Supplementary Material

The main purpose of this supplementary section is to provide proofs for Theorems 1 and 2.

Preliminaries

Here we shall give a proof of (1) as well as preliminary results that will be needed to complete the proofs of Theorems 1 and 2.

Proposition 1. *The fixed point error probability $p_{e,fx}$ is upper bounded as shown in (1).*

Proof. From the definitions of $p_{e,fx}$, $p_{e,fl}$, and p_m,

$$\begin{aligned} &p_{e,fx} \\ &= \Pr\{\hat{Y}_{fx} \neq Y\} \\ &= \Pr\{\hat{Y}_{fx} \neq Y, \hat{Y}_{fx} = \hat{Y}_{fl}\} + \Pr\{\hat{Y}_{fx} \neq Y, \hat{Y}_{fx} \neq \hat{Y}_{fl}\} \\ &= \Pr\{\hat{Y}_{fl} \neq Y, \hat{Y}_{fx} = \hat{Y}_{fl}\} + \Pr\{\hat{Y}_{fx} \neq Y, \hat{Y}_{fx} \neq \hat{Y}_{fl}\} \\ &\leq p_{e,fl} + p_m. \end{aligned}$$

□

Next is a simple result that allows us to replace the problem of upper bounding p_m by several smaller and easier problems by virtue of the union bound.

Proposition 2. *In a M-class classification problem, the total mismatch probability can be upper bounded as follows:*

$$p_m \leq \sum_{j=1}^{M} \sum_{i=1, i\neq j}^{M} \Pr(\hat{Y}_{fx} = i | \hat{Y}_{fl} = j) \Pr(\hat{Y}_{fl} = j) \tag{15}$$

Proof.

$$\begin{aligned} p_m &= \Pr(\hat{Y}_{fx} \neq \hat{Y}_{fl}) = \Pr\left(\bigcup_{j=1}^{M} (\hat{Y}_{fx} \neq j, \hat{Y}_{fl} = j)\right) \\ &\leq \sum_{j=1}^{M} \Pr(\hat{Y}_{fx} \neq j, \hat{Y}_{fl} = j) \\ &= \sum_{j=1}^{M} \Pr(\hat{Y}_{fx} \neq j | \hat{Y}_{fl} = j) \Pr(\hat{Y}_{fl} = j) \\ &= \sum_{j=1}^{M} \Pr\left(\bigcup_{i=1, i\neq j}^{M} \hat{Y}_{fx} = i \middle| \hat{Y}_{fl} = j\right) \Pr(\hat{Y}_{fl} = j) \\ &\leq \sum_{j=1}^{M} \sum_{i=1, i\neq j}^{M} \Pr(\hat{Y}_{fx} = i | \hat{Y}_{fl} = j) \Pr(\hat{Y}_{fl} = j) \end{aligned}$$

where both inequalities are due to the union bound. □

The next result is also straightforward, but quite useful in obtaining upper bounds that are fully determined by averages.

Proposition 3. *Given a random variable X and an event $\mathcal{E}$, we have:*

$$\mathbb{E}\left[X \cdot \mathbb{1}_{\mathcal{E}}\right] = \mathbb{E}\left[X | \mathcal{E}\right] \Pr(\mathcal{E}) \tag{16}$$

where $\mathbb{1}_{\mathcal{E}}$ denotes the indicator function of the event $\mathcal{E}$.

Proof. By the law of total expectation,

$$\begin{aligned} &\mathbb{E}\left[X \cdot \mathbb{1}_{\mathcal{E}}\right] \\ &= \mathbb{E}\left[X \cdot \mathbb{1}_{\mathcal{E}} \mid \mathcal{E}\right] \Pr(\mathcal{E}) + \mathbb{E}\left[X \cdot \mathbb{1}_{\mathcal{E}} \mid \mathcal{E}^c\right] \Pr(\mathcal{E}^c) \\ &= \mathbb{E}\left[X \cdot 1 \mid \mathcal{E}\right] \Pr(\mathcal{E}) + \mathbb{E}\left[X \cdot 0 \mid \mathcal{E}^c\right] \Pr(\mathcal{E}^c) \\ &= \mathbb{E}\left[X | \mathcal{E}\right] \Pr(\mathcal{E}). \end{aligned}$$

□

Proof of Theorem 1

Let us define $p_{m,j\to i}$ for $i \neq j$ as follows.

$$p_{m,j\to i} = \Pr\{\hat{Y}_{fx} = i \mid \hat{Y}_{fl} = j\} \tag{17}$$

We first prove the following Lemma.

