(1)$
Choi, K 2023	SC	L-smooth	$\alpha(t) \equiv \alpha$	$\|x_i(t) - x_*\|$	$O(e^{-t}) + O(\sqrt{\alpha})$
Choi, K 2023	SC	L-smooth	$\alpha(t) = ct^{-p}$	$\|x_i(t) - x_*\|$	$O(t^{-p/2})$ if $p \in (0, 1]$
Choi, K 2023	1-d example	L-smooth	$\alpha(t) \equiv \alpha$	$\|x_i(t) - x_*\|$	$O(e^{-t}) + O(\alpha)$
Choi, K 2023	Half-space example	L-smooth	$\alpha(t) \equiv \alpha$	$\|x_i(t) - x_*\|$	$O(e^{-t}) + O(\alpha)$

Decentralized gradient descent

Recall that

$$x_i(t+1) = \sum_{j=1}^{m} w_{ij} x_j(t) - \alpha \nabla f_i(x_i(t)).$$

Summing over the range from $i = 1$ to m, it follows that

$$\bar{x}(t+1) = \bar{x}(t) - \frac{\alpha}{m} \sum_{i=1}^{m} \nabla f_i(x_i(t)),$$

where $\bar{x}(t) = \frac{1}{m}\sum_{i=1}^{m} x_i(t)$. Then we can easily obtain the following inequality.

$$\begin{aligned}
\|\bar{x}(t+1) - x_*\| &\le \Big(1 - \frac{\mu L}{\mu + L}\alpha\Big)\|\bar{x}(t) - x_*\| + \frac{L\alpha}{n}\sum_{i=1}^{m}\|x_i(t) - \bar{x}(t)\| \\
&\le (1 - c\alpha)\|\bar{x}(t) - x_*\| + C\alpha^2 \\
&\le (1 - c\alpha)^{t+1}\|\bar{x}(0) - x_*\| + O(\alpha)
\end{aligned}$$

Decentralized projected gradient descent

Decentralized projected gradient descent is written as follows:

$$x_i(t+1) = \mathcal{P}_{\Omega_i}\left[\sum_{j=1}^{m} w_{ij} x_j(t) - \alpha(t)\nabla f_i(x_i(t))\right].$$

Here, $\mathcal{P}_{\Omega_i}$ represents the projection of a vector y onto the set Ω_i, defined as

$$\mathcal{P}_{\Omega_i}[y] = \arg\min_{x\in\Omega_i} \|x - y\|.$$

We assume that $\Omega = \Omega_i$ for all $i \in \{1, 2, \cdots, m\}$.
Unlike the DGD algorithm, we have

$$\bar{x}(t+1) \neq \mathcal{P}_{\Omega}\left[\bar{x}(t) - \frac{\alpha(t)}{m}\sum_{i=1}^{m} \nabla f_i(x_i(t))\right].$$

$O(1)$-convergence

We can write the DPG in the following way:

$$x_i(t+1) = \sum_{j=1}^{m} w_{ij} x_j(t) - \alpha \nabla f_i(x_i(t)) + \phi_i(t)$$

where $\phi_i(t)$ is the difference between DGD and DPG defined as follows:

$$\phi_i(t) = \underbrace{\sum_{j=1}^{m} w_{ij} x_j(t) - \alpha(t) \nabla f_i(x_i(t))}_{\text{DGD}} - \underbrace{\mathcal{P}_\Omega \left[\sum_{j=1}^{m} w_{ij} x_j(t) - \alpha(t) \nabla f_i(x_i(t)) \right]}_{\text{DPG}}.$$

Averaging (1) for $1 \leq k \leq n$ one has

$$\bar{x}(t+1) = \bar{x}(t) - \frac{\alpha}{m} \sum_{i=1}^{m} \nabla f_i(x_i(t)) + \frac{1}{m} \sum_{i=1}^{m} \phi_i(t).$$

C. Liu et al., Distributed event-triggered gradient method for constrained convex optimization, IEEE Transactions on Automatic Control, 65(2):778-785, 2019

