Fully Nested Neural Network for Adaptive Compression and Quantization - Proof

Yufei Cui[1], Ziquan Liu[1], Wuguannan Yao[2], Qiao Li[1], Antoni B. Chan[1], Tei-wei Kuo[1], and Chun Jason Xue[1]

[1]Department of Computer Science, City University of Hong Kong
[2]Department of Mathematics, City University of Hong Kong

Proposition 1. *Using the settings and notations in Section 2.3, maximizing the data log-likelihood is equivalent to maximizing the mutual information between ground-truth y and the prediction $\hat{y}$, i.e.,*

$$\max_{\Theta} \mathbb{E}_{\mathbf{x},y\sim p_{\mathbf{x},y}} \log p_{\Theta}(y|\mathbf{x}) \Leftrightarrow \max_{\Theta} \mathbb{I}(y,\hat{y}). \tag{1}$$

Proof. The structure of $p_{\Theta}(y|\mathbf{x})$ is given by

$$\begin{aligned} p_{\Theta}(y|\mathbf{x}) &= \int p_{\Theta}(y|\hat{y}) p_{\Theta}(\hat{y}|\mathbf{x}) d\hat{y} \\ &= \int p_{\Theta}(y|\hat{y}) \delta_{f_{M,\Theta}(\mathbf{x})}(\hat{y}) d\hat{y} \\ &= p_0(y|f_{M,\Theta}(\mathbf{x})) \end{aligned} \tag{2}$$

where $\delta_{f_{M,\Theta}(\mathbf{x})}(\hat{y}) = \delta(\hat{y} - f_{M,\Theta}(\mathbf{x}))$, $f_{M,\Theta}(\mathbf{x}) = \sum_{i=1}^{M} \mathbf{v}_i \sigma(\mathbf{U}_i^T \mathbf{x})$. That is, given an input $\mathbf{x}_0$, $p_{\Theta}(y|\mathbf{x} = \mathbf{x}_0) = p_0(y|\hat{y} = f_{M,\Theta}(\mathbf{x}_0))$ are the same. The joint probability

$$p_{\Theta}(y,\hat{y}) = p_{\Theta}(y|\hat{y}) p_{\Theta}(\hat{y}) = p_0(y|\hat{y}) p_{\Theta}(\hat{y}) \tag{3}$$

The MLE objective is,

$$\max_{\Theta} \mathbb{E}_{\mathbf{x},y\sim p_{\mathbf{x},y}} \log p_{\Theta}(y|\mathbf{x}). \tag{4}$$

The mutual information between ground y and prediction $\hat{y}$ can be written as,

$$\begin{aligned} \max_{\Theta} \mathbb{I}(y,\hat{y}) &= \max_{\Theta} \mathbb{H}(y) - \mathbb{H}(y|\hat{y}) \\ &= \max_{\Theta} -\mathbb{H}(y|\hat{y}) \end{aligned} \tag{5}$$

We can write,

$$\begin{aligned}\mathbb{H}(y|\hat{y}) &= \mathbb{E}_{y,\hat{y}\sim p_\Theta(y,\hat{y})} - \log p_\Theta(y|\hat{y}) \\ &= \mathbb{E}_y \mathbb{E}_{y|\hat{y}} - \log p_0(y|\hat{y}) \\ &= \mathbb{E}_y \mathbb{E}_{\mathbf{x}|y} \mathbb{E}_{\hat{y}|\mathbf{x}} - \log p_0(y|\hat{y}) \end{aligned} \tag{6}$$

where the second equation is due to Eq. 3.

Note that $\hat{y}|\mathbf{x} \sim \delta_{f_{M,\Theta}}$ leads to

$$\mathbb{E}_{\hat{y}|\mathbf{x}} - \log p_0(y|\hat{y}) = -\log p_0(y|f_{M,\Theta}(\mathbf{x})) = -\log p_\Theta(y|\mathbf{x}) \tag{7}$$

where the second equation is due to Eq. 2 again.

Now, we see that

$$\max_\Theta \mathbb{E}_{\mathbf{x},y\sim p_{\mathbf{x},y}} \log p_\Theta(y|\mathbf{x}) \Leftrightarrow \max_\Theta -\mathbb{H}(y|\hat{y}) \Leftrightarrow \max_\Theta \mathbb{I}(y,\hat{y}). \tag{8}$$

Thus Eq. 1 holds. □

Corollary 1. *Using the setting and notations in Section 2.3, by applying ordered dropout on the element of* $\mathbf{v}$*, the maximum likelihood objective (LHS Eq. 1) is equivalent to*

$$\max_\Theta \mathbb{I}_1 + \frac{1}{M}\sum_{c=2}^{M}(M-c)(\mathbb{I}_c - \mathbb{I}_{c-1}), \tag{9}$$

where $\mathbb{I}_c = \mathbb{I}(y, f_c(\mathbf{x}))$*,* $f_c(\mathbf{x}) = \sum_i^c b(\mathbf{x}; \mathbf{U}_i, \mathbf{v}_i) = \sum_i^c \mathbf{v}_i \sigma(\mathbf{U}_i^T \mathbf{x})$.

Proof. By assigning the $\mathcal{C}(\cdot)$ over the indices of elements in $\mathbf{v}$, Eq. 1 is written as

$$\max_\Theta \mathbb{E}_{c\sim\mathcal{C}} \mathbb{E}_{\mathbf{x},y\sim p_{\mathbf{x},y}} \log p_\Theta(y|\mathbf{x}) \Leftrightarrow \max_\Theta \mathbb{E}_{c\sim\mathcal{C}} \mathbb{I}(y, f_c(\mathbf{x})) \tag{10}$$

Let the $\mathcal{C}(\cdot)$ be with uniform probability parameter $\frac{1}{M}$, the objective becomes

$$\max_\Theta \sum_c \frac{1}{M}\mathbb{I}_c, \tag{11}$$

which is expanded as

$$\max_\Theta \quad \mathbb{I}_1 + (1-\frac{1}{M})(\mathbb{I}_2 - \mathbb{I}_1) + (1-\frac{2}{M})(\mathbb{I}_3 - \mathbb{I}_2) \cdots + (1-\frac{M-1}{M})(\mathbb{I}_M - \mathbb{I}_{M-1}) \tag{12}$$

$$\Leftrightarrow \max_\Theta \quad \mathbb{I}_1 + \frac{1}{M}\sum_{c=2}^{M}(M-c)(\mathbb{I}_c - \mathbb{I}_{c-1}) \tag{13}$$

□