

## Part I: Pen and paper

For questions in this group, show your numerical results with 5 decimals or scientific notation. Hint: we highly recommend the use of numpy (e.g., `linalg.pinv` for inverse) or other programmatic facilities to support the calculus involved in both questions (1) and (2).

Below is a training dataset  $D$  composed by two input variables and two output variables, one of which is numerical ( $y_{\text{num}}$ ) and the other categorical ( $y_{\text{class}}$ ). Consider a polynomial basis function  $\phi(y_1, y_2) = y_1 \times y_2$  that transforms the original space into a new one-dimensional space.

D	$y_1$	$y_2$	$y_{\text{num}}$	$y_{\text{class}}$
$x_1$	1	1	1.25	B
$x_2$	1	3	7.0	A
$x_3$	3	2	2.7	C
$x_4$	3	3	3.2	A
$x_5$	2	4	5.5	B

1. Learn a regression model on the transformed feature space using the OLS closed form solution to predict the continuous output variable  $y_{\text{num}}$ .

The first step will be to calculate the transformed feature space for each observation using the polynomial basis function:

$$\phi(y_1, y_2) = y_1 \times y_2 \tag{1}$$

By applying (1) to each observation, we get the following transformed feature space:

$$\begin{aligned} \phi(x_1) &= 1 \times 1 = 1 & \phi(x_2) &= 1 \times 3 = 3 & \phi(x_3) &= 3 \times 2 = 6 \\ \phi(x_4) &= 3 \times 3 = 9 & \phi(x_5) &= 2 \times 4 = 8 \end{aligned}$$

We end up with the following transformed feature space:  $\phi(y_1, y_2) = [1, 3, 6, 9, 8]$

The regression model in the transformed feature space is given by:

$$z = w_0 + w_1 \cdot \phi(y_1, y_2)$$

The next step is to learn a regression model using the OLS closed form solution which, for this exercise, is given by:

$$W = (\Phi^T \Phi)^{-1} \Phi^T y_{\text{num}}$$

where  $\Phi$  is the design matrix and  $y_{\text{num}}$  is the continuous output variable.

$$\Phi = \begin{bmatrix} 1 & 1 \\ 1 & 3 \\ 1 & 6 \\ 1 & 9 \\ 1 & 8 \end{bmatrix} \quad y_{\text{num}} = \begin{bmatrix} 1.25 \\ 7.0 \\ 2.7 \\ 3.2 \\ 5.5 \end{bmatrix}$$

As suggested in the hint, we will use `numpy` to calculate the weights  $W$ :

```
1 import numpy as np
2
3 M = np.array([[1, 1], [1, 3], [1, 6], [1, 9], [1, 8]])
4 y_num = np.array([1.25, 7.0, 2.7, 3.2, 5.5])
5
6 W = np.linalg.inv(M.T @ M) @ M.T @ y_num
7
8 W = np.round(W, 5) # round to 5 decimal digits
9 print(W)
```

The output of the code above is `[3.315930.11372]`.

Therefore, the weights are  $w_0 = 3.31593$  and  $w_1 = 0.11372$ .

The regression model is then given by:

$$y_{\text{num}} = 3.31593 + 0.11372 \cdot \phi(y_1, y_2)$$

2. Repeat the previous exercise, but this time learn a Ridge regression with penalty factor  $\lambda = 1$ . Compare the learnt coefficients with the ones from the previous exercise and discuss how regularization affects them.

The Ridge regression model is given by:

$$W = (\Phi^T \Phi + \lambda I)^{-1} \Phi^T y_{\text{num}}$$

where  $\Phi$  is the design matrix,  $y_{\text{num}}$  is the continuous output variable (which remain the same as the previous exercise), and  $\lambda$  is the penalty factor (this being the only difference compared to the first exercise).

```
1 import numpy as np
2
3 M = np.array([[1, 1], [1, 3], [1, 6], [1, 9], [1, 8]])
4 lambda_value = 1.0
5 I = np.array([[1, 0], [0, 1]]) # Identity matrix
6 y_num = np.array([1.25, 7.0, 2.7, 3.2, 5.5])
7
8 W = np.linalg.inv(M.T @ M + lambda_value * I) @ M.T @ y_num
9
10 W = np.round(W, 5) # round to 5 decimal digits
11 print(W)
```

The output of the code above is [1.818090.32376].

Therefore, the weights are  $w_0 = 1.81809$  and  $w_1 = 0.32376$ .