$O(1)$-convergence

Using the contraction property of the projection operator, it follows that

$$
\begin{aligned}
&\|\bar{x}(t+1) - x_*\| \\
&\leq \underbrace{\left\| \bar{x}(t) - \frac{\alpha}{m} \sum_{i=1}^{m} \nabla f_i(x_i(t)) - x_* - \frac{\alpha}{m} \sum_{i=1}^{n} \nabla f_i(x_*) \right\|}_{\text{DGD part}} + \underbrace{\left\| \frac{1}{m} \sum_{i=1}^{m} \phi_i(t) \right\|}_{\text{Error part} < O(\alpha)}
\end{aligned}
$$

. Subsequently, applying L-smoothness and strong convexity, we have

$$
\begin{aligned}
\|\bar{x}(t+1) - x_*\| &\leq (1 - c\alpha)\|\bar{x}(t) - x_*\| + \frac{L\alpha}{n} \sum_{i=1}^{n} \|\bar{x}(t) - x_i(t)\| + C\alpha \\
&\leq \underbrace{(1 - c\alpha)^{t+1}\|\bar{x}(0) - x_*\| + O(\alpha)}_{DGD} + \underbrace{O(1)}_{Error}
\end{aligned}
$$

C. Liu et al., Distributed event-triggered gradient method for constrained convex optimization, IEEE Transactions on Automatic Control, 65(2):778-785, 2019

The idea for the $O(\sqrt{\alpha})$-convergence

In unconstrained DGD, we follow these steps:

1. Average the DGD algorithm
2. Estimate Gradient Descent (Linear convergence).
3. Estimate Consensus ($O(\alpha)$).

In DPG, we follow these steps:

1. Average the DPG algorithm
2. Estimate Gradient Descent. (Linear convergence)
3. Estimate Consensus. ($O(\alpha)$)
4. Estimate error between the DGD and DPG ($O(1)$)

Question) Instead of averaging, can we directly analyze the DPG algorithm?
$\Rightarrow$ Estimate $\|\mathrm{x}(t) - \mathrm{x}_*\|^2$ instead of $\|\bar{\mathrm{x}}(t) - \mathrm{x}_*\|^2$.

The idea for the $O(\sqrt{\alpha})$ convergence

Using the fact $x_* = \mathcal{P}_\Omega[x_* - \alpha \nabla f(x_*)]$ and contraction property of Projection operator, it follows that

$$\begin{aligned}\|\mathrm{x}(t+1) - \mathrm{x}_*\|^2 &\leq \sum_{i=1}^m \left\| \sum_{j=1}^m w_{ij} x_j(t) - \alpha \nabla f_i(x_i(t)) - x_* - \alpha \nabla f(x_*) \right\|^2 \\ &\leq C\alpha^2 \|\mathrm{x}(t) - \bar{\mathrm{x}}(t)\|^2 + (1 - C\alpha)\|\bar{\mathrm{x}}(t) - \mathrm{x}_*\|^2 + C\alpha^2\end{aligned}$$

Also using $\|\mathrm{x}(t) - \mathrm{x}_*\|^2 = \|\mathrm{x}(t) - \bar{\mathrm{x}}(t)\|^2 + \|\bar{\mathrm{x}}(t) - \mathrm{x}_*\|^2$, we can obtain the following inequality.

$$\begin{aligned}\|\mathrm{x}(t+1) - \mathrm{x}_*\|^2 &\leq (1 - c\alpha)\|\mathrm{x}(t) - \mathrm{x}_*\|^2 + C\alpha^2 \\ &\leq (1 - c\alpha)^{t+1}\|\mathrm{x}(0) - \mathrm{x}_*\|^2 + \frac{C}{c}\alpha,\end{aligned}$$

which implies $O(\sqrt{\alpha})$-convergence.

W. Choi and J. Kim, 2023. On the convergence analysis of the decentralized projected gradient descent, arXiv:2303.08412.

Towards $O(\alpha)$ convergence

- In the unconstrained case, we can separately estimate $\|x(t) - \bar{x}(t)\|$ and $\|\bar{x}(t) - x_*\|$.
- Conversely, in the constrained case, we simultaneously estimate $\|x(t) - \bar{x}(t)\|$ and $\|\bar{x}(t) - x_*\|$ using the relation

$$\|x(t) - x_*\|^2 = \|x(t) - \bar{x}(t)\|^2 + \|\bar{x}(t) - x_*\|^2$$

- To circumvent this challenge, we have devised a new approach that analyzes the sequence $x_k(t)$ by partitioning the coordinates into two distinct segments: one influenced by the projection operator and the other unaffected by it.

