Question 1. Depths, Widths, and Learning Rates.

a) b) c) Please see the code.

d) We were able to observe that the increase in the depth of our model produced a moderate improvement in accuracy. Width increase proved less reliable which is in line with the theoretical results that emphasize "power" of depth compared to width (Lu, Zhou, Hongming Pu, Feicheng Wang, Zhiqiang Hu, and Liwei Wang. "The Expressive Power of Neural Networks: A View from the Width." In Advances in Neural Information Processing Systems, 2017-Decem:623240, 2017.).

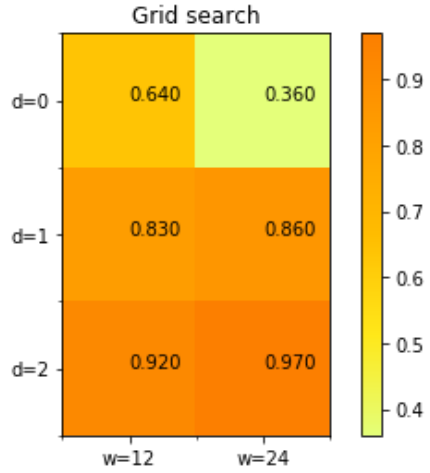
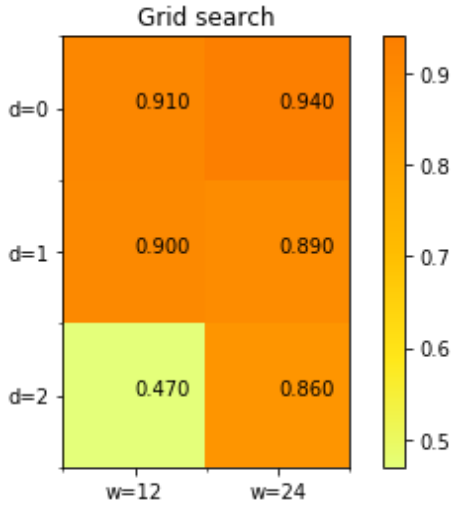
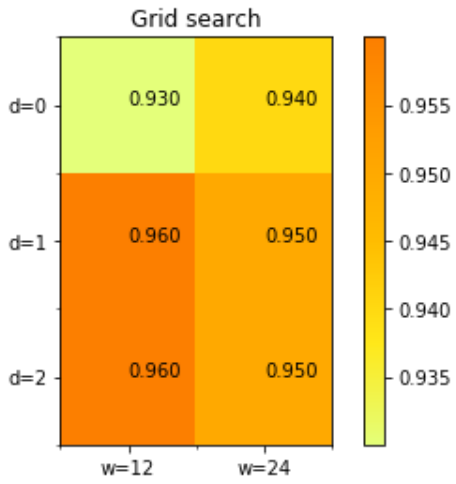


Figure 1: Grid search on the non-normalised dataset with the learning rate of 0.01

e) After normalizing we observed something quite surprising: shallower networks did much better than the deep ones. This effect reminded us of what happened in grid-search2 upon changing the learning rate and lead us to believe that the standardization effectively changed the timescale of the cost-function. Indeed, changing the learning-rate to 0.1 brought about a similar pattern as we saw before the normalization.



(a) Grid search on the normalised dataset with the learning rate of 0.01



(b) Grid search on the normalised dataset with the learning rate **changed to 0.1**

f) As one would expect, accuracy at the learning rate value of 0.1 was better than that at 1, possibly to the rather coarse convergence of the latter. Accuracy also increased with increasing width for the learning rate of 0.1. However, the performance of the model with the learning rate of 0.001 was worse. This was due to a shortage of training epochs, for with the smaller learning rate the model parameters did not have enough time to converge for optimal performance.

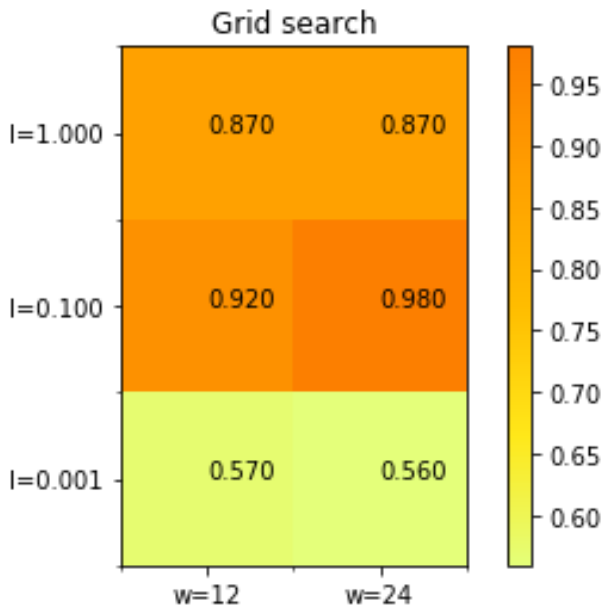


Figure 3: Grid search on the normalised dataset over a range of widths and learning rates with the depth set to 1.