ECE 761: Quiz 7

Varsha Pendyala

Fall 2018

Theorem to Prove:

Let $h: [0,1]^d \to R$ be a continuous function. Then for any $\epsilon > 0$ there exists a three-layer ReLU network g such that

$$\left|\left|g - h\right|\right|^2 = \int_{[0,1]^d} |g(x) - h(x)|^2 \, dx \leq \epsilon$$

Proof:

Let $B \subset [0,1]^d$ be any "box" of the form $B = [a_1, b_1] \, X \, [a_2, b_2] \, X \dots [a_d, b_d]$ where $0 \leq a_j < b_j \leq 1$ for j = 1, …, d and $I_B(x) = \mathbf{1}_{\{x \in B\}}$

Assume that there exists a three-layer ReLU network $g_B(x)$ such that $\left|\left|I_B - g_B\right|\right| \leq \epsilon$ for any $\epsilon > 0$ (***It will be proved later)***

$$\left|\left|g - h\right|\right|^2 = \int_{[0,1]^d} |g(x) - h(x)|^2 \, dx$$

$$\sum_j \int_{B_j} |g(x) - h(x)|^2 \, dx$$

Since any continuous function **h** can be approximated using histogram partitioning, for $x \in B_j$:

$$\left|h(x) - h_j\right| \leq \epsilon_j$$

$$where \; h_j = \frac{\int_{B_j} h(x) dx}{\int_{B_j} dx}$$

i.e, h_j is the average of h(x) over B_j

Consider $\left|\left|I_B - g_B\right|\right| \leq \epsilon$:

$$\Rightarrow \forall x \in B_j, |g(x) - 1| \leq \epsilon$$

$$\Rightarrow |h_j||g(x) - 1| \leq |h_j|\epsilon$$

$$\Rightarrow \left|h_j g(x) - h_j\right| \leq \left|h_j\right|\epsilon$$

$$\Rightarrow \left|h_j g(x) - h_j\right|^2 \leq h_j^2 \epsilon^2$$

Let $g(x) = h_j g_B(x),\ \forall\, x \in B_j$:

$\forall\, \boldsymbol{x} \in \boldsymbol{B_j}$:

$$\begin{aligned}|g(x) - h(x)|^2 &= \left|\left(g(x) - h_j\right) - \left(h(x) - h_j\right)\right|^2 \\ &= \left|\left(h_j g_B(x) - h_j\right) - \left(h(x) - h_j\right)\right|^2 \\ &\leq \left|h_j g_B(x) - h_j\right|^2 + \left|h(x) - h_j\right|^2\end{aligned}$$

It implies, $|g(x) - h(x)|^2 \leq h_j^2 \epsilon^2 + \epsilon_j^2$

$$\Rightarrow \int_{B_j} |g(x) - h(x)|^2\, dx \leq \int_{B_j} h_j^2 \epsilon^2 + \epsilon_j^2\ dx$$

$$= \epsilon^2 \int_{B_j} h_j^2\, dx + \int_{B_j} \epsilon_j^2\, dx$$

$$\Rightarrow \sum_j \int_{B_j} |g(x) - h(x)|^2\, dx \leq \sum_j (h_j^2 \epsilon^2 + \epsilon_j^2) \int_{B_j} dx$$

Let

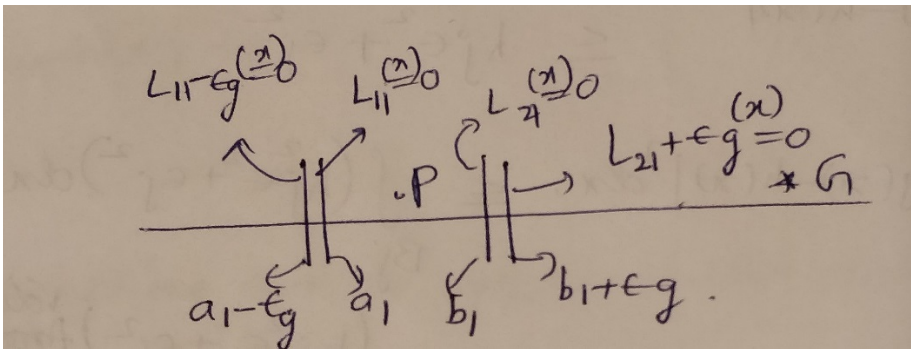
$$\epsilon_M^2 = \epsilon^2 \textstyle\sum_j h_j^2 \int_{B_j} dx + \sum_j \epsilon_j^2 \int_{B_j} dx \qquad \text{— (1)}$$

Proving that there exists a three-layer ReLU network $g_B(x)$ such that $||I_B - g_B|| \leq \epsilon$, for any $\epsilon > 0$:

Partition $[0,1]^d$ into "boxes" of certain width.

Each box, thus becomes a polytope which can be represented by linear inequalities.