Lemma 1. *Given B_X and B_F, if the output of the floating-point network is $\hat{Y}_{fl} = j$, then that of the fixed-point network would be $\hat{Y}_{fx} = i$ with a probability upper bounded as follows:*

$$\begin{aligned} p_{m,j\to i} \leq\ & \frac{\Delta_A^2}{24} \mathbb{E}\left[\frac{\sum_{h\in\mathcal{A}} \left|\frac{\partial(Z_i - Z_j)}{\partial A_h}\right|^2}{|Z_i - Z_j|^2} \middle| \hat{Y}_{fl} = j\right] \\ &+ \frac{\Delta_W^2}{24} \mathbb{E}\left[\frac{\sum_{h\in\mathcal{W}} \left|\frac{\partial(Z_i - Z_j)}{\partial w_h}\right|^2}{|Z_i - Z_j|^2} \middle| \hat{Y}_{fl} = j\right]. \end{aligned} \tag{18}$$

Proof. We can claim that, if $i \neq j$:

$$p_{m,j\to i} \leq \Pr\{Z_i + q_{Z_i} > Z_j + q_{Z_j} \mid \hat{Y}_{fl} = j\} \tag{19}$$

where the equality holds for $M = 2$.

From the law of total probability,

$$p_{m,j\to i} \leq \int f_{\mathbf{X}}(\mathbf{x}) \Pr\left(z_i + q_{z_i} > z_j + q_{z_j} \mid \hat{Y}_{fl} = j, \mathbf{x}\right) d\mathbf{x}, \tag{20}$$

where $\mathbf{x}$ denotes the input of the network, or equivalently an element from the dataset and $f_{\mathbf{X}}()$ is the distribution of the input data. But for one specific $\mathbf{x}$ given $\hat{Y}_{fl} = j$, we have:

$$\Pr\left(z_i + q_{z_i} > z_j + q_{z_j}\right) = \frac{1}{2}\Pr\left(\left|q_{z_i} - q_{z_j}\right| > |z_j - z_i|\right)$$

where the $\frac{1}{2}$ term is due to the symmetry of the distribution of the quantization noise around zero per output. By (7), we can claim that

$$q_{z_i} - q_{z_j} = \sum_{h\in\mathcal{A}} q_{a_h}\frac{\partial(z_i - z_j)}{\partial a_h} + \sum_{h\in\mathcal{W}} q_{w_h}\frac{\partial(z_i - z_j)}{\partial w_h}. \tag{21}$$

Note that $q_{z_i} - q_{z_j}$ is a zero mean random variable with the following variance

$$\frac{\Delta_A^2}{12}\sum_{h\in\mathcal{A}}\left|\frac{\partial(z_i - z_j)}{\partial a_h}\right|^2 + \frac{\Delta_W^2}{12}\sum_{h\in\mathcal{W}}\left|\frac{\partial(z_i - z_j)}{\partial w_h}\right|^2.$$

By Chebyshev's inequality, we obtain

$$\Pr\left(z_i + q_{z_i} > z_j + q_{z_j}\right) \leq \frac{\Delta_A^2\sum_{h\in\mathcal{A}}\left|\frac{\partial(z_i - z_j)}{\partial a_h}\right|^2 + \Delta_W^2\sum_{h\in\mathcal{W}}\left|\frac{\partial(z_i - z_j)}{\partial w_h}\right|^2}{24\left|z_i - z_j\right|^2}. \tag{22}$$

From (20) and (22), we can derive (18). □

Plugging (18) of Lemma 1 into (15) and using (16),

$$p_m \leq \sum_{j=1}^{M}\sum_{i=1,i\neq j}^{M}\left(\frac{\Delta_A^2}{24}\mathbb{E}\left[\frac{\sum_{h\in\mathcal{A}}\left|\frac{\partial(Z_i - Z_j)}{\partial A_h}\right|^2}{|Z_i - Z_j|^2}\mathbb{1}_{\hat{Y}_{fl}=j}\right] + \frac{\Delta_W^2}{24}\mathbb{E}\left[\frac{\sum_{h\in\mathcal{W}}\left|\frac{\partial(Z_i - Z_j)}{\partial w_h}\right|^2}{|Z_i - Z_j|^2}\mathbb{1}_{\hat{Y}_{fl}=j}\right]\right) \tag{23}$$

which can be simplified into (8) in Theorem 1.

Proof of Theorem 2

We start with the following lemma.

Lemma 2. *Given B_A and B_W, $p_{m,j\to i}$ is upper bounded as follows:*

$$p_{m,j\to i} \leq \mathbb{E}\left[e^{-T\cdot V}\prod_{h\in\mathcal{A}}\frac{\sinh\left(T\cdot D_{A,h}\right)}{T\cdot D_{A,h}}\cdot \prod_{h\in\mathcal{W}}\frac{\sinh\left(T\cdot D_{W,h}\right)}{T\cdot D_{W,h}}\middle|\hat{Y}_{fl} = j\right] \tag{24}$$

where $T = \frac{3V}{\sum_{h\in\mathcal{A}}\Delta_{A,h}^2 + \sum_{h\in\mathcal{W}}\Delta_{W,h}^2}$, $V = Z_j - Z_i$, $D_{A,h} = \frac{\Delta_A}{2}\cdot\frac{\partial(Z_i - Z_j)}{\partial A_h}$, *and* $D_{W,h} = \frac{\Delta_W}{2}\cdot\frac{\partial(Z_i - Z_j)}{\partial W_h}$.

