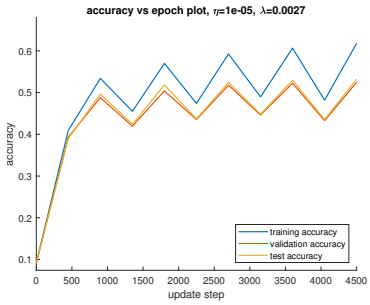
1 Optimizing the performance of the network

For this task, the full data set, minus 1000 data points for validation, was used to train the network, and the `test_batch.mat` was used for testing. The data was manipulated to be zero-mean, and the regularisation parameter λ was set to 0.027, which was the optimal value from the previous tasks in assignment 2. The following ways to optimise the network were examined:

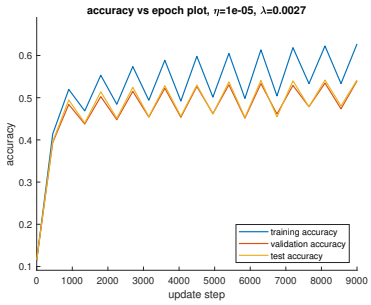
- Save W and b after each cycle and for use as ensemble, with majority voting.
- Increase number of hidden nodes from 50 to 500, with some values in between.
- Apply dropout to regularise the high number of hidden nodes.

Table 1: Best accuracies for different number of hidden nodes, with varying degree of dropout, and with ensemble majority vote accuracy on the full test data set. Note that the best accuracy does not necessarily correspond to the accuracy at the end of the last cycle.

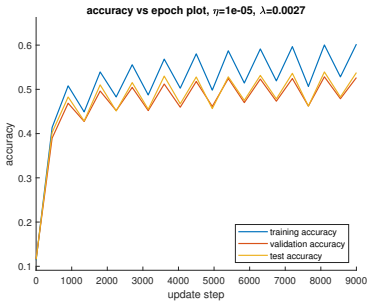
Dropout rate	Hidden nodes	Ensemble test accuracy	Test accuracy	Number of cycles
20 %	500	56.28 %	55.99 %	10
40 %	500	54.74 %	54.35 %	10
30 %	250	54.20 %	54.56 %	10
40 %	250	53.98 %	53.93 %	10
0 %	**500**	**56.44** %	**56.37** %	**10**
0 %	50	53.27 %	50.44 %	2
0 %	100	53.22 %	54.26 %	10
Previous	assignment	-	51.98 %	6



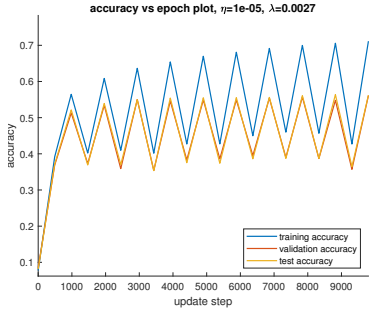
(a) Accuracy vs update step, 100 hidden nodes without dropout.



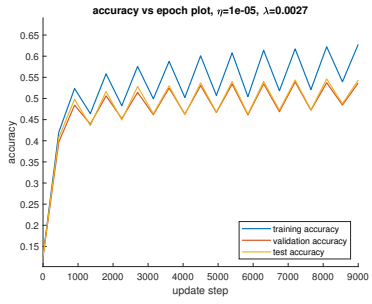
(b) Accuracy vs update step, 250 hidden nodes with 30% dropout.



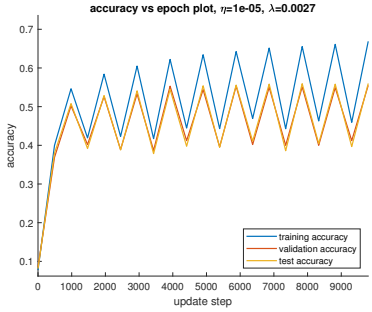
(c) Accuracy vs update step, 250 hidden nodes with 40% dropout.



