Communication-Efficient Decentralized and Stochastic Optimization

LCCC Focus Period, Lund University
June 5th, 2017

Soomin Lee
soomin.lee@isye.gatech.edu
Industrial and Systems Engineering
Georgia Tech

Joint work with **Guanghui (George) Lan** and **Yi Zhou**

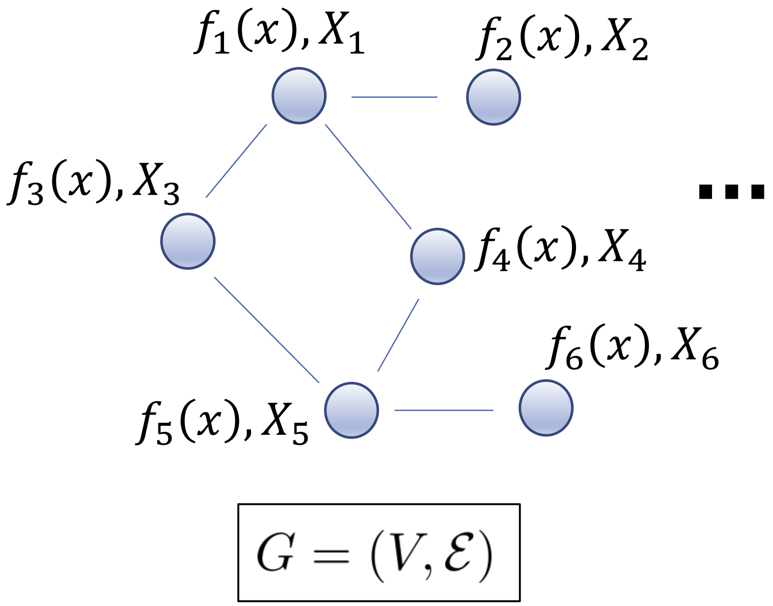
Decentralized Optimization

Optimization problem defined over ***complex*** multi-agent systems

- No central authority
- Time-varying topology

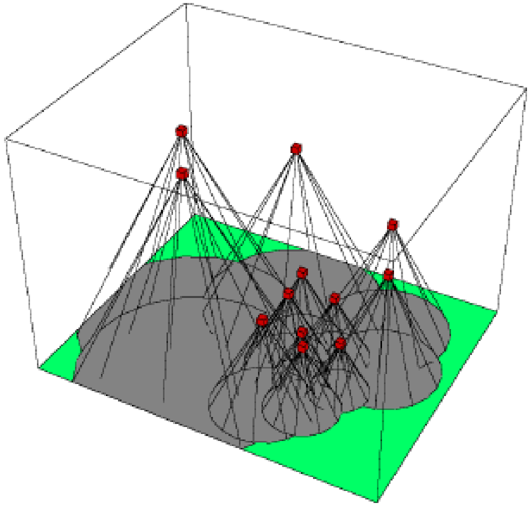
The m agents **collaboratively** solve:

$$\min_x f(x) := \sum_{i=1}^m f_i(x) \quad \text{s.t.} \quad x \in \bigcap_{i=1}^m X_i$$



Communication is an important factor, but can be very expensive

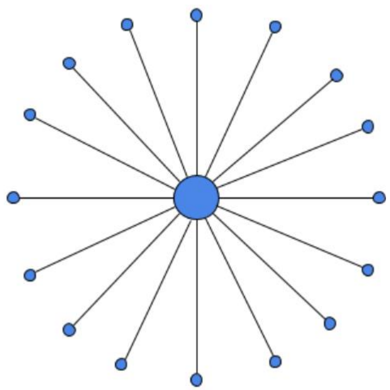
Why interested?

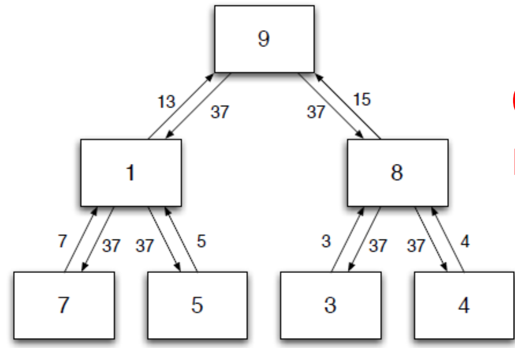


- Lots of potential applications: swarming robots, drones…
- Strength in numbers
- Privacy preserving
- Distributed data mining/processing
- Convergence analysis and algorithms
- Scientifically interesting!!

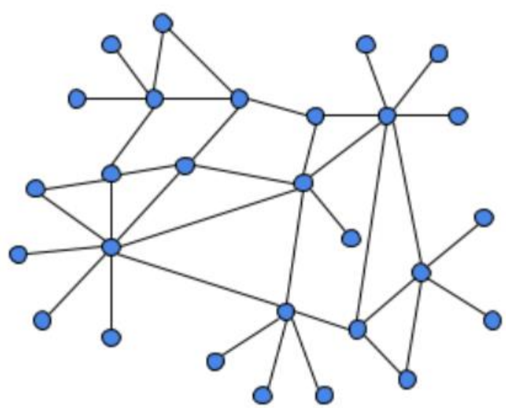
How to Handle Decentralized Structure?

$$\min_x f(x) := \sum_{i=1}^m f_i(x) \quad \text{s.t. } x \in \bigcap_{i=1}^m X_i$$





Only the central node maintains x



Everybody maintains a local copy of x

How to Handle Decentralized Structure?

$$\min_x f(x) := \sum_{i=1}^m f_i(x) \quad \text{s.t. } x \in \bigcap_{i=1}^m X_i$$

