

Evaluating Activation Functions and Optimizers in Feed-Forward Neural Networks: A Study on the Franke Function and Breast Cancer Classification

Edvard B. Rørnes* and Isak O. Rukan†

*Institute of Physics, University of Oslo,
0371 Oslo, Norway*

(Dated: November 5, 2024)

In recent years, neural networks have become increasingly important for advancements in various scientific fields. In this report, we develop a program^a that consists of both linear and logistic regression methods and a neural network. In particular, the program implements gradient descent, stochastic gradient descent, the optimization methods AdaGrad, RMSprop and Adam, and a simple feed-forward neural network with backpropagation. The neural network is evaluated using the Adam algorithm, and compared against linear regression on the Franke function, and against logistic regression using `sklearn`'s breast cancer data [1]. We explore three activation functions: Sigmoid, ReLU, and LeakyReLU. The analysis involves tuning the number of epochs, hidden nodes, regularization strength λ and the learning rate η to optimal values. On the Franke function, our own implementation of the neural network achieved minimal MSE of 3×10^{-4} using ReLU, compared to `tensorflow.keras`' [2] 2.5×10^{-4} and 1.5×10^{-2} for gradient decent with RMSprop. For the cancer data, the highest accuracies achieved with ReLU, LeakyReLU, and Sigmoid activation functions were all 99.3%, compared to 97.9% for logistic regression with stochastic gradient decent. Notably, LeakyReLU activation function exhibited minimal sensitivity to variations in both λ and η , whereas ReLU and Sigmoid showed significant sensitivity, causing worse performance outside of optimal values. Whilst ReLU yielded the best performance on the Franke function, and Sigmoid managed to perform well on the breast cancer data, we still find that overall the LeakyReLU activation function contains the most benefits in these applications, performing consistently well throughout.

1. INTRODUCTION

Over the last few years, machine learning and neural networks have become an increasingly important part of data analysis with an enormous range of applications. From image recognition to predictive analytics and scientific simulations, these techniques are reshaping the way the scientific community tackles complicated problems. Linear and logistic regression play a fundamental role in machine learning, providing robust ways for modeling linear and logistic relationships in data. They also serve as important building blocks in understanding more advanced techniques, such as neural networks.

Neural networks excel at handling complex, nonlinear relationships in data. Their flexibility in approximating obfuscated patterns has made them indispensable across diverse fields, including biology, engineering, finance, and physics. For just a few examples, see [3–5].

The main goal of this project is to gain a deeper understanding of neural networks by first exploring the fundamentals of linear and logistic regression. To achieve this, we implement and experiment with various optimization techniques, comparing the performance of linear and logistic regression to that of neural networks. We use data generated from the Franke function to evaluate linear regression, and the binary breast cancer dataset from `sklearn`'s datasets [1] to assess logistic regression.

We first introduce the theory behind linear/logistic regression, introduce various gradient decent algorithms, and then give a basic introduction to neural networks. Further, we explain how these techniques are implemented and finally look at the results. The performance of linear/logistic regression is compared to the performance of the implemented neural network with various different activation functions. Our own neural network is also tested against `tensorflow.keras` [2] on the Franke function to give us an estimate of its performance.

2. METHODS

In this section we present the various methods used in this report. For reference, the Franke function which we will be using to test our linear regression methods is:

$$\begin{aligned} f(x, y) = & \frac{3}{4} \exp\left(-\frac{(9x - 2)^2}{4} - \frac{(9y - 2)^2}{4}\right) \\ & + \frac{3}{4} \exp\left(-\frac{(9x + 1)^2}{49} - \frac{(9y + 1)^2}{10}\right) \\ & + \frac{1}{2} \exp\left(-\frac{(9x - 7)^2}{4} - \frac{(9y - 3)^2}{4}\right) \\ & - \frac{1}{5} \exp(-(9x - 4)^2 - (9y - 7)^2) + \varepsilon, \quad (1) \end{aligned}$$

where $\varepsilon \sim \mathcal{N}(0, 0.01)$ is a normal distributed error to simulate the irreducible error of measured data. In the domain $x, y \in [0, 1]$ then the Franke function has the range $\sim [-0.2, 1.05]$. Thus this corresponds to a reasonable error of roughly 0.8%.

* e.b.rornes@fys.uio.no

† icrukan@uio.no

^a Github

2.1. Linear Regression

As discussed in a previous project [6], linear regression is the simplest method for fitting a continuous function to a given data set. The data set is approximated by $\mathbf{y} = \mathbf{X}\beta$ and the β coefficients are found by minimizing the cost function. For this project we consider the two regression methods:

$$C_{\text{OLS}}(\beta) = \frac{2}{n}(\mathbf{y} - \mathbf{X}\beta)^2, \quad (2)$$

$$C_{\text{Ridge}}(\beta) = C_{\text{OLS}}(\beta) + \lambda\|\beta\|_2^2. \quad (3)$$

We then insist that the derivative of these w.r.t. β is 0, and choose the resulting β coefficients as our model. Doing this we arrive at:

$$\beta_{\text{OLS}} = (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{y}, \quad (4)$$

$$\beta_{\text{Ridge}} = (\mathbf{X}^T \mathbf{X} + \lambda \mathbf{I})^{-1} \mathbf{X}^T \mathbf{y}. \quad (5)$$

2.2. Regularization Terms

Regularization is a technique to prevent overfitting by adding a penalty to the cost function that discourages complex models. Overfitting occurs when a model learns the noise in the training data rather than the underlying patterns, leading to poor generalization on unseen data. In the context of neural networks, regularization plays a critical role, especially when working with architectures that have many parameters.

The two common regularization methods that we inspected previously are Ridge and LASSO regularization. In this project, we will only be considering Ridge regularization, where the cost function is given by

$$C_{\text{Ridge}}(\beta) = C_{\text{OLS}}(\beta) + \lambda\|\beta\|_2^2, \quad (6)$$

where $C_{\text{OLS}}(\beta)$ is the ordinary least squares cost function, β represents the model parameters, and the hyperparameter λ controls the magnitude of the penalty to large coefficients. For more details on regularization in linear regression, see [6].

In the context of neural networks, regularization techniques are crucial for controlling model complexity and improving generalization. Ridge regularization can be implemented in neural networks by applying ℓ^2 regularization to the weights of the network. This approach adds the penalty term $\lambda\|\mathbf{W}\|_2^2$ to the loss function, where \mathbf{W} is the weight matrix of the neural network. The modified cost function for a neural network with ℓ^2 regularization becomes

$$C_{\text{NN}} = C_{\text{loss}} + \lambda\|\mathbf{W}\|_2^2, \quad (7)$$

where C_{loss} will be the MSE for regression tasks and cross-entropy loss for classification tasks.

Incorporating regularization in neural networks offers several important benefits. Firstly, ℓ^2 regularization effectively applies a weight decay, which shrinks the

weights during training and helps prevent the model from becoming overly complex. This leads to smoother decision boundaries and reduces the risk of overfitting, allowing the model to generalize better to unseen data. Secondly, regularized models tend to exhibit increased robustness to variations in the data, as the penalty encourages the learning of simpler models that do not fit the noise in the training set. Lastly, by tuning the hyperparameter λ , we can effectively control the model's complexity, striking a balance between bias and variance of the model to the specific problem.

