On the Linear Speedup Analysis of Communication Efficient Momentum SGD for Distributed Non-Convex Optimization

Poster @ Pacific Ballroom #182

Hao Yu, Rong Jin, Sen Yang
Machine Intelligence Technology
Alibaba Group (US) Inc., Bellevue, WA

Distributed Non-Convex Stochastic Opt

- Consensus non-convex stochastic optimization

$$\min_{x \in \mathscr{R}^m} \frac{1}{N} \sum_{i=1}^{N} \underbrace{\mathbb{E}_{\zeta_i}[F_i(x; \zeta_i)]}_{\triangleq f_i(x)}$$

Distributed Non-Convex Stochastic Opt

- Consensus non-convex stochastic optimization

$$\min_{x \in \mathscr{R}^m} \frac{1}{N} \sum_{i=1}^{N} \underbrace{\mathbb{E}_{\zeta_i}[F_i(x; \zeta_i)]}_{\triangleq f_i(x)}$$

  - N parallel nodes, each has its own non-convex objective

Distributed Non-Convex Stochastic Opt

- Consensus non-convex stochastic optimization

$$\min_{x \in \mathscr{R}^m} \frac{1}{N} \sum_{i=1}^{N} \underbrace{\mathbb{E}_{\zeta_i}[F_i(x; \zeta_i)]}_{\triangleq f_i(x)}$$

  - N parallel nodes, each has its own non-convex objective

  - Need to find a consensus solution in a distributed environment

Distributed Non-Convex Stochastic Opt

- Consensus non-convex stochastic optimization

$$\min_{x \in \mathscr{R}^m} \frac{1}{N} \sum_{i=1}^{N} \underbrace{\mathbb{E}_{\zeta_i}[F_i(x; \zeta_i)]}_{\triangleq f_i(x)}$$

  - N parallel nodes, each has its own non-convex objective
  - Need to find a consensus solution in a distributed environment

- Applications:

Distributed Non-Convex Stochastic Opt

- Consensus non-convex stochastic optimization

$$\min_{x \in \mathscr{R}^m} \frac{1}{N} \sum_{i=1}^{N} \underbrace{\mathbb{E}_{\zeta_i}[F_i(x; \zeta_i)]}_{\triangleq f_i(x)}$$

  - N parallel nodes, each has its own non-convex objective
  - Need to find a consensus solution in a distributed environment

