(A) Rewrite the general form of the backpropagation equations given in class for MLPs for the following specific activation functions i) ReLU, ii) hyperbolic tangent and iii) sigmoid. For each activation function write down the range of the gradients.

(B) Download the MNIST data set and use it to train a fully-connected neural network to recognise handwritten digits. The MNIST data set can be loaded from *keras.datasets*. Make sure that your data is normalized. Each hidden layer should have 32 units and the output layer should have a softmax activation function. Compile your model to use the standard SGD optimizer with a learning rate of 0.01 and the categorical crossentropy loss function. Use (i) 5, (ii) 20 and (iii) 40 layers. For each choice (i), (ii) and (iii), use the (a) ReLU, (b) hyperbolic tangent and (c) sigmoid activation function on all hidden layers. Report the test scores for each model. What are your observations?

(C) For each of the models in (B), trained for 3 epochs on the MNIST data set, compute for each layer the maximum value of the gradient on a given mini-batch and create a plot of "layer depth vs. max gradient". Organize your plots as a grid so that your results for the different activation functions for each depth choice appear on the same subplot. Can you explain your observations? What insight do you gain for the observations in (B)?

(D) Train a model using the topology given in (B) and the activation function

$$LeCun(x) = 1.7159\ tanh(\frac{2}{3}x) + 0.01x.$$

Compare the learning curves of the models using LeCun and hyperbolic tangent activation functions. Write down the backpropagation equations and the gradient range for the LeCun activation function. Plot the gradients for the choices of depth given above for an untrained model using LeCun and hyperbolic tangent activations.