Quiz 7

Quiz Rules: Due December 12, 2018. You may discuss basic concepts related to the quiz with classmates, but your answers should be developed independently. Typset your solutions, ideally in Latex, and submit a pdf of your solutions.

On my honor, I pledge that I have neither given nor received unpermitted aid on this quiz.

signature:

print name:

For this "take-home" quiz you will prove the following theorem and corollary showing that ReLU neural networks can approximate any continuous function. Your proof should be formal and logically sound, in style of proofs we have discussed in class (e.g., Theorem 1 in `mle_asymp.pdf`, Theorem 1 in `sgd_notes.pdf`, Theorem 1 in `kernels_theory.pdf`, and Theorem 1 in `soft_threshold_analysis.pdf`).

Consider neural networks of the form

$$g(\boldsymbol{x}) \;=\; \boldsymbol{W}_L f(\boldsymbol{W}_{L-1} \cdots f(\boldsymbol{W}_1 \boldsymbol{x} + \boldsymbol{b}_1) + \cdots + \boldsymbol{b}_{L-1}) + \boldsymbol{b}_L$$

where the input $\boldsymbol{x} \in [0,1]^d$, $f(z) = \max(0, z)$ is the Rectified Linear Unit (ReLU) activation function, and L is the number of layers.

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

$$\|g - h\|^2 \;=\; \int_{[0,1]^d} |g(\boldsymbol{x}) - h(\boldsymbol{x})|^2 \, d\boldsymbol{x} \;\leq\; \epsilon \;.$$

Hints: *Recall that continuous functions can be approximated using a histogram partition. Let $B \subset [0,1]^d$ be any "box" of the form $B = [a_1, b_1] \times [a_2, b_2] \times \cdots \times [a_d, b_d]$, where $0 \leq a_j < b_j \leq 1$ for $j = 1, \ldots, d$, and define the function $I_B(\boldsymbol{x}) := \mathbb{1}_{\{\boldsymbol{x} \in B\}}$. Show that for any $\epsilon > 0$ there exists a three-layer ReLU network $g_B(\boldsymbol{x})$ such that $\|I_B - g_B\| \leq \epsilon$.*

Now suppose that h is a L-Lipschitz function; i.e., for any $\boldsymbol{x}_1, \boldsymbol{x}_2 \in [0,1]^d$

$$|h(\boldsymbol{x}_1) - h(\boldsymbol{x}_2)| \;\leq\; L\|\boldsymbol{x}_1 - \boldsymbol{x}_2\| \;,$$

where $L > 0$ is a constant. The constant L controls how rapidly h can vary.

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

$$\|g - h\|^2 \;=\; \int_{[0,1]^d} |g(\boldsymbol{x}) - h(\boldsymbol{x})|^2 \, d\boldsymbol{x} \;\leq\; \epsilon \;.$$

ECE761: Quiz 7

Zirui Tao

December 2018

1 Theorem 1.

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

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

Proof.
Establish the proof by two parts, as referenced by [Tel18]:

1. $\forall B \in [0,1]^d$ there exists a network with two ReLU layers: $g_B = f(W_2 f(W_1 x + b_1) + b_2)$ that approximates the indicator function $I_B := \mathbb{1}_{\{x \in B\}}$ with $\|I_B - g_B\| \leq \tau, \forall \tau > 0$

2. By the conclusion from 1, the last layer $(W_3 * (\cdot) + b_3)$ performs the affine combination layer to compute the as linear combination of piece-wise constant over all boxes, thus completing the proof.

First we establish a lemma that proof the statement of 1.

Lemma 2 *For any* $B \subset [0,1]^d$ *and* $\tau > 0$*, for a function class* $\mathcal{F}$*, there exists a two-layer ReLU networks* $g_B \in \mathcal{F}$ *such that*

$$\|I_B - g_B\|^2 \leq \tau$$

Proof.
It is equivalent to prove: $\|I_B - g_B\|_1 \leq \tau_1$ for some $\tau_1 > 0$.

Construct a set of function g_i for $i \in 1, ..., d$, which can be regarded as the output of the first hidden layer with arbitrary positive δ, $0 \leq a_i \leq b_i \leq 1$:

$$g_i(x) := f(\frac{x_i - a_i}{\delta} + 1) - f(\frac{x_i - a_i}{\delta}) - f(\frac{x_i - b_i}{\delta}) + f(\frac{x_i - b_i - \delta}{\delta})$$

function f is ReLU activation: $f(z) = max(0, z)$

Therefore, we g_i is equivalent to be written as:

$$g_i(x) := \begin{cases} 1 & x_i \in [a_i, b_i] \\ 0 & x_i \in (-\infty, a_i - \delta] \cup (b_i + \delta, \infty] \\ \in [0,1] & otherwise \end{cases}$$

This g_i function, for i $\in \{1, 2, ...d\}$ can be implemented by having 4*d nodes for each box $B \in$[a_i, a_j]^d for $0 \leq a_i \leq b_i \leq 1$ on the first hidden layer and each group of four node corresponds to compute the g_i for input x_i and zero-off other inputs.

