2802ICT Assignment 2

# Introduction

The purpose of this document is to provide my understanding of the mechanisms of a neural network, including forward, back propagation and the results of manual calculation, and to compare the results of a digit classification by implementing this with different hyper-parameter settings. To enhance the performance, Kaiming initialisation (He et al., 2015[[1]](#footnote-1)) and learning rate decay are also implemented in this assignment.

The report is structured as follows: First, manual calculation of the simple neural network model is undertaken. Second, the network is implemented in Python to perform a digit recognition task based on mnist data. The implementation method is addressed while the results with the default hyper-parameters and with different settings are compared in this section. Lastly, the results with the cross-entropy loss function is compared to those with the quadratic cost function.

# Part 1: Manual calculation for a small neural network

The forward and backward processes in a small neural network is manually calculated using a calculator. The forward process consists of the linear and sigmoid functions for a hidden layer and an output layer.

The first forward process is a linear function as follows:

The manual calculation of the sample X1 and X2 is:

Then, results are passed through a sigmoid function:

It is noted that the results from the hidden layer of the second input of and the first input of is same till .

With the similar process from a hidden layer, the results of the output layer are as follow:

After the forward process, the final loss based on the quadratic cost function with the two samples is calculated as:

With the loss of the backpropagation results of the output layer are calculated as follow:

Individually, the chain rule above can be broken down to the following three forms:

Calculating all this together provides:

The loss of is then back-propagated to calculate the changes in weights of a hidden layer:

In particular, the previously calculated values can be used for the parts highlighted in yellow.

Calculating a hidden layer through a calculator provides:

After one manual backpropagation where is 0.1, the gradients are adjusted as follows:

# Part 2: Implementation in Python

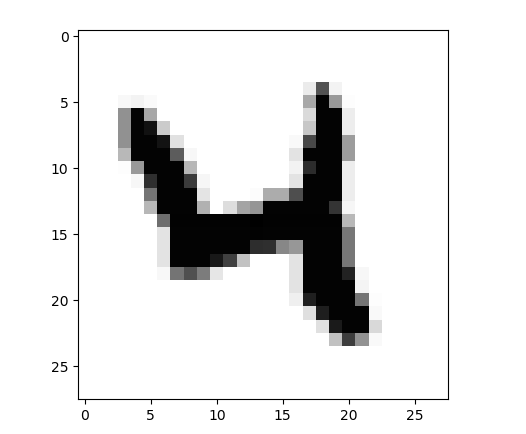
This part addresses the methods and results of the implementation in Python with the quadratic cost function.

## Implementation

The implementation of the neural network is done through matrix calculation for each batch. While an element-wise multiplication is used in some cases, a dot product is mainly utilised during the forward and backpropagation process where two examples in a batch need to be multiplied and added together.

Input matrix’s first dimension is the example size in a batch (e.g. default used 20 in mnist or 2 in Part A test case). Input matrix’s second dimension is 784 (28x28) representing pixels to make up a digit. The example of the input can be visualised by using the “visualise\_individual\_data” function in the source code submitted, displaying the digit in a plot (see Figure 1).

Figure 1. An example of the input data in a plot



Output matrix’s first dimension is also the example size in a batch. Output matrix’s second dimension is the number of classes/labels (e.g. 10 in mnist or 2 in Part A test case).

Input values are shuffled so that different digit numbers can be mixed within a training batch. However, a testing set does not get shuffled. Without standardising input values, the learning process is too slow. Thus, input values are standardised around mean 0.0 and standard deviation 1.0 for each batch prior to forwarding the network.

Similarly, without standardising weights and biases, the learning process is tedious. Therefore, random weights and biases are initialised with mean 0.0 and standard deviation 1.0. To aim for a better initialisation, Kaiming initialisation is implemented by multiplying to the random initialisation. The comparison of the performance is listed in the result section.

A function to reduce the learning rate as the epoch increases is implemented to ensure that the gradient steps are reduced when it is closer to the optimum. Output labels are vectorized through an one-hot encoding method to compare it with the prediction and obtain accuracy.

