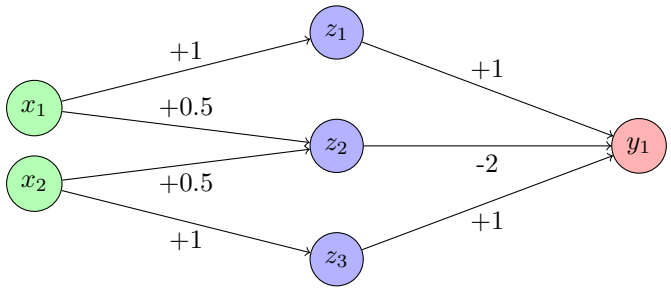
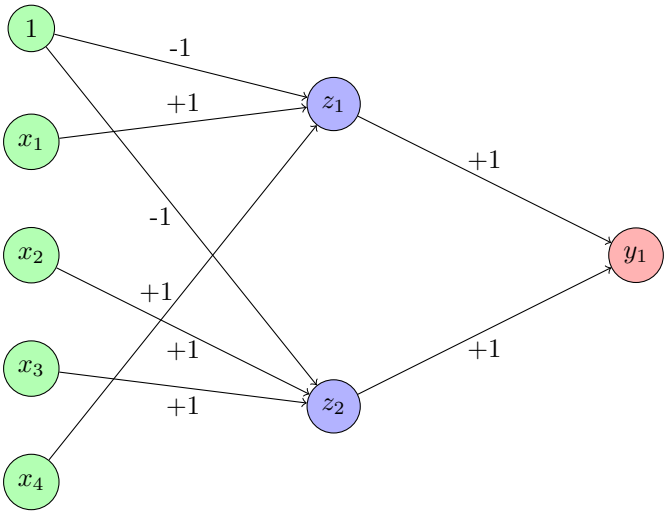
1 Exercise *(Multi-Layer Perceptron (8p))*

1. Draw multi-layer perceptron to solve each given logical function below.
   (We assume a threshold $\Theta = 1$ for each neuron. That is, neurons fire if they have an activation of 1 or higher, and they do not fire if they have an activation lower than 1.)

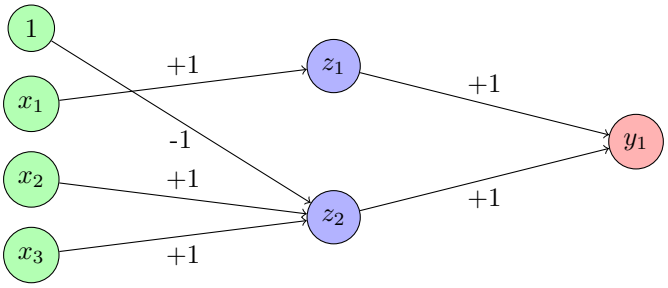
   (a) $x1 \oplus x2$



   (b) $(x_1 \wedge x_4) \vee (x_2 \wedge x_3)$



   (c) $x_1 \vee (x_2 \wedge x_3)$



2. Calculate the following derivatives:

(a)

$$
\begin{aligned}
f(x) &= \frac{1}{1+exp(-\lambda x)} = (1+exp(-\lambda x))^{-1} \\
f'(x) &= (-1) * (1+exp(-\lambda x))^{-2} * exp(-\lambda x) * (-\lambda) \\
&= \frac{\lambda}{1+exp(-\lambda x)} * \frac{exp(-\lambda x)}{1+exp(-\lambda x)} \\
&= \frac{\lambda}{1+exp(-\lambda x)} * \frac{1+exp(-\lambda x)-1}{1+exp(-\lambda x)} \\
&= \frac{\lambda}{1+exp(-\lambda x)} * (1 - \frac{1}{1+exp(-\lambda x)}) \\
&= \lambda * f(x) * (1-f(x))
\end{aligned}
$$

