I Pen-and-paper

1.

	y_1	y_1^2	y_1^3
x_1	0.8	0.64	0.512
x_2	1	1	1
x_3	1.2	1.44	1.728
x_4	1.4	1.96	2.744
x_5	1.6	2.56	4.096

$$w = (X^T \cdot X + \lambda I)^{-1} \cdot X^T \cdot z$$

$$= \left(\begin{bmatrix} 1 & 0.8 & 0.64 & 0.512 \\ 1 & 1 & 1 & 1 \\ 1 & 1.2 & 1.44 & 1.728 \\ 1 & 1.4 & 1.96 & 2.744 \\ 1 & 1.6 & 2.56 & 4.096 \end{bmatrix}^T \cdot \begin{bmatrix} 1 & 0.8 & 0.64 & 0.512 \\ 1 & 1 & 1 & 1 \\ 1 & 1.2 & 1.44 & 1.728 \\ 1 & 1.4 & 1.96 & 2.744 \\ 1 & 1.6 & 2.56 & 4.096 \end{bmatrix} + 2I \right)^{-1} \cdot \begin{bmatrix} 1 & 0.8 & 0.64 & 0.512 \\ 1 & 1 & 1 & 1 \\ 1 & 1.2 & 1.44 & 1.728 \\ 1 & 1.4 & 1.96 & 2.744 \\ 1 & 1.6 & 2.56 & 4.096 \end{bmatrix}^T \cdot \begin{bmatrix} 24 \\ 20 \\ 10 \\ 13 \\ 12 \end{bmatrix}$$

$$= \left(\begin{bmatrix} 5 & 6 & 7.6 & 10.08 \\ 6 & 7.6 & 10.08 & 13.8784 \\ 7.6 & 10.08 & 13.8784 & 19.68 \\ 10.08 & 13.8784 & 19.68 & 28.55488 \end{bmatrix} + 2I \right)^{-1} \cdot \begin{bmatrix} 1 & 0.8 & 0.64 & 0.512 \\ 1 & 1 & 1 & 1 \\ 1 & 1.2 & 1.44 & 1.728 \\ 1 & 1.4 & 1.96 & 2.744 \\ 1 & 1.6 & 2.56 & 4.096 \end{bmatrix}^T \cdot \begin{bmatrix} 24 \\ 20 \\ 10 \\ 13 \\ 12 \end{bmatrix}$$

$$= \left(\begin{bmatrix} 7 & 6 & 7.6 & 10.08 \\ 6 & 9.6 & 10.08 & 13.8784 \\ 7.6 & 10.08 & 15.8784 & 19.68 \\ 10.08 & 13.8784 & 19.68 & 30.55488 \end{bmatrix} \right)^{-1} \cdot \begin{bmatrix} 1 & 0.8 & 0.64 & 0.512 \\ 1 & 1 & 1 & 1 \\ 1 & 1.2 & 1.44 & 1.728 \\ 1 & 1.4 & 1.96 & 2.744 \\ 1 & 1.6 & 2.56 & 4.096 \end{bmatrix}^T \cdot \begin{bmatrix} 24 \\ 20 \\ 10 \\ 13 \\ 12 \end{bmatrix}$$

$$= \begin{bmatrix} 0.34168753 & -0.1214259 & -0.07490231 & -0.00932537 \\ -0.1214259 & 0.3892078 & -0.09667718 & -0.07445624 \\ -0.07490231 & -0.09667718 & 0.37257788 & -0.17135047 \\ -0.00932537 & -0.07445624 & -0.17135047 & 0.17998796 \end{bmatrix} \cdot \begin{bmatrix} 1 & 0.8 & 0.64 & 0.512 \\ 1 & 1 & 1 & 1 \\ 1 & 1.2 & 1.44 & 1.728 \\ 1 & 1.4 & 1.96 & 2.744 \\ 1 & 1.6 & 2.56 & 4.096 \end{bmatrix}^T \cdot \begin{bmatrix} 24 \\ 20 \\ 10 \\ 13 \\ 12 \end{bmatrix}$$

$$= \begin{bmatrix} 0.19183474 & 0.13603395 & 0.07200288 & -0.00070608 & -0.08254055 \\ 0.08994535 & 0.09664848 & 0.07774793 & 0.02966982 & -0.05115977 \\ -0.00152564 & 0.02964793 & 0.04950363 & 0.04981662 & 0.02236208 \\ -0.08640083 & -0.07514413 & -0.03439835 & 0.04447593 & 0.17011812 \end{bmatrix} \cdot \begin{bmatrix} 24 \\ 20 \\ 10 \\ 13 \\ 12 \end{bmatrix}$$

$$= \begin{bmatrix} 7.0450759 \\ 4.64092765 \\ 1.96734046 \\ -1.30088142 \end{bmatrix}$$