Half-space example

We investigate the case where the domain is a half-space:

$$\Omega = \{(\tilde{x}, x[d]) \,|\, \tilde{x} \in \mathbb{R}^{d-1}, x[d] \geq 0\} \subseteq \mathbb{R}^d.$$

Here $x[k] \in \mathbb{R}$ denotes the *k-th* component of the vector $x \in \mathbb{R}^d$. Let $x_* = (\tilde{x}_*, x_*[d]) \in \mathbb{R}^d$ be a solution, i.e. $x_* = \arg\min_{x \in \Omega} f(x)$. We make the following assumption for a solution.

Assumption
The minimizer of f is on the boundary of Ω, specifically, $x_* = (\tilde{x}_*, 0)$ with some $\tilde{x}_* \in \mathbb{R}^{d-1}$ and the minimizer $x_* = (\tilde{x}_*, 0)$ satisfies

$$\partial_d f(x_*) = \frac{1}{m} \sum_{i=1}^{m} \partial_d f_i(x_*) \geq \omega > 0.$$

Half-space example

Theorem

Suppose that the domain Ω is the half space $\mathbb{R}^{d-1} \times \mathbb{R}_+$ with any dimension $d \geq 1$. If the stepsize $\alpha > 0$ satisfies $\alpha \leq \frac{2}{L+\mu}$, we have

$$\lim_{t \to \infty} \|x_k(t) - x_*\| = O(\alpha) \quad \forall\, 1 \leq k \leq n.$$

W. Choi and J. Kim, in preparation

Proof

Lemma

There exists a time T such that for all $t \geq T$, one of the following two cases holds.

- *Case 1.* Assume that $\sum_{j=1}^{m} w_{kj} x_j(t) - \eta \nabla f_k(x_k(t))$ belongs to Ω for all $1 \leq k \leq m$. Then we have

$$\bar{x}(t+1)[d] \leq \bar{x}(t)[d] - \frac{\omega\alpha}{2}.$$

- *Case 2.* Assume that $\sum_{j=1}^{m} w_{kj} x_j(t) - \eta \nabla f_k(x_k(t))$ does not belong to Ω for some $1 \leq k \leq m$. Then we have

$$\bar{x}(t+1)[d] \leq \left(\beta^{t+1} \|\mathrm{x}(0) - \bar{\mathrm{x}}(0)\|^2 + \frac{3(L^2R^2 + nD^2)\alpha^2}{(1-\beta)^2}\right)^{1/2}.$$

This lemma implies that there exists a time $\tilde{T}$ such that for all $t \geq \tilde{T}$,

$$\bar{x}(t+1)[d] \leq O(\alpha).$$

Proof (Continued)

Lemma

There exists $c_1, c_2 > 0$ *such that for* $t \geq \tilde{T}$ *we have*

$$\left\|y(t) - \tilde{x}_*\right\|^2 \leq (1 - c_1\alpha)^t \left\|y(0) - \tilde{x}_*\right\|^2 + \frac{c_2}{c_1}\alpha^2,$$

where $y(t) \in \mathbb{R}^{d-1}$ *denotes the first* $d-1$ *coordinates of* $\bar{x}(t)$*.*

Therefore, we conclude that

$$\begin{aligned} \|\bar{x}(t) - x_*\|^2 &= \|y(t) - \tilde{x}_*\|^2 + |x_k(t)[d] - 0|^2 \\ &\leq (1 - c_1\alpha)^t \left\|y(0) - \tilde{x}_*\right\|^2 + O(\alpha^2). \end{aligned}$$

Remarks for the DPG

- There exist algorithms that linearly and exactly converge to the optimal point with a constant stepsize, such as the P-DIGing.
- P-DIGing is a version of the DIGing algorithm for the constrained problem.
- What is a point of view to study the DPG? What is the strength of the DPG?