Proof. The setup is similar to that of Lemma 1. Denote $v = z_j - z_i$. By the Chernoff bound,

$$\Pr\left(q_{z_i} - q_{z_j} > v\right) \leq e^{-tv}\mathbb{E}\left[e^{t(q_{z_i} - q_{z_j})}\right]$$

for any $t > 0$. Because quantizations noise terms are independent, by (21),

$$\mathbb{E}\left[e^{t(q_{z_i} - q_{z_j})}\right] = \prod_{h\in\mathcal{A}}\mathbb{E}\left[e^{tq_{a_h}d'_{a_h}}\right]\prod_{h\in\mathcal{W}}\mathbb{E}\left[e^{tq_{w_h}d'_{w_h}}\right]$$

where $d'_{a_h} = \frac{\partial(z_i - z_j)}{\partial a_h}$ and $d'_{w_h} = \frac{\partial(z_i - z_j)}{\partial w_h}$. Also, $\mathbb{E}\left[e^{tq_{a_h}d'_{a_h}}\right]$ is given by

$$\begin{aligned}\mathbb{E}\left[e^{tq_{a_h}d'_{a_h}}\right] &= \frac{1}{\Delta_A}\int_{-\frac{\Delta_A}{2}}^{\frac{\Delta_A}{2}} e^{tq_{a_h}d'_{a_h}}\,dq_{a_h} \\ &= \frac{2}{td'_{a_h}\Delta_A}\sinh\left(\frac{td'_{a_h}\Delta_A}{2}\right) \\ &= \frac{\sinh\left(td_{a_h}\right)}{td_{a_h}}\end{aligned}$$

where $d_{a_h} = \frac{d'_{a_h}\Delta_A}{2}$. Similarly, $\mathbb{E}\left[e^{tq_{w_h}d'_{w_h}}\right] = \frac{\sinh\left(td_{w_h}\right)}{td_{w_h}}$ where $d_{w_h} = \frac{d'_{w_h}\Delta_W}{2}$.

Hence,

$$\Pr\left(q_{z_i} - q_{z_j} > v\right) \leq e^{-tv}\prod_{h\in\mathcal{A}}\frac{\sinh\left(td_{a,h}\right)}{td_{a,h}}\prod_{h\in\mathcal{W}}\frac{\sinh\left(td_{w,h}\right)}{td_{w,h}}. \tag{25}$$

By taking logarithms, the right-hand-side is given by

$$-tv + \sum_{h\in\mathcal{A}}\left(\ln\sinh\left(td_{a,h}\right) - \ln\left(td_{a,h}\right)\right) + \sum_{h\in\mathcal{W}}\left(\ln\sinh\left(td_{w,h}\right) - \ln\left(td_{w,h}\right)\right).$$

This term corresponds to a linear function of t added to a sum of log-moment generating functions. It is hence convex in t. By taking derivative with respective to t and setting to zero,

$$v+\frac{|\mathcal{A}|+|\mathcal{W}|}{t}=\sum_{h\in\mathcal{A}}\frac{d_{a,h}}{\tanh(td_{a,h})}+\sum_{h\in\mathcal{W}}\frac{d_{w,h}}{\tanh(td_{w,h})}.$$

But $\tanh(x)=x-\frac{1}{3}x^3+\mathbf{o}\left(x^5\right)$, so dropping fifth order terms yields:

$$v+\frac{|\mathcal{A}|+|\mathcal{W}|}{t}= \sum_{h\in\mathcal{A}}\frac{1}{t(1-\frac{(td_{a,h})^2}{3})}+\sum_{h\in\mathcal{W}}\frac{1}{t(1-\frac{(td_{w,h})^2}{3})}.$$

Note, for the terms inside the summations, we divided numerator and denominator by $d_{a,h}$ and $d_{w,h}$, respectively, then factored the denominator by t. Now, me multiply both sides by t to get:

$$tv+|\mathcal{A}|+|\mathcal{W}|= \sum_{h\in\mathcal{A}}\frac{1}{1-\frac{(td_{a,h})^2}{3}}+\sum_{h\in\mathcal{W}}\frac{1}{1-\frac{(td_{w,h})^2}{3}}.$$

Also $\frac{1}{1-x^2}=1+x^2+\mathbf{o}(x^4)$, so we drop fourth order terms:

$$tv+|\mathcal{A}|+|\mathcal{W}| =\sum_{h\in\mathcal{A}}\left(1+\frac{(td_{a,h})^2}{3}\right)+\sum_{h\in\mathcal{W}}\left(1+\frac{(td_{w,h})^2}{3}\right)$$

which yields:

$$t=\frac{3v}{\sum_{h\in\mathcal{A}}(d_{a,h})^2+\sum_{h\in\mathcal{W}}(d_{w,h})^2} \tag{26}$$

By plugging (25) into (26) and using the similar method of Lemma 1, we can derive (24) of Lemma 2. □

Theorem 2 is obtained by plugging (24) of Lemma 2 into (15) and using (16). Of course, $D_{A_h}^{(i,j)}$ is the random variable of $d_{a,h}$ when $\hat{y}_{fx}=i$ and $\hat{y}_{fl}=j$, and the same applies to $D_{w_h}^{(i,j)}$ and $d_{w,h}$. We dropped the superscript (i,j) in the Lemma as it was not needed for the consistency of the definitions.