(b)

$$
\begin{aligned}
f(x) &= \frac{2}{1+exp(-x)} - 1 = 2 * (1+exp(-x))^{-1} \\
f'(x) &= (-2) * (1+exp(-x))^{-2} * exp(-x) * (-1) \\
&= 2 * (1+exp(-x))^{-2} * exp(-x) \\
&= \frac{2}{1+exp(-x)} * \frac{exp(-x)}{1+exp(-x)} \\
&= \frac{2}{1+exp(-x)} * \frac{1+exp(-x)-1}{1+exp(-x)} \\
&= \frac{2}{1+exp(-x)} * (1 - \frac{1}{1+exp(-x)}) \\
&= \frac{1}{2}(\frac{2}{1+exp(-x)}) * (2 - \frac{2}{1+exp(-x)}) \\
&= \frac{1}{2}(1 + \frac{2}{1+exp(-x)} - 1) * (1 - \frac{2}{1+exp(-x)} + 1) \\
&= \frac{1}{2}(1+f(x)) * (1-f(x))
\end{aligned}
$$

3. Write down a general sigmoid function and its derivative.

$$
\begin{aligned}
f(x) &= \frac{|b-a|}{1+exp(-x)} + a = |b-a| * (1+exp(-x))^{-1} + a \\
f'(x) &= -|b-a| * (1+exp(-x))^{-2} * exp(-x) * (-1) \\
&= \frac{|b-a|}{1+exp(-x)} * \frac{exp(-x)}{1+exp(-x)} \\
&= \frac{|b-a|}{1+exp(-x)} * \frac{1+exp(-x)-1}{1+exp(-x)} \\
&= \frac{|b-a|}{1+exp(-x)} * (1 - \frac{1}{1+exp(-x)}) \\
&= \frac{1}{|b-a|}(\frac{|b-a|}{1+exp(-x)}) * (|b-a| - \frac{|b-a|}{1+exp(-x)}) \\
&= \frac{1}{|b-a|}(a + \frac{|b-a|}{1+exp(-x)} - a) * (b - \frac{|b-a|}{1+exp(-x)} - a) \text{ for } b \geq a \\
&= \frac{1}{|b-a|}(-a + f(x)) * (b - f(x))
\end{aligned}
$$

2 Exercise *(Backpropagation (4p))*

1. How to avoid local minima in backpropagation?

   If we repeat the training procedure several times, we will have gotten multiple different minima. Depending on the amount of times we have repeated it, the minimum of the local minima is then likely to be the global minimum as well.
   Another method to avoid local minima is simulated annealing: We add random noise to the weights to be able to "hop" out of local minima. Reducing the noise over time will help to not pass over the actual global minimum.

2. Explain the generalization and avoiding overfitting.

   Generalization is basically a measure for the goodness of the fit of our weights to the test data: How well do our weights perform for another data set with the same underlying distribution?
   After several updates during the training procedure, our weights will adapt very strongly to the training examples, and it may come to overfitting: Instead of accounting for noise in the training data and giving a general result, our weights can only explain the training data but not any given test data. Hence, the error for our training data set will steadily decrease while the error for the test data set may increase again after a certain

amount of weight updates.
Overfitting can be avoided using weight decay. If we stop the weights from growing indefinitely, they can no longer reach numbers that are highly spezialized to the examples in our training data set without giving good results for a test data set as well.

3. To prevent overly large weights which cause the high sensitivity of inputs, we apply the quadratic regularization term in the error function. Use gradient descent to minimize this error function.

   (See slide 48 of "ML-7 Neural Networks" for gradient descent on first summand.)

$$
\begin{aligned}
E[\{w\}] &= \frac{1}{2} \sum_{i=1..|D|} (t^i - y(x^i))^2 + \frac{\beta}{2} \sum_{w \in \{w\}} w^2 \\
\Delta w_{jk}(x) &= -\epsilon \frac{\delta E}{\delta w_{jk}(m,n)} \\
&= -\epsilon * - \sum_{i=1..|D|} (t^i - y(x^i))^2 \frac{\delta}{\delta w_{jk}(m,n)} y(x^i) + \sum_{w \in \{w\}} \beta w
\end{aligned}
$$