2.

$$E = \frac{1}{5}\sum_{i=1}^{5}(z_i - \hat{z}_i)^2$$

$$= \frac{1}{5}\left[\left(24 - \begin{bmatrix} 7.0450759 \\ 4.64092765 \\ 1.96734046 \\ -1.30088142 \end{bmatrix} \cdot \begin{bmatrix} 1 & 0.8 & 0.64 & 0.512 \end{bmatrix}\right)^2\right.$$

$$+ \left(20 - \begin{bmatrix} 7.0450759 \\ 4.64092765 \\ 1.96734046 \\ -1.30088142 \end{bmatrix} \cdot \begin{bmatrix} 1 & 1 & 1 & 1 \end{bmatrix}\right)^2$$

$$+ \left(10 - \begin{bmatrix} 7.0450759 \\ 4.64092765 \\ 1.96734046 \\ -1.30088142 \end{bmatrix} \cdot \begin{bmatrix} 1 & 1.2 & 1.44 & 1.728 \end{bmatrix}\right)^2$$

$$+ \left(13 - \begin{bmatrix} 7.0450759 \\ 4.64092765 \\ 1.96734046 \\ -1.30088142 \end{bmatrix} \cdot \begin{bmatrix} 1 & 1.4 & 1.96 & 2.744 \end{bmatrix}\right)^2$$

$$\left.+ \left(12 - \begin{bmatrix} 7.0450759 \\ 4.64092765 \\ 1.96734046 \\ -1.30088142 \end{bmatrix} \cdot \begin{bmatrix} 1 & 1.6 & 2.56 & 4.096 \end{bmatrix}\right)^2\right]$$

$$= 46.83068498017306$$

$$\text{RMSE} = \sqrt{E} = \sqrt{46.83068498017306} = 6.843294892094967$$

3. x_1)

- Forward Propagation

$$\text{net}^{[1]} = w^{[1]} \cdot a^{[0]} + b^{[1]} = \begin{bmatrix}1\\1\end{bmatrix} \times 0.8 + \begin{bmatrix}1\\1\end{bmatrix} = \begin{bmatrix}1.8\\1.8\end{bmatrix}$$

$$\text{a}^{[1]} = e^{0.1\times\text{net}^{[1]}} = \begin{bmatrix}1.19721736\\1.19721736\end{bmatrix}$$

$$\text{net}^{[2]} = w^{[2]} \cdot a^{[1]} + b^{[2]} = \begin{bmatrix}1 & 1\end{bmatrix} \cdot \begin{bmatrix}1.19721736\\1.19721736\end{bmatrix} + 1 = 3.3944347262436203$$

$$\text{a}^{[2]} = e^{0.1\times\text{net}^{[2]}} = 1.404166$$

- Backward Propagation

$$\frac{\partial l}{\partial a^{[2]}} = a^{[2]} - z = -22.595834083700154$$

$$\frac{\partial a^{[2]}}{\partial \text{net}^{[2]}} = 0.1 \times e^{0.1\times\text{net}^{[2]}} = 0.1404165916299847$$

$$\frac{\partial \text{net}^{[2]}}{\partial a^{[1]}} = (w^{[2]})^T = \begin{bmatrix}1\\1\end{bmatrix}$$

$$\frac{\partial \text{a}^{[1]}}{\partial \text{net}^{[1]}} = \begin{bmatrix}0.1 \times e^{0.1\times\text{net}_1^{[1]}} & 0\\ 0 & 0.1 \times e^{0.1\times\text{net}_2^{[1]}}\end{bmatrix} = \begin{bmatrix}0.11972174 & 0\\ 0 & 0.11972174\end{bmatrix}$$

$$\delta^{[2]} = \frac{\partial l}{\partial \text{net}^{[2]}} = \frac{\partial a^{[2]}}{\partial \text{net}^{[2]}} \cdot \frac{\partial l}{\partial \text{a}^{[2]}} = -3.1728300070698143$$