- **Dual decomposition (explicit)**

$$\min_{\mathbf{x}} F(\mathbf{x}) := \sum_{i=1}^m f_i(x_i)$$

$$\text{s.t. } x_1 = \cdots = x_m$$

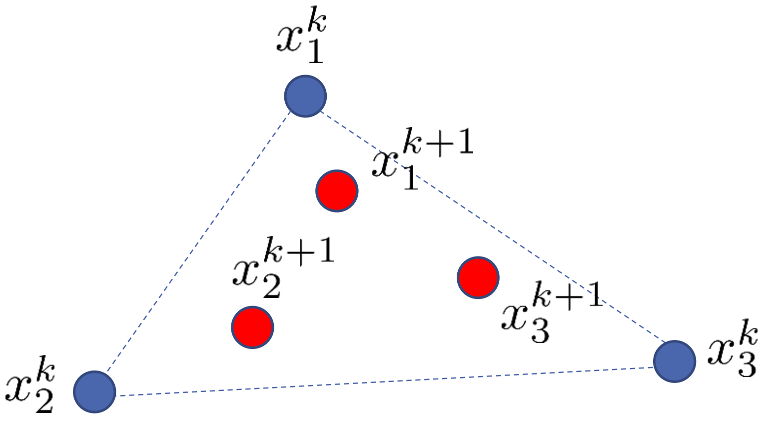
$$x_i \in X_i, \; i = 1, \ldots, m$$

$$\mathbf{x} := [x_1^\top \cdots x_m^\top]^\top$$

- **Consensus (implicit)**

$$\min_{\mathbf{x}} F(\mathbf{x}) := \sum_{i=1}^m f_i(x_i)$$

$$\text{s.t. } x_i \in X_i, \; i = 1, \ldots, m$$



How to Handle Decentralized Structure?

- **Dual decomposition: Pros and Cons**
  - Need to solve nontrivial Lagrangian related local subproblem
  - Requires a fewer number of communications

- **Consensus: Pros and Cons**
  - Inexpensive local subgradient update in primal space
  - Requires lots of inter-node communications

- **Our Goal**

Dual based decentralized methods (optimal communication) whose **local subproblems** can be solved easily through **linearizations**

This Talk

1. Decentralized Communication Sliding Method (**DCS**)
   - Subproblems solved **approximately** using **exact** subgradients

2. Stochastic Decentralized Communication Sliding (**SDCS**)
   - Subproblems solved **approximately** using **noisy** subgradients

Decentralized Optimization for Nonsmooth Functions

$$\frac{\mu}{2}\|x-y\|^2 \le f_i(x) - f_i(y) - \langle f_i'(y), x-y \rangle \le M\|x-y\|, \quad \forall x, y \in X_i$$
$$\text{for some } M, \mu \ge 0 \text{ and } f_i'(y) \in \partial f_i(y)$$

μ: strong convexity, M: Lipschitz constant

Convergence Rates

- Iteration complexity to find a solution $\bar{x}$ such that $f(\bar{x}) - f^* \leq \epsilon$

Algorithm	Requirement	Communication	Gradient Computation
DCS	Exact subgradient Convexity	$1/\epsilon$	$1/\epsilon^2$
	Exact subgradient Strong convexity	$1/\sqrt{\epsilon}$	$1/\epsilon$
SDCS	Noisy subgradient Convexity	$1/\epsilon$	$1/\epsilon^2$
	Noisy subgradient Strong convexity	$1/\sqrt{\epsilon}$	$1/\epsilon$

Comparable to the best known results in centralized mirror descent

Highlights of Our Contributions

Communication is about 1000 times more expensive!!

- Communication over TCP/IP: 10KB/ms + a few ms for startup
- CPUs read/write from/to memory:10KB/μs

Algorithm	Requirement	Communication	Gradient computation
ADMM / GD+Backtraking	Smoothness Strong convexity Unconstrained	$\log 1/\epsilon$	$\log 1/\epsilon$
(Proximal) AGD + multistep consensus	Smoothness Unconstrained	$\frac{1}{\sqrt{\epsilon}} \log \frac{1}{\epsilon}$ $(1/\epsilon)$	$1/\sqrt{\epsilon}$
Decentralized Stochastic MD*	Strong convexity	$1/\epsilon$	$1/\epsilon$
DCS/SDCS	Convexity	$1/\epsilon$	$1/\epsilon^2$
DCS/SDCS*	Strong convexity	$1/\sqrt{\epsilon}$	$1/\epsilon$

*Noisy subgradient can be used

Background: Bregman Distance Function

- **Distance generating function**

  $\omega: X \to \mathbb{R}$, differentiable and strongly convex with modulus $\nu > 0$

  $$\text{e.g. } \omega(x) = \frac{1}{2}\|x\|^2, \quad -\sum_i \log x_i$$

- **Prox-function** or **Bregman distance function** induced by ω

  $$V(x,u) \equiv V_\omega(x,u) := \omega(u) - [\omega(x) + \langle \nabla\omega(x), u - x\rangle].$$

- For all agent i, we assume $\nu = 1$

  $$V_i(x_i, u_i) \geq \frac{1}{2}\|x_i - u_i\|_{X_i}^2, \quad \forall x_i, u_i \in X_i$$

  $$\mathbf{V}(\mathbf{x}, \mathbf{u}) := \sum_{i=1}^m V_i(x_i, u_i), \ \forall \mathbf{x}, \mathbf{u} \in X^m$$

- We also assume $\mathbf{V}$ is **growing quadratically** with constant $\mathcal{C}$

  $$V_i(x_i, u_i) \leq \frac{\mathcal{C}}{2}\|x_i - u_i\|_{X_i}^2, \quad \forall x_i, u_i \in X_i$$

