# 2004 Paper 9 Question 8

## Artificial Intelligence

In the following, $N$ is a feedforward neural network architecture taking a vector

$$\mathbf{x}^T = \begin{pmatrix} x_1 & x_2 & \cdots & x_n \end{pmatrix}$$

of $n$ inputs. The complete collection of weights for the network is denoted $\mathbf{w}$ and the output produced by the network when applied to input $\mathbf{x}$ using weights $\mathbf{w}$ is denoted $N(\mathbf{w}, \mathbf{x})$. The number of outputs is arbitrary. We have a sequence $\mathbf{s}$ of $m$ labelled training examples

$$\mathbf{s} = ((\mathbf{x}_1, \mathbf{l}_1), (\mathbf{x}_2, \mathbf{l}_2), \ldots, (\mathbf{x}_m, \mathbf{l}_m))$$

where the $\mathbf{l}_i$ denote vectors of desired outputs. Let $E(\mathbf{w}; (\mathbf{x}_i, \mathbf{l}_i))$ denote some measure of the error that $N$ makes when applied to the $i$th labelled training example. Assuming that each node in the network computes a weighted summation of its inputs, followed by an activation function, such that the node $j$ in the network computes a function

$$g\left(w_0^{(j)} + \sum_{i=1}^{k} w_i^{(j)} \text{input}(i)\right)$$

of its $k$ inputs, where $g$ is some activation function, derive in full the backpropagation algorithm for calculating the gradient

$$\frac{\partial E}{\partial \mathbf{w}} = \begin{pmatrix} \frac{\partial E}{\partial w_1} & \frac{\partial E}{\partial w_2} & \cdots & \frac{\partial E}{\partial w_W} \end{pmatrix}^T$$

for the $i$th labelled example, where $w_1, \ldots, w_W$ denotes the complete collection of $W$ weights in the network.

[20 marks]