The Ridge regression model is then given by:

$$y_{\text{num}} = 1.81809 + 0.32376 \cdot \phi(y_1, y_2)$$

In order to compare the weights obtained in the OLS and Ridge regression models, we must first normalize the data:

normalization for the OLS model:

$$w_{0\text{normalized}} = \frac{3.31593}{3.31593 + 1.81809} = 0.64587$$

$$w_{1\text{normalized}} = \frac{0.11372}{0.11372 + 0.32376} = 0.25994$$

Normalization for the Ridge Regression model:

$$w_{0\text{normalized}} = \frac{1.81809}{3.31593 + 1.81809} = 0.35413$$

$$w_{1\text{normalized}} = \frac{0.32376}{0.11372 + 0.32376} = 0.74006$$

Comparing the normalized weights obtained in the OLS and Ridge regression models, we can see that in the Ridge regression model  $w_0$  is smaller and  $w_1$  is larger than the corresponding OLS model weights.

The regularization term in the Ridge regression model penalizes large weights, but it allows for the weights to increase or decrease relative to one another based on the data. In this case, the regularization reduces the bias term  $w_0$ , as the model learns to rely more on the feature  $w_1$  to explain the variance in the target variable. This reduces the need for a large bias term, helping to prevent overfitting and stabilizing the model's predictions.

These results suggest that the feature has a strong positive correlation with the target variable. As a result,  $w_1$  increases to better capture this relationship, allowing  $w_0$  to decrease.

### 3. Given three new test observations and their corresponding output

$\mathbf{x}_6 = (2, 2, 0.7)$ ,  $\mathbf{x}_7 = (1, 2, 1.1)$ , and  $\mathbf{x}_8 = (5, 1, 2.2)$ , compare the train and test RMSE of the two models obtained in (1) and (2). Explain if the results go according to what is expected.

The Root Mean Squared Error (RMSE) is given by:

$$\text{RMSE} = \sqrt{\frac{1}{N} \sum_{i=1}^N (z_i - \tilde{z}_i)^2}$$

where  $\tilde{z}_i$  is the predicted value,  $z_i$  is the true value and  $N$  is the number of observations (in this case,  $N = 8$ ).

OLS model:

Using the values of the weights obtained in the first exercise, we can calculate the predicted values for the test observations:

$$\underbrace{\Phi = \begin{bmatrix} 1 & 1 \\ 1 & 3 \\ 1 & 6 \\ 1 & 9 \\ 1 & 8 \end{bmatrix} \wedge w = \begin{bmatrix} 3.31593 \\ 0.11372 \end{bmatrix} \Rightarrow \tilde{z} = \begin{bmatrix} 3.42965 \\ 3.65709 \\ 3.99825 \\ 4.33941 \\ 4.22569 \end{bmatrix}}_{\text{train observations}} \quad \underbrace{\Phi = \begin{bmatrix} 1 & 4 \\ 1 & 2 \\ 1 & 5 \end{bmatrix} \wedge w = \begin{bmatrix} 3.31593 \\ 0.11372 \end{bmatrix} \Rightarrow \tilde{z} = \begin{bmatrix} 3.77081 \\ 3.54337 \\ 3.88453 \end{bmatrix}}_{\text{test observations}}$$

Finally, we take  $z = [1.25, 7.0, 2.7, 3.2, 5.5, 0.7, 1.1, 2.2]$  along with the predicted values to calculate the two RMSE for the OLS model.

$$\text{RMSE}_{\text{train}(x_1-x_5)} = \sqrt{\frac{1}{8} \sum_{i=1}^8 (z_i - \tilde{z}_i)^2} = \sqrt{\frac{1}{8} \times 20.53350} = 1.60209$$

$$\text{RMSE}_{\text{test}(x_6-x_8)} = \sqrt{\frac{1}{8} \sum_{i=1}^8 (z_i - \tilde{z}_i)^2} = \sqrt{\frac{1}{8} \times 18.23757} = 1.50987$$

As expected, the RMSE for the test observations is slightly lower than the RMSE for the train observations, being close to one another. This is a good sign that the model is generalizing well to new data.

Ridge Regression model:

Repeating the exact same process for the Ridge model, we get the following predicted values:

$$\underbrace{\Phi = \begin{bmatrix} 1 & 1 \\ 1 & 3 \\ 1 & 6 \\ 1 & 9 \\ 1 & 8 \end{bmatrix} \wedge w = \begin{bmatrix} 1.81809 \\ 0.32376 \end{bmatrix} \Rightarrow \tilde{z} = \begin{bmatrix} 2.14185 \\ 2.78937 \\ 3.76065 \\ 4.73193 \\ 4.40817 \end{bmatrix}}_{\text{train observations}} \quad \underbrace{\Phi = \begin{bmatrix} 1 & 4 \\ 1 & 2 \\ 1 & 5 \end{bmatrix} \wedge w = \begin{bmatrix} 1.81809 \\ 0.32376 \end{bmatrix} \Rightarrow \tilde{z} = \begin{bmatrix} 3.11313 \\ 2.46561 \\ 3.43689 \end{bmatrix}}_{\text{test observations}}$$

