Ben-Gurion University - IEM Department
Introduction to Deep Learning (364-2-1071)
2023 Semester B
Assignment 1
Ofir azulay and Moshiko cohen

Question 1

a)

$$x^T Bx = \sum_{i,j} x_j B_{ij} x_i$$

$$[\frac{\partial x^T Bx}{\partial x_k}]_k := \frac{\partial}{\partial x_k} \sum_{i,j} x_j B_{ij} x_i = \sum_{i,j} \frac{\partial}{\partial x_k} x_j B_{ij} x_i = \sum_{i,j} \frac{\partial x_j}{\partial x_k} B_{ij} x_i + \sum_{i,j} x_i B_{ij} \frac{\partial x_j}{\partial x_k} =$$
$$\sum_{i,j} \delta_{jk} B_{ij} x_i + \sum_{i,j} x_j B_{ij} \delta_{ik} = \sum_i B_{ik} x_i + \sum_j x_j B_{ki} = [Bx^T]_k + [X^T B]_k = [X^T B^T]_k + [X^T B]_k$$

final result:

$$\boxed{\frac{\partial \mathbf{x^T Bx}}{\partial \mathbf{x}} = \mathbf{x^T}(\mathbf{B^T} + \mathbf{B})}$$

b)

$$[\frac{\partial tr(VxW)}{\partial x}]_{pq} = \frac{\sum_{ilk} V_{ik} X_{kl} W_{li}}{\partial X_{pq}} = \sum_{ikl} V_{ik} \frac{\partial X_{kl}}{\partial X_{pq}} W_{li} = \sum_{ilk} V_{ik} \delta_{kp} \delta_{lq} W_{li} =$$
$$\sum_{il} V_{ip} \delta_{pp} \delta_{lq} W_{li} = \sum_i V_{ip} \delta_{qq} W_{qi} = \sum_i V_{ip} W_{qi}$$

final result:

$$\boxed{\frac{\partial \mathbf{tr(VxW)}}{\partial \mathbf{x}} = \mathbf{V^T W^T}}$$

c)

$$\frac{\partial \|\mathbf{W}\|}{\partial W} := [\frac{\partial \sqrt{W^T W}}{\partial W}]_k = \frac{1}{2\sqrt{\sum_i (W_i)^2}} \frac{\sum_i (W_i)^2}{\partial_k} = \frac{1}{2\sqrt{\sum_i (W_i)^2}} 2W_k = \frac{1}{\sqrt{\sum_i (W_i)^2}} W_k = \frac{W_k}{\|\mathbf{W}\|}$$

final result:

$$\boxed{\frac{\partial \|\mathbf{W}\|}{\partial \mathbf{W}} = \frac{\mathbf{W}}{\|\mathbf{W}\|}}$$

d)

$$\frac{\partial tr(S)}{\partial S} := [\frac{\partial tr(S)}{\partial S}]_{pq} = \frac{\sum_i S_{ii}}{\partial S_{pq}} = \sum_i \frac{S_{ii}}{\partial S_{pq}} = \sum_i \delta_{ip}\delta_{iq} = \delta_{pp}\delta_{qq} = 1$$

final result:

$$\boxed{\frac{\partial \mathbf{tr}(\mathbf{S})}{\partial \mathbf{S}} = \mathbf{I}}$$

Question 2

a.i

$$\frac{\partial L}{\partial W} := [\frac{\partial L}{\partial W}]_{pq} = \frac{\partial L}{\partial W_{pq}} = \sum_{cd} \frac{\partial L}{\partial Y_{cd}} \frac{\partial_{cd}}{\partial W_{pq}}$$

$$\frac{\partial Y_{cd}}{\partial W_{pq}} = \frac{\partial \sum_i X_{ci} W^T_{id}}{\partial W_{pq}} = \sum_i X_{ci} \frac{\partial W_{di}}{\partial W_{pq}} = \sum_i X_{ci}\delta_{dp}\delta_{iq} = X_{cq}\delta_{dp}$$

