Quantized Decentralized Stochastic Learning over Directed Graphs

Hossein Taheri[1]

Joint work with Aryan Mokhtari[2], Hamed Hassani[3], and Ramtin Pedarsani[1]

[1]University of California, Santa Barbara
[2]University of Texas, Austin
[3]University of Pennsylvania

Thirty-seventh International Conference on Machine Learning (ICML), 2020

Decentralized Optimization

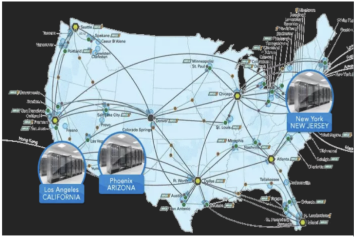
- Decentralized Stochastic Learning involves multiple agents or nodes that collect data, and want to learn an ML model collaboratively.

Decentralized Optimization

- Decentralized Stochastic Learning involves multiple agents or nodes that collect data, and want to learn an ML model collaboratively.
- Applications including federated learning, multi-agent robotics systems, sensor networks, etc.

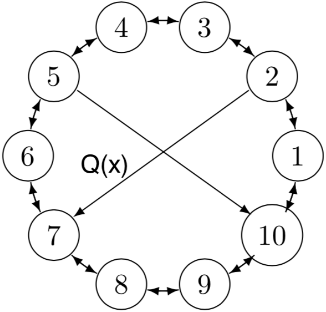
Decentralized Optimization

- Decentralized Stochastic Learning involves multiple agents or nodes that collect data, and want to learn an ML model collaboratively.
- Applications including federated learning, multi-agent robotics systems, sensor networks, etc.
- In many cases, communication links are asymmetric due to failures and bottlenecks and communication is done over a directed graph [Tsianos et al. 2012, Nedic et al. 2014, Assran et al. 2020].



This Talk

- Link failure: Nodes communicate over a directed graph
- High communication cost: Nodes communicate compressed information $Q(x)$
  Compression operator $Q : \mathbb{R}^d \to \mathbb{R}^d$



Introduction: Push-sum Algorithm

- Decentralized optimization over *directed* graphs with *exact* communication:

$$
\begin{cases}
\mathbf{x}_i(t+1) & = \sum_{j=1}^{n} w_{ij}\,\mathbf{x}_j(t) - \alpha(t)\nabla f_i\left(\mathbf{z}_i(t)\right) \\
y_i(t+1) & = \sum_{j=1}^{n} w_{ij}\,y_j(t) \\
\mathbf{z}_i(t+1) & = \mathbf{x}_i(t+1)/y_i(t+1)
\end{cases}
$$

Introduction: Push-sum Algorithm

- Decentralized optimization over *directed* graphs with *exact* communication:

$$\begin{cases} \mathbf{x}_i(t+1) & = \sum_{j=1}^{n} w_{ij}\, \mathbf{x}_j(t) - \alpha(t) \nabla f_i\left(\mathbf{z}_i(t)\right) \\ y_i(t+1) & = \sum_{j=1}^{n} w_{ij}\, y_j(t) \\ \mathbf{z}_i(t+1) & = \mathbf{x}_i(t+1)/y_i(t+1) \end{cases}$$

- [Nedic et al. 2014] prove that for convex, Lipschitz objectives and $\alpha(t) = \mathcal{O}(1/\sqrt{T}) \Rightarrow \|f(\widetilde{\mathbf{z}}_i(T)) - f^\star\| = \mathcal{O}(1/\sqrt{T})$, $\widetilde{\mathbf{z}}_i(T) = \frac{1}{T}\sum_{t=1}^{T} \mathbf{z}_i(t)$

Introduction: Push-sum Algorithm

- Decentralized optimization over *directed* graphs with *exact* communication:

$$\begin{cases} \mathbf{x}_i(t+1) & = \sum_{j=1}^n w_{ij}\,\mathbf{x}_j(t) - \alpha(t)\nabla f_i\left(\mathbf{z}_i(t)\right) \\ y_i(t+1) & = \sum_{j=1}^n w_{ij}\,y_j(t) \\ \mathbf{z}_i(t+1) & = \mathbf{x}_i(t+1)/y_i(t+1) \end{cases}$$

- [Nedic et al. 2014] prove that for convex, Lipschitz objectives and $\alpha(t) = \mathcal{O}(1/\sqrt{T}) \Rightarrow \|f(\widetilde{\mathbf{z}}_i(T)) - f^\star\| = \mathcal{O}(1/\sqrt{T})$, $\widetilde{\mathbf{z}}_i(T) = \frac{1}{T}\sum_{t=1}^T \mathbf{z}_i(t)$

- How can we incorporate quantized message exchanging for this setting?