  $\mathcal{C}$: growth constant

Background: Laplacian L

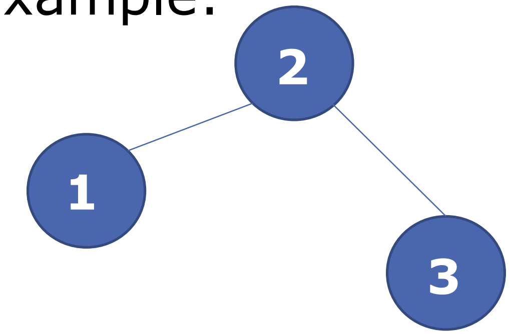
Let N_i denote the **set of neighbors** of agent i:

$$N_i = \{j \in V \mid (i,j) \in \mathcal{E}\} \cup \{i\}$$

Then, the **Laplacian** $L \in \mathbb{R}^{m \times m}$ of a graph $G = (V, \mathcal{E})$ is defined as:

$$L_{ij} = \begin{cases} |N_i| - 1 & \text{if } i = j \\ -1 & \text{if } i \neq j \text{ and } (i,j) \in \mathcal{E} \\ 0 & \text{otherwise.} \end{cases}$$

For example:



$$L = \begin{bmatrix} 1 & -1 & 0 \\ -1 & 2 & -1 \\ 0 & -1 & 1 \end{bmatrix}$$

$$L\mathbf{1} = \mathbf{0}$$

"Agreement Subspace"

Problem Reformulation

$$\min_{\mathbf{x}} \; F(\mathbf{x}) := \sum_{i=1}^{m} f_i(x_i)$$

$$\text{s.t.} \; x_1 = \cdots = x_m$$

$$x_i \in X_i, \; i = 1, \ldots, m$$

(=)

$$\min_{\mathbf{x}} \; F(\mathbf{x}) := \sum_{i=1}^{m} f_i(x_i)$$

$$\text{s.t.} \; x_i = x_j, \; \forall (i,j) \in \mathcal{E}$$

$$x_i \in X_i, \; i = 1, \ldots, m$$

If G is **connected**

(=)

$$\min_{\mathbf{x}} \; F(\mathbf{x}) := \sum_{i=1}^{m} f_i(x_i)$$

$$\text{s.t.} \; \mathbf{L}\mathbf{x} = \mathbf{0}$$

$$x_i \in X_i, \; i = 1, \ldots, m$$

Using **Laplacian** L, consistency constraints can be **compactly** rewritten

$$\mathbf{L} := L \otimes I_d$$

(=)

$$\min_{\mathbf{x} \in X^m} F(\mathbf{x}) + \max_{\mathbf{y} \in \mathbb{R}^{md}} \langle \mathbf{L}\mathbf{x}, \mathbf{y} \rangle$$

Equivalent **Saddle Point** form

Decentralized Primal-Dual (DPD): Vector Form

$$\min_{\mathbf{x}\in X^m} F(\mathbf{x}) + \max_{\mathbf{y}\in\mathbb{R}^{md}} \langle \mathbf{L}\mathbf{x}, \mathbf{y}\rangle$$

$$\mathbf{x} := [x_1^\top \cdots x_m^\top]^\top$$
$$\mathbf{y} := [y_1^\top \cdots y_m^\top]^\top$$

Let $\boldsymbol{x}^0 = \boldsymbol{x}^{-1} \in X^m$, $\boldsymbol{y} \in \mathbb{R}^{md}$, $\{\alpha_t\}, \{\tau_t\}, \{\eta_t\}$ and $\{\theta_t\}$ be given.

For $t = 1, \ldots, N$, update $\boldsymbol{z}^t = (\boldsymbol{x}^t, \boldsymbol{y}^t)$

$$\tilde{\mathbf{x}}^t = \alpha_t(\mathbf{x}^{t-1} - \mathbf{x}^{t-2}) + \mathbf{x}^{t-1}$$
$$\mathbf{y}^t = \operatorname{argmin}_{\mathbf{y}\in\mathbb{R}^{md}} \langle -\mathbf{L}\tilde{\mathbf{x}}^t, \mathbf{y}\rangle + \tfrac{\tau_t}{2}\|\mathbf{y} - \mathbf{y}^{t-1}\|^2$$
$$\mathbf{x}^t = \operatorname{argmin}_{\mathbf{x}\in X^m} \langle \mathbf{L}\mathbf{y}^t, \mathbf{x}\rangle + F(\mathbf{x}) + \eta_t \mathbf{V}(\mathbf{x}^{t-1}, \mathbf{x})$$

Return $\bar{\mathbf{z}}^N = (\sum_{t=1}^N \theta_t)^{-1} \sum_{t=1}^N \theta_t \mathbf{z}^t$

DPD: Agent i's point of view

Let $x_i^0 = x_i^{-1} \in X_i,\ y_i \in \mathbb{R}^d$, $\{\alpha_t\}, \{\tau_t\}, \{\eta_t\}$ and $\{\theta_t\}$ be given.

For $t = 1, \ldots, N$, update $z_i^t = (x_i^t, y_i^t)$

$$\tilde{x}_i^t = \alpha_t(x_i^{t-1} - x_i^{t-2}) + x_i^{t-1}$$

$$v_i^t = \textstyle\sum_{j \in N_i} L_{ij} \tilde{x}_j^t$$ ← Communication of updated primal

$$y_i^t = y_i^{t-1} + \tfrac{1}{\tau_t} v_i^t$$

$$w_i^t = \textstyle\sum_{j \in N_i} L_{ij} y_j^t$$ ← Communication of updated dual

$$x_i^t = \operatorname{argmin}_{x_i \in X_i}\ \langle w_i^t, x_i \rangle + f_i(x_i) + \eta_t V_i(x_i^{t-1}, x_i)$$

Return $\bar{z}_i^N = (\sum_{t=1}^N \theta_t)^{-1} \sum_{t=1}^N \theta_t z_i^t$

The algorithm is **Decentralized**!

