Expressiveness of Rectifier Networks (Supplementary Material)

Xingyuan Pan XPAN@CS.UTAH.EDU
Vivek Srikumar SVIVEK@CS.UTAH.EDU
The University of Utah, Salt Lake City, UT 84112, USA

1. Proof of Theorem 1

In this section, we prove our main theorem, Theorem 1, which is repeated here for convenience.

Theorem 1. *Consider a two-layer rectifier network with n hidden units represented in its general form (Eq. (3)). Then, for any input $\mathbf{x}$, the following conditions are equivalent:*

1. *The network classifies the example $\mathbf{x}$ as positive.*
2. *There exists a subset $\mathcal{S}_1$ of $\mathcal{P}$ such that, for every subset $\mathcal{S}_2$ of $\mathcal{N}$, we have $w_0 + \sum_{k\in\mathcal{S}_1} a_k(\mathbf{x}) - \sum_{k\in\mathcal{S}_2} a_k(\mathbf{x}) \geq 0$.*
3. *For every subset $\mathcal{S}_2$ of $\mathcal{N}$, there exists a subset $\mathcal{S}_1$ of $\mathcal{P}$ such that $w_0 + \sum_{k\in\mathcal{S}_1} a_k(\mathbf{x}) - \sum_{k\in\mathcal{S}_2} a_k(\mathbf{x}) \geq 0$.*

Step 1: Equivalence of conditions 1 and 2

Let us prove that condition 1 implies condition 2 first. We can construct a subset $\mathcal{S}_1^*$ of $\mathcal{P}$

$$\mathcal{S}_1^* = \{k : k \in \mathcal{P} \text{ and } a_k(\mathbf{x}) \geq 0\},$$

such that

$$\sum_{k\in\mathcal{P}} R(a_k(\mathbf{x})) = \sum_{k\in\mathcal{S}_1^*} a_k(\mathbf{x}).$$

The example $\mathbf{x}$ is classified as positive implies that

$$w_0 + \sum_{k\in\mathcal{S}_1^*} a_k(\mathbf{x}) \geq \sum_{k\in\mathcal{N}} R(a_k(\mathbf{x})).$$

For any subset $\mathcal{S}_2$ of $\mathcal{N}$, we have

$$\sum_{k\in\mathcal{N}} R(a_k(\mathbf{x})) \geq \sum_{k\in\mathcal{S}_2} R(a_k(\mathbf{x})) \geq \sum_{k\in\mathcal{S}_2} a_k(\mathbf{x}).$$

Therefore, for any subset $\mathcal{S}_2$ of $\mathcal{N}$,

$$w_0 + \sum_{k\in\mathcal{S}_1^*} a_k(\mathbf{x}) \geq \sum_{k\in\mathcal{S}_2} a_k(\mathbf{x}).$$

Now, we need to show that condition 2 implies condition 1. Assume there is a subset $\mathcal{S}_1$ of $\mathcal{P}$ such that for any subset $\mathcal{S}_2$ of $\mathcal{N}$, $w_0 + \sum_{k\in\mathcal{S}_1} a_k(\mathbf{x}) \geq \sum_{k\in\mathcal{S}_2} a_k(\mathbf{x})$. Let us define a specific subset $\mathcal{S}_2^*$ of $\mathcal{N}$,

$$\mathcal{S}_2^* = \{k : k \in \mathcal{N} \text{ and } a_k(\mathbf{x}) \geq 0\},$$

such that

$$\sum_{k\in\mathcal{N}} R(a_k(\mathbf{x})) = \sum_{k\in\mathcal{S}_2^*} a_k(\mathbf{x}).$$

We know that

$$\sum_{k\in\mathcal{P}} R(a_k(\mathbf{x})) \geq \sum_{k\in\mathcal{S}_1} R(a_k(\mathbf{x})) \geq \sum_{k\in\mathcal{S}_1} a_k(\mathbf{x})$$

and

$$w_0 + \sum_{k\in\mathcal{S}_1} a_k(\mathbf{x}) \geq \sum_{k\in\mathcal{S}_2^*} a_k(\mathbf{x}).$$

Therefore,

$$w_0 + \sum_{k\in\mathcal{P}} R(a_k(\mathbf{x})) \geq \sum_{k\in\mathcal{N}} R(a_k(\mathbf{x}))$$

which means that the decision function y in Eq. (3) is positive.