- Applications:
  - Parallel training of deep neural networks with N workers

Distributed Non-Convex Stochastic Opt

- Consensus non-convex stochastic optimization

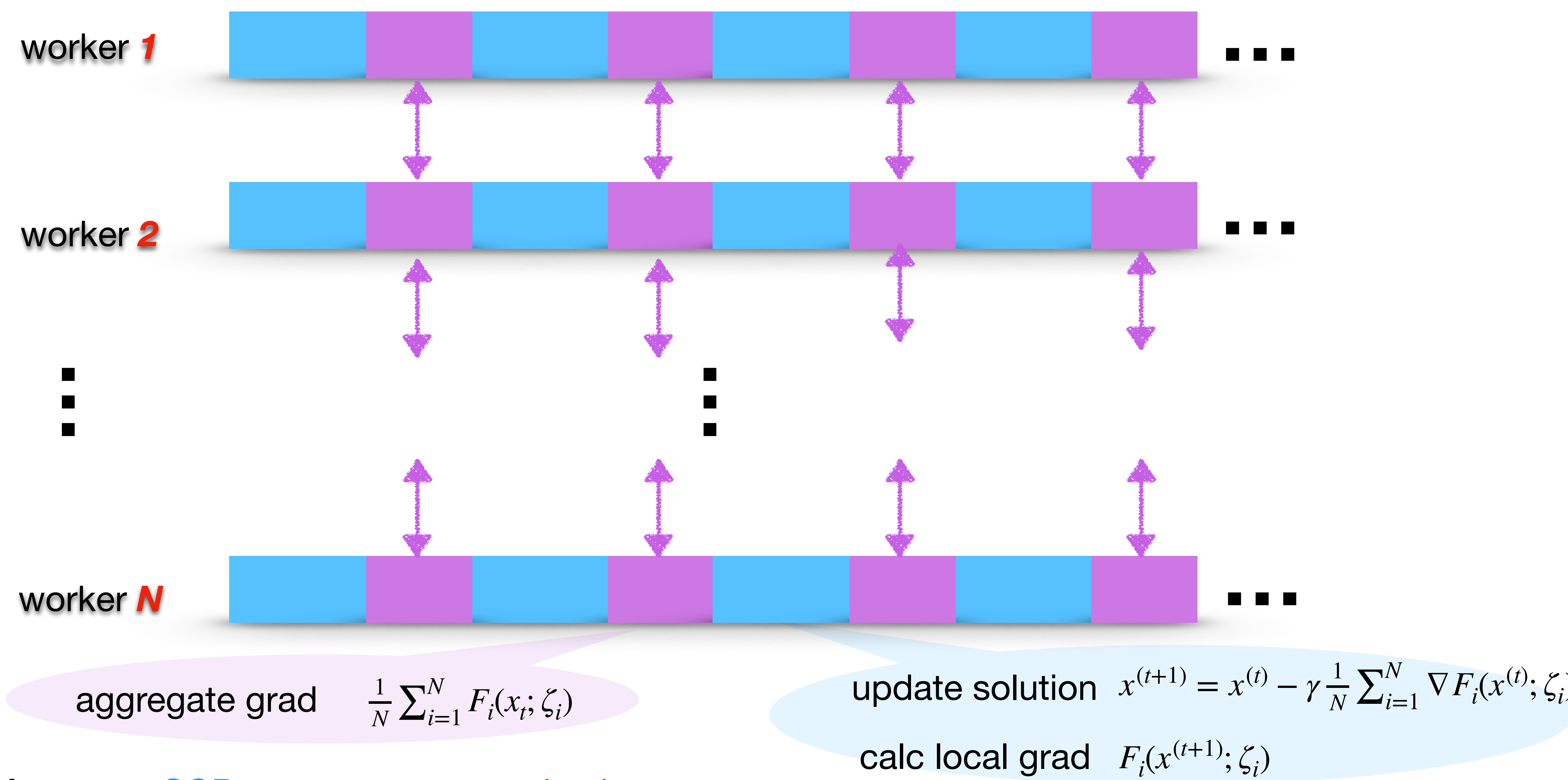
$$\min_{x \in \mathscr{R}^m} \frac{1}{N} \sum_{i=1}^{N} \underbrace{\mathbb{E}_{\zeta_i}[F_i(x; \zeta_i)]}_{\triangleq f_i(x)}$$

  - N parallel nodes, each has its own non-convex objective
  - Need to find a consensus solution in a distributed environment

- Applications:
  - Parallel training of deep neural networks with N workers
  - ***Federated Learning***: learn a common ML model with intermittent communication where each user possesses non-identical private data