Now if the second hidden layer $f(W_2(\cdot) + b_2)$ performs the operation on $g_i$as:

$$g_B := f((\sum_i g_i(x)) - (d - 1))$$

Then, g_B can be written as:

$$g_B(x) := \begin{cases} 1 & x \in \mathbb{R}, \\ 0 & \inf_{y \in B} \|x - y\|_\infty \geq \delta, \\ \in [0,1] & otherwise \end{cases}$$

This can be achieved by setting all the weights that are associated with terms with B W_2 to 1 and bias to be $-(d-1)$.Only $4d$ nodes are used for g_B. Therefore, for total number of N boxes partitioned, there are N copies of such "effective connections" so first hidden layer contains $4 * d * N$ nodes)

Let $\Delta_i := b_i - a_i$,

$$\begin{aligned} \|g_B - I_B\|_1 &\leq vol(\times_{i=1}^d [a_i - \delta, b_i + \delta] \setminus \times_{i=1}^d [a_i, b_i]) \\ &= vol(\times_{i=1}^d [a_i - \delta, b_i + \delta]) - vol(\times_{i=1}^d [a_i, b_i]) \\ &= \prod_{i=1}^d (\Delta_i + 2\delta) - \prod_{i=1}^d (\Delta_i) \\ &\leq \sum_{i=1}^d \binom{d}{i} (2\delta)^i := \xi, \because \Delta_i \leq 1 \end{aligned}$$

Therefore, $\|g_B - I_B\|_1 \to 0$ as $\delta \to 0$, $\exists \delta > 0$ sufficiently small such that $\|g_B - I_B\|_1 \leq \xi \leq \tau_1 = \sqrt{\tau}$, so $\|g_B - I_B\|^2 \leq \tau, \forall \tau > 0$. ■

Then we use Lemma 2 to prove the statement 2.

Proof.

Applying the piecewise constant approximation, let $(B_1, B_2, ...B_N)$ be a partition of $[0,1]^d$ and $h_0 = \sum_i^N \alpha_i I_{B_i}$ be a piecewise constant function so that $\|h_0 - f\|_\infty^2 \leq \epsilon/2$. Let $A := \sum_i |\alpha_i|^2$, ; then $h \in span(\mathscr{F})$ and satisfies

$$\|h_0 - f\|^2 = \int_{[0,1]^d} |h_0(x) - f(x)|^2 d\mathbf{x} \leq \int_{[0,1]^d} \|h_0 - f\|_\infty^2 \, d\mathbf{x} \leq \frac{\epsilon}{2}$$

if $A = 0$ then it suffices to show that:

$$\|h - f\|^2 = \|h_0 - f\|^2 \leq \int_{[0,1]^d} |h_0(x) - f(x)|^2 dx \leq \int_{[0,1]^d} \|h_0 - f\|_\infty^2 \, dx \leq \frac{\epsilon}{2}$$

else if $A \neq 0$, $\frac{\epsilon}{2A} > 0$, $\forall i \in \{1, ..., N\}$, according to Lemma 2, by the assumption on $\mathscr{F}$, select g_B so that $\|g_B - I_{Bi}\|^2 \leq \frac{\epsilon}{2A}$. Let $g := \sum_i |\alpha_i|^2 g_B$,

$$\begin{aligned} \|g - f\|^2 &\leq \|g - h_0\|^2 + \|h_0 - f\|^2 \\ &= \int_{[0,1]^d} |\sum_i \alpha_i g_i - \sum_i \alpha_i I_{Bi}|^2 dx + \int_{[0,1]^d} |h_0(x) - f(x)|^2 dx \\ &\leq \int_{[0,1]^d} \sum_i |\alpha_i|^2 |g_i - I_{Bi}|^2 dx + \int_{[0,1]^d} \|h_0 - f\|_\infty^2 \, dx \\ &\leq \sum_i |a_i|^2 * (\frac{\epsilon}{2A}) + \frac{\epsilon}{2} = \epsilon \end{aligned}$$

■

In conclusion, we can construct a 2-layer RuLU network g_B. In addition, by Lemma 2, we can always control the error between I_B and g_B by arbitrary small positive $\tau := \frac{\epsilon}{2A}$, then the last layer approximates continuous function f by linearly combining the output of each box indicator function approximator g_B through its affine combination $(W_3(\cdot) + b_3)$. We can also control the approximation error between histogram classifier h_0 and continuous function f by taking "sufficient" amount of partitions in $[0,1]^d$ such that $\|h_0 - f\|^2 \leq \frac{\epsilon}{2}$. By triangle inequality, the total approximate error can be bounded by the sum of: 1.approximate error between 3-layer ReLU network g and piecewise constant box function h_0. 2. The error between h_0 and continuous function f. ■

2 Corollary 1. (Changed bounding constant ϵ to ϵ^2 for desired bounding)

Now suppose that h is a L-Lipschitz function; i.e., for any $\mathbf{x}_1, \mathbf{x}_2 \in [0,1]^d$

$$|h(\mathbf{x}_1) - h(\mathbf{x}_2)| \leq L \|\mathbf{x}_1 - \mathbf{x}_2\|,$$

where $L > 0$ is a constant. The constant L controls how rapidly h can vary.