$$\delta^{[1]} = \frac{\partial l}{\partial \text{net}^{[1]}} = \frac{\partial a^{[1]}}{\partial \text{net}^{[1]}} \cdot \frac{\partial \text{net}^{[2]}}{\partial \text{a}^{[1]}} \times \delta^{[2]} = \begin{bmatrix}-0.37985672\\-0.37985672\end{bmatrix}$$

$$\frac{\partial l}{\partial b^{[2]}} = \delta^{[2]} = -3.1728300070698143$$

$$\frac{\partial l}{\partial b^{[1]}} = \delta^{[1]} = \begin{bmatrix}-0.37985672\\-0.37985672\end{bmatrix}$$

$$\frac{\partial l}{\partial w^{[2]}} = \delta^{[2]} \cdot (a^{[1]})^T = \begin{bmatrix}-3.79856717 & -3.79856717\end{bmatrix}$$

$$\frac{\partial l}{\partial w^{[1]}} = \delta^{[1]} \cdot x^T = \begin{bmatrix}-0.30388537\\-0.30388537\end{bmatrix}$$

$x_2)$

- Forward Propagation

$$\text{net}^{[1]} = w^{[1]} \cdot a^{[0]} + b^{[1]} = \begin{bmatrix} 1 \\ 1 \end{bmatrix} \times 1 + \begin{bmatrix} 1 \\ 1 \end{bmatrix} = \begin{bmatrix} 2 \\ 2 \end{bmatrix}$$

$$\text{a}^{[1]} = e^{0.1 \times \text{net}^{[1]}} = \begin{bmatrix} 1.22140276 \\ 1.22140276 \end{bmatrix}$$

$$\text{net}^{[2]} = w^{[2]} \cdot a^{[1]} + b^{[2]} = \begin{bmatrix} 1 & 1 \end{bmatrix} \cdot \begin{bmatrix} 1.22140276 \\ 1.22140276 \end{bmatrix} + 1 = 3.4428055163203397$$

$$\text{a}^{[2]} = e^{0.1 \times \text{net}^{[2]}} = 1.410974431163945$$

- Backward Propagation

$$\frac{\partial l}{\partial a^{[2]}} = a^{[2]} - z = -18.589025568836057$$

$$\frac{\partial a^{[2]}}{\partial \text{net}^{[2]}} = 0.1 \times e^{0.1 \times \text{net}^{[2]}} = 0.1410974431163945$$

$$\frac{\partial \text{net}^{[2]}}{\partial a^{[1]}} = (w^{[2]})^T = \begin{bmatrix} 1 \\ 1 \end{bmatrix}$$

$$\frac{\partial \text{a}^{[1]}}{\partial \text{net}^{[1]}} = \begin{bmatrix} 0.1 \times e^{0.1 \times \text{net}_1^{[1]}} & 0 \\ 0 & 0.1 \times e^{0.1 \times \text{net}_2^{[1]}} \end{bmatrix} = \begin{bmatrix} 0.12214028 & 0 \\ 0 & 0.12214028 \end{bmatrix}$$

$$\delta^{[2]} = \frac{\partial l}{\partial \text{net}^{[2]}} = \frac{\partial a^{[2]}}{\partial \text{net}^{[2]}} \cdot \frac{\partial l}{\partial \text{a}^{[2]}} = -2.6228639777880485$$

$$\delta^{[1]} = \frac{\partial l}{\partial \text{net}^{[1]}} = \frac{\partial a^{[1]}}{\partial \text{net}^{[1]}} \cdot \frac{\partial \text{net}^{[2]}}{\partial \text{a}^{[1]}} \times \delta^{[2]} = \begin{bmatrix} -0.32035733 \\ -0.32035733 \end{bmatrix}$$

$$\frac{\partial l}{\partial b^{[2]}} = \delta^{[2]} = -2.6228639777880485$$

$$\frac{\partial l}{\partial b^{[1]}} = \delta^{[1]} = \begin{bmatrix} -0.32035733 \\ -0.32035733 \end{bmatrix}$$

$$\frac{\partial l}{\partial w^{[2]}} = \delta^{[2]} \cdot (a^{[1]})^T = \begin{bmatrix} -3.2035733 & -3.2035733 \end{bmatrix}$$

$$\frac{\partial l}{\partial w^{[1]}} = \delta^{[1]} \cdot x^T = \begin{bmatrix} -0.32035733 \\ -0.32035733 \end{bmatrix}$$

x_3)