Proposed Algorithm: Quantized Push-sum

- We propose the quantized Push-sum algorithm for *stochastic optimization*

$\mathbf{q}_i(t) = Q\left(\mathbf{x}_i(t) - \widehat{\mathbf{x}}_i(t)\right)$
for all nodes $k \in \mathcal{N}_i^{out}$ and $j \in \mathcal{N}_i^{in}$ **do**
  send $\mathbf{q}_i(t)$ and $y_i(t)$ to k and receive $\mathbf{q}_j(t)$ and $y_j(t)$ from j.
  $\widehat{\mathbf{x}}_j(t+1) = \widehat{\mathbf{x}}_j(t) + \mathbf{q}_j(t)$
end for
$\mathbf{v}_i(t+1) = \mathbf{x}_i(t) - \widehat{\mathbf{x}}_i(t+1) + \sum_{j \in \mathcal{N}_i^{in}} w_{ij}\widehat{\mathbf{x}}_j(t+1)$
$y_i(t+1) = \sum_{j \in \mathcal{N}_i^{in}} w_{ij} y_j(t)$
$\mathbf{z}_i(t+1) = \mathbf{v}_i(t+1)/y_i(t+1)$
$\mathbf{x}_i(t+1) = \mathbf{v}_i(t+1) - \alpha(t+1)\nabla F_i(\mathbf{z}_i(t+1))$

Proposed Algorithm: Quantized Push-sum

- We propose the quantized Push-sum algorithm for *stochastic optimization*

$\mathbf{q}_i(t) = Q\left(\mathbf{x}_i(t) - \widehat{\mathbf{x}}_i(t)\right)$
for all nodes $k \in \mathcal{N}_i^{out}$ and $j \in \mathcal{N}_i^{in}$ **do**
  send $\mathbf{q}_i(t)$ and $y_i(t)$ to k and receive $\mathbf{q}_j(t)$ and $y_j(t)$ from j.
  $\widehat{\mathbf{x}}_j(t+1) = \widehat{\mathbf{x}}_j(t) + \mathbf{q}_j(t)$
end for
$\mathbf{v}_i(t+1) = \mathbf{x}_i(t) - \widehat{\mathbf{x}}_i(t+1) + \sum_{j \in \mathcal{N}_i^{in}} w_{ij} \widehat{\mathbf{x}}_j(t+1)$
$y_i(t+1) = \sum_{j \in \mathcal{N}_i^{in}} w_{ij} y_j(t)$
$\mathbf{z}_i(t+1) = \mathbf{v}_i(t+1) / y_i(t+1)$
$\mathbf{x}_i(t+1) = \mathbf{v}_i(t+1) - \alpha(t+1) \nabla F_i(\mathbf{z}_i(t+1))$

- $\widehat{\mathbf{x}}_j(t)$ is stored in all out-neighbors of node j

Proposed Algorithm: Quantized Push-sum

- We propose the quantized Push-sum algorithm for *stochastic optimization*

$\mathbf{q}_i(t) = Q\left(\mathbf{x}_i(t) - \widehat{\mathbf{x}}_i(t)\right)$
for all nodes $k \in \mathcal{N}_i^{out}$ and $j \in \mathcal{N}_i^{in}$ **do**
  send $\mathbf{q}_i(t)$ and $y_i(t)$ to k and receive $\mathbf{q}_j(t)$ and $y_j(t)$ from j.
  $\widehat{\mathbf{x}}_j(t+1) = \widehat{\mathbf{x}}_j(t) + \mathbf{q}_j(t)$
end for
$\mathbf{v}_i(t+1) = \mathbf{x}_i(t) - \widehat{\mathbf{x}}_i(t+1) + \sum_{j\in\mathcal{N}_i^{in}} w_{ij}\widehat{\mathbf{x}}_j(t+1)$
$y_i(t+1) = \sum_{j\in\mathcal{N}_i^{in}} w_{ij} y_j(t)$
$\mathbf{z}_i(t+1) = \mathbf{v}_i(t+1)/y_i(t+1)$
$\mathbf{x}_i(t+1) = \mathbf{v}_i(t+1) - \alpha(t+1)\nabla F_i(\mathbf{z}_i(t+1))$