We take  $z = [1.25, 7.0, 2.7, 3.2, 5.5, 0.7, 1.1, 2.2]$  along with the predicted values to calculate the RMSE for the Ridge model.

$$\text{RMSE}_{\text{train}(x_1-x_5)} = \sqrt{\frac{1}{8} \sum_{i=1}^8 (z_i - \tilde{z}_i)^2} = \sqrt{\frac{1}{8} \times 23.18868} = 1.70252$$

$$\text{RMSE}_{\text{test}(x_6-x_8)} = \sqrt{\frac{1}{8} \sum_{i=1}^8 (z_i - \tilde{z}_i)^2} = \sqrt{\frac{1}{8} \times 9.21798} = 1.07343$$

The RMSE for the test observations is significantly lower than the RMSE for the train observations, which is a sign of underfitting. This suggests that the Ridge regression model is not capturing the underlying patterns in the data as well as the OLS model. The regularization term in the Ridge regression model may be too strong, causing the model to oversimplify the relationship between the features and the target variable.

4. **Consider an MLP to predict the output  $y_{\text{class}}$  characterized by the weights**

$$W^{[1]} = \begin{bmatrix} 0.1 & 0.1 \\ 0.1 & 0.2 \\ 0.2 & 0.1 \end{bmatrix}, \quad b^{[1]} = \begin{bmatrix} 0.1 \\ 0 \\ 0.1 \end{bmatrix}, \quad W^{[2]} = \begin{bmatrix} 1 & 2 & 2 \\ 1 & 2 & 1 \\ 1 & 1 & 1 \end{bmatrix}, \quad b^{[2]} = \begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix}$$

**the output activation function**

$$\text{softmax}(z_c^{[\text{out}]}) = \frac{e^{z_c^{[\text{out}]}}}{\sum_{l=1}^{|C|} e^{z_l^{[\text{out}]}}}$$

**, no activations on the hidden layer(s) and the cross-entropy loss:**

$$\text{CE} = - \sum_{i=1}^N \sum_{l=1}^{|C|} t_l^{(i)} \log(X_l^{[\text{out}](i)}) \quad (2)$$

**Consider also that the output layer of the MLP gives the predictions for the classes A, B and C in this order. Perform one stochastic gradient descent update to all the weights and biases with learning rate  $\eta = 0.1$  using the training observation  $x_1$ .**

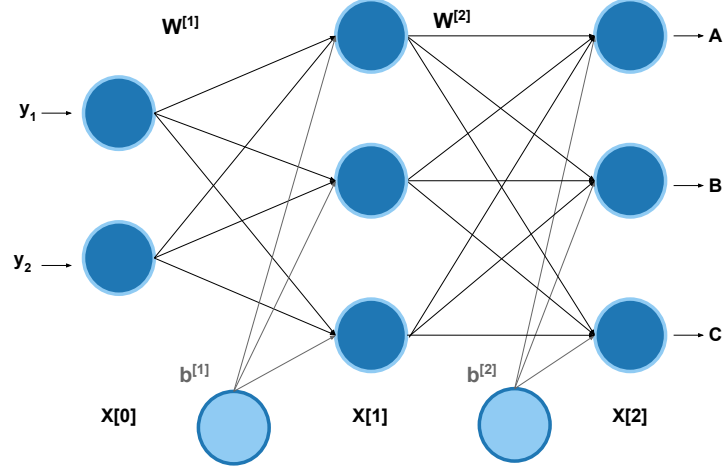


Figure 1: MLP architecture

We may also draw the corresponding scheme:

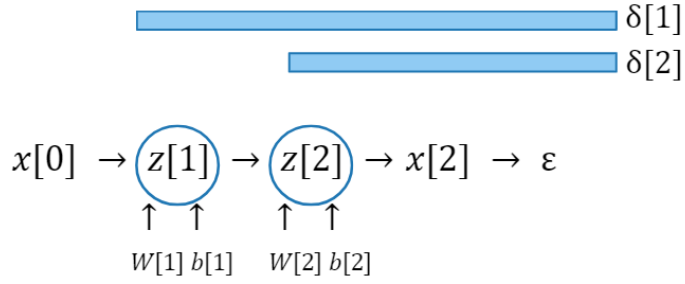


Figure 2: MLP Scheme

To begin, we initiate the process with the **forward propagation**. Below are the key equations to calculate the values of each layer,  $x_i^{[p]}$ :

$$z_i^{[p]} = W^{[p]} x_i^{[p-1]} + b^{[p]} \quad x_i^{[p]} = f(z_i^{[p]})$$

$f$  is the activation function, for every layer, provided in the statement.

$$f = \text{softmax}(z_c^{[\text{out}]}) = \frac{e^{z_c^{[\text{out}]}}}{\sum_{l=1}^{|C|} e^{z_l^{[\text{out}]}}}$$

In this notation, the value  $p$  is the index of the MLP layer.

