Trading Redundancy for Communication: Speeding up Distributed SGD for Non-convex Optimization

Farzin Haddadpour
Joint work with

Mohammad Mahdi Kamani  Mehrdad Mahdavi  Viveck Cadambe
Goal: Solving $\min \ f(x) \triangleq \sum_i f_i(x)$
Goal: Solving $\min f(x) \triangleq \sum_i f_i(x)$

SGD

$x^{(t+1)} = x^{(t)} - \eta \frac{1}{|\xi^{(t)}|} \nabla f(x^{(t)}; \xi^{(t)})$
Goal: Solving \( \min f(x) \triangleq \sum_i f_i(x) \)

\[ x^{(t+1)} = x^{(t)} - \eta \frac{1}{|\xi^{(t)}|} \nabla f(x^{(t)}; \xi^{(t)}) \]

Parallelization due to computational cost

\[ x^{(t+1)} = x^{(t)} - \frac{\eta}{p} \sum_{j=1}^{p} \frac{1}{|\xi_j^{(t)}|} \nabla f(x^{(t)}; \xi_j^{(t)}) \]

SGD

Distributed SGD
Goal: Solving \( \min f(x) \triangleq \sum_i f_i(x) \)

**SGD**

\[ x^{(t+1)} = x^{(t)} - \eta \frac{1}{|\xi(t)|} \nabla f(x^{(t)}; \xi^{(t)}) \]

Parallelization due to computational cost

**Distributed SGD**

\[ x^{(t+1)} = x^{(t)} - \frac{\eta}{p} \sum_{j=1}^p \frac{1}{|\xi_j^{(t)}|} \nabla f(x^{(t)}; \xi_j^{(t)}) \]

Communication is bottleneck
Communication

Number of bits per iteration

Gradient compression based techniques
Communication

- Number of bits per iteration
  - Gradient compression based techniques
- Number of rounds
  - Local SGD with periodic averaging
Local SGD with periodic averaging

\[
\begin{align*}
x_j^{(t+1)} &= \frac{1}{p} \sum_{j=1}^{p} \left[ x_j^{(t)} - \eta \tilde{g}_j^{(t)} \right] \quad \text{if } \tau | T \\
x_j^{(t+1)} &= x_j^{(t)} - \eta \tilde{g}_j^{(t)} \quad \text{otherwise},
\end{align*}
\]

Averaging step (a)

Local update (b)
Local SGD with periodic averaging

\[ x_j^{(t+1)} = \frac{1}{p} \sum_{j=1}^{p} \left[ x_j^{(t)} - \eta \tilde{g}_j^{(t)} \right] \] if \( \tau \mid T \)

\[ x_j^{(t+1)} = x_j^{(t)} - \eta \tilde{g}_j^{(t)} \] otherwise,

\[ p = 3, \tau = 1 \]
Local SGD with periodic averaging

\[
\begin{align*}
    x_j^{(t+1)} &= \frac{1}{p} \sum_{j=1}^{p} \left[ x_j^{(t)} - \eta \tilde{g}_j^{(t)} \right] \quad \text{if } \tau | T \\
    x_j^{(t+1)} &= x_j^{(t)} - \eta \tilde{g}_j^{(t)} \quad \text{otherwise},
\end{align*}
\]

- **Averaging step (a)**
- **Local update (b)**

\[p = 3, \tau = 1\]

\[p = 3, \tau = 3\]
Convergence Analysis of Local SGD with periodic averaging

Table 1: Comparison of different SGD based algorithms.

<table>
<thead>
<tr>
<th>Strategy</th>
<th>Convergence error</th>
<th>Assumptions</th>
<th>Com-round($T/\tau$)</th>
</tr>
</thead>
<tbody>
<tr>
<td>SGD</td>
<td>$O(1/\sqrt{pT})$</td>
<td>i.i.d. &amp; b.g</td>
<td>$T$</td>
</tr>
<tr>
<td>[Yu et.al.]</td>
<td>$O(1/\sqrt{pT})$</td>
<td>i.i.d. &amp; b.g</td>
<td>$O(p^{3/4}T^{1/4})$</td>
</tr>
<tr>
<td>[Wang &amp; Joshi]</td>
<td>$O(1/\sqrt{pT})$</td>
<td>i.i.d.</td>
<td>$O(p^{3/2}T^{1/2})$</td>
</tr>
</tbody>
</table>

b.g: Bounded gradient $\|g_i\|_2^2 \leq G$

Unbiased gradient estimation $\mathbb{E}[\tilde{g}_j] = g_j$
Convergence Analysis of Local SGD with periodic averaging

Table 1: Comparison of different SGD based algorithms.

