Can Decentralized Algorithms Outperform Centralized Algorithms? A Case Study for Decentralized Parallel Stochastic Gradient Descent

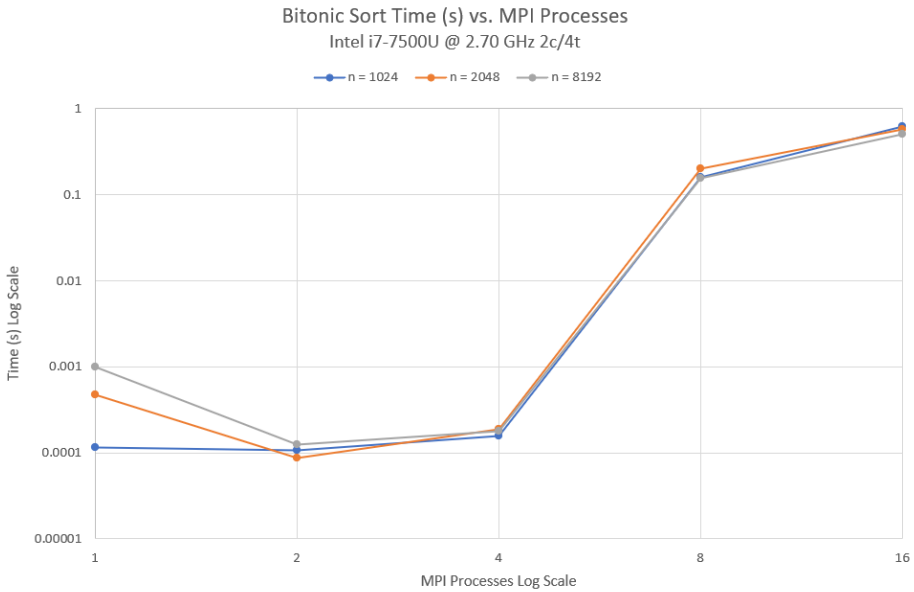
Xiangru Lian, Ce Zhang, Huan Zhang, Cho-Jui Hsieh, Wei Zhang, and Ji Liu

Po-Hao Chen

May 3, 2022

Motivation

- ML workload becoming increasingly large.
- Mainstream distributed training frameworks (i.e, Tensorflow, Torch, and CNTK) are built in a centralized fashion.
- Communication overheads



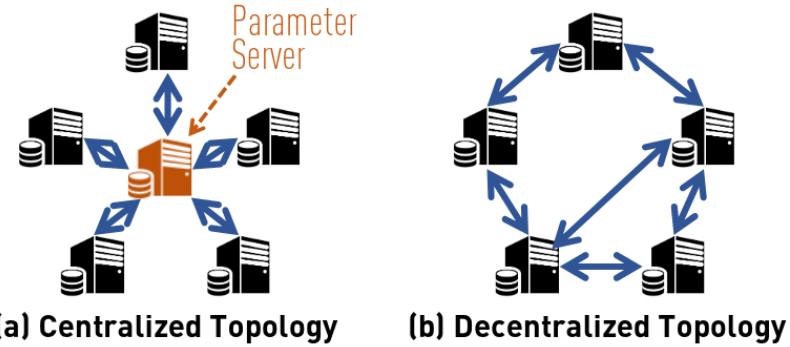
Stochastic Optimization Problem

$$\min_{x \in R^n} f(x) := \mathop{\mathbb{E}}_{\xi \sim D} [F(x; \xi)]$$

- R^n feasible set
- Find x to minimize cost function F.
- Denote ξ as a random variable drawn from arbitrary distribution D
- ξ available after x is chosen, we optimize expectation $\mathbb{E}_{\xi \sim D} [F(x; \xi)]$ instead of F directly.

