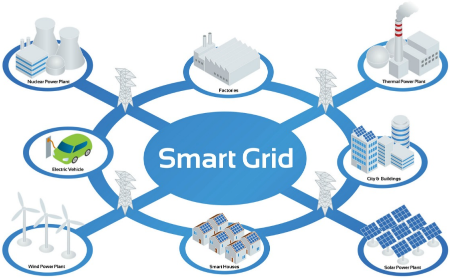
UC-Lab Center for Distribution System Cybersecurity

UCSB Presentation - Ramtin Pedarsani

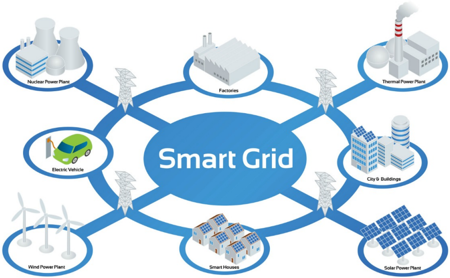
March 2019

- Future power grid is smart.
- The cyber-physical system is inherently geometrically distributed and has heterogeneous communications capability.



Decentralized Smart Grid: Motivation

- Future power grid is smart.
- The cyber-physical system is inherently geometrically distributed and has heterogeneous communications capability.



- New energy management schemes need to be robust to network node and link failures.

Welfare Maximization for Generation/Demand Response

- The goal is to maximize the welfare at time t for all generations and consumers.
- The ideal solution to the welfare maximization problem should have the following properties:
  1. Manages the demand side and the generation side simultaneously.
  2. Performs simple local optimization at each iteration and exchanges information with only neighbors.
  3. Privacy of information is guaranteed. No information about the utility/cost functions should be disclosed.
  4. Convergence to global optimum.
  5. The algorithm should be scalable for large networks.

Problem Formulation

- We can formulate a constrained utility maximization problem:

Welfare maximization problem formulation

$$\max_{\mathbf{d},\mathbf{g}} \sum_{j\in\mathcal{J}} U_j(\mathbf{d}_j) - \sum_{v\in\mathcal{V}} C_v(\mathbf{g}_v)$$

$$\text{s.t.} \quad \mathbf{1}^T\mathbf{d} = \mathbf{1}^T\mathbf{g}$$

$$d_{i,\min} \le d_i \le d_{i,\max}$$

$$0 \le g_i \le g_{i,\max}$$

- The t^{th} entry of the vectors of demand and generation correspond to period t (different times of the day).

Dual Decomposition

- Next, for simplicity, we focus on one single period.
- Let $\Delta = (\sum_{j \in \mathcal{J}} d_j - \sum_{v \in \mathcal{V}} g_v)$ be the power mismatch.

Dual Decomposition

$$J = \sum_{v \in \mathcal{V}} C_v(g_v) - \sum_{j \in \mathcal{J}} U_j(d_j) + p\Delta$$

$$d_i^{(k)} = \arg \min_{d_{i,\min} \leq d_i \leq d_{i,\max}} (p^{(k)} d_i - U_i(d_i))$$

$$g_i^{(k)} = \arg \min_{0 \leq g_i \leq g_{i,\max}} (C_i(d_i) - p^{(k)} g_i)$$

$$p^{(k+1)} = p^{(k)} + \eta \Delta^{(k)}$$

Consensus Algorithm

- Global parameters are Δ and price p that are not private.
- In the consensus algorithm, each node can only share the global parameters with the neighbors.
- W is the weight matrix that also determines the communication graph.

Distributed Estimation of Global Information

$$i \in \mathcal{V}: \; \hat{\Delta}_i^{(k+1)} = \hat{\Delta}_i^{(k)} + \sum_{j \in \mathcal{N}_i} w_{ij}(\hat{\Delta}_j^{(k)} - \hat{\Delta}_i^{(k)}) + g_i^{(k)} - g_i^{(k+1)}$$

$$i \in \mathcal{J}: \; \hat{\Delta}_i^{(k+1)} = \hat{\Delta}_i^{(k)} + \sum_{j \in \mathcal{N}_i} w_{ij}(\hat{\Delta}_j^{(k)} - \hat{\Delta}_i^{(k)}) - d_i^{(k)} + d_i^{(k+1)}$$

$$i \in \mathcal{V} \cup \mathcal{J}: \; p_i^{(k+1)} = p_i^{(k)} + \sum_{j \in \mathcal{N}_i} w_{ij}(p_j^{(k)} - p_i^{(k)}) + \eta \hat{\Delta}_i^{(k)}$$

- We consider the attack that the adversary jams the communication links.
- The bandwidth on links will be significantly reduced. Can we solve the decentralized optimization problem?
- Key idea is to use quantization!

Decentralized Gradient Decent (DGD)

Decentralized Gradient Decent (DGD)

- Network $\mathcal{G}$ with n nodes, weight matrix $W = [w_{ij}]_{n \times n}$

Decentralized Gradient Decent (DGD)

- Network $\mathcal{G}$ with n nodes, weight matrix $W = [w_{ij}]_{n \times n}$
- At iteration t, node i:

Decentralized Gradient Decent (DGD)

- Network $\mathcal{G}$ with n nodes, weight matrix $W = [w_{ij}]_{n \times n}$
- At iteration t, node i:
  - sends $\mathbf{x}_{i,t}$ to neighbors $\mathcal{N}_i$ and receives $\mathbf{x}_{j,t}$

Decentralized Gradient Decent (DGD)

- Network $\mathcal{G}$ with n nodes, weight matrix $W = [w_{ij}]_{n \times n}$
- At iteration t, node i:
  - sends $\mathbf{x}_{i,t}$ to neighbors $\mathcal{N}_i$ and receives $\mathbf{x}_{j,t}$
  - updates

$$\mathbf{x}_{i,t+1} = \underbrace{\sum_{j \in \mathcal{N}_i} w_{ij} \mathbf{x}_{j,t}}_{\substack{\text{average of local} \\ \text{and neighboring models}}} \underbrace{- \alpha \nabla f_i(\mathbf{x}_{i,t})}_{\text{local gradient descent}}$$