Classical Parallel SGD for Non-Convex Opt

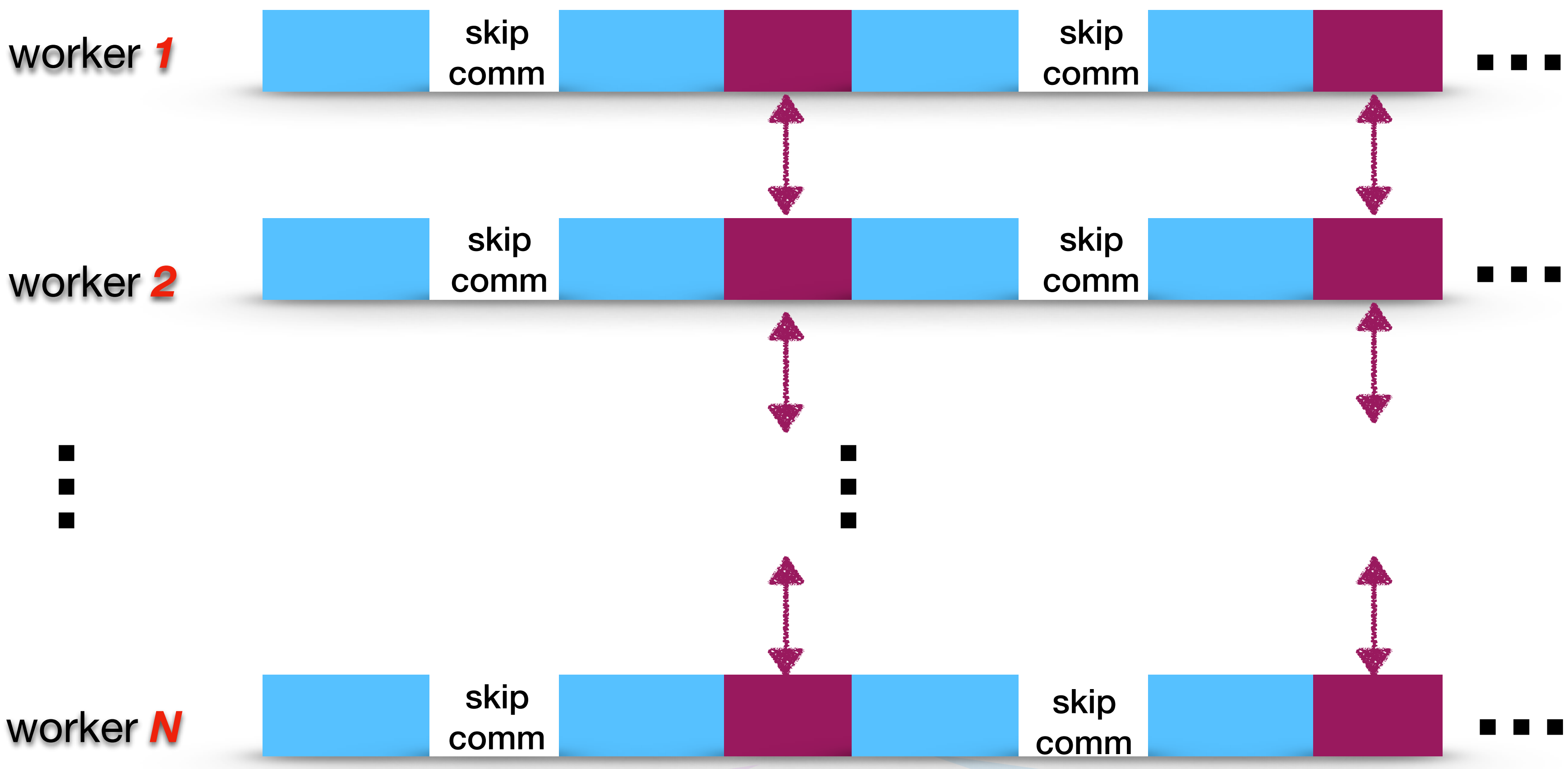
- Classical Parallel mini-batch SGD (PSGD) achieves $O(1/\sqrt{NT})$ convergence (linear speedup) with N workers.



After every **SGD** step, use communication to aggregate gradients

Skip Comm: Parallel Restarted SGD with momentum (ext from [Zhou&Cong'18][Stich'18][Yu et.al.'18][Wang&Joshi'18][Jiang&Agrawal'18])

- Skipping communication rounds so that aggregate models every I (I>1) iterations
- Generalize SGD to momentum SGD (to improve model quality)



aggregate **momentum** $u_i^{(t)} \leftarrow \frac{1}{N}\sum_{i=1}^{N} u_i^{(t)}$

and **model** $x_i^{(t)} \leftarrow \frac{1}{N}\sum_{i=1}^{N} x_i^{(t)}$

update **local momentum** $u_i^{(t+1)} = \beta u_i^{(t)} + \nabla F_i(x_i^{(t)}; \zeta_i)$

and **solution** $x_i^{(t+1)} = x_i^{(t)} - \gamma u_i^{(t)}$

Parallel Restarted SGD with momentum

- Main Result: Converge as fast as PSGD with I times fewer comm rounds

Parallel Restarted SGD with momentum

- Main Result: Converge as fast as PSGD with I times fewer comm rounds
  - If workers access identical training sets, by choosing $\gamma = \frac{\sqrt{N}}{\sqrt{T}}$ and $I = O(\frac{T^{1/2}}{N^{3/2}})$, PR-SGD-Momentum has $O(1/\sqrt{NT})$ convergence
  - If workers use non-identical training sets, by choosing $\gamma = \frac{\sqrt{N}}{\sqrt{T}}$ and $I = O(\frac{T^{1/4}}{N^{3/4}})$, PR-SGD-Momentum has $O(1/\sqrt{NT})$ convergence