Parallel Stochastic Gradient Descent (PSGD)

- Large-scale dataset
- PSGD aggregates stochastic gradient computed by all nodes.
- Bottleneck in centralized topology



Contribution

Open Question: **Could decentralized methods have advantages over centralized algorithms?**

- Prior work implicitly suggests computation and communication complexity is much worse for decentralized methods.
- Lian et al. [2]: *decentralized PSGD*
  - can be faster than centralized counterpart
  - scales linearily w.r.t computational complexity.
  - empirical analysis across frameworks

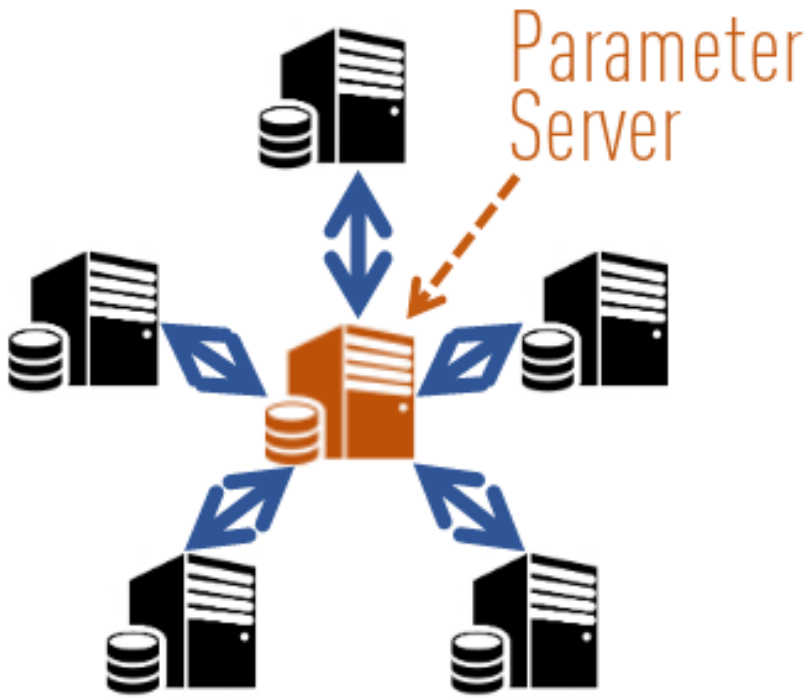
Algorithm	Comm. Complexity	Comp. Complexity
C-PSGD	$O(n)$	$O(\frac{n}{\epsilon} + \frac{1}{\epsilon^2})$
D-PSGD	O (deg(network))	$O(\frac{n}{\epsilon} + \frac{1}{\epsilon^2})$

n := number of nodes ϵ-approximation solution

Related Work

Assume K, n is the number of iterations and nodes, respectively.

- **Stochastic Gradient Descent**
  Convergence Rate:
  - Convex: $O(1/\sqrt{K})$
  - Strongly Convex: $O(1/K)$
  - Non-Convex (ergodic): $O(1/\sqrt{K})$
- **Centralized PSGD**
  - Implementation based on parameter server topology.
  - Agarwal and Duchi [1] proved C-PSGD converges in $O(1/\sqrt{Kn})$
  - Linear speedup over SGD



(a) Centralized Topology

Related Work

"Computational Complexity":

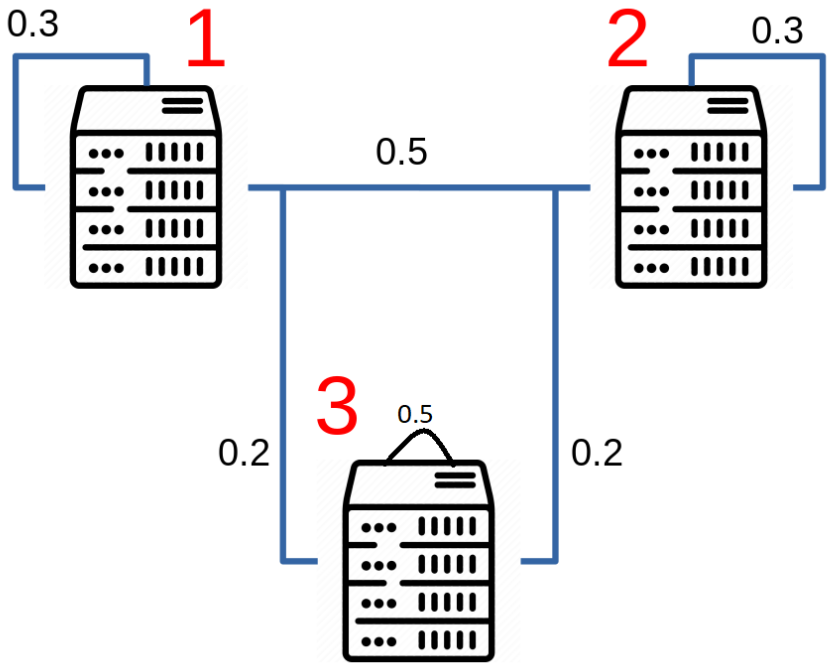
- Decentralized Parallel Stochastic Algorithms
  - Convex and Strongly Convex $O(n/\epsilon^2)$, $O(n/\epsilon)$
  - HogWild++ shown superior to C-PSGD.
  - No clear **convergence** rate analysis existed
- Other Decentralized Algorithms
  - consensus, privacy, network
  - No clear advantage over centralized counterpart.