Decentralized Gradient Decent (DGD)

- Network $\mathcal{G}$ with n nodes, weight matrix $W = [w_{ij}]_{n \times n}$
- At iteration t, node i:
  - sends $\mathbf{x}_{i,t}$ to neighbors $\mathcal{N}_i$ and receives $\mathbf{x}_{j,t}$
  - updates

$$\mathbf{x}_{i,t+1} = \underbrace{\sum_{j \in \mathcal{N}_i} w_{ij} \mathbf{x}_{j,t}}_{\substack{\text{average of local} \\ \text{and neighboring models}}} \underbrace{-\alpha \nabla f_i(\mathbf{x}_{i,t})}_{\text{local gradient descent}}$$

[Nedić, Ozdaglar,'07]

Decentralized Gradient Decent (DGD)

- Network $\mathcal{G}$ with n nodes, weight matrix $W = [w_{ij}]_{n\times n}$
- At iteration t, node i:
  - sends $\mathbf{x}_{i,t}$ to neighbors $\mathcal{N}_i$ and receives $\mathbf{x}_{j,t}$
  - updates

$$\mathbf{x}_{i,t+1} = \underbrace{\sum_{j\in\mathcal{N}_i} w_{ij}\mathbf{x}_{j,t}}_{\substack{\text{average of local}\\ \text{and neighboring models}}} \underbrace{-\alpha\nabla f_i(\mathbf{x}_{i,t})}_{\text{local gradient descent}}$$

[Nedić, Ozdaglar,'07]

Theorem (Yuan, Ling, Yin '16)

Under A1,2,3, $\mathbf{x}_{i,t}$ geometrically converges to an $\mathcal{O}\left(\frac{\alpha}{1-\beta}\right)$-neighborhood of the unique solution $\mathbf{x}^*$. $(1-\beta$: spectral gap of $W)$

Decentralized Gradient Decent (DGD)

- Network $\mathcal{G}$ with n nodes, weight matrix $W = [w_{ij}]_{n\times n}$
- At iteration t, node i:
  - sends $\mathbf{x}_{i,t}$ to neighbors $\mathcal{N}_i$ and receives $\mathbf{x}_{j,t}$
  - updates

$$\mathbf{x}_{i,t+1} = \underbrace{\sum_{j\in\mathcal{N}_i} w_{ij}\mathbf{x}_{j,t}}_{\substack{\text{average of local}\\ \text{and neighboring models}}} \underbrace{-\alpha\nabla f_i(\mathbf{x}_{i,t})}_{\text{local gradient descent}}$$

[Nedić, Ozdaglar,'07]

Theorem (Yuan, Ling, Yin '16)

Under A1,2,3, $\mathbf{x}_{i,t}$ geometrically converges to an $\mathcal{O}\left(\frac{\alpha}{1-\beta}\right)$-neighborhood of the unique solution $\mathbf{x}^*$. ($1-\beta$: spectral gap of W)

Related work in quantized setting: [Nedic et al., 2009], [Rabbat & Novak, 2005], [Li et al., 2016], [Zhang et al., 2019]

Decentralized Gradient Decent (DGD)

- Network $\mathcal{G}$ with n nodes, weight matrix $W = [w_{ij}]_{n \times n}$
- At iteration t, node i:
  - sends $\mathbf{x}_{i,t}$ to neighbors $\mathcal{N}_i$ and receives $\mathbf{x}_{j,t}$
  - updates

$$\mathbf{x}_{i,t+1} = \underbrace{\sum_{j \in \mathcal{N}_i} w_{ij} \mathbf{x}_{j,t}}_{\substack{\text{average of local} \\ \text{and neighboring models}}} \quad \underbrace{- \alpha \nabla f_i(\mathbf{x}_{i,t})}_{\text{local gradient descent}}$$

[Nedić, Ozdaglar,'07]

Theorem (Yuan, Ling, Yin '16)

Under A1,2,3, $\mathbf{x}_{i,t}$ geometrically converges to an $\mathcal{O}\left(\frac{\alpha}{1-\beta}\right)$-neighborhood of the unique solution $\mathbf{x}^*$. ($1-\beta$: spectral gap of W)

Related work in quantized setting: [Nedic et al., 2009], [Rabbat & Novak, 2005], [Li et al., 2016], [Zhang et al., 2019]

No "EXACT" convergence!

Assumptions

Assumptions

A1 Local objectives f_i are differentiable & L-smooth:

Assumptions

A1 Local objectives f_i are differentiable & L-smooth:

$$\|\nabla f_i(\mathbf{x}) - \nabla f_i(\mathbf{y})\| \leq L \|\mathbf{x} - \mathbf{y}\| \qquad \forall \mathbf{x}, \mathbf{y}$$

Assumptions

A1 Local objectives f_i are differentiable & L-smooth:

$$\|\nabla f_i(\mathbf{x}) - \nabla f_i(\mathbf{y})\| \leq L \|\mathbf{x} - \mathbf{y}\| \qquad \forall \mathbf{x}, \mathbf{y}$$

A2 Local objectives f_i are μ-strongly convex:

Assumptions

A1 Local objectives f_i are differentiable & L-smooth:

$$\|\nabla f_i(\mathbf{x}) - \nabla f_i(\mathbf{y})\| \leq L \|\mathbf{x} - \mathbf{y}\| \qquad \forall \mathbf{x}, \mathbf{y}$$

A2 Local objectives f_i are μ-strongly convex:

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

Assumptions

A1 Local objectives f_i are differentiable & L-smooth:

$$\|\nabla f_i(\mathbf{x}) - \nabla f_i(\mathbf{y})\| \leq L \|\mathbf{x} - \mathbf{y}\| \qquad \forall \mathbf{x}, \mathbf{y}$$