2.3. Logistic Regression

Whilst linear regression is quite successful in fitting continuous data, when the output is supposed to be discrete it fails. Linear regression predicts values across a continuous spectrum, resulting in predictions outside the range of valid class labels, such as giving negative probabilities. Logistic regression on the other hand is specifically designed for binary classification problems, and is thus ideal when dealing with discrete outcomes.

Logistic regression models the probability that an input belongs to a particular class by mapping real-valued inputs to a range between 0 and 1 using an **activation function**. Typically, the sigmoid function σ is used, converting the linear prediction into a probability in a smooth manner. Given an input vector X and a set of weights β , the predicted probability that the class label y equals 1 is expressed as:

$$P(y = 1 | \mathbf{X}) = \sigma(\mathbf{X}\beta) = \frac{1}{1 + e^{-\mathbf{X}\beta}}. \quad (8)$$

For each sample i , we refer to this probability as \hat{y}_i . To optimize the weights, logistic regression minimizes the cross-entropy loss function:

$$C(\beta) = -\frac{1}{n} \sum_{i=1}^n (y_i \ln(\hat{y}_i) + (1 - y_i) \ln(1 - \hat{y}_i)), \quad (9)$$

where y_i is the class label.

Furthermore, to penalize overfitting, a regularization term may be added to (9). This terms adds a penalty for large weights, trying to keep the weights (relatively) small. Adding the ℓ^2 regularization term results in

$$C(\beta) = -\frac{1}{n} \sum_{i=1}^n (y_i \ln(\hat{y}_i) + (1 - y_i) \ln(1 - \hat{y}_i)) + \lambda \sum_{i=1}^n w_j^2. \quad (10)$$

2.4. Resampling Methods

Resampling methods are used to estimate the accuracy of predictive models by splitting the data into training and testing sets or by generating multiple datasets. A resampling method that we will use is bootstrapping. This

involves sampling with replacement from the dataset to create multiple training sets. This helps assess model stability and generalizability on unseen data. For more details see [6].

2.5. Gradient Descent

Gradient descent (GD) is an essential optimization algorithm in machine learning, commonly used to minimize cost functions by adjusting model parameters iteratively. Given model parameters θ and a cost function $C(\theta)$, the GD update rule adjusts parameters in the opposite direction of the gradient:

$$\theta_i^{(j+1)} = \theta_i^{(j)} - \eta \frac{\partial C}{\partial \theta_i}, \quad (11)$$

where η is the **learning rate**. Batch gradient descent (BGD) calculates the gradient over the entire dataset:

$$\theta^{(j+1)} = \theta^{(j)} - \eta \nabla_{\theta} C. \quad (12)$$

BGD is computationally expensive for large datasets but provides smooth convergence toward the minimum.

The learning rate η does not necessarily need to be constant, and can change with each iteration. There are several ways to implement a varying learning rate. In this report, we either use a constant learning rate or a learning rate on the form

$$\eta(e, i; N; b; t_0; t_1) = \frac{t_0}{e \cdot N/b + i + t_1}, \quad (13)$$

where e is the current epoch, N the data-size, b the batch size (see sec. 2.6) and i the current batch-iteration. The parameter t_0 is related to the initial magnitude of the learning rate, allowing larger learning rates at the beginning of the algorithm. Keeping this parameter fairly large can be beneficial in scenarios where multiple local minima are present, as the learning rate will ‘wait’ a bit before it starts converging on a solution. The parameter t_1 on the other hand influences how quickly the learning rate decreases over ‘time’. For example, in scenarios where the data is sensitive to small changes, keeping t_1 small can help increase the accuracy of the model near the end of the training.

2.6. Stochastic Gradient Descent

Stochastic Gradient Descent (SGD) is a variation of gradient descent where each parameter update is performed on a single data point or a small batch. The update rule for SGD is:

$$\theta_i^{(j+1)} = \theta_i^{(j)} - \eta \frac{\partial C^{(i)}}{\partial \theta_i}$$

where $C^{(i)}$ is the cost function evaluated at a single data point i . While SGD introduces noise in the updates, it

often converges faster for large datasets, and helps escape local minima, making it well suited for training neural networks.

‘Plain’ gradient descent or stochastic gradient descent may also keep track of a so-called ‘momentum’ parameter m , which is supposed to push the descent algorithm in the correct direction. This parameter helps build up speed towards a solution, and can be helpful to overcome local minima. It is implemented in the following way:

$$m_i = \beta m_i + (1 - \beta)(\nabla_{\theta} C)_i, \quad (14)$$

and modifies the new θ_{i+1} like

$$\theta_{i+1} = \theta_i - \eta m_i. \quad (15)$$

2.7. Optimization Algorithms

To reach the global minima of the cost function, there exists several optimization algorithms which can help speed up the process and becoming trapped in local minima. These algorithms essentially modify the learning rate by analyzing the magnitude and behavior of the gradients. This section gives a brief summary of three optimization algorithms; the adaptive gradient (AdaGrad) algorithm, the root mean squared propagation (RMSprop) algorithm and the adaptive moment estimation (Adam) algorithm.

2.7.1. AdaGrad

The AdaGrad algorithm modifies the learning rate by keeping track of how large contributions from the gradients build up over time;

$$\eta_i \rightarrow \eta_i \frac{1}{\epsilon + \sqrt{\sum_{j=1}^i (\nabla_{\theta} C)_j^2}}, \quad (16)$$

where ϵ is a small parameter to avoid division by zero.

2.7.2. RMSprop

The RMSprop optimization algorithm has a similar goal as AdaGrad, minimizing the negative effects of large gradients. However, RMSprop does this a bit differently, by calculating a ‘decaying average of the squared gradients’. Specifically, it keeps track of a parameter G_i which represents how the average of the squared gradients change, and uses a parameter r which controls the ‘rate of decay’. Typically, r is set very close to 1, and we have used $r = 0.99$ throughout. The RMSprop algorithm modifies the learning rate as such:

$$G_i = rG_{i-1} + (1 - r) \cdot (\nabla_{\theta} C)_i^2, \quad (17)$$

$$\eta_i \rightarrow \frac{\epsilon + \eta_i}{\sqrt{G_i}}, \quad (18)$$

where ϵ is again introduced to avoid zero-division.

2.7.3. Adam

The Adam algorithm is perhaps the most advanced optimization algorithm of the ones we present here. It works by essentially combining RMSprop and momentum. It adjusts the learning rate by computing estimates of the mean ('first momentum' m) and the variance ('second momentum' v) of the gradients. The two momenta are updated in each iteration:

$$m_i = \beta_1 m_{i-1} + (1 - \beta_1) \nabla_\theta C, \quad (19)$$

$$v_i = \beta_2 v_{i-1} + (1 - \beta_2) (\nabla_\theta C)^2, \quad (20)$$

with the parameters β_1 and β_2 which are typically close to one. Given these values for β_1, β_2 , eq. (19) implies that m and v initially starts out close to zero. To account for this, Adam includes additional corrections terms:

$$\hat{m}_i = \frac{m_i}{1 - \beta_1^i}, \quad (21)$$

$$\hat{v}_i = \frac{v_i}{1 - \beta_2^i}. \quad (22)$$

Adam then calculates the next θ_{i+1} as such:

$$\theta_{i+1} = \theta_i - \frac{\eta}{\sqrt{\hat{v}_i} + \epsilon} \hat{m}_i \quad (23)$$

2.8. Neural Networks

Neural networks are computational models inspired by the human brain, designed to recognize patterns and relationships within data. They consist of layers of interconnected neurons or nodes, where each neuron applies a transformation to the input data. In each layer, neurons take a weighted sum of inputs, apply an *activation function* to introduce non-linearity, and pass the result to the next layer. The final layer produces the output, serving as the network's prediction.