Decentralized Communication Sliding (DCS)

Q: Is the **subproblem** always **easy** to solve?

$$x_i^t = \text{argmin}_{x_i \in X_i} \langle w_i^t, x_i \rangle + f_i(x_i) + \eta_t V_i(x_i^{t-1}, x_i)$$

A: No, solve this **iteratively** using **linearization** of $f_i(x_i)$

Let $u^0 = \hat{u}^0 = x_i^{t-1}$, $\{\beta_k\}$ and $\{\lambda_k\}$ be given.

For $k = 1, \dots, K_t$

$$h^{k-1} \in \partial f_i(u^{k-1})$$

$$u^k = \arg\min_{u \in X_i} \langle h^{k-1} + w_i^t, u \rangle + \eta_t V_i(x_i^{t-1}, u) + \eta_t \beta_k V_i(u^{k-1}, u)$$

Return $x_i^t = u^{K_t}$ and $\hat{x}_i^t = \left(\sum_{k=1}^{K_t} \lambda_k \right)^{-1} \sum_{k=1}^{K_t} \lambda_k u^k$

The same w_i^t is used, communication is skipped!

There are two outputs x_i^t and $\hat{x}_i^t$

Decentralized Communication Sliding (DCS)

Let $x_i^0 = x_i^{-1} \in X_i$, $y_i \in \mathbb{R}^d$, $\{\alpha_t\}, \{\tau_t\}, \{\eta_t\}, \{\theta_t\}$ and $\{K_t\}$ be given.

For $t = 1, \ldots, N$, update $z_i^t = (\hat{x}_i{}^t, y_i^t)$

$$\tilde{x}_i^t = \alpha_t(\hat{x}_i^{t-1} - x_i^{t-2}) + x_i^{t-1}$$

$$v_i^t = \sum_{j \in N_i} L_{ij} \tilde{x}_j^t$$ ← Communication of updated primal

$$y_i^t = y_i^{t-1} + \frac{1}{\tau_k} v_i^t$$

$$w_i^t = \sum_{j \in N_i} L_{ij} y_j^t$$ ← Communication of updated dual

$$(x_i^t, \hat{x}_i^t) = \text{Inner loop for } K_t \text{ times}$$ ← **No Communication!**

Return $z_i^N = (\hat{x}_i{}^N, y_i^N)$

DCS: Convergence for Convex Cases

Theorem 1

Set parameters for $t = 1, \ldots, N$ and $k = 1, \ldots, K_t$

$\boldsymbol{\alpha_t}$	Primal prediction	1	$\boldsymbol{\beta_k}$	Inner-loop projection	$k/2$
$\boldsymbol{\tau_t}$	Dual projection	$\|\mathbf{L}\|$	$\boldsymbol{\lambda_k}$	Inner-loop averaging	$k+1$
$\boldsymbol{\eta_t}$	Primal Projection	$2\|\mathbf{L}\|$	$\boldsymbol{K_t}$	# inner-loop iterations	$\left\lceil \frac{mM^2N}{\|\mathbf{L}\|^2\tilde{D}} \right\rceil$
$\boldsymbol{\theta_t}$	Outer-loop averaging	1			

Then, iteration complexity to find a solution $\hat{\mathbf{x}}^N = \left(\sum_{t=1}^N \theta_t\right)^{-1} \sum_{t=1}^N \theta_t \mathbf{x}^t$ such that $F(\hat{\boldsymbol{x}}^N) - F^* \leq \epsilon$ and $\|\mathbf{L}\hat{\boldsymbol{x}}^N\| \leq \epsilon$

$$O\left(\frac{\|\mathbf{L}\| D_{X^m}^2}{\epsilon}\right)$$ for communications

$$O\left(\frac{mM^2 D_{X^m}^2}{\epsilon^2}\right)$$ for gradient computations

DCS: Convergence for Strongly Convex Cases

Theorem 2

Set parameters for $t = 1, \ldots, N$ and $k = 1, \ldots, K_t$

$\boldsymbol{\alpha_t}$	Primal prediction	$\frac{t}{t+1}$	$\boldsymbol{\beta_k}$	Inner-loop projection	$\frac{(k+1)\mu}{2\eta_t C} + \frac{k-1}{2}$
$\boldsymbol{\tau_t}$	Dual projection	$\frac{4\|\mathbf{L}\|^2 C}{(t+1)\mu}$	$\boldsymbol{\lambda_k}$	Inner-loop averaging	k
$\boldsymbol{\eta_t}$	Primal Projection	$\frac{t\mu}{2C}$	$\boldsymbol{K_t}$	# inner-loop iterations	$\left\lceil \sqrt{\frac{2m}{\tilde{D}}} \frac{CMN}{\mu} \max\left\{ \sqrt{\frac{2m}{\tilde{D}}} \frac{4CM}{\mu}, 1 \right\} \right\rceil$
$\boldsymbol{\theta_t}$	Outer-loop averaging	$t+1$			

Then, iteration complexity to find a solution $\hat{\mathbf{x}}^N = \left(\sum_{t=1}^N \theta_t\right)^{-1} \sum_{t=1}^N \theta_t \mathbf{x}^t$ such that $F(\hat{\mathbf{x}}^N) - F^* \le \epsilon$ and $\|\mathbf{L}\hat{\mathbf{x}}^N\| \le \epsilon$

