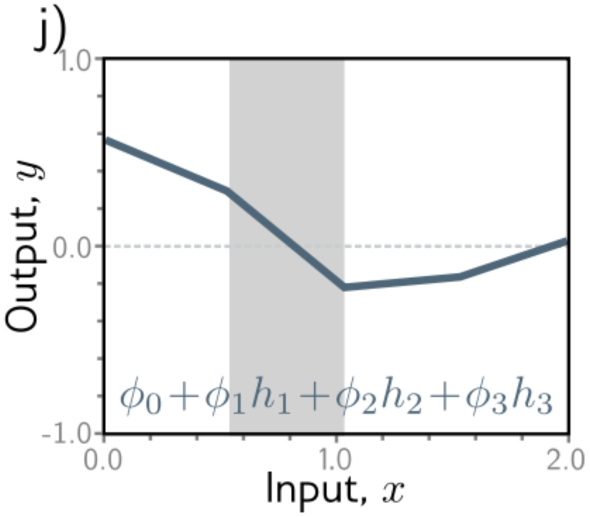
3 Shallow Neural Networks

Problems

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 $a[z] = \psi_0 + \psi_1 z$? What king of mapping would be created if the activation function was removed, so $a[z] = z$?

If the activation is a linear mapping, the overall transformation from the input x to $a[\theta_0 + \theta_1 x]$ would be overall linear. Therefore, the total mapping from x to y, which is again the linear combination of outputs from the activation would be also linear. Similar interpretation can be applied to the case where the activation is the identity function.

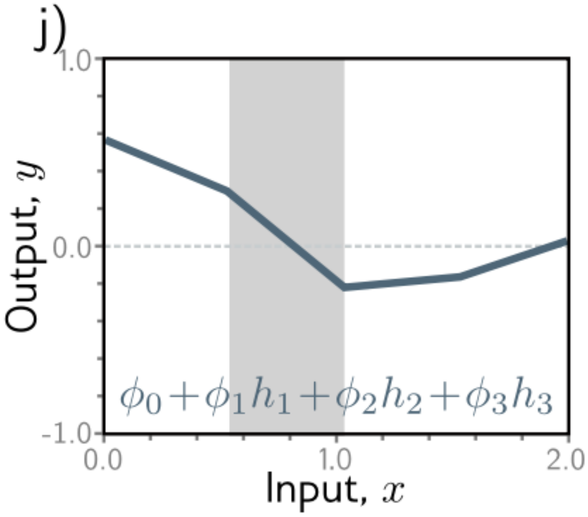
Problem 3.2 For each of the four linear regions in figure 3.3j, indicate which hidden units are inactive and which are active (i.e., which do and do not clip their inputs)



- In the first region, only h_3 is active.
- In the second region, h_1 and h_3 are active.

- In the third region, all h_1, h_2, and h_3 are active.
- In the last region, h_1 and h_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 ϕ and the input x. Derive expressions for the slopes of the four linear regions.



Note that the equation for the shallow neural network is

$$y = \phi_0 + \phi_1 a[\theta_{10} + \theta_{11}x] + \phi_2 a[\theta_{20} + \theta_{21}x] + \phi_3 a[\theta_{30} + \theta_{31}x]$$

- In the first region, only h_3 is active so the slope is $\phi_3\theta_{31}$.
- The first joint is derived from the first hidden unit h_1, so $x_1^* = -\theta_{10}/\theta_{11}$.
- In the second region, h_1 and h_3 are active, so the slope is $\phi_1\theta_{11} + \phi_3\theta_{31}$ here.
- The second joint is where the input to h_2 becomes zero, so $x_2^* = -\theta_{20}/\theta_{21}$.
- In the third region, all the hidden units are active and the slope is $\phi_1\theta_{11} + \phi_2\theta_{21} + \phi_3\theta_{31}$.

- The third join is at $x_3^* = -\theta_{30}/\theta_{31}$.
- The slope of the last region is $\phi_1\theta_{11} + \phi_2\theta_{21}$.

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.

