Refined Convergence and Topology Learning for Decentralized Optimization with Heterogeneous Data

CAp 2022

Batiste Le Bars

Magnet, Inria Lille

Joint work with: A. Bellet (Inria), M. Tommasi (Inria), E. Lavoie (EPFL), AM. Kermarrec (EPFL)

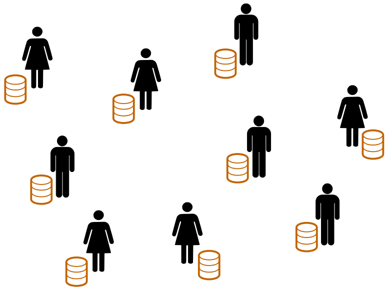
Thursday, July 7th, 2022

Background and motivations

Fully decentralized learning

Framework

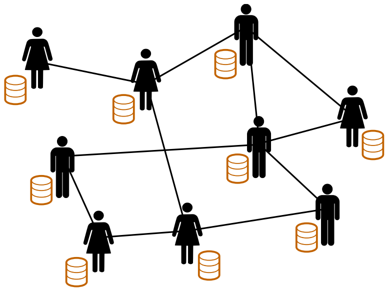
- Decentralized data
- Centralization is not allowed



Fully decentralized learning

Framework

- Decentralized data
- Centralization is not allowed

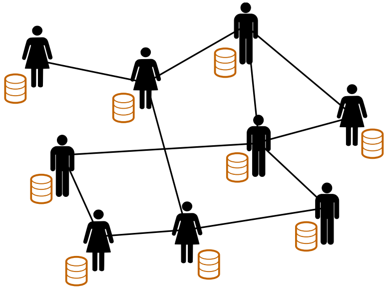


- Agents can collaborate
- Communication according to a graph topology

Fully decentralized learning

Framework

- Decentralized data
- Centralization is not allowed



- Agents can collaborate
- Communication according to a graph topology

Problem setting

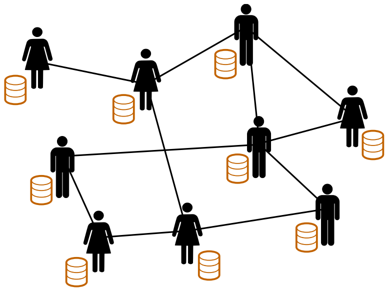
- n agents (nodes) seeking to optimize

$$\min_{\theta\in\mathbb{R}^d} \left[f(\theta) \triangleq \frac{1}{n}\sum_{i=1}^{n} f_i(\theta)\right],$$

Fully decentralized learning

Framework

- Decentralized data
- Centralization is not allowed



- Agents can collaborate
- Communication according to a graph topology

Problem setting

- n agents (nodes) seeking to optimize

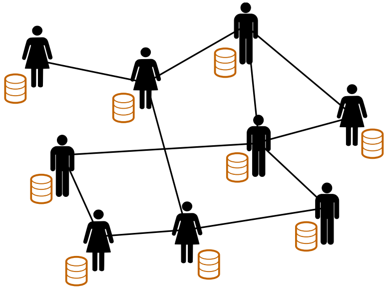
$$\min_{\theta \in \mathbb{R}^d} \left[f(\theta) \triangleq \frac{1}{n} \sum_{i=1}^{n} f_i(\theta) \right],$$

- $f_i(\theta) \triangleq \mathbb{E}_{Z_i \sim \mathcal{D}_i}[F_i(\theta; Z_i)]$
- $F_i =$ local loss function
- $\mathcal{D}_i =$ local data distribution (heterogeneity)

Fully decentralized learning

Framework

- Decentralized data
- Centralization is not allowed



- Agents can collaborate
- Communication according to a graph topology

Problem setting

- n agents (nodes) seeking to optimize

$$\min_{\theta\in\mathbb{R}^d}\left[f(\theta)\triangleq\frac{1}{n}\sum_{i=1}^{n}f_i(\theta)\right],$$

