Distributed Optimization in Cluster-Based Systems

Noa Schiller  Supervised by Hagit Atiya
Computer Science Department, Technion
noa.schiller@cs.technion.ac.il

Cluster-Based Model

Stochastic Gradient Descent

Processes have the same data distribution $D$, loss function $\ell(x, z)$ and stochastic gradient oracle $G(x, z)$
Collectively minimize the cost function $Q(x) \triangleq \mathbb{E}_{z \sim D}[\ell(x, z)]$
Each cluster $c$ starts at the same $x_1$ and outputs $x^c$:
Convergence: $\mathbb{E}[\|\nabla Q(x^c)\|_2^2] \leq \epsilon$
Bounded diameter: $\mathbb{E}[\max_{c, c'} \|x^c - x^{c'}\|_2] \leq \delta$

Strongly-Convex Cost Functions

Theorem. For a $\mu$-strongly convex $Q$ and any $\text{thresh}$:
$$\max_{c} \mathbb{E}[\|x^c_{t+1} - x^c_t\|_2^2] \leq (1 - \eta_t \mu) \max_{c} \mathbb{E}[\|x^c_t - x^*\|_2^2] + \frac{2\eta^2_t \sigma^2}{\text{thresh}}$$

Non-Convex Cost Functions

Theorem. For smooth non-convex cost function $Q$,
$$\mathbb{E}[Q(x^c_{t+1})] \leq \mathbb{E}[Q(x^c_t)] - \frac{\eta_t}{4} \mathbb{E}[\nabla Q(x^c_t)]^2 + \frac{4\eta_t \sigma^2}{\text{thresh}} + \left(\frac{2}{\eta_t} + 4\eta_t L^2\right) \mathbb{E}[\Delta_{cw}(x^c_t)]$$

In addition, if $\text{thresh} > \frac{m}{2}$ then $\mathbb{E}[\Delta_{cw}(x^c_t)] = O(C\eta_t \epsilon_{cluster})$

Diameter contraction happens explicitly

Problem constants

Diameter

Cluster fails when all its processes fail

Decentralized Algorithm

Process $p_i$, part of cluster $c$, in iteration $t = 1, \ldots, T$:
1. Compute stochastic gradient $g^c_t$ at $x^c_t$
2. Update $x^c_{t+1} \leftarrow x^c_t - \eta_t g^c_t$
3. Send $(t, x^c_{t+1})$ to all processes
4. Collect round-$t$ messages from at least $\text{thresh}$ clusters
5. Set using CAS $x^c_{t+1} \leftarrow \text{avg}$(received round-$t$ parameters)

Partitioning

$\forall p, p_j$ in the same cluster
$$\mathbb{E}[\|g_i - g_j\|_2^2] \leq \epsilon_{\text{cluster}}$$

Happens when at least half of the clusters fail

No convergence with partitioning

Convergence with no partitioning

Informal) Theorem. There is non-convex cost function such that any algorithm running in the presence of partitioning results in $\mathbb{E}[\Delta_2(x)] > \delta$

Coordinate-wise diameter