We may now calculate the values of each node in the network for the training observation:

$$\boxed{x_1}$$

$$z^{[1]} = W^{[1]} x^{[0]} + b^{[1]} = \begin{bmatrix} 0.1 & 0.1 \\ 0.1 & 0.2 \\ 0.2 & 0.1 \end{bmatrix} \begin{bmatrix} 1 \\ 1 \end{bmatrix} + \begin{bmatrix} 0.1 \\ 0 \\ 0.1 \end{bmatrix} = \begin{bmatrix} 0.3 \\ 0.3 \\ 0.4 \end{bmatrix}$$

$$z^{[2]} = W^{[2]} z^{[1]} + b^{[2]} = \begin{bmatrix} 1 & 2 & 2 \\ 1 & 2 & 1 \\ 1 & 1 & 1 \end{bmatrix} \begin{bmatrix} 0.3 \\ 0.3 \\ 0.4 \end{bmatrix} + \begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix} = \begin{bmatrix} 2.7 \\ 2.3 \\ 2.0 \end{bmatrix}$$

$$x^{[2]} = f(z^{[2]}) = \begin{bmatrix} 0.461 \\ 0.309 \\ 0.229 \end{bmatrix}$$

Now, we can calculate the **backward propagation** to update the weights and biases. Firstly, we calculate the derivatives:

- Derivative of the cross-entropy loss function:

$$\frac{\partial \text{CE}}{\partial Z_l^{[\text{out}]}} = - \sum_{i=1}^N \frac{t_l^{(i)}}{X_l^{[\text{out}](i)}} = x^{[2]} - t$$

- Derivative of the softmax function:

$$\frac{\partial f}{\partial z_c^{[\text{out}]}} = f(z_c^{[\text{out}]}) \cdot (1 - f(z_c^{[\text{out}]}))$$

- Derivatives for the  $z[i]$ :

$$\begin{aligned} \frac{\partial x^{[i]}(z^{[i]})}{\partial z^{[i]}} &= f(z^{[i]})(1 - f(z^{[i]})) & \frac{\partial z^{[i]}(W^{[i]}, b^{[i]}, x^{[i-1]})}{\partial W^{[i]}} &= x^{[i-1]} \\ \frac{\partial z^{[i]}(W^{[i]}, b^{[i]}, x^{[i-1]})}{\partial b^{[i]}} &= 1 & \frac{\partial z^{[i]}(W^{[i]}, b^{[i]}, x^{[i-1]})}{\partial x^{[i-1]}} &= W^{[i]} \end{aligned}$$

To start the recursion for the **backward propagation**, we calculate the deltas (for the output layer and the hidden layer):

$$\begin{aligned} \delta^{[2]} &= \frac{\partial E}{\partial \mathbf{x}^{[2]}} \circ \frac{\partial \mathbf{x}^{[2]}}{\partial \mathbf{z}^{[2]}} = (\mathbf{x}^{[2]} - t) \circ f(\mathbf{z}^{[2]})(1 - f(\mathbf{z}^{[2]})) \\ &= \left( \begin{pmatrix} 0.461 \\ 0.309 \\ 0.229 \end{pmatrix} - \begin{pmatrix} 0 \\ 1 \\ 0 \end{pmatrix} \right) \circ f \left( \begin{pmatrix} 2.7 \\ 2.3 \\ 2.0 \end{pmatrix} \right) \left( 1 - f \left( \begin{pmatrix} 2.7 \\ 2.3 \\ 2.0 \end{pmatrix} \right) \right) = \begin{pmatrix} 0.461 \\ -0.691 \\ 0.229 \end{pmatrix} \circ \begin{pmatrix} 0.248 \\ 0.214 \\ 0.177 \end{pmatrix} = \begin{pmatrix} 0.114 \\ -0.148 \\ 0.052 \end{pmatrix} \end{aligned}$$