$O\left(\sqrt{\frac{\mu D_X^2 m}{C\epsilon}}\right)$ for communications

$O\left(\frac{mM^2 C}{\mu\epsilon}\right)$ for gradient computations

Outline of Convergence Analysis

- **Inner loop**

$$(x_i^t, \hat{x}_i^t) = \text{Inner loop for } K_t \text{ times}$$

$$u^k = \arg\min_{u \in X_i} \langle h^{k-1} + w_i^t, x_i \rangle + \eta_t V_i(x_i^{t-1}, u) + \eta_t \beta_k V_i(u^{k-1}, u)$$

- Recursive relation

$$\left(\textstyle\sum_{k=1}^{K_t} \lambda_k\right)^{-1} \left[\eta_t (1 + \beta_{K_t}) \lambda_{K_t} V_i(u^{K_t}, u)\right] + \Phi_i^t(\hat{u}^{K_t}) - \Phi_i^t(u)$$
$$\leq \left(\textstyle\sum_{k=1}^{K_t} \lambda_k\right)^{-1} \left[(\eta_t \beta_1 - \mu/\mathcal{C}) \lambda_1 V_i(u^0, u) + \textstyle\sum_{k=1}^{K_t} \frac{M^2 \lambda_k}{2 \eta_t \beta_k}\right],$$

$$\text{where } \Phi_i^t(u) := \langle w_i^t, u \rangle + f_i(u) + \eta_t V_i(x_i^{t-1}, u)$$

- $x_i^{t-1} = u^0$ and $x_i^t = u^{K_t}$ is used for telescoping sum
- $\hat{x}_i^t = \hat{u}^{K_t}$ is used for actual perturbed primal-dual gap evaluation

Outline of Convergence Analysis

- **Outer loop**

$$Q(\hat{\mathbf{z}}^N;\mathbf{z}) \le \left(\sum_{t=1}^{N}\theta_t\right)^{-1}\left[\frac{(K_1+1)(K_1+2)\theta_1\eta_1}{K_1(K_1+3)}\mathbf{V}(\mathbf{x}^0,\mathbf{x}) + \frac{\theta_1\tau_1}{2}\|\mathbf{y}^0\|^2 + \langle\hat{\mathbf{s}},\mathbf{y}\rangle + \sum_{t=1}^{N}\frac{4mM^2\theta_t}{(K_t+3)\eta_t}\right]$$

Primal distance: $\mathbf{V}(\mathbf{x}^0,\mathbf{x})$

Dual distance: $\frac{\theta_1\tau_1}{2}\|\mathbf{y}^0\|^2$

Perturbation term: $\langle\hat{\mathbf{s}},\mathbf{y}\rangle$

Accumulated inner-loop error: $\sum_{t=1}^{N}\frac{4mM^2\theta_t}{(K_t+3)\eta_t}$

Stochastic DCS (SDCS)

- **Stochastic** Decentralized Optimization

$$f_i(x) := \mathbb{E}_{\xi_i}[F_i(x;\xi_i)],$$

where ξ_i models agent i's uncertainty and $\mathbb{P}(\xi_i)$ not known.

- As a special case, **sum of many components**

$$f_i(x) := \sum_{j=1}^{l} f_i^j(x)$$

- Only **noisy** first-order information $G_i(\cdot, \xi_i^t)$ is available

$$\mathbb{E}[G_i(u^t, \xi_i^t)] = f_i'(u^t) \in \partial f_i(u^t),$$

$$\mathbb{E}[\|G_i(u^t, \xi_i^t) - f_i'(u^t)\|_*^2] \le \sigma^2$$

Stochastic DCS (SDCS)

Q: Is the **subproblem** always **easy** to solve?

$$x_i^t = \mathrm{argmin}_{x_i \in X_i} \langle w_i^t, x_i \rangle + f_i(x_i) + \eta_t V_i(x_i^{t-1}, x_i)$$

A: No, solve this **iteratively** using **linearization** of $f_i(x_i)$

Let $u^0 = \hat{u}^0 = x_i^{t-1}$, $\{\beta_k\}$ and $\{\lambda_k\}$ be given.

For $k = 1, \dots, K_t$

$$h^{k-1} \in G_i(u^{k-1}, \xi_i^{k-1})$$ ← One data sample (Stochastic)!

$$u^k = \arg\min_{u \in X_i} \langle h^{k-1} + w_i^t, u \rangle + \eta_t V_i(x_i^{t-1}, u) + \eta_t \beta_k V_i(u^{k-1}, u)$$

Return $x_i^t = u^{K_t}$ and $\hat{x}_i^t = \left(\sum_{k=1}^{K_t} \lambda_k\right)^{-1} \sum_{k=1}^{K_t} \lambda_k u^k$

The same w_i^t is used, communication is skipped!

There are two outputs x_i^t and $\hat{x}_i^t$

SDCS: Convergence for Convex Cases

Theorem 3

Set parameters for $t = 1, \dots, N$ and $k = 1, \dots, K_t$

$\boldsymbol{\alpha_t}$	Primal prediction	1	$\boldsymbol{\beta_k}$	Inner-loop projection	$k/2$
$\boldsymbol{\tau_t}$	Dual projection	$\lVert \mathbf{L} \rVert$	$\boldsymbol{\lambda_k}$	Inner-loop averaging	$k+1$
$\boldsymbol{\eta_t}$	Primal Projection	$2\lVert \mathbf{L} \rVert$	$\boldsymbol{K_t}$	# inner-loop iterations $\left\lceil \frac{m(M^2+\sigma^2)N}{\lVert \mathbf{L} \rVert^2 \tilde{D}} \right\rceil$	
$\boldsymbol{\theta_t}$	Outer-loop averaging	1			

