How Learning Rate and Delay Affect Minima Selection in Asynchronous Training of Neural Networks

Niv Giladi*, Mor Shpigel Nacson*, Elad Hoffer, Daniel Soudry
{giladiniv, mor.shpigel, elad.hoffer, daniel.soudry}@gmail.com

Background

Asynchronous SGD (A-SGD) can speed up training and improve scalability of neural networks. Training with A-SGD without hyperparameter adjustment hurts the generalization whereas adjusting the hyperparameters can improve generalization.

Related Result - Dynamical Stability:

- The dynamical stability approach was used in [2, 3], to study which minima are accessible under a specific choice of optimization algorithm and hyperparameters.

Goal

We use dynamical stability to analyze the dynamics of asynchronous SGD. Using this approach, the main question we try to tackle is

How do learning rate and delay interact and affect the minima selection process?

Theory

A-SGD:

\[ f(x) = \frac{N}{i=1} f_i(x) , \]

\[ x_{t+1} = x_t - \eta \nabla f_{n(t-\tau)}(x_{t-\tau}) \]

- \(\eta\) - learning rate, \(N\) - number of samples, \(\tau\) - delay
- \(n(t)\) - random selection process of a sample from \(\{1,2,\ldots,N\}\) at iteration \(t\)

We examine the expectation of the linearized dynamics around some minimum point \(x^*\). In order to ensure stability it is sufficient to require that the following equation is stable:

\[ x_{t+1} - x_t + \eta a x_{t-\tau} = 0 \]  

(1)

where \(a = \lambda_{\max}(H)\) is a "sharpness term" and \(H = \frac{1}{N} \sum_{i=1}^{N} \nabla^2 f_i(x^*)\). The characteristic equation is

\[ z^{\tau+1} - z^\tau + \eta a = 0 \]  

(2)

In order to ensure stability of eq. 2 it is required that \(a\eta = 2\sin\left(\frac{1}{\tau+1}\right)\). For large \(\tau\) we can use Taylor approximation and get

\[ \frac{1}{a\eta} = \frac{1}{\pi}(2\tau + 1) + O\left(\frac{1}{\tau}\right) \]

Implication:

- In order to maintain stability for a given minimum point, the learning rate should be kept inversely proportional to the delay.

Results

Figure 2: (a): Stability threshold is maintained when \(\eta \propto 1/\tau\). We show the number of epochs it takes VGG-11 trained with CIFAR10 to diverge from a minimum as a function of the learning rate \(\eta\), scaled by the delay \(\tau\). The black circle is the stability threshold — below which, we do not escape the minimum. (b): For each value of delay \(\tau\) we show 1/\(\eta\) where \(\eta\) is the maximal learning rate in which we did not diverge. (c): Comparison of synchronous [1] and asynchronous training of ResNet-50 on ImageNet. Solid lines are validation error and dashed lines are train error. A-SGD training reaches 28.13% validation error (blue) while synchronous training reaches 23.74% (black). Training asynchronously using the proposed scaling rules reaches 25.49% (orange). (d): Zoom in of the last part of training.

References