$$\begin{aligned}
\delta^{[1]} &= \left( \frac{\partial \mathbf{z}^{[2]}}{\partial \mathbf{x}^{[1]}} \right)^T \cdot \delta^{[2]} \circ \mathbf{z}^{[1]} = (\mathbf{W}^{[2]})^T \cdot \delta^{[2]} \circ \mathbf{z}^{[1]} \\
&= \begin{pmatrix} 1 & 1 & 1 \\ 2 & 2 & 1 \\ 2 & 1 & 1 \end{pmatrix} \cdot \begin{pmatrix} 0.114 \\ -0.148 \\ 0.052 \end{pmatrix} \circ \begin{pmatrix} 0.3 \\ 0.3 \\ 0.4 \end{pmatrix} = \begin{pmatrix} 0.018 \\ -0.016 \\ 0.132 \end{pmatrix} \circ \begin{pmatrix} 0.3 \\ 0.3 \\ 0.4 \end{pmatrix} = \begin{pmatrix} 0.0054 \\ -0.0048 \\ 0.0528 \end{pmatrix}
\end{aligned}$$

Finally, we can **update the weights and biases** with the gradient descent:

$$\begin{aligned}
\frac{\partial E}{\partial \mathbf{W}^{[1]}} &= \delta^{[1]} \cdot \frac{\partial \mathbf{z}^{[1]}}{\partial \mathbf{W}^{[1]}} = \delta^{[1]} \cdot (\mathbf{x}^{[0]})^T \\
&= \begin{pmatrix} 0.0054 \\ -0.0048 \\ 0.0528 \end{pmatrix} \cdot (1 \ 1) = \begin{pmatrix} 0.0054 & 0.0054 \\ -0.0048 & -0.0048 \\ 0.0528 & 0.0528 \end{pmatrix}
\end{aligned}$$

$$\mathbf{W}^{[1]} = \mathbf{W}^{[1]} - \eta \cdot \frac{\partial E}{\partial \mathbf{W}^{[1]}} = \begin{pmatrix} 0.1 & 0.1 \\ 0.1 & 0.2 \\ 0.2 & 0.1 \end{pmatrix} - 0.1 \cdot \begin{pmatrix} 0.0054 & 0.0054 \\ -0.0048 & -0.0048 \\ 0.0528 & 0.0528 \end{pmatrix} = \begin{pmatrix} 0.09946 & 0.09946 \\ 0.09952 & 0.19952 \\ 0.19472 & 0.09472 \end{pmatrix}$$

$$\frac{\partial E}{\partial \mathbf{b}^{[1]}} = \delta^{[1]} \cdot \frac{\partial \mathbf{z}^{[1]}}{\partial \mathbf{b}^{[1]}} = \delta^{[1]}$$

$$\mathbf{b}^{[1]} = \mathbf{b}^{[1]} - \eta \cdot \frac{\partial E}{\partial \mathbf{b}^{[1]}} = \begin{pmatrix} 0.1 \\ 0 \\ 0.1 \end{pmatrix} - 0.1 \cdot \begin{pmatrix} 0.0054 \\ -0.0048 \\ 0.0528 \end{pmatrix} = \begin{pmatrix} 0.1 \\ 0 \\ 0.1 \end{pmatrix} - \begin{pmatrix} 0.00054 \\ -0.00048 \\ 0.00528 \end{pmatrix} = \begin{pmatrix} 0.09946 \\ 0.00048 \\ 0.09472 \end{pmatrix}$$

$$\begin{aligned}
\frac{\partial E}{\partial \mathbf{W}^{[2]}} &= \delta^{[2]} \cdot \frac{\partial \mathbf{z}^{[2]}}{\partial \mathbf{W}^{[2]}} = \delta^{[2]} \cdot (\mathbf{z}^{[1]})^T \\
&= \begin{pmatrix} 0.114 \\ -0.148 \\ 0.052 \end{pmatrix} \cdot (0.3 \ 0.3 \ 0.4) = \begin{pmatrix} 0.0342 & 0.0342 & 0.0456 \\ -0.0444 & -0.0444 & -0.0592 \\ 0.0156 & 0.0156 & 0.0208 \end{pmatrix}
\end{aligned}$$

$$\begin{aligned}
\mathbf{W}^{[2]} &= \mathbf{W}^{[2]} - \eta \cdot \frac{\partial E}{\partial \mathbf{W}^{[2]}} = \begin{pmatrix} 1 & 2 & 2 \\ 1 & 2 & 1 \\ 1 & 1 & 1 \end{pmatrix} - 0.1 \cdot \begin{pmatrix} 0.0342 & 0.0342 & 0.0456 \\ -0.0444 & -0.0444 & -0.0592 \\ 0.0156 & 0.0156 & 0.0208 \end{pmatrix} \\
&= \begin{pmatrix} 1 & 2 & 2 \\ 1 & 2 & 1 \\ 1 & 1 & 1 \end{pmatrix} - \begin{pmatrix} 0.00342 & 0.00342 & 0.00456 \\ -0.00444 & -0.00444 & -0.00592 \\ 0.00156 & 0.00156 & 0.00208 \end{pmatrix} = \begin{pmatrix} 0.99658 & 1.99658 & 1.99544 \\ 1.00444 & 2.00444 & 1.00592 \\ 0.99844 & 0.99844 & 0.99792 \end{pmatrix}
\end{aligned}$$