- $f_i(\theta)\triangleq\mathbb{E}_{Z_i\sim\mathcal{D}_i}[F_i(\theta;Z_i)]$
- F_i = local loss function
- $\mathcal{D}_i$ = local data distribution (heterogeneity)
- Communication topology is $W\in[0,1]^{n\times n}$
- $W_{ij}=0$ (no edge) $\Leftrightarrow$ node i and j cannot communicate

Decentralized Stochastic Gradient Descent (D-SGD)

Algorithm

- $W \in [0,1]^{n \times n}$ is doubly stochastic
- It can change across iterations t

D-SGD (Lian et al., 2017)

Input: $\theta_i^{(0)} = \theta^{(0)} \in \mathbb{R}^d$, stepsizes $\{\eta_t\}_{t=0}^{T-1}$, mixing $\{W^{(t)}\}_{t=0}^{T-1}$
for $t = 0, \ldots, T-1$ **do**
  for each node $i = 1, \ldots, n$ **do**
    Sample $Z_i^{(t)} \sim \mathcal{D}_i$
    1. $\theta_i^{(t+\frac{1}{2})} \leftarrow \theta_i^{(t)} - \eta_t \nabla F_i(\theta_i^{(t)}, Z_i^{(t)})$
    2. $\theta_i^{(t+1)} \leftarrow \sum_{j=1}^n W_{ij}^{(t)} \theta_j^{(t+\frac{1}{2})}$
  end for
end for

Decentralized Stochastic Gradient Descent (D-SGD)

Algorithm

- $W \in [0,1]^{n \times n}$ is doubly stochastic
- It can change across iterations t

D-SGD (Lian et al., 2017)

Input: $\theta_i^{(0)} = \theta^{(0)} \in \mathbb{R}^d$, stepsizes $\{\eta_t\}_{t=0}^{T-1}$, mixing $\{W^{(t)}\}_{t=0}^{T-1}$

for $t = 0, \dots, T-1$ **do**

  for each node $i = 1, \dots, n$ **do**

    Sample $Z_i^{(t)} \sim \mathcal{D}_i$

   1. $\theta_i^{(t+\frac{1}{2})} \leftarrow \theta_i^{(t)} - \eta_t \nabla F_i(\theta_i^{(t)}, Z_i^{(t)})$

   2. $\theta_i^{(t+1)} \leftarrow \sum_{j=1}^n W_{ij}^{(t)} \theta_j^{(t+\frac{1}{2})}$

  end for

end for

Impact of the topology

Decentralized Stochastic Gradient Descent (D-SGD)

Algorithm

- $W \in [0,1]^{n \times n}$ is doubly stochastic
- It can change across iterations t

D-SGD (Lian et al., 2017)

Input: $\theta_i^{(0)} = \theta^{(0)} \in \mathbb{R}^d$, stepsizes $\{\eta_t\}_{t=0}^{T-1}$, mixing $\{W^{(t)}\}_{t=0}^{T-1}$
for $t = 0, \dots, T-1$ **do**
  for each node $i = 1, \dots, n$ **do**
    Sample $Z_i^{(t)} \sim \mathcal{D}_i$
    1. $\theta_i^{(t+\frac{1}{2})} \leftarrow \theta_i^{(t)} - \eta_t \nabla F_i(\theta_i^{(t)}, Z_i^{(t)})$
    2. $\theta_i^{(t+1)} \leftarrow \sum_{j=1}^n W_{ij}^{(t)} \theta_j^{(t+\frac{1}{2})}$
  end for
end for

Impact of the topology

- Communication costs (maximum degree)
  $\rightarrow$ W should be sparse

Decentralized Stochastic Gradient Descent (D-SGD)

Algorithm

- $W \in [0,1]^{n\times n}$ is doubly stochastic
- It can change across iterations t

D-SGD (Lian et al., 2017)

Input: $\theta_i^{(0)} = \theta^{(0)} \in \mathbb{R}^d$, stepsizes $\{\eta_t\}_{t=0}^{T-1}$, mixing $\{W^{(t)}\}_{t=0}^{T-1}$
for $t = 0, \dots, T-1$ **do**
  for each node $i = 1, \dots, n$ **do**
    Sample $Z_i^{(t)} \sim \mathcal{D}_i$
    1. $\theta_i^{(t+\frac{1}{2})} \leftarrow \theta_i^{(t)} - \eta_t \nabla F_i(\theta_i^{(t)}, Z_i^{(t)})$
    2. $\theta_i^{(t+1)} \leftarrow \sum_{j=1}^n W_{ij}^{(t)} \theta_j^{(t+\frac{1}{2})}$
  end for
end for