2.8.1. Feed Forward Neural Networks

Feed-forward neural networks (FFNNs) are the simplest type of neural network, where data flows forward from input to output without forming cycles. These networks contain one or more hidden layers that apply an activation function to capture complex, nonlinear patterns in the data. The training process adjusts the weights of each connection to minimize a cost function, typically using gradient descent.

2.8.2. Activation Functions

Activation functions play a critical role in neural networks by introducing non-linearity, which enables the network to approximate more complex functions beyond

simple linear mappings. In this project, we use three different activation functions: sigmoid, ReLU, and Leaky ReLU. Each function has different properties that can impact training performance and convergence.

- The sigmoid function, suitable for binary classification, takes an input $z \in \mathbb{R}$ and outputs a value in the range $(0, 1)$. This makes it useful for probabilistic interpretations. As mentioned prior, it is given by:

$$\sigma(z) = \frac{1}{1 + e^{-z}}. \quad (24)$$

- The ReLU (Rectified Linear Unit) function activates only positive values:

$$R(z) = \begin{cases} 0 & \text{if } z \leq 0 \\ z & \text{if } z > 0 \end{cases}. \quad (25)$$

This reduces the number of calculations that the network has to perform and can speed up the training.

- The Leaky ReLU (LReLU) function is a variation of ReLU that allows a small gradient when $z \leq 0$. This can help mitigate an issue known as 'dying ReLU' [7] where neurons become inactive due to consistently receiving negative inputs, helping to mitigate issues with inactive neurons. It is given by:

$$LR(z) = \begin{cases} \alpha z & \text{if } z \leq 0 \\ z & \text{if } z > 0 \end{cases} \quad (26)$$

where α is some small number. In this project we only consider $\alpha = 0.01$.

As we will see later, selecting appropriate activation functions for each layer, FFNNs can effectively capture complex data patterns which can heavily enhance model performance.

2.8.3. Backpropagation

Backpropagation is the key algorithm for training neural networks by optimizing weights to minimize the cost function. It works by propagating the error backward from the output layer to the input layers, computing gradients for each weight based on the error. These gradients are then used to update the weights, enabling the network to learn from its errors and make more accurate predictions over time.

3. IMPLEMENTATION

3.1. Linear and Logistic Regression

Linear and Logistic Regression was used to study data from the Franke function and the breast cancer data,

respectively. Both methods rely on some sort of optimization algorithm when applying the gradient descent algorithm, see sec. 2.7. The optimization algorithms (plain GD/SGD, Adagrad, RMSprop or Adam) are implemented as classes inheriting from a parent class `Optimizer`. This class defaults to the parameters $\eta = 0.01, m = 0.9, \epsilon = 1e - 8, \beta_1 = 0.9, \beta_2 = 0.999, r = 0.9$, with the possibility of η being a callable, see sec. 2.5. In this report, when comparing constant and varying learning rates, we have set $t_0 = 2$ and $t_1 = 2/\eta$ (for constant η) in the equation for varying learning rate (13).

After an optimization algorithm has been chosen, the class `DescentSolver` is used together with a chosen gradient, the Ridge-gradient (linear regression) or a logistic-gradient (logistic regression), to compute GD or SGD. To analyze the results of `DescentSolver`, the class `DescentAnalyzer` is used. This class computes metrics, MSE/R^2 (linear regression) or accuracy score (logistic regression), for an (equally sized) grid of λ, η -values. The metrics, together with all other parameters, are saved as pickle-files, which can be read at a later time from `DescentAnalyzer.load_data` giving a dictionary of the data.

3.2. Neural Network

The FFNN class was implemented to analyze both the Franke function and the breast cancer data. The FFNN is structured with an input layer, one or more hidden layers, and an output layer, with each layer utilizing an activation function, and currently only supports the Adam optimizer. This was due to unknown performance issues when trying to incorporate the `Optimize` class. Due to this we are only considering dynamical learning rates for the FFNN. The architecture is defined by specifying the input size, the number of neurons in hidden layers, and the output size. In the case of the Franke function, the input layer is size 2, one for x and one for y , while for the breast cancer data, the input layer corresponds to the number of features, i.e. 30, representing several properties of the tumor, such as circumference, radius, density, etc.

We tested various configurations for layering the hidden layers. Ultimately, for the Franke function, we settled on a pyramid-like architecture for the hidden layers [4, 8, 16, 32, 16, 8, 4, 2], while for the breast cancer data, we used [15, 30, 15, 8, 4, 2]. These choices were based on the sizes of both the input and output layers, where the output layer is 1 in both cases. The weights and biases are initialized using random values scaled by the number of input neurons, which aids in faster convergence during training. The network supports ReLU, Sigmoid, and Leaky ReLU as activation functions.

The forward propagation computes the output of the network by applying the chosen activation function to the weighted sum of inputs at each layer. Depending on the task, the FFNN uses mean squared error (MSE) as the

loss function for regression tasks (like the Franke function) and binary cross-entropy (BCE) for classification tasks (like breast cancer data). The backward propagation algorithm updates the weights and biases using gradient descent. We also implemented an ℓ^2 regularization scheme to mitigate overfitting.

The training process includes splitting the data into training and test sets, followed by iterating through a specified number of epochs during which the network adjusts its weights to minimize the error. After training, the network can predict outputs for new data, allowing for evaluation against known values from either the Franke function or the breast cancer target. The overall performance of the model is assessed using MSE for the Franke function and accuracy for the breast cancer data. To test our FFNN, for the Franke function analysis we also included `tensorflow.keras` with the LReLU activation function as a benchmark for our own implementation.

3.3. Data Parameters

We tried various different learning rates η and λ -values. Specifically, we used $\eta_{\text{const}}, \lambda \in [10^{-10}, 10^1]$. For varying learning rate, we settled on $t_0 = 2$ and $t_1 = 2/\eta_{\text{const}}$, see (13). We sampled 20, 250 datapoints from the Franke function and looked at epoch sizes $e \in \{10, 100\}$, and batch sizes $b \in \{4, 5, 10\}$. For linear regression we used 30 bootstraps, while for logistic regression we used 4. Both for linear and logistic regression we decided to study a polynomial degree of 5, suggested by the results in [6]. We used a test size of 25% throughout.

On the cancer data we use StandardScaling to scale the data after splitting it to avoid data leakage, whereas the Franke function was normalized to the range [0, 1] using min-max scaling.

4. RESULTS & DISCUSSION

We represent the results and compare the various methods. The larger plots have been placed in Appendix A to preserve space for the sake of readability. More figures can be found in the `Figures` folder [8].

