Chapter 3: Shallow neural networks

Problem 3.1

What kind of mapping from input to output would be created if the activation function in equation (3.1) was linear so that $\mathrm{a}[z] = \psi_0 + \psi_1 z$? What kind of mapping would be created if the activation function was removed so $\mathrm{a}[z] = z$?

$$\begin{aligned} y &= \mathrm{f}[x, \boldsymbol{\phi}] \\ &= \phi_0 + \phi_1 \mathrm{a}[\theta_{10} + \theta_{11} x] + \phi_2 \mathrm{a}[\theta_{20} + \theta_{21} x] + \phi_3 \mathrm{a}[\theta_{30} + \theta_{31} x] \end{aligned} \tag{3.1}$$

In both scenarios, the resulting mapping would be linear. With no activation function, the linearity of equation (3.1) is preserved. If the activation function is linear, the mapping is also linear but shifted and scaled by the parameters ψ_0 and ψ_1. This is shown in the jupyter notebook.

Problem 3.2

For each of the four linear regions in figure 3.3j, indicate which hidden units are inactive and which are active (i.e., which do and do not clip their inputs)

- Region 1: only hidden unit 1 is active
- Region 2: hidden units 1 and 3 are active
- Region 3: all hidden units are active
- Region 4: hidden units 1 and 2 are active

Problem 3.3

Derive expressions for the positions of the "joints" in function in figure 3.3j in terms of the ten parameters $\boldsymbol{\phi}$ and the input x. Derive expressions for the slopes of the four linear regions

From figures 3.3a - 3.3c, the $\boldsymbol{\theta}$ parameters are given as:

$$\begin{aligned}\theta_{10} &= -0.2, \theta_{11} = 0.3\\ \theta_{20} &= -0.9, \theta_{21} = 0.85\\ \theta_{30} &= 1.2, \theta_{31} = -0.75\end{aligned}$$

(by inspection)

The joints are given by the points where the hidden units switch from clipping to passing the input (i.e, when the functions cross the line $y = 0$). The joints are given by:

$$\begin{aligned}\text{Joint 1} &= \frac{\theta_{10}}{\theta_{11}} = \frac{0.2}{0.3} = 0.67\\ \text{Joint 2} &= \frac{\theta_{20}}{\theta_{21}} = \frac{0.9}{0.85} = 1.06\\ \text{Joint 3} &= \frac{\theta_{30}}{\theta_{31}} = \frac{1.2}{0.75} = 1.6\end{aligned}$$

Problem 3.4

> Draw a version of figure 3.3 where the y-intercept and slope of the third hidden unit have changed as in figure 3.14c. Assume that the remaining parameters remain the same.

Assuming $\theta_{30} = 1.1$ and $\theta_{31} = -0.7$, the new figure is shown below:

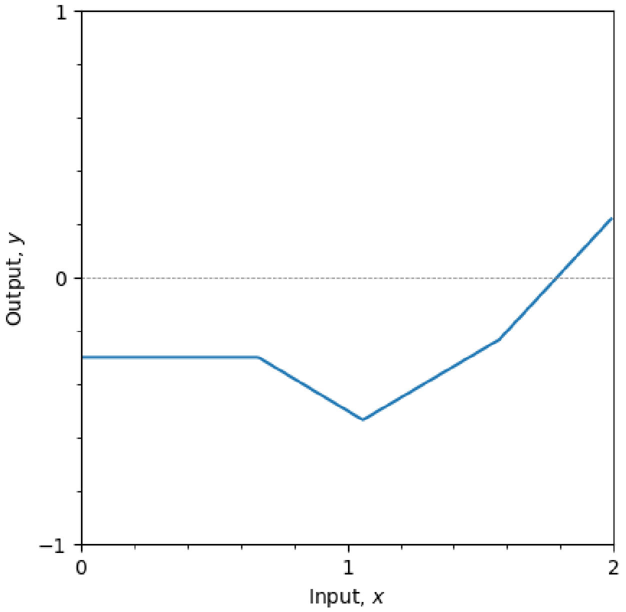

Figure 1: Modified version of figure 3.3 with $\theta_{30} = 1.1$ and $\theta_{31} = -0.7$

Problem 3.5

Prove that the following property holds for $\alpha \in \mathbb{R}^+$:

$$\text{ReLU}[\alpha z] = \alpha\text{ReLU}[z]$$

This is known as the *non-negative homogeneity* property of the ReLU function.