$$\frac{\partial L}{\partial W_{pq}} = \sum_{cd} \frac{\partial L}{\partial Y_{cd}} X_{cq}\delta_{dp} = \sum_c \frac{\partial L}{\partial Y_{cp}} X_{cp}$$

final result:

$$\boxed{\frac{\partial \mathbf{L}}{\partial \mathbf{W}} = (\frac{\partial \mathbf{L}}{\partial \mathbf{Y}})^{\mathbf{T}}\mathbf{X}}$$

a.ii

$$\frac{\partial L}{\partial b} := [\frac{\partial L}{\partial b}]_i = \frac{\partial L}{\partial b_i} = \sum_{pq} \frac{\partial L}{\partial Y_{pq}} \frac{\partial Y_{pq}}{\partial b_i}$$

$$\frac{\partial Y_{pq}}{\partial b_i} = \frac{\partial [\sum_k X_{pk} W^T_{kq} + b_{pq}]}{\partial b_i} = \frac{\partial \sum_k X_{pk} W^T_{pk}}{\partial b_i} + \frac{\partial b_{pq}}{\partial b_i} = \frac{\partial b_{pq}}{\partial b_i} = \frac{\partial b_q}{\partial b_i} = \delta_{qi}$$

$$\frac{\partial L}{\partial b_i} = \sum_{pq} \frac{\partial L}{\partial Y_{pq}} \delta_{qi} = \sum_p \frac{\partial L}{\partial Y_{pi}} \delta_{ii} = \sum_p \frac{\partial L}{\partial Y_{pi}} = \sum_p [\frac{\partial L}{\partial Y}]_{pi}$$

final result:

$$\boxed{\frac{\partial \mathbf{L}}{\partial \mathbf{b}} = (\mathbf{all} - \mathbf{one} - \mathbf{vector})(\frac{\partial \mathbf{L}}{\partial \mathbf{Y}})}$$

a.iii

$$\frac{\partial L}{\partial X} := [\frac{\partial L}{\partial X}]_{pq} = \frac{\partial L}{\partial X_{pq}} = \sum_{cd} \frac{\partial L}{\partial Y_{cd}} \frac{\partial Y_{cd}}{\partial X_{pq}}$$

$$\frac{\partial Y_{cd}}{\partial X_{pq}} = \frac{\partial (\sum_i X_{ci} W^T_{id} + B_{cd})}{\partial X_{pq}} = \frac{\partial \sum_i X_{ci} W^T_{id}}{\partial X_{pq}} + \frac{\partial B_{cd}}{\partial X_{pq}} =$$

$$\sum_i \frac{\partial X_{ci}}{\partial X_{pq}} W^T_{id} = \sum_i \delta_{cp} \delta_{iq} W_{di} = \delta_{cp} \delta_{qq} W_{dq} = \delta_{cd} W_{dq}$$

$$\frac{\partial L}{\partial X_{pq}} = \sum_{cd} \frac{\partial L}{cd} \delta_{cp} W_{dq} = \sum_d \frac{\partial L}{\partial Y_{dd}} \delta_{pp} W_{dq} = \sum_d \frac{\partial L}{\partial Y_{pd}} W_{dq} = \sum_d [\frac{\partial L}{\partial Y}]_{pd} W_{dq}$$

final result:

$$\boxed{\frac{\partial \mathbf{L}}{\partial \mathbf{X}} = \frac{\partial \mathbf{L}}{\partial \mathbf{Y}} \mathbf{W}}$$

b.i

$$\frac{\partial L}{\partial X} := [\frac{\partial L}{\partial X}]_{pq} = \frac{\partial L}{\partial X_{pq}} = \sum_{cd} \frac{\partial L}{\partial Y_{cd}} \frac{\partial Y_{cd}}{\partial X_{pq}}$$

$$\frac{\partial Y_{cd}}{\partial X_{pq}} = \frac{\partial h(X_{cd})}{\partial X_{pq}} \delta_{cp} \delta_{dq}$$