A2 Local objectives f_i are μ-strongly convex:

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

A3 Weight matrix W is non-negative doubly stochastic:

Assumptions

A1 Local objectives f_i are differentiable & L-smooth:

$$\|\nabla f_i(\mathbf{x}) - \nabla f_i(\mathbf{y})\| \leq L \|\mathbf{x} - \mathbf{y}\| \qquad \forall \mathbf{x}, \mathbf{y}$$

A2 Local objectives f_i are μ-strongly convex:

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

A3 Weight matrix W is non-negative doubly stochastic:

$$w_{ij} \geq 0 \quad \& \quad W = W^\top \quad \& \quad W\mathbf{1} = \mathbf{1} \quad \& \quad \text{null}(I - W) = \text{span}(\mathbf{1})$$

Quantized-DGD

- Network $\mathcal{G}$ with n nodes, weight matrix $W \in \mathbb{R}_+^{n \times n}$
- At iteration t, node i:
  - sends $\mathbf{z}_{i,t} = Q(\mathbf{x}_{i,t})$ to neighbors $j \in \mathcal{N}_i$ and receives $\mathbf{z}_{j,t}$
  - updates

$$\mathbf{x}_{i,t+1} = \underbrace{(1 - \epsilon + \epsilon w_{ii})\mathbf{x}_{i,t}}_{\text{noiseless local model}} + \underbrace{\epsilon \sum_{j \in \mathcal{N}_i \setminus \{i\}} w_{ij} \mathbf{z}_{i,t}}_{\text{average of noisy neighboring models}} \underbrace{- \alpha \epsilon \nabla f_i(\mathbf{x}_{i,t})}_{\text{local gradient descent}}$$

Quantized-DGD

- Network $\mathcal{G}$ with n nodes, weight matrix $W \in \mathbb{R}_+^{n \times n}$
- At iteration t, node i:
  - sends $\mathbf{z}_{i,t} = Q(\mathbf{x}_{i,t})$ to neighbors $j \in \mathcal{N}_i$ and receives $\mathbf{z}_{j,t}$
  - updates

$$\mathbf{x}_{i,t+1} = \underbrace{(1-\epsilon+\epsilon w_{ii})\mathbf{x}_{i,t}}_{\text{noiseless local model}} + \underbrace{\epsilon \sum_{j \in \mathcal{N}_i \setminus \{i\}} w_{ij} \mathbf{z}_{i,t}}_{\text{average of noisy neighboring models}} \underbrace{-\alpha\epsilon \nabla f_i(\mathbf{x}_{i,t})}_{\text{local gradient descent}}$$

Theorem (QDGD with variance-bounded quantization)

- A1,2,3,4 ✓

Quantized-DGD

- Network $\mathcal{G}$ with n nodes, weight matrix $W \in \mathbb{R}_{+}^{n \times n}$
- At iteration t, node i:
  - sends $\mathbf{z}_{i,t} = Q(\mathbf{x}_{i,t})$ to neighbors $j \in \mathcal{N}_i$ and receives $\mathbf{z}_{j,t}$
  - updates

$$\mathbf{x}_{i,t+1} = \underbrace{(1-\epsilon+\epsilon w_{ii})\mathbf{x}_{i,t}}_{\text{noiseless local model}} + \underbrace{\epsilon \sum_{j \in \mathcal{N}_i \setminus \{i\}} w_{ij} \mathbf{z}_{i,t}}_{\text{average of noisy neighboring models}} \underbrace{-\alpha\epsilon \nabla f_i(\mathbf{x}_{i,t})}_{\text{local gradient descent}}$$

Theorem (QDGD with variance-bounded quantization)

- A1,2,3,4 ✓
- fix $\delta \in (0,1)$ & large enough T

Quantized-DGD

- Network $\mathcal{G}$ with n nodes, weight matrix $W \in \mathbb{R}_+^{n\times n}$
- At iteration t, node i:
  - sends $\mathbf{z}_{i,t} = Q(\mathbf{x}_{i,t})$ to neighbors $j \in \mathcal{N}_i$ and receives $\mathbf{z}_{j,t}$
  - updates

$$\mathbf{x}_{i,t+1} = \underbrace{(1-\epsilon+\epsilon w_{ii})\mathbf{x}_{i,t}}_{\text{noiseless local model}} + \underbrace{\epsilon \sum_{j\in\mathcal{N}_i\setminus\{i\}} w_{ij}\mathbf{z}_{i,t}}_{\text{average of noisy neighboring models}} \underbrace{-\alpha\epsilon\nabla f_i(\mathbf{x}_{i,t})}_{\text{local gradient descent}}$$

Theorem (QDGD with variance-bounded quantization)

- A1,2,3,4 ✓
- fix $\delta \in (0,1)$ & large enough T
- pick $\epsilon = \frac{c_1}{T^{\frac{3}{4}(1-\delta)}}$, $\alpha = \frac{c_2}{T^{\frac{1}{4}(1-\delta)}}$

Quantized-DGD

- Network $\mathcal{G}$ with n nodes, weight matrix $W \in \mathbb{R}_+^{n\times n}$
- At iteration t, node i:
  - sends $\mathbf{z}_{i,t} = Q(\mathbf{x}_{i,t})$ to neighbors $j \in \mathcal{N}_i$ and receives $\mathbf{z}_{j,t}$
  - updates

$$\mathbf{x}_{i,t+1} = \underbrace{(1-\epsilon+\epsilon w_{ii})\mathbf{x}_{i,t}}_{\text{noiseless local model}} + \underbrace{\epsilon \sum_{j\in\mathcal{N}_i\setminus\{i\}} w_{ij}\mathbf{z}_{i,t}}_{\text{average of noisy neighboring models}} \underbrace{-\alpha\epsilon\nabla f_i(\mathbf{x}_{i,t})}_{\text{local gradient descent}}$$

Theorem (QDGD with variance-bounded quantization)