Consider **1 dimensional** case:



Those hyperplanes (*lines in 1D case*) are such that:

$$L_{11-\epsilon_g}(P) > 0$$

$$L_{11}(P) > 0$$

$$L_{21}(P) < 0$$

$$L_{21+\epsilon_g}(P) < 0$$

whereas for point G (*i.e for a point outside the boundary)*:

$$L_{11-\epsilon_g}(G) > 0$$

$$L_{11}(G) > 0$$

$$L_{21}(G) > 0$$

$$L_{21+\epsilon_g}(G) > 0$$

Consider

$$L_{11-\epsilon_g} \equiv w_1 x + a_1 + \epsilon_g = 0$$

$$L_{11} \equiv w_1 x + a_1 = 0$$

$$L_{21} \equiv w_1 x + b_1 = 0$$

$$L_{21+\epsilon_g} \equiv w_1 x + b_1 - \epsilon_g = 0$$

and

$$g_1(x) = \frac{\left[\left(f\left(w_1 x + a_1 + \epsilon_g\right) - f(w_1 x + a_1)\right) - \left(f(w_1 x + b_1) - f\left(w_1 x + b_1 - \epsilon_g\right)\right)\right]}{\epsilon_g}$$

where f is a ReLU function.

w_1 is a unit vector (so that it is easy to compute the distance of a point from hyperplane just by substituting the point in hyperplane equation)

i. $\forall\, x$ between the hyperplanes (i.e, points like P):

$$g_1(x) = 1$$

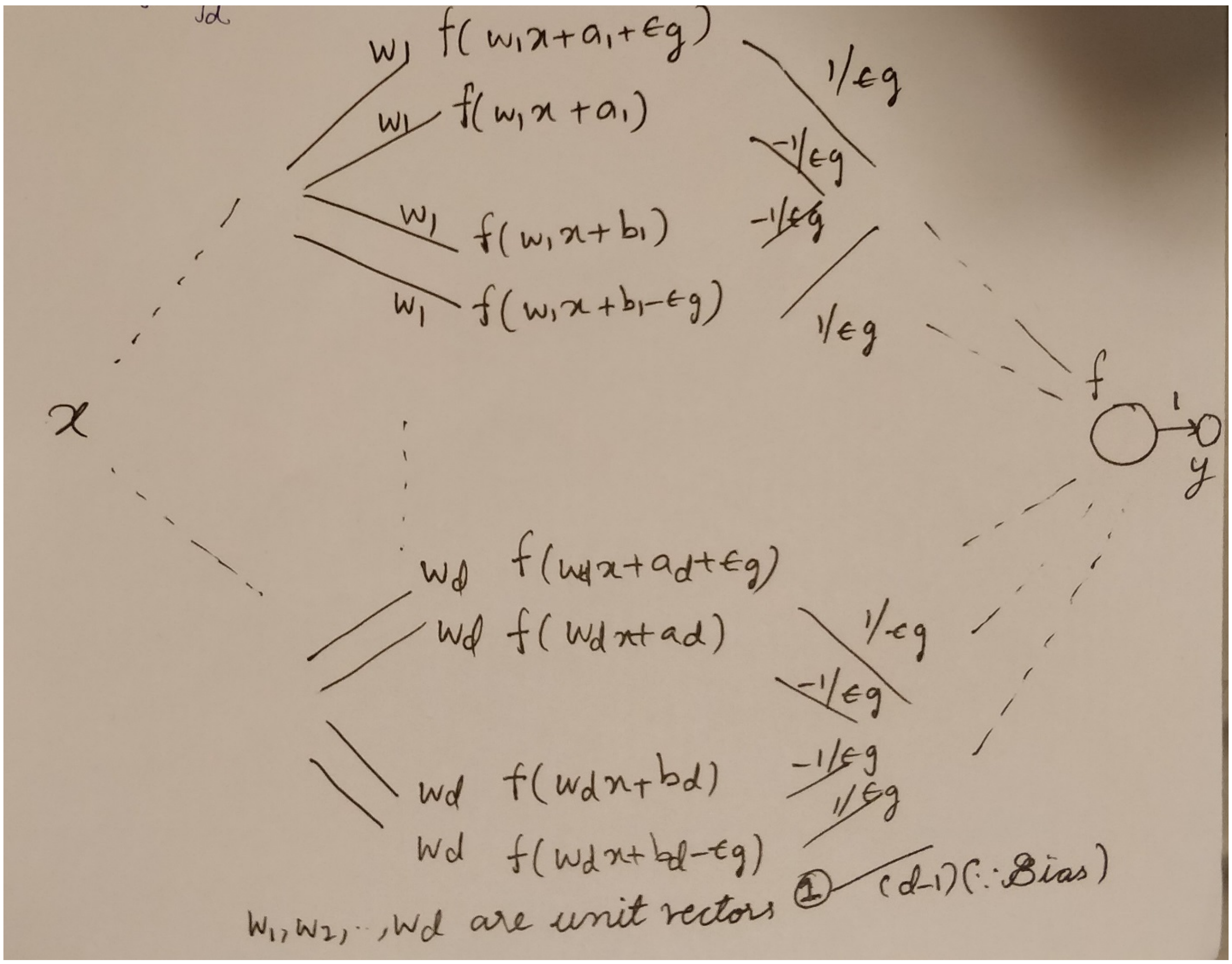
ii. $\forall\, x$ outside the boundary (i.e, points like G):

$$g_1(x) = 0$$

Generalizing to d dimensions:

By observing the 1-D case, we can see that for each box in d-dimensional space, 4 hyperplanes are required to represent both the boundaries along each dimension.