$$\frac{\partial L}{\partial X_{pq}} = \sum_{cd} \frac{\partial L}{\partial Y_{cd}} \frac{\partial h(X_{cd})}{\partial X_{pq}} \delta_{cp} \delta_{dp} = \frac{\partial L}{\partial Y_{pq}} \frac{\partial h(X_{pq})}{\partial X_{pq}}$$

final result:

$$\boxed{\frac{\partial \mathbf{L}}{\partial \mathbf{X}} = \frac{\partial \mathbf{L}}{\partial \mathbf{Y}} \circ \frac{\partial \mathbf{h}(\mathbf{X})}{\partial \mathbf{X}}}$$

b.ii

$$\frac{\partial L}{\partial X} := [\frac{\partial L}{\partial X}]_{pq} = \frac{\partial L}{\partial Y} \frac{\partial h(X)}{\partial X}$$

$$U(X) = \begin{cases} 1, & \text{x>0} \\ 0, & \text{x<=0} \end{cases}$$

final result:

$$\boxed{\frac{\partial \mathbf{L}}{\partial \mathbf{X}} = \frac{\partial \mathbf{L}}{\partial \mathbf{Y}} \circ \mathbf{U}(\mathbf{X})}$$

c.i

$$\frac{\partial L}{\partial X} := [\frac{\partial L}{\partial X}]_{pq} = \frac{\partial L}{\partial X_{pq}} = \sum_{cd} \frac{\partial L}{\partial Y_{cd}} \frac{Y_{cd}}{\partial X_{pq}}$$

$$\frac{\partial Y_{cd}}{\partial X_{pq}} = \frac{\partial \frac{\exp X_{cd}}{\sum_k \exp X_{ck}}}{\partial X_{pq}} = \frac{\frac{\partial \exp X_{cd}}{\partial X_{pq}} \sum_k \exp X_{ck} - [\exp X_{cd} \frac{\partial \sum_k \exp X_{ck}}{\partial X_{pq}}]}{(\sum_k \exp X_{ck})^2}$$

$$\frac{\exp X_{cd}\delta_{cp}\delta_{dq} \sum_k \exp X_c k - [\exp X_c \sum_k \exp X_{ck}\delta_{cp}\delta_{kp}]}{(\sum_k \exp X_{ck})^2}$$

$$\exp X_{cd}\delta_{cp}\delta_{pd} - \frac{\exp X_{cd} \exp X_{cq}\delta_{cp}}{(\sum_k \exp X_{ck})^2}$$

$$\frac{\partial L}{\partial X} = \sum_{cd} \frac{\partial L}{\partial Y_{cd}} [\frac{\exp X_{cd}\delta_{cp}\delta_{dq}}{\sum_k \exp X_{ck}} - \frac{\exp X_{cd}\delta_{cp}\delta_{dq}}{(\sum_k \exp X_{ck})^2}]$$

$$= \sum_{d} \frac{\partial L}{\partial Y_{pd}} [\frac{\exp X_{pd}\delta_{dq}}{\sum_k \exp X_{pk}} - \frac{\exp X_{pd} \exp X_{pq}}{(\sum_k \exp X_{pk})^2}]$$

$$= \sum_{d} \frac{\partial L}{\partial Y_{pd}} \frac{\exp X_{pq}\delta_{dq}}{\sum_k \exp X_{pk}} - \sum_{d} \frac{\partial L}{\partial Y_{pd}} \frac{\exp X_{pd} \exp X_{pq}}{(\sum_k \exp X_{pk})^2}$$

$$= [\frac{\partial L}{\partial Y}]_{pq} \frac{\exp X_{pq}}{\sum_k \exp X_{pk}} - \sum_{d} \frac{\partial L}{\partial Y_{pd}} \frac{\exp X_{pd}}{\sum_k \exp X_{pk}} \frac{\exp X_{pq}}{\sum_k \exp X_{pk}}$$