P-DIGing: Z. Dong, S. Mao, W. Du, and Y. Tang, Distributed constrained optimization with linear convergence rate, in 2020 IEEE 16th International Conference on Control & Automation (ICCA), IEEE, 2020, pp. 937–942.

DIGing: A. Nedic, A. Olshevsky, and W. Shi, Achieving geometric convergence for distributed optimization over time-varying graphs, SIAM Journal on Optimization, 27 (2017), pp. 2597–2633.

Numerical Results

We consider the following decentralized non-negative least squares problem with n agents

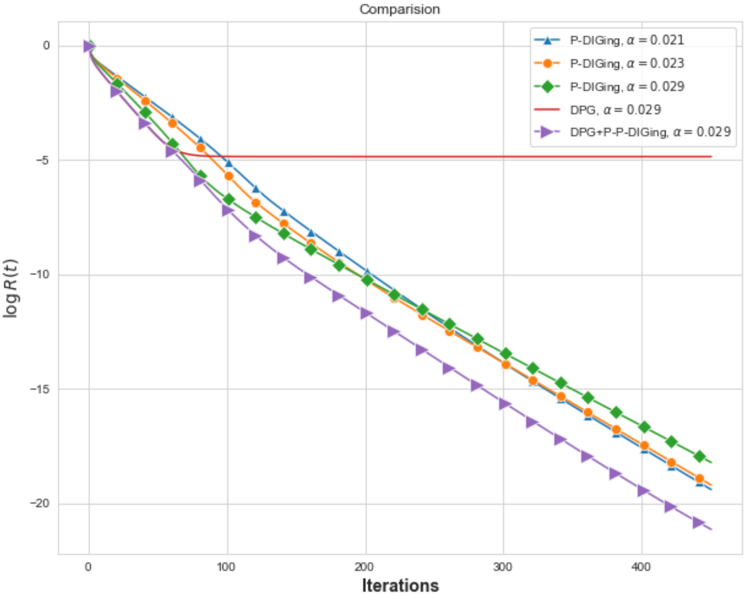
$$\min_{x\in\Omega} \frac{1}{m}\sum_{i=1}^{m}\frac{1}{2}\|q_i - p_i^T x\|^2.$$

- $\Omega = \{(x[1], x[2], \cdots, x[d]) \mid x[k] \in \mathbb{R}_+\}$.
- The variables $p_i \in \mathbb{R}^{d\times p}$ and $q_i \in \mathbb{R}^p$ are randomly chosen from the uniform distribution on $[0,1]$.
- We set $d = 10$, $p = 5$, and $n = 30$
- We use the Watts and Strogatz model to construct a connected graph.
- We measure

$$R(t) = \frac{\sum_{i=1}^m \|x_i(t) - x_*\|^2}{\sum_{i=1}^m \|x_i(0) - x_*\|^2}$$

D. J. Watts and S. H. Strogatz, Collective dynamics of small-world networks, nature, 393 (1998), pp. 440-442.

Numerical results



Numerical Results

We consider the following decentralized logistic regression problem with the MNIST:

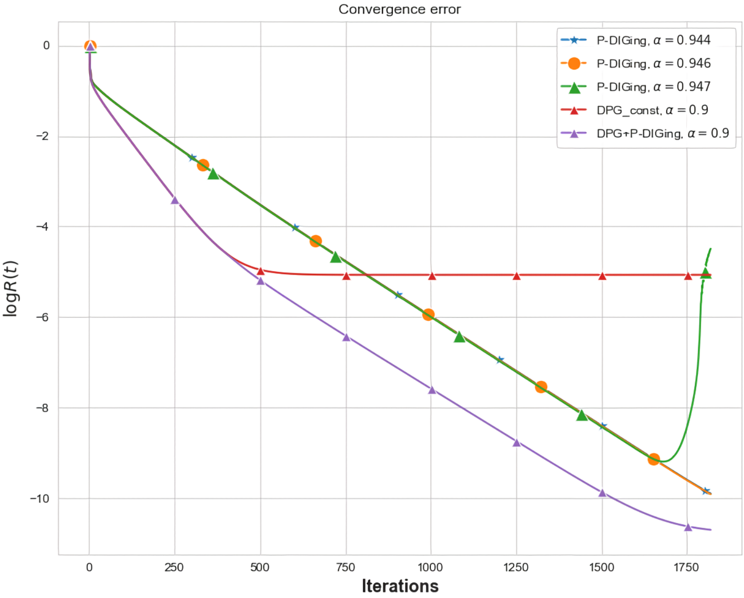
$$\min_{x\in\Omega} \sum_{i=1}^{n} f_i(x),$$

