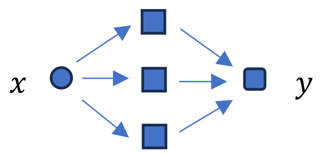
**AIS5103 Foundations of Deep Learning**

Midterm Test, 6 March 2026, closed book, 1 hr 45 min

1. Define or explain the following concepts briefly:
    a. What is a synapse, and how is information transmitted between biological neurons?
    b. Consider linear regression, $y = Wx$, here $W$ is a learnable matrix of dimension $E \times D$. Let $X$ be the design matrix and $e$ the corresponding error matrix between the model and the data. What is the meaning of the expression $e^T X$?
    c. Define cross-entropy and show that maximizing the likelihood is equivalent to minimizing the cross-entropy.
    d. What is a differential, and how is it related to the gradient?

2. As an illustration of the universal approximation theorem, consider the following simple two-layer neural network with one input unit $x$, one output unit $y$, and three hidden units. We train this network to fit $\sin x$, $|x|$, or Heaviside step function $\theta(x)$, in the closed interval $x \in [-1,1]$. Implement the training process in PyTorch using torch.nn, using any one of the given functions. Define the model class, implement the forward() method, use the mean-squared error loss, apply stochastic gradient descent (SGD) for optimization, and write the training loop.




3. The Kullback-Leibler divergence between two probability distributions $P$ and $Q$ can be expressed as $D(P||Q) = \langle \ln \frac{P}{Q} \rangle_P$. The angular brackets denote expectation $\langle f \rangle_P = E_P f$. Consider the die experiment in which we assume in theory (for $Q$) each of the six outcomes of 1 to 6 is equally probable. The actual probability $P$ may be different.
    a. Show that if $P = Q$, then $D(P||Q) = 0$.
    b. Compute the Kullback-Leibler divergence if $P = (1,0,0,0,0,0)$, i.e., the probability of getting 1 is 1, and all others are 0.
    c. Prove that $D(P||Q) \geq 0$, i.e., the Kullback-Leibler divergence cannot be negative.

4. Consider an error function that consists simply of the quadratic regularizer

$$E(\mathbf{w}) = \frac{1}{2}\mathbf{w}^T\mathbf{w}$$

together with the gradient update formula

$$\mathbf{w}^{(\tau+1)} = \mathbf{w}^{(\tau)} - \eta \nabla E(\boldsymbol{w}),$$

here $\mathbf{w}$ is a vector. By considering the limit of infinitesimal updates, write down the solution of this equation starting from an initial value $\mathbf{w}_0$, and show that the elements of $\mathbf{w}$ decay exponentially to zero.

- The end of Paper -