- Forward Propagation

$$\text{net}^{[1]} = w^{[1]} \cdot a^{[0]} + b^{[1]} = \begin{bmatrix} 1 \\ 1 \end{bmatrix} \times 1.2 + \begin{bmatrix} 1 \\ 1 \end{bmatrix} = \begin{bmatrix} 2.2 \\ 2.2 \end{bmatrix}$$

$$\text{a}^{[1]} = e^{0.1 \times \text{net}^{[1]}} = \begin{bmatrix} 1.24607673 \\ 1.24607673 \end{bmatrix}$$

$$\text{net}^{[2]} = w^{[2]} \cdot a^{[1]} + b^{[2]} = \begin{bmatrix} 1 & 1 \end{bmatrix} \cdot \begin{bmatrix} 1.24607673 \\ 1.24607673 \end{bmatrix} + 1 = 3.4921534611747616$$

$$\text{a}^{[2]} = e^{0.1 \times \text{net}^{[2]}} = 1.4179545084644258$$

- Backward Propagation

$$\frac{\partial l}{\partial a^{[2]}} = a^{[2]} - z = -8.582045491535574$$

$$\frac{\partial a^{[2]}}{\partial \text{net}^{[2]}} = 0.1 \times e^{0.1 \times \text{net}^{[2]}} = 0.14179545084644 26$$

$$\frac{\partial \text{net}^{[2]}}{\partial a^{[1]}} = (w^{[2]})^T = \begin{bmatrix} 1 \\ 1 \end{bmatrix}$$

$$\frac{\partial \text{a}^{[1]}}{\partial \text{net}^{[1]}} = \begin{bmatrix} 0.1 \times e^{0.1 \times \text{net}_1^{[1]}} & 0 \\ 0 & 0.1 \times e^{0.1 \times \text{net}_2^{[1]}} \end{bmatrix} = \begin{bmatrix} 0.12460767 & 0 \\ 0 & 0.12460767 \end{bmatrix}$$

$$\delta^{[2]} = \frac{\partial l}{\partial \text{net}^{[2]}} = \frac{\partial a^{[2]}}{\partial \text{net}^{[2]}} \cdot \frac{\partial l}{\partial \text{a}^{[2]}} = -1.2168950096569668$$

$$\delta^{[1]} = \frac{\partial l}{\partial \text{net}^{[1]}} = \frac{\partial a^{[1]}}{\partial \text{net}^{[1]}} \cdot \frac{\partial \text{net}^{[2]}}{\partial \text{a}^{[1]}} \times \delta^{[2]} = \begin{bmatrix} -0.15163446 \\ -0.15163446 \end{bmatrix}$$

$$\frac{\partial l}{\partial b^{[2]}} = \delta^{[2]} = -1.2168950096569668$$

$$\frac{\partial l}{\partial b^{[1]}} = \delta^{[1]} = \begin{bmatrix} -0.15163446 \\ -0.15163446 \end{bmatrix}$$

$$\frac{\partial l}{\partial w^{[2]}} = \delta^{[2]} \cdot (a^{[1]})^T = \begin{bmatrix} -1.51634456 & -1.51634456 \end{bmatrix}$$

$$\frac{\partial l}{\partial w^{[1]}} = \delta^{[1]} \cdot x^T = \begin{bmatrix} -0.18196135 \\ -0.18196135 \end{bmatrix}$$

- Batch Gradient Descent Update

$$
\begin{aligned}
b^{[2]} &= b^{[2]} - \eta \sum_{i=1}^{3} \frac{\partial l}{\partial b^{[2]}} \\
&= 1 - 0.1\,(-3.1728300070698143 + -2.6228639777880485 + -1.2168950096569668) \\
&= 1.701258899451483
\end{aligned}
$$

$$
\begin{aligned}
b^{[1]} &= b^{[1]} - \eta \sum_{i=1}^{3} \frac{\partial l}{\partial b^{[1]}} \\
&= \begin{bmatrix} 1 \\ 1 \end{bmatrix} - 0.1 \left(\begin{bmatrix} -0.37985672 \\ -0.37985672 \end{bmatrix} + \begin{bmatrix} -0.32035733 \\ -0.32035733 \end{bmatrix} + \begin{bmatrix} -0.15163446 \\ -0.15163446 \end{bmatrix} \right) \\
&= \begin{bmatrix} 1.08518485 \\ 1.08518485 \end{bmatrix}
\end{aligned}
$$

