Assignment 1- Introduction to Deep Learning

Yarden Cohen, Yotam Komash

June 21, 2024

Contents

1 Question 1

1.1 a) Given a vector $x \in \mathbb{R}^n$ and square matrix $B \in \mathbb{R}^{n \times n}$, evaluate $\frac{\partial x^\top Bx}{\partial x}$

Let $r = x^\top Bx$.

$$\left[\frac{\partial r}{\partial \mathbf{x}}\right]_i = \frac{\partial r}{\partial x_i} = \frac{\partial}{\partial x_i}(x^\top Bx) = \frac{\partial}{\partial x_i}\left(\sum_{m,n} x_m B_{mn} x_n\right) = \sum_{m,n} B_{mn}\frac{\partial}{\partial x_i}(x_m x_n)$$

Using the product rule ($\frac{d}{dx}(f(x)g(x)) = f'(x)g(x) + f(x)g'(x)$) we get:

$$\frac{\partial}{\partial x_i}(x_m x_n) = \frac{\partial x_m}{\partial x_i}x_n + x_m\frac{\partial x_n}{\partial x_i} = x_m\delta_{in} + x_n\delta_{im}$$

where δ is the Kronecker delta.

Thus:

$$\frac{\partial r}{\partial x_i} = \sum_{m,n} B_{mn}(x_m\delta_{in} + x_n\delta_{im}) = \sum_m B_{mi}x_m + \sum_n B_{in}x_n = (Bx + B^\top x)_i$$

Hence:

$$\frac{\partial x^\top Bx}{\partial x} = Bx + B^\top x$$

1.2 b) Given matrices $V \in \mathbb{R}^{n\times m}$, $X \in \mathbb{R}^{m\times p}$, and $W \in \mathbb{R}^{p\times n}$, find an expression for $\frac{\partial \operatorname{tr}(VXW)}{\partial X}$

Let $s = \operatorname{tr}(VXW)$.

$$s = \operatorname{tr}(VXW) = \sum_i (VXW)_{ii} = \sum_i \sum_{j,k} V_{ij} X_{jk} W_{ki}$$

Thus:

$$\frac{\partial s}{\partial X_{mn}} = \frac{\partial}{\partial X_{mn}} \left(\sum_{i,j,k} V_{ij} X_{jk} W_{ki} \right) = \sum_{i,j,k} V_{ij} W_{ki} \frac{\partial X_{jk}}{\partial X_{mn}}$$

Using the Kronecker delta:

$$\frac{\partial X_{jk}}{\partial X_{mn}} = \delta_{jm}\delta_{kn}$$

Thus:

$$\frac{\partial s}{\partial X_{mn}} = \sum_{i,j,k} V_{ij} W_{ki} \delta_{jm}\delta_{kn} = \sum_i V_{im} W_{ni} = (VW)_{mn}$$

Hence:

$$\frac{\partial \operatorname{tr}(VXW)}{\partial X} = VW$$

1.3 c) For a vector $w \in \mathbb{R}^n$ and its Euclidean norm $\|w\| := \sqrt{w^\top w}$, calculate $\frac{\partial \|w\|}{\partial w}$

Let $s = \|w\| = \sqrt{w^\top w}$.

$$s = (w^\top w)^{1/2}$$

Using the chain rule:

$$\frac{\partial s}{\partial w_i} = \frac{\partial (w^\top w)^{1/2}}{\partial w_i} = \frac{1}{2}(w^\top w)^{-1/2} \frac{\partial (w^\top w)}{\partial w_i}$$

Since:

$$\frac{\partial (w^\top w)}{\partial w_i} = \frac{\partial (\sum_j w_j^2)}{\partial w_i} = 2w_i$$

Thus:

$$\frac{\partial s}{\partial w_i} = \frac{1}{2}(w^\top w)^{-1/2} \cdot 2w_i = \frac{w_i}{(w^\top w)^{1/2}} = \frac{w_i}{\|w\|}$$

