7 Gradients and Initialization

Problem 7.1 A two-layer network with two hidden units in each layer can be defined as:

$$
\begin{aligned}
y = \phi_0 + \phi_1 a[\psi_{01} + \psi_{11} a[\theta_{01} + \theta_{11} x] + \psi_{21} a[\theta_{02} + \theta_{12} x]] \\
+ \phi_2 a[\psi_{02} + \psi_{12} a[\theta_{01} + \theta_{11} x] + \psi_{22} a[\theta_{02} + \theta_{12} x]]
\end{aligned}
$$

where the functions $a[\bullet]$ are ReLU functions. Compute the derivatives of the output y with respect to each of the 13 parameters $\phi_{\bullet}$, $\theta_{\bullet\bullet}$, and $\psi_{\bullet\bullet}$ directory (i.e. not using the backpropagation algorithm). The derivative of the ReLU function with respect to the its input $\partial a[z]/\partial z$ is the indicator function $\mathbb{I}[z > 0]$, which returns one if the argument is greater than zero and zero otherwise (figure 7.6).

$$
\begin{aligned}
\frac{\partial y}{\partial \phi_0} =& 1 \\
\frac{\partial y}{\partial \phi_1} =& a[\psi_{01} + \psi_{11} a[\theta_{01} + \theta_{11} x] + \psi_{21} a[\theta_{02} + \theta_{12} x]] \\
\frac{\partial y}{\partial \phi_2} =& a[\psi_{02} + \psi_{12} a[\theta_{01} + \theta_{11} x] + \psi_{22} a[\theta_{02} + \theta_{12} x]] \\
\frac{\partial y}{\partial \psi_{01}} =& \phi_1 \mathbb{I}[\psi_{01} + \psi_{11} a[\theta_{01} + \theta_{11} x] + \psi_{21} a[\theta_{02} + \theta_{12} x] > 0] \\
\frac{\partial y}{\partial \psi_{11}} =& \phi_1 \mathbb{I}[\psi_{01} + \psi_{11} a[\theta_{01} + \theta_{11} x] + \psi_{21} a[\theta_{02} + \theta_{12} x] > 0] a[\theta_{01} + \theta_{11} x] \\
\frac{\partial y}{\partial \psi_{21}} =& \phi_1 \mathbb{I}[\psi_{01} + \psi_{11} a[\theta_{01} + \theta_{11} x] + \psi_{21} a[\theta_{02} + \theta_{12} x] > 0] a[\theta_{02} + \theta_{12} x] \\
\frac{\partial y}{\partial \psi_{02}} =& \phi_2 \mathbb{I}[\psi_{02} + \psi_{12} a[\theta_{01} + \theta_{11} x] + \psi_{22} a[\theta_{02} + \theta_{12} x] > 0] \\
\frac{\partial y}{\partial \psi_{11}} =& \phi_2 \mathbb{I}[\psi_{02} + \psi_{12} a[\theta_{01} + \theta_{11} x] + \psi_{22} a[\theta_{02} + \theta_{12} x] > 0] a[\theta_{01} + \theta_{11} x] \\
\frac{\partial y}{\partial \psi_{21}} =& \phi_2 \mathbb{I}[\psi_{02} + \psi_{12} a[\theta_{01} + \theta_{11} x] + \psi_{22} a[\theta_{02} + \theta_{12} x] > 0] a[\theta_{02} + \theta_{12} x] \\
\frac{\partial y}{\partial \theta_{01}} =& \phi_1 \psi_{11} \mathbb{I}[\psi_{01} + \psi_{11} a[\theta_{01} + \theta_{11} x] + \psi_{21} a[\theta_{02} + \theta_{12} x] > 0] \mathbb{I}[\theta_{01} + \theta_{11} x > 0] \\
& + \phi_2 \psi_{12} \mathbb{I}[\psi_{02} + \psi_{12} a[\theta_{01} + \theta_{11} x] + \psi_{22} a[\theta_{02} + \theta_{12} x] > 0] \mathbb{I}[\theta_{01} + \theta_{11} x > 0] \\
\frac{\partial y}{\partial \theta_{11}} =& \phi_1 \psi_{11} x \mathbb{I}[\psi_{01} + \psi_{11} a[\theta_{01} + \theta_{11} x] + \psi_{21} a[\theta_{02} + \theta_{12} x] > 0] \mathbb{I}[\theta_{01} + \theta_{11} x > 0] \\
& + \phi_2 \psi_{12} x \mathbb{I}[\psi_{02} + \psi_{12} a[\theta_{01} + \theta_{11} x] + \psi_{22} a[\theta_{02} + \theta_{12} x] > 0] \mathbb{I}[\theta_{01} + \theta_{11} x > 0] \\
\frac{\partial y}{\partial \theta_{02}} =& \phi_1 \psi_{21} \mathbb{I}[\psi_{01} + \psi_{11} a[\theta_{01} + \theta_{11} x] + \psi_{21} a[\theta_{02} + \theta_{12} x] > 0] \mathbb{I}[\theta_{02} + \theta_{12} x > 0] \\
& + \phi_2 \psi_{22} \mathbb{I}[\psi_{02} + \psi_{12} a[\theta_{01} + \theta_{11} x] + \psi_{22} a[\theta_{02} + \theta_{12} x] > 0] \mathbb{I}[\theta_{02} + \theta_{12} x > 0] \\
\frac{\partial y}{\partial \theta_{12}} =& \phi_1 \psi_{21} x \mathbb{I}[\psi_{01} + \psi_{11} a[\theta_{01} + \theta_{11} x] + \psi_{21} a[\theta_{02} + \theta_{12} x] > 0] \mathbb{I}[\theta_{02} + \theta_{12} x > 0] \\
& + \phi_2 \psi_{22} x \mathbb{I}[\psi_{02} + \psi_{12} a[\theta_{01} + \theta_{11} x] + \psi_{22} a[\theta_{02} + \theta_{12} x] > 0] \mathbb{I}[\theta_{02} + \theta_{12} x > 0]
\end{aligned}
$$