(d) Accuracy vs update step, 500 hidden nodes without dropout.



(e) Accuracy vs update step, 500 hidden nodes with 40% dropout.



(f) Accuracy vs update step, 500 hidden nodes with 20% dropout.

Figure 1: Accuracy vs update steps for different rates of dropout, and number of hidden nodes.

From what we can see in Table 1, both increasing the number of hidden nodes, and performing ensemble voting improved the accuracy greatly, generally with at least 2 percentage units. We also see that by increasing the number of hidden nodes we gain the most in accuracy, but at 500 hidden nodes without any dropout (best accuracy), the netwprk also tends to overfit quite heavily. By introducing dropout to regularise the network, we lose some accuracy, but do indeed decrease overfitting. 40% does seem to be a bit too much dropout for this case with regard to accuracy, so 20% dropout with 500 hidden nodes seem to give the best accuracy.

With regard to ensemble models, the ensemble tended to increase the accuracy as compared to no improvement with 1-2 percentage units, and also gave more "reliable" results. When combining ensemble with more hidden nodes, the improvement over normal best test accuracy was still present, although not a super large improvement.

So to summarise, increase number of hidden nodes, use dropout, and use ensemble.

2 Optimising cyclical learning rate

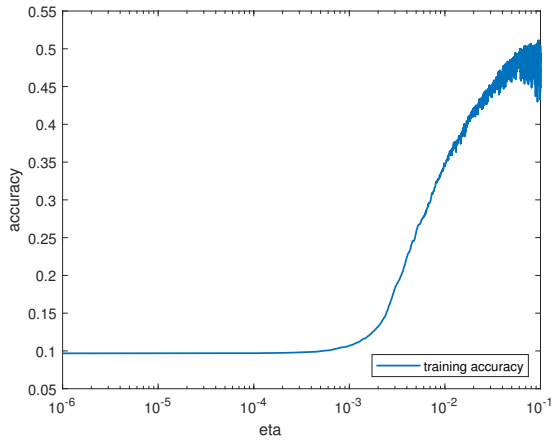
Following the guidelines in Smith (2015), η_{min} and η_{max} were set to $1e-6$ and 1.0 respectively, and trained the network for half a cycle (so that η increases monotonically) with 2000 updates per cycle. Then, training accuracy was plotted against η, to see for which η the accuracy began increasing, and for which η the accuracy would begin to decrease. For this experiment, 100 hidden nodes were used, with 10% dropout rate was used to train the network. The full dataset of 50000 samples was used to find an optimal interval for η. Additionally, an experiment with 500 hidden nodes with 50% dropout was also done, to find optimal learning rate parameters for a larger number of hidden nodes. As with the prevous task (5.1), λ was set to 0.0027.

What we can see from the accuracy plots in Figure 2 is that the accuracy only start to climb significantly for $\eta > 1e-4$, and plateaus or starts to decrease for $\eta > .05$. The accuracy plots for both 100 and 500 hidden nodes were fairly similar, despite different dropout rates and number of hidden nodes. As per the suggestions of Smith (2015), η_{min} and η_{max} were chosen from the accuracy vs η plots for when accuracy starts to increase, and plateau, respectively. Using this method, I selected $\eta_{min} = 0.0065$, $\eta_{max} = 0.08$ and $\eta_{min} = 0.0060$, $\eta_{max} = 0.07$ for the network with 100 and 500 hidden nodes respectively.