Step 2: Equivalence of conditions 1 and 3

That condition 1 implies condition 3 holds by virtue of the first part of the previous step. We only need to prove that condition 3 implies 1 here. Assume for all subset $\mathcal{S}_2$ of $\mathcal{N}$ there is a subset $\mathcal{S}_1$ of $\mathcal{P}$ such that $w_0 + \sum_{k\in\mathcal{S}_1} a_k(\mathbf{x}) - \sum_{k\in\mathcal{S}_2} a_k(\mathbf{x}) \geq 0$. Use the same $\mathcal{S}_2^*$ defined in previous step

$$\begin{aligned} w_0 + \sum_{k\in\mathcal{P}} R(a_k(\mathbf{x})) &\geq w_0 + \sum_{k\in\mathcal{S}_1} R(a_k(\mathbf{x})) \\ &\geq w_0 + \sum_{k\in\mathcal{S}_1} a_k(\mathbf{x}) \\ &\geq \sum_{k\in\mathcal{S}_2^*} a_k(\mathbf{x}) \\ &= \sum_{k\in\mathcal{N}} R(a_k(\mathbf{x})) \end{aligned}$$

Therefore, the decision function y in Eq. (3) is positive. □

2. Proof of Theorem 2

The first part of this theorem says that every rectifier network with n hidden units that are all positive or negative can be represented by a threshold network with $2^n - 1$ hidden units. This is a direct consequence of the main theorem.

The second part of the theorem says that for any n, there are families of rectifier networks whose equivalent threshold network will need an exponential number of hidden threshold units. We prove this assertion constructively by providing one such rectifier network. Consider the decision function of a two-layer rectifier network

$$y = \text{sgn}[-1 + R(x_1) + R(x_2) + \cdots + R(x_n)]$$

where x_i is the i^{th} component of the input vector (recall that the dimensionality of the input $d \geq n$). From Theorem 1 the decision boundary of this network can be determined by $2^n - 1$ hyperplanes of the form $-1 + \sum_{i \in \mathcal{S}} x_i = 0$, each of which corresponds a *non-empty* subset $\mathcal{S} \in [n]$. The output y is positive if any of these hyperplanes classify the example as positive.

To prove that we need at least $2^n - 1$ threshold units to represent the same decision boundary, it suffices to prove that each of the $2^n - 1$ hyperplanes is needed to define the decision boundary.

Consider any hyperplane defined by *non-empty* subset $\mathcal{S} \in [n]$ whose cardinality $s = |\mathcal{S}|$. Let $\bar{\mathcal{S}}$ be the complement of $\mathcal{S}$. Consider an input vector $\mathbf{x}$ that satisfies the following condition if $s > 1$:

$$\begin{cases} 1/s < x_i < 1/(s-1), & \text{if } i \in \mathcal{S} \\ x_i < -1/(s-1), & \text{if } i \in \bar{\mathcal{S}} \end{cases}$$

For those subsets $\mathcal{S} = \{x_j\}$ whose cardinality is one, we can let $x_j > 1$ and all the other components $x_i < -x_j$. Clearly, the rectifier network will classify the example as positive. Furthermore, by construction, it is clear that $-1 + \sum_{i \in \mathcal{S}} x_i > 0$, and for all other subset $\mathcal{S}' \in [n]$, $-1 + \sum_{i \in \mathcal{S}'} x_i < 0$. In other words, *only* the selected hyperplane will classify the input as a positive one. That is, this hyperplane is required to define the decision boundary because without it, the examples in the above construction will be incorrectly classified by the threshold network.

Therefore we we have identified the decision boundary of the given rectifier network as an polytope with *exactly* $2^n - 1$ faces, by constructing $2^n - 1$ hyperplanes using $2^n - 1$ *non-empty* subsets $\mathcal{S} \in [n]$. To complete the proof, we need to show that other construction methods cannot do better than our construction, i.e., achieving same decision boundary with less hidden threshold units. To see this, note that for each face of the decision polytope, one needs a hidden threshold unit to represent it. Therefore no matter how we construct the threshold network, we need at least $2^n - 1$ hidden threshold units. □

3. Proof of Lemma 2

In this section, we provide a simple example of a two-layer threshold network, whose decision boundary cannot be represented by a two-layer rectifier network with fewer hidden units. Consider a threshold network

$$y = \text{sgn}[n - 1 + \text{sgn}(x_1) + \text{sgn}(x_2) + \cdots + \text{sgn}(x_n)],$$

where x_i is the i^{th} component of the input vector (here we assume the dimensionality of the input $d \geq n > 1$). It is easy to see that the decision function of this network is positive, if, and only if at least one of the component x_i is non-negative. From a geometric point of view, each x_i defines a hyperplane $x_i = 0$. Let H_1 be the set of n such hyperplanes,

$$H_1 = \{x_i = 0 : i \in [n]\}$$

These n hyperplanes form 2^n hyperoctants and only one hyperoctant gets negative label.