- A1,2,3,4 ✓
- fix $\delta \in (0,1)$ & large enough T
- pick $\epsilon = \frac{c_1}{T^{\frac{3}{4}(1-\delta)}}$, $\alpha = \frac{c_2}{T^{\frac{1}{4}(1-\delta)}}$
- then

$$\mathbb{E}\Big[\left\|\mathbf{x}_{i,T}-\mathbf{x}^*\right\|^2\Big] \le \mathcal{O}\left(\frac{(1-\beta)^{-2}+n\sigma^2\left\|W-W_D\right\|^2}{T^{\frac{1-\delta}{2}}}\right)$$

A1 Local objectives f_i are differentiable & L-smooth:

$$\|\nabla f_i(\mathbf{x}) - \nabla f_i(\mathbf{y})\| \leq L \|\mathbf{x} - \mathbf{y}\| \qquad \forall \mathbf{x}, \mathbf{y}$$

A2 Local objectives f_i are μ-strongly convex:

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

A3 Weight matrix W is non-negative doubly stochastic:

$$w_{ij} \geq 0 \quad \& \quad W = W^\top \quad \& \quad W\mathbf{1} = \mathbf{1} \quad \& \quad \text{null}(I - W) = \text{span}(\mathbf{1})$$

A4 Random quantizer $Q(\cdot)$ is unbiased & variance-bounded:

$$\mathbb{E}\left[Q(\mathbf{x})|\mathbf{x}\right] = \mathbf{x} \quad \& \quad \mathbb{E}\left[\|Q(\mathbf{x}) - \mathbf{x}\|^2 |\mathbf{x}\right] \leq \sigma^2 \qquad \forall \mathbf{x}$$

. . . Assumptions

A1 Local objectives f_i are differentiable & L-smooth:

$$\|\nabla f_i(\mathbf{x}) - \nabla f_i(\mathbf{y})\| \leq L \|\mathbf{x} - \mathbf{y}\| \qquad \forall \mathbf{x}, \mathbf{y}$$

A2 Local objectives f_i are μ-strongly convex:

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

A3 Weight matrix W is non-negative doubly stochastic:

$$w_{ij} \geq 0 \quad \& \quad W = W^\top \quad \& \quad W\mathbf{1} = \mathbf{1} \quad \& \quad \text{null}(I - W) = \text{span}(\mathbf{1})$$

A4 Random quantizer $Q(\cdot)$ is unbiased & variance-bounded:

$$\mathbb{E}\left[Q(\mathbf{x})|\mathbf{x}\right] = \mathbf{x} \quad \& \quad \mathbb{E}\left[\|Q(\mathbf{x}) - \mathbf{x}\|^2 |\mathbf{x}\right] \leq \sigma^2 \qquad \forall \mathbf{x}$$

A1 Local objectives f_i are differentiable & L-smooth:

$$\|\nabla f_i(\mathbf{x}) - \nabla f_i(\mathbf{y})\| \leq L \|\mathbf{x} - \mathbf{y}\| \qquad \forall \mathbf{x}, \mathbf{y}$$

A2 Local objectives f_i are μ-strongly convex:

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

A3 Weight matrix W is non-negative doubly stochastic:

$$w_{ij} \geq 0 \quad \& \quad W = W^\top \quad \& \quad W\mathbf{1} = \mathbf{1} \quad \& \quad \text{null}(I - W) = \text{span}(\mathbf{1})$$

A4 Random quantizer $Q(\cdot)$ is unbiased & variance-bounded:

$$\mathbb{E}\left[Q(\mathbf{x})|\mathbf{x}\right] = \mathbf{x} \quad \& \quad \mathbb{E}\left[\|Q(\mathbf{x}) - \mathbf{x}\|^2 |\mathbf{x}\right] \leq \sigma^2 \qquad \forall \mathbf{x}$$

Proof Sketch I

- To solve the main problem:

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

Proof Sketch I

- To solve the main problem:

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

- Solve an equivalent:

$$\begin{aligned} \min_{\mathbf{x}\in\mathbb{R}^{np}} &\quad F(\mathbf{x}) = \sum_{i=1}^{n} f_i(\mathbf{x}_i) \\ \text{s.t.} &\quad \mathbf{x}_1 = \cdots = \mathbf{x}_n \end{aligned} \qquad \text{where } \mathbf{x} = \begin{pmatrix} \mathbf{x}_1 \\ \vdots \\ \mathbf{x}_n \end{pmatrix} \in \mathbb{R}^{np}$$

Proof Sketch I

- To solve the main problem:

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

- Solve an equivalent:

By A3: $\min_{\mathbf{x}\in\mathbb{R}^{np}} \quad F(\mathbf{x}) = \sum_{i=1}^{n} f_i(\mathbf{x}_i)$ where $\mathbf{W} = W \otimes I_p \in \mathbb{R}^{np\times np}$

s.t. $(\mathbf{I}-\mathbf{W})^{1/2}\mathbf{x} = 0$

Proof Sketch I

- To solve the main problem:

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

- Solve an equivalent:

By A3: $$\min_{\mathbf{x}\in\mathbb{R}^{np}} \quad F(\mathbf{x}) = \sum_{i=1}^{n} f_i(\mathbf{x}_i) \qquad \text{where } \mathbf{W} = W \otimes I_p \in \mathbb{R}^{np\times np}$$

$$\text{s.t.} \quad (\mathbf{I}-\mathbf{W})^{1/2}\mathbf{x} = 0$$

- Penalty function:

$$\text{Define } \forall\alpha : \qquad h_\alpha(\mathbf{x}) = \frac{1}{2}\mathbf{x}^\top(\mathbf{I}-\mathbf{W})\mathbf{x} + \alpha F(\mathbf{x})$$

Proof Sketch I

- To solve the main problem:

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

- Solve an equivalent:

By A3: $\min_{\mathbf{x}\in\mathbb{R}^{np}} \quad F(\mathbf{x}) = \sum_{i=1}^{n} f_i(\mathbf{x}_i)$ where $\mathbf{W} = W \otimes I_p \in \mathbb{R}^{np\times np}$