$$
\begin{aligned}
w^{[2]} &= w^{[2]} - \eta \sum_{i=1}^{3} \frac{\partial l}{\partial w^{[2]}} \\
&= \begin{bmatrix} 1 & 1 \end{bmatrix} - 0.1 \Big(\begin{bmatrix} -3.79856717 & -3.79856717 \end{bmatrix} + \\
&\qquad \begin{bmatrix} -3.2035733 & -3.2035733 \end{bmatrix} + \\
&\qquad \begin{bmatrix} -1.51634456 & -1.51634456 \end{bmatrix} \Big) \\
&= \begin{bmatrix} 1.8518485 & 1.8518485 \end{bmatrix}
\end{aligned}
$$

$$
\begin{aligned}
w^{[1]} &= w^{[1]} - \eta \sum_{i=1}^{3} \frac{\partial l}{\partial w^{[1]}} \\
&= \begin{bmatrix} 1 \\ 1 \end{bmatrix} - 0.1 \left(\begin{bmatrix} -0.30388537 \\ -0.30388537 \end{bmatrix} + \begin{bmatrix} -0.32035733 \\ -0.32035733 \end{bmatrix} + \begin{bmatrix} -0.18196135 \\ -0.18196135 \end{bmatrix} \right) \\
&= \begin{bmatrix} 1.08062041 \\ 1.08062041 \end{bmatrix}
\end{aligned}
$$

II Programming

4. Ridge MAE = 0.162829976437694
   MLP1 MAE = 0.0680414073796843
   MLP2 MAE = 0.0978071820387748

5.