## Comparison with manual calculation

Testing to see whether the network is implemented correctly is undertaken by comparing the results of the manual calculation to the outputs of the network implemented in Python below. Except for the different precision, the results are consistent between the manual calculation and the outputs from the implementation.

Table 1. Comparison of the manual calculation and the outputs from the implementation after 1 epoch

| Result | Manual calculation | Output |
| --- | --- | --- |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |

Table 2 provides the weights and biases after each epoch up to 3 epochs per requirement in the assignment instruction document.

Table 2. Weights and biases after each epoch up to 3 epochs

| Weights and biases | Epoch 1 | Epoch 2 | Epoch 3 |
| --- | --- | --- | --- |
|  |  |  |  |
|  |  |  |  |
|  |  |  |  |
|  |  |  |  |
|  |  |  |  |
|  |  |  |  |
|  |  |  |  |
|  |  |  |  |
|  |  |  |  |
|  |  |  |  |
|  |  |  |  |
|  |  |  |  |

## Results

### Result with the default hyper-parameters

The default hyper-parameters for the result includes:

the number of input = 784

the number nodes in a hidden layer = 30

the number of output = 10

the number of epochs = 30

the number in a batch = 20

learning rate = 3.0

cost function: the quadratic cost function

Table 3. Accuracy result with the aforementioned default hyper-parameters

|  |  |  |
| --- | --- | --- |
| Cost function | Maximum accuracy using Random initialisation | Maximum accuracy using Kaiming initialisation |
| Quadratic cost function | 94.44 % | 94.81% |

Given all other parameter are equivalent, Kaiming initialisation provides marginally better accuracy. The network with Kaiming initialisation starts at higher accuracy and improves more quickly. However, regardless of different initialisation methods, the gap between the training and test accuracy becomes wider as the number of epochs increases. This may be due to overfitting starting to occur with the training data.

Figure 2. Accuracy over epoch – random initialisation

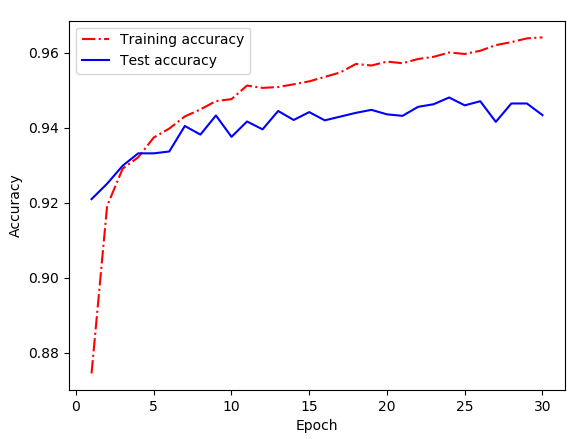
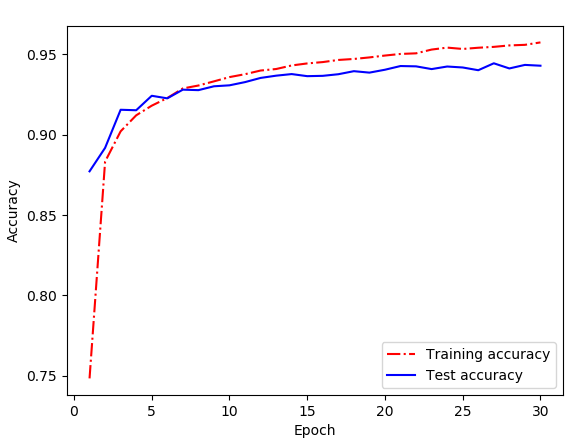


Figure 3. Accuracy over epoch – Kaiming initialisation



### Different learning rates

This section documents the experiments with different learning rates. The networks with random and Kaiming initialisation are examined.