Impact of the topology

- Communication costs (maximum degree)
  $\rightarrow$ W should be sparse
- Convergence speed
  $\rightarrow$ W should be sufficiently connected

Previous work and open questions

Based on the spectral gap of W

- Most common analysis (e.g. Koloskova et al. (2020); Lian et al. (2017); Wang et al. (2019))
- Small spectral gap $\Rightarrow$ dense matrix $W \Rightarrow$ D-SGD closer to centralized SGD
- Problem: convergence rates heavily impacted by heterogeneity!

$\rightarrow$ *Can we exhibit a better quantity?*

Previous work and open questions

Based on the spectral gap of W

- Most common analysis (e.g. Koloskova et al. (2020); Lian et al. (2017); Wang et al. (2019))
- Small spectral gap $\Rightarrow$ dense matrix $W \Rightarrow$ D-SGD closer to centralized SGD
- Problem: convergence rates heavily impacted by heterogeneity!

$\rightarrow$ *Can we exhibit a better quantity?*

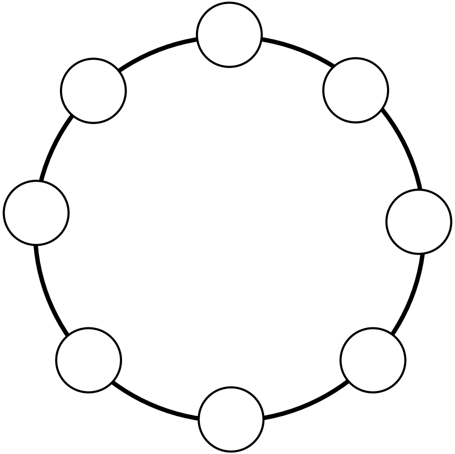
Data-dependent topology?

- D-cliques (Bellet et al., 2021)
- Topology that compensates data-heterogeneity
- Problem: Only empirical results, not flexible topology

$\rightarrow$ *Can we propose a data-dependent topology that is theoretically understood?*

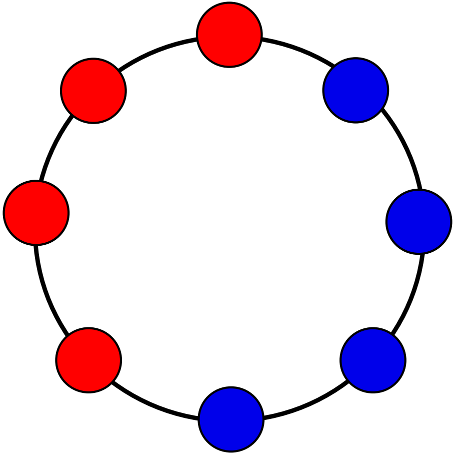
Toy example

- Half nodes have blue distribution, other half have red distribution
- Ring graph



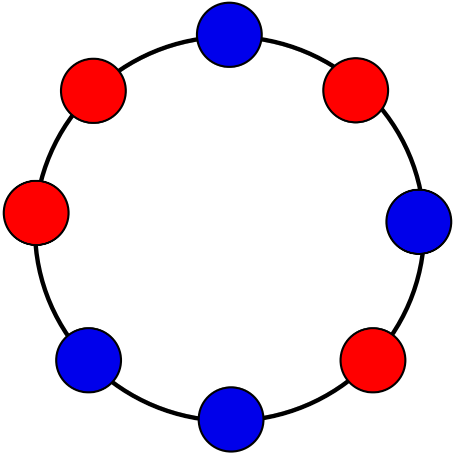
Toy example

- Half nodes have blue distribution, other half have red distribution
- Ring graph