Consider the case when $z \geq 0$:

$$\text{ReLU}[\alpha z] = \max(0, \alpha z) = \alpha z = \alpha \max(0, z) = \alpha\text{ReLU}[z]$$

Similarly, when $z < 0$:

$$\text{ReLU}[\alpha z] = \max(0, \alpha z) = 0 = \alpha \max(0, z) = \alpha\text{ReLU}[z]$$

Problem 3.6

Following on from problem 3.5, what happens to the shallow network defined in equations 3.3 and 3.4 when we multiply the parameters θ_{01} and θ_{11} by a positive constant α and divide the slope ϕ_1 by the same parameter α? What happens if α is negative?

When multiplying the parameters θ_{01} and θ_{11} by a positive constant α and dividing the slope ϕ_1 by the same parameter α, the network remains unchanged. This is because the ReLU activation function is non-negative homogeneous. (see jupyter notebook)

If α is negative, the network will change. The ReLU activation function is not homogeneous for negative values of α.

Problem 3.7

Consider fitting the model in equation 3.1 using a least squares loss function. Does this loss function have a unique minimum? i.e., is there a single "best" set of parameters?

The loss function in equation 3.1 is a linear function of the parameters $\boldsymbol{\phi}$, and assuming a ReLU activation function, the resulting loss function will be unlikely be convex, resulting in multiple local minima. This is because the ReLU activation function introduces non-linearities into the model, leading to a non-convex optimization problem.

Problem 3.8

Consider replacing the ReLU activation function with (i) the Heaviside step function heaviside$[z]$, (ii) the hyperbolic tangent function tanh$[z]$, and (iii) the rectangular function rect$[z]$, where:

$$\text{heaviside}[z] = \begin{cases} 0 & \text{if } z < 0 \\ 1 & \text{otherwise} \end{cases}$$

$$\text{rect}[z] = \begin{cases} 0 & \text{if } z < 0 \\ 1 & \text{if } 0 \leq z \leq 1 \\ 0 & \text{if } z > 1 \end{cases}$$

Redraw a version of figure 3.3 for each of these functions. The original parameters were: $\boldsymbol{\phi} = \{\phi_0, \phi_1, \phi_2, \phi_3, \theta_{01}, \theta_{11}, \theta_{20}, \theta_{21}, \theta_{30}, \theta_{31}, \}$ ={-0.23, -1.3, 1.3, 0.66, -0.2, 0.4, -0.9, 0.9, 1.1, -0.7}. Provide an informal description of the family of functions that can be created by neural networks with one input, three hidden units, and one output for each activation function.

Using the heaviside function:

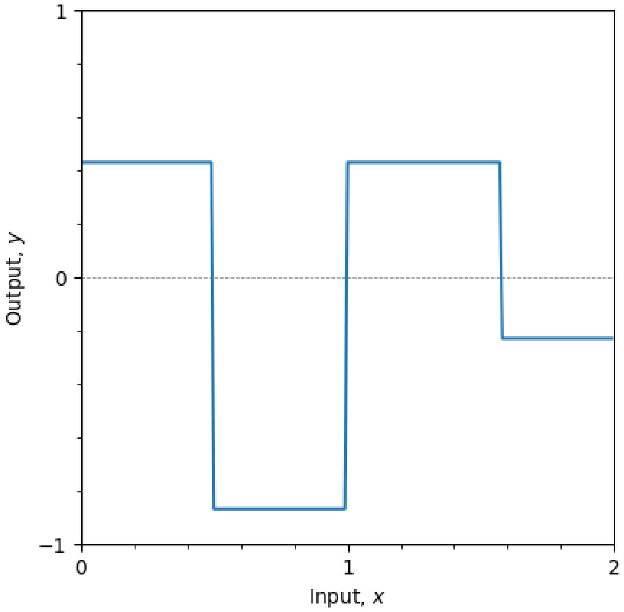

Figure 2: Modified version of figure 3.3 with the Heaviside step function

Using the hyperbolic tangent function:

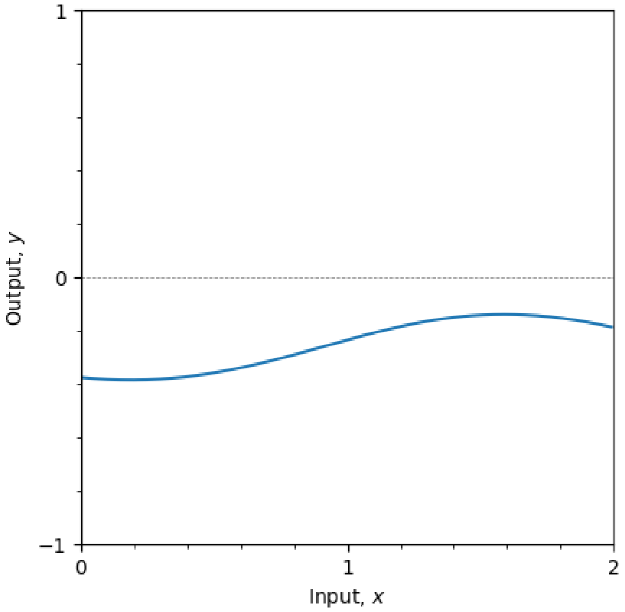

Figure 3: Modified version of figure 3.3 with the hyperbolic tangent function

Using the rectangular function:

