Question 2

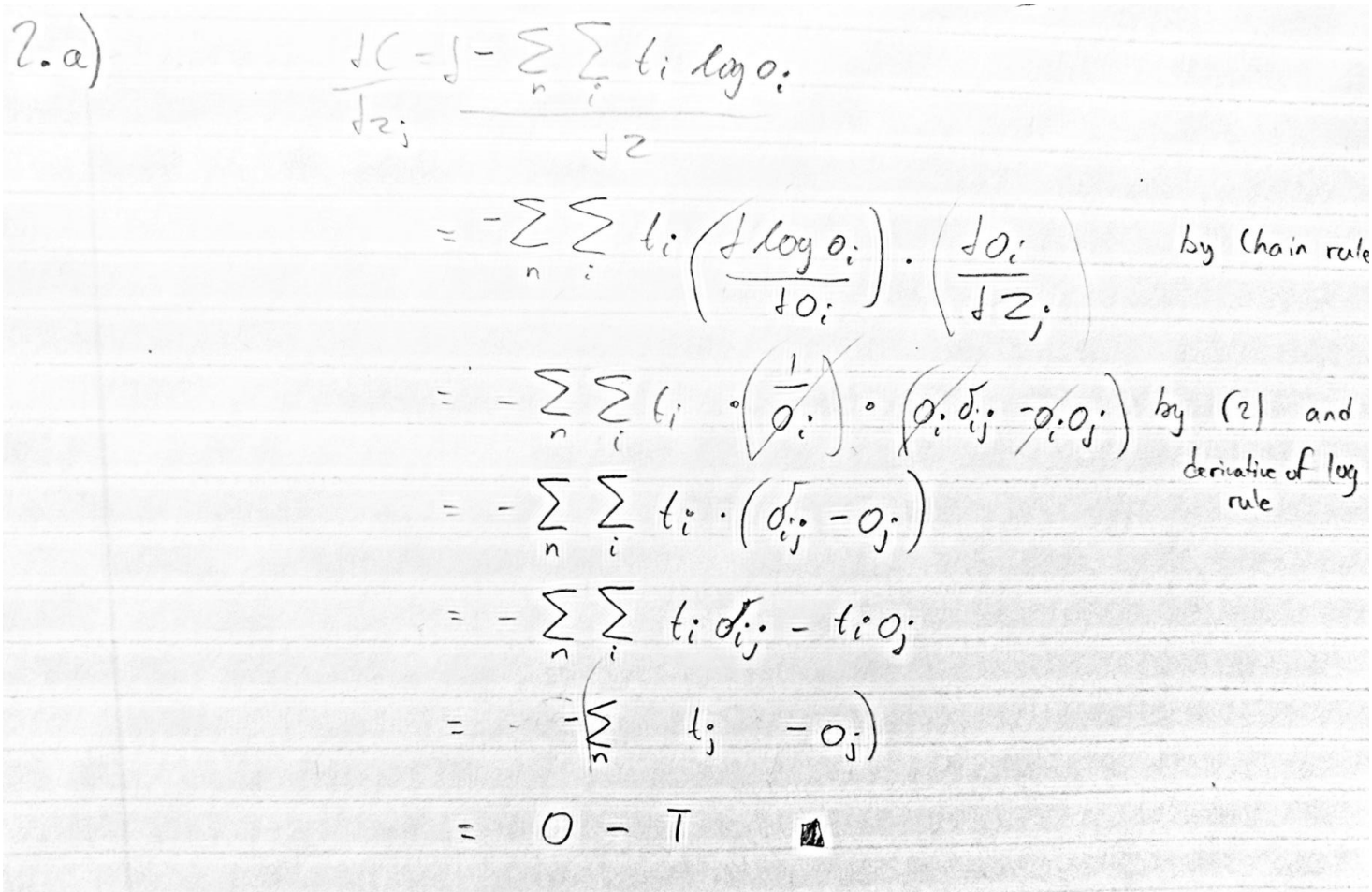
1(b) The decision boundary is a straight line because an MLP Classifier with only one hidden layer behaves as a linear classifier and does not have enough complexity capacity to model a nonlinear decision boundary.

1(f) The best test accuracies in parts (c), (d), and (e) should all be less than in Question 4(c) of Assignment 2 because the MLPClassifiers in parts (c), (d), and (e) are still limited in the complexity of the decision curves they model by the relatively low number of hidden units of two, three, and four respectively. With this number of hidden units, the Gaussian Bayes classifier in Question 4(c) of Assignment 2 can model more complex curves and will lead to a higher accuracy.

1(g) You get so many different decision boundaries in parts (c), (d), and (e) because the weight matrices get initialized with random values at the beginning of every training instance and can descend into different local minimums than each other, resulting in different decision curves.

Question 2

2(a)



2(h)

```
dCdw0 = np.sum(dCdZ, axis=0)
```

Question 3

3(b) The accuracy is lower than in part (a) because batch gradient descent (BGD) takes longer to reach the minimum than stochastic gradient descent (SGD) in part (a). Another explanation is due to the fact that SGD deals better with non-convex and non-smooth error space than BGD. While BGD moves fairly accurately towards the local or global minimum in its region of error space, SGD moves more randomly because of the noisy mini-batches of samples and can be jerked out of a local minimum into hopefully the global minimum.

3(d) Using the average gradient means that the optimal learning rate does not change much when the size of the training set changes because averaging the gradient keeps the gradient magnitude independent of batch size.