DINGO: Distributed Newton-Type Method for Gradient-Norm Optimization

Rixon Crane*, Fred Roosta*
*School of Mathematics and Physics, University of Queensland, Australia.

The Problem
Centralized Computing Environment

Use Case: Big Data Regimes

Distributively Working With a Very Large Dataset $\{\{x_i\}\}_{i=1}^\infty$

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

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

$$\min_{w \in \mathbb{R}^d} \left\{ \|\nabla f(w)\|^2 \right\} \leq \frac{1}{\lambda_d} \sum_{i=1}^n f_i(w)$$

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

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>$\times$</td>
<td>$\times$</td>
<td>$\times$</td>
<td>$\checkmark$</td>
<td>$\checkmark$</td>
</tr>
<tr>
<td>DISCO</td>
<td>$\times$</td>
<td>$\checkmark$</td>
<td>$\checkmark$</td>
<td>$\checkmark$</td>
<td>$\checkmark$</td>
</tr>
<tr>
<td>DANE</td>
<td>$\checkmark$</td>
<td>$\checkmark$</td>
<td>$\checkmark$</td>
<td>$\times$</td>
<td>$\checkmark$</td>
</tr>
<tr>
<td>InexactDANE</td>
<td>$\checkmark$</td>
<td>$\checkmark$</td>
<td>$\checkmark$</td>
<td>$\checkmark$</td>
<td></td>
</tr>
<tr>
<td>AIDE</td>
<td>$\checkmark$</td>
<td>$\checkmark$</td>
<td>$\checkmark$</td>
<td>$\checkmark$</td>
<td>$\checkmark$</td>
</tr>
<tr>
<td>DINGO</td>
<td>$\checkmark$</td>
<td>$\checkmark$</td>
<td>$\checkmark$</td>
<td>$\checkmark$</td>
<td>$\checkmark$</td>
</tr>
</tbody>
</table>

Benefits

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

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

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

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

Each Iteration of DINGO

- **Update Direction:** $p_k$
  - **Case 1:** $\langle \sum_{i=1}^n H_i p_k, p_k \rangle \geq \|\nabla f_i(x_k)\|^2$ then $p_k = -\sum_{i=1}^n H_i p_k$ with $p_j = -H_j^{-1} \nabla f_j(x_k)$
  - **Case 2:** $\langle \sum_{i=1}^n H_i p_k, p_k \rangle \geq \|\nabla f_i(x_k)\|^2$ then $p_k = -\sum_{i=1}^n H_i p_k$ with $p_j = -H_j^{-1} \nabla f_j(x_k)$
  - **Case 3:** $\langle \sum_{i=1}^n H_i p_k, p_k \rangle \geq \|\nabla f_i(x_k)\|^2$ then $p_k = -\sum_{i=1}^n H_i p_k$ with $p_j = -H_j^{-1} \nabla f_j(x_k)$

- **Line Search:** $r_k$
  - Distributively choose $r_k \in (0,1]$ such that $\nabla f_i(x_k + r_k p_k) \leq \|\nabla f_i(x_k)\|^2 + 2r_k \|\nabla f_i(x_k)\|^2$

- **Update:** $x_{k+1} = x_k + r_k p_k$