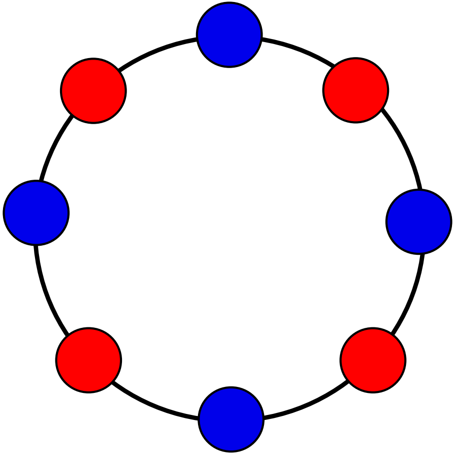
Toy example

- Half nodes have blue distribution, other half have red distribution
- Ring graph



Toy example

- Half nodes have blue distribution, other half have red distribution
- Ring graph



Refined convergence with Neighborhood Heterogeneity

Basic assumptions

Assumption 1 (*L-smoothness*)

$\exists L > 0$ s.t. $\forall Z \in \Omega_i, \theta, \tilde{\theta} \in \mathbb{R}^d$: $\quad \|\nabla F_i(\theta, Z) - \nabla F_i(\tilde{\theta}, Z)\| \leq L\|\theta - \tilde{\theta}\|$

Basic assumptions

Assumption 1 (*L-smoothness*)

$\exists L > 0$ s.t. $\forall Z \in \Omega_i, \theta, \tilde{\theta} \in \mathbb{R}^d$: $\quad \|\nabla F_i(\theta, Z) - \nabla F_i(\tilde{\theta}, Z)\| \leq L\|\theta - \tilde{\theta}\|$

Assumption 2 (*Bounded variance*)

$\forall i = 1, \ldots, n, \exists\, \sigma_i^2 > 0$ s.t. $\forall \theta \in \mathbb{R}^d$: $\quad \mathbb{E}_{Z\sim\mathcal{D}_i}\left[\, \|\nabla F_i(\theta, Z) - \nabla f_i(\theta)\|_2^2 \,\right] \leq \sigma_i^2$

Basic assumptions

Assumption 1 (*L-smoothness*)

$\exists L > 0$ s.t. $\forall Z \in \Omega_i, \theta, \tilde{\theta} \in \mathbb{R}^d$: $\quad \|\nabla F_i(\theta, Z) - \nabla F_i(\tilde{\theta}, Z)\| \leq L\|\theta - \tilde{\theta}\|$

Assumption 2 (*Bounded variance*)

$\forall i = 1, \ldots, n, \exists\, \sigma_i^2 > 0$ s.t. $\forall \theta \in \mathbb{R}^d$: $\quad \mathbb{E}_{Z \sim \mathcal{D}_i}\left[\, \|\nabla F_i(\theta, Z) - \nabla f_i(\theta)\|_2^2 \,\right] \leq \sigma_i^2$

Assumption 3 (*Mixing parameter*)

$\exists p \in [0, 1]$ s.t. $\forall M \in \mathbb{R}^{d \times n}$: $\quad \|MW^{\mathsf{T}} - \overline{M}\|_F^2 \leq (1 - p)\|M - \overline{M}\|_F^2$, with $\overline{M} = M \cdot \frac{1}{n}\mathbf{1}\mathbf{1}^{\mathsf{T}}$.

- p linked with spectral gap of W

Local vs Neighborhood heterogeneity

Previously: Bounded *local* heterogeneity assumption i.e. $\exists\, \bar{\zeta}^2 > 0$ s.t.
$\frac{1}{n}\sum_{i=1}^{n} \mathbb{E}\left\|\nabla F_i(\theta, Z_i) - \frac{1}{n}\sum_{j=1}^{n} \nabla F_j(\theta, Z_j)\right\|_2^2 \leq \bar{\zeta}^2, \quad \forall \theta \in \mathbb{R}^d.$

Local vs Neighborhood heterogeneity

Previously: Bounded *local* heterogeneity assumption i.e. $\exists\, \bar{\zeta}^2 > 0$ s.t.
$\frac{1}{n}\sum_{i=1}^{n} \mathbb{E}\left\|\nabla F_i(\theta, Z_i) - \frac{1}{n}\sum_{j=1}^{n} \nabla F_j(\theta, Z_j)\right\|_2^2 \le \bar{\zeta}^2, \quad \forall \theta \in \mathbb{R}^d.$

