Training Quantized Nets: A Deeper Understanding

Hao Li*, Soham De*, Zheng Xu†, Christoph Studer†, Hanan Samet†, Tom Goldstein†
University of Maryland, College Park†, Cornell University‡

Convergence of Stochastic Rounding

• Theorem 1 Assume that is µ-strongly convex and the learning rates are given by .

SR converges until it reaches an "accurate floor," which is determined by the quantization error .

Convergence of BinaryConnect

• Theorem 4 Assume that is µ-strongly convex and the learning rates are given by .

BC converges until it reaches an "accurate floor," which is determined by the quantization error and .

Goal

Study quantized training methods from a theoretical perspective.

Understand the differences in behavior, and reasons for success or failure, of various methods.

Main Results

We proved that both BinaryConnect (BC) and Stochastic Rounding (SR) are capable of solving convex discrete problems up to a level of accuracy that depends on the quantization level.

We address the issue of why algorithms that maintain floating-point representations, like BC, work so well, while fully quantized training methods like SR stall before training is complete.

BC, like SGD, can efficiently concentrate on minimizers.

SR cannot concentrate on minimizers on decreasing step size.

Continuous-valued SGD & BC

• When the learning rate is large, the algorithm explores by moving quickly between states.

• When the learning rate is small, exploitation happens.

Continuous-valued SGD & BC

Theorem 5 Let denote the distribution function of the th entry in the stochastic gradient . Assume , and both for all , and both and for constants and .

Define .

where is the largest constant that makes non-negative. Suppose has a stationary distribution . Then, for sufficiently small , has stationary distribution and .

This satisfies for any state and is not concentrated on minimizers of .

Exploration-Exploitation Tradeoff

Stochastic Rounding lacks this important tradeoff.

As the stepsize gets small and the algorithm slows down, the quality of the iterates does not improve.

Effect of shrinking the learning rate in SR vs BC on a toy problem. Histograms plot the distribution of the quantized weights over 10^6 iterations. As the learning rate shrinks, the BC distribution concentrates on a minimizer, while the SR distribution stagnates.

SR starts at some state , and moves to a new state with some transition probability , that depends only on and the learning rate .

For fixed , this is a Markov process with transition matrix .

As gets smaller, the distribution of the perturbation gets "squished," making the algorithm less likely to move. The "squishing" effect does not affect the relative probability of moving left or right.

Our theory predicts that we can improve the performance of SR by increasing the batch size, which shrink the variance of the gradient distribution without changing the mean and concentrates more of the gradient distribution towards downhill directions, making the algorithm more greedy.

Table 1: Top 4 test error after training with bit precision (ADAM); quantized weights (SR-ADAM, BC-ADAM); and quantized weights with big batch size (Big SR-ADAM).

<table>
<thead>
<tr>
<th>Model</th>
<th>4-bit</th>
<th>8-bit</th>
<th>16-bit</th>
</tr>
</thead>
<tbody>
<tr>
<td>ADAM</td>
<td>6.2</td>
<td>6.8</td>
<td>7.1</td>
</tr>
<tr>
<td>BC-ADAM</td>
<td>6.36</td>
<td>8.21</td>
<td>10.67</td>
</tr>
<tr>
<td>Big SR-ADAM</td>
<td>6.95</td>
<td>16.77</td>
<td>19.84</td>
</tr>
<tr>
<td>SR-ADAM</td>
<td>23.33</td>
<td>20.56</td>
<td>26.49</td>
</tr>
</tbody>
</table>

References