4.1. Franke

4.1.1. Linear Regression

MSE scores from linear regression for plain SGD and the three more advanced optimization methods (Adagrad, RMSprop and Adam), for different epoch and batch sizes, and for both constant and varying learning rates, look for files on the form *LinReg...pdf* in [8]. Figs. 5. and 6. are composed of one figure for each of the optimization methods, for 10 epochs with 20 and 250 samples and a batch size of 4 and 50 respectively.

For plain SGD with a varying η , increasing epoch sizes and decreasing batch sizes lowered the MSE scores, suggesting a stable convergence. However, with a constant learning rate, the MSE scores blew up, implying that the algorithm skipped the global minima and diverged. This can be seen from plots in [8] (look for *LinReplainSGD_constEta...pdf*), where all MSE scores are ~ 0.207 . This feature was not seen in any of other optimization methods. Both RMSprop and Adam showed a clear convergence when increasing (decreasing) the epoch size (batch size). However, RMSprop did not converge for constant learning rates. For varying learning rates, RMSprop gave relatively low MSE scores with $t_1 = 2/\eta_{\text{const}} \sim [2, 2 \cdot 10^2]$. In this range, RMSprop gave quite good results, averaging an MSE score of $\sim [0.1, 0.01]$.

On the other hand, Adam converged for both varying and constant learning rates, and essentially outperformed all the other algorithms by giving better MSE scores overall. In fact, Adam seemed to treat varying and constant learning rates the same, in the sense that for $\eta_{\text{const}} \sim [2 \cdot 10^{-2}, 1]$, it averaged an MSE score of $\lesssim 0.1$.

The odd one out was AdaGrad. Giving the largest MSE scores, AdaGrad performed the worst out of all the optimization methods. In fact, AdaGrad seemed to not converge at all for too high batch sizes, rather producing sort of random data. We found this somewhat strange, as on the other hand, AdaGrad showed good results for logistic regression (see sec. 4.2.1), and for small batch sizes, e.g. 4 in Fig. 5b. it gave MSE scores $\lesssim 0.1$ for varying learning rates with $t_1 \sim [1.8 \cdot 10^{-2}, 3.5]$.

Together, the results show an overall better performance for varying learning rate, compared to that of a constant learning rate. Plain SGD is especially effected by this, compared to Adam which seems to converge for both varying and constant learning rates. Common for all optimization methods, except plain SGD, was that an increase (decrease) in number of epochs (batch size) gave better MSE scores.

4.1.2. Regression with a Neural Network

The MSE and R^2 as a function of η with 1000 epochs on the Franke function are given in Fig. 1. Here we show our own FFNN with different activation functions and `keras` NN with the LReLU activation function, all using the Adam optimizer. In this particular plot we are not using any regularization terms. The reason for this is simply due to `keras` being very slow, making the parameter scan take too long. This is however likely due to our initialization, but we still wanted to include it as a reference. The optimal learning rates for $\lambda = 0$ can easily be read off this plot: $\{10^{-3}, 10^{-3}, 3 \times 10^{-3}, 2 \times 10^{-2}\}$ for {ReLU, Sigmoid, LReLU, `keras`}.

Further we used the result for the best learning rate to plot the training loss after each epoch, given in Fig.

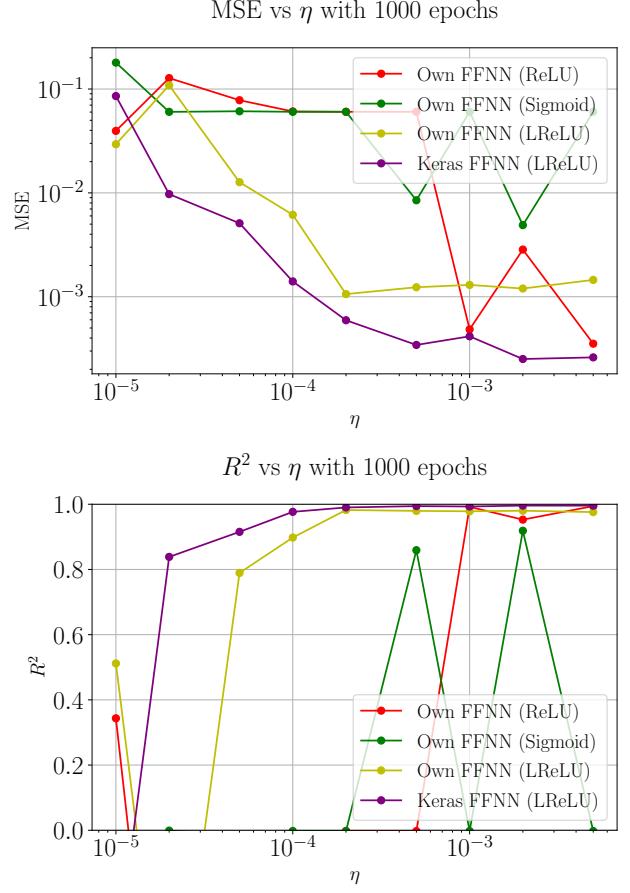


Fig. 1: MSE and R^2 regression results for the FFNN as a function of η with 1000 epochs.

2. Here we see that `keras` performs much better than our implementation for low epochs, but eventually our own FFNN begin to catch up. At the end, all seem to achieve a relatively good MSE as the number of epochs increase, with the exception of our own implementation of the Sigmoid function. This makes sense due to it being more suited for binary classification tasks compared to regression.

Further we plotted the predictions of the four variants in Fig. 7a. ReLU, LReLU and `keras` are all quite close to recreating the Franke function, whilst Sigmoid performs appreciably worse.

We then did the above again, but now with a regularization parameter λ and excluded `keras` due to time constraints. The MSE with 250 epochs for various combinations of η and λ are given in Fig. 8. The data corresponding to the best η and λ for each activation function are then saved, and the resulting MSE over epochs and prediction are given in Figs. 3. and 7b. Clearly the sigmoid activation function performs worse overall once

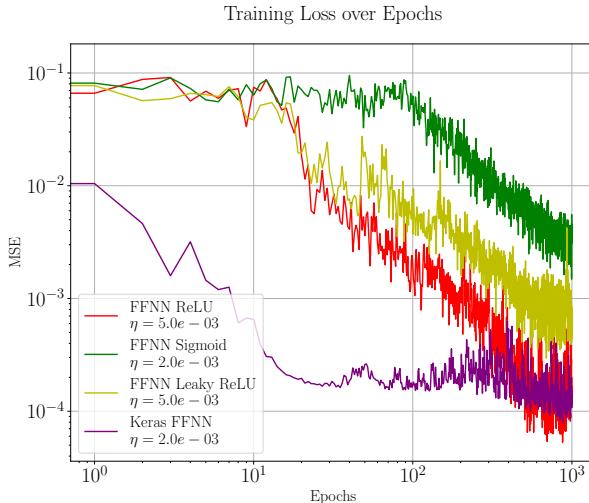


Fig. 2: The MSE regression results for the FFNN as a function of number of epochs for ReLU, LReLU, Sigmoid and keras.

again, whilst ReLU and LReLU have a close competition between them. The convergence is noticeably faster here, suggesting that the regularization parameter is improving our performance.