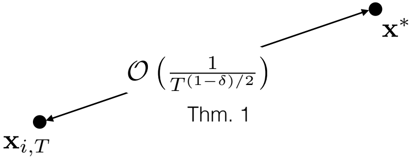
s.t. $\quad (\mathbf{I}-\mathbf{W})^{1/2}\mathbf{x} = 0$

- Penalty function:

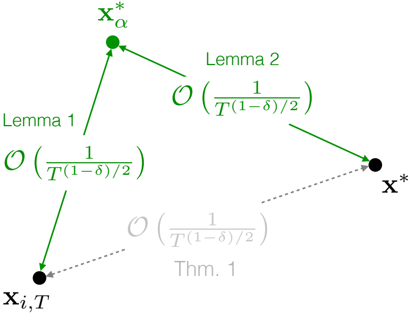
Define $\forall\alpha$: $\quad h_\alpha(\mathbf{x}) = \frac{1}{2}\mathbf{x}^\top\big(\mathbf{I}-\mathbf{W}\big)\mathbf{x} + \alpha F(\mathbf{x})$

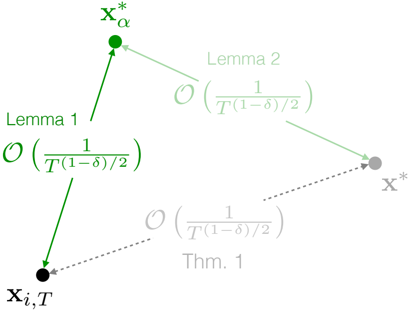
$$\mathbf{x}_\alpha^* = \underset{\mathbf{x}\in\mathbb{R}^{np}}{\operatorname{argmin}}\, h_\alpha(\mathbf{x})$$

Proof Sketch II



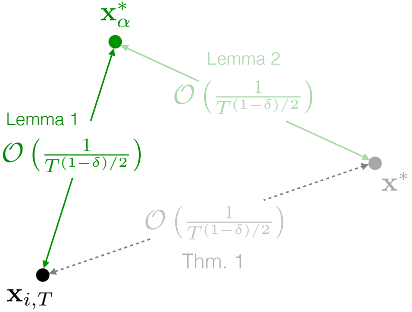
Proof Sketch II





Stochastic gradient descent on penalty function $h_\alpha(\mathbf{x}_t)$:

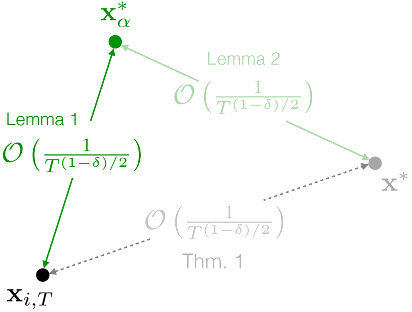
Proof Sketch II



Stochastic gradient descent on penalty function $h_\alpha(\mathbf{x}_t)$:

proposed update rule: $$\mathbf{x}_{t+1} = \mathbf{x}_t - \epsilon\Big(\underbrace{(\mathbf{I} - \mathbf{W}_D)\mathbf{x}_t + (\mathbf{W}_D - \mathbf{W})\mathbf{z}_t + \alpha \nabla F(\mathbf{x}_t)}_{\widetilde{\nabla} h_\alpha(\mathbf{x}_t) \quad \& \quad \mathbb{E}[\widetilde{\nabla} h_\alpha] = \nabla h_\alpha} \Big)$$

Proof Sketch II

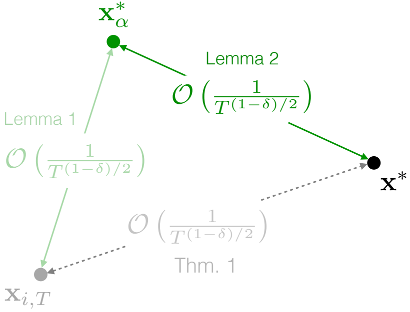


Stochastic gradient descent on penalty function $h_\alpha(\mathbf{x}_t)$:

proposed update rule: $$\mathbf{x}_{t+1} = \mathbf{x}_t - \epsilon\Big(\underbrace{(\mathbf{I} - \mathbf{W}_D)\mathbf{x}_t + (\mathbf{W}_D - \mathbf{W})\mathbf{z}_t + \alpha \nabla F(\mathbf{x}_t)}_{\widetilde{\nabla} h_\alpha(\mathbf{x}_t) \quad \& \quad \mathbb{E}[\widetilde{\nabla} h_\alpha] = \nabla h_\alpha}\Big)$$

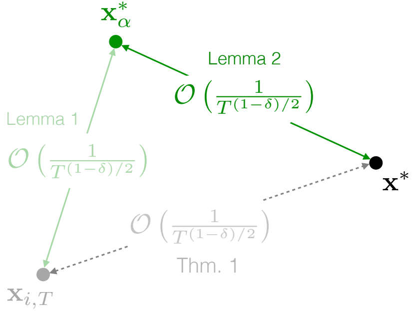
$$\Rightarrow \mathbb{E}\Big[\|\mathbf{x}_T - \mathbf{x}_\alpha^*\|^2\Big] \leq \mathcal{O}\left(\frac{c_1 n \sigma^2 \|W - W_D\|^2}{\mu c_2} \frac{1}{T^{(1-\delta)/2}}\right)$$ Lemma 1 ✓

Proof Sketch II



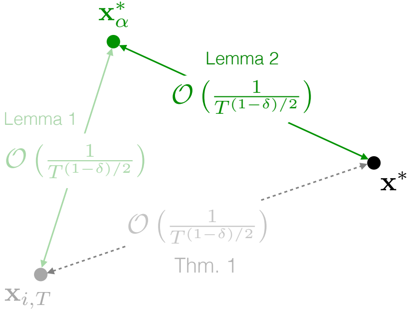
Gradient descent on penalty function $h_\alpha(\mathbf{x}_t)$:

Proof Sketch II



Gradient descent on penalty function $h_\alpha(\mathbf{x}_t)$:

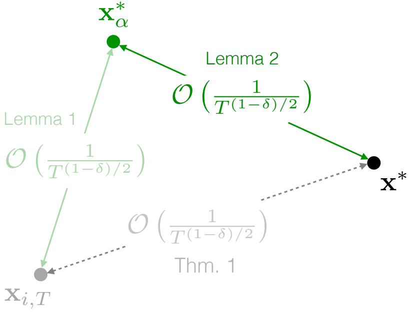
$$\begin{cases} \mathbf{u}_{t+1} = \mathbf{u}_t - 1 \cdot \nabla h_\alpha(\mathbf{u}_t) \\ \\ \end{cases}$$



Gradient descent on penalty function $h_\alpha(\mathbf{x}_t)$:

$$\left\{ \mathbf{u}_{t+1} = \mathbf{u}_t - 1 \cdot \nabla h_\alpha(\mathbf{u}_t) \Rightarrow \left\|\mathbf{u}_T - \mathbf{x}_\alpha^*\right\|^2 \leq e^{-c_2 T^{(3+\delta)/4}} \left\|\mathbf{u}_0 - \mathbf{x}_\alpha^*\right\|^2 \right.$$

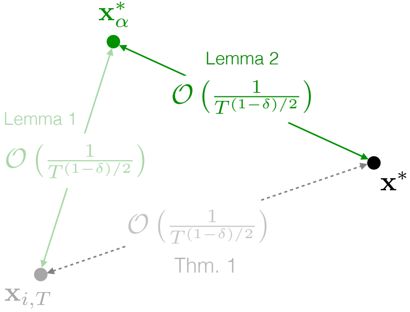
Proof Sketch II



Gradient descent on penalty function $h_\alpha(\mathbf{x}_t)$:

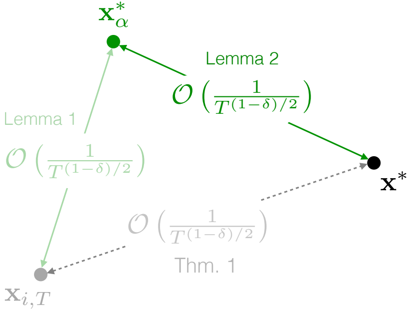
$$\begin{cases} \mathbf{u}_{t+1} = \mathbf{u}_t - 1 \cdot \nabla h_\alpha(\mathbf{u}_t) \Rightarrow \left\|\mathbf{u}_T - \mathbf{x}_\alpha^*\right\|^2 \leq e^{-c_2 T^{(3+\delta)/4}} \left\|\mathbf{u}_0 - \mathbf{x}_\alpha^*\right\|^2 \\ \mathbf{u}_{t+1} = \mathbf{W}\mathbf{u}_t - \alpha \nabla F(\mathbf{u}_t) \end{cases}$$

Proof Sketch II



Gradient descent on penalty function $h_\alpha(\mathbf{x}_t)$:

$$\begin{cases} \mathbf{u}_{t+1} = \mathbf{u}_t - 1 \cdot \nabla h_\alpha(\mathbf{u}_t) \Rightarrow \|\mathbf{u}_T - \mathbf{x}_\alpha^*\|^2 \leq e^{-c_2 T^{(3+\delta)/4}} \|\mathbf{u}_0 - \mathbf{x}_\alpha^*\|^2 \\ \mathbf{u}_{t+1} = \mathbf{W}\mathbf{u}_t - \alpha \nabla F(\mathbf{u}_t) \quad \text{Y-L-Y'16} \Rightarrow \quad \|\mathbf{u}_T - \mathbf{x}^*\| \leq \mathcal{O}\left(\frac{\alpha}{1-\beta}\right) \end{cases}$$



Gradient descent on penalty function $h_\alpha(\mathbf{x}_t)$:

$$\begin{cases} \mathbf{u}_{t+1} = \mathbf{u}_t - 1 \cdot \nabla h_\alpha(\mathbf{u}_t) \Rightarrow \|\mathbf{u}_T - \mathbf{x}_\alpha^*\|^2 \leq e^{-c_2 T^{(3+\delta)/4}} \|\mathbf{u}_0 - \mathbf{x}_\alpha^*\|^2 \\ \mathbf{u}_{t+1} = \mathbf{W}\mathbf{u}_t - \alpha \nabla F(\mathbf{u}_t) \quad \text{Y-L-Y'16} \Rightarrow \quad \|\mathbf{u}_T - \mathbf{x}^*\| \leq \mathcal{O}\left(\frac{\alpha}{1-\beta}\right) \end{cases}$$

$$\Rightarrow \|\mathbf{x}_\alpha^* - \mathbf{x}^*\|^2 = \|\mathbf{x}_\alpha^* - \mathbf{u}_T + \mathbf{u}_T - \mathbf{x}^*\|^2 \leq \mathcal{O}\left(\frac{\alpha^2}{(1-\beta)^2}\right) = \mathcal{O}\left(\frac{(1-\beta)^{-2}}{T^{(1-\delta)/2}}\right)$$

Lemma 2 ✓

More quantizers

A4 $\mathbb{E}\left[Q(\mathbf{x})|\mathbf{x}\right] = \mathbf{x} \quad \& \quad \mathbb{E}\left[\|Q(\mathbf{x}) - \mathbf{x}\|^2 \,|\mathbf{x}\right] \leq \sigma^2$ for all $\mathbf{x}$

More quantizers

A4 $\mathbb{E}\left[Q(\mathbf{x})|\mathbf{x}\right] = \mathbf{x}$ & $\mathbb{E}\left[\|Q(\mathbf{x}) - \mathbf{x}\|^2 |\mathbf{x}\right] \leq \sigma^2$ for all $\mathbf{x}$