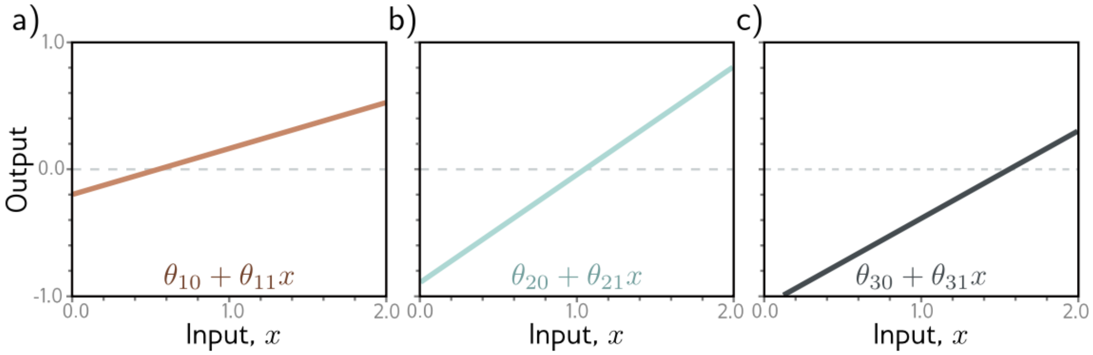
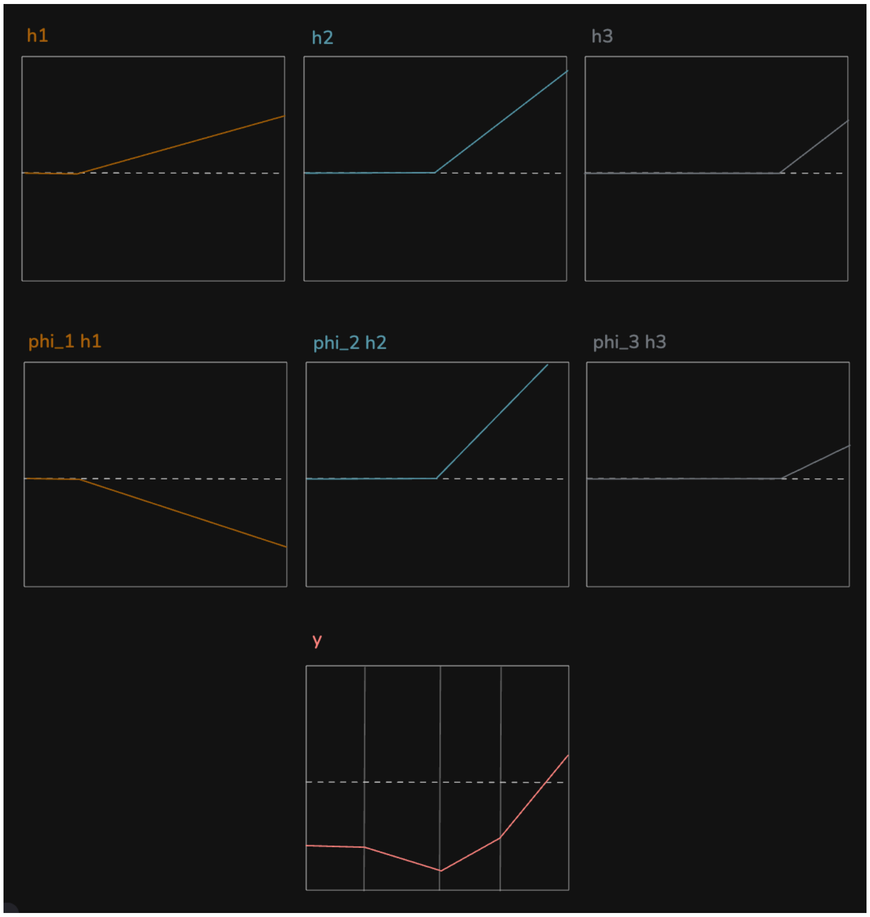


Figure 3.14 Processing in network with one input, three hidden units, and one output for problem 3.4. a–c) The input to each hidden unit is a linear function of the inputs. The first two are the same as in figure 3.3, but the last one differs.



Problem 3.5 Prove that the following property holds for $\alpha \in \mathbb{R}^{+}$:

$$\mathrm{ReLU}[\alpha \cdot z] = \alpha \cdot \mathrm{ReLU}[z]$$

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

We consider the case $z < 0$ and $z \geq 0$.