where $f_i(x) \sum_{j=1}^{k} \log[1 + \exp((-x^T \tau_j)\phi_j] + \frac{\alpha}{2}\|x\|^2$.

- $\Omega = [-1, 5]^{784}$
- $\tau_j \in \mathbb{R}^{784}$ is the feature vector and $\phi_j \in [-1, 1]$.
- We set $n = 20$, $k = 50$ and $\alpha = 0.01$.
- We use the Watts and Strogatz model to construct a connected graph.

D. J. Watts and S. H. Strogatz, Collective dynamics of small-world networks, nature, 393 (1998), pp. 440-442.

Numerical results



Chapter 5: Gradient-push algorithm with Event-triggered Communication

-This chapter is based on the work 'Gradient-push algorithm for distributed optimization with event-triggered communication', IEEE Access, vol 11 (2023), pp. 517-534, with Woocheol Choi

Time-varying directed graph and the mixing matrix

- The communication pattern is characterized by a time-varying directed graph $\mathcal{G}(t) = (\mathcal{V}, \mathcal{E}(t))$, which is uniformly strongly connected, i.e. there exists a value $B \in \mathbb{N}$ such that the graph with edge set $\cup_{i=kB}^{(k+1)B-1} \mathcal{E}(i)$ is connected for any $k \geq 0$.

- Define the mixing matrix $A(t)$ such that $[A(t)]_{ij} = a_{ij}(t)$, where

$$a_{ij}(t) = \begin{cases} 1/d_j^{\text{out}}(t), & \text{if } i \in N_j^{\text{out}}(t) \\ 0, & \text{otherwise,} \end{cases}$$

  where $N_i^{\text{out}}(t) = \{j | (i, j) \in \mathcal{E}(t)\} \cup \{i\}$ and $d_i^{\text{out}}(t) = |N_i^{\text{out}}(t)|$.

- The mixing matrix $A(t)$ is a column stochastic matrix, i.e. $1^T A(t) = 1^T$ for any $j \in \mathcal{V}$.

Gradient-push algorithm

One fundamental algorithm is the Gradient-Push algorithm [A. Nedić and A. Olshevsky (2014)].

Choose $y(0) = [1, \cdots, 1]$ and for $t = 0, 1, \cdots$

$$w_i(t+1) = \sum_{j \in N_i^{\text{in}}(t)} a_{ij}(t) x_j(t)$$

$$y_i(t+1) = \sum_{j \in N_i^{\text{in}}(t)} a_{ij}(t) y_j(t)$$

$$z_i(t+1) = \frac{w_i(t+1)}{y_i(t+1)}$$

$$x_i(t+1) = w_i(t+1) - \alpha(t+1) \nabla f_i(z_i(t+1))$$

Nedić, Angelia, and Alex Olshevsky. "Distributed optimization over time-varying directed graphs." IEEE Transactions on Automatic Control 60.3 (2014): 601-615.

Simple case

- Let A be a column stochastic matrix, i.e. $1^T A = 1^T$.
- There exists $\pi = (\pi_1, \cdots, \pi_m)$ such that

$$A^\infty = \lim_{t\to\infty} A^t = \begin{bmatrix} \pi_1 & \pi_1 & \cdots & \pi_1 \\ \pi_2 & \pi_2 & \cdots & \pi_2 \\ \vdots & \vdots & \ddots & \vdots \\ \pi_m & \pi_m & \cdots & \pi_m \end{bmatrix},$$

  where $\sum_{i=1}^m \pi_i = 1$.
- For a doubly stochastic matrix W, we have

$$\lim_{t\to\infty} W^t = \frac{1}{m} 11^T.$$

Perron-Frobenius theorem

Simple case

Consider the dynamic

$$x(t+1) = Ax(t)$$
$$y(t+1) = Ay(t),$$

with $x(0) = (x_1, \cdots, x_n)^T$ and $y(0) = (1, \cdots, 1)^T$. Then we have

$$x(\infty) = \lim A^t x(0) = A^\infty x(0) = [\pi_1 \sum_{i=1}^{n} x_i, \cdots, \pi_n \sum_{i=1}^{n} x_i]^T$$
$$y(\infty) = \lim A^t y(0) = A^\infty y(0) = [n\pi_1, n\pi_2, \cdots, n\pi_n]^T.$$