Parallel Restarted SGD with momentum

- Main Result: Converge as fast as PSGD with I times fewer comm rounds

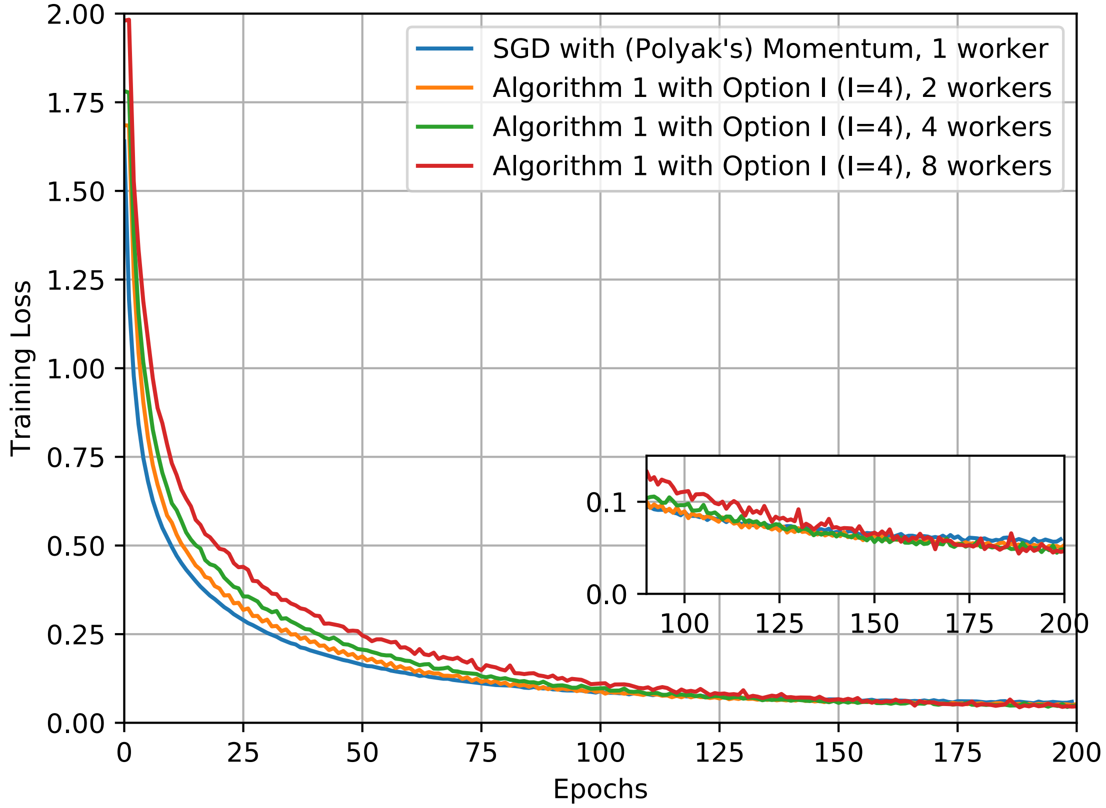
  - If workers access identical training sets, by choosing $\gamma = \frac{\sqrt{N}}{\sqrt{T}}$ and $I = O(\frac{T^{1/2}}{N^{3/2}})$, PR-SGD-Momentum has $O(1/\sqrt{NT})$ convergence