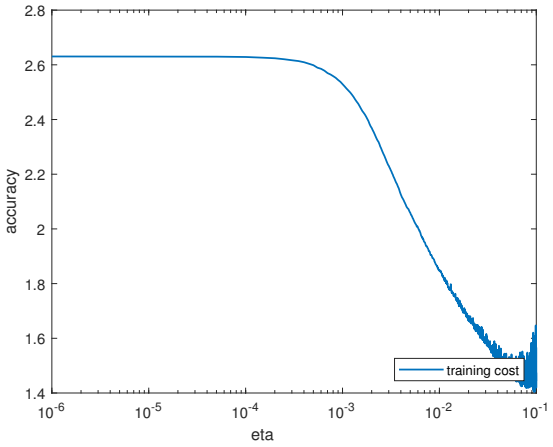
Table 2: Best test accuracies using improved parmeter settings for η_{min} and η_{max}, for different number of hidden nodes and dropout rates. The networks were trained for 10 cycles. Note that the best test accuracy does not necessarily correspond to the accuracy at the end of training.

Hidden nodes	Droupout	η_{min}	η_{max}	Ensemble test accuracy	Test accuracy
100	0 %	0.00065	0.008	57.71 %	54.99 %
100	10 %	0.00065	0.008	57.93 %	55.15 %
100	20 %	0.00065	0.008	56.65 %	55.00 %
500	**0 %**	**0.00060**	**0.007**	**59.36 %**	**57.71 %**
500	20 %	0.00060	0.007	58.37 %	57.87 %
500	40 %	0.00060	0.007	57.90 %	57.20 %

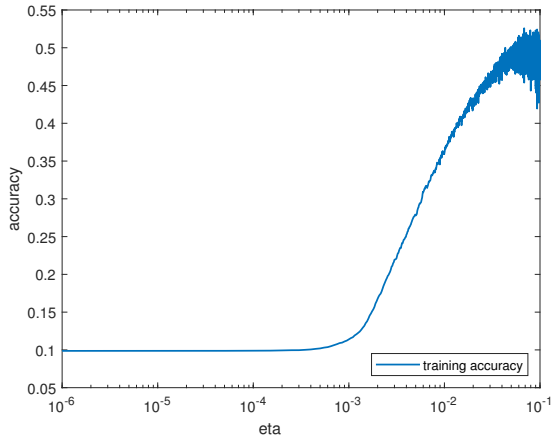
After selecting appropriate η_{min} and η_{max}, networks were trained using 100 or 500 hidden nodes, with variable dropout rates and otherwise similar settings as in task 5.1. The best test accuracies can



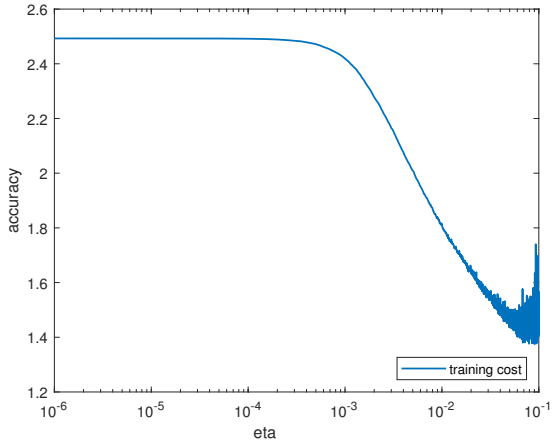
(a) Accuracy (training) plot when training network for half a cycle with 100 hidden nodes.



(b) Cost (training) plot when training network for half a cycle with 100 hidden nodes.



(c) Accuracy (training) plot when training network for half a cycle with 500 hidden nodes.

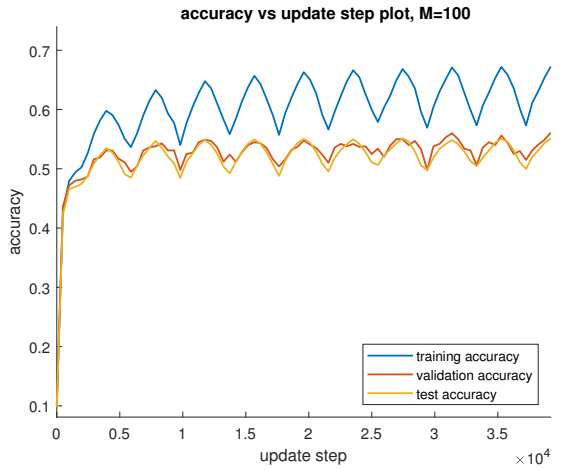


(d) Cost (training) plot of when training network for half a cycle with 500 hidden nodes.

