Supplementary Material for NeurIPS Rebuttal

Correpondance between the quantized activation and the quantized objective function

For the answer of the question 1 and 2, we insist the quantization of an objective function and the activation function is almost equivalent.

First, we consider the following quantization for an objective function $f : \mathbf{R}^d \rightarrow \mathbf{R}$ with respect to an activation function as follows:

$$f = \sum_{k=0}^{\infty} f_k b^{n-k} = \sum_{k=0}^{m-1} f_k b^{n-k} + \sum_{k=m}^{\infty} f_k b^{n-k} = f^Q + O(b^m), \quad \because f^Q = \sum_{k=0}^{m} f_k b^{n-k}, \tag{1}$$

where $b \in \mathbf{Z}^+$ denotes the base of the number system for quantization, and $f_k \in \mathbf{Z}^+[0, b)$ denotes a coefficient for b^{n-k}. Hence considereing a binary number system such that $b = 2$, we note that $f_i \in \{0, 1\} \forall i \in \mathbf{Z}^+$.

Assume that there exist a neural network contains l layers which contain the map $\boldsymbol{h}^l : \mathbf{R}^m \rightarrow \mathbf{R}^d$ consisted with the activation function $h_i^l : \mathbf{R} \rightarrow \mathbf{R}^d$ at the i-th node in the l-th Layer such that $h_i^l(y) \triangleq h_i^l(\boldsymbol{w}_i \boldsymbol{h}^{l-1}), \; \boldsymbol{h} = [h_i^l]_{i=1}^d$ where $\boldsymbol{w}_i \in \mathbf{R}^m, \; [w_i^j] \in \mathbf{R}^{d \times m}$ denotes the weight vector for the i th node.

Additionally, we let a quantized activation function ${\boldsymbol{h}^l}_{s_q}^Q$ with the quantization step defined as the reciprocal of the quantization parameter $\boldsymbol{Q}_p^{-1} \in \mathbf{Q}^d$ such that ${\boldsymbol{h}^l}_{s_q}^Q = {\boldsymbol{h}^l}_0^Q + s_q \boldsymbol{Q}_p^{-1}, \; s_q \in \mathbf{Z}$, where each component of $\boldsymbol{Q}_p^{-1}$, i.e. $Q_{p,i}^{-1}$ represents one of the elements to the set $\{-Q_p^{-1}, 0, Q_p^{-1}\}$.

Consider the second-order Taylor series for the objective function f. Particularly, we set $s_q = 1$ for convenience, then

$$\begin{aligned} f({\boldsymbol{h}^l}_1^Q) &= f({\boldsymbol{h}^l}_0^Q) + \nabla_{\boldsymbol{h}} f({\boldsymbol{h}^l}_0^Q) \cdot \boldsymbol{Q}_p^{-1} + \frac{1}{2} \boldsymbol{Q}_p^{-1} \cdot \nabla_h^2 f({\boldsymbol{h}^l}_0^Q) \cdot \boldsymbol{Q}_p^{-1} + O(\|\boldsymbol{Q}_p^{-1}\|^3) \\ &\approx \bar{f}({\boldsymbol{h}^l}_1^Q) + O(Q_p^{-3}), \end{aligned} \tag{3}$$

where $\bar{f}({\boldsymbol{h}^l}_1^Q) \triangleq f({\boldsymbol{h}^l}_0^Q) + \nabla_{\boldsymbol{h}} f({\boldsymbol{h}^l}_0^Q) \cdot \boldsymbol{Q}_p^{-1} + \frac{1}{2} \boldsymbol{Q}_p^{-1} \cdot \nabla_h^2 f({\boldsymbol{h}^l}_0^Q) \cdot \boldsymbol{Q}_p^{-1}$.

Assume that the quantization step Q_p^{-1} is sufficiently small such that $[O(\|Q_p^{-1}\|^3)]^Q = 0$. Then, we calculate the Taylor expansion of the quantization of f such that

$$
\begin{aligned}
f^{Q}(\boldsymbol{h}_{1}^{l\,Q}) &= f^{Q}(\boldsymbol{h}^{l\,Q}_{\,0}) + \left[\nabla_{\boldsymbol{h}} f(\boldsymbol{h}^{l\,Q}_{\,0})\right]^{Q} \cdot \boldsymbol{Q}_{p}^{-1} + \frac{1}{2}\boldsymbol{Q}_{p}^{-1} \cdot \left[\nabla_{h}^{2} f(\boldsymbol{h}^{l\,Q}_{\,0})\right]^{Q} \cdot \boldsymbol{Q}_{p}^{-1} + [O(\|\boldsymbol{Q}_{p}^{-1}\|^{3})]^{Q} \\
&= f(\boldsymbol{h}^{l\,Q}_{\,0}) + \varepsilon_{q} Q_{p}^{-1} + \left(\nabla_{\boldsymbol{h}} f(\boldsymbol{h}^{l\,Q}_{\,0}) + \boldsymbol{\varepsilon}_{q} Q_{p}^{-1}\right) \cdot \boldsymbol{Q}_{p}^{-1} + \frac{1}{2}\boldsymbol{Q}_{p}^{-1} \cdot \left(\nabla_{h}^{2} f(\boldsymbol{h}^{l\,Q}_{\,0}) + \boldsymbol{\varepsilon}_{q} Q_{p}^{-1}\right)^{Q} \cdot \boldsymbol{Q}_{p}^{-1} + [O(\|\boldsymbol{Q}_{p}^{-1}\|^{3})]^{Q} \\
&= f(\boldsymbol{h}_{0}^{l\,Q}) + \nabla_{\boldsymbol{h}} f(\boldsymbol{h}^{l\,Q}_{\,0}) \cdot \boldsymbol{Q}_{p}^{-1} + \frac{1}{2}\boldsymbol{Q}_{p}^{-1} \cdot \nabla_{h}^{2} f(\boldsymbol{h}^{l\,Q}_{\,0}) \cdot \boldsymbol{Q}_{p}^{-1} + \varepsilon_{q} Q_{p}^{-1} + O(|\boldsymbol{\varepsilon}_{q} \cdot \boldsymbol{Q}_{p}^{-1}|^{2}) \\
&= \bar{f}(\boldsymbol{h}^{l\,Q}_{\,1}) + \varepsilon_{q} Q_{p}^{-1} + O(|\boldsymbol{\varepsilon}_{q} \cdot \boldsymbol{Q}_{p}^{-1}|^{2}) \\
&\approx \bar{f}^{Q}(\boldsymbol{h}^{l\,Q}_{\,1}) + O(Q_{p}^{-2}).
\end{aligned}
\tag{4}
$$

As shown in (3) and (4), we get

$$
|f(\boldsymbol{h}^{l\,Q}_{\,1}) - \bar{f}(\boldsymbol{h}^{l\,Q}_{\,1})| \approx |f^{Q}(\boldsymbol{h}_{1}^{l\,Q}) - \bar{f}^{Q}(\boldsymbol{h}^{l\,Q}_{\,1})| + O(Q_{p}^{-2}). \tag{2}
$$

Consequently, if Q_p^{-1} is sufficiently small, the objective function calculated from the quantized activation is almost equivalent to the quantization of the objective function.

This result demonstrates that we can develop a learning equation based on the quantized objective function that is equivalent to the learning equation based on the quantized activation.