(b) Decentralized Topology

Decentralized PSGD (D-PSGD)

- Setup: consider a computing cluster with n nodes.
- We represent our topology as an undirected graph $G(V, W)$
- $V = \{1, 2, ..., n\}$
- $W = \begin{bmatrix} 0.3 & 0.5 & 02 \\ 0.5 & 0.3 & 0.2 \\ 0.3 & 0.2 & 0.5 \end{bmatrix}$
- W is a doubly stochastic symmetric matrix
- V is a vertex set

Example:



Decentralized PSGD (D-PSGD)

Original Problem:

$$\min_{x \in R^n} f(x) := \mathop{\mathbb{E}}_{\xi \sim D} [F(x; \xi)]$$

Recall: F is cost function, ξ drawn from distribution D of entire dataset.

- We want to distribute workload to each node $i \subseteq [n]$

$$\min_{x \in R^n} f(x) := \frac{1}{n} \sum_{i=1}^{n} \mathop{\mathbb{E}}_{\xi \sim D_i} [F_i(x; \xi)]$$

Both approaches are sufficient for $F_i = F$:

- Shared database: $D_i = D$
- Partition into n local storage, define D_i to have same distribution as D over local data.

Decentralized PSGD (D-PSGD)

The following algorithm is synchronous, where each node i executes the program concurrently.

- Input: learning rate γ, stochastic matrix W, number of iterations K
- Initialize $x_{0,i} = x_0$
- **For** $k = 0, 1, 2, ..., K-1$:

Decentralized PSGD (D-PSGD)

The following algorithm is synchronous, where each node i executes the program concurrently.

- Input: learning rate γ, stochastic matrix W, number of iterations K
- Initialize $x_{0,i} = x_0$
- **For** $k = 0, 1, 2, ..., K-1$:
  - Randomly sample $\xi_{k,i}$ from local data

Decentralized PSGD (D-PSGD)

The following algorithm is synchronous, where each node i executes the program concurrently.

- Input: learning rate γ, stochastic matrix W, number of iterations K
- Initialize $x_{0,i} = x_0$
- **For** $k = 0, 1, 2, ..., K-1$:
  - Randomly sample $\xi_{k,i}$ from local data
  - Compute stochastic gradient $\nabla F_i(x_{k,i}; \xi_{k,i}), \forall i \leq n$

we're using local variable $x_{k,i}$ and exchange it with the neighbors once synchronization is met.

Decentralized PSGD (D-PSGD)

The following algorithm is synchronous, where each node i executes the program concurrently.

- Input: learning rate γ, stochastic matrix W, number of iterations K
- Initialize $x_{0,i} = x_0$
- **For** $k = 0, 1, 2, ..., K-1$:
  - Randomly sample $\xi_{k,i}$ from local data
  - Compute stochastic gradient $\nabla F_i(x_{k,i}; \xi_{k,i}), \forall i \leq n$
  - Aggregate neighborhood and compute weighted average $x_{k+\frac{1}{2},i} = \sum_{j=1}^{n} W_{ij} x_{k,j}$

We use $x_{k+\frac{1}{2},i}$ as intermediate variable

The previous two steps can be run in parallel

Decentralized PSGD (D-PSGD)

The following algorithm is synchronous, where each node i executes the program concurrently.