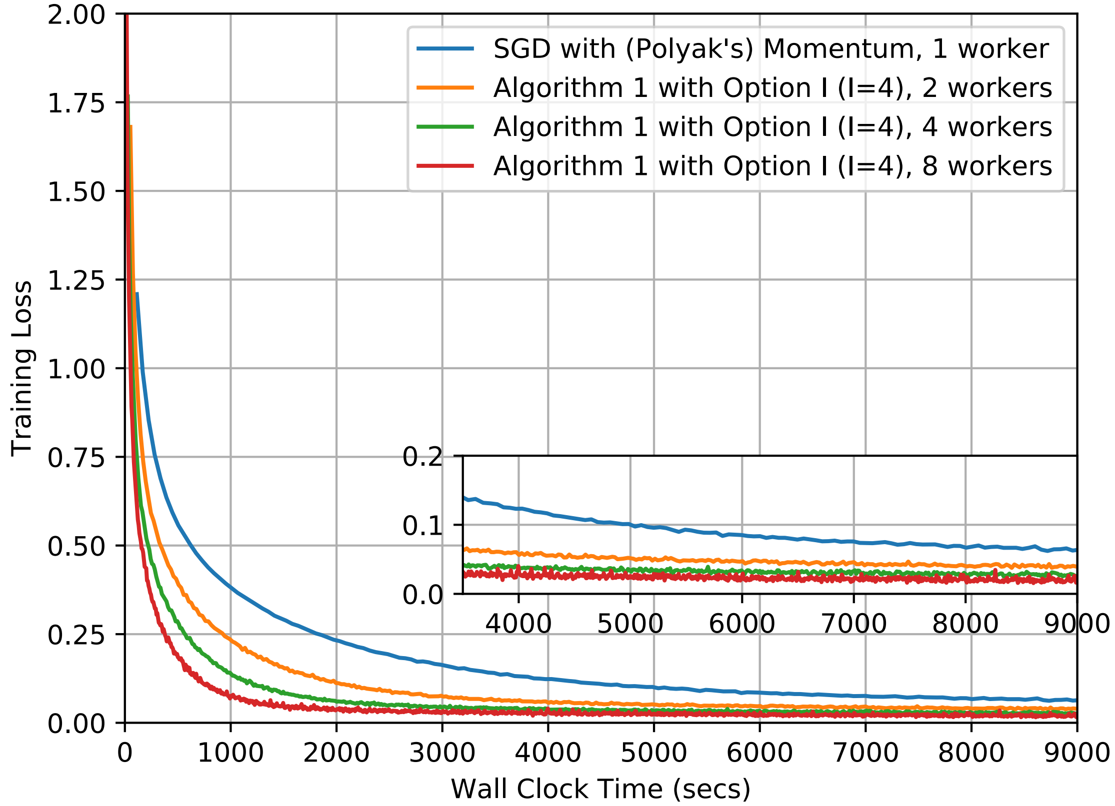
  - If workers use non-identical training sets, by choosing $\gamma = \frac{\sqrt{N}}{\sqrt{T}}$ and $I = O(\frac{T^{1/4}}{N^{3/4}})$, PR-SGD-Momentum has $O(1/\sqrt{NT})$ convergence

- The results with zero momentum (reducing to PR-SGD) improves the analysis in [Yu et.al.'18][Wang&Joshi'18][Jiang&Agrawal'18].