- $\widehat{\mathbf{x}}_j(t)$ is stored in all out-neighbors of node j
- $\widehat{\mathbf{x}}_j(t) \to \mathbf{x}_j(t)$ therefore $\mathbf{q}_j(t) \to \mathbf{0}$ (Similar to [Koloskova et al. 2018])

Assumptions

Assumptions on graph and connectivity

Assumptions

Assumptions on graph and connectivity

- Strongly connected graph and $W_{ij} \geq 0$, $W_{ii} > 0$, $\forall i, j \in [n]$

Assumptions

Assumptions on graph and connectivity

- Strongly connected graph and $W_{ij} \geq 0$, $W_{ii} > 0$, $\forall i, j \in [n]$

Note that this results in $\|W^t - \phi \mathbf{1}'\| \leq C\lambda^t$, $\forall t \geq 1$
where $\phi \in \mathbb{R}^n$, $0 < \lambda < 1$

Assumptions

Assumptions on graph and connectivity

- Strongly connected graph and $W_{ij} \geq 0$, $W_{ii} > 0$, $\forall i, j \in [n]$

Note that this results in $\|W^t - \phi \mathbf{1}'\| \leq C\lambda^t$, $\forall t \geq 1$
where $\phi \in \mathbb{R}^n$, $0 < \lambda < 1$

Assumptions on local objectives

- Lipschitz Local Gradients,
  $\Big\|\nabla f_i(\mathbf{y}) - \nabla f_i(\mathbf{x})\Big\| \leq L\Big\|\mathbf{y} - \mathbf{x}\Big\|$, $\forall \mathbf{x}, \mathbf{y} \in \mathbb{R}^d$

Assumptions

Assumptions on graph and connectivity

- Strongly connected graph and $W_{ij} \geq 0$, $W_{ii} > 0$, $\forall i, j \in [n]$

Note that this results in $\|W^t - \phi \mathbf{1}'\| \leq C\lambda^t$, $\forall t \geq 1$
where $\phi \in \mathbb{R}^n$, $0 < \lambda < 1$

Assumptions on local objectives

- Lipschitz Local Gradients,
  $\Big\|\nabla f_i(\mathbf{y}) - \nabla f_i(\mathbf{x})\Big\| \leq L\Big\|\mathbf{y} - \mathbf{x}\Big\|$, $\forall \mathbf{x}, \mathbf{y} \in \mathbb{R}^d$

- Bounded Stochastic Gradients,
  $\mathbb{E}_{\zeta_i \sim \mathcal{D}_i}\Big\|\nabla F_i(\mathbf{x}, \zeta_i)\Big\|^2 \leq D^2$, $\forall \mathbf{x} \in \mathbb{R}^d$

Assumptions

Assumptions on graph and connectivity

- Strongly connected graph and $W_{ij} \geq 0,\ W_{ii} > 0,\ \forall i, j \in [n]$

Note that this results in $\|W^t - \phi \mathbf{1}'\| \leq C\lambda^t$, $\forall t \geq 1$
where $\phi \in \mathbb{R}^n$, $0 < \lambda < 1$

Assumptions on local objectives

- Lipschitz Local Gradients,
  $\Big\|\nabla f_i(\mathbf{y}) - \nabla f_i(\mathbf{x})\Big\| \leq L\Big\|\mathbf{y} - \mathbf{x}\Big\|,\ \forall \mathbf{x}, \mathbf{y} \in \mathbb{R}^d$

- Bounded Stochastic Gradients,
  $\mathbb{E}_{\zeta_i \sim \mathcal{D}_i}\Big\|\nabla F_i(\mathbf{x}, \zeta_i)\Big\|^2 \leq D^2,\ \forall \mathbf{x} \in \mathbb{R}^d$

- Bounded Variance,
  $\mathbb{E}_{\zeta_i \sim \mathcal{D}_i}\Big\|\nabla F_i(\mathbf{x}, \zeta_i) - \nabla f_i(\mathbf{x})\Big\|^2 \leq \sigma^2,\ \forall \mathbf{x} \in \mathbb{R}^d$

Assumptions

Assumption on quantization function

The quantization function $Q : \mathbb{R}^d \to \mathbb{R}^d$ satisfies for all $\boldsymbol{x} \in \mathbb{R}^d$,

$$\mathbb{E}_Q \left[\left\| Q(\boldsymbol{x}) - \boldsymbol{x} \right\|^2 \right] \leq \omega^2 \left\| \boldsymbol{x} \right\|^2 , \tag{1}$$

where $0 \leq \omega < 1$.