- Input: learning rate γ, stochastic matrix W, number of iterations K
- Initialize $x_{0,i} = x_0$
- **For** $k = 0, 1, 2, ..., K-1$:
  - Randomly sample $\xi_{k,i}$ from local data
  - Compute stochastic gradient $\nabla F_i(x_{k,i}; \xi_{k,i}), \forall i \leq n$
  - Aggregate neighborhood and compute weighted average $x_{k+\frac{1}{2},i} = \sum_{j=1}^{n} W_{ij} x_{k,j}$
  - Update $x_{k+1,i} = X_{k+\frac{1}{2},i} - \gamma \nabla F_i(x_{k,i;\xi_{k,i}})$
- Output: $\frac{1}{n} \sum_{i=1}^{n} x_{K,i}$

Convergence Rate Analysis

$$\min_{x \in R^n} f(x) := \frac{1}{n} \sum_{i=1}^{n} \mathop{\mathbb{E}}_{\xi \sim D_i} [F_i(x; \xi)]$$

Define

$$\partial f(X_k) := [\nabla f_1(x_{k,1}), \nabla f_2(x_{k,2}), ..., \nabla f_n(x_{k,n})] \in R^{Nxn}$$

where $f_i(x) := \mathbb{E}_{\xi \sim D_i} [F_i(x; \xi)]$, $X_k := [x_{k,1}, ..., x_{k,n}]$.
$x_{k,i}$:= local variable on node i at iterate k

common assumptions:

- f_i is L-Lipschitz $||\nabla f_i(x) - \nabla f_i(y)|| \leq L||x - y||, \forall x, y \in R^n$
- spectral gap of W, $\rho := (max\{|\lambda_2(W)|, |\lambda_n(W)|\})^2 < 1$
- bounded variance: can we get the "bias" to decay to small value with more iterations?
- Start from 0: $X_0 = 0$ simplifies proof W.L.O.G

Convergence Rate Analysis

We define constant σ, ς
By our assumption, variance of stochastic gradient
$\mathbb{E}_{i\sim\mathcal{U}([n])}\left[\mathbb{E}_{\xi\sim D_I}\left[||\nabla F_i(x;\xi)-\nabla f(x)||^2\right]\right]$ is bounded.
$\Rightarrow \mathbb{E}_{\xi\sim D_i}\left[||\nabla F_i(x;\xi)-\nabla f_i(x)||^2\right]\leq\sigma^2, \forall i, \forall x$
and $\mathbb{E}_{i\sim\mathcal{U}([n])}\left[||\nabla f_i(x)-\nabla f(x)||^2\right]\leq\varsigma^2, \forall x$

Interpretation: variance of distribution in node partitions is bounded by σ^2 and variance of output distribution is bounded by ς^2.

Convergence Rate Analysis

$$C_1 := (\frac{1}{2} - \frac{9\gamma^2 L^2 n}{(1-\sqrt{\rho})^2 C_2}) \qquad C_2 := (1 - \frac{18\gamma^2}{(1-\sqrt{\rho})^2} nL^2)$$

Theorem

$$\frac{1}{K}(\frac{1-\gamma L}{2}\sum_{k=0}^{K-1}\mathbb{E}\left[||\frac{\partial f(X_k)1_n}{n}||^2\right] + C_1\sum_{k=0}^{K-1}\mathbb{E}\left[||\nabla f(\frac{X_k 1_n}{n})||^2\right])$$
$$\leq \frac{f(0)-f^*}{\gamma K} + \frac{\gamma L}{2n}\sigma^2 + \frac{\gamma^2 L^2 n\sigma^2}{(1-\rho)C_2} + \frac{9\gamma^2 L^2 n\varsigma^2}{(1-\sqrt{\rho})^2 C_2}$$

Characterizes the convergence of of the average of all local variable $X_{k,i}$.

Convergence Rate Analysis

Lemma

$||\frac{1_n}{n} - W^k e_i||^2 \leq \rho^k, \forall i \in \{1, 2, ..., n\}, k \in \mathcal{N}$ **(Spectral Gap Bound)**

Proof: $||\frac{1_n}{n} - W^k e_i||^2 = ||(W^\infty - W^k)e_i||^2 \leq ||W^\infty - W^k||^2 ||e_i||^2 \leq \rho^k$

Convergence Rate Analysis

Lemma

$$\mathbb{E}\left[||\partial f(X_j)||^2\right] \leq \sum_{h=1}^{n} 3\,\mathbb{E}\left[L^2||\frac{\sum_{i'=1}^{n} x_{j,i'}}{n} - x_{j,h}||^2\right] + 3n\varsigma^2 + 3\,\mathbb{E}\left[||\nabla f(\frac{X_j 1_n}{n})1_n^T||^2\right]$$