Eventually, consensus is attained, meaning that

$$z(\infty) = \lim z(t) = \lim \frac{x(t)}{y(t)} = \frac{x(\infty)}{y(\infty)} = \begin{bmatrix} \frac{1}{n}\sum_{i=1}^{n} x_i \\ \vdots \\ \frac{1}{n}\sum_{i=1}^{n} x_i \end{bmatrix}$$

Gradient-push algorithm

Choose $y(0) = [1, \cdots, 1]$ and for $t = 0, 1, \cdots$

$$w_i(t+1) = \sum_{j=1}^{m} a_{ij}(t) x_j(t) \text{ (Communication)}$$

$$y_i(t+1) = \sum_{j=1}^{m} a_{ij}(t) y_j(t) \text{ (Communication)}$$

$$z_i(t+1) = \frac{w_i(t+1)}{y_i(t+1)} \text{ (Consensus)}$$

$$x_i(t+1) = w_i(t+1) - \alpha(t+1) \nabla f_i(z_i(t+1)) \text{ (Optimization).}$$

Each agent needs to communicate with its neighbors at every iteration. $\Rightarrow$ Power Consumption

Shnayder, Victor, et al. "Simulating the power consumption of large-scale sensor network applications." Proceedings of the 2nd international conference on Embedded networked sensor systems. 2004.

Event-triggered communication

- Let $\tau(t), \zeta(t) \geq 0$ be the thresholds.
- For each time t, each agent i sends the states $x_i(t+1)$ and $y_i(t+1)$ to its neighbors respectively if

$$\begin{aligned} \|x_i(t+1) - \hat{x}_i(t)\| &\geq \tau(t) \\ |y_i(t+1) - \hat{y}_i(t)| &\geq \zeta(t), \end{aligned}$$

where $\hat{x}_i(t)$ and $\hat{y}_i(t)$ the latest sent states.

Gradient-push algorithm with event-triggered

Choose $y(0) = [1, \cdots, 1]$ and for $t = 0, 1, \cdots$

$$w_i(t+1) = \sum_{j=1}^{m} a_{ij}(t)\hat{x}_j(t)$$

$$y_i(t+1) = \sum_{j=1}^{m} a_{ij}(t)\hat{y}_j(t)$$

$$z_i(t+1) = \frac{w_i(t+1)}{y_i(t+1)}$$

$$x_i(t+1) = w_i(t+1) - \alpha(t+1)\nabla f_i(z_i(t+1))$$

if $\|x_i(t+1) - \hat{x}_i(t)\| \geq \tau(t+1)$ then set $\hat{x}_i(t+1) = x_i(t+1)$.
else set $\hat{x}_i(t+1) = \hat{x}_i(t)$ (do not send).
if $|y_i(t+1) - \hat{y}_i(t)| \geq \zeta(t+1)$ then set $\hat{y}_i(t+1) = y_i(t+1)$.
else set $\hat{y}_i(t+1) = \hat{y}_i(t)$ (do not send).

Assumptions

(A) For each $i \in \{1, \cdots, m\}$, there exists $D_i > 0$ such that

$$\|\nabla f_i(x)\| \le D_i \quad \forall x \in \mathbb{R}^d.$$

We set $D = \max_{1 \le i \le m} D_i$.

(B) The sequence of graph $\{\mathcal{G}(t)\}_{t \ge 0}$ is uniformly strongly connected.

