Shift gears a bit and study results that are more directly applicable to classification

First paper: Montúfar, Pascanu, Cho, Bengio, 2014
"On the number of linear regions in deep neural networks"

See also: Pascanu, Montufar, Bengio, 2013
"On the # of response regions of deep feed forward networks w/ piecewise linear activations"
and more recent papers!

Idea: Count the # of piecewise linear parts of a function $f(x;\theta)$ that a ReLU network can implement

Summary of result: Deep ReLU networks can implement functions $f(x;\theta)$ with exponentially more piecewise linear regions than shallow networks w/ the same # of hidden neurons.

Now show Figure 1 from the paper

Types of networks we will study: $f(\cdot;\theta): \mathbb{R}^d \to \mathbb{R}^{out}$

(i) Shallow, one layer:

$$f(x;\theta) = f_{out} \circ \sigma \circ A(x), \qquad A(x) = Wx + b$$

where: $W \in \mathbb{R}^{d_1 \times d}$, $b \in \mathbb{R}^{d_1}$, i.e., d_1 neurons

$\sigma(z) = \max(0, z) = \text{ReLU}(z)$

$f_{out}: \mathbb{R}^{d_1} \to \mathbb{R}^{out}$, e.g.

- Regression, $f_{out}: \mathbb{R}^{d_1} \to \mathbb{R}$ and $f_{out}(z) = \langle z, \alpha \rangle$
- Classification, $f_{out}: \mathbb{R}^{d_1} \to \mathbb{R}^{\#\text{classes}}$ and $f_{out} = \text{softmax}$

(ii) Deep, L layers:

$$f(x;\theta) = f_{out} \circ \sigma \circ A_L \circ \cdots \circ \sigma \circ A_1(x)$$

where

$$A_\ell(x) = W_\ell x + b_\ell, \quad W_\ell \in \mathbb{R}^{d_\ell \times d_{\ell-1}}, \quad d_0 = d, \quad b_\ell \in \mathbb{R}^{d_\ell}$$

$f_{out}: \mathbb{R}^{d_L} \to \mathbb{R}^{out}$

Rest same as shallow network.

Def: Let $f : \mathbb{R}^d \to \mathbb{R}^m$ be a piecewise linear function. A linear region of f is a maximal connected subset of the input space $\mathbb{R}^d$ on which f is linear.

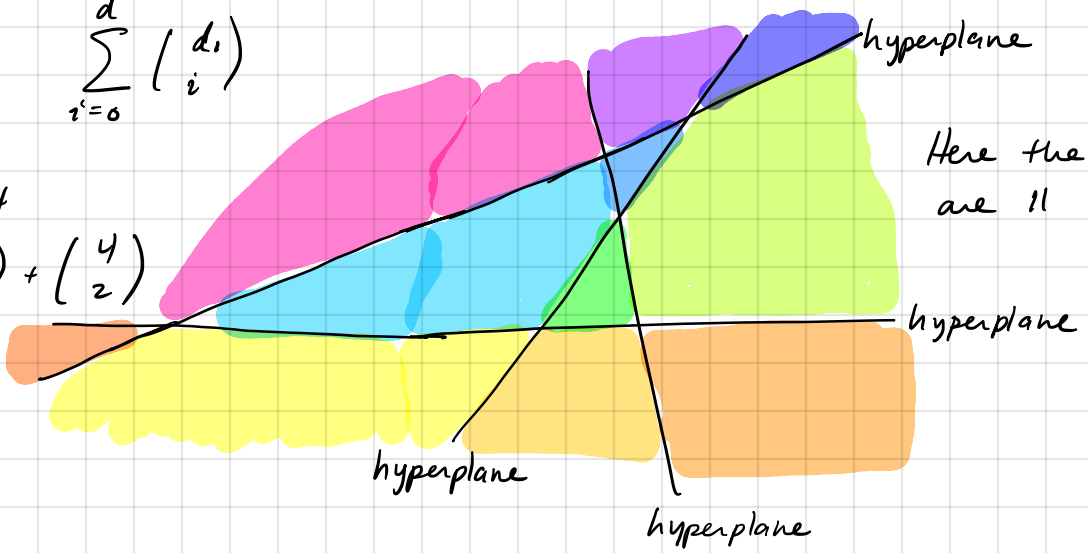
Note: For ReLU networks $f(x;\theta)$, the dimension of each linear region is d.

Results for shallow networks

Let $f(x;\theta)$ be a shallow ReLU network w/ one hidden layer and d_1 neurons. Then the maximum # of linear regions of $f(x;\theta)$ is

$$\text{max \# of linear regions} = \sum_{i=0}^{d} \binom{d_1}{i}$$

Idea of proof: Suppose you have d_1 hyperplanes in $\mathbb{R}^d$. The maximum # of regions these hyperplanes can divide $\mathbb{R}^d$ into is $\sum_{i=0}^{d} \binom{d_1}{i}$



In this example:

$d = 2$, $d_1 = 4$

$$\text{max \#} = \binom{4}{0} + \binom{4}{1} + \binom{4}{2} = 1 + 4 + 6 = 11$$

Here there are 11 regions

Results for deep networks

Main result: Let $f(x;\theta)$ be deep ReLU network w/ L hidden layers and $d_\ell \geq d$, $1 \leq \ell \leq L$, neurons in each layer. Then the maximal # of linear regions in $f(x;\theta)$ is at least:

$$\text{max \# of linear regions} \geqslant \left(\prod_{\ell=1}^{L-1} \left\lfloor \frac{d_\ell}{d} \right\rfloor^d \right) \sum_{i=0}^{d} \binom{d_L}{i} \qquad (*)$$

Come back to this at the end

This part just follows from one layer analysis

Asymptotic comparison of shallow to deep networks

Let the input dimension d be fixed.

(i) For the shallow network, give it $d_1 = Ln$ neurons

(ii) For the deep network w/ L layers, let each layer have $d_\ell = n$ neurons

Therefore both networks have the same # of neurons.

Using the shallow network result, we see:

$$\max \#_{\text{linear regions}} \text{ shallow network} = O(L^d n^d)$$

← bounded from above asymptotically as $L, n \to \infty$

polynomial in n
polynomial in L

Using the deep network result, we see:

$$\max \#_{\text{linear regions}} \text{ deep network} = \Omega\left(\left(\frac{n}{d}\right)^{(L-1)d} n^d\right)$$

← bounded from below asymptotically as $L, n \to \infty$

polynomial in n (same as shallow)
exponential in L (larger than shallow!)

Back to (*)

$$(*): \max \#\text{ linear regions for deep network} \geq \underbrace{\left(\prod_{\ell=1}^{L-1} \left\lfloor \frac{d_\ell}{d} \right\rfloor^d\right)}_{(**)} \sum_{i=0}^{d} \binom{d_L}{i}$$

Idea behind (**):

(1) Each layer can divide a single linear region into $\left\lfloor \frac{d_\ell}{d} \right\rfloor^d$ linear regions

(2) So the first layer can create $\left\lfloor \frac{d_1}{d} \right\rfloor^d$ linear regions

(3) The 2nd layer can create $\left\lfloor \frac{d_2}{d} \right\rfloor^d$ linear regions from each of the linear regions created by the 1st layer.

Thus the total # of linear regions is:

$$\left\lfloor \frac{d_1}{d} \right\rfloor^d \cdot \left\lfloor \frac{d_2}{d} \right\rfloor^d$$

(4) Continue for the first $L-1$ layers, you get (**).