Stochastic Gradient Push for Distributed Deep Learning

Mido Assran, Nicolas Loizou, Nicolas Ballas, Mike Rabbat
parallel Stochastic Gradient Descent

\[ x^{(k+1)} = x^{(k)} - \gamma^{(k)} \left( \frac{1}{n} \sum_{i=1}^{n} \nabla \tilde{f}_i(x) \right) \]

inter-node average

\[ x^{(k+1)} = \frac{1}{n} \sum_{i=1}^{n} \left( x^{(k)} - \gamma^{(k)} \nabla \tilde{f}_i(x) \right) \]
Data Parallel Training

Existing Approaches

1. Parallel SGD (AllReduce gradient aggregation, all nodes)
Data Parallel Training

Existing Approaches

1. Parallel SGD \((\text{AllReduce gradient aggregation, all nodes})\)
Data Parallel Training

Existing Approaches

1. Parallel SGD (AllReduce gradient aggregation, all nodes)
2. D-PSGD (PushPull parameter aggregation, neighboring nodes)
3. AD-PSGD (PushPull parameter aggregation, pairs of nodes)

Data Parallel Training

Existing Approaches

1. **Parallel SGD** (*AllReduce gradient aggregation, all nodes*)

2. **D-PSGD** (*PushPull parameter aggregation, neighboring nodes*)

3. **AD-PSGD** (*PushPull parameter aggregation, pairs of nodes*)

Blocks subsets of nodes and requires deadlock avoidance

---

Data Parallel Training

Existing Approaches

1. **Parallel SGD** (*AllReduce gradient aggregation, all nodes*)

2. **D-PSGD** (*PushPull parameter aggregation, neighboring nodes*)

3. **AD-PSGD** (*PushPull parameter aggregation, pairs of nodes*)

   Blocks subsets of nodes and requires deadlock avoidance

Proposed Approach

**Stochastic Gradient Push** (*PushSum parameter aggregation*)

nonblocking, no deadlock avoidance required
Stochastic Gradient Push

Enables optimization over directed and time-varying graphs

Stochastic Gradient Push

Enables optimization over directed and time-varying graphs

... naturally enables asynchronous implementations

Stochastic Gradient Push

Time

Local Optimization
Nonblocking Communication
Local Optimization
Nonblocking Communication
Local Optimization
Nonblocking Communication
Local Optimization
Nonblocking Communication

facebook Artificial Intelligence Research
Distributed Stochastic Optimization

ImageNet, ResNet 50

32 nodes (256 GPUs) interconnected via 10 Gbps Ethernet
Stochastic Gradient Push

Data Parallelism

Algorithm features:

* nonblocking communication
Stochastic Gradient Push

Data Parallelism

Algorithm features:

* nonblocking communication
* convergence guarantees for smooth non-convex functions with arbitrary (bounded) message staleness

code: github.com/facebookresearch/stochastic_gradient_push
poster: Pacific Ballroom #183