Communication-Efficient Distributed Second-Order Optimization Methods for Generalized Convex Problems

Fred Roosta (Lecturer, School of Mathematics and Physics, The University of Queensland)
Rixon Crane (PhD Student, School of Mathematics and Physics, The University of Queensland)

Why use Second-Order Methods?
Second-order methods employ curvature (Hessian matrix) information to transform the gradient so that it is a more suitable direction to follow.

Benefits
- Perform more computations per iteration
- May take full advantage of available distributed computational resources
- May require significantly less communication costs
- Often require far fewer iterations to achieve similar results

Every Minute There Are:

- ~500,000 Tweets Sent
- ~50,000 Photos Posted
- ~100,000 Hours Streamed
- ~4,000,000 Searches
- ~750,000 Songs Streamed
- ~2,000,000 Snaps

Our Method: DINGO
Derived by optimization of the gradient's norm as a surrogate function, i.e.,

\[
\min_{w \in \mathbb{R}^d} \left\{ \| \nabla f(w) \|_2^2 = \frac{1}{m} \sum_{i=1}^{m} \| \nabla f_i(w) \|_2^2 \right\}
\]

DINGO is for "Distributed Newton-type method for Gradient-norm Optimization". DINGO is particularly suitable for inexact objectives. A strict linear-rate reduction in the gradient norm is always guaranteed.

Each Iteration of DINGO

Update Direction: \( p_t \)

Case 1

If \( \langle \nabla f, \nabla^2 f \rangle \geq \theta \) then \( p_t = -\nabla f - \lambda \nabla^2 f \)

Case 2

Else if \( \langle \nabla f, \nabla^2 f \rangle \geq \theta \) then \( p_t = -\nabla f - \lambda \nabla^2 f \)

Case 3

Else: \( p_t = -\nabla f - \lambda \nabla^2 f \) for all \( i \neq t \) \( \Rightarrow \{ i = 1, \ldots, m \} \cap \{ i, \nabla^2 f_i \} < \| \nabla^2 f \|_F^2 \}

Send \( H_{ij} p_t \) to all workers \( i \neq t \)

Let \( a_t = \max(1, \theta) \) such that \( \langle \nabla^2 f, \nabla^2 (w_t + a_t p_t) \rangle \leq \| \nabla^2 f \|_F^2 + 2 \alpha p_t \phi H_{ij} \)

Update

\( w_{t+1} = w_t + a_t p_t \)

The constants \( \theta, \phi > 0 \) and \( \rho \in (0,1) \) are hyper-parameters. The vector \( w_t \in \mathbb{R}^d \) denotes the point at iteration \( t \). For notational convenience, we denote \( \beta_{ij} = \nabla^2 f_i(w_t), H_{ij} = \nabla^2 f_i(w_t), \beta_i = \nabla^2 f_i(w_t), H_i = \nabla^2 f_i(w_t) \). We also let \( \beta_{ij} \leq \| \beta_{ij} \|_F^2 = \| \beta_{ij} \|_F^2 \leq \| \beta_{ij} \|_F^2 \leq \| \beta_{ij} \|_F^2 \leq \| \beta_{ij} \|_F^2 \leq \| \beta_{ij} \|_F^2 \)

Softmax regression, with regularization, problem on the CIFAR10 dataset.

Related Work

<table>
<thead>
<tr>
<th>Method</th>
<th>Applicable to Non-Convex Functions</th>
<th>Arbitrary Data Distribution</th>
<th>Arbitrary Form of ( f_i )</th>
<th>Simple Sub-Problems</th>
<th>Not Sensitive to Hyper-Parameters</th>
</tr>
</thead>
<tbody>
<tr>
<td>GIANT</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>DISCO</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>✓</td>
</tr>
<tr>
<td>DANE</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>InexactDANE</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>AIDE</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>DINGO</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
</tbody>
</table>

Convex
- Hessian is positive semidefinite.
- Local minima are global minima.

Inexact
- Hessian can be indefinite and singular.
- Local minima are global minima.

Non-Convex
- Hessian can be indefinite and singular.
- Not all local minima are global minima.

References