Then, iteration complexity to find a solution $\hat{\mathbf{x}}^N = \left(\sum_{t=1}^N \theta_t\right)^{-1} \sum_{t=1}^N \theta_t \mathbf{x}^t$ such that $F(\hat{\boldsymbol{x}}^N) - F^* \le \epsilon$ and $\lVert \mathbf{L}\hat{\boldsymbol{x}}^N \rVert \le \epsilon$

$$O\left(\frac{\lVert \mathbf{L} \rVert D_{X^m}^2}{\epsilon}\right)$$ for communications

$$O\left(\frac{m(M^2+\sigma^2)D_{X^m}^2}{\epsilon^2}\right)$$ for gradient computations

SDCS: Convergence for Strongly Convex Cases

Theorem 4

Set parameters for $t = 1, \dots, N$ and $k = 1, \dots, K_t$

$\boldsymbol{\alpha_t}$	Primal prediction	$\frac{t}{t+1}$	$\boldsymbol{\beta_k}$	Inner-loop projection	$\frac{(k+1)\mu}{2\eta_t C} + \frac{k-1}{2}$
$\boldsymbol{\tau_t}$	Dual projection	$\frac{4\|\mathbf{L}\|^2 C}{(t+1)\mu}$	$\boldsymbol{\lambda_k}$	Inner-loop averaging	k
$\boldsymbol{\eta_t}$	Primal Projection	$\frac{t\mu}{2C}$	$\boldsymbol{K_t}$	# inner-loop iterations $\left\lceil \sqrt{\frac{m(M^2+\sigma^2)}{\tilde{D}}\frac{2NC}{\mu}} \max\left\{\sqrt{\frac{m(M^2+\sigma^2)}{\tilde{D}}\frac{8C}{\mu}}, 1\right\}\right\rceil$	
$\boldsymbol{\theta_t}$	Outer-loop averaging	$t+1$			

Then, iteration complexity to find a solution $\hat{\mathbf{x}}^N = \left(\sum_{t=1}^N \theta_t\right)^{-1} \sum_{t=1}^N \theta_t \mathbf{x}^t$ such that $F(\hat{\mathbf{x}}^N) - F^* \le \epsilon$ and $\|\mathbf{L}\hat{\mathbf{x}}^N\| \le \epsilon$

$$O\left(\sqrt{\frac{\mu D_X^2 m}{C\epsilon}}\right) \quad \text{for communications}$$

$$O\left(\frac{m(M^2+\sigma^2)C}{\mu\epsilon}\right) \quad \text{for gradient computations}$$

Summary of Convergence Results

Complexity for obtaining **ϵ-optimal** and **ϵ-feasible** solution

Algorithm	# of communications	# of subgradient evaluations
DCS: Convex	$\mathcal{O}\left\{\frac{\|\mathbf{L}\|\mathcal{D}_X^2 m}{\epsilon}\right\}$	$\mathcal{O}\left\{\frac{mM^2\mathcal{D}_X^2 m}{\epsilon^2}\right\}$
DCS: Strongly convex	$\mathcal{O}\left\{\sqrt{\frac{\mu\mathcal{D}_X^2 m}{\mathcal{C}\epsilon}}\right\}$	$\mathcal{O}\left\{\frac{mM^2\mathcal{C}}{\mu\epsilon}\right\}$
SDCS: Convex	$\mathcal{O}\left\{\frac{\|\mathbf{L}\|\mathcal{D}_X^2 m}{\epsilon}\right\}$	$\mathcal{O}\left\{\frac{m(M^2+\sigma^2)\mathcal{D}_X^2 m}{\epsilon^2}\right\}$
SDCS: Strongly convex	$\mathcal{O}\left\{\sqrt{\frac{\mu\mathcal{D}_X^2 m}{\mathcal{C}\epsilon}}\right\}$	$\mathcal{O}\left\{\frac{m(M^2+\sigma^2)\mathcal{C}}{\mu\epsilon}\right\}$

Comparable with **centralized** mirror-descent method

Conclusions

- First-order **Decentralized Primal-Dual** algorithm for convex **nonsmooth deterministic/stochastic** problems
- **Primal subproblems approximately** solved using **linearizations**
- **Communication Sliding** to **reduce communication overhead**

- The **most communication efficient** algorithm until now in nonsmooth decentralized optimization
- Subgradient computation complexity **comparable with centralized** mirror-descent

- Ongoing work: Implementation, time-varying networks

Boundedness of y^*

Theorem 4

Let $\boldsymbol{x}^*$ be an optimal solution. Then, $\exists \boldsymbol{y}^*$ such that

$$\|\mathbf{y}^*\| \leq \frac{\sqrt{m}M}{\tilde{\sigma}_{min}(\mathbf{L})},$$

where $\tilde{\sigma}_{min}(\boldsymbol{L})$ denotes the smallest nonzero singular value of $\boldsymbol{L}$.