Convergence Results (Convex objectives)

- Define $\gamma := \|W - \mathbb{I}\|_2$ and $C(\lambda, \gamma) := \frac{1}{\sqrt{6(1+\frac{6C^2}{(1-\lambda)^2})(1+\gamma^2)}}$

Theorem 1

Assume local objectives f_i are convex for all $i \in [n]$. By choosing $\omega \leq C(\lambda, \gamma)$ and $\alpha = \frac{\sqrt{n}}{8L\sqrt{T}}$, for all $T \geq 1$, it holds that,

$$\mathbb{E} f\left(\frac{1}{T}\sum_{t=1}^{T} \mathbf{z}_i(t+1)\right) - f^\star = \mathcal{O}\left(\frac{1}{\sqrt{nT}}\right)$$

Convergence Results (Convex objectives)

- Define $\gamma := \|W - \mathbb{I}\|_2$ and $C(\lambda, \gamma) := \frac{1}{\sqrt{6(1+\frac{6C^2}{(1-\lambda)^2})(1+\gamma^2)}}$

Theorem 1

Assume local objectives f_i are convex for all $i \in [n]$. By choosing $\omega \leq C(\lambda, \gamma)$ and $\alpha = \frac{\sqrt{n}}{8L\sqrt{T}}$, for all $T \geq 1$, it holds that,

$$\mathbb{E} f\left(\frac{1}{T}\sum_{t=1}^{T} \mathbf{z}_i(t+1)\right) - f^\star = \mathcal{O}\left(\frac{1}{\sqrt{nT}}\right)$$

- Time average of local parameters $\mathbf{z}_i$ converges to the exact solution!

Convergence Results (Convex objectives)

- Define $\gamma := \|W - \mathbb{I}\|_2$ and $C(\lambda, \gamma) := \frac{1}{\sqrt{6(1+\frac{6C^2}{(1-\lambda)^2})(1+\gamma^2)}}$

Theorem 1

Assume local objectives f_i are convex for all $i \in [n]$. By choosing $\omega \leq C(\lambda, \gamma)$ and $\alpha = \frac{\sqrt{n}}{8L\sqrt{T}}$, for all $T \geq 1$, it holds that,

$$\mathbb{E} f\left(\frac{1}{T}\sum_{t=1}^{T} \mathbf{z}_i(t+1)\right) - f^\star = \mathcal{O}\left(\frac{1}{\sqrt{nT}}\right)$$

- Time average of local parameters $\mathbf{z}_i$ converges to the exact solution!
- The convergence rate is the same as the case of undirected graphs with exact communication (e.g. [Yuan et al. 2016])

Convergence Results (Convex objectives)

- Define $\gamma := \|W - \mathbb{I}\|_2$ and $C(\lambda, \gamma) := \frac{1}{\sqrt{6(1+\frac{6C^2}{(1-\lambda)^2})(1+\gamma^2)}}$

Theorem 1

Assume local objectives f_i are convex for all $i \in [n]$. By choosing $\omega \leq C(\lambda, \gamma)$ and $\alpha = \frac{\sqrt{n}}{8L\sqrt{T}}$, for all $T \geq 1$, it holds that,

$$\mathbb{E} f\left(\frac{1}{T}\sum_{t=1}^{T} \mathbf{z}_i(t+1)\right) - f^\star = \mathcal{O}\left(\frac{1}{\sqrt{nT}}\right)$$

- Time average of local parameters $\mathbf{z}_i$ converges to the exact solution!
- The convergence rate is the same as the case of undirected graphs with exact communication (e.g. [Yuan et al. 2016])
- Error is proportional to $1/\sqrt{n}$

Convergence Results (Non-Convex objectives)

Theorem 2

Let $\omega \leq C(\lambda, \gamma)$ and $\alpha = \frac{\sqrt{n}}{L\sqrt{T}}$. Then after sufficiently large number of iterations, ($T \geq 4n$), it holds that

$$\frac{1}{T}\sum_{t=1}^{T} \mathbb{E}\left\|\nabla f\left(\frac{1}{n}\sum_{i=1}^{n}\mathbf{x}_i(t)\right)\right\|^2 = \mathcal{O}\left(\frac{1}{\sqrt{nT}}\right)$$

Convergence Results (Non-Convex objectives)

Theorem 2

Let $\omega \leq C(\lambda, \gamma)$ and $\alpha = \frac{\sqrt{n}}{L\sqrt{T}}$. Then after sufficiently large number of iterations, ($T \geq 4n$), it holds that

$$\frac{1}{T} \sum_{t=1}^{T} \mathbb{E} \left\| \nabla f \left(\frac{1}{n} \sum_{i=1}^{n} \mathbf{x}_i(t) \right) \right\|^2 = \mathcal{O} \left(\frac{1}{\sqrt{nT}} \right)$$

- Average of local parameters $\mathbf{x}_i(t)$ converges a stationary point!