- $z < 0$: $\mathrm{ReLU}[z] = \alpha \cdot \mathrm{ReLU}[z] = 0$ by definition,
  and since
  $\alpha \geq 0, \alpha \cdot z < 0$ so $\mathrm{ReLU}[\alpha \cdot z] = 0.$
- $z \geq 0$: $\alpha \cdot \mathrm{ReLU}[z] = \alpha \cdot z$
  and
  $\mathrm{ReLU}[\alpha \cdot z] = \alpha \cdot z.$

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_{10} 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?

The expression $\phi_1\mathrm{ReLU}[\theta_{10}+\theta_{11}x]$ with parameters θ_{10}, θ_{11} multiplied by α and ϕ_1 divided by α can be computed as

$$\begin{aligned}(\phi_1/\alpha)\mathrm{ReLU}[\alpha(\theta_{10}+\theta_{11}x)] &= (\phi_1/\alpha)(\alpha\mathrm{ReLU}[\theta_{10}+\theta_{11}x]) \\ &= \phi_1\mathrm{ReLU}[\theta_{10}+\theta_{11}x]\end{aligned}$$

so the end result is the same.

The homogeneity property does not hold for $\alpha < 0$, so the result differ.

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?

We have shown in problem 3.6 that the parameters ϕ_i, θ_{i0}, and θ_{i1} can be scaled with a nonnegative α to produce the equivalent model. Therefore, there does not exist a unique set of parameters producing the minimal square loss.

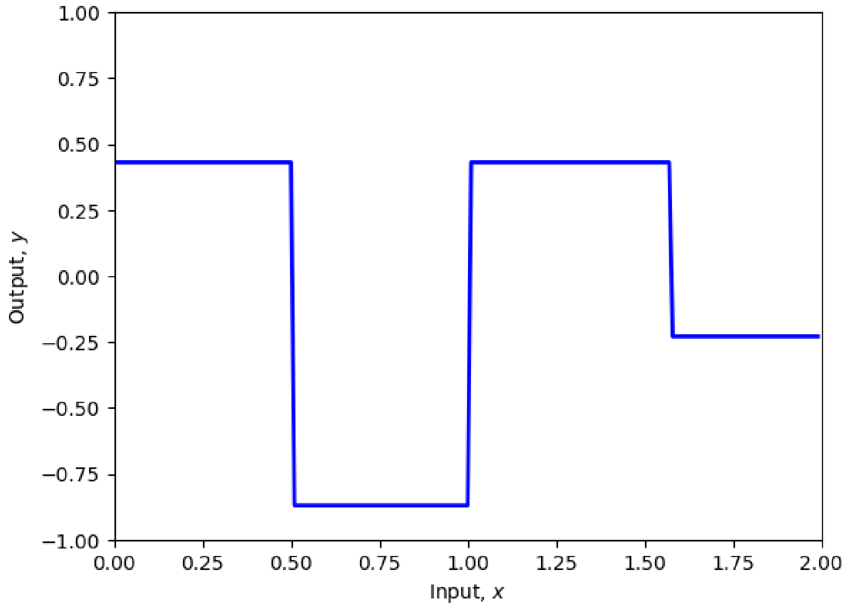
Problem 3.8 Consider replacing the ReLU activation function with (i) the Heaviside step function $\mathrm{heaviside}[z]$, (ii) the hyperbolic tangent function $\tanh[z]$, and (iii) the rectangular function $\mathrm{rect}[z]$, where:

$$\mathrm{heaviside}[z] = \begin{cases} 0 & z < 0 \\ 1 & z \geq 0 \end{cases}$$

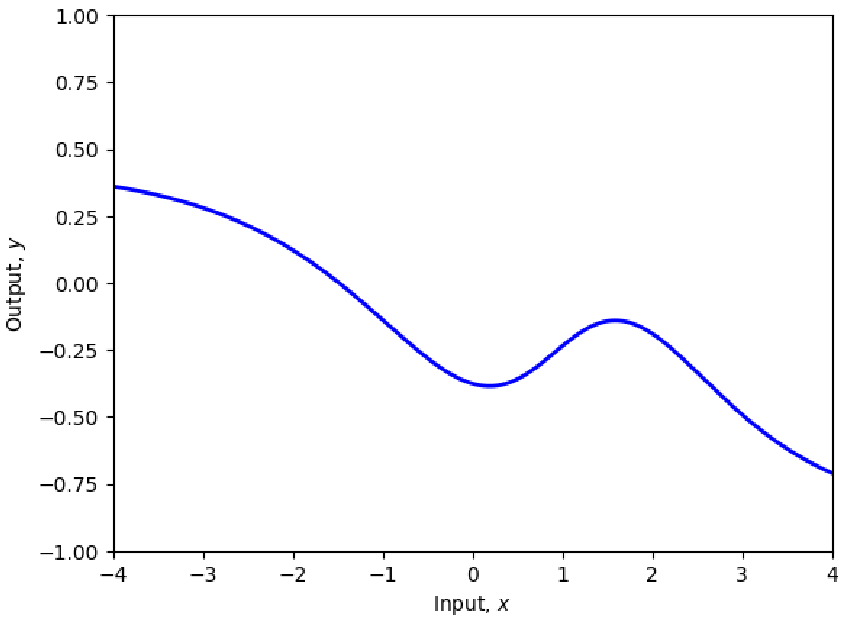
$$\mathrm{rect}[z] = \begin{cases} 0 & z < 0 \\ 1 & 0 \leq z \leq 1 \\ 0 & z > 1 \end{cases}$$

Redraw a version of figure 3.3 for each of these functions. The original parameters were: $\phi = \{\phi_0, \phi_1, \phi_2, \phi_3, \theta_{10}, \theta_{11}, \theta_{20}, \theta_{21}, \theta_{30}, \theta_{31}\} = \{-0.23, -1.3, 1.3, 0.66, -0.2, 0.4, -0.9, 0.9, 1.1, -0.7\}$. 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.

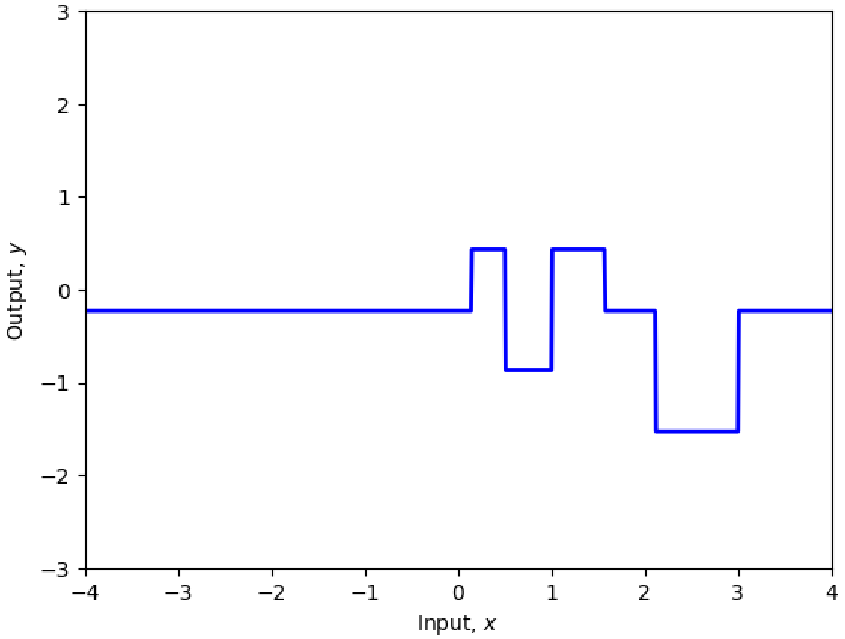
heaviside: Each hidden layer allows a single discontinuity, thus the family of functions would consist of four constant regions.



tanh: The gradient of the function should converge to zero as the input goes to either side of infinity. The function can have at most three inflection points.



rect: As the input goes to either side of infinity, the output becomes zero. Within the finite region, the function can have at most 5 constant regions with distinct values.