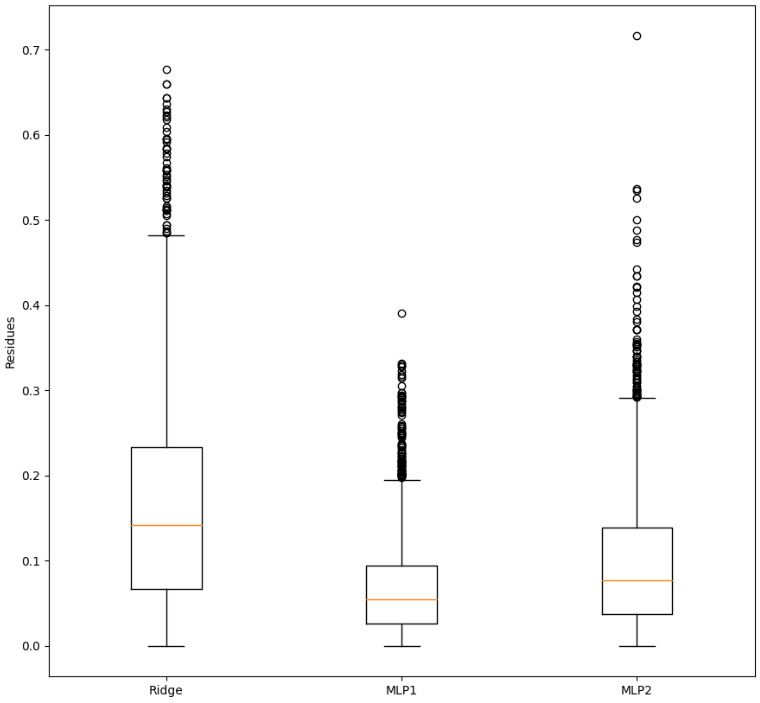

Figure 1: Boxplot

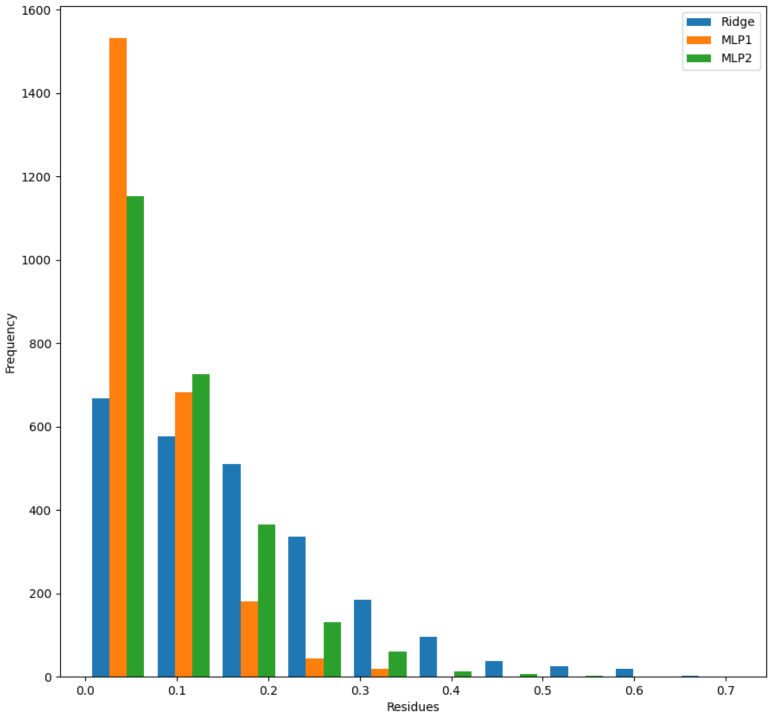

Figure 2: Histogram

6. The MLP1 converged in 452 iterations.
   The MLP2 converged in 77 iterations.

7. The two MLP differ in the stopping criteria. The MLP1 sets aside 10% of the training data as validation, terminating when the validation score doesn't improve significantly for some number of consecutive epochs. The MLP2 terminates when the training loss does not improve for some number of consecutive epochs.

   The training loss function converges faster than the validation score. The MLP2 goes through the training set, fitting to it rather quickly. Therefore, the training loss converges after few iterations, not improving by much between epochs, triggering the training termination. On the other hand, the MLP1 is confronted with the validation set. The validation score converges slowly as the MLP1 is confronted with data that it didn't yet observed.

   The gathered results support this conclusion, as the MLP2 performs better in the testing set, relative to the MLP1. Therefore, the MLP1, despite fitting better to the training set, performs poorly on the testing set. Therefore, MLP1 suffers from overfitting.

III Appendix

```python
from scipy.io.arff import loadarff
from sklearn import metrics
import pandas as pd
from sklearn.model_selection import train_test_split
from sklearn.linear_model import Ridge
from sklearn.neural_network import MLPRegressor
import matplotlib.pyplot as plt

if __name__ == "__main__":
    data = loadarff("kin8nm.arff")
    df = pd.DataFrame(data[0])
    X, y = df.drop("y", axis=1), df["y"]

    X_train, X_test, y_train, y_test = train_test_split(X,
                                                        y,
                                                        train_size=0.7,
                                                        random_state=0)

    regressors = [
        Ridge(alpha=0.1),
        MLPRegressor(hidden_layer_sizes=(10, 10),
                     max_iter=500,
                     random_state=0,
                     activation="tanh",
                     early_stopping=True),
        MLPRegressor(hidden_layer_sizes=(10, 10),
                     max_iter=500,
                     random_state=0,
                     activation="tanh",
                     early_stopping=False)
    ]

    mae = []
    for regressor in regressors:
        regressor.fit(X_train, y_train)
        y_pred = regressor.predict(X_test)
        mae.append(metrics.mean_absolute_error(y_test, y_pred))

    print("Ridge MAE: {}.".format(mae[0]))
    print("MLP1 MAE: {}.".format(mae[1]))
    print("MLP2 MAE: {}.".format(mae[2]))

    print("MLP1 converged in {} iterations.".format(regressors[1].n_iter_))
    print("MLP2 converged in {} iterations.".format(regressors[2].n_iter_))

    plt.boxplot([
        abs(y_test - y_pred)
        for y_pred in [regressor.predict(X_test) for regressor in regressors]
    ])
    plt.xticks([1, 2, 3], ["Ridge", "MLP1", "MLP2"])
    plt.ylabel("Residues")
    plt.show()

    plt.hist([
        abs(y_test - y_pred)
        for y_pred in [regressor.predict(X_test) for regressor in regressors]
    ])
    plt.legend(["Ridge", "MLP1", "MLP2"])
    plt.xlabel("Residues")
    plt.ylabel("Frequency")
    plt.show()
```