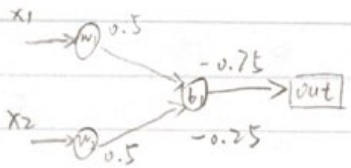
PS2 Yang Tian

1.1

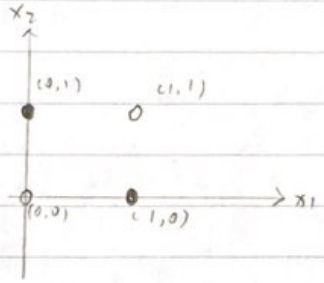


$$W_{AND} = \begin{bmatrix} 0.5 \\ 0.5 \end{bmatrix} \quad b_{AND} = -0.75$$

$$W_{OR} = \begin{bmatrix} 0.5 \\ 0.5 \end{bmatrix} \quad b_{OR} = -0.25$$

1.2

x_1	x_2	$f_{XOR}(x)$
0	0	0
0	1	1
1	0	1
1	1	0



$f(x) = W^T x + b$ is a linear model. from left figure, we can find that no trivial linear separator (line) can separate two classes.

2.1 Since $f: \mathbb{R}^d \to \mathbb{R}$ is differentiable and Lipschitz continuous

$$|f(x_1) - f(x_0)| \le k \|x_1 - x_0\|$$

$$\frac{\|f(x_1) - f(x_0)\|}{\|x_1 - x_0\|} \le k$$

According to mean value theorem, for $\forall x, y \in [a, b]$, $x < y$ $\exists c \in (x, y) \subseteq (a, b)$ s.t.

$$\left| \frac{f(y) - f(x)}{y - x} \right| = |f'(c)| \le L$$

Furthermore, we could argue that lipschitz constant $K \ge \sup_{x \in (x_0, x_1)} |f'(x)|$

Therefore $|\nabla f(x_0)| \le K$

$$|R(x)| = |f(x) - f(x_0) - \nabla f(x_0)^T (x - x_0)|$$

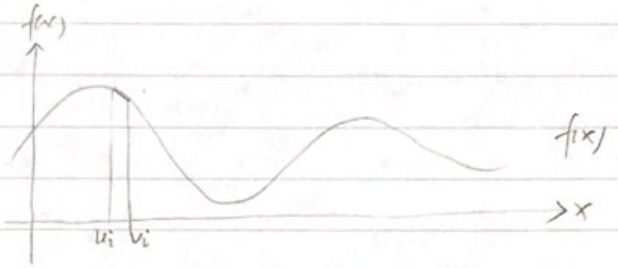
$$\le |f(x) - f(x_0)| + |\nabla f(x_0)^T (x - x_0)|$$

$$\le k|x - x_0| + k|x - x_0| \le 2kC$$

2.2. Activation function

$$g(h, u, v)_i = \begin{cases} 0 & h_i < u_i \\ h_i & u_i < h_i < v_i \\ 0 & v_i < h_i \end{cases}$$

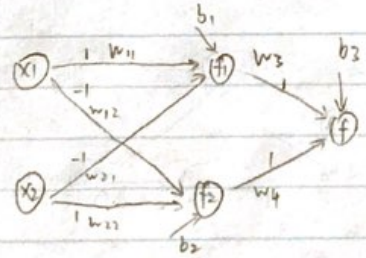
From the figure, we know it is a bumper function, whose bumper slope could further adjusted by $W^{(2)}$. Using a bunch of these functions, we could approximate the differentiable and lipschitz continuous function $f(x)$ to any desired error ε.



For better approximation ~~and~~, we could use a larger but finite number of hidden units. each unit only approximates $f(x)$, $u_i < x \leq v_i$ Therefore the number of hidden units and corresponding weights ~~depends~~ depend on the total range of x and approximation error.

Theoretically speaking we could think of approximation as wavelet transform. The activation function is mother wave. Using wavelet transform, we could find the parameters for each wave and the total number of ~~needed~~ waves needed, given the desired approximation error.