Hence:

$$\frac{\partial \|w\|}{\partial w} = \frac{w}{\|w\|}$$

1.4 d) Let S be a square matrix, find an expression for $\frac{\partial \operatorname{tr}(S)}{\partial S}$

Let $r = \operatorname{tr}(S)$.

$$r = \operatorname{tr}(S) = \sum_i S_{ii}$$

Thus:

$$\frac{\partial r}{\partial S_{mn}} = \frac{\partial}{\partial S_{mn}} \left(\sum_i S_{ii} \right) = \delta_{mn}$$

Hence:

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

2 Question 2

2.1 2.a Linear Module

Consider a linear module with input features X and output features Y.
The relationship is given by

$$Y = XW^\top + b$$

where W is the weight matrix and b is the bias row vector.
Given the gradients of the loss with respect to the output features $\frac{\partial L}{\partial Y}$, find closed-form expressions for $\frac{\partial L}{\partial W}$, $\frac{\partial L}{\partial b}$, and $\frac{\partial L}{\partial X}$.

Solution

Given:

$$L = L(Y)$$
$$Y = Y(X, W, b)$$

The derivatives of Y with respect to X, W, b are:

$$\frac{\partial Y}{\partial W} = \frac{\partial (XW^\top + b)}{\partial W} = X$$

$$\frac{\partial Y}{\partial b} = \frac{\partial (XW^\top + b)}{\partial b} = 1$$

$$\frac{\partial Y}{\partial X} = \frac{\partial (XW^\top + b)}{\partial X} = W$$

where 1 is an all-ones n vector.
Thus:

$$\frac{\partial L}{\partial W} = \frac{\partial L}{\partial Y} \cdot X$$

$$\frac{\partial L}{\partial b} = \frac{\partial L}{\partial Y} \cdot 1$$

$$\frac{\partial L}{\partial X} = \frac{\partial L}{\partial Y} \cdot W$$

2.2 2.b Activation Module

Consider an element-wise activation function h. The activation module has input features X and output features Y, such that $Y = h(X)$, where $Y_{ij} = h(X_{ij})$. Given the gradient of the loss with respect to the output features $\frac{\partial L}{\partial Y}$, find a closed-form expression for $\frac{\partial L}{\partial X}$.

Solution

i) For a generic activation function h:

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

where $\circ$ denotes the Hadamard (element-wise) product and $h'(X)$ is the element-wise derivative of h with respect to X.

ii) For the ReLU activation function $h(x) = \max(0, x)$:

$$\frac{\partial L}{\partial X} = \frac{\partial L}{\partial Y} \circ \mathbb{I}(X > 0)$$

where $\mathbb{I}(X > 0)$ is an indicator function that is 1 if $X > 0$ and 0 otherwise.

2.3 2.c Softmax and Loss Modules

2.3.1 i)

Consider a softmax module such that $Y_{ij} = [\text{softmax}(X)]_{ij}$, where X is the input and Y is the output of the module. Find a closed-form expression for $\frac{\partial L}{\partial X}$ in terms of $\frac{\partial L}{\partial Y}$.

Solution

The softmax function is given by:

$$Y_{ij} = \frac{e^{X_{ij}}}{\sum_k e^{X_{ik}}}$$

$$\frac{\partial Y_{ij}}{\partial X_{pq}} = \frac{\partial(\frac{e^{X_{ij}}}{\sum_k e^{X_{ik}}})}{\partial X_{pq}}$$

1. When $i \neq p$:

$$\frac{\partial Y_{ij}}{\partial X_{pq}} = 0$$

2. When $i = p$:

$$\frac{\partial Y_{pj}}{\partial X_{pq}} = \frac{\partial(\frac{e^{X_{pj}}}{\sum_k e^{X_{pk}}})}{\partial X_{pq}}$$

For a function $\frac{u}{v}$ the quotient rule is given by:

$$\frac{\partial}{\partial x}\left(\frac{u}{v}\right) = \frac{v\frac{\partial u}{\partial x} - u\frac{\partial v}{\partial x}}{v^2}$$

In our case, $u = e^{X_{pj}}$ and $v = \sum_k e^{X_{pk}}$.