Now, suppose we construct a two-layer network with m rectifier units that has the same decision boundary as the above threshold network, and it has the form

$$y = \text{sgn}\left(w_0 + \sum_{k \in \mathcal{P}} R\left(a_k(\mathbf{x})\right) - \sum_{k \in \mathcal{N}} R\left(a_k(\mathbf{x})\right)\right)$$

using the same notations as in Theorem 1. From the theorem, we know the decision boundary of the above equation is determined by 2^m hyperplane equations and these 2^m hyperplanes form a set H_2,

$$H_2 = \left\{w_0 + \sum_{k \in \mathcal{S}_1} a_k(\mathbf{x}) - \sum_{k \in \mathcal{S}_2} a_k(\mathbf{x}) = 0 : \mathcal{S}_1 \subseteq \mathcal{P}, \mathcal{S}_2 \subseteq \mathcal{N}\right\}$$

Because we assume the rectifier network has the same decision function as the threshold network, we have

$$H_1 \subseteq H_2.$$

Note that the hyperplanes in H_1 have normal vectors independent of each other, which means there are at least n hyperplanes in H_2 such that their normal vectors are independent. Recall that $a_k(\mathbf{x}) = \mathbf{u}_k \cdot \mathbf{x} + b_k$, so all normal vectors for hyperplanes in H_2 can be expressed using linear combinations of m vectors $\mathbf{u}_1, \ldots, \mathbf{u}_m$. Since these m vectors together define the n orthogonal hyperplanes in H_1, it is impossible to have $m < n$. In other words, for this threshold network, the number of hidden units cannot be reduced by any conversion to a ReLU network.

□

4. Proof of Theorem 3

In this section we prove our theorem about hidden layer equivalence, Theorem 3, which is repeated here for convenience.

Theorem 3. *If the true concept of a 2^n-class classifier is given by a two-level threshold network in Eq. (7), then we can learn a two-layer rectifier network with only n hidden units of the form in Eq. (9) that is hidden layer equivalent to it, if for any example $\mathbf{x}$, we have*

$$\|(V - UT)^T\|_\infty \leq \frac{\gamma(\mathbf{x})}{2\|\mathbf{x}\|_\infty},$$

where $\gamma(\mathbf{x})$ is the multiclass margin for $\mathbf{x}$, defined as the difference between its highest score and second-highest scoring classes.

Let us define $\epsilon = \|(V - UT)^T\|_\infty \leq \frac{\gamma(x)}{2\|x\|_\infty}$. From the definition of the L_∞ vector norm, we have

$$\|(V - UT)^T x\|_\infty \geq |((V - UT)^T x)_k|$$

for all x and all k. The subscript k labels the k^{th} component of the vector. From the definition of the induced norm we have

$$\|(V - UT)^T x\|_\infty \leq \epsilon\|x\|.$$

Combining the above two inequalities we have

$$|((UT)^T x)_k - (V^T x)_k| \leq \epsilon\|x\|$$

for all x and all k. Assuming k^* is the highest scoring unit, for k^* we have

$$(V^T x)_{k^*} - ((UT)^T x)_{k^*} \leq \epsilon\|x\|,$$

For any other $k' \neq k^*$, we have

$$((UT)^T x)_{k'} - (V^T x)_{k'} \leq \epsilon\|x\|.$$

From the definition of the margin $\gamma(x)$, we also know that

$$(V^T x)_{k^*} - (V^T x)_{k'} \geq \gamma(x) \geq 2\epsilon\|x\|.$$

Combining the above three inequalities, we have

$$((UT)^T x)_{k'} \leq ((UT)^T x)_{k^*}$$

which means if k^* is the correct class with the highest score according to the weight parameters V, it will still be the highest scoring class according to the weight parameters UT, even if $V \neq UT$. □