Proof: $\mathbb{E}\left[||\partial f(X_j)||^2\right] \leq 3\,\mathbb{E}\left[||\partial f(X_j) - \partial f(\frac{X_j 1_n}{n}1_n^T)||^2\right] +$

(Triangle Inequality $\rightarrow$) $3\mathbb{E}\left[\partial f(\frac{X_j 1_n}{n}1_n^T) - \nabla f(\frac{X_j 1_n}{n})1_n^T||^2\right] +$

(Bounded Variance $\downarrow$) $3\mathbb{E}\left[||\nabla f(\frac{X_j 1_n}{n})1_n^T||^2\right] \leq$

$3\,\mathbb{E}\left[||\partial f(X_j) - \partial f(\frac{X_j 1_n}{n}1_n^T)||_F^2\right] + 3n\varsigma^2 + 3\,\mathbb{E}\left[||\nabla f(\frac{X_j 1_n}{n})1_n^T||^2\right]$

$\leq \sum_{h=1}^{n} 3\,\mathbb{E}\left[L^2||\frac{\sum_{i'=1}^{n} x_{j,i'}}{n} - x_{j,h}||^2\right] + 3n\varsigma^2 + 3\,\mathbb{E}\left[||\nabla f(\frac{X_j 1_n}{n})1_n^T||^2\right]$

Proof of Convergence Theorem of D-PSGD

begin with $f(\frac{X_{k+1}1_n}{n})$:

$$\mathbb{E}\left[f(\frac{X_{k+1}1_n}{n})\right] =$$

$$\mathbb{E}\left[f(\frac{X_k W 1_n}{n} - \gamma\frac{\partial F(X_k;\xi_k)1_n}{n})\right] \quad \textbf{(last step of D-PSGD)}$$

$$\overset{\rho}{=}{}^{\leq 1} \mathbb{E}\left[f(\frac{X_k 1_n}{n}) - \gamma\frac{\partial F(X_k;\xi_k)1_n}{n})\right] \leq \quad \textbf{(Descent Lemma)}$$

$$\mathbb{E}\left[f(\frac{X_k 1_n}{n})\right] - \gamma\,\mathbb{E}\left[\langle\nabla f(\frac{X_k 1_n}{n}), \frac{\partial f(X_k)1_n}{n}\rangle\right] + \frac{\gamma^2 L}{2}\,\mathbb{E}\left[||\frac{\sum_{i=1}^n \nabla F_i(x_{k,i};\xi_{k,i})}{n}||^2\right]$$

$$= \mathbb{E}\left[f(\frac{X_k 1_n}{n})\right] - \gamma\,\mathbb{E}\left[\langle\nabla f(\frac{X_k 1_n}{n}), \frac{\partial f(X_k)1_n}{n}\rangle\right] +$$

$$\frac{\gamma^2 L}{2}(\mathbb{E}\left[||\frac{\sum_{i=1}^n \nabla F_i(x_{k,i};\xi_{k,i}) - \sum_{i=1}^n \nabla f_i(x_{k,i})}{n}||^2\right] + \mathbb{E}\left[||\frac{\sum_{i=1}^n \nabla f_i(x_{k,i})}{n}||^2\right])$$

Proof of Converge Theorem of D-PSGD (Cont.)

So far we have: $\mathbb{E}\left[f(\frac{X_{k+1}1_n}{n})\right] = \mathbb{E}\left[f(\frac{X_k 1_n}{n})\right] - \gamma\,\mathbb{E}\left[\langle \nabla f(\frac{X_k 1_n}{n}), \frac{\partial f(X_k)1_n}{n}\rangle\right] + \underbrace{\frac{\gamma^2 L}{2}\,\mathbb{E}\left[||\frac{\sum_{i=1}^n \nabla F_i(x_{k,i};\xi_{k,i}) - \sum_{i=1}^n \nabla f_i(x_{k,i})}{n}||^2\right]}_{\leq \frac{\gamma^2 L}{2n}\sigma^2} + \frac{\gamma^2 L}{2}\,\mathbb{E}\left[||\frac{\sum_{i=1}^n \nabla f_i(x_{k,i})}{n}||^2\right]$

Note: $2\langle a, b\rangle = ||a||^2 + ||b||^2 - ||a-b||^2$. Let $a = \nabla f(\frac{X_k 1_n}{n})$, $b = \frac{\partial f(X_k)1_n}{n}$