With a small learning rate (lr) such as 0.001, the learning process is very slow (in particular, for the model with random initialisation) and it still displays underfitting symptoms after 30 epochs for both models. The lr 0.1 improves slower than the performance with the learning rate 3.0 but this learning rate seems to be appropriate (95.18%) as the accuracy reaches to a higher performance compared to the performance of 94.81% with lr = 3.0 for the network with Kaiming initialisation. The lr 1.0 is similarly effective while the lr 10 performs as worse as lr = 0.001.

When lr is 100, the network performance does not improve better than that of the smaller learning rates for both models. Lr of 100 performs terribly because the network cannot perform the gradient descent properly. With the too large learning rate, the model’s convergence cannot be made as it only goes back and forth near the similar points instead of descending towards the optimum.

Figure 4. Conceptual representation of the movement of the large learning rate

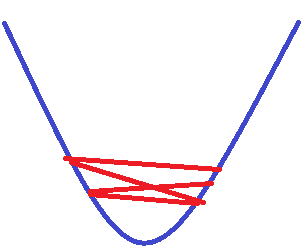
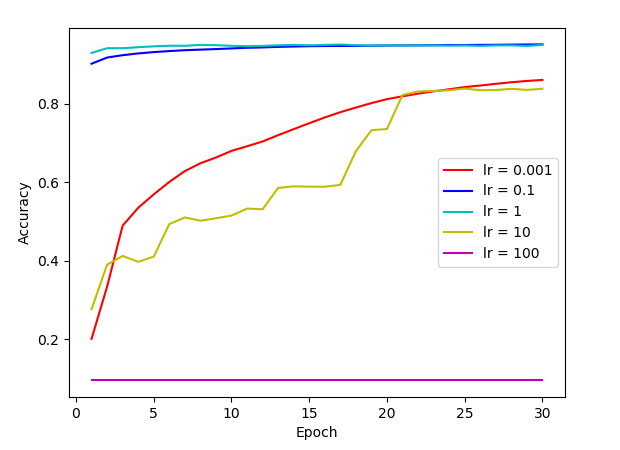


Table 4. Maximum accuracy result with different learning rates

| Learning Rate | Maximum accuracy using Random initialisation | Maximum accuracy using Kaiming initialisation |
| --- | --- | --- |
| 0.001 | 21.87% | 86.13% |
| 0.1 | 88.37% | 95.18% |
| 1.0 | 93.09% | 95.14% |
| 10 | 94.17% | 83.93% |
| 100 | 10.96% | 9.58% |

Overall, the learning rate of 0.1 and 1 provides the better accuracy for the model with Kaiming initialisation while the learning rate of 1 and 10 works better for the model with random initialisation.

Figure 5. Accuracy over epoch with different learning rates – Kaiming initialisation



### Mini-batch sizes

This section addresses the experiments with different batch sizes. The networks with random and Kaiming initialisation are compared.

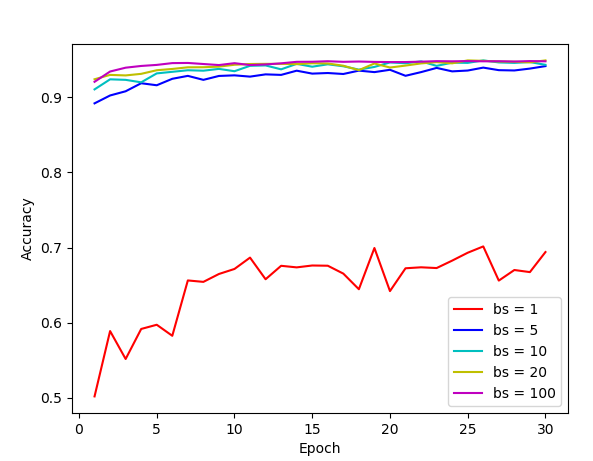
For both networks, the batch size of 1 displays a slow learning process and underfitting behaviours. For both models, as the size of each batch grows, the accuracy improves marginally until the batch size of 20. The best accuracy is shown with the batch size of 10 or 20. However, the batch size of 100 displays worse performance for both models.

