A. Appendix

A.1. Lipschitz Constants

The Lipschitz constant describes: when input changes, how much does the output change correspondingly. For a function \( f : X \to Y \), if it satisfies
\[
\| f(x_1) - f(x_2) \|_Y \leq L \| x_1 - x_2 \|_X, \quad \forall x_1, x_2 \in X
\]
for \( L \geq 0 \), and norms \( \| \cdot \|_X \) and \( \| \cdot \|_Y \) on their respective spaces, then we call \( f \) Lipschitz continuous and \( L \) is the known as the Lipschitz constant of \( f \).

For a one layer network, full precision network \( f_p \) has Lipschitz constant \( L \), which satisfies
\[
L \leq C_\sigma \| W_p \| \text{ for } C_\sigma = \frac{d\sigma}{dx}.
\]
This bound is immediate from the fact that \( \nabla f_p(x) = \sigma'((W_p)x) \cdot [W_{1,1} \ldots W_{d,1}] \), and \( L \leq \max_x \| \nabla f_p(x) \| \).

A.2. Proofs and Additional Lemmas

Lemma 1. Let \( f_p \) be an \( m \) layer network, and each layer has Lipschitz constant \( L_i \). Assume that quantizing each layer leads to a maximum pointwise error of \( \delta_i \), and results in a quantized \( m \) layer network \( f_q \). Then for any two points \( x, y \in X \), \( f_q \) satisfies
\[
\| f_q(x) - f_q(y) \| < \left( \prod_{j=1}^{m} L_j \right) \| x - y \| + 2\Delta_{m,L},
\]
where \( \Delta_{m,L} = \delta_m + \sum_{i=1}^{m-1} \left( \prod_{j=i+1}^{m} L_j \right) \delta_i \).

Proof of Lemma 1. Let \( \phi_q^{(i)} \) be the quantized \( i \)th layer of the network. From Section A.1, we know that
\[
\| \phi_q^{(i)}(x) - \phi_q^{(i)}(y) \| \leq L_i \| x - y \| + 2\delta_i.
\]
Similarly, we know that feeding in the previous layer’s quantized output yields
\[
\| \phi_q^{(2)} \circ \phi_q^{(1)}(x) - \phi_q^{(2)} \circ \phi_q^{(1)}(y) \| \leq L_2 \| \phi_q^{(1)}(x) - \phi_q^{(1)}(y) \| + 2\delta_2
\]
\[
\leq L_2 L_1 \| x - y \| + 2L_2 \delta_1 + 2\delta_2.
\]
By chaining together the \( i \) layers inductively up to \( m \), we complete the desired inequality. \( \square \)

Proof of Theorem 2. We know that \( \| \phi_q^{(1)}(x) - \phi^{(1)}(x) \| < \delta_1 \). This means \( \phi^{(2)} \) receives different inputs depending on whether \( \phi^{(1)} \) was quantized or not, and thus requires the Lipschitz bound. Thus
\[
\| \phi_q^{(2)}(\phi_q^{(1)}(x)) - \phi^{(2)}(\phi^{(1)}(x)) \| \leq \| \phi_q^{(2)}(\phi_q^{(1)}(x)) - \phi_q^{(2)}(\phi^{(1)}(x)) \| + \| \phi_q^{(2)}(\phi^{(1)}(x)) - \phi^{(2)}(\phi^{(1)}(x)) \|
\]
\[
\leq \left( L_2 \| \phi_q^{(1)}(x) - \phi^{(1)}(x) \| + 2\delta_2 \right) + \delta_2
\]
\[
\leq 2L_2 \delta_1 + 3\delta_2,
\]
where the second inequality comes from Lemma 1. Chaining the argument for the \( i \)th layer inductively up to \( m \), we arrive at the desired inequality. \( \square \)

Proof of Theorem 2. From the guarantee of Lemma 1, we know
\[
\| f_q(x + \eta) - f_q(x) \| \leq L \| (x + \eta) - x \| + 2\Delta_{m,L}.
\]
If we consider a full precision network $f_{fp}$ that classifies $x_i$ correctly with output margin $r_i > 0$, then we must simply apply a triangle inequality to attain

$$
\|f_q(x_i + \eta) - f_{fp}(x_i)\| \leq \|f_q(x_i + \eta) - f_q(x_i)\| + \|f_q(x_i) - f_{fp}(x_i)\|
$$

$$\leq L \|(x_i + \eta) - x_i\| + 2\Delta_m.L + 3\Delta_m.L.
$$

Thus for $\eta$ such that $\|\eta\| < \frac{r_i - 5\Delta_m.L}{L}$, we will attain $\|f_q(x_i + \eta) - f_{fp}(x_i)\| < r_i$.

