Supplementary Material - HyperBlock Floating Point: Generalised Quantization Scheme for Gradient and Inference Computation

Marcelo Gennari do Nascimento
University of Oxford - Active Vision Lab
Parks Road, Oxford OX1 3PJ
marcelo@robots.ox.ac.uk

Victor Adrian Prisacariu
University of Oxford - Active Vision Lab
Parks Road, Oxford OX1 3PJ
victor@robots.ox.ac.uk

Roger Fawcett
Intel Corporation
Eclipse, Globe Park, Marlow SL7 1YL
roger.fawcett@intel.com

Martin Langhammer
Intel Corporation
Eclipse, Globe Park, Marlow SL7 1YL
martin.langhammer@intel.com

1. Stochastic Rounding Equivalence Proof

Stochastic rounding is an umbrella term that encompasses many different methods. The most common method is setting the probability of rounding up to be proportional to the normalized distance between the value and its ceiling. Mathematically, let x be a real number, $\lceil x \rceil$ be its rounded-up (ceiling) value, and $\lfloor x \rfloor$ be its rounded-down (floor) value. The most common stochastic rounding procedure used in deep learning consists of the following function:

$$f(x) = \begin{cases} \lceil x \rceil & w.p. \quad p = \frac{x - \lfloor x \rfloor}{\lceil x \rceil - \lfloor x \rfloor} \\ \lfloor x \rfloor & w.p. \quad p = \frac{\lceil x \rceil - x}{\lceil x \rceil - \lfloor x \rfloor} \end{cases} \tag{1}$$

assuming for example that $x \in [0, 1]$, then if $x = 0.8$, the probability that it will round to 1 is 80%. In hardware a stright-forward way of implementing this operation is by simply adding a number sampled from a uniform distribution, and flooring its value:

$$\begin{aligned} \nu &\sim U(0, \lceil x \rceil - \lfloor x \rfloor) \\ f(x) &= \lfloor x + \nu \rfloor \end{aligned} \tag{2}$$

where $U(0, a)$ is a uniform distribution from 0 to a. This has a close relation to the idea of *dithering* [2, 1], in which noise is intentionally added to the signal in order to suppress error. In the following part, we proof this equivalence.

Theorem 1.1. *The function $f(x) = \lfloor x + \nu \rfloor$ for $\nu \sim U(0, \lceil x \rceil - \lfloor x \rfloor)$ is equivalent to stochastic rounding as defined in equation 1.*

Proof. The function $f(x)$ can be rewritten in the following format:

$$f(x) = \begin{cases} \lceil x \rceil & w.p. \quad p(x + \nu \geq \lceil x \rceil) \\ \lfloor x \rfloor & w.p. \quad p(x + \nu < \lceil x \rceil) \end{cases}$$

Note that for any value of x and any quantization gap $\lceil x \rceil - \lfloor x \rfloor$, we have that $\lfloor x \rfloor \leq x \leq \lceil x \rceil$.

$$p(x + \nu \geq \lceil x \rceil) = \int_{\lceil x \rceil}^{\infty} x + U(0, \lceil x \rceil - \lfloor x \rfloor) dy \tag{3}$$

$$= \int_{\lceil x \rceil}^{\infty} U(x, x + \lceil x \rceil - \lfloor x \rfloor) dy \tag{4}$$

$$= \int_{\lceil x \rceil}^{\infty} U(x, x + \lceil x \rceil - \lfloor x \rfloor) dy \tag{5}$$

$$= \int_{\lceil x \rceil}^{\infty} \frac{1}{x + \lceil x \rceil - \lfloor x \rfloor - x} \mathbb{1}_{y \in (x, x + \lceil x \rceil - \lfloor x \rfloor)} dy \tag{6}$$

where $\mathbb{1}_{y \in (a,b)}$ is the indicator function, which is equal to 1 when the condition $y \in (a, b)$ is satisfied, otherwise it is equal to 0.

By definition we have that $x < \lceil x \rceil$, therefore:

$$p(x + \nu \geq \lceil x \rceil) = \int_{\lceil x \rceil}^{x + \lceil x \rceil - \lfloor x \rfloor} \frac{1}{\lceil x \rceil - \lfloor x \rfloor} dy \tag{7}$$

$$= \frac{x + \lceil x \rceil - \lfloor x \rfloor - \lceil x \rceil}{\lceil x \rceil - \lfloor x \rfloor} \tag{8}$$

$$= \frac{x - \lfloor x \rfloor}{\lceil x \rceil - \lfloor x \rfloor} \tag{9}$$

The proof for $p(x + \nu < \lceil x \rceil)$ is analogous. □

Theorem 1.2. *The function $f(x) = \lfloor x + \nu \rfloor$ for $\nu \sim U(0, \lceil x \rceil - \lfloor x \rfloor)$ is unbiased in expectation*