Now: Bounded neighborhood heterogeneity

Assumption 4 *(Bounded neighborhood heterogeneity)*

$\exists\, \bar{\tau}^2 > 0$ s.t.

$$H \triangleq \frac{1}{n}\sum_{i=1}^{n} \mathbb{E}\left\|\sum_{j=1}^{n} W_{ij}\nabla F_j(\theta, Z_j) - \frac{1}{n}\sum_{j=1}^{n} \nabla F_j(\theta, Z_j)\right\|_2^2 \le \bar{\tau}^2, \quad \forall \theta \in \mathbb{R}^d.$$

Local vs Neighborhood heterogeneity

Previously: Bounded *local* heterogeneity assumption i.e. $\exists\, \bar{\zeta}^2 > 0$ s.t.
$\frac{1}{n}\sum_{i=1}^{n} \mathbb{E}\left\|\nabla F_i(\theta, Z_i) - \frac{1}{n}\sum_{j=1}^{n} \nabla F_j(\theta, Z_j)\right\|_2^2 \leq \bar{\zeta}^2, \quad \forall \theta \in \mathbb{R}^d.$

Now: Bounded neighborhood heterogeneity

Assumption 4 *(Bounded neighborhood heterogeneity)*

$\exists\, \bar{\tau}^2 > 0$ s.t.

$$H \triangleq \frac{1}{n}\sum_{i=1}^{n} \mathbb{E}\left\|\sum_{j=1}^{n} W_{ij}\nabla F_j(\theta, Z_j) - \frac{1}{n}\sum_{j=1}^{n} \nabla F_j(\theta, Z_j)\right\|_2^2 \leq \bar{\tau}^2, \quad \forall \theta \in \mathbb{R}^d.$$

- Less restrictive (see Le Bars et al. (2022))
- Jointly quantifies the impact of W and heterogeneity
- Now W can compensate heterogeneity

Convergence result

Convergence Theorem (Informal)

Error ε achieved after T iterations with

Convex case:

$$T \geq \mathcal{O}\left(\frac{\bar{\sigma}^2}{n\varepsilon^2} + \frac{\sqrt{L}\bar{\tau}}{p\varepsilon^{\frac{3}{2}}} + \frac{L}{p\varepsilon}\right) r_0 \, ,$$

Non-convex case:

$$T \geq \mathcal{O}\left(\frac{L\bar{\sigma}^2}{n\varepsilon^2} + \frac{L\bar{\tau}}{p\varepsilon^{\frac{3}{2}}} + \frac{L}{p\varepsilon}\right) f_0 \, ,$$

- $r_0 = \|\theta^{(0)} - \theta^\star\|_2^2$, $f_0 = f(\theta^{(0)}) - f^\star$ and $\mathcal{O}(\cdot)$ hides the numerical constants.

Convergence result

Convergence Theorem (Informal)

Error ε achieved after T iterations with

Convex case:

$$T \geq \mathcal{O}\left(\frac{\bar{\sigma}^2}{n\varepsilon^2} + \frac{\sqrt{L\bar{\tau}}}{p\varepsilon^{\frac{3}{2}}} + \frac{L}{p\varepsilon}\right) r_0 \,,$$

Non-convex case:

$$T \geq \mathcal{O}\left(\frac{L\bar{\sigma}^2}{n\varepsilon^2} + \frac{L\bar{\tau}}{p\varepsilon^{\frac{3}{2}}} + \frac{L}{p\varepsilon}\right) f_0 \,,$$

- Green terms come from standard SGD

Convergence result

Convergence Theorem (Informal)

Error ε achieved after T iterations with

Convex case:

$$T \geq \mathcal{O}\left(\frac{\bar{\sigma}^2}{n\varepsilon^2} + \frac{\sqrt{L\bar{\tau}}}{p\varepsilon^{\frac{3}{2}}} + \frac{L}{p\varepsilon}\right) r_0 \,,$$