A4' $\mathbb{E}\left[Q(\mathbf{x})|\mathbf{x}\right] = \mathbf{x}$ & $\mathbb{E}\left[\|Q(\mathbf{x}) - \mathbf{x}\|^2 |\mathbf{x}\right] \leq \eta^2 \|\mathbf{x}\|^2$ for all $\mathbf{x}$

More quantizers

A4 $\mathbb{E}\left[Q(\mathbf{x})|\mathbf{x}\right] = \mathbf{x} \quad \& \quad \mathbb{E}\left[\left\|Q(\mathbf{x}) - \mathbf{x}\right\|^2 |\mathbf{x}\right] \leq \sigma^2$ for all $\mathbf{x}$

A4' $\mathbb{E}\left[Q(\mathbf{x})|\mathbf{x}\right] = \mathbf{x} \quad \& \quad \mathbb{E}\left[\left\|Q(\mathbf{x}) - \mathbf{x}\right\|^2 |\mathbf{x}\right] \leq \eta^2 \left\|\mathbf{x}\right\|^2$ for all $\mathbf{x}$

Theorem

Under A1,2,3,4', the same rate is achieved.

More quantizers

A4 $\mathbb{E}[Q(\mathbf{x})|\mathbf{x}] = \mathbf{x}$ & $\mathbb{E}\left[\|Q(\mathbf{x}) - \mathbf{x}\|^2 |\mathbf{x}\right] \leq \sigma^2$ for all $\mathbf{x}$

A4' $\mathbb{E}[Q(\mathbf{x})|\mathbf{x}] = \mathbf{x}$ & $\mathbb{E}\left[\|Q(\mathbf{x}) - \mathbf{x}\|^2 |\mathbf{x}\right] \leq \eta^2 \|\mathbf{x}\|^2$ for all $\mathbf{x}$

Theorem

Under A1,2,3,4', the same rate is achieved.

- Example: Low-precision Q.

$$Q_i^{\text{LP}}(\mathbf{x}) = \|\mathbf{x}\| \cdot \text{sign}(x_i) \cdot \xi_i(\mathbf{x}) \quad \& \quad \xi_i(\mathbf{x}) \text{ is a Bernoulli r.v. with parameter } \frac{|x_i|}{\|\mathbf{x}\|}$$

More quantizers

A4 $\mathbb{E}\left[Q(\mathbf{x})|\mathbf{x}\right] = \mathbf{x} \quad \& \quad \mathbb{E}\left[\|Q(\mathbf{x}) - \mathbf{x}\|^2 |\mathbf{x}\right] \leq \sigma^2$ for all $\mathbf{x}$

A4' $\mathbb{E}\left[Q(\mathbf{x})|\mathbf{x}\right] = \mathbf{x} \quad \& \quad \mathbb{E}\left[\|Q(\mathbf{x}) - \mathbf{x}\|^2 |\mathbf{x}\right] \leq \eta^2 \|\mathbf{x}\|^2$ for all $\mathbf{x}$

Theorem

Under A1,2,3,4', the same rate is achieved.

- Example: Low-precision Q.

$$Q_i^{\text{LP}}(\mathbf{x}) = \|\mathbf{x}\| \cdot \text{sign}(x_i) \cdot \xi_i(\mathbf{x}) \quad \& \quad \xi_i(\mathbf{x}) \text{ is a Bernoulli r.v. with parameter } \frac{|x_i|}{\|\mathbf{x}\|}$$

$$\mathbf{x} = \begin{pmatrix} + \\ - \\ - \\ \vdots \\ + \end{pmatrix}$$

More quantizers

A4 $\mathbb{E}\left[Q(\mathbf{x})|\mathbf{x}\right] = \mathbf{x} \quad \& \quad \mathbb{E}\left[\|Q(\mathbf{x}) - \mathbf{x}\|^2 |\mathbf{x}\right] \leq \sigma^2$ for all $\mathbf{x}$

A4' $\mathbb{E}\left[Q(\mathbf{x})|\mathbf{x}\right] = \mathbf{x} \quad \& \quad \mathbb{E}\left[\|Q(\mathbf{x}) - \mathbf{x}\|^2 |\mathbf{x}\right] \leq \eta^2 \|\mathbf{x}\|^2$ for all $\mathbf{x}$

Theorem

Under A1,2,3,4', the same rate is achieved.

- Example: Low-precision Q.

$$Q_i^{\text{LP}}(\mathbf{x}) = \|\mathbf{x}\| \cdot \text{sign}(x_i) \cdot \xi_i(\mathbf{x}) \quad \& \quad \xi_i(\mathbf{x}) \text{ is a Bernoulli r.v. with parameter } \frac{|x_i|}{\|\mathbf{x}\|}$$

$$\mathbf{x} = \begin{pmatrix} + \\ - \\ - \\ \vdots \\ + \end{pmatrix} \xrightarrow{Q^{\text{LP}}} \begin{pmatrix} \|\mathbf{x}\| \\ 0 \\ -\|\mathbf{x}\| \\ \vdots \\ 0 \end{pmatrix}$$

More quantizers

A4 $\mathbb{E}\left[Q(\mathbf{x})|\mathbf{x}\right] = \mathbf{x}$ & $\mathbb{E}\left[\|Q(\mathbf{x}) - \mathbf{x}\|^2 \,|\mathbf{x}\right] \leq \sigma^2$ for all $\mathbf{x}$

A4' $\mathbb{E}\left[Q(\mathbf{x})|\mathbf{x}\right] = \mathbf{x}$ & $\mathbb{E}\left[\|Q(\mathbf{x}) - \mathbf{x}\|^2 \,|\mathbf{x}\right] \leq \eta^2 \|\mathbf{x}\|^2$ for all $\mathbf{x}$

Theorem

Under A1,2,3,4', the same rate is achieved.