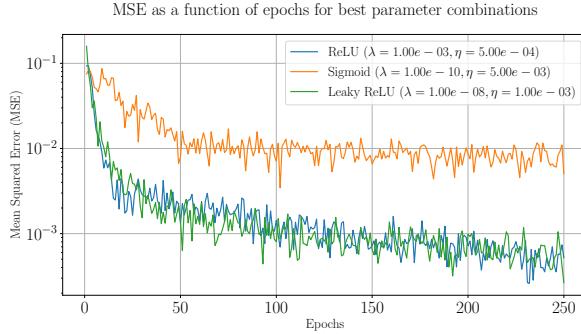


Fig. 3: The MSE regression results for the FFNN as a function of number of epochs for ReLU, LReLU and Sigmoid with the best performing combination of λ and η .

4.1.3. Comparison

From just the MSE alone we can see that the only real competition on larger number of epochs between the gradient decent methods and the FFNN is when we compare the GD methods to the Sigmoid activation function. Both the ReLU variants completely outclass all GD methods, reaching MSE scores which are an order of magnitude lower. This is of course neglecting AdaGrad which performed much worse than simply plotting the average of the Franke function, i.e. $z = 0.5$, as this would give an

MSE of roughly 0.066 given our particular test data sample. Some of GD methods, in particular RMSprop and at times plain SGD, showed divergent behavior when the number of epochs was too high. In contrast, Adam continued to improve for increasing epochs, although not as fast as FFNN. For small numbers of epoch counts, the competition between the GD methods and FFNN was more balanced; however, FFNN still came out on top in both speed and accuracy in these cases as well. For the relevant low epoch FFNN figures, see [8].

4.2. Cancer Data

4.2.1. Logistic Regression

Fig. 9. shows the accuracy score of (plain stochastic) logistic regression for batch size $N \in \{10, 100\}$ and a batch size $b = 50$, more figures can be found in the **Figures** folder. Note that the figures to the right in Fig. 9. are zoomed in versions of the one to the right, and the varying learning rate has been used. This figure shows an average accuracy score of ~ 0.97 for $t_1 = 2/\eta_{\text{const}} \sim [10^{-2}, 10^{-1}]$. The more advanced optimization methods were briefly looked into, showing similar results. However, Adam diverged for $N = 100$, and AdaGrad showed promising results (accuracy scores above 90%), the opposite behavior of what we saw for linear regression. The overall best performance was found using plain SGD, keeping in mind that this was studied the most. In particular, for 100 epochs (and a batch size of 50), the accuracy score came close to $\sim 98\%$ for $\lambda \sim [10^{-10}, 3.2 \cdot 10^{-3}], \eta \sim [8.3 \cdot 10^{-3}, 10^{-1}]$, see fig. 9.

4.2.2. Binary Classification with a Neural Network

The accuracy of the FFNN using three activation function across different values of η and λ is presented in Fig. 10.

All three activation functions manage to achieve an incredible maximal accuracy of 99.3%, getting only one mistake out of the 143 samples in the test set. We see however that of the 3, LReLU is the least sensitive to changes in both η and λ . The other two get appreciably worse results whenever the combination of η and λ is suboptimal. The poor accuracies showing up are seemingly correlated with the number of actual positives and negatives in the test data, as if the neural network simply decides to only guess positive or negative in certain runs. This can be clearly seen to be the strategy for Sigmoid when η and λ are suboptimal, where roughly half the parameter space results the percentages where it sticks to one guess throughout. We hypothesize that the reason ReLU and LReLU handle a wider range of learning rates and regularization more effectively is due to their piecewise linear nature, which prevents gradient vanishing issues. For Sigmoid this is not the case,

and is often affected by aggressive learning rates which may cause vanishing gradients [9]. This shows the importance of choosing the right activation function for the given task.

Fig. 4 shows the confusion matrix of the best runs. This figure illuminates where the NN failed on their best runs. All 3 only fail by giving a single false positive. Previously, all three activation functions consistently had 3 false positives, yielding an accuracy of 97.9%. We then assumed that this was due to a few difficult outliers in the test data, or lack of similar cases in the test data for the network to train on. Realizing this, we then ran the test again with a different random state on the `train_test_split` function, to get the result showed here. Of course in a general application, having difficult outliers is something that a NN must deal with, but when running over multiple different random states, we still find that more often than not, the best accuracy achieved was still 99.3%, even achieving 100% with certain seeds.

4.2.3. Comparison

When comparing the results from the GD methods to the neural networks (NNs), both methods demonstrate the ability to reach high accuracies, with SGD peaking at $\sim 98\%$ and the NNs with ReLU or LReLU achieving relatively higher scores. Both methods showed a clear dependence on the learning rate η , and the hyperparameter λ . Additionally, both approaches occasionally encountered convergence issues, with the Adam optimization method diverging for large batch sizes, and the neural networks showed performance variability when using the sigmoid function, especially for higher learning rates.

The overall performance of the NN, along with a better flexibility in its parameter space, makes it clear that it is more suitable than GD for this specific task. The insensitivity of η and λ was especially noticeable when using ReLU and LReLU activation functions, suggesting that they can handle suboptimal parameter values, without suffering from issues like gradient vanishing; this is often a large problem for GD methods.

5. CONCLUSION

On the Franke function, we found that Sigmoid is not overly sensitive to changes in the learning rate and hy-

perparameter λ , but also struggles to find good results. On the other hand ReLU has the ability to get great results, but only for very specific combinations of η and λ . LReLU seems to be the best of both worlds, being relatively insensitive to getting the exact right combination of η and λ , whilst still managing to obtain very good MSE scores like ReLU. When comparing the MSE scores to the MSE scores of the GD methods, we found that the NNs, in particular those utilizing ReLU and LReLU, consistently delivered an overall better performance. The best performing GD method; Adam showed convergence for various different parameters, and lower MSE scores compared to those of plain SGD, AdaGrad and RMSprop. However, they were still outperformed by the FFNN. In particular, for increasing epochs, the NNs performance consistently increased, a trend which only Adam shared, though with lower metric scores.

For the FFNN on the cancer data, overall assessments here are similar for the Franke function. The best performing activation function was again LReLU, followed by ReLU and finally the Sigmoid function. The latter performed much better on this classification problem compared to regression as one would expect from its definition, but once again still falls short to the ReLU variants. It must also be noted that the inputs to the Sigmoid function were clipped to prevent overflow, which might have impacted its accuracy. The slightly higher accuracy of LReLU over ReLU suggests it may provide an advantage by maintaining small gradients for negative inputs, offering a more stable training process. At their best, the FFNN gave accuracy scores of about 99%, beating plain SGD at $\sim 98\%$.

There are plentiful of avenues to pursue which can extend and improve our analysis. A few examples would be: testing out different values of α in LReLU, adjusting β_1 and β_2 and testing out different methods for initializing the weights and biases. All of these may have potential benefits, both in performance and computational efficiency. Similarly, whilst we did test different hidden layer setups, this was not done extensively. The odds of us achieving an optimal setup with relatively few attempts compared to possible combinations, is next to null. Finally, thoroughly testing out other optimizers such as Adagrad and RMSprop for the FFNN may also yield benefits, but due to time constraints we did not implement this properly.

-
- [1] F. Pedregosa, G. Varoquaux, *et. al.*, *Scikit-learn: Machine learning in python*, *Journal of Machine Learning Research* **12** (2011) 2825–2830.
<http://jmlr.org/papers/v12/pedregosa11a.html>.
 - [2] M. Abadi, A. Agarwal, *et. al.*, *TensorFlow: Large-scale machine learning on heterogeneous systems*, 2015.