Proof. From the saddle point inequality, we have

$$\mathcal{L}(\mathbf{x}^*, \mathbf{y}^*) \leq \mathcal{L}(\mathbf{x}, \mathbf{y}^*) \quad \Longrightarrow \quad F(\mathbf{x}^*) - F(\mathbf{x}) \leq \langle -\mathbf{L}^\top \mathbf{y}^*, \mathbf{x} - \mathbf{x}^* \rangle$$

From the definition of the subgradient, $-\mathbf{L}^\top \mathbf{y}^* \in \partial F(\mathbf{x}^*)$

Everything can be represented in primal terms

DPD: Convergence Results

Theorem 1

Let $\boldsymbol{x}^*$ be an optimal point, and

$$\alpha_k = \theta_k = 1,\ \ \eta_k = 2\|\mathbf{L}\|,\ \ and\ \tau_k = \|\mathbf{L}\|, \quad \forall k = 1, \ldots, N.$$

Then, for any $N \geq 1$,

$$F(\bar{\mathbf{x}}^N) - F(\mathbf{x}^*) \leq \frac{\|\mathbf{L}\|}{N}\left[2\mathbf{V}(\mathbf{x}^0, \mathbf{x}^*) + \tfrac{1}{2}\|\mathbf{y}^0\|^2\right]$$

$$\|\mathbf{L}\bar{\mathbf{x}}^N\| \leq \frac{2\|\mathbf{L}\|}{N}\left[3\sqrt{\mathbf{V}(\mathbf{x}^0, \mathbf{x}^*)} + 2\|\mathbf{y}^* - \mathbf{y}^0\|\right],$$

where $\bar{\mathbf{x}}^N = \frac{1}{N}\sum_{k=1}^N \mathbf{x}^k$.

$O\left(\frac{1}{\epsilon}\right)$ **iterations** for **ϵ-optimal** and **ϵ-feasible** solution

\# of required **communication** is also $O\left(\frac{1}{\epsilon}\right)$

Outline of Convergence Analysis

Definition: Primal-dual gap function $Q(z; \bar{z})$

Given a pair of feasible solutions $\boldsymbol{z} = (\boldsymbol{x}, \boldsymbol{y})$ and $\bar{\boldsymbol{z}} = (\bar{\boldsymbol{x}}, \bar{\boldsymbol{y}})$,

$$Q(\mathbf{z}; \bar{\mathbf{z}}) := F(\mathbf{x}) + \langle \mathbf{L}\mathbf{x}, \bar{\mathbf{y}} \rangle - [F(\bar{\mathbf{x}}) + \langle \mathbf{L}\bar{\mathbf{x}}, \mathbf{y} \rangle]$$

If $\boldsymbol{z}^* = (\boldsymbol{x}^*, \boldsymbol{y}^*)$ is a saddle point, $Q(\boldsymbol{z}^*; \bar{\boldsymbol{z}}) \leq 0$ for any $\bar{\boldsymbol{z}} \in X^m \times Y$

$$\sup_{\bar{\mathbf{z}} \in X^m \times Y} Q(\mathbf{z}; \bar{\mathbf{z}})$$

Definition: Perturbed Primal-dual gap function $g_Y(s, z)$

$$g_Y(\mathbf{s}, \mathbf{z}) := \sup_{\bar{\mathbf{y}} \in Y} Q(\mathbf{z}; \mathbf{x}^*, \bar{\mathbf{y}}) - \langle \mathbf{s}, \bar{\mathbf{y}} \rangle$$

Proposition: ϵ-optimal and ϵ-feasible solution

If $g_Y(\mathbf{L}\boldsymbol{x}, \boldsymbol{z}) < \epsilon$ and $\|\mathbf{L}\boldsymbol{x}\| < \epsilon$, where $\boldsymbol{z} \in X^m \times Y$,
then x is an ϵ-optimal and ϵ-feasible solution

DCS: Convergence for Convex Cases

Theorem 2

Let $\boldsymbol{x}^*$ be an optimal point, and

$$\alpha_k = \theta_k = 1,\ \eta_k = 2\|\mathbf{L}\|,\ \tau_k = \|\mathbf{L}\|,\ and\ T_k = \left\lceil \frac{mM^2 N}{\|\mathbf{L}\|^2 \tilde{D}} \right\rceil, \quad \forall k = 1, \ldots, N,$$

Then, for any $N \geq 1$,

$$F(\hat{\mathbf{x}}^N) - F(\mathbf{x}^*) \leq \frac{\|\mathbf{L}\|}{N} \left[3\mathbf{V}(\mathbf{x}^0, \mathbf{x}^*) + \frac{1}{2}\|\mathbf{y}^0\|^2 + 2\tilde{D} \right]$$

$$\|\mathbf{L}\hat{\mathbf{x}}^N\| \leq \frac{\|\mathbf{L}\|}{N} \left[3\sqrt{6\mathbf{V}(\mathbf{x}^0, \mathbf{x}^*) + 4\tilde{D}} + 4\|\mathbf{y}^* - \mathbf{y}^0\| \right],$$

where $\hat{\mathbf{x}}^N = \frac{1}{N}\sum_{k=1}^{N} \hat{\mathbf{x}}^k$.

$\boldsymbol{O}\left(\frac{1}{\epsilon}\right)$ **iterations** for **ϵ-optimal** and **ϵ-feasible** solution

# of required **communications** is also $\boldsymbol{O}\left(\frac{1}{\epsilon}\right)$

of **subgradient evaluations** is $\boldsymbol{O}\left(\frac{1}{\epsilon^2}\right)$

DCS: Convergence for Strongly Convex Cases

Theorem 3

Let $\boldsymbol{x}^*$ be an optimal point, and

$$\alpha_k = \frac{k}{k+1},\ \theta_k = k+1,\ \eta_k = \frac{k\mu}{2\mathcal{C}},\ \tau_k = \frac{4\|\mathbf{L}\|^2\mathcal{C}}{(k+1)\mu},$$

$$and\ T_k = \left\lceil \sqrt{\frac{2m}{\tilde{D}}}\frac{\mathcal{C}MN}{\mu}\max\left\{\sqrt{\frac{2m}{\tilde{D}}}\frac{4\mathcal{C}M}{\mu}, 1\right\}\right\rceil,$$

