Moshpit SGD: Communication-Efficient Decentralized Training on Heterogeneous Unreliable Devices

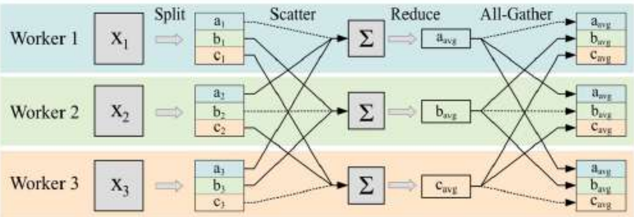
Max Ryabinin*, Eduard Gorbunov*,
Vsevolod Plokhotnyuk, Gennady Pekhimenko

The Paper in Brief

- We develop a simple method for distributed training with unstable (i.e. frequently joining and leaving) workers
- It combines the communication efficiency of All-Reduce with the fault tolerance of Gossip-based methods
- Has strong theoretical guarantees both for convergence to the actual average and stochastic optimization
- In practice, allows distributed training on preemptible instances and outperforms standard data-parallel training at a fraction of cost for ResNet and ALBERT

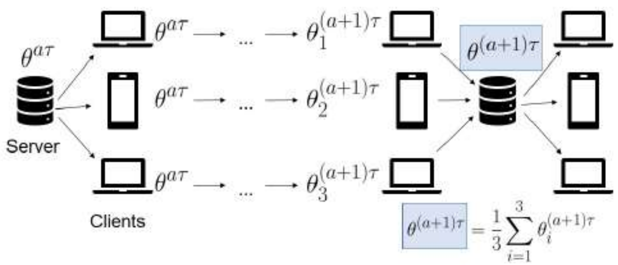
Background: Data-Parallel Training

- Most popular approach to distributed training: split batches across devices, average gradients, run the SGD step
- Naively sending all gradients is slow; more efficient versions (Ring, Butterfly) are used in practice
- However, standard All-Reduce fails in congested/high-latency networks
- Gossip fares better, but loses efficiency (sends all data to each neighbor)



Background: Local SGD

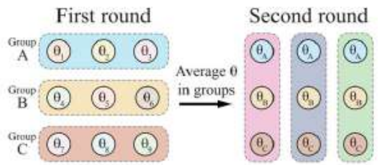
Iteration $a\tau$ τ local steps of SGD Iteration $(a+1)\tau$



$$\theta^{(a+1)\tau} = \frac{1}{3}\sum_{i=1}^{3}\theta_i^{(a+1)\tau}$$

Moshpit All-Reduce

Special case: exact averaging in d steps for a full "grid"



$N = 9$

$d = 2$

General case:

Algorithm 1 Moshpit All-Reduce (for i-th peer)

Input: parameters $\{\theta_j\}_{j=1}^N$, number of peers N, d, M, number of iterations T, peer index i
$\theta_i^0 := \theta_i$
$C_i^0 :=$ `get_initial_index(i)`
for $t \in 1 \ldots T$ **do**
  `DHT`$[C_i^{t-1}, t]$`.add(address`$_i$`)`
  `Matchmaking()` // wait for peers to assemble
  `peers`$_t :=$ `DHT.get(`$[C_i^{t-1}, t]$`)`
  $\theta_i^t, c_i^t :=$ `AllReduce(`θ_i^{t-1}`, peers`$_t$`)`
  $C_i^t := (C_i^{t-1}$`[1:]`$, c_i^t)$ // same as eq. (1)
end for
Return θ_i^T

- Group key:

$$C_i^t := \left(c_i^{t-d+1}, c_i^{t-d+2}, \ldots, c_i^t\right)$$

- get_initial_index(i) =

$$\left(\left\lfloor i/M^{d-1} \right\rfloor \mod M\right)_{j\in\{1,\ldots,d\}}$$

- **Key property:** if two peers were in the same group in round t, they choose different groups in round t+1

Theoretical guarantees:

1. **Correctness:** If all workers have a non-zero probability of successfully running a communication round and the order of peers$_t$ is random, then all local vectors converge to the global average with probability 1:

$$\forall i, \left\|\theta_i^t - \frac{1}{N}\sum_i \theta_i^0\right\|_2^2 \xrightarrow[t\to\infty]{} 0$$