Table 5. Maximum accuracy result with different batch sizes

| Batch size | Maximum accuracy using Random initialisation | Maximum accuracy using Kaiming initialisation |
| --- | --- | --- |
| 1 | 62.55% | 70.15% |
| 5 | 93.42% | 94.16% |
| 10 | 94.57% | **94.92%** |
| 20 | 93.97% | **94.92%** |
| 100 | 91.87% | 94.83% |

Overall, the model with Kaiming initialisation seems to deal with the larger the batch size better. In addition, the batch size of 1 is slowest in the matrix operation of this neural network.

Figure 6. Accuracy over epoch with different batch sizes – Kaiming initialisation



### Results with combinations of different hyper-parameters

This section provides the results of the combinations of different hyper-parameters for the model with Kaiming initialisation. The quadratic cost function is used for all cases.

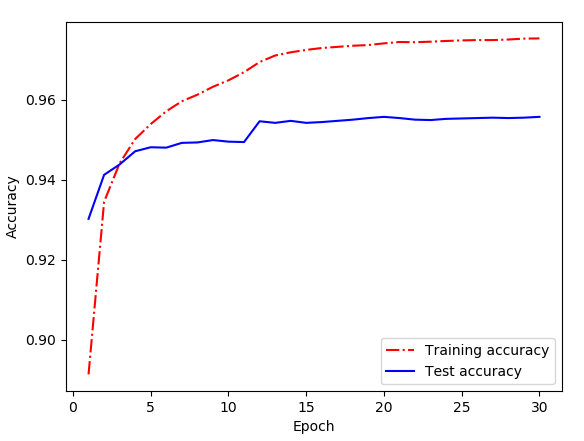
To enhance the maximum accuracy, the learning rate decay mechanism is adopted. The learning rate is decreased every 10th epoch to ensure when it approaches the optimal point, the gradient descent takes a smaller step so that it will not pass over the optimum.

Table 6. Maximum accuracy with different combinations of hyper-parameters

|  |  |
| --- | --- |
| Combination | Maximum accuracy using Kaiming initialisation |
| Default: epochs=30, lr=3.0, bs=20, no lr decay | 94.81% |
| epochs=30, lr=3.0, bs=20, lr decay | 95.04% |
| epochs=30, lr=1.0, bs=20, lr decay | 95.56% |
| epochs=30, lr=0.5, bs=20, lr decay | 95.37% |
| epochs=30, lr=1.0, bs=25, lr decay | **95.57%** |
| epochs=30, lr=1.0, bs=50, lr decay | 95.41% |
| epochs=30, lr=1.0, bs=100, lr decay | 95.13% |

In summary, **the maximum accuracy, 95.5700% can be achieved with the hyper-parameter settings of epochs=30, lr=1.0, bs=25, and lr decay**. However, the overfitting pattern seems to be noticeable around this maximum accuracy as seen in Figure 7.

Figure 7. Accuracy over epoch with epochs=30, lr=1.0, bs=25, and lr decay.



# Part 3: An alternate cost function

### Results with the default hyper-parameters

The default hyper-parameters for this result equal to those from the model with the quadratic cost function. However, in this experiment, the cross-entropy cost function is used. The cross-entropy cost function is defined as:

The both networks achieve a similar accuracy with the cross-entropy cost function. Despite small improvements in the accuracy for the model with Kaiming initialisation, according to Nielsen (2018)[[2]](#footnote-2), the cross-entropy cost function can be resilient to the issues associated with the quadratic cost function’s slowdown of learning when using many-layer and multi-neuron network. Hence, using the cross-entropy cost function makes sense when there is a neural network with many layers and nodes.

