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

$$\|f(x_1) - f(x_2)\|_Y \leq L\|x_1 - x_2\|_X, \ \forall x_1, x_2 \in X$$

for $L \geq 0$, and norms $\|\cdot\|_X$ and $\|\cdot\|_Y$ on their respective spaces, then we call h Lipschitz continuous and L is the known as the Lipschitz constant of h.

For a one layer network, full precision network f_{fp} has Lipschitz constant L, which satisfies

$$L \leq C_\sigma \|W_{fp}\| \text{ for } C_\sigma = \frac{d\sigma}{dx}.$$

This bound is immediate from the fact that $\nabla f_{fp}(x) = \sigma'(W_{fp}x) \cdot \begin{bmatrix} W_{\cdot,1} & \dots & W_{\cdot,d} \end{bmatrix}$, and $L \leq \max_x \|\nabla f_{fp}(x)\|$.

A.2. Proofs and Additional Lemmas

Lemma 1. *Let f_{fp} 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 δ_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

$$\|\phi_q^{(i)}(x) - \phi_q^{(i)}(y)\| < L_i\|x - y\| + 2\delta_i.$$

Similarly, we know that feeding in the previous layer's quantized output yields

$$\begin{aligned} \|\phi_q^{(2)} \circ \phi_q^{(1)}(x) - \phi_q^{(2)} \circ \phi_q^{(1)}(y)\| &\leq L_2\|\phi_q^{(1)}(x) - \phi_q^{(1)}(y)\| + 2\delta_2 \\ &\leq L_2 L_1\|x - y\| + 2L_2\delta_1 + 2\delta_2. \end{aligned}$$

By chaining together the i layers inductively up to m, we complete the desired inequality. □

Proof of Theorem 1. 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

$$\begin{aligned} \|\phi_q^{(2)}((\phi_q^{(1)}(x)) - \phi^{(2)}(\phi^{(1)}(x))\| &\leq \|\phi_q^{(2)}(\phi_q^{(1)}(x)) - \phi_q^{(2)}(\phi^{(1)}(x))\| + \|\phi_q^{(2)}(\phi^{(1)}(x)) - \phi^{(2)}(\phi^{(1)}(x))\| \\ &\leq \left(L_2\|\phi_q^{(1)}(x) - \phi^{(1)}(x)\| + 2\delta_2\right) + \delta_2 \\ &\leq 2L_2\delta_1 + 3\delta_2, \end{aligned}$$

where the second ineuqlity comes from Lemma 1. Chaining the argument for the i^{th} layer inductively up to m, we arrive at the desired inequality. □

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

$$\begin{aligned}\|f_q(x_i+\eta)-f_{fp}(x_i)\| &\leq \|f_q(x_i+\eta)-f_q(x_i)\|+\|f_q(x_i)-f_{fp}(x_i)\| \\ &\leq L\|(x_i+\eta)-x_i\|+2\Delta_{m,L}+3\Delta_{m,L}.\end{aligned}$$

Thus for η 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 $z\in\mathbb{R}^K$, 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

$$\begin{aligned}f_q(x_i+\eta)_k-f_q(x_i+\eta)_j &= f_q(x_i+\eta)_k-f_q(x_i+\eta)_j\pm f_{fp}(x_i)_k\pm f_{fp}(x_i)_j \\ &=(f_q(x_i+\eta)_k-f_{fp}(x_i)_k)-(f_q(x_i+\eta)_j-f_{fp}(x_i)_j)+(f_{fp}(x_i)_k-f_{fp}(x_i)_j) \\ &>-r_i-r_i+2r_i \\ &\geq 0.\end{aligned}$$

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

Proof of Theorem 3. Let $\widehat{y}_{i,fp}$ be the estimated class of x_i using f_{fp} and $\widehat{y}_{i,q}$ 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

$$\begin{aligned}e_q &= \Pr(\widehat{y}_{i,q}\neq y_i) \\ &=\Pr(\widehat{y}_{i,q}\neq y_i \text{ and } \widehat{y}_{i,fp}\neq y_i)+\Pr(\widehat{y}_{i,q}\neq y_i \text{ and } \widehat{y}_{i,fp}=y_i) \\ &\leq\Pr(\widehat{y}_{i,fp}\neq y_i)+\Pr(\widehat{y}_{i,fp}=y_i \text{ and } \widehat{y}_{i,q}\neq\widehat{y}_{i,fp}) \\ &\leq e_{fp}+\Pr(\widehat{y}_{i,fp}=y_i)\Pr(\widehat{y}_{i,q}\neq\widehat{y}_{i,fp}|\widehat{y}_{i,fp}=y_i) \\ &\leq e_{fp}+(1-e_{fp})\Pr(\widehat{y}_{i,q}\neq\widehat{y}_{i,fp}|\widehat{y}_{i,fp}=y_i) \\ &=e_{fp}+(1-e_{fp})(1-\Pr(\widehat{y}_{i,q}=\widehat{y}_{i,fp}|\widehat{y}_{i,fp}=y_i))\end{aligned}$$

All that remains is lower bounding the final conditional probability of matching. However, this can be done using Theorem 2. We know that $\widehat{y}_{i,q}=\widehat{y}_{i,fp}$ 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

$$\begin{aligned}\Pr(\widehat{y}_{i,q}=\widehat{y}_{i,fp}|\widehat{y}_{i,fp}=y_i) &= \Pr(\|f_q(x_i)+f_{fp}(x_i)\|_\infty<r_i|\widehat{y}_{i,fp}=y_i) \\ &\geq\Pr(\exists\delta\geq 0,\forall\|\eta\|<\delta,\|f_q(x_i+\eta)+f_{fp}(x_i)\|_\infty<r_i|\widehat{y}_{i,fp}=y_i) \\ &\geq\Pr\left(\frac{r_i-5\Delta_{m,L}}{L}>0\middle|\widehat{y}_{i,fp}=y_i\right) \\ &=\mathbb{E}_{x_i\in X}\left[\mathbf{1}_{r_i>5\Delta_{m,L}}\middle|\widehat{y}_{i,fp}=y_i\right].\end{aligned}$$

Combining these terms, we arrive at

$$\begin{aligned}e_q &\leq e_{fp}+(1-e_{fp})\left(1-\mathbb{E}_{x_i\in X}\left[\mathbf{1}_{r_i>5\Delta_{m,L}}\middle|\widehat{y}_{i,fp}=y_i\right]\right) \\ &=e_{fp}+(1-e_{fp})\mathbb{E}_{x_i\in X}\left[\mathbf{1}_{r_i\leq 5\Delta_{m,L}}\middle|\widehat{y}_{i,fp}=y_i\right].\end{aligned}$$

□