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

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

Proof. The number of nodes per layer is bounded by the max number of nodes between two hidden layers since the input layer is fixed as d and output layer is fixed as 1.

According to the construction illustrated in the proof of Theorem 1, the number of nodes in the hidden layer is $4*d$. Here we find the number of nodes in the second layer, which is number of binning partitions such that $\|g - h\| \leq \epsilon$.

Let $(B_1, B_2, ...B_N)$ be a partition of $[0,1]^d$. Define h_0 as histogram classifier: $h_0 = \sum_i^N \alpha_i I_B$. By tranglular inequality we have:

$$\|g - h\|^2 \leq \|g - h_0\|^2 + \|h_0 - h\|^2$$

Let $\|g - h_0\|^2 \leq \epsilon_1^2$, $\|h_0 - h\|^2 \leq \epsilon_2^2$ for some $\epsilon_1, \epsilon_2 > 0$ and $\epsilon_1^2 + \epsilon_2^2 \leq \epsilon^2$. By Lemma 2, we can have $\|g - h_0\|^2 \leq \epsilon_1^2$, now we analyze the condition for achieving $\|h_0 - h\|^2 \leq \epsilon_2^2 \leq \epsilon^2 - \epsilon_1^2$.

$$\|h_0 - h\|^2 = \int_{[0,1]^d} |h_0(x) - h(x) dx| \leq \int_{[0,1]^d} \|h_0 - h\|_\infty^2 \, dx \leq \|h_0 - h\|_\infty^2 \leq \epsilon_2^2 \leq \epsilon^2 - \delta, \tag{1}$$

Equally partition $[0,1]^d$ to N bins, for each bin B_i, we have:

$$\|B_i\|^2 = d * N^{-2/d} \tag{2}$$

Since h is L-Lipschitz, assuming for each bin, h_0 takes the center value of h as the piece-wise constant $h_0(x) = h(\bar{x}), \bar{x} = \frac{\int_{B_i} h(x)dx}{vol(B_i)}, \forall x \in B_i$. By the L-Lipschitz assumption of h, we can bound the error in each bin as:

$$|h_0(x_1) - h(x_1)|^2 \leq (L \|x_2 - \bar{x}\|)^2 \leq (\frac{L \|B_i\|^2}{2})^2, \forall x_1, x_2 \in B_i \subset [0,1]^d$$

Since the geometry of each bin B_i is identical, then:

$$\|h_0 - h\|_\infty^2 = (\frac{L \|B_i\|}{2})^2 \tag{3}$$

Substitute e.q 2 into e.q 3:

$$\|h_0 - h\|_\infty^2 = (\frac{L}{2})^2 d N^{-\frac{2}{d}} \tag{4}$$

Substitute e.q 4 into e.q 1:

$$\|h_0 - h\|^2 \leq (\frac{L}{2})^2 d N^{-\frac{2}{d}} \leq \epsilon^2 - \epsilon_1^2 < \epsilon^2$$

Therefore, $N > ((\frac{L}{2})^2 \frac{d}{\epsilon^2})^{d/2}$, since approximation error between g and h_0 ($\|g - h_0\|^2$), by Lemma 2, can be bounded by a arbitrarily small positive value (ϵ_1^2), then upon arbitrarily small positive δ, $N \geq \lceil \delta + ((\frac{L}{2})^2 \frac{d}{\epsilon^2})^{d/2} \rceil = O((\frac{d}{\epsilon^2})^{d/2})$, we can bound the approximation error $\|g - h\|^2$ within ϵ^2.

In conclusion, with $O(max(4d, (\frac{d}{\epsilon^2})^{d/2}))$, or, when $(\frac{L}{2})^2(\frac{d}{\epsilon^2})^{d/2} \gg 4d$, that is, $d \gg \frac{16}{L}^{1-\frac{d}{2}}\epsilon^{1+\frac{d}{2}}$, with $O((\frac{d}{\epsilon^2})^{d/2})$ nodes per layer, there exists a three-layer ReLU network g such that:

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

■

References

[Tel18] Matus Telgarsky. Machine Learning Theory 2018 [CS 598 Tel], 2018. URL: http://mjt.web.engr.illinois.edu/courses/mlt-f18/. Last visited on 2018/12/08.