$\leq \mathbb{E}\left[f(\frac{X_k 1_n}{n})\right] - \frac{\gamma - \gamma^2 L}{2}\,\mathbb{E}\left[||\frac{\partial f(X_k)1_n}{n}||^2\right] - \frac{\gamma}{2}\,\mathbb{E}\left[||\nabla f(\frac{X_k 1_n}{n})||^2\right] + \frac{\gamma^2 L}{2}\frac{\sigma^2}{n} + \underbrace{\frac{\gamma}{2}\,\mathbb{E}\left[||\nabla f(\frac{X_k 1_n}{n}) - \frac{\partial f(X_k)1_n}{n}||^2\right]}_{:=T_1}$

Proof of Convergence Theorem of D-PSGD (Cont.)

$$T_1 = \mathbb{E}\left[||\nabla f(\frac{X_k 1_n}{n}) - \frac{\partial f(X_k) 1_n}{n}||^2\right] \leq \frac{L^2}{n}\sum_{i=1}^{n} \underbrace{\mathbb{E}\left[||\frac{\sum_{i'=1}^{n} x_{k,i'}}{n} - x_{k,i}||\right]}_{Q_{k,i}}$$

$Q_{k,i}$ is the squared distance of the local optimization variable on i-th node from average local optimization variable.

$$Q_{k,i} \leq 2\gamma^2 \underbrace{\mathbb{E}\left[||\sum_{j=0}^{k-1}(\partial F(X_j;\xi_j) - \partial f(X_j))(\frac{1_n}{n} - W^{k-j-1}e_i)^2||^2\right]}_{:=T_2} +$$

$$\underbrace{2\gamma^2 \mathbb{E}\left[||\sum_{j=0}^{k-1}\partial f(X_j)(\frac{1_n}{n} - W^{k-j-1}e_i)||^2\right]}_{:=T_3}$$

Proof of Convergence Theorem of D-PSGD (Cont.)

$$
\begin{aligned}
T_2 &= \mathbb{E}\left[|| \sum_{j=0}^{k-1} (\partial F(X_j;\xi_j) - \partial f(X_j))(\tfrac{1_n}{n} - W^{k-j-1} e_i) ||^2 \right] \\
&\leq \sum_{j=0}^{k-1} \mathbb{E}\left[||\partial F(X_j;\xi_j) - \partial f(X_j)||_F^2 \right] || \tfrac{1_n}{n} - W^{k-j-1} e_i ||^2 \\
&\leq n\sigma^2 \sum_{j=0}^{k-1} \rho^{k-j-1} \leq \tfrac{n\sigma^2}{1-\rho}
\end{aligned}
$$

Proof of Convergence Theorem of D-PSGD (Cont.)

$$T_3 = \mathbb{E}\left[||\sum_{j=0}^{k-1} \partial f(X_j)(\tfrac{1_n}{n} - W^{k-j-1}e_i)||^2\right]$$

$$= \underbrace{\sum_{j=0}^{k-1} \mathbb{E}\left[||\partial f(X_j)(\frac{1_n}{n} - W^{k-j-1}e_i)||^2\right]}_{:=T_4}$$

$$+ \underbrace{\sum_{j\neq j'} \mathbb{E}\left[\langle \partial f(X_j)(\frac{1n}{n} - W^{k-j-1}e_i), \partial f(X_j')(\frac{1n}{n} - W^{k-j'-1}e_i)\rangle\right]}_{:=T_5}$$

Proof of Convergence Theorem of D-PSGD (Cont.)

$$T_4 = \sum_{j=0}^{k-1} \mathbb{E}\left[||\partial f(X_j)(\frac{1_n}{n} - W^{k-j-1}e_i)||^2\right]$$
$$\leq \sum_{j=0}^{k-1} \mathbb{E}\left[||\partial f(X_j)||^2\right] ||\frac{1_n}{n} - W^{k-j}e_i||^2$$
Apply Spectral gap bound and Inequality Lemma for $\mathbb{E}\left[||\partial f(X_j)||^2\right]$
$$\leq 3\sum_{j=0}^{k-1}\sum_{j=1}^{n} \mathbb{E}\left[L^2 Q_{j,h}||\frac{1_n}{n} - W^{k-j-1}e_i||^2\right] + 3n\varsigma^2\frac{1}{1-\rho}$$
$$+3\sum_{j=0}^{k-1} \mathbb{E}\left[||\nabla f(\frac{X_j 1_n}{n})1_n^T||^2\right] ||\frac{1_n}{n} - W^{k-j-1}e_i||^2$$