Table 7. Maximum accuracy with the aforementioned default hyper-parameters

|  |  |  |
| --- | --- | --- |
| Cost function | Maximum accuracy using Random initialisation | Maximum accuracy using Kaiming initialisation |
| Quadratic cost function | 94.44 % | 94.81% |
| Cross-entropy cost function | 94.13% | 95.61% |

Figure 8. Accuracy over epoch – random initialisation, cross-entropy cost function

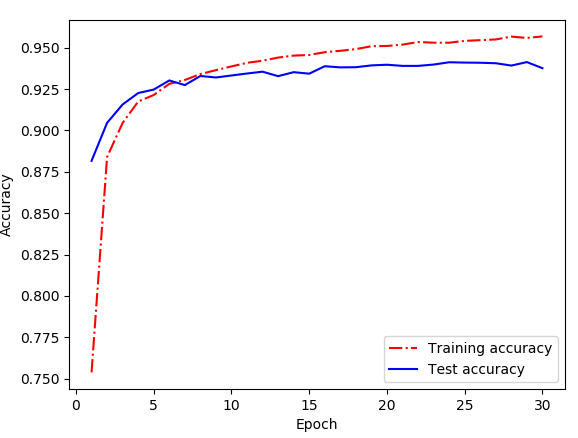
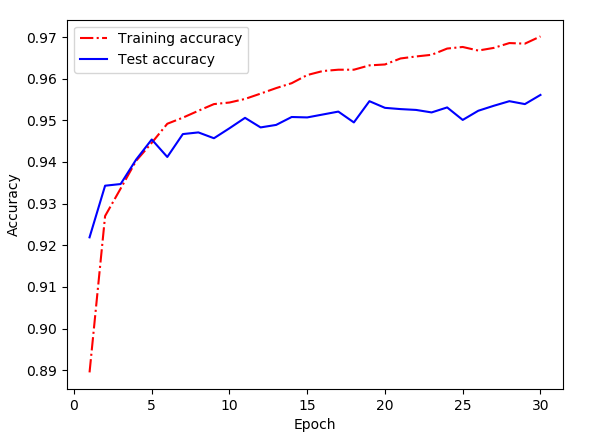


Figure 9. Accuracy over epoch – Kaiming initialisation, cross-entropy cost function



### Different learning rates with cross-entropy cost function

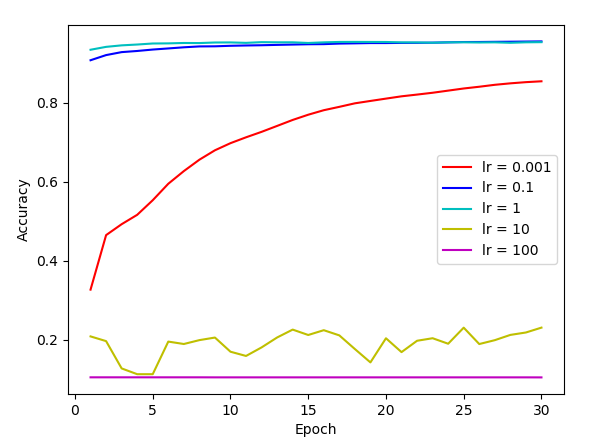
This section explores the performance of the model with a new cost function by using different learning rates. All other hyper-parameters are set per requirements: the number of input = 784, the number nodes in a hidden layer = 30, the number of output = 10, the number of epochs = 30, and the number in a batch = 20.

With a learning rate (lr) of 0.001, the learning process is slower and there is still room to train more on the data set after 30 epochs. lr 0.1 and 1.0 provides the best performance with a new cost function, which is consistent with the result from the quadratic cost function. However, if lr is larger than 10, the performance dramatically declines as shown in Table 6.

Table 8. Accuracy result with different learning rates with cross-entropy cost function

| Learning Rate | Maximum accuracy using Kaiming initialisation |
| --- | --- |
| 0.001 | 85.38% |
| 0.1 | 95.46% |
| 1.0 | 95.27% |
| 10 | 23.12% |
| 100 | 10.56% |

Figure 10. Accuracy over epoch with different learning rates – cross-entropy cost function