(C) The sequence of stepsize $\{\alpha(t)\}_{t \in \mathbb{N}}$ is monotonically non-increasing and satisfies

$$\sum_{t=1}^{\infty} \alpha(t) = \infty, \ \sum_{t=1}^{\infty} \alpha(t)^2 < \infty.$$

Assumptions (continued)

(D) The sequence of event-triggering thresholds $\{\tau(t)\}_{t\in\mathbb{N}}$ is non-increasing and satisfies

$$\sum_{t=0}^{\infty} \tau(t) < \infty.$$

(E) The sequence of event-triggering thresholds $\{\zeta(t)\}_{t\in\mathbb{N}}$ is monotonically non-increasing and satisfies

$$\sum_{t=0}^{\infty} t^{3/2}\zeta(t) < \infty, \quad \sum_{t=0}^{\infty} \zeta(t) < 1.$$

Main results

Theorem (asymptotic convergence)
The sequence $\{z_i(t)\}_{t\in\mathbb{N}}$ generated by Gradient-push algorithm with event-triggered communication satisfies

$$\lim_{t\to\infty} z_i(t) = x^* \text{ for all } i \text{ and for some } x^* \in X^*.$$

Theorem (Informal)
Define $\alpha(t) = \frac{1}{\sqrt{t}}$ and $\tilde{z}_i(t) = \sum_{s=0}^{t} a_t(s) z_i(s)$, where $\sum_{s=0}^{t} a_t(s) = 1$. Then, we obtain the following estimate:

$$f(\tilde{z}_i(T+1)) - f(x^*) \leq O(\log(T)/\sqrt{T}).$$

J. Kim and W. Choi, Gradient-push algorithm for distributed optimization with event-triggered communications, IEEE Access, vol 11 (2023) 517-534

Sketch of the proof

Non-event-triggered

- (Consensus) $\|z_i(t) - \bar{x}(t)\| \to 0$ as $t \to \infty$
- (Convergence) $\|\bar{x}(t) - x_*\| \to 0$ as $t \to \infty$

Event-triggered

- (Consensus) $\|z_i(t) - C\bar{x}(t)\| \to 0$ as $t \to \infty$
- (Convergence) $\|C\bar{x}(t) - x_*\| \to 0$ as $t \to \infty$

It is worth mentioning that $C = 1$ when $\tau(t), \zeta(t) \equiv 0$.

Consensus

Lemma

Suppose that Assumptions (D) and (E) hold. Then there exists a stochastic vector $\phi(t)$ such that

$$\|y(t) - m_{\zeta}\phi(t)\| \leq \beta(t).$$

Here $m_{\zeta} = m + \sum_{s=1}^{\infty} 1_m^T \theta(s)$, where $\theta(s) = \hat{y}(s) - y(s)$ and $\beta(t)$. In addition, we have

$$\lim_{t\to\infty} t^{3/2}\beta(t) = 0.$$

If $\zeta(t) \equiv 0$ which is equivalent to non-event-triggered case, then this lemma can be written as

$$\|y(t) - m\phi(t)\| \leq C\lambda^t.$$

For the non-time varying case, we can write this lemma as

$$|y_i(t) - m\pi_i| \leq C\lambda^t$$

Consensus

Lemma

Suppose that Assumptions (A), (B),(D),(E) hold. Then for any $t \geq 1$ we have

$$\|z_i(t+1) - B_\zeta \bar{x}(t)\| \leq C \left(\zeta(t) + \tau(t) + \sum_{s=0}^{t-1} \lambda^{t-s-1} \alpha(t) \right),$$

where $B_\zeta = \frac{m_\zeta}{m}$.

This Lemma implies the consensus is achieved, i.e.

$$\lim_{t \to \infty} \|z_i(t+1) - B_\zeta \bar{x}(t)\| = 0.$$

If $\tau(t), \zeta(t) \equiv 0$, it follows that

$$\lim_{t \to \infty} \|z_i(t+1) - \bar{x}(t)\| = 0.$$

Convergence

Lemma

Suppose Assumptions (A) and (B) hold. Then for any $t \geq 0$, we have

$$\sum_{i=1}^{m}(f_i(B_\zeta \bar{x}(t)) - f_i(x)) \leq \frac{m}{2\alpha(t+1)B_\zeta}(\|B_\zeta \bar{x}(t) - x\|^2 - \|B_\zeta \bar{x}(t+1) - x\|^2)$$
$$+ \frac{B_\zeta}{2\alpha(t+1)}(2\alpha(t+1)^2 D^2 + 2\tau(t)^2) + \frac{m\tau(t)}{\alpha(t+1)}\|B_\zeta \bar{x}(t) - x\|$$
$$+ 2D\sum_{i=1}^{m}\|z_i(t+1) - B_\zeta \bar{x}(t)\|$$