Then, for any $N \geq 1$,

$$F(\hat{\mathbf{x}}^N) - F(\mathbf{x}^*) \leq \frac{2}{N(N+3)}\left[\frac{\mu}{\mathcal{C}}\mathbf{V}(\mathbf{x}^0, \mathbf{x}^*) + \frac{2\|\mathbf{L}\|^2\mathcal{C}}{\mu}\|\mathbf{y}^0\|^2 + \frac{2\mu\tilde{D}}{\mathcal{C}}\right],$$

$$\|\mathbf{L}\hat{\mathbf{x}}^N\| \leq \frac{8\|\mathbf{L}\|}{N(N+3)}\left[3\sqrt{2\tilde{D} + \mathbf{V}(\mathbf{x}^0, \mathbf{x}^*)} + \frac{7\|\mathbf{L}\|\mathcal{C}}{\mu}\|\mathbf{y}^* - \mathbf{y}^0\|\right],$$

where $\hat{\mathbf{x}}^N = \frac{2}{N(N+3)}\sum_{k=1}^{N}(k+1)\hat{\mathbf{x}}^k.$

$\boldsymbol{O}\left(\frac{1}{\sqrt{\epsilon}}\right)$ **iterations** for **ϵ-optimal** and **ϵ-feasible** solution

\# of required **communications** is also $\boldsymbol{O}\left(\frac{1}{\sqrt{\epsilon}}\right)$

\# of **subgradient evaluations** is $\boldsymbol{O}\left(\frac{1}{\epsilon}\right)$

SDCS: Convergence for Convex Cases

Theorem 5

Let $\boldsymbol{x}^*$ be an optimal point, and

$$\alpha_k = \theta_k = 1,\ \eta_k = 2\|\mathbf{L}\|,\ \tau_k = \|\mathbf{L}\|,\ and\ T_k = \left\lceil \frac{m(M^2+\sigma^2)N}{\|\mathbf{L}\|^2\tilde{D}} \right\rceil, \quad \forall k = 1,\ldots,N,$$

Then, for any $N \geq 1$,

$$\mathbb{E}[F(\hat{\mathbf{x}}^k) - F(\mathbf{x}^*)] \leq \frac{\|\mathbf{L}\|}{N}\left[3\mathbf{V}(\mathbf{x}^0,\mathbf{x}^*) + \frac{1}{2}\|\mathbf{y}^0\|^2 + 4\tilde{D}\right],$$

$$\mathbb{E}[\|\mathbf{L}\hat{\mathbf{x}}^N\|] \leq \frac{\|\mathbf{L}\|}{N}\left[3\sqrt{6\mathbf{V}(\mathbf{x}^0,\mathbf{x}^*) + 8\tilde{D}} + 4\|\mathbf{y}^* - \mathbf{y}^0\|\right].$$

where $\hat{\mathbf{x}}^N = \frac{1}{N}\sum_{k=1}^{N}\hat{\mathbf{x}}^k.$

$\boldsymbol{O}\left(\frac{1}{\epsilon}\right)$ **iterations** for **ϵ-optimal** and **ϵ-feasible** solution

\# of required **communications** is also $\boldsymbol{O}\left(\frac{1}{\epsilon}\right)$

\# of **subgradient evaluations** is $\boldsymbol{O}\left(\frac{1}{\epsilon^2}\right)$

SDCS: Convergence for Strongly Convex Cases

Theorem 6

Let $\boldsymbol{x}^*$ be an optimal point, and

$$\alpha_k = \frac{k}{k+1},\ \theta_k = k+1,\ \eta_k = \frac{k\mu}{2\mathcal{C}},\ \tau_k = \frac{4\|\mathbf{L}\|^2\mathcal{C}}{(k+1)\mu},\ \textit{and}$$

$$T_k = \left\lceil \sqrt{\frac{m(M^2+\sigma^2)}{\tilde{D}}} \frac{2N\mathcal{C}}{\mu} \max\left\{ \sqrt{\frac{m(M^2+\sigma^2)}{\tilde{D}}} \frac{8\mathcal{C}}{\mu}, 1 \right\} \right\rceil, \quad \forall k = 1, \ldots, N,$$

Then, for any $N \geq 1$,

$$\mathbb{E}[F(\bar{\mathbf{x}}^N) - F(\mathbf{x}^*) \leq \frac{2}{N(N+3)} \left[\frac{\mu}{\mathcal{C}} \mathbf{V}(\mathbf{x}^0, \mathbf{x}^*) + \frac{2\|\mathbf{L}\|^2\mathcal{C}}{\mu} \|\mathbf{y}^0\|^2 + \frac{2\mu\tilde{D}}{\mathcal{C}} \right],$$

$$\mathbb{E}[\|\mathbf{L}\hat{\mathbf{x}}^N\|] \leq \frac{8\|\mathbf{L}\|}{N(N+3)} \left[3\sqrt{2\tilde{D} + \mathbf{V}(\mathbf{x}^0, \mathbf{x}^*)} + \frac{7\|\mathbf{L}\|\mathcal{C}}{\mu} \|\mathbf{y}^* - \mathbf{y}^0\| \right],$$

where $\hat{\mathbf{x}}^N = \frac{2}{N(N+3)} \sum_{k=1}^{N} (k+1)\hat{\mathbf{x}}^k$.

$\boldsymbol{O}\left(\frac{1}{\sqrt{\epsilon}}\right)$ **iterations** for **ϵ-optimal** and **ϵ-feasible** solution

\# of required **communications** is also $\boldsymbol{O}\left(\frac{1}{\sqrt{\epsilon}}\right)$

\# of **subgradient evaluations** is $\boldsymbol{O}\left(\frac{1}{\epsilon}\right)$