Proof Convergence Theorem of D-PSGD (Cont.)

$$T_5 = \sum_{j\neq j'}^{k-1} \mathbb{E}\left[\langle \partial f(X_j)(\tfrac{1_n}{n} - W^{k-j-1}e_i), \partial f(X'_j)(\tfrac{1_n}{n} - W^{k-j'-1}e_i)\rangle\right]$$

$$\leq \sum_{j\neq j'}^{k-1} \underbrace{\mathbb{E}\left[||\partial f(X_j)(\frac{1_n}{n} - W^{k-j-1}e_i)||\right]}_{\leq \mathbb{E}\left[||\partial f(X_j)||\right]\cdot||\frac{1_n}{n} - W^{k-j-1}e_i||} \underbrace{||\partial f(x'_j)(\frac{1_n}{n} - w^{k-j'-1}e_i)||}_{\leq ||\partial f(x'_j)||\cdot||(\frac{1_n}{n} - w^{k-j'-1}e_i)||}$$

$$\leq \sum_{j\neq j'}^{k-1} \mathbb{E}\left[\frac{||\partial f(X_j)||^2}{2} + \frac{||\partial f(X_{j'})||^2}{2}\right] \rho^{k-\frac{j-j'}{2}-1}$$

$$\underbrace{\leq 3\sum_{j\neq j'}^{k-1}(\sum_{h=1}^{n} \mathbb{E}\left[L^2 Q_{j,h}\right] + \mathbb{E}\left[||\nabla f(\frac{X_j 1_n}{n} 1_n^T)||^2\right])\rho^{k-\frac{j+j'}{2}-1}}_{=:T_6}$$

$$+ \underbrace{\sum_{j\neq j'}^{k-1} 3n\varsigma^2\rho^{k-1-\frac{j+j'}{2}}}_{=:T_7}$$

Proof Convergence Theorem of D-PSGD (Cont.)

Plugging T_6, T_7 into T_5 and T_5, T_4 into T_3. Then, we plug T_2, T_3 into the $Q_{k,i}$ bound to obtain the following:

$$Q_{k,i} \leq \frac{2\gamma^2 n\sigma^2}{1-\rho} + \frac{18\gamma^2 n\varsigma^2}{(1-\sqrt{\rho})^2} + 6\gamma^2 \sum_{j=0}^{k-1} \mathbb{E}\left[||\nabla f(\frac{X_j 1_n}{n})1_n^Y||^2\right](\rho^{k-j-1} + \frac{2\sqrt{\rho^{k-j-1}}}{1-\sqrt{\rho}}) + 6\gamma^2 \sum_{j=0}^{k-1}\sum_{h=1}^{n} \mathbb{E}\left[L^2 Q_{j,h}(\frac{2\sqrt{\rho^{k-j-1}}}{1-\sqrt{\rho}} + \rho^{k-j-1})\right]$$

Define M_k as average of $Q_{k,i}$ on all nodes $:= \frac{\mathbb{E}[\sum Q_{k,i}]}{n}$.

We can bound T_1 with M_k. $\mathbb{E}[T_1] \leq \frac{L^2}{n}\sum_{i=1}^{n} \mathbb{E}[Q_{k,i}] = L^2\,\mathbb{E}[M_k]$

Proof of Convergence Theorem of D-PSGD (Cont.)

