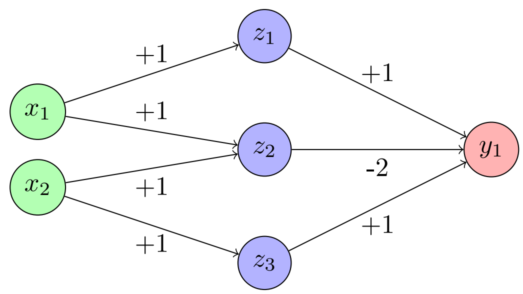
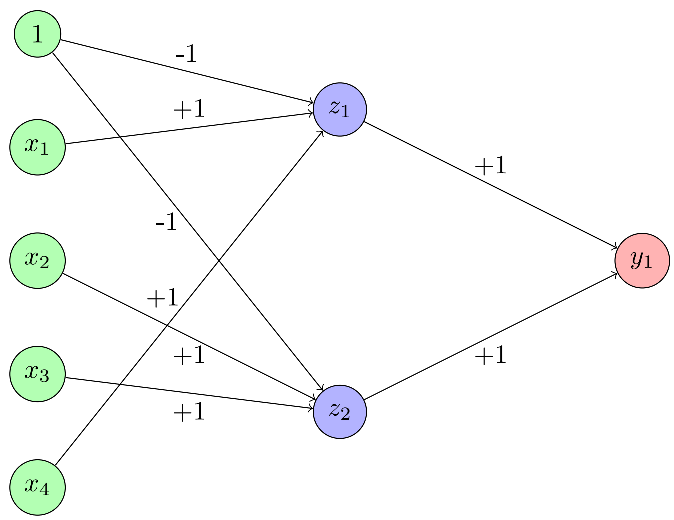
1 Exercise *(Multi-Layer Perceptron (8p))*

1. Draw multi-layer perceptron to solve each given logical function below. (We assume a threshold Θ with $0 < \Theta < 1$ for each neuron.)

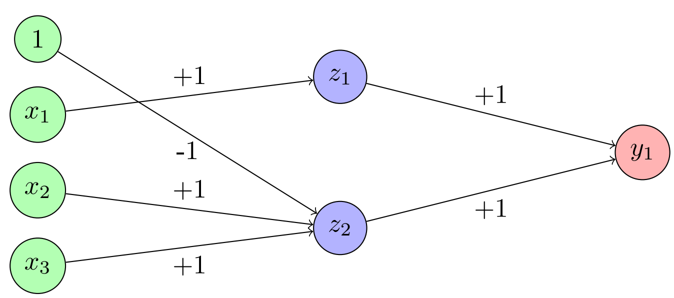
   (a) $x1 \oplus x2$



   (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)

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

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

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

2 Exercise *(Backpropagation (4p))*

1. How to avoid local minima in backpropagation?
2. Explain the generalization and avoiding overfitting.
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.