Motivation

- Nonconvex optimization is quite common in Machine Learning (Deep Learning, Natural Language Processing, Recommendation System, etc.)
- Asynchronous Stochastic Gradient (AsySG) is a powerful method in solving large scale machine learning problems.
- However, the theoretical analysis is still limited for nonconvex optimization.

Background

Consider the nonconvex optimization (deep learning, NLP, Recommendation):

- \( f(x) \in \mathbb{R} \) is a random variable.
- \( f(x) \) is a smooth but not necessarily convex function.
- Example: \( [1.2, 3.4, \cdots, N] \) is an index set of all training samples and \( f(x) \) is the corresponding loss function.

Asynchronous vs Synchronous

<table>
<thead>
<tr>
<th>Time</th>
<th>Child Node</th>
<th>Thr Node</th>
<th>Machine</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Idle</td>
<td>Computing</td>
<td>Update</td>
</tr>
<tr>
<td>2</td>
<td>Computing</td>
<td>Update</td>
<td>Computing</td>
</tr>
<tr>
<td>3</td>
<td>Update</td>
<td>Computing</td>
<td>Idle</td>
</tr>
</tbody>
</table>

Key challenges in analysis

1. \( x_k \neq x_i \): Different implementations \( \Rightarrow \) Different forms of \( x_i \).
2. \( x_k \neq x_i \): \( \Rightarrow \) Different implementations \( \Rightarrow \) Different forms of \( x_i \).

Two Implementations of AsySG (AsySG-con & AsySG-incon)

The procedure of AsySG:

A central node or a shared memory maintains the optimization variable \( x \).

All child nodes/threads run the following procedure concurrently:

1. (Read): read the parameter \( x_k \) from the central node/shared memory.
2. (Compute): sample \( m \) training data \((x_1, y_1), \cdots, (x_m, y_m)\) and compute a batch of the stochastic gradient \( G_k, \bar{G}_k = \nabla f(x_k, \xi) \) locally.
3. (Update): Update parameter \( x \) in the central node/shared memory without locks.

Convergence Rate for AsySG

Theorem

Assume that certain assumptions hold and \( \mathbb{E} (\| f(x, t) - f(x) \|^2) \leq \sigma^2 \).

Set the step length to be a constant \( \gamma = \frac{1}{\mathbb{E} \| f(x, t) - f(x) \|^2} \), then the output of AsySG-con satisfies the following ergodic convergence rate:

\[
\min_{k \geq 1} \mathbb{E} (\| f(x_k, t) - f(x) \|^2) \leq \frac{1}{N \gamma} \mathbb{E} (\| f(x_k) - f(x) \|^2) \leq O \left( \frac{\sigma}{\gamma \sqrt{N}} \right).
\]

- Consistent convergence rate with SGD.
- Linear speedup up to \( \mathcal{O}(N) \) machines.
- Better linear speedup property than existing work (linear speedup up to \( \mathcal{O}(k^{1/2}) \) machines in [2]).
- For AsySG-incon, we have a similar convergence rate.

References