### Mini-batch sizes with cross-entropy cost function

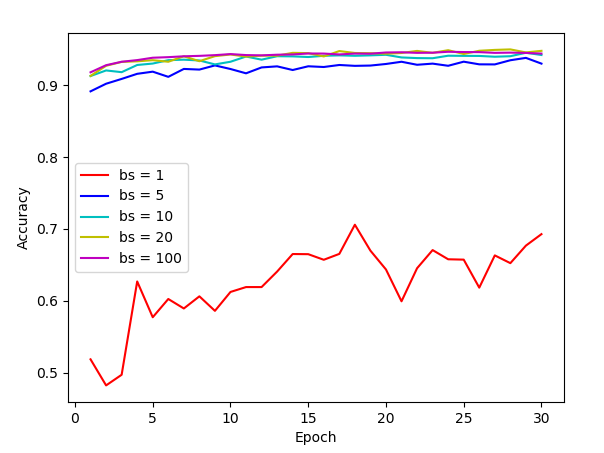
The various mini-batch sizes are tested to identify the batch size contributing to the best performance for a new cost function.

Similar to the experiments undertaken with the quadratic cost function, the accuracy for batch size of 10 or 20 is better while batch size of 1, 5 and 100 are performing worse. The batch size of 1 performs slowest for the network.

Table 9. Accuracy result with different batch sizes with cross-entropy cost function

|  |  |
| --- | --- |
| Batch size | Maximum accuracy using Kaiming initialisation |
| 1 | 69.27% |
| 5 | 93.81% |
| 10 | 94.52% |
| 20 | **94.99%** |
| 100 | 94.66% |

Figure 11. Accuracy over epoch with different batch sizes – cross-entropy cost function



### Results with combinations of different hyper-parameters

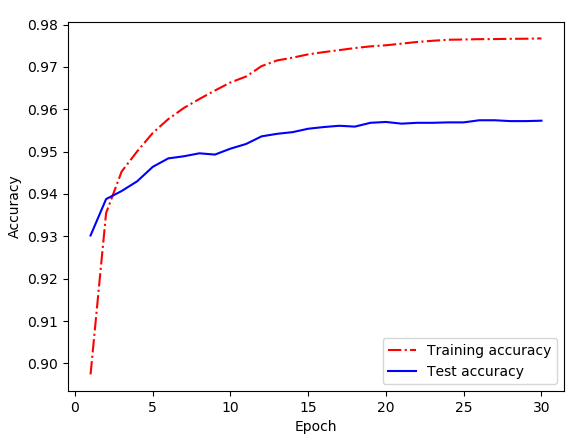
This section shows attempts to enhance the maximum accuracy with additional changes to the hyper-parameters: the cost function and the number of epochs. The best accuracy is achieved through the increased hidden nodes of 90 instead of 30.

Table 10. Maximum accuracy with different combinations of hyper-parameters

| Combination | Maximum accuracy using Kaiming initialisation |
| --- | --- |
| Previous best: epochs=30, lr=1.0, bs=25, lr decay with the quadratic cost function | 95.57% |
| epochs=30, lr=5.0, bs=20, lr decay with the cross-entropy cost function | 94.74% |
| epochs=30, lr=3.0, bs=20, lr decay with the cross-entropy cost function | 95.26% |
| epochs=30, lr=1.0, bs=20, lr decay with the cross-entropy cost function | **95.74%** |
| epochs=30, lr=0.5, bs=20, lr decay with the cross-entropy cost function | 95.71% |
| epochs=30, lr=1.0, bs=25, lr decay with the cross-entropy cost function | 95.46% |
| epochs=30, lr=1.0, bs=20, lr decay, hidden nodes=60 with the cross-entropy cost function | 96.87% |
| epochs=60, lr=1.0, bs=20, lr decay, hidden nodes=60, with the cross-entropy cost function | 96.89% |
| epochs=60, lr=1.0, bs=20, lr decay, hidden nodes=90, with the cross-entropy cost function | **97.20%** |

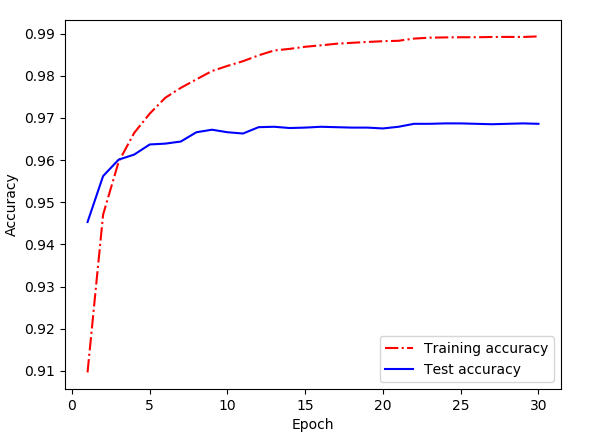
**Without increasing the number of nodes in a hidden layer, the maximum accuracy is 95.74% with epochs=30, lr=1.0, bs=20, and lr decay with the cross-entropy cost function**, as shown below.

