Intra-iteration Approximation for Large Scale Parallel Gradient Descent Optimization
Chengjie Qin & Florin Rusu
EECS, University of California, Merced

Motivation & Goal
- Per iteration time of large-scale training can be long even if parallel training is deployed.
- A user can do nothing while an iteration is running.
- We introduce intra-iteration techniques to speed up large scale parallel training and to allow interactive parameter tuning.

(Stochastic) Gradient Descent
- Problem definition:
  \[ \min_{\mathbf{w} \in \mathbb{R}^d} \mathcal{L}(\mathbf{w}) = \sum_{i=1}^N f(\mathbf{w}, \mathbf{x}_i; y_i) + \mu R(\mathbf{w}) \]  
- \( \mathbf{w} \) is the parameter to be learned, \( \mathbf{x}_i \) is feature vector, \( y_i \) is label.
- Gradient Descent
  \[ \mathbf{w}^{(k+1)} = \mathbf{w}^{(k)} - \alpha^{(k)} \nabla \mathcal{L}(\mathbf{w}^{(k)}) \]
- Stochastic Gradient Descent
  \[ \mathbf{w}^{(k+1)} = \mathbf{w}^{(k)} - \alpha^{(k)} \nabla f_i(\mathbf{w}^{(k)}) \]

Datasets & Tasks

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Dimension</th>
<th># Examples</th>
<th>Size</th>
</tr>
</thead>
<tbody>
<tr>
<td>splice</td>
<td>13M</td>
<td>50M</td>
<td>3TB</td>
</tr>
</tbody>
</table>

Query Overlapping & Intra-Iteration Approximation Parallel (Stochastic) Gradient Descent in GLADE

Iteration Time Comparison
- SVM on splice 50m examples, 13m features
- Single nodes: 16 cores @ 2GHz; 20GB RAM; 4 disks @ 110MB/s throughput/disk

OLA PGD Loss Comparison
- SVM on splice, 9 nodes

OLA PSGD Loss Comparison
- SVM on splice, 9 nodes

Figure: Loss comparison between Query Overlapping + PGD and OLA + Query Overlapping + PGD

Figure: Loss comparison between Query Overlapping + PSGD and OLA + Query Overlapping + PSGD