Non-convex case:

$$T \geq \mathcal{O}\left(\frac{L\bar{\sigma}^2}{n\varepsilon^2} + \frac{L\bar{\tau}}{p\varepsilon^{\frac{3}{2}}} + \frac{L}{p\varepsilon}\right) f_0 \,,$$

- Middle red term comes from decentralization

Convergence result

Convergence Theorem (Informal)

Error ε achieved after T iterations with

Convex case:

$$T \geq \mathcal{O}\left(\frac{\bar{\sigma}^2}{n\varepsilon^2} + \frac{\sqrt{L\bar{\tau}}}{p\varepsilon^{\frac{3}{2}}} + \frac{L}{p\varepsilon}\right) r_0 \,,$$

Non-convex case:

$$T \geq \mathcal{O}\left(\frac{L\bar{\sigma}^2}{n\varepsilon^2} + \frac{L\bar{\tau}}{p\varepsilon^{\frac{3}{2}}} + \frac{L}{p\varepsilon}\right) f_0 \,,$$

- Middle red term comes from decentralization
- Smaller constant in the middle term (see e.g. Koloskova et al. (2020))

Convergence result

Convergence Theorem (Informal)

Error ε achieved after T iterations with

Convex case:

$$T \geq \mathcal{O}\left(\frac{\bar{\sigma}^2}{n\varepsilon^2} + \frac{\sqrt{L\bar{\tau}}}{p\varepsilon^{\frac{3}{2}}} + \frac{L}{p\varepsilon}\right) r_0 \,,$$

Non-convex case:

$$T \geq \mathcal{O}\left(\frac{L\bar{\sigma}^2}{n\varepsilon^2} + \frac{L\bar{\tau}}{p\varepsilon^{\frac{3}{2}}} + \frac{L}{p\varepsilon}\right) f_0 \,,$$

- W controls p AND $\bar{\tau}$!

Convergence result

Convergence Theorem (Informal)

Error ε achieved after T iterations with

Convex case:

$$T \geq \mathcal{O}\Big(\frac{\bar{\sigma}^2}{n\varepsilon^2} + \frac{\sqrt{L\bar{\tau}}}{p\varepsilon^{\frac{3}{2}}} + \frac{L}{p\varepsilon}\Big) r_0 \,,$$

Non-convex case:

$$T \geq \mathcal{O}\Big(\frac{L\bar{\sigma}^2}{n\varepsilon^2} + \frac{L\bar{\tau}}{p\varepsilon^{\frac{3}{2}}} + \frac{L}{p\varepsilon}\Big) f_0 \,,$$

- W controls p AND $\bar{\tau}$!

Can we propose a topology W than minimizes $\bar{\tau}$ and the middle term?

Data-based topology learning

Model and objective

- Minimizing directly H not possible: need additional knowledge!
- $Z = (X, Y)$ with $Y = 1, \ldots, K$
- $\mathcal{D}_i = P(X|Y)P_i(Y)$ (label-skew)
- Assume $\Pi_{ik} = P_i(Y = k)$ is known

Model and objective

- Minimizing directly H not possible: need additional knowledge!
- $Z = (X, Y)$ with $Y = 1, \ldots, K$
- $\mathcal{D}_i = P(X|Y)P_i(Y)$ (label-skew)
- Assume $\Pi_{ik} = P_i(Y = k)$ is known

Proposition

$\exists\, \lambda > 0$ s.t. neighborhood heterogeneity H is upper bounded by

$$H \leq g(W) \triangleq \frac{1}{n}\left\| W\Pi - \frac{\mathbf{1}\mathbf{1}^{\mathsf{T}}}{n}\Pi \right\|_F^2 + \frac{\lambda}{n}\left\| W - \frac{\mathbf{1}\mathbf{1}^{\mathsf{T}}}{n} \right\|_F^2$$

Model and objective

- Minimizing directly H not possible: need additional knowledge!
- $Z = (X, Y)$ with $Y = 1, \ldots, K$
- $\mathcal{D}_i = P(X|Y)P_i(Y)$ (label-skew)
- Assume $\Pi_{ik} = P_i(Y = k)$ is known