First, compute $\frac{\partial u}{\partial x}$:

$$\frac{\partial e^{X_{pj}}}{\partial X_{pq}} = e^{X_{pj}} \cdot \frac{\partial X_{pj}}{\partial X_{pq}} = e^{X_{pj}} \cdot \delta_{jq}$$

Next, compute $\frac{\partial v}{\partial x}$:

$$\frac{\partial \left(\sum_k e^{X_{pk}}\right)}{\partial X_{pq}} = \sum_k \frac{\partial e^{X_{pk}}}{\partial X_{pq}} = \sum_k e^{X_{pk}} \frac{\partial X_{pk}}{\partial X_{pq}} = \sum_k e^{X_{pk}} \cdot \delta_{kq} = e^{X_{pq}}$$

Now, applying the quotient rule:

$$\frac{\partial Y_{pj}}{\partial X_{pq}} = \frac{\left(\sum_k e^{X_{pk}}\right)\left(e^{X_{pj}} \cdot \delta_{jq}\right) - e^{X_{pj}}\left(e^{X_{pq}}\right)}{\left(\sum_k e^{X_{pk}}\right)^2}$$

Simplifying, we get:

$$\frac{\partial Y_{pj}}{\partial X_{pq}} = \frac{e^{X_{pj}} \cdot \delta_{jq} \cdot \sum_k e^{X_{pk}} - e^{X_{pj}} \cdot e^{X_{pq}}}{\left(\sum_k e^{X_{pk}}\right)^2} = \frac{e^{X_{pj}} \cdot \delta_{jq}}{\sum_k e^{X_{pk}}} - \frac{e^{X_{pj}}}{\sum_k e^{X_{pk}}} \cdot \frac{e^{X_{pq}}}{\sum_k e^{X_{pk}}}$$

Since $Y_{mn} = \frac{e^{X_{mn}}}{\sum_k e^{X_{mk}}}$, we have derived:

$$\frac{\partial Y_{pj}}{\partial X_{pq}} = Y_{pj} \cdot \delta_{jq} - Y_{pj} \cdot Y_{pq} = Y_{pj}(\delta_{jq} - Y_{pq})$$

Using the chain rule, the gradient of the loss with respect to X is:

$$\frac{\partial L}{\partial X_{pq}} = \sum_j \frac{\partial L}{\partial Y_{pj}} \cdot \frac{\partial Y_{pj}}{\partial X_{pq}} = \sum_j \frac{\partial L}{\partial Y_{pj}} \cdot Y_{pj}(\delta_{jq} - Y_{pq}) = \frac{\partial L}{\partial Y_{pq}} \cdot Y_{pq}(1 - Y_{pq})$$

For the entire matrix X, this can be written compactly as:

$$\frac{\partial L}{\partial X} = \frac{\partial L}{\partial Y} \cdot Y(1 - Y)$$

where 1 is an all-ones $m \times n$ matrix.

2.3.2 ii)

The categorical cross-entropy loss function for a given example i and class k is:

$$L_i = -\sum_k T_{ik} \log(X_{ik})$$

Were, T represents the target labels and X represents the predicted probabilities from the softmax

output.
The total loss over all examples is:

$$L = \frac{1}{S} \sum_i L_i = -\frac{1}{S} \sum_{i,k} T_{ik} \log(X_{ik})$$

To find the gradient of the loss L with respect to the input X, we take the partial derivative of L with respect to X_{ij}.

$$\frac{\partial L}{\partial X_{ij}} = \frac{\partial}{\partial X_{ij}} \left(-\frac{1}{S} \sum_{i,k} T_{ik} \log(X_{ik}) \right) = -\frac{1}{S} \sum_k T_{ik} \frac{\partial \log(X_{ik})}{\partial X_{ij}}$$