$$\frac{\partial E}{\partial \mathbf{b}^{[2]}} = \delta^{[2]} \cdot \frac{\partial \mathbf{z}^{[2]T}}{\partial \mathbf{b}^{[2]}} = \delta^{[2]}$$

$$\mathbf{b}^{[2]} = \mathbf{b}^{[2]} - \eta \cdot \frac{\partial E}{\partial \mathbf{b}^{[2]}} = \begin{pmatrix} 1 \\ 1 \\ 1 \end{pmatrix} - 0.1 \cdot \begin{pmatrix} 0.114 \\ -0.148 \\ 0.052 \end{pmatrix} = \begin{pmatrix} 1 \\ 1 \\ 1 \end{pmatrix} - \begin{pmatrix} 0.0114 \\ -0.0148 \\ 0.0052 \end{pmatrix} = \begin{pmatrix} 0.9886 \\ 1.0148 \\ 0.9948 \end{pmatrix}$$

## Part II: Programming

Consider the parkinsons.csv dataset (available at the course's webpage), where the goal is to predict a patient's score on the Unified Parkinson's Disease Rating Scale based on various biomedical measurements. To answer question (5), average the performance of the models over 10 separate runs. In each run, use a different 80 – 20 train test split by setting a `random_state = i`, with `i = 1...10`.

5. Train a Linear Regression model, an MLP Regressor with 2 hidden layers of 10 neurons each and no activation functions, and another MLP Regressor with 2 hidden layers of 10 neurons each using ReLU activation functions. (Use `random_state=0` on the MLPs, regardless of the run). Plot a boxplot of the test MAE of each model.

```

1 import matplotlib.pyplot as plt, pandas as pd
2 from sklearn.model_selection import train_test_split
3 from sklearn.linear_model import LinearRegression
4 from sklearn.metrics import mean_absolute_error
5 from sklearn.neural_network import MLPRegressor
6
7 # Read the dataset
8 df = pd.read_csv("./parkinsons.csv")
9 X, y = df.drop("target", axis=1), df["target"]
10
11 linear_mae = []
12
13 # Linear Regression
14 for i in range(1, 11):
15     X_train, X_test, y_train, y_test = train_test_split(X, y, test_size
16                                                         =0.2, random_state=i)
17
18     lr_model = LinearRegression()
19     lr_model.fit(X_train, y_train)
20
21     y_pred = lr_model.predict(X_test)
22     mae = mean_absolute_error(y_test, y_pred)
23     linear_mae.append(mae)
24
25 mlp_no_activation_mae = []

```

```

26 # mlps with no activation function
27 for i in range(1, 11):
28     X_train, X_test, y_train, y_test = train_test_split(X, y, test_size
29                                                         =0.2, random_state=i)
30     mlp_no_activation = MLPRegressor(hidden_layer_sizes=(10, 10),
31                                     activation='identity',
32                                     random_state=0, validation_fraction
33                                     =0.2)
34     mlp_no_activation.fit(X_train, y_train)
35
36     y_pred = mlp_no_activation.predict(X_test)
37     mae = mean_absolute_error(y_test, y_pred)
38     mlp_no_activation_mae.append(mae)
39
40 mlp_relu_mae = []
41
42 # mlps with ReLU activation function
43 for i in range(1, 11):
44     X_train, X_test, y_train, y_test = train_test_split(X, y, test_size
45                                                         =0.2, random_state=i)
46     mlp_relu = MLPRegressor(hidden_layer_sizes=(10, 10), activation='relu',
47                             random_state=0, validation_fraction=0.2)
48     mlp_relu.fit(X_train, y_train)
49
50     y_pred = mlp_relu.predict(X_test)
51     mae = mean_absolute_error(y_test, y_pred)
52     mlp_relu_mae.append(mae)
53
54 # Data for plotting
55 b_plot = plt.boxplot(
56     [linear_mae, mlp_no_activation_mae, mlp_relu_mae], patch_artist=True,
57     labels=['Linear Regression', 'MLP No Activation', 'MLP ReLU']
58 )
59
60 colors = ["#1f77b4", "#E40071"]
61 for patch, color in zip(b_plot["boxes"], colors):
62     patch.set_facecolor(color)
63 for median in b_plot["medians"]:
64     median.set_color("black")
65
66 # Create boxplot
67 plt.title('Test MAE of Models')
68 plt.ylabel('Mean Absolute Error')
69 plt.grid(axis="y")
70 plt.show()