Finally, putting it all together. $\mathbb{E}\left[f(\frac{X_{k+1}1_n}{n})\right] \leq \mathbb{E}\left[f(\frac{X_k 1_n}{n})\right] - \frac{\gamma-\gamma^2 L}{2}\mathbb{E}\left[||\frac{\partial f(X_k)1_n}{n}||^2\right] - \frac{\gamma}{2}\mathbb{E}\left[\nabla f(\frac{X_k 1_n}{n})||^2\right] + \frac{\gamma^2 L}{2n}\sigma^2 + \frac{\gamma}{2}L^2\mathbb{E}\left[M_k\right]$

Summing from k=0 to K-1

$\frac{\gamma-\gamma^2 L}{2}\mathbb{E}\left[||\frac{\partial f(X_k)1_n}{n}||^2\right] + \frac{\gamma}{2}\mathbb{E}\left[\nabla f(\frac{X_k 1_n}{n})||^2\right] \leq$

$f(0) - f^* + \frac{\gamma^2 KL}{2n}\sigma^2 + \frac{\gamma}{2}L^2\sum_{k=0}^{K-1}\mathbb{E}\left[M_k\right] \leq f(0) - f^* + \frac{\gamma^2 KL}{2n}\sigma^2$

$$+ \frac{\gamma^3 L^2 n\sigma^2}{(1-\rho)\left(1-\frac{18}{(1-\sqrt{\rho})^2}\gamma^2 nL^2\right)}K + \frac{9\gamma^3 L^2 n\varsigma^2}{(1-\sqrt{\rho})^2\left(1-\frac{18}{(1-\sqrt{\rho})^2}\gamma^2 nL^2\right)}K$$

$$+ \frac{9n\gamma^3 L^2}{(1-\sqrt{\rho})^2\left(1-\frac{18}{(1-\sqrt{\rho})^2}\gamma^2 nL^2\right)}\sum_{k=0}^{K-1}\mathbb{E}\left\|\nabla f\left(\frac{X_k \mathbf{1}_n}{n}\right)\right\|^2$$

Proof of Convergence Theorem

Rearrange the derivation:

$$\frac{\gamma-\gamma^2 L}{2}\sum_{k=0}^{K-1}\mathbb{E}\left[||\frac{\partial f(X_k)1_n}{n}||^2\right]+\frac{\gamma}{2}\sum_{k=0}^{K-1}\mathbb{E}\left[||\gamma f(\frac{X_k 1_n}{n})||^2\right]$$
$$\leq \frac{f(0)-f^*}{\gamma K}+\frac{\gamma L}{2n}\sigma^2+\frac{\gamma^2 L^2 n\sigma^2}{(1-\rho)C_2}+\frac{9\gamma^2 L^2 n\varsigma^2}{(1-\sqrt{\rho})^2 C_2}\ \square$$

Convergence Rate Analysis

We appropriately pick a learning rate to derive the final convergence rate $\gamma = \frac{1}{2L+\sigma\sqrt{K/n}}$

Corollary

$$\frac{\sum k=0^{K-1}\, \mathbb{E}\left[||\nabla f(\frac{X_k 1_n}{n})||^2\right]}{K} \leq \frac{8(f(0)-f^*)L}{K} + \frac{(8f(0)-8f^*+4L)\sigma}{\sqrt{Kn}}$$

if K is sufficiently large, in particular,

$K \geq \frac{4L^4 n^5}{\sigma^6(f(0)-f^*+L)^2}(\frac{\sigma^2}{1-\rho} + \frac{9\varsigma^2}{(1-\sqrt{\rho})^2})^2$, *and* $K \geq \frac{72L^2n^2}{\sigma^2(1-\sqrt{\rho})^2}$

Interpretation: D-PSGD convergence rate is $O(\frac{1}{K} + \frac{1}{\sqrt{nK}})$ given we have enough iterations.

Implication

D-PSGD convergence rate: $O(\frac{1}{K} + \frac{1}{\sqrt{nK}})$

- Linear Speedup:
  - $\frac{1}{\sqrt{nK}}$ dominates $\frac{1}{K}$ for sufficiently large K
  - total comp. complexity $O(1/\epsilon^2)$, single node $O(1/n\epsilon^2)$
- Better than C-PSGD
  - same comp. complexity
  - avoid traffic jam, maximum comm. cost is only $O(deg(G))$ where $O(n)$ for C-PSGD

Note 1: deg of network could be much smaller than $O(n)$, or even $O(1)$ in special case of ring.

Note 2: $K^{-1}(\sum_{k=0}^{k=1} \mathbb{E}\left[||\nabla f(\frac{x_k 1_n}{n})||^2\right]) \leq \epsilon$ means ϵ-approx.

Empirical Data

Validate theory with 112 GPUS (7 GPUs example below).

- Dataset and models:
  - image classification with ResNet on CIFAR-10.
- Setups:
  - CNTK
  - C-SGD
  - D-SGD