The derivative of the logarithm function $\log(X_{ik})$ with respect to X_{ij} is:

$$\frac{\partial \log(X_{ik})}{\partial X_{ij}} = \frac{1}{X_{ik}} \delta_{kj}$$

Substituting this back into our expression, we get:

$$\frac{\partial L}{\partial X_{ij}} = -\frac{1}{S} \sum_k T_{ik} \frac{1}{X_{ik}} \delta_{jk} = -\frac{1}{S} \frac{T_{ij}}{X_{ij}}$$

For the entire matrix X, this can be written compactly as:

$$\frac{\partial L}{\partial X} = -\frac{1}{S} T \oslash X$$

where $\oslash$ is the division is element-wise, i.e., $[A \oslash B]_{ij} = \frac{A_{ij}}{B_{ij}}$.

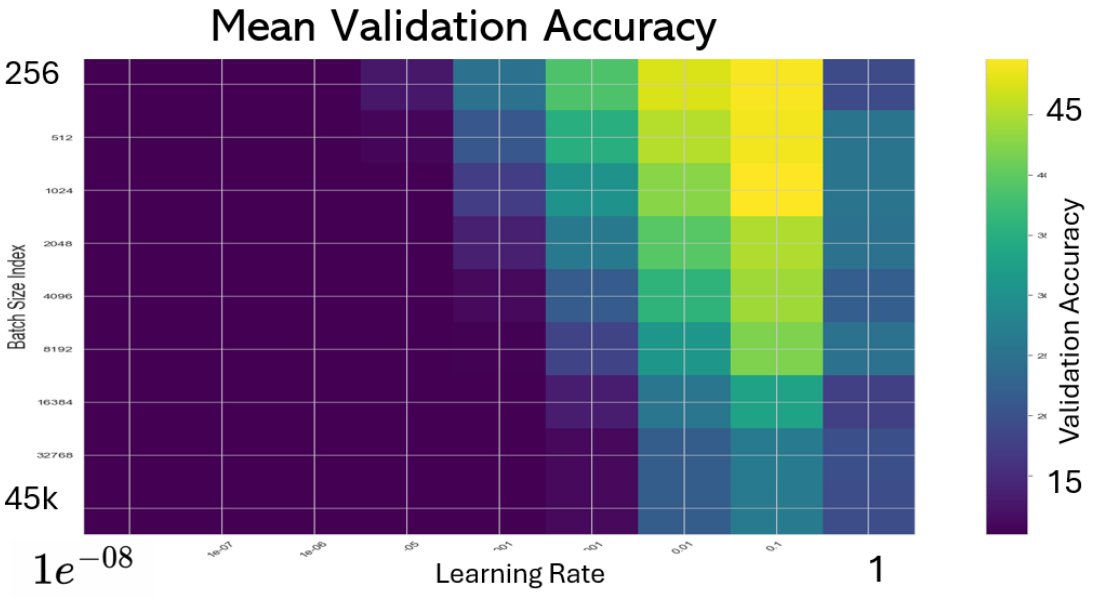

Figure 1: First looking at the results, looks like we need to explore smaller batch sizes

3 Question 3

3.1 Impact of Learning Rate and Batch Size on Convergence

The plots in Figure **??** showcase the train and validation accuracy per epoch for various combinations of learning rates and batch sizes. In the train accuracy plots, we observe that for smaller learning rates (0.000001 - 0.01), smaller batch sizes (16, 32, 64) generally lead to better convergence and higher final accuracy. However, as the learning rate increases to around 0.05 and above, the performance advantage of small batch sizes diminishes, and larger batch sizes (256, 512, 1024) start to outperform.

The highest Validation accuracy of 51.8% is achieved with a learning rate of 0.05. Figure **??** showcase The graphs clearly illustrate the interplay between learning rate and batch size in terms of convergence behavior and final accuracy.