Experiments

Train ResNet56 over Cifar10 with N={2,4,8} workers. I=4; $\gamma = 0.01$

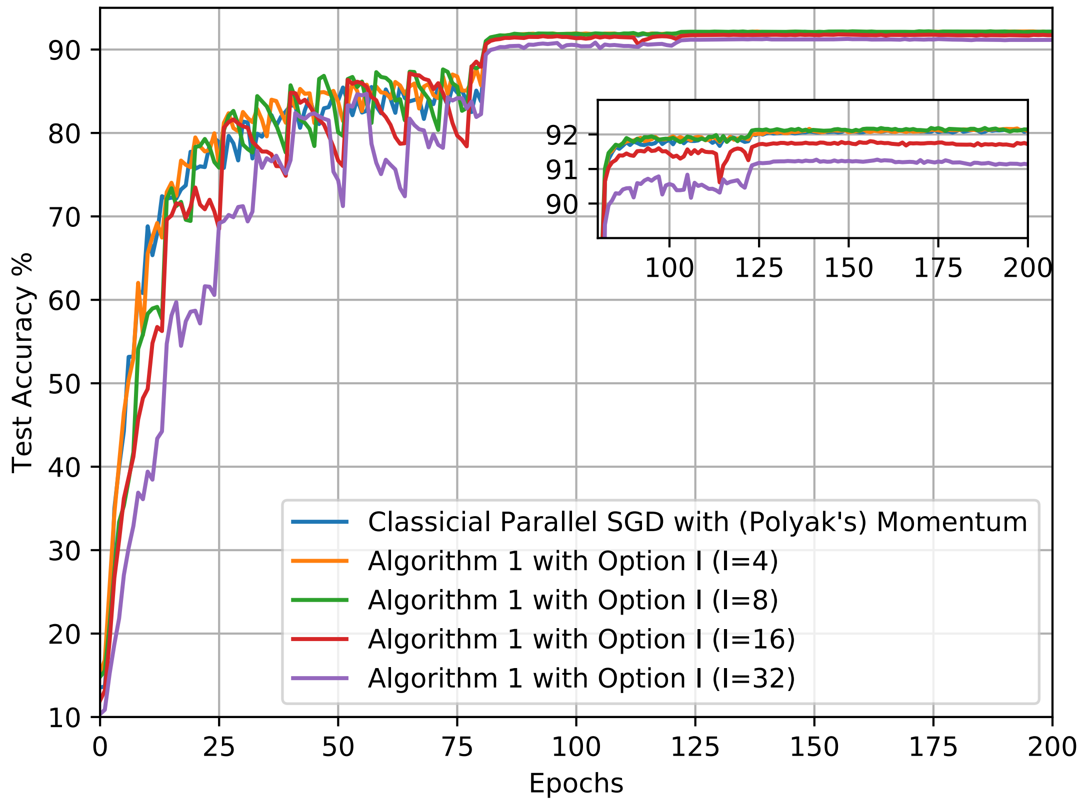


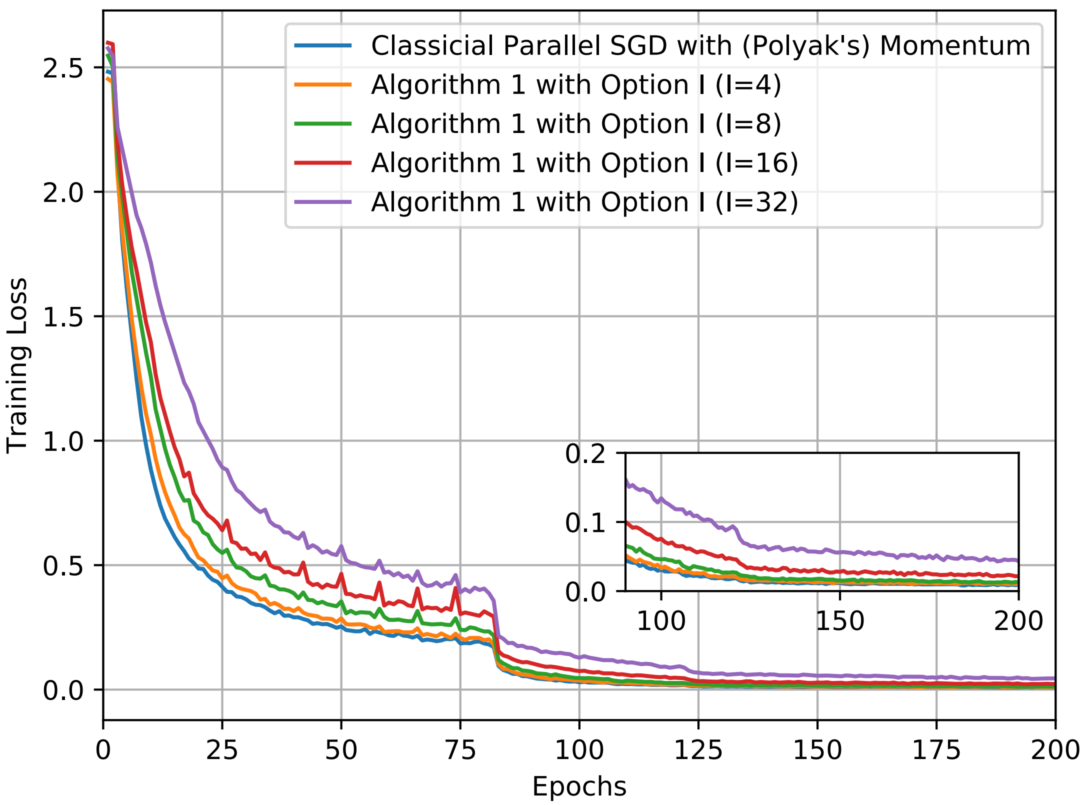


# of epochs **jointly** accessed by **all** workers

Experiments

Train ResNet56 over Cifar10 with 8 workers. I=4,8,16,32; periodically decayed learning rates in [He et.al.'16]
Similar observation for Imagenent. (see supplement in our paper)



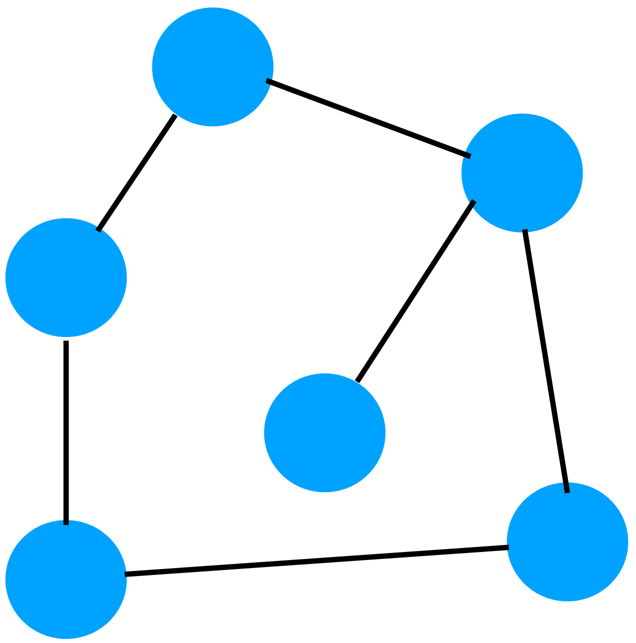


# of epochs **jointly** accessed by **all** workers

Extension: Distributed Momentum SGD with decentralized communication

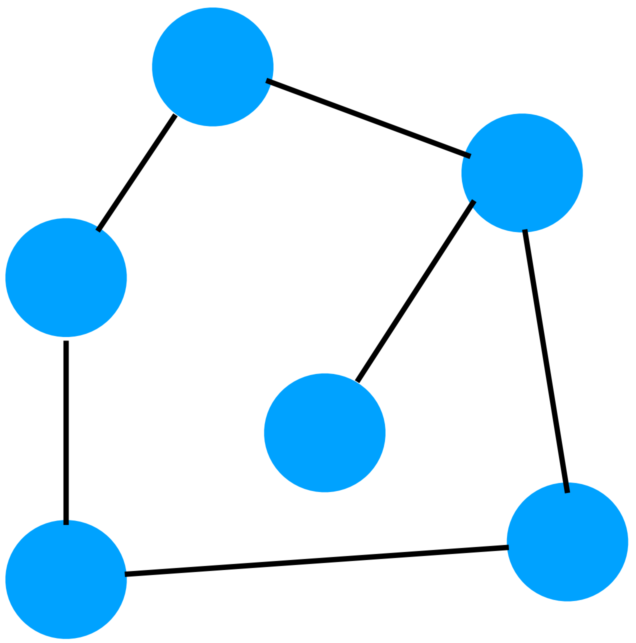
- PR-SGD-Momentum requires to average/aggregate models from all workers.

Extension: Distributed Momentum SGD with decentralized communication



- PR-SGD-Momentum requires to average/aggregate models from all workers.
- What if workers are only allowed to communicate with neighbors?

Extension: Distributed Momentum SGD with decentralized communication



- PR-SGD-Momentum requires to average/aggregate models from all workers.

- What if workers are only allowed to communicate with neighbors?

- This paper shows momentum SGD with decentralized communication has $O(1/\sqrt{NT})$ convergence. Its zero-momentum case degrades to the results in [Lian et.al.'17].

Thanks!

Poster on Wed Jun 12th 06:30 -- 09:00 PM @ Pacific Ballroom #182