```

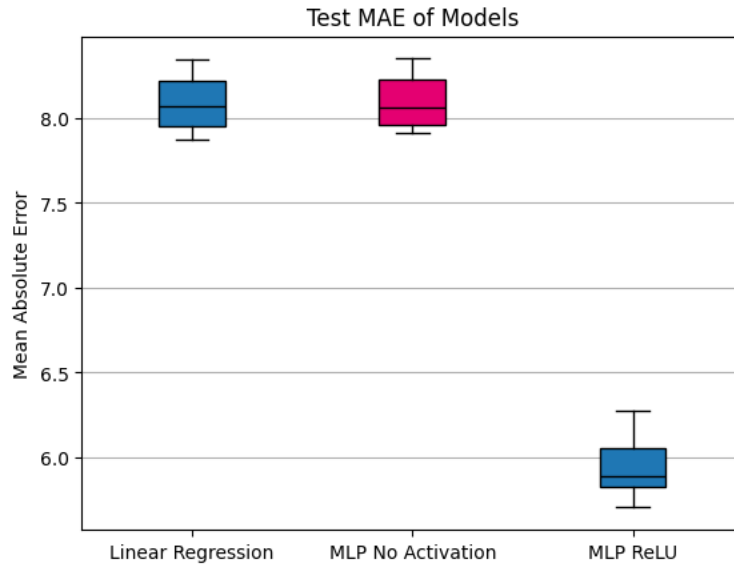


Figure 3: Boxplot of the test MAE for each model

6. Compare a Linear Regression with a MLP with no activations, and explain the impact and the importance of using activation functions in a MLP. Support your reasoning with the results from the boxplots.

As we can see in the previous plot, the MLP without activation functions behaves similarly to Linear Regression, while the MLP with ReLU activation function performs better than the other two models. Linear Regression and MLP without activation functions have a high MAE, centered around 8.0. On the other hand, the MLP with ReLU activation function has a lower MAE, centered around 6.0 and this shows the power of activation functions. Activation functions enable the MLP to learn from non-linear relationships in the data. Without them, each layer is just performing a linear operation, which severely limits the network's ability to generalize and fit real-world data, especially when the underlying relationships are complex or non-linear.

7. Using a 80 – 20 train-test split with `random_state=0`, use a Grid Search to tune the hyperparameters of an MLP regressor with two hidden layers (size 10 each). The parameters to search over are: (i) L2 penalty, with the values  $\{0.0001, 0.001, 0.01\}$ ; (ii) learning rate, with the values  $\{0.001, 0.01, 0.1\}$ ; and (iii) batch size, with the values  $\{32, 64, 128\}$ . Plot the test MAE for each combination of hyperparameters, report the best combination, and discuss the trade-offs between the combinations.

```

1 import pandas as pd
2 from sklearn.model_selection import train_test_split, GridSearchCV
3 from sklearn.neural_network import MLPRegressor
4 from sklearn.metrics import mean_absolute_error
5 import matplotlib.pyplot as plt
6 import seaborn as sns
7