Proposition

$\exists\, \lambda > 0$ s.t. neighborhood heterogeneity H is upper bounded by

$$H \leq g(W) \triangleq \frac{1}{n}\left\| W\Pi - \frac{\mathbf{1}\mathbf{1}^\mathsf{T}}{n}\Pi \right\|_F^2 + \frac{\lambda}{n}\left\| W - \frac{\mathbf{1}\mathbf{1}^\mathsf{T}}{n} \right\|_F^2$$

Objective: Minimize $g(W)$ s.t. W **doubly stochastic**

Model and objective

- Minimizing directly H not possible: need additional knowledge!
- $Z = (X, Y)$ with $Y = 1, \dots, K$
- $\mathcal{D}_i = P(X|Y)P_i(Y)$ (label-skew)
- Assume $\Pi_{ik} = P_i(Y = k)$ is known

Proposition

$\exists\, \lambda > 0$ s.t. neighborhood heterogeneity H is upper bounded by

$$H \leq g(W) \triangleq \frac{1}{n}\left\| W\Pi - \frac{\mathbf{1}\mathbf{1}^\top}{n}\Pi \right\|_F^2 + \frac{\lambda}{n}\left\| W - \frac{\mathbf{1}\mathbf{1}^\top}{n} \right\|_F^2$$

Objective: Minimize $g(W)$ s.t. W **doubly stochastic**

- Avoid trivial (dense) solution $W = \frac{1}{n}\mathbf{1}\mathbf{1}^\top$
- Find W sparse instead: using Frank-Wolfe!

Optimization with Frank-Wolfe

Algorithm (STL-FW)

Input: $\widehat{W}^{(0)} = I_n, \Pi \in [0,1]^{n \times K}$ and $\lambda > 0$
for $l = 0, \ldots, L$ **do**

1. $P^{(l+1)} = \arg\min_{P \in \mathcal{S}} \langle P, \nabla g(\widehat{W}^{(l)}) \rangle$ {Find best doubly-stochastic matrix}
2. $\gamma^{(l+1)} = \arg\min_{\gamma \in [0,1]} g\big((1-\gamma)\widehat{W}^{(l)} + \gamma P^{(l+1)}\big)$ {Line-search}
3. $\widehat{W}^{(l+1)} = (1 - \gamma^{(l+1)})\widehat{W}^{(l)} + \gamma^{(l+1)} P^{(l+1)}$ {Convex update}

end for

- Optimal solution of line 1. is sparse
- Closed-form solution for line 2.

Properties of the algorithm

Theorem (informal)

STL-FW converges to the optimal solution at a rate $\mathcal{O}(\frac{1}{t})$ and at the end of the t-th iteration, each node have at most t neighbors.

- Approximately minimizes an upper-bound over H
- Controls the level of sparsity

Setup

- Datasets: MNIST and CIFAR10 ($K = 10$ classes)

Setup

- Datasets: MNIST and CIFAR10 ($K = 10$ classes)
- Models: Linear and Group Normalized LeNet

Setup

- Datasets: MNIST and CIFAR10 ($K = 10$ classes)
- Models: Linear and Group Normalized LeNet
- $n = 100$, 1-4 classes per node

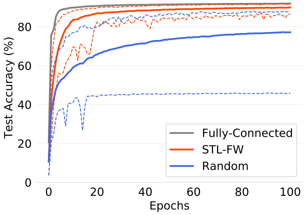
Setup

- Datasets: MNIST and CIFAR10 ($K = 10$ classes)
- Models: Linear and Group Normalized LeNet
- $n = 100$, 1-4 classes per node
- 3 competitors: Random, D-Cliques and Exponential graph

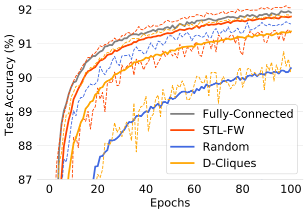
Setup

- Datasets: MNIST and CIFAR10 ($K = 10$ classes)
- Models: Linear and Group Normalized LeNet
- $n = 100$, 1-4 classes per node
- 3 competitors: Random, D-Cliques and Exponential graph
- Different level of sparsity (degree max). $d_{\text{max}} = 2, 5, 10$