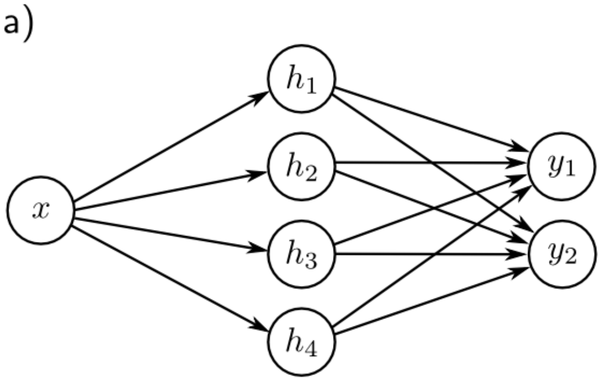
Problem 3.9 Show that the third linear region in figure 3.3 has a slope that is the sum of the slopes of the first and fourth linear regions.

We already computed the slopes in problem 3.3. The slope of the third region is $\phi_1\theta_{11} + \phi_2\theta_{21} + \phi_3\theta_{31}$, which is equal to the sum of the slope of the first region ($\phi_3\theta_{31}$) and the last region ($\phi_1\theta_{11} + \phi_2\theta_{21}$).

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?

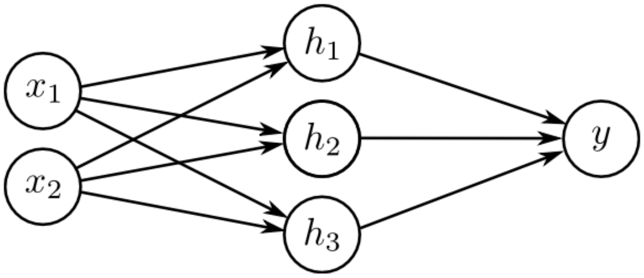
The joints could be formed at the same position (e.g. $\theta_{10}/\theta_{11} = \theta_{20}/\theta_{21}$).

Problem 3.11 How many parameters does the model in figure 3.6 have?



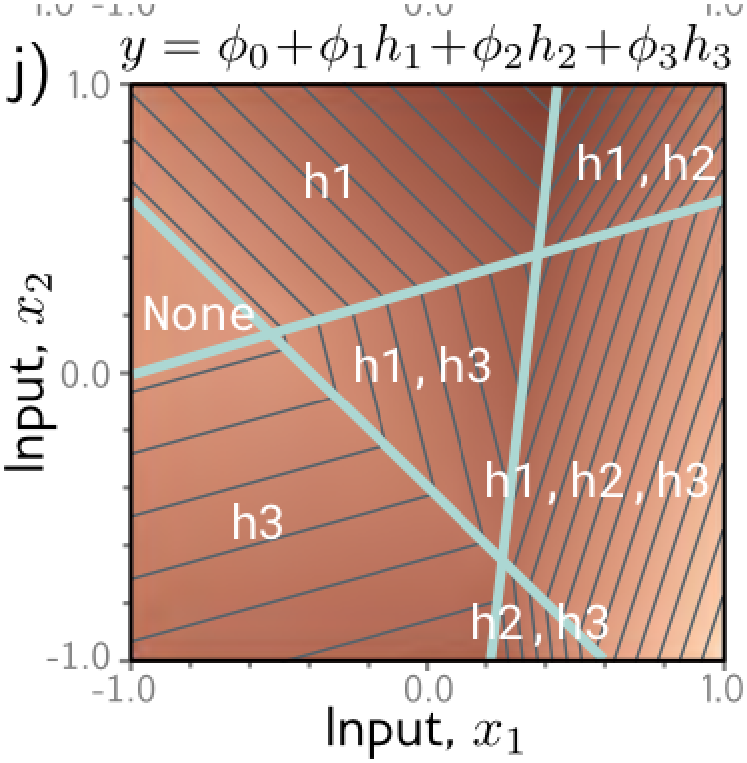
There are 12 slopes and 6 offsets, so 18 parameters in total.

Problem 3.12 How many parameters does the model in figure 3.7 have?



9 slopes and 4 offsets, so 13 parameters in total.

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?



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.