$$[\frac{\partial L}{\partial Y}]_{pq} Y_{pq} - \sum_{d} [\frac{\partial L}{\partial Y}]_{pd} Y_{pd} Y_{pq}$$

final result:

$$\boxed{\frac{\partial \mathbf{L}}{\partial \mathbf{X}} = \frac{\partial \mathbf{L}}{\partial \mathbf{Y}} \circ \mathbf{Y} - (\frac{\partial \mathbf{L}}{\partial \mathbf{Y}} \circ \mathbf{Y} * \mathbf{1}) \circ \mathbf{Y}}$$

c.ii

$$\frac{\partial L}{\partial X} := [\frac{\partial L}{\partial X}]_{pq} = \frac{\partial L}{\partial X_{pq}} = \frac{\partial [-\frac{1}{s} \sum_{ik} T_{ik} log(X_{ik})]}{\partial X_{pq}}$$

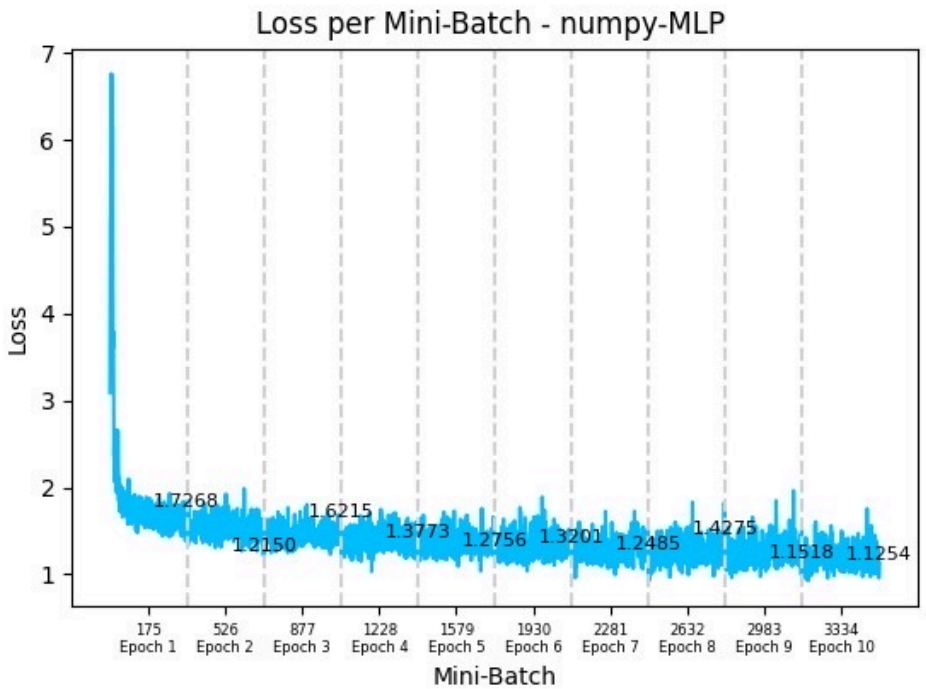
$$= -\frac{1}{s} \sum_{ik} T_{ik} \frac{\partial (log(X_{ik}))}{\partial X_{pq}} = -\frac{1}{s} \sum_{ik} T_{ik} \frac{1}{X_{ik}} \delta_{ip} \delta_{kp} = -\frac{1}{s} [T_{pq} \frac{1}{X_{pq}}]$$

final result:

$$\boxed{\frac{\partial \mathbf{L}}{\partial \mathbf{X}} = -\frac{\mathbf{1}}{\mathbf{s}} \mathbf{T} \oslash \mathbf{X}}$$