2. **Exponential convergence to the average:** for a version of Moshpit All-Reduce with random splitting into r groups at each step, we have

$$\mathbb{E}\left[\frac{1}{N}\sum_{i=1}^N \left\|\theta_i^T - \bar{\theta}\right\|^2\right] = \left(\frac{r-1}{N} + \frac{r}{N^2}\right)^T \frac{1}{N}\sum_{i=1}^N \|\theta_i - \bar{\theta}\|^2$$

Moshpit SGD & Its Convergence

Moshpit SGD = Local SGD + Moshpit All-Reduce

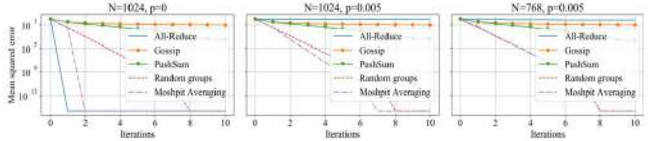
Theoretical guarantees: under the standard assumptions of bounded variance of stochastic gradients, reasonable number of iterations of Moshpit All-Reduce, and the bounded effect of peers' vanishing we recover:

- The best known rates from (Khaled et al., 2020; Woodworth et al., 2020) in **convex and strongly convex** cases
- The best known rates from (Koloskova et al., 2020; Li et al., 2019) in the **non-convex** case

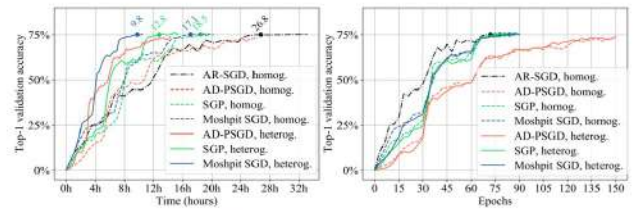
Experiments

Averaging performance

- We compare per-iteration convergence in a simulated setup
- All-Reduce takes too long to average with non-zero failure probability
- Gossip/SGP converge much slower (>10 iterations for target precision)



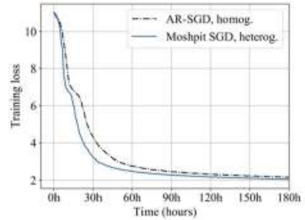
ResNet-50 on ImageNet



- We evaluate Moshpit-SGD and several baselines in two environments
- (16 nodes with 1xV100 and 64 workers with 81 different GPUs)
- Comparable to All-Reduce in terms of iterations, faster in terms of time
- Decentralized methods run faster, but achieve worse results

ALBERT on BookCorpus

- Baseline: All-Reduce on 8 V100
- Moshpit SGD: 66 preemptible GPUs
- For standard DDP, latency and failures make it impossible to run
- Costs of spot instances are much smaller, yet we converge 1.5x faster



References

Ahmed Khaled, Konstantin Mishchenko, and Peter Richtárik. Tighter Theory for Local SGD on Identical and Heterogeneous Data. In IAISTATS, pages 4519–4529. PMLR, 2020.

Blake Woodworth, Kumar Kshitij Patel, Sebastian Stich, Zhen Dai, Brian Bullins, Brendan Mcmahan, Ohad Shamir, and Nathan Srebro. Is Local SGD Better Than Minibatch SGD? In ICML, pages 10334–10343. PMLR, 2020.

Anastasia Koloskova, Nicolas Loizou, Sadra Boreiri, Martin Jaggi, and Sebastian Stich. A Unified Theory of Decentralized SGD with Changing Topology and Local Updates. In IICML, pages 5381–5393. PMLR, 2020.

Xiang Li, Wenhao Yang, Shusen Wang, and Zhihua Zhang. Communication Efficient Decentralized Training with Multiple Local Updates. arXiv:1910.09126, 2019.

Zhenzhong Lan, Mingda Chen, Sebastian Goodman, Kevin Gimpel, Piyush Sharma, Radu Soricut. ALBERT: A Lite BERT for Self-supervised Learning of Language Representations. ICML, 2020.