Consider following neural network:



For points inside $\boldsymbol{B_j}$:

Output at node A:

$$f\big(d-(d-1)\big)=1$$

For points outside $\boldsymbol{B_j}$:

Let $D=\{1,2,\dots,d\}$ and $D_s\subset D$ is a proper subset of D.

$\forall\ \boldsymbol{d_j}\in\boldsymbol{D_s}$ **:**

$$\frac{\left[\left(f\left(w_j x + a_j + \epsilon_g\right) - f\left(w_j x + a_j\right)\right) - \left(f\left(w_j x + b_j\right) - f\left(w_j x + b_j - \epsilon_g\right)\right)\right]}{\epsilon_g} = 1$$

$\forall\, \boldsymbol{d_j} \in \boldsymbol{D} - \boldsymbol{D_s}$: **(*it is a non-empty set*)**

$$\frac{\left[\left(f\left(w_j x + a_j + \epsilon_g\right) - f\left(w_j x + a_j\right)\right) - \left(f\left(w_j x + b_j\right) - f\left(w_j x + b_j - \epsilon_g\right)\right)\right]}{\epsilon_g} = 0$$

Output at node A:

$$f\left(\left(\sum\nolimits_{d_j \in D_S} 1 + \sum_{d_j \in D - D_S} 0\right) - (d-1)\right) = 0$$

Thus, above neural network gives zero approximation error for indicator function $\boldsymbol{I_B}$. Hence, $\boldsymbol{\epsilon}$ in equation (1) is "zero".

Thus, equation (1) simplifies to:

$$\epsilon_M^2 = \sum\nolimits_j \epsilon_j^2 \int_{B_j} dx \qquad \_\_\ (2)$$

$\epsilon_j' s$ can be made arbitrarily small by partitioning $[0,1]^d$ to large number of boxes. It implies, ϵ_M can be made arbitrarily smaller.

Thus, there exists a neural network $g(x)$ such that

$g(x) = h_j g_B(x)$ when $x \in B_j$,

for which $\int_{[0,1]^d} |g(x) - h(x)|^2\, dx \leq \epsilon_M^2$ for any $\epsilon_M > 0$

Hence proved Theorem 1.

Corollary to be proved:

Let $h: [0,1]^d \to R$ be a L-Lipschitz function. Then for any $\epsilon > 0$ there exists a three-layer ReLU network g with $N = O\left(\left(\frac{d}{\epsilon^2}\right)^{\frac{d}{2}}\right)$ nodes per layer such that

$$\left|\left|g - h\right|\right|^2 = \int_{[0,1]^d} |g(x) - h(x)|^2 \, dx \leq \epsilon$$

Proof:

For a L-Lipschitz function h, for any $x_1, x_2 \in [0,1]^d$:

$$|h(x_1) - h(x_2)| \leq L\left|\left|x_1 - x_2\right|\right|$$

where $L > 0$ is a constant.

When $x_1, x_2 \in B_j$, and $h(x_2) = h_j$ (It is possible for some $x_2 \in B_j$ from mean value theorem for continuous functions)

$$|h(x_1) - h(x_2)| \leq L\left|\left|x_1 - x_2\right|\right| \leq L.sidelength(B_j)\sqrt{d}$$

$$\Rightarrow \left|h(x) - h_j\right| \leq L.sidelength(Bj)\sqrt{d}, \forall\, x \in B_j$$

Thus equation (2) simplifies to:

$$\epsilon_M^2 = \sum_j \epsilon_j^2 \int_{B_j} dx$$

If all boxes are of same dimensions:

$$\epsilon_M = L.sidelength\sqrt{d} \Rightarrow sidelength = \frac{\epsilon_M}{L\sqrt{d}}$$

$sidelength.M = 1,$ where M is number of boxes along one dimension. Thus $M = \frac{L\sqrt{d}}{\epsilon_M}$

Hence, total number of boxes is $M^d = L^d \left(\frac{d}{\epsilon_M^2}\right)^{\frac{d}{2}}$

To construct $g(x)$ such that $g(x) = h_j g_B(x)$ when $x \in B_j$, we need hidden nodes equal to the number of boxes. Thus the number of nodes needed to achieve following approximation is $O\left(L^d \left(\frac{d}{\epsilon^2}\right)^{\frac{d}{2}}\right)$: $\left|\left|g - h\right|\right|^2 = \int_{[0,1]^d} |g(x) - h(x)|^2 \, dx \leq \epsilon^2$