Question 3

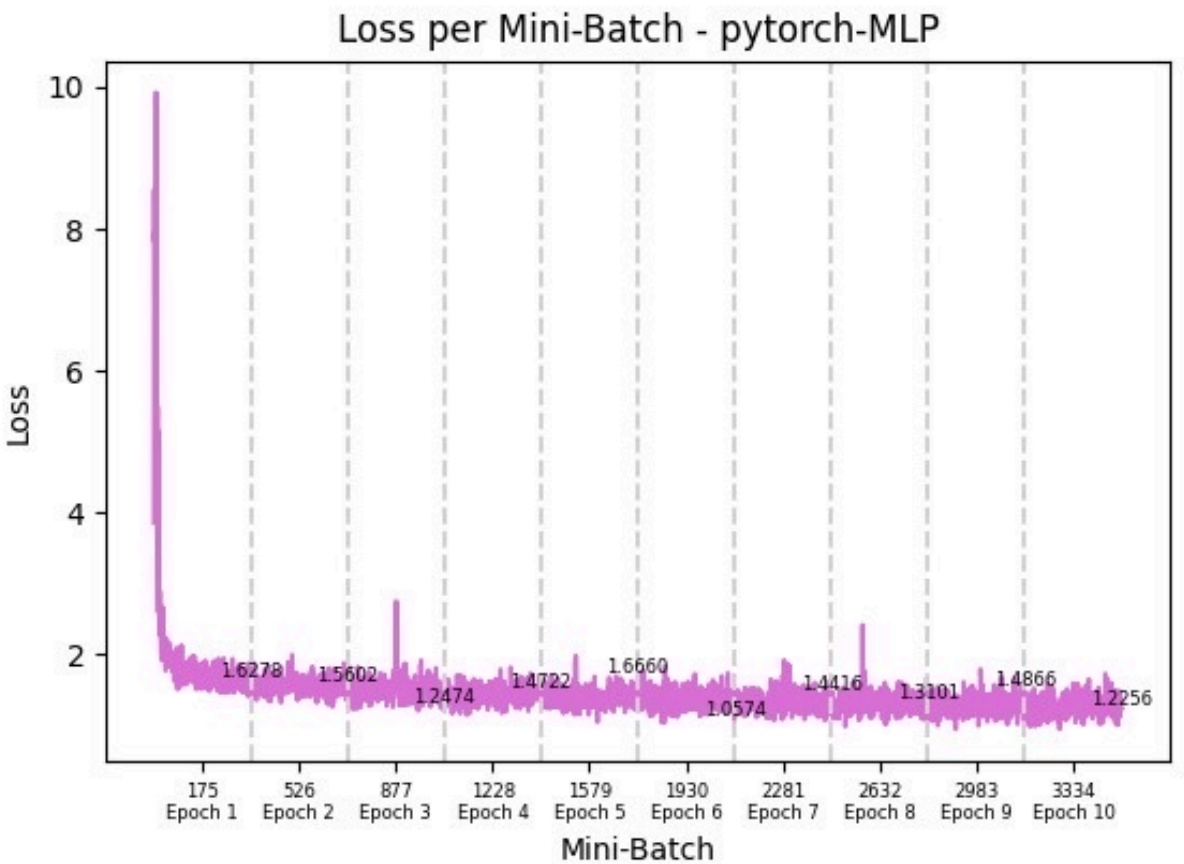
Implement a multi-layer perceptron by Numpy.
the following plot show the values of the loss function over time (for each batch in each epoch) for the model implemented with the default values of parameters (one layer, 128 hidden units, 10 epochs, no batch normalization, learning rate 0.1) in numpy.




```
validation accuracies from epoch 1 to epoch 10 is:
[0.4306, 0.4726, 0.4728, 0.4692, 0.4896, 0.4856, 0.4962, 0.4776, 0.494, 0.4932]
 Validation best accuracy is  :  0.4962
Test Accuracy With Best Model: 0.4912
```

Question 4

a) the following plot show the values of the loss function over time (for each batch in each epoch) for the model implemented with the default values of parameters (one layer, 128 hidden units, 10 epochs, no batch normalization, learning rate 0.1) in pytorch.