Since we also have that $\|z\|_\infty \leq \|z\|_2$ for any $z \in \mathbb{R}^X$, this means that $\|f_q(x_i + \eta) - f_{fp}(x_i)\|_\infty < r_i$. If $f_{fp}$ classifies $x_i$ as class $k$, this means that

$$
f_{fp}(x_i)_k - f_{fp}(x_i)_j \geq 2r_i, \forall j \neq k.
$$

By the triangle inequality, we get

$$
f_q(x_i + \eta)_k - f_q(x_i + \eta)_j = f_q(x_i + \eta)_k - f_q(x_i + \eta)_j + f_{fp}(x_i)_k + f_{fp}(x_i)_j
$$

$$\geq (f_q(x_i + \eta)_k - f_{fp}(x_i)_k) - (f_q(x_i + \eta)_j - f_{fp}(x_i)_j) + (f_{fp}(x_i)_k - f_{fp}(x_i)_j)
$$

$$> -r_i - r_j + 2r_i
$$

$$\geq 0.
$$

Since this difference is strictly greater than 0, $f_q$ classifies $x + \eta$ correctly.

**Proof of Theorem** Let $\hat{y}_{i,fp}$ be the estimated class of $x_i$ using $f_{fp}$ and $\hat{y}_{i,q}$ be the estimated class of $x_i$ using $f_q$. We use basic probabilistic bounds (where the probability is a uniform distribution over the dataset) to arrive at

$$
e_q = \Pr(\hat{y}_{i,q} \neq y_i) = \Pr(\hat{y}_{i,q} \neq y_i \text{ and } \hat{y}_{i,fp} \neq y_i) + \Pr(\hat{y}_{i,q} \neq y_i \text{ and } \hat{y}_{i,fp} = y_i)
$$

$$\leq \Pr(\hat{y}_{i,fp} = y_i) + \Pr(\hat{y}_{i,q} \neq y_i \text{ and } \hat{y}_{i,fp} = y_i)
$$

$$\leq \epsilon_{fp} + \Pr(\hat{y}_{i,fp} = y_i) \Pr(\hat{y}_{i,q} \neq \hat{y}_{i,fp} | \hat{y}_{i,fp} = y_i)
$$

$$\leq \epsilon_{fp} + (1 - \epsilon_{fp}) \Pr(\hat{y}_{i,q} \neq \hat{y}_{i,fp} | \hat{y}_{i,fp} = y_i)
$$

$$= \epsilon_{fp} + (1 - \epsilon_{fp}) (1 - \Pr(\hat{y}_{i,q} = \hat{y}_{i,fp} | \hat{y}_{i,fp} = y_i))
$$

All that remains is lower bounding the final conditional probability of matching. However, this can be done using Theorem 2. We know that $\hat{y}_{i,q} = \hat{y}_{i,fp}$ so long as $\|f_q(x_i) + f_{fp}(x_i)\|_\infty < r_i$. From Theorem 2, a sufficient condition for this is for $r_i - 5\Delta_m.L > 0$, as this implies one can construct a neighborhood of positive radius $\|\eta\| < \frac{r_i - 5\Delta_m.L}{L}$ such that $\|f_q(x_i + \eta) + f_{fp}(x_i)\|_\infty < r_i$. In particular, this implies $\|f_q(x_i) + f_{fp}(x_i)\|_\infty < r_i$ by choosing $\eta = 0$. This gives us

$$\Pr(\hat{y}_{i,q} = \hat{y}_{i,fp} | \hat{y}_{i,fp} = y_i) = \Pr(\|f_q(x_i) + f_{fp}(x_i)\|_\infty < r_i | \hat{y}_{i,fp} = y_i)
$$

$$\geq \Pr(3\delta \geq 0, \forall \|\eta\| < \|\hat{y}_{i,q} - f_{fp}(x_i)\|_\infty < r_i | \hat{y}_{i,fp} = y_i)
$$

$$\geq \Pr \left( \frac{r_i - 5\Delta_m.L}{L} > 0 | \hat{y}_{i,fp} = y_i \right)
$$

$$= \mathbb{E}_{x_i \in X} \left[ 1_{r_i > 5\Delta_m.L} | \hat{y}_{i,fp} = y_i \right].
$$

Combining these terms, we arrive at

$$e_q \leq \epsilon_{fp} + (1 - \epsilon_{fp}) \left( 1 - \mathbb{E}_{x_i \in X} \left[ 1_{r_i > 5\Delta_m.L} | \hat{y}_{i,fp} = y_i \right] \right)
$$

$$= \epsilon_{fp} + (1 - \epsilon_{fp}) \mathbb{E}_{x_i \in X} \left[ 1_{r_i \leq 5\Delta_m.L} | \hat{y}_{i,fp} = y_i \right].
$$