Software available from tensorflow.org
<https://www.tensorflow.org/>.

- [3] A. Dawid, J. Arnold, *et. al.*, *Modern applications of machine learning in quantum sciences*, 2023.
<https://arxiv.org/abs/2204.04198>.
- [4] V. Thapar, *Applications of machine learning to*

- modelling and analysing dynamical systems*, 2023.
<https://arxiv.org/abs/2308.03763>.
- [5] F. G. Mohammadi, F. Shenavarmasouleh, and H. R. Arabnia, *Applications of machine learning in healthcare and internet of things (iot): A comprehensive review*, 2022. <https://arxiv.org/abs/2202.02868>.
- [6] I. Rukan and E. Rørnes, *Application of regression and resampling on usgs terrain data*, 2024.
<https://github.com/EdvardRornes/FYS-STK4155/blob/main/Project1/project1.pdf>.
- [7] L. L. Lu Lu, Y. S. Yeonjong Shin, Y. S. Yanhui Su, and G. E. K. George Em Karniadakis, *Dying relu and initialization: Theory and numerical examples*, *Communications in Computational Physics* **28** (2020) 1671–1706.
- <http://dx.doi.org/10.4208/cicp.OA-2020-0165>.
- [8] I. Rukan and E. Rørnes, 2024. Project 2 additional figures <https://github.com/EdvardRornes/FYS-STK4155/tree/main/Project2/Figures>.
- [9] S. R. Dubey, S. K. Singh, and B. B. Chaudhuri, *Activation functions in deep learning: A comprehensive survey and benchmark*, 2022. <https://arxiv.org/abs/2109.14545>.

Appendix A: Large Figures

A selection of figures, more can be found in the **Figures** folder.

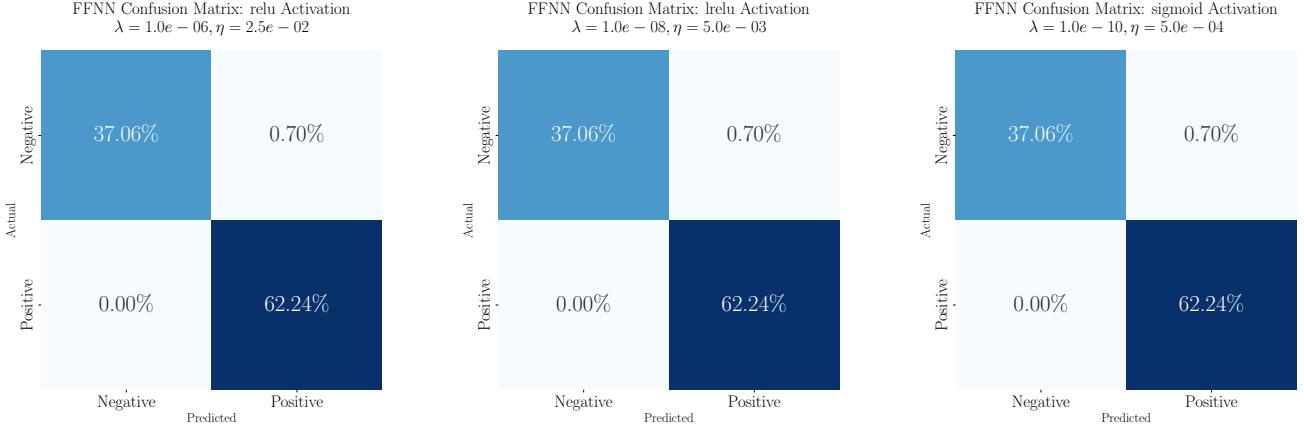


Fig. 4: Confusion matrix of the FFNN results with 100 epochs and optimal η and λ . The vertical axis show the correct answer, whilst the horizontal axis shows the FFNN's guess for all 3 activation functions. 100% accuracy corresponds to the off diagonals being 0% and the diagonals summing to 100%.

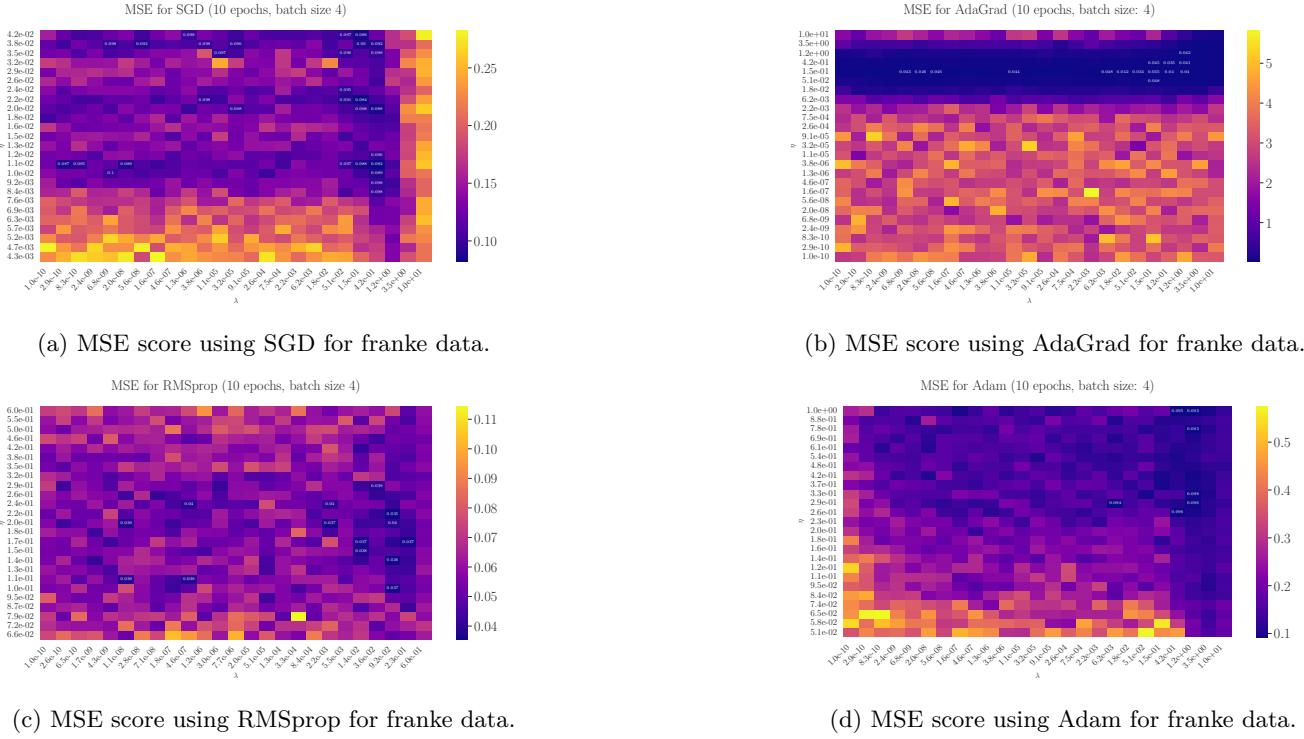


Fig. 5: MSE scores for linear regression on 20 data samples from the Franke function, for number of epochs $N = 10$ and a batch size of 4. A varying learning rate is used, and the y -axis denotes η_{const} , which is used to set t_1 in (13).