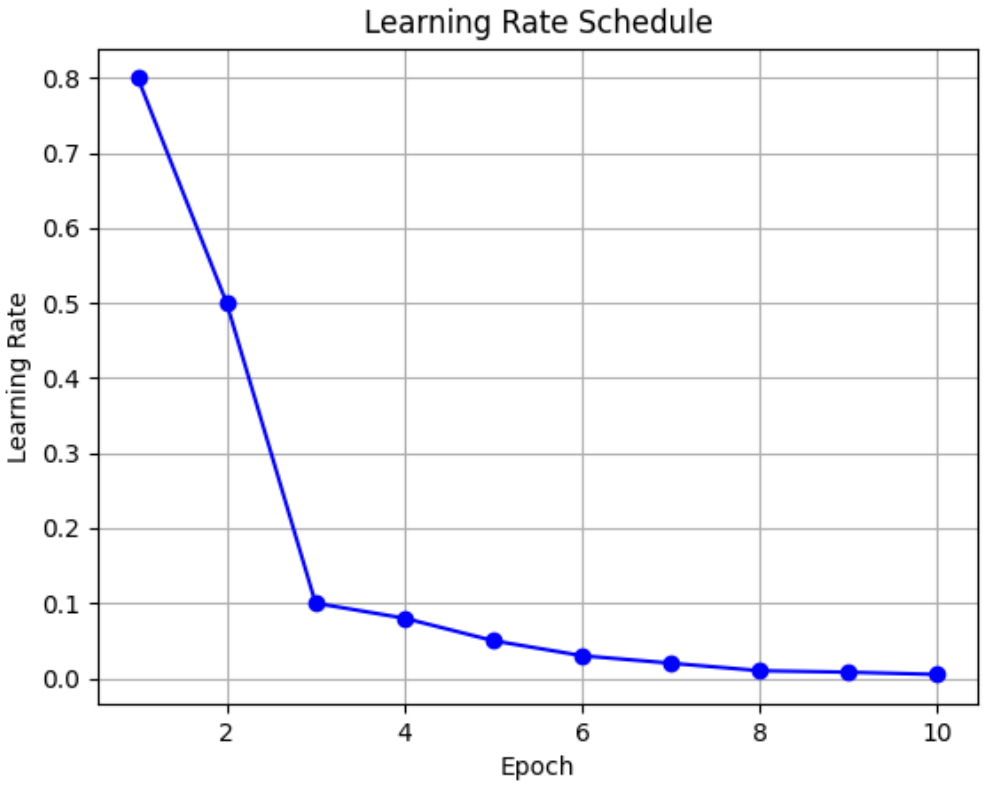

```
validation accuracies from epoch 1 to epoch 10 is:
[0.4234, 0.46, 0.456, 0.4714, 0.4734, 0.4676, 0.492, 0.4882, 0.4972, 0.5012]
new current best accuracy :  0.5012 and it found in epoch:  10
 Validation best accuracy is  :  0.5012
Test Accuracy With Best Model: 0.4974
```

b.i) The learning rate controls the step size at which the model's parameters are updated during the optimization process, such as gradient descent. The choice of learning rate can significantly impact the training process and the performance of the model. Therefore, the learning rate may be the most important parameter when configuring the neural network. too low learning rate may result in slow convergence, requiring more time and epochs to train the model. the model might stay at a local minimum point leading to lower accuracy on both the training and validation sets. The value of the loss function decreases at a low rate and may converge to a relatively high value. However, if we set the learning rate too high the model's parameter updates can be too large. This may cause the model to overshoot the optimal solution resulting in instability or divergence (We may miss the minimum point). The loss function may fluctuate significantly and the loss increases rather than decreases during training. it can lead to lower accuracy on both the training and validation sets. In summary, selecting the right learning rate is crucial for effectively training deep neural networks. It can significantly impact the loss, accuracy, and training time.