```

```

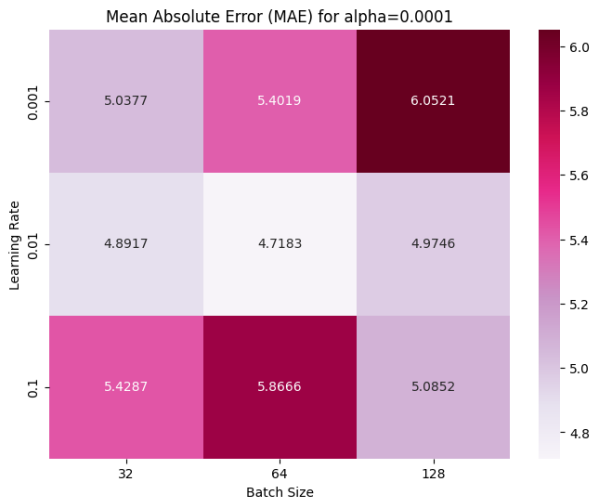
8 # Load the dataset (Assuming the file is named 'parkinsons.csv')
9 data = pd.read_csv('parkinsons.csv')
10
11 # Separating features and target
12 df = pd.read_csv("./parkinsons.csv")
13 X, y = df.drop("target", axis=1), df["target"]
14
15 # Split the data into training and testing sets
16 X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2,
17 random_state=0)
18
19 # Define the MLPRegressor with two hidden layers (size 10 each)
20 mlp = MLPRegressor(hidden_layer_sizes=(10, 10), random_state=0)
21
22 # Define the hyperparameters to search
23 param_grid = {
24     'alpha': [0.0001, 0.001, 0.01], # L2 penalty
25     'learning_rate_init': [0.001, 0.01, 0.1], # Learning rate
26     'batch_size': [32, 64, 128], # Batch size
27 }
28
29 # Perform grid search
30 grid_search = GridSearchCV(mlp, param_grid, cv=3, scoring='
31 neg_mean_absolute_error')
32 grid_search.fit(X_train, y_train)
33
34 # Get the best parameters and best model
35 best_params = grid_search.best_params_
36 best_model = grid_search.best_estimator_
37
38 # Make predictions on the test set
39 y_pred = best_model.predict(X_test)
40
41 # Evaluate the model using Mean Absolute Error (MAE)
42 test_mae = mean_absolute_error(y_test, y_pred)
43 print(f"Test MAE: {test_mae}")
44 print(f"Best Hyperparameters: {best_params}")
45
46 # Visualize the test MAE for each combination of hyperparameters, grouped
47 by 'alpha'
48 results = pd.DataFrame(grid_search.cv_results_)
49 results['mean_test_mae'] = -results['mean_test_score'] # Convert to
50 positive MAE
51
52 # Plot a heatmap for each value of alpha
53 alphas = results['param_alpha'].unique()
54
55 for alpha in alphas:
56     subset = results[results['param_alpha'] == alpha]
57     pivot_table = subset.pivot(index='param_learning_rate_init', columns='
58 param_batch_size', values='mean_test_mae')
59
60     # Plot the heatmap
61     fig, ax = plt.subplots(figsize=(8, 6))
62     sns.heatmap(pivot_table, annot=True, fmt=".4f", cmap="PuRd", ax=ax)
63     plt.title(f'Mean Absolute Error (MAE) for alpha={alpha}')

```

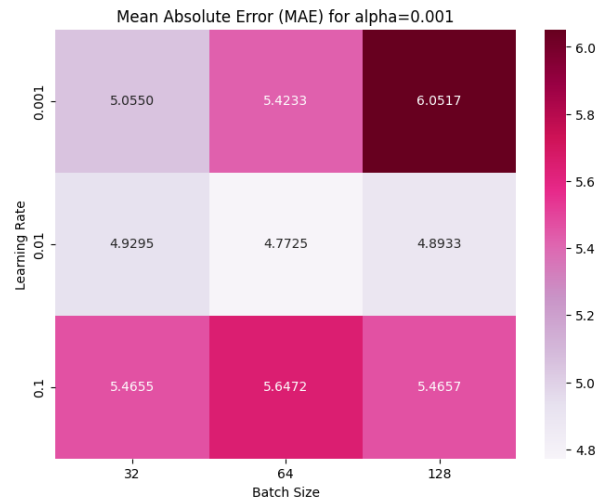
```

59 plt.xlabel('Batch Size')
60 plt.ylabel('Learning Rate')
61 plt.show()

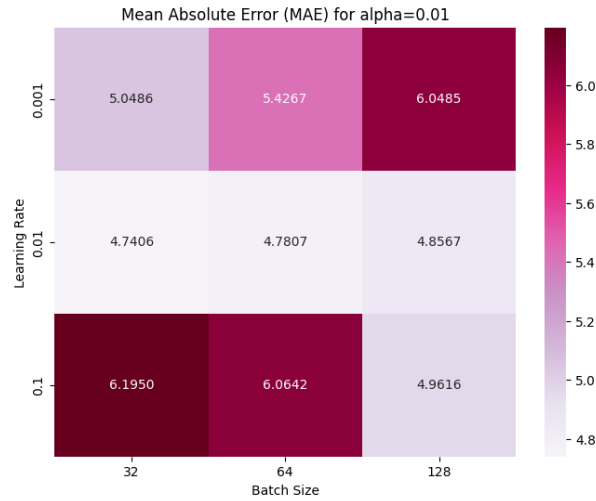
```



(a) MAE plot for L2 penalty = 0.0001



(b) MAE plot for L2 penalty = 0.001



(c) MAE plot for L2 penalty = 0.01

Figure 4: MAE plot for each combination of hyperparameters

When analyzing MAE values, the smaller the MAE the better the models performance. With that being said, the best combination for heatmap is:

- Figure 4 (a): Learning Rate = 0.01 and Batch Size = 64
- Figure 4 (b): Learning Rate = 0.01 and Batch Size = 64
- Figure 4 (c): Learning Rate = 0.01 and Batch Size = 32