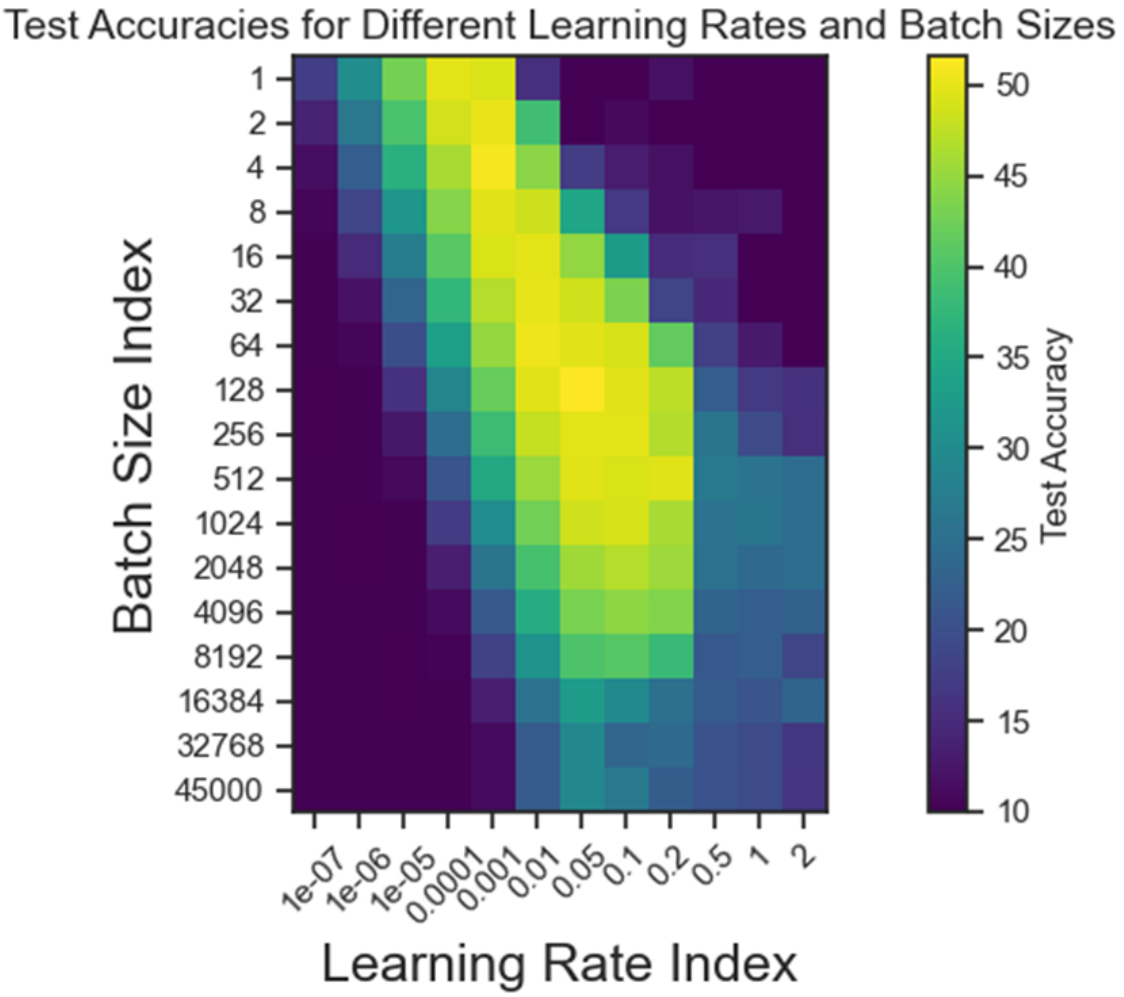

Figure 2: Mean validation accuracy of a range of batch size and learning rate combinations. from yellow (high accuracy) to dark blue (low accuracy).