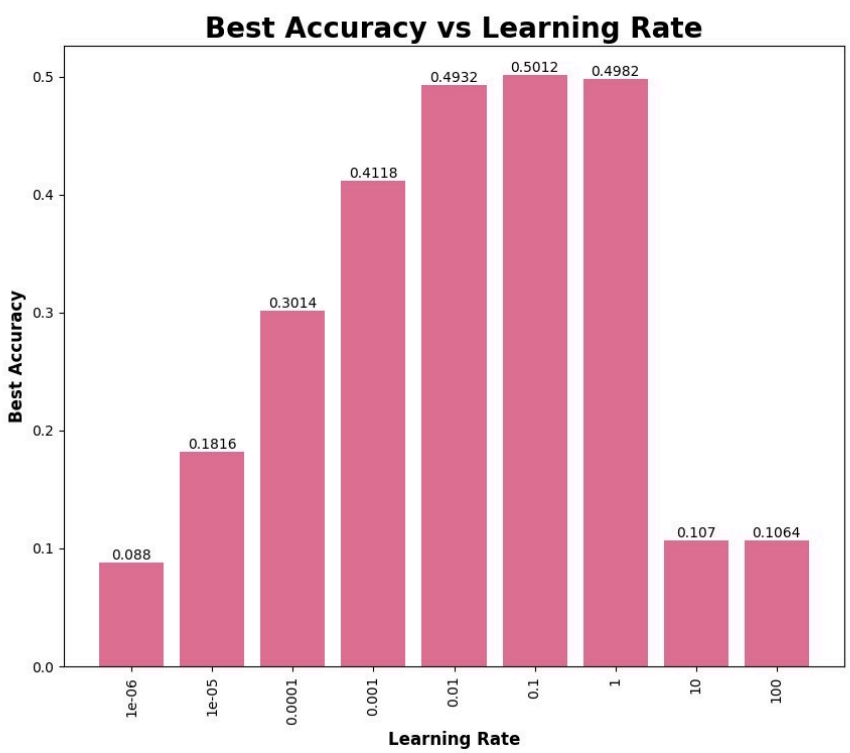
b.ii) The optimal strategy for adjusting the learning rate during training can vary depending on the nature of the problem being tackled such as the characteristics of the dataset, and the complexity of the model architecture. there is no one solution that works for all scenarios.
However, in general, it is advisable to start the training process with a relatively large learning rate what allows to do significant strides towards minimizing the loss function. As the training progresses, gradually decreasing the learning rate becomes essential to converge towards the global minimum and prevent overshooting.
One commonly employed technique is Time-based Decay, which involves reducing the learning rate based on the number of training steps or epochs. This approach often utilizes the formula:

$$new \quad learning \quad rate = \frac{initial \quad learning \quad rate}{(1 + decay \quad rate * epoch)}$$

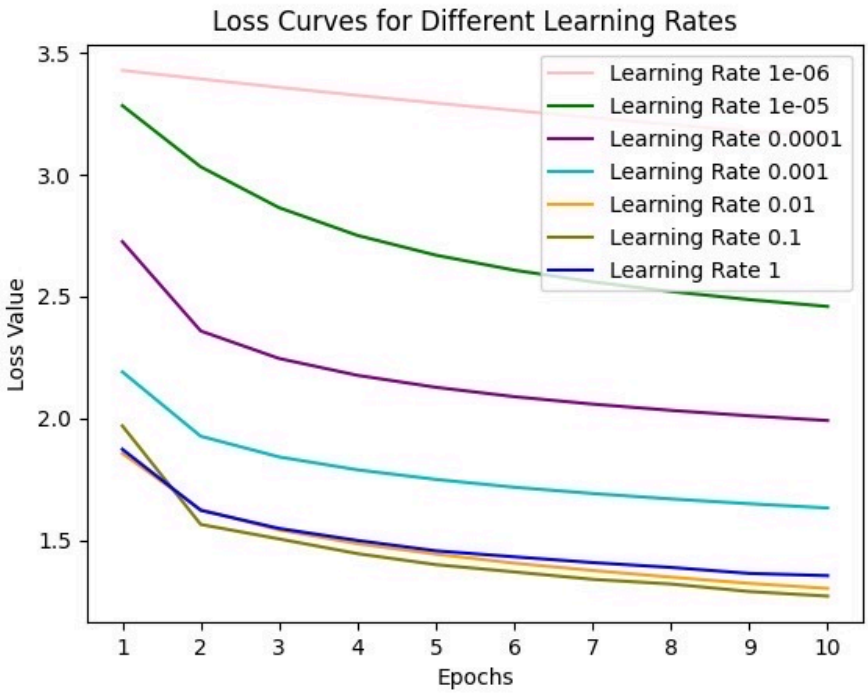
The decay rate parameter governs the rate at which the learning rate decreases over time. Ultimately, finding the most effective learning rate schedule requires experimentation and adaptation to the specific problem.
In the figure below you can see graph that illustrate the explanation theoretically the idea of a schedule.

