**Why is it generally preferable to use a Logistic Regression classifier rather than a classical Perceptron (i.e., a single layer of linear threshold units trained using the Perceptron training algorithm)? How can you tweak a Perceptron to make it equivalent to a Logistic Regression classifier?**

A classical Perceptron will converge only if the dataset is linearly separable, and it won't be able to estimate class probabilities.

In contrast, a Logistic Regression classifier will converge to a good solution even if the dataset is not linearly separable, and it will output class probabilities

**Why was the logistic activation function a key ingredient in training the first MLPs?**

The logistic activation function was a key ingredient in training the first MLPs because its derivative is always nonzero, so Gradient Descent can always roll down the slope. When the activation function is a step function, Gradient Descent cannot move, as there is no slope at all.

**Name three popular activation functions. Can you draw them?**

The step function, the logistic function, the hyperbolic tangent, the rectified linear unit

**Suppose you have an MLP composed of one input layer with 10 passthrough neurons, followed by one hidden layer with 50 artificial neurons, and finally one output layer with 3 artificial neurons. All artificial neurons use the ReLU activation function.**

**What is the shape of the input matrix X?**

**What about the shape of the hidden layer’s weight vector W*h*, and the shape of its bias vector b*h*?**

**What is the shape of the output layer’s weight vector W*o*, and its bias vector b*o*?**

**What is the shape of the network’s output matrix Y?**

**Write the equation that computes the network’s output matrix Y as a function of X, W*h*, b*h*, W*o* and b*o*.**

**How many neurons do you need in the output layer if you want to classify email into spam or ham? What activation function should you use in the output layer? If instead you want to tackle MNIST, how many neurons do you need in the output layer, using what activation function?**

To classify email into spam or ham, you just need one neuron in the output layer of a neural network — for example, indicating the probability that the email is spam.

**What is backpropagation and how does it work? What is the difference between backpropagation and reverse-mode autodiff?**

In contrast, reverse-mode auto diff is simply a technique used to compute gradients efficiently and it happens to be used by backpropagation. This is a sort of recursive definition: backpropagation consists of multiple backpropagation steps.

**Can you list all the hyperparameters you can tweak in an MLP? If the MLP overfits the training data, how could you tweak these hyperparameters to try to solve the problem?**

number of hidden layers

number of neurons in each hidden layer

activation function used in each hidden layer and in the output layer

the weight initialization logic

If the MLP overfits the training data, you can try reducing the number of hidden layers and reducing the number of neurons per hidden layer.

**Train a deep MLP on the MNIST dataset and see if you can get over 98% precision. Try adding all the bells and whistles (i.e., save checkpoints, restore the last checkpoint in case of an interruption, add summaries, plot learning curves using TensorBoard, and so on).**