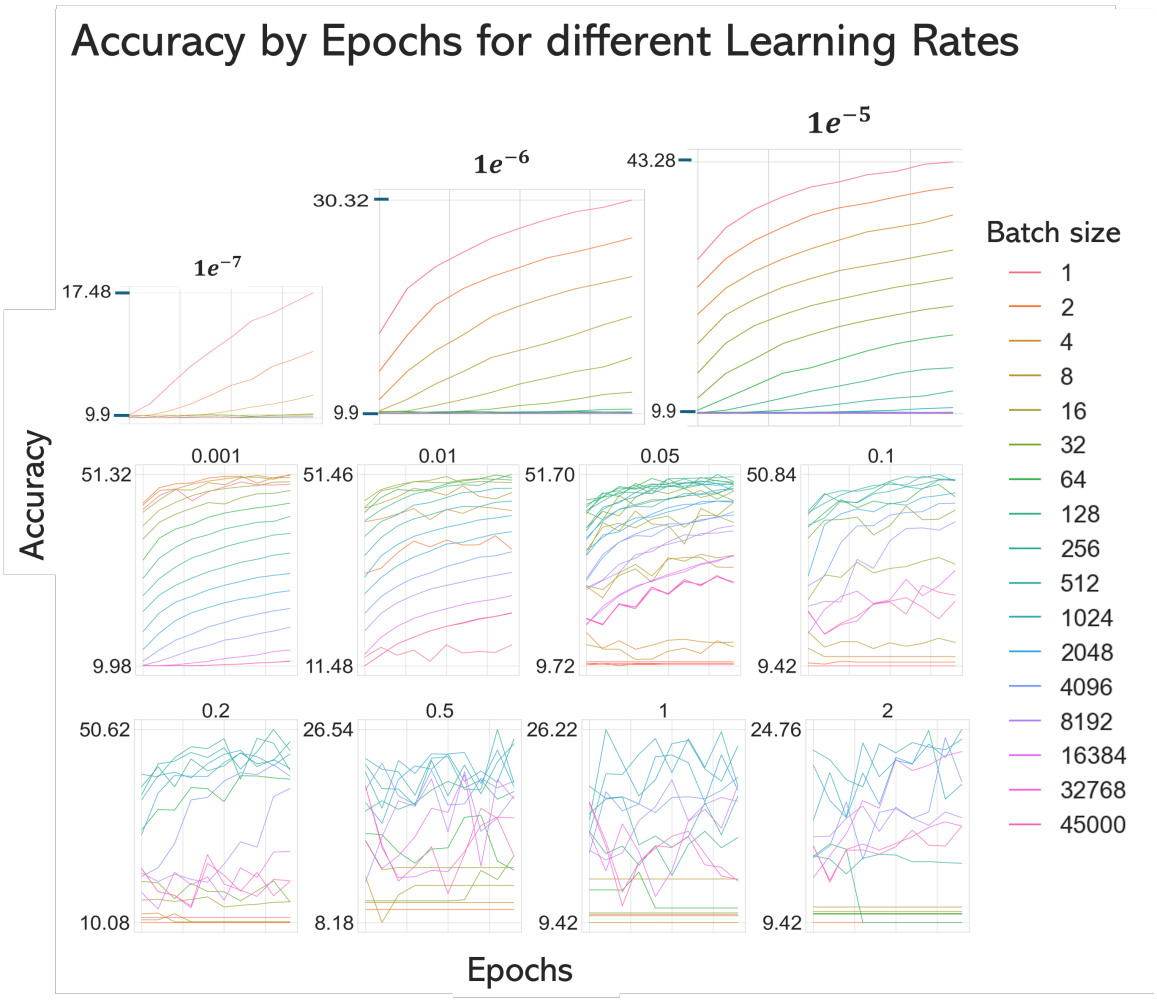

Figure 3: Impact of Learning Rate and Batch Size on Convergence of the Train set