Convergence Results (Non-Convex objectives)

Theorem 2

Let $\omega \leq C(\lambda, \gamma)$ and $\alpha = \frac{\sqrt{n}}{L\sqrt{T}}$. Then after sufficiently large number of iterations, $(T \geq 4n)$, it holds that

$$\frac{1}{T}\sum_{t=1}^{T} \mathbb{E}\left\|\nabla f\left(\frac{1}{n}\sum_{i=1}^{n} \mathbf{x}_i(t)\right)\right\|^2 = \mathcal{O}\left(\frac{1}{\sqrt{nT}}\right)$$

- Average of local parameters $\mathbf{x}_i(t)$ converges a stationary point!
- Again, the convergence rate is the same as the case of undirected graphs with exact communication(e.g. [Lian et al. 2017])

Convergence Results (Non-Convex objectives)

Theorem 2

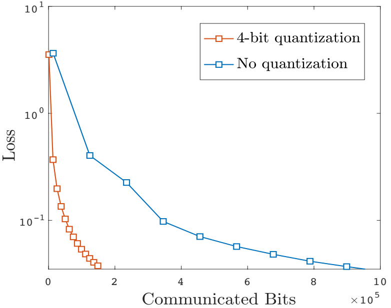
Let $\omega \leq C(\lambda, \gamma)$ and $\alpha = \frac{\sqrt{n}}{L\sqrt{T}}$. Then after sufficiently large number of iterations, ($T \geq 4n$), it holds that

$$\frac{1}{T}\sum_{t=1}^{T} \mathbb{E}\left\|\nabla f\left(\frac{1}{n}\sum_{i=1}^{n}\mathbf{x}_i(t)\right)\right\|^2 = \mathcal{O}\left(\frac{1}{\sqrt{nT}}\right)$$

- Average of local parameters $\mathbf{x}_i(t)$ converges a stationary point!
- Again, the convergence rate is the same as the case of undirected graphs with exact communication(e.g. [Lian et al. 2017])
- Error is proportional to $1/\sqrt{n}$

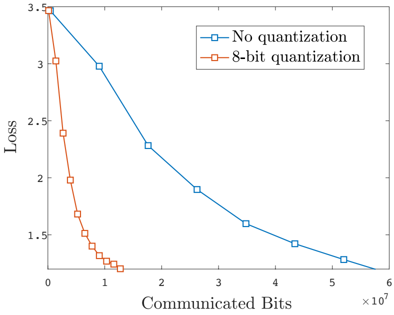
Numerical Experiments

- $f(\mathbf{x}) = \frac{1}{2nm} \sum_{i=1}^{n} \sum_{j=1}^{m} \left\| \mathbf{x} - \zeta_j^i \right\|^2,$
- Data-set size=100, mini-batch size =1, dimension= 256, n=10.

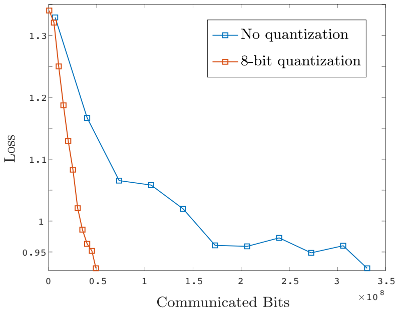


- 5x speedup in communication time.

- Neural network with one hidden layer with 10 hidden units
- Mini-batch size = 10 (Left) & 100 (Right), $n = 10$



(a) MNIST dataset



(b) CIFAR-10 dataset

- 5x speed up in communication time.

Conclusion

- We proposed the quantized push-sum algorithm for collaborative optimization.
- The proposed algorithm converges with optimal convergence rates w.r.t. vanilla push-sum protocol.

Conclusion

- We proposed the quantized push-sum algorithm for collaborative optimization.
- The proposed algorithm converges with optimal convergence rates w.r.t. vanilla push-sum protocol.
- Interesting future directions: Communication-efficient algorithms for collaborative optimization with "asynchrony" or "periodic averaging".