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

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 the 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.

Background: Local SGD

- Iteration $\alpha \mathcal{T}$: local steps of SGD
- Iteration $(\alpha + 1)\mathcal{T}$

Moshpit All-Reduce

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

<table>
<thead>
<tr>
<th>First round</th>
<th>Second round</th>
</tr>
</thead>
<tbody>
<tr>
<td>$G_1$</td>
<td>$G_2$</td>
</tr>
<tr>
<td>$G_3$</td>
<td>$G_4$</td>
</tr>
<tr>
<td>$G_5$</td>
<td>$G_6$</td>
</tr>
<tr>
<td>$G_7$</td>
<td>$G_8$</td>
</tr>
<tr>
<td>$G_9$</td>
<td>$G_{10}$</td>
</tr>
<tr>
<td>$G_{11}$</td>
<td></td>
</tr>
</tbody>
</table>

General case:

1. Group key: $c_i^d = (c_i^1, \ldots, c_i^d)$
2. Group leader: $\text{get领导者}(i) = (i/1, \ldots, M/1)$
3. Key property: if two peers are 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 is random, then all local vectors converge to the global average with probability 1.
2. Exponential convergence to the average: for a version of Moshpit All-Reduce with random splitting into $r$ groups at each step, we have

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 performance in a simulated setup.
- All-Reduce takes too long to average with non-zero failure probability.
- Gossip/SGD converge much slower (>10 iterations for target precision).

ResNet-50 on ImageNet

ALBERT on BookCorpus
- Baseline: All-Reduce on 8 V100
- Moshpit SGD: 64 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