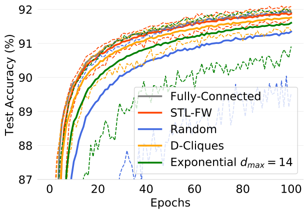
Results



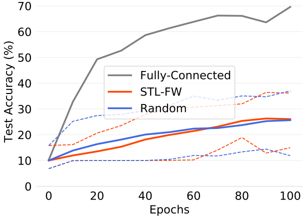
MNIST $d_{\max} = 2$



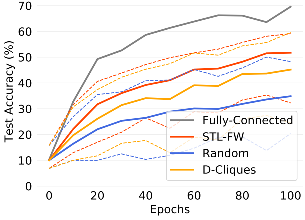
MNIST $d_{\max} = 5$



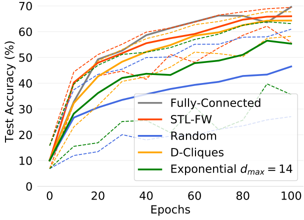
MNIST $d_{\max} = 10$



CIFAR10 $d_{\max} = 2$



CIFAR10 $d_{\max} = 5$



CIFAR10 $d_{\max} = 10$

Figure: Convergence of D-SGD with STL-FW (our approach) and alternative topologies.

Conclusion

Conclusion

Full paper: Le Bars, B., Bellet, A., Tommasi, M., Lavoie, E., and Kermarrec, A. (2022). *Refined convergence and topology learning for decentralized optimization with heterogeneous data*

Future directions:

- Explore more general framework
- Topology learning during D-SGD: using gradient knowledge
- Topology learning in presence of adversarial nodes
- ...

References

Bellet, A., Kermarrec, A.-M., and Lavoie, E. (2021). D-cliques: Compensating noniidness in decentralized federated learning with topology. *arXiv:2104.07365*.

Koloskova, A., Loizou, N., Boreiri, S., Jaggi, M., and Stich, S. U. (2020). A unified theory of decentralized sgd with changing topology and local updates. In *ICML*.

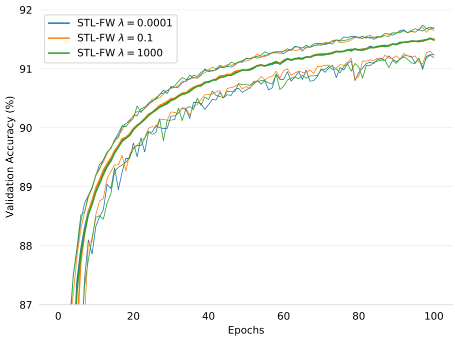
Le Bars, B., Bellet, A., Tommasi, M., Lavoie, E., and Kermarrec, A. (2022). Refined convergence and topology learning for decentralized optimization with heterogeneous data.

Lian, X., Zhang, C., Zhang, H., Hsieh, C.-J., Zhang, W., and Liu, J. (2017). Can decentralized algorithms outperform centralized algorithms? a case study for decentralized parallel stochastic gradient descent. *NIPS*.

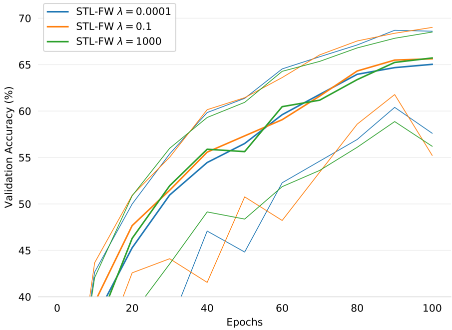
Wang, J., Sahu, A. K., Yang, Z., Joshi, G., and Kar, S. (2019). Matcha: Speeding up decentralized sgd via matching decomposition sampling. In *ICC*.

Appendix

Impact of λ



MNIST



CIFAR10

Figure: Effect of the hyperparameter λ of STL-FW on the convergence speed of D-SGD with 100 nodes, $d_{max} = 10$.