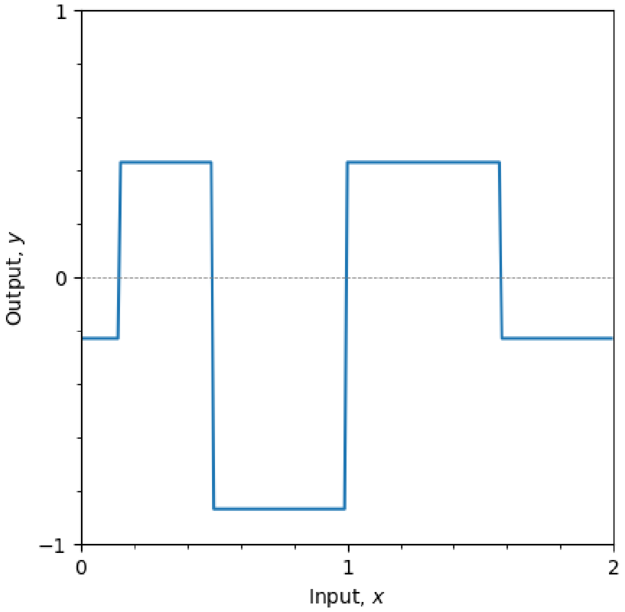

Figure 4: Modified version of figure 3.3 with the rectangular function

Problem 3.9

Show that the third linear region in figure 3.3 has a slope that is the sum of the slopes of the first and fourth linear regions.

In the third linear region, all hidden units are active, therefore the slope is given by:

$$\text{Slope}_3 = \phi_1\theta_{11} + \phi_2\theta_{21} + \phi_3\theta_{31} = -1.3 \times 0.4 + 1.3 \times 0.9 + 0.66 \times -0.7 = 0.188$$

In the fourth region, the slope is given by a linear combination of the slopes of the first and second hidden units (multiplied by $\boldsymbol{\phi}$):

$$\text{Slope}_4 = \phi_1\theta_{11} + \phi_2\theta_{21} = -1.3 \times 0.4 + 1.3 \times 0.9 = 0.65$$

In the first region, the slope is given by the slope of the third hidden unit (multiplied by $\boldsymbol{\phi}$):

$$\text{Slope}_1 = \phi_3\theta_{31} = 0.66 \times -0.7 = -0.462$$

Therefore, the slope of the third region is the sum of the slopes of the first and fourth regions ($0.188 = 0.65 - 0.462$).

Problem 3.10

> Consider a neural network with one input, one output, and three hidden units. The construction in figure 3.3 shows how this creates four linear regions. Under what circumstances could this network produce a function with fewer than four linear regions?

If two or more ReLU units have weights and biases configured such that their kinks (transitions from zero to a linear positive slope) occur at the same input value, the expected number of kinks (and thus linear regions) in the output will reduce. alternatively, If the weight or bias of any hidden unit results in it never activating (i.e., the output is always zero regardless of the input), it effectively does not contribute to segmenting the input space.

A third option for reducing the number of linear regions occurs if two or more hidden units cancel each other out. This can happen if the weights and biases of two hidden units are configured such that their activations are always equal and opposite, leading to a net zero contribution to the output.

Problem 3.11

> How many parameters does the model in figure 3.6 have?

It has $1 \times 4 + 4 \times 2 = 12$ slopes, and $4 + 2 = 6$ biases, giving a total of $12 + 6 = 18$ parameters.

Problem 3.12

How many parameters does the model in figure 3.7 have?

It has $2 \times 3 + 3 \times 1 = 9$ slopes, and $3 + 1 = 4$ biases, giving a total of $9 + 4 = 13$ parameters.

Problem 3.13

What is the activation pattern for each of the seven regions in figure 3.8? In other words, which hidden units are active (pass the input) and which are inactive (clip the input) for each region?

The bottom left region is defined by the third hidden layer. The one directly to the right of this by hidden layers 2 and 3, and the region to the right by all of the hidden layers. The middle left region is empty. The top middle region is defined by hidden layer 1, the top right region by layers 1 and 2. The middle region (triangle) is defined by hidden layers 1 and 3.