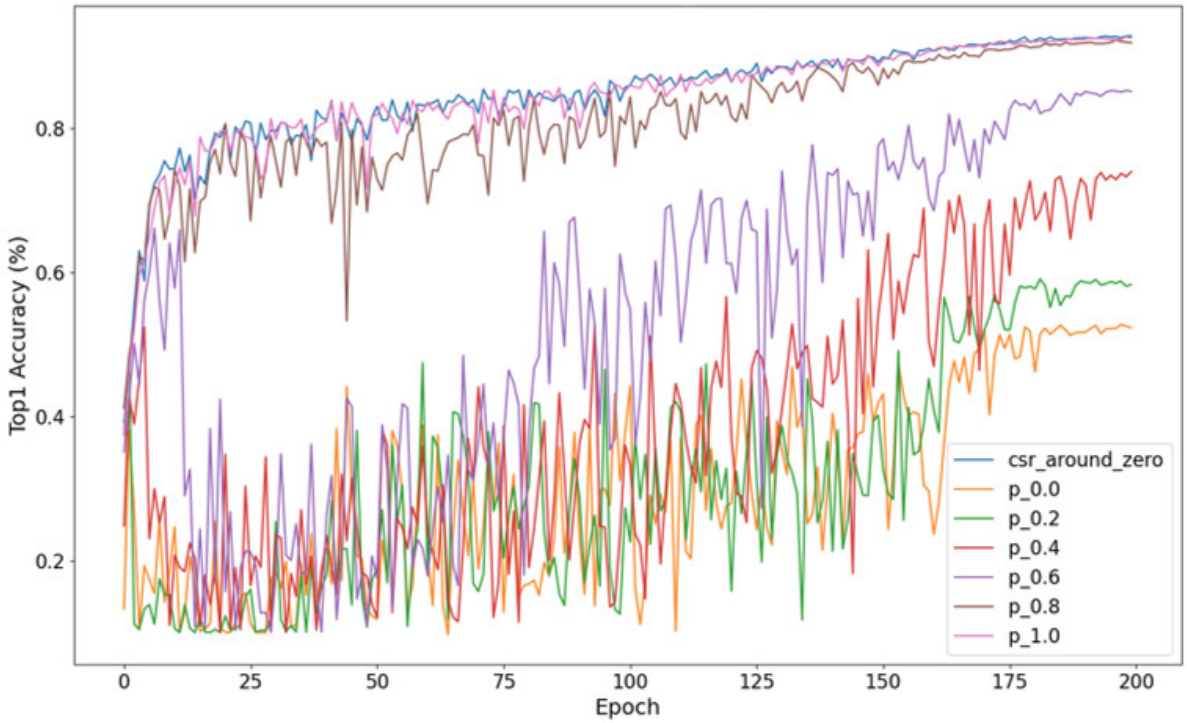

Figure 1. Stochastic Rounding using ResNet18 on CIFAR10 being applied at different percentages of tensor values, and around zero values.

Proof.

$$E[f(x)] = \lceil x \rceil \frac{x - \lfloor x \rfloor}{\lceil x \rceil - \lfloor x \rfloor} + \lfloor x \rfloor \frac{\lceil x \rceil - x}{\lceil x \rceil - \lfloor x \rfloor} \tag{10}$$

$$= \frac{\lceil x \rceil x - \lceil x \rceil \lfloor x \rfloor + \lfloor x \rfloor \lceil x \rceil - \lfloor x \rfloor x}{\lceil x \rceil - \lfloor x \rfloor} \tag{11}$$

$$= x \tag{12}$$

□

2. Stagnation Problem Analysis

There are two main ideas for why to use Stochastic Rounding when doing quantization for training: i) it is an *unbiased* estimator of the real value of x; ii) it avoids *stagnation*, which happens when gradients get rounded to zero and therefore do not contribute to the learning of the neural network.

We have briefly investigated the reasons why stochastic rounding works and we have assembled the results in Figure 1. The figure shows the accuracy when training ResNet18 on CIFAR10 for different configurations of stochastic quantization. The lines labelled "Prob 0.x" means that at any iteration, only a random set of $x\%$ of the tensor values are being stochastically rounded, and the rest are rounding to the nearest value. For example, "Prob 0.2" means that a random 20% of the values are being stochastically rounded. The line labelled "csr_around_zero" means that only the values that would be rounded to zero are stochastically rounded.

The results show that using stochastic rounding on values around zero, which means avoiding *stagnation*, performs as well as using stochastic rounding in all of the values ("Prob 1.0"). When a low percentage of values are stochastically rounded, such as 20%, the training becomes too unstable and achieves significantly lower final accuracy. This indicates that the most likely reason why stochastic rounding works is because it avoids the weights of the neural network from getting stagnated at a certain value due to the gradients being rounded to zero. We believe this provides some insight into the reasons why stochastic rounding seems to be so effective when quantizing gradients specifically.

References

[1] L. Roberts. Picture coding using pseudo-random noise. *IRE Transactions on Information Theory*, 8(2):145–154, 1962.

[2] L. Schuchman. Dither signals and their effect on quantization noise. *IEEE Transactions on Communication Technology*, 12(4):162–165, 1964.