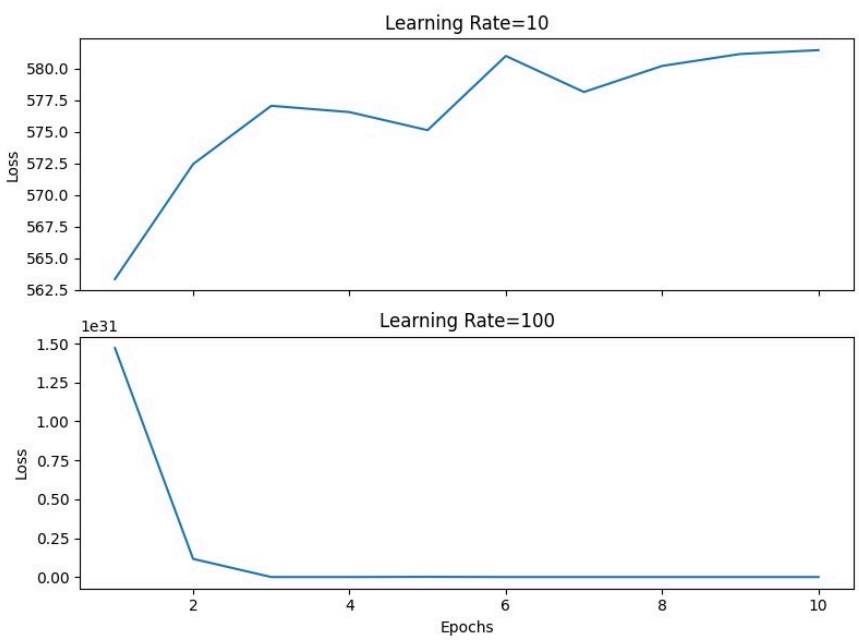


b.iv.1) the following plot show the Best validation accuracy as function of the learning rates



b.iv.2) the following plots show the values of the loss function over time (for each epoch) for each learning rates. For learning rates 1 10 and 100 batch normalization was performed because the loss function values were so high that their value was Nan in Python.
Batch normalization is a technique that helps mitigate the negative effects of high learning rates and enables their use without resulting in excessive loss values.
Even after batch normalization ,for learning rates 10 and 100 the values of the loss function were very high (but defined and not Nan) therefore the curve for learning rates 10 and 100 presented in a separate plot due to the constraints of the proportion of the graph.





Question 5

random variable f –> E(f)=0 , V(f)=σ^2

$b = ReLU(f) = max(f, 0)$

$b^2 = (max(f, 0))^2$

$E(b^2) = E(max(f, 0))^2$

$E(ReLU(f))^2) = V(ReLU(f)) + (E(ReLU(f)))^2 = \frac{\sigma^2}{2} + 0^2 = \frac{\sigma^2}{2}$

$E(b^2) = \frac{\sigma^2}{2}$

f is symmetric, the expectation is zero, and its variance is σ^2. Therefore, in positive values, the variance is half of the general variance.