2.3



$w_{11} = 1$
$w_{12} = -1$
$w_{21} = -1$
$w_{22} = 1$

$b_1 = -1$

$f_1 = \begin{cases} 1 & x \geq 0 \\ 0 & x < 0 \end{cases}$

$b_2 = -1$

$f_2 = \begin{cases} 1 & x \geq 0 \\ 0 & x < 0 \end{cases}$

$w_3 = 1$
$w_4 = 1$

$b_3 = -0.1$

$f = \begin{cases} 1 & x \geq 0 \\ 0 & x < 0 \end{cases}$

0	0	-1	0	0
		-1	0	
		-1	0	
1	1	-1	0	0
		-2	0	
0	1	0	1	1
		0	1	
1	0	-2	0	1

3. $h(x) = W^{(3)} \max\{0, W^{(2)} \max\{0, W^{(1)}x + b^{(1)}\} + b^{(2)}\} + b^{(3)}$

3.1 $x = 1$

$h(x) = W^{(3)} \max\{0, W^{(2)} \begin{bmatrix} 0.5 \\ 1.5 \end{bmatrix} + b^{(2)}\} + b^{(3)}$

$= W^{(3)} \max\{0, \begin{bmatrix} 2 \\ 2 \end{bmatrix}\} + b^{(3)}$

$= 5$

$\frac{dh}{dx} = [1,1] \begin{bmatrix} 1 & 1 \\ 1 & 1 \end{bmatrix} \begin{bmatrix} 0.5 \\ 0.5 \end{bmatrix} = 2$ $\quad h(x) = Wx + b$, $W = 2$ $b = 3$

3.2 $x = -1$

$h(x) = W^{(3)} \max\{0, W^{(2)} \begin{bmatrix} 0 \\ 0.5 \end{bmatrix} + b^{(2)}\} + b^{(3)}$

$= W^{(3)} \max\{0, \begin{bmatrix} 0.5 \\ 0.5 \end{bmatrix}\} + b^{(3)}$

$= 2$

$\frac{dh}{dx} = [1,1] \begin{bmatrix} 1 & 1 \\ 1 & 1 \end{bmatrix} \begin{bmatrix} 0 & 0 \\ 0 & 1 \end{bmatrix} \begin{bmatrix} 0.5 \\ 0.5 \end{bmatrix} = 1$ $\quad h(x) = Wx + b$, $W = 1$ $b = 3$

3.3 $x = -0.5$

$h(x) = W^{(3)} \max\{0, W^{(2)} \begin{bmatrix} 0 \\ 0.75 \end{bmatrix} + b^{(2)}\} + b^{(3)}$

$= W^{(3)} \max\{0, \begin{bmatrix} 0.75 \\ 0.75 \end{bmatrix}\} + b^{(3)}$

$= 2.5$

$\frac{dh}{dx} = 1$

$h(x) = Wx + b$

$W = 1$ $\quad b = 3$

4.(a) $W = 2 \cdot I = \begin{bmatrix} 2 & & 0 \\ & \ddots & \\ 0 & & 2 \end{bmatrix}$ $b = \begin{bmatrix} -1 \\ -1 \\ \vdots \\ -1 \end{bmatrix}$

For d-dimension space, there are 2^d regions of input are identified onto $O = (0,1)^d$ by $f_1(x) = | W^{(1)} x_0 + b^{(1)})$

when $d=1$ we can easily find $R_1 = (0.5, 1)$ and $R_2 = (0, 0.5)$
when $d=2$, we need to choose either R_1 or R_2 for each dimension. so ~~or~~ there are $2^2 = 4$ ~~possibility~~ regions
Generally speaking, for d-dimension, there are two possibilities for each dimension, so there are 2^d regions.

(b) Without loss of ~~gener~~ generality, we could assume $g(\cdot)$ is applied on input first, then $f(\cdot)$.
For function $f(\cdot)$, there are n_f regions that could be identified onto $(0,1)^d$, each of which, obtained by applying $g(\cdot)$ on input field, has n_g identified ~~reg~~ regions.
As a result, totally speaking there are $n_f \cdot n_g$ regions identified onto final $[0,1]^d$.

(c) Let's start backwards from layer L.
Output region is $[0,1]^d$. So h_L will have 2^d identified regions. Normally not all of them (or even ~~none~~ none of them) are still $[0,1]^d$. But the result from question (a) doesn't have any restriction on output region. So for each of 2^d regions before layer L, there are 2^d regions at layer L-1. And this goes on and on, we will find down to the beginning, there are 2^{Ld} regions identified to $[0,1]^d$, which is exponential to the depth of networks and dimension of space