Problem 3.14

Write out the equations that define the network in figure 3.11. There should be three equations to compute the three hidden units from the inputs and two equations to compute the outputs from the hidden units.

The equations for the hidden layers are given by:

$$\begin{aligned}
h_1 &= \mathrm{a}\left[\theta_{10} + \theta_{11}x_1 + \theta_{12}x_2 + \theta_{13}x_3\right] \\
h_2 &= \mathrm{a}\left[\theta_{20} + \theta_{21}x_1 + \theta_{22}x_2 + \theta_{23}x_3\right] \\
h_3 &= \mathrm{a}\left[\theta_{30} + \theta_{31}x_1 + \theta_{32}x_2 + \theta_{33}x_3\right]
\end{aligned}$$

The equations for the output layer are given by:

$$\begin{aligned}
y_1 &= \phi_{10} + \phi_{11}h_1 + \phi_{12}h_2 + \phi_{13}h_3 \\
y_2 &= \phi_{20} + \phi_{21}h_1 + \phi_{22}h_2 + \phi_{23}h_3
\end{aligned}$$

Problem 3.15

What is the maximum possible number of 3D linear regions that can be created by the network in figure 3.11?

A shallow network with D_i inputs, D hidden units, and D_o outputs can create a maximum of $\sum_{j=0}^{D_i} \binom{D}{j}$ regions. For the network in figure 3.11, with $D_i = 3$, $D = 3$, and $D_o = 2$, the maximum number of regions is $\sum_{j=0}^{3} \binom{3}{j} = 1 + 3 + 3 + 1 = 8$.

Problem 3.16

Write out the equations for a network with two inputs, four hidden units, and three outputs. Draw this model in the style of figure 3.11.

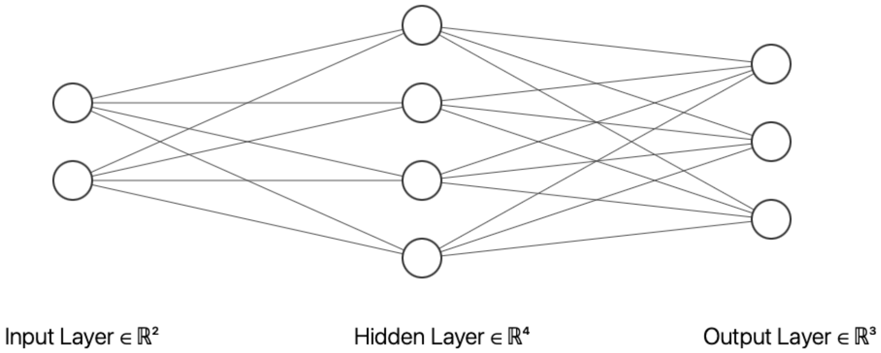

Figure 5: Shallow NN with 2 inputs, 4 hidden units, and 3 outputs

Problem 3.17

Equations (3.11) and (3.12) define a general neural network with D_i inputs, one hidden layer containing D hidden units, and D_o outputs.

$$h_d = \mathrm{a}\left[\theta_{d0} + \sum_{i=1}^{D_i} \theta_{di} x_i\right] \tag{3.11}$$

$$y_j = \phi_{j0} + \sum_{d=1}^{D} \phi_{jd} h_d \tag{3.12}$$

Find an expression for the number of parameters in the model in terms of D_i, D, and D_o.

Given that there are D hidden units and $1 + D_i$ parameters per hidden unit, there are a total of $D(1 + D_i)$ parameters for the hidden layer. The output layer has $D_o(1 + D)$ parameters. Therefore, the total number of parameters is given by:

$$D(1 + D_i) + D_o(1 + D)$$

Problem 3.18

Show that the maximum number of regions created by a shallow network with $D_i = 2$-dimensional input, $D_o = 1$-dimensional output, and $D = 3$ hidden units is seven, as in figure 3.8j. Use the result of Zaslavsky (1975) that the maximum number of regions created by partitioning a D_i-dimensional space with D hyperplanes is $\sum_{j=0}^{D_i} \binom{D}{j}$. What is the maximum number of regions if we add two more hidden units to this model, so $D = 5$?

Defining the binomial coefficient:

$$\binom{D}{j} = \frac{D!}{j!(D-j)!}$$

With three hidden units and two inputs, the computation is:

$$\begin{aligned} N &= \binom{3}{0} + \binom{3}{1} + \binom{3}{2} \\ &= 1 + 3 + 3 && = 7 \end{aligned}$$

With five units and two inputs, the computation is:

$$
\begin{aligned}
N &= \binom{5}{0} + \binom{5}{1} + \binom{5}{2} \\
&= 1 + 5 + 10 && = 16
\end{aligned}
$$