- Example: Low-precision Q.

$$Q_i^{\text{LP}}(\mathbf{x}) = \|\mathbf{x}\| \cdot \text{sign}(x_i) \cdot \xi_i(\mathbf{x}) \quad \& \quad \xi_i(\mathbf{x}) \text{ is a Bernoulli r.v. with parameter } \frac{|x_i|}{\|\mathbf{x}\|}$$

$$\mathbf{x} = \begin{pmatrix} + \\ - \\ - \\ \vdots \\ + \end{pmatrix} \xrightarrow{Q^{\text{LP}}} \begin{pmatrix} \|\mathbf{x}\| \\ 0 \\ -\|\mathbf{x}\| \\ \vdots \\ 0 \end{pmatrix}$$

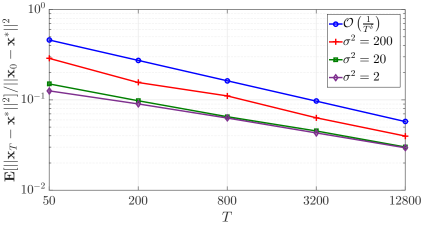
[1-bit SGD, Seide et al., 14], [QSGD, Alistarh, et al.,'17]

Numerical results: synthetic data

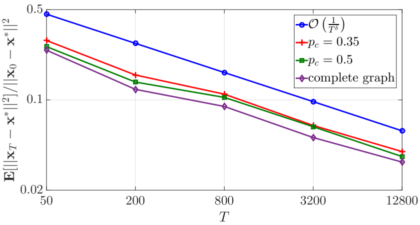
- Decentralized quadratic: $\min_{\mathbf{x} \in \mathbb{R}^p} \; f(\mathbf{x}) = \sum_{i=1}^{n} \frac{1}{2} \mathbf{x}^\top \mathbf{A}_i \mathbf{x} + \mathbf{b}_i^\top \mathbf{x}$, $p = 20$
- Network: Erdős-Rényi graph, $n = 50$ nodes, connectivity prob. $p_c = 0.35$
- Weight matrix: $W = I - \frac{2}{3\lambda_{\mathsf{max}}(\mathbf{L})} \mathbf{L}$ where $\mathbf{L}$ is the Laplacian

Numerical results: synthetic data

- Decentralized quadratic: $\min_{\mathbf{x}\in\mathbb{R}^p} \; f(\mathbf{x}) = \sum_{i=1}^{n} \frac{1}{2}\mathbf{x}^\top \mathbf{A}_i \mathbf{x} + \mathbf{b}_i^\top \mathbf{x}$, $p = 20$
- Network: Erdős-Rényi graph, $n = 50$ nodes, connectivity prob. $p_c = 0.35$
- Weight matrix: $W = I - \frac{2}{3\lambda_{\max}(\mathbf{L})}\mathbf{L}$ where $\mathbf{L}$ is the Laplacian



varying Q. noise



varying graph

Numerical results: total communication cost

- Decentralized least squared: $\min_{\mathbf{x}\in\mathbb{R}^p} f(\mathbf{x}) = \sum_{i=1}^{n} \frac{1}{2} \left\| \mathbf{A}_i \mathbf{x} - \mathbf{b}_i \right\|^2$
- Normal data-set $p = 200$, Erdős-Rényi $n = 50$, $p_c = 0.35$, $b = 64$
- Quantizer: Low-precision Q.: $Q_i^{\mathsf{LP}}(\mathbf{x}) = \|\mathbf{x}\| \cdot \mathsf{sign}(x_i) \cdot \xi_i(\mathbf{x}, s)$, $s = 1, 2, \cdots$

Numerical results: total communication cost

- Decentralized least squared: $\min_{\mathbf{x}\in\mathbb{R}^p} f(\mathbf{x}) = \sum_{i=1}^{n} \frac{1}{2} \|\mathbf{A}_i\mathbf{x} - \mathbf{b}_i\|^2$
- Normal data-set $p = 200$, Erdős-Rényi $n = 50$, $p_c = 0.35$, $b = 64$
- Quantizer: Low-precision Q.: $Q_i^{\text{LP}}(\mathbf{x}) = \|\mathbf{x}\| \cdot \text{sign}(x_i) \cdot \xi_i(\mathbf{x}, s)$, $s = 1, 2, \cdots$

# quantization levels	# iterations ($\times 10^3$)	code length per vector (bits)	communication cost (bits) ($\times 10^8$)
$s = 1$	614.2	216.9	66.6
$s = 10$	11.69	678.2	3.96
$s^* = 50$	2.3	949.8	1.09
$s = 70$	2.14	1037	1.11

communication cost to reach 0.01 of the optimal

Numerical results: total communication cost

- Decentralized least squared: $\min_{\mathbf{x}\in\mathbb{R}^p} f(\mathbf{x}) = \sum_{i=1}^{n} \frac{1}{2} \|\mathbf{A}_i\mathbf{x} - \mathbf{b}_i\|^2$
- Normal data-set $p = 200$, Erdős-Rényi $n = 50$, $p_c = 0.35$, $b = 64$
- Quantizer: Low-precision Q.: $Q_i^{\text{LP}}(\mathbf{x}) = \|\mathbf{x}\| \cdot \text{sign}(x_i) \cdot \xi_i(\mathbf{x}, s)$, $s = 1, 2, \cdots$

# quantization levels	# iterations ($\times 10^3$)	code length per vector (bits)	communication cost (bits) ($\times 10^8$)
$s = 1$	614.2	216.9	66.6
$s = 10$	11.69	678.2	3.96
$s^* = 50$	2.3	949.8	1.09
$s = 70$	2.14	1037	1.11

communication cost to reach 0.01 of the optimal

Conclusion

- We considered the attack that the adversary reduces communication bandwidth on the links, and proposed an exact decentralized gradient decent algorithm for quantized communications.
- Many interesting directions to continue:
  - Numerical study for IEEE 39-bus power network is ongoing.
  - best quantizer?
  - adversarial nodes?
  - link failures?
  - most resilient network topology?