<table>
<thead>
<tr>
<th>Strategy</th>
<th>Convergence error</th>
<th>Assumptions</th>
<th>Com-round($T/\tau$)</th>
</tr>
</thead>
<tbody>
<tr>
<td>SGD</td>
<td>$O(1/\sqrt{pT})$</td>
<td>i.i.d. &amp; b.g</td>
<td>$T$</td>
</tr>
<tr>
<td>[Yu et.al.]</td>
<td>$O(1/\sqrt{pT})$</td>
<td>i.i.d. &amp; b.g</td>
<td>$O(p^{3/4}T^{1/4})$</td>
</tr>
<tr>
<td>[Wang &amp; Joshi]</td>
<td>$O(1/\sqrt{pT})$</td>
<td>i.i.d.</td>
<td>$O(p^{3/2}T^{1/2})$</td>
</tr>
</tbody>
</table>

b.g: Bounded gradient $\|g_i\|_2^2 \leq G$

Unbiased gradient estimation $\mathbb{E}[\tilde{g}_j] = g_j$

A. Residual error is observed in practice but theoretical understanding is missing?
B. How can we capture this in convergence analysis?
C. Any solution to improve it?
A. Residual error is observe in practice but theoretical understanding is missing?

Unbiased gradient estimation does not hold
A. Residual error is observe in practice but theoretical understanding is missing?

- Unbiased gradient estimation does not hold

B. How to capture this in convergence analysis?

- Our work
- Analysis based on biased gradients
Insufficiency of convergence analysis

A. Residual error is observed in practice but theoretical understanding is missing?

Unbiased gradient estimation does not hold

B. How to capture this in convergence analysis?

Our work

Analysis based on biased gradients

C. Any solution to improve it?

Our work

Redundancy
Redundancy infused local SGD (RI-SGD)

\[ \mathcal{D} = \mathcal{D}_1 \cup \mathcal{D}_2 \cup \mathcal{D}_3 \]

Local SGD \( p = 3, \tau = 3 \)
Redundancy infused local SGD (RI-SGD)

\[ D = D_1 \cup D_2 \cup D_3 \]

Local SGD \( p = 3, \tau = 3 \)

RI-SGD \( q = 2, p = 3, \tau = 3 \)

Explicit redundancy
Comparing RI-SGD with other schemes

Assumption: Bounded inner product of gradients $\langle g_i, g_j \rangle \leq \beta$

Biased gradients

Redundancy: Number of data chunks at each worker node $q$
Comparing RI-SGD with other schemes

Assumption → b.d: Bounded inner product of gradients $\langle g_i, g_j \rangle \leq \beta$

Biased gradients

Redundancy → q: Number of data chunks at each worker node

Table 1: Comparison of different SGD based algorithms.

<table>
<thead>
<tr>
<th>Strategy</th>
<th>Convergence error</th>
<th>Assumptions</th>
<th>Com-round($T/\tau$)</th>
</tr>
</thead>
<tbody>
<tr>
<td>SGD</td>
<td>$O(1/\sqrt{pT})$</td>
<td>i.i.d. &amp; b.g</td>
<td>$T$</td>
</tr>
<tr>
<td>[Yu et.al.]</td>
<td>$O(1/\sqrt{pT})$</td>
<td>i.i.d. &amp; b.g</td>
<td>$O(p^{3/4} T^{1/4})$</td>
</tr>
<tr>
<td>[Wang &amp; Joshi]</td>
<td>$O(1/\sqrt{pT})$</td>
<td>i.i.d.</td>
<td>$O(p^{3/2} T^{1/2})$</td>
</tr>
<tr>
<td>RI-SGD ($\tau, q$)</td>
<td>$O(1/\sqrt{pT}) + O((1 - q/p)\beta)$</td>
<td>non-i.i.d. &amp; b.d.</td>
<td>$O(p^{3/2} T^{1/2})$</td>
</tr>
</tbody>
</table>
Advantages of RI-SGD:

1. Speed up not only due to larger effective mini-batch size, but also due to increasing intra-gradient diversity.
2. Fault-tolerance.
3. Extension to heterogeneous mini-batch size and possible application to federated optimization.
Faster convergence: Experiments over Image-net (top figures) and Cifar-100 (bottom figures)
Increasing intra-gradient diversity: Experiments over Cifar-10
Fault-Tolerance: Experiments over Cifar-10

Graphs showing the error rate over iterations and wall clock time for different values of $\mu$. The graphs compare the performance with and without node failures for $\mu = 0.0$ and $\mu = 0.5$.
For more details please come to my poster session
Wed Jun 12th 06:30 -- 09:00 PM @ Pacific Ballroom #185

Thanks for your attention!