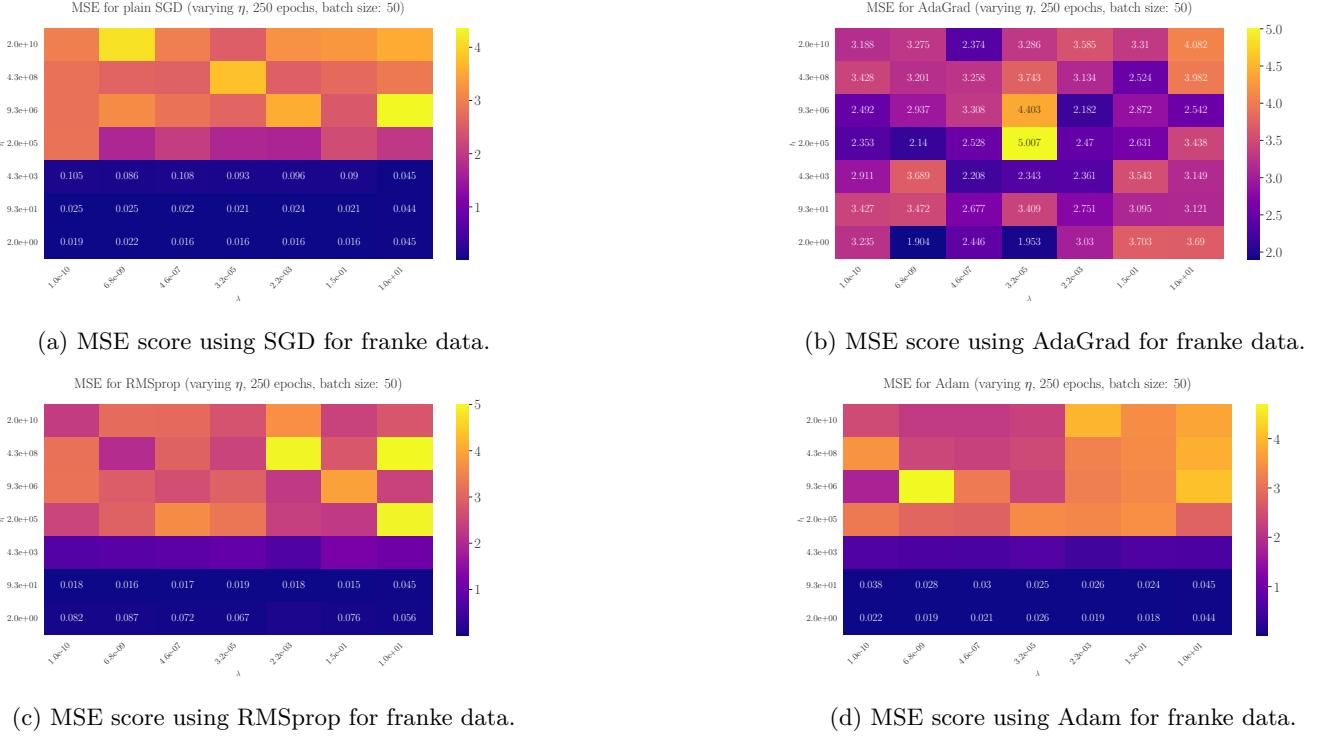


Fig. 6: MSE scores for linear regression on 250 data samples from the Franke function, for number of epochs $N = 250$ and a batch size of 50. A varying learning rate is used, and the y -axis denotes $t_1 = 2/\eta_{\text{const}}$ in (13).

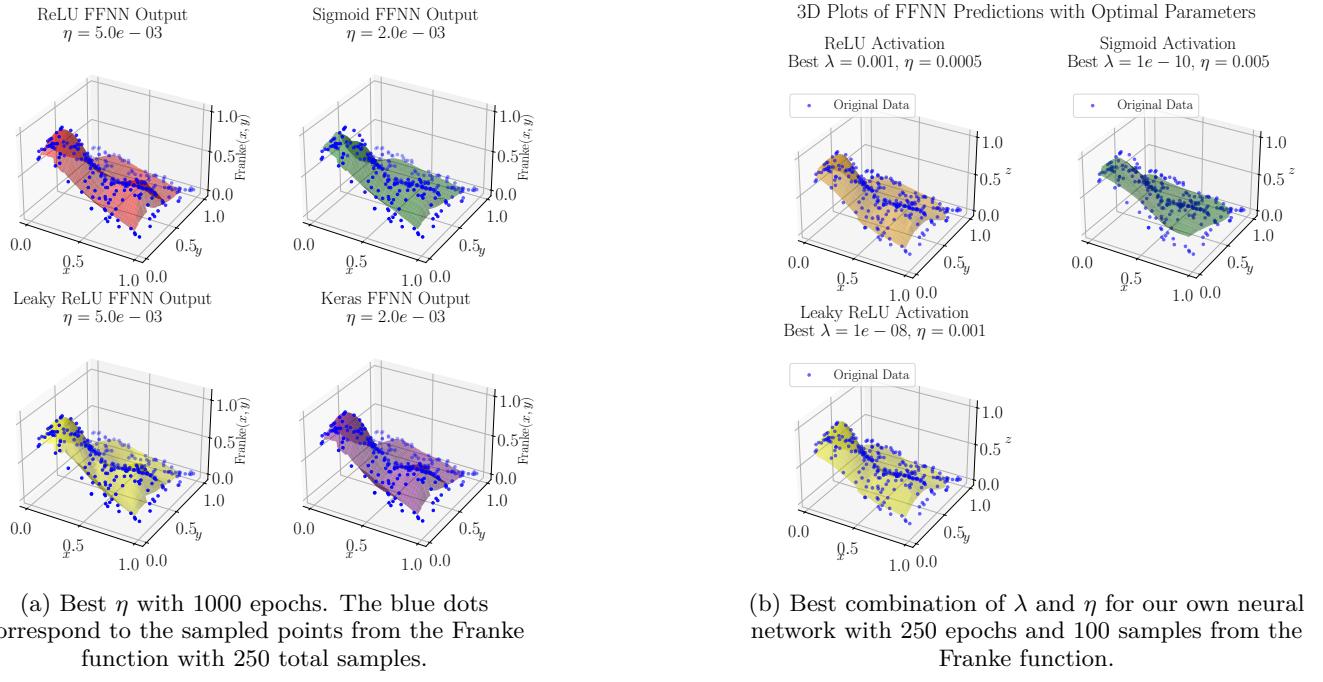


Fig. 7: 3D plots with (right) and without (left) regularization. The left figure also contains keras' prediction for comparison.