Rearranging the above inequality together with the consensus result, for any $t \geq 0$, we have

$$\|B_\zeta \bar{x}(t+1) - x\|^2 \leq (1+\tau(t))\|B_\zeta \bar{x}(t) - x\|^2 - C\alpha(t)(f(B_\zeta \bar{x}(t)) - f(x)) + c(t).$$

Convergence

Lemma
Consider a minimization problem

$$\min_{x \in \mathbb{R}^d} f(x),$$

where $f : \mathbb{R}^d \to \mathbb{R}$ is a continuous function. Assume that the solution X^* of the problem is nonempty. Let $\{x(t)\}_{t\in\mathbb{N}}$ be a sequence such that for all $x \in X^*$ and for all $t \geq 0$,

$$\|x(t+1) - x\|^2 \leq (1 + b(t))\|x(t) - x\|^2 - a(t)(f(x(t)) - f(x)) + c(t)$$

where $b(t) \geq 0$, $a(t) \geq 0$ and $c(t) \geq 0$ for all $t \geq 0$ with $\sum_{t=0}^{\infty} b(t) < \infty$,$\sum_{t=0}^{\infty} a(t) = \infty$ and $\sum_{t=0}^{\infty} c(t) < \infty$. Then the sequence $\{x(t)\}_{t\in\mathbb{N}}$ converges to some solution $x^* \in X^*$

Nedić, Angelia, and Alex Olshevsky. "Distributed optimization over time-varying directed graphs." IEEE Transactions on Automatic Control 60.3 (2014): 601-615.

Simulation

$$\min_{x \in \mathbb{R}^d} \sum_{i=1}^{m} f_i(x) \quad \text{with} \quad f_i(x) = \|q_i - p_i^T x\|^2,$$

where $p_i \in \mathbb{R}^{d \times p}$ is the input data and the variable $q_i \in \mathbb{R}^p$ is the output data.

- $d = 5$, $p = 1$, and $m = 50$
- We use connected directed graph where every node has four out neighbors
- We measure

$$R_d(t) = \frac{\sum_{i=1}^{m} \|z_i(t) - x^*\|}{\sum_{i=1}^{m} \|z_i(0) - x^*\|}$$

Simulation

- We fix $\alpha(t) = 1/\sqrt{t}$
- We take several choices of $\tau(t)$ and $\zeta(t)$.
- κ_f is the first time $k \in \mathbb{N}$ when $R_d(k) < 10^{-2}$
- N_x and N_y are the average of a total number of triggers for all agents until the termination time.

$\tau(t)$ $\zeta(t)$	0 0	0 $1/(3t^3)$	$1/t^{1.5}$ 0	$1/t^{1.5}$ $1/(3t^3)$
N_x	11425	11305	8860	8767
N_y	11425	26	11644	26
κ_f	11425	11305	11644	11514

Conclusions

Chapter 3: We study the decentralized gradient descent.

- Without convexity assumption for a local function, we demonstrate that the DGD algorithm achieves exact convergence to an optimal point when utilizing a diminishing stepsize.
- We present an example highlighting the sharpness of the stepsize condition for obtaining uniform boundedness.

Chapter 4: We study the decentralized projected gradient descent.

- We obtain that the DPG algorithm converges exponentially fast to an $O(\sqrt{\alpha})$-neighborhood of an optimal point.
- We present a half-space example, which achieves $O(\alpha)$-neighborhood as in the DGD.

Chapter 5: We study the gradient-push algorithm with event-triggered communication.

- We achieve asymptotic convergence results under suitable decays and summability conditions on the stepsize and triggering thresholds.
- We also derive an exact convergence rate for $\alpha(t) = 1/\sqrt{t}$.

Thank you for your attention!