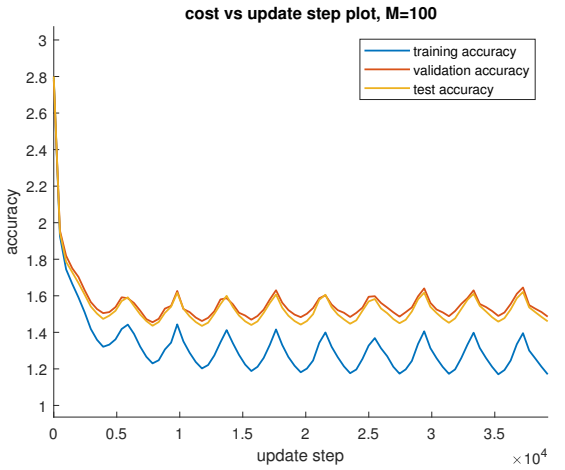
Figure 2: Plots accuracy and cost of training data as function of η while training half a cycle to measure which η_{min} and η_{max} to use for cyclical learning. The x-axis is on a logarithmic scale.

be seen in Table 2. Clearly, the chosen values for η_{min} and η_{max} improved classification accuracy significantly, and overall improving test accuracy with 2 percentage units across the board if compared with Table 1.

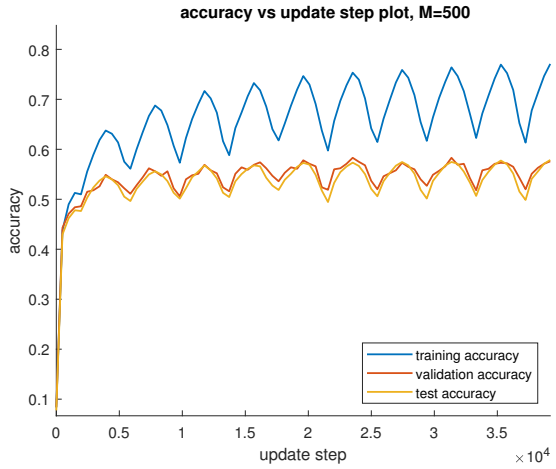
The jaggedness of the plots in Figure 3 can be explained by having a higher resolution on the X axis for this task. Otherwise the results and the plots are hardly surprising. I am very satisified with the results produced by my network.



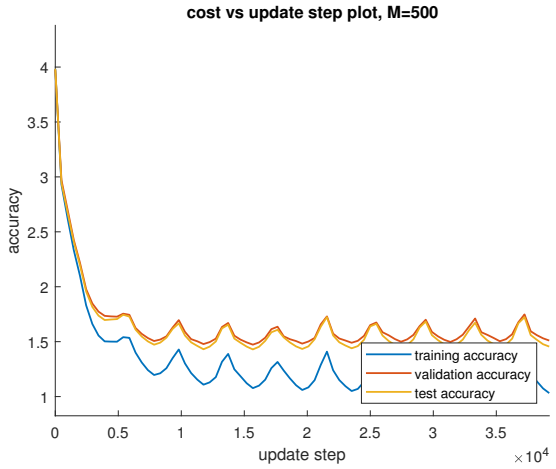
(a) Accuracy plot when training network for 10 cycles with 100 hidden nodes.



(b) Cost plot when training network for 10 cycles with 100 hidden nodes.



(c) Accuracy plot when training network 10 cycles with 500 hidden nodes.



(d) Cost plot of when training network for 10 cycles with 500 hidden nodes.

Figure 3: Plots accuracy and cost of training, validation and test data when training for 10 cycles, using improved values for η_{min} and η_{max}.