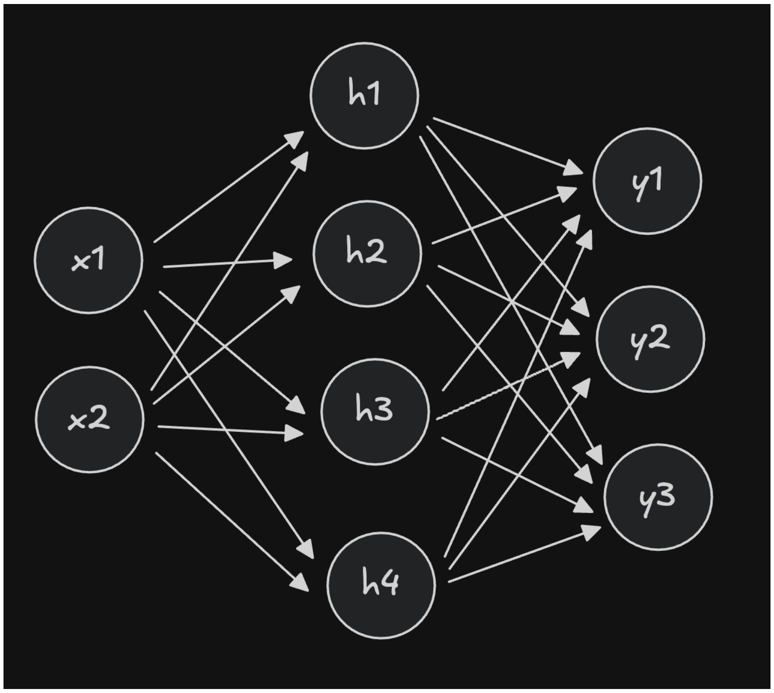
$$
\begin{aligned}
h_1 &= a[\theta_{10} + \theta_{11}x_1 + \theta_{12}x_2 + \theta_{13}x_3] \\
h_2 &= a[\theta_{20} + \theta_{21}x_1 + \theta_{22}x_2 + \theta_{23}x_3] \\
h_3 &= a[\theta_{30} + \theta_{31}x_1 + \theta_{32}x_2 + \theta_{33}x_3] \\
y_1 &= \phi_{10} + \phi_{11}h_1 + \phi_{12}h_2 + \phi_{13}h_3 \\
y_2 &= \phi_{20} + \phi_{21}h_1 + \phi_{22}h_2 + \phi_{23}h_3
\end{aligned}
$$

Problem 3.15 What is the maximum possible number of 3D linear regions that can be created by the network in figure 3.11?

Each hidden unit defines a 2D plane which divides the space into two 3D linear regions. Given three hidden units, there can be at most eight linear regions.

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.

$$
\begin{aligned}
h_1 &= a[\theta_{10} + \theta_{11}x_1 + \theta_{12}x_2] \\
h_2 &= a[\theta_{20} + \theta_{21}x_1 + \theta_{22}x_2] \\
h_3 &= a[\theta_{30} + \theta_{31}x_1 + \theta_{32}x_2] \\
h_4 &= a[\theta_{40} + \theta_{41}x_1 + \theta_{42}x_2] \\
y_1 &= \phi_{10} + \phi_{11}h_1 + \phi_{12}h_2 + \phi_{13}h_3 + \phi_{14}h_4 \\
y_2 &= \phi_{20} + \phi_{21}h_1 + \phi_{22}h_2 + \phi_{23}h_3 + \phi_{24}h_4 \\
y_3 &= \phi_{30} + \phi_{31}h_1 + \phi_{32}h_2 + \phi_{33}h_3 + \phi_{34}h_4
\end{aligned}
$$



Problem 3.17 Equations 3.11 and 3.14 define a general neural network with D_i inputs, one hidden layer containing D hidden units, and D_o outputs. Find an expression for the number of parameters in the model interms of D_i, D, and D_o.

There are $(D_i + D_o)D$ slopes and $D + D_o$ offset parameters. Thus in total, $(D_i + D_o + 1)D + D_o$ parameters.

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 $D = 5$?

Using the result of Zaslavsky, the maximum number of linear regions is

$$\sum_{j=0}^{2} \binom{3}{j} = \binom{3}{0} + \binom{3}{1} + \binom{3}{2} = 1 + 3 + 3 = 7$$

In case of $D = 5$, the maximum number of linear regions is

$$\sum_{j=0}^{2} \binom{5}{j} = \binom{5}{0} + \binom{5}{1} + \binom{5}{2} = 1 + 5 + 10 = 16$$