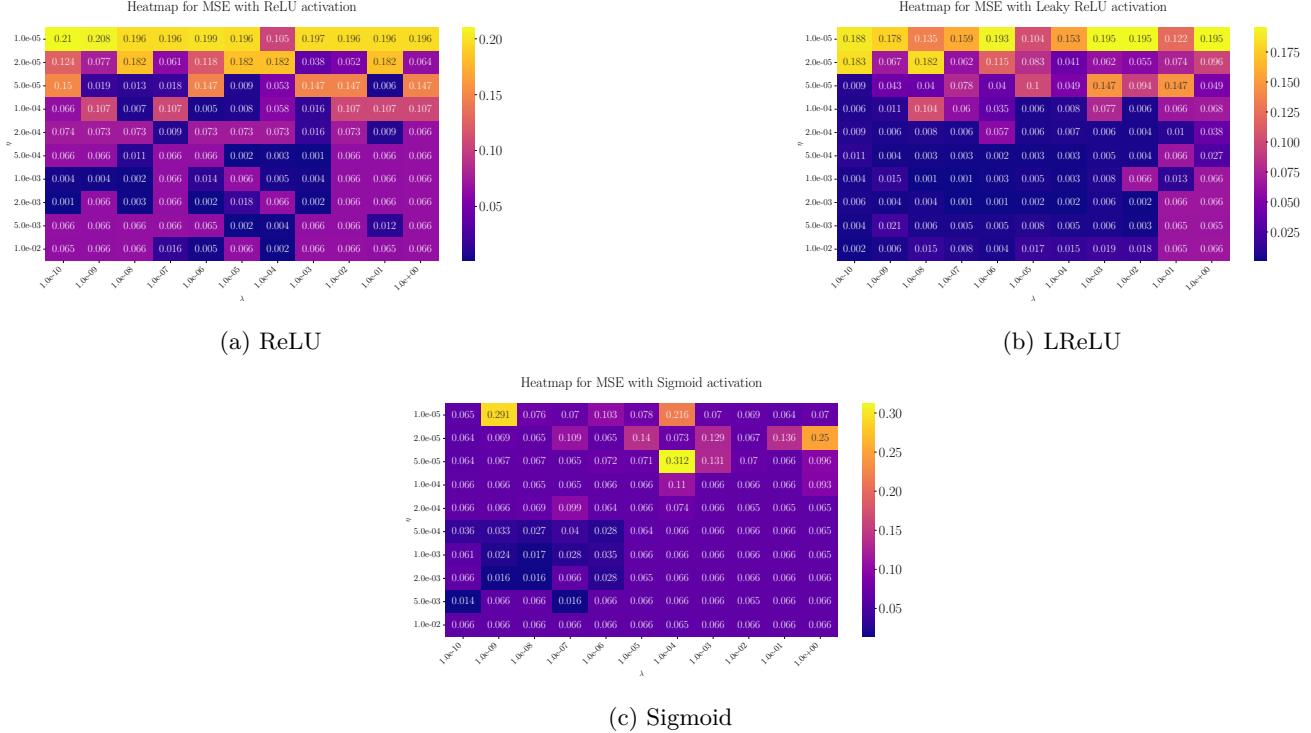


Fig. 8: MSE for various combinations of η and λ for ReLU, LReLU and Sigmoid activation functions with 250 epochs on the Franke function.

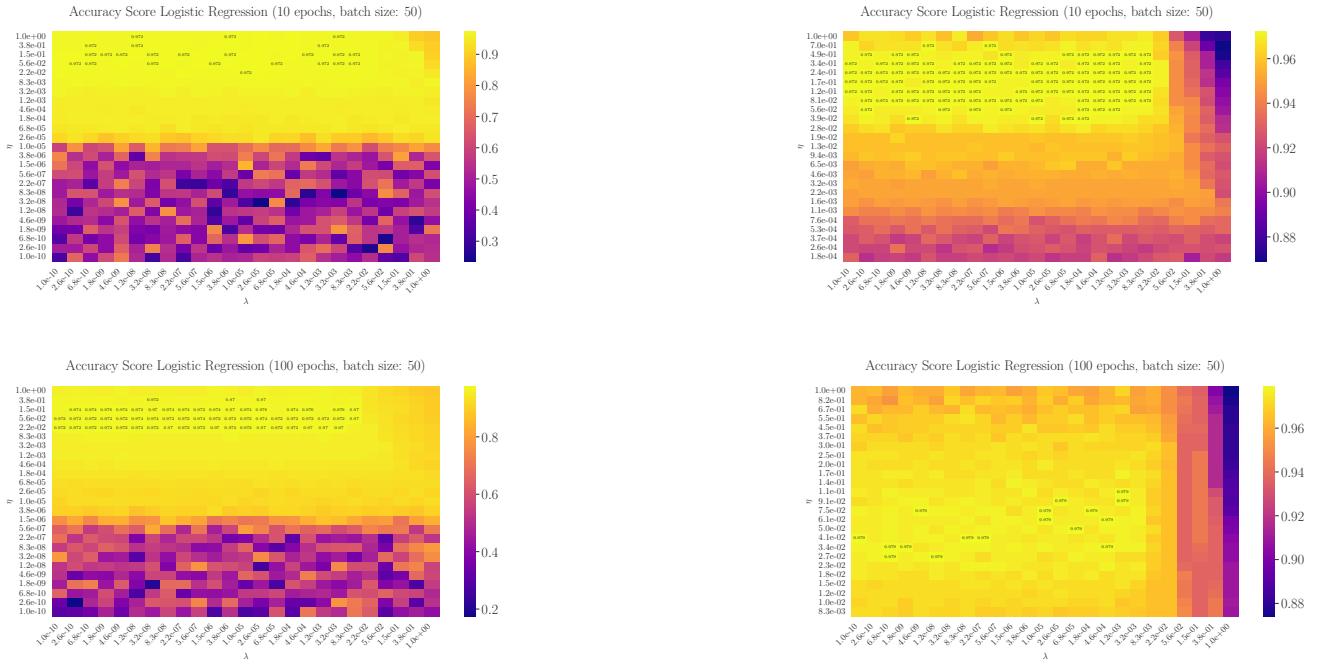


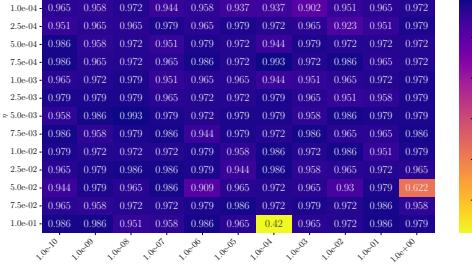
Fig. 9: Accuracy score for logistic regression, for number of epochs $N = 10$ (upper), $N = 100$ (lower). The figures to the right are zoomed in versions of the ones on the left. A varying learning rate has been used, with $t_0 = 2$, $t_1 = 2/\eta_{\text{const}} \in [10^{-10}, 1]$

Accuracy for FFNN with activation relu, 100 epochs and 50 batch size



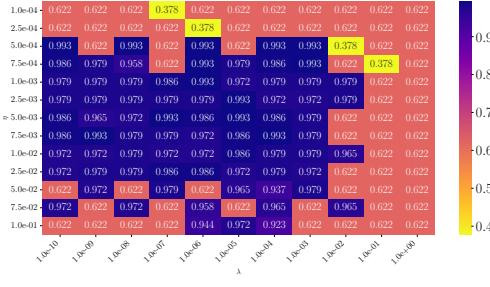
(a) ReLU

Accuracy for FFNN with activation lrelu, 100 epochs and 50 batch size



(b) LReLU

Accuracy for FFNN with activation sigmoid, 100 epochs and 50 batch size



(c) Sigmoid

Fig. 10: Accuracy for various combinations of η and λ for ReLU, LReLU and Sigmoid activation functions with 100 epochs and 50 batch size on the breast cancer dataset.