Figure 12. Accuracy over epoch with epochs=30, lr=1.0, bs=20, and lr decay with the cross-entropy cost function



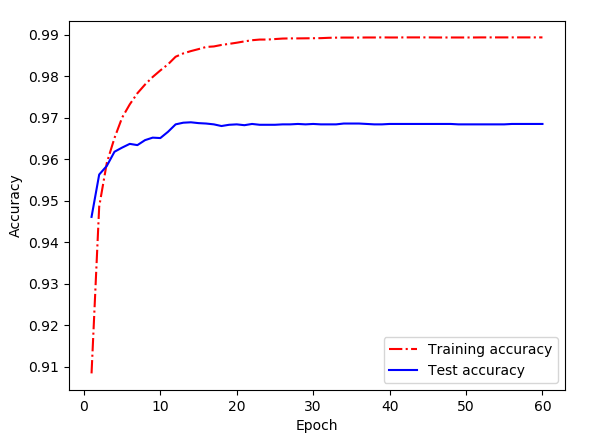
The larger the number of hidden nodes is, the better the maximum accuracy is.

Figure 13. Accuracy over epoch with epochs=30, lr=1.0, bs=20, lr decay, hidden nodes=60, with the cross-entropy cost function



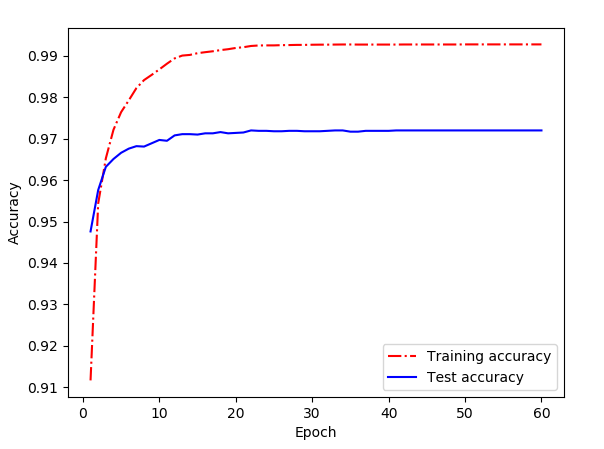
Increasing the number of epochs does not seem to make a larger difference due to continual overfitting.

Figure 14. Accuracy over epoch with epochs=60, lr=1.0, bs=20, lr decay, hidden nodes=60, with the cross-entropy cost function



However, **the increased hidden node of 90 further provides better accuracy of 97.20% on the test dataset**. The accuracy of 97.20% is reached within 30 epochs, thus, the number of epochs over 30 does not make a difference.

Figure 15. Accuracy over epoch with epochs=60, lr=1.0, bs=20, lr decay, hidden nodes=90, with the cross-entropy cost function



# Conclusion

In summary, it seems to be important to identify appropriate hyper-parameters for a given data set and a model for a task through experiments. For the mnist dataset, the neural network with one hidden layer and one output layer performs better with a cross-entropy cost function, Kaiming initialisation, learning rate decay, and the settings of epochs=30, lr=1.0, bs=20, and hidden nodes=90.

1. He, K, Zhang, X., Ren, S, and Sun, J, Delving Deep into Rectifiers: Surpassing Human-Level Performance on ImageNet Classification (https://arxiv.org/abs/1502.01852) [↑](#footnote-ref-1)
2. http://neuralnetworksanddeeplearning.com/chap3.html [↑](#footnote-ref-2)