Problem 7.2 Find an expression for the final term in each of the five chains of derivatives in equation 7.12.

$$
\begin{aligned}
\frac{\partial l_i}{\partial f_2} &= \frac{\partial h_3}{\partial f_2}\frac{\partial f_3}{\partial h_3}\frac{\partial l_i}{\partial f_3} = (-\sin[f_2])(\omega_3)(2f_3) \\
\frac{\partial l_i}{\partial h_2} &= \frac{\partial f_2}{\partial h_2}\frac{\partial h_3}{\partial f_2}\frac{\partial f_3}{\partial h_3}\frac{\partial l_i}{\partial f_3} = (\omega_2)(-\sin[f_2])(\omega_3)(2f_3) \\
\frac{\partial l_i}{\partial f_1} &= \frac{\partial h_2}{\partial f_1}\frac{\partial f_2}{\partial h_2}\frac{\partial h_3}{\partial f_2}\frac{\partial f_3}{\partial h_3}\frac{\partial l_i}{\partial f_3} = (\exp[f_1])(\omega_2)(-\sin[f_2])(\omega_3)(2f_3) \\
\frac{\partial l_i}{\partial h_1} &= \frac{\partial f_1}{\partial h_1}\frac{\partial h_2}{\partial f_1}\frac{\partial f_2}{\partial h_2}\frac{\partial h_3}{\partial f_2}\frac{\partial f_3}{\partial h_3}\frac{\partial l_i}{\partial f_3} = (\omega_1)(\exp[f_1])(\omega_2)(-\sin[f_2])(\omega_3)(2f_3) \\
\frac{\partial l_i}{\partial f_0} &= \frac{\partial h_1}{\partial f_0}\frac{\partial f_1}{\partial h_1}\frac{\partial h_2}{\partial f_1}\frac{\partial f_2}{\partial h_2}\frac{\partial h_3}{\partial f_2}\frac{\partial f_3}{\partial h_3}\frac{\partial l_i}{\partial f_3} = (\cos[f_0])(\omega_1)(\exp[f_1])(\omega_2)(-\sin[f_2])(\omega_3)(2f_3)
\end{aligned}
$$

Problem 7.3 What size are each of the terms in equation 7.19?

$$
\frac{\partial l_i}{\partial \mathbf{f}_0} = \frac{\partial \mathbf{h}_1}{\partial \mathbf{f}_0}\frac{\partial \mathbf{f}_1}{\partial \mathbf{h}_1}\frac{\partial \mathbf{h}_2}{\partial \mathbf{f}_1}\frac{\partial \mathbf{f}_2}{\partial \mathbf{h}_2}\frac{\partial \mathbf{h}_3}{\partial \mathbf{f}_2}\frac{\partial \mathbf{f}_3}{\partial \mathbf{h}_3}\frac{\partial l_i}{\partial \mathbf{f}_3}
$$

Sizes of each term are:

$$
(D_1 \times 1) = (D_1 \times D_1)(D_1 \times D_2)(D_2 \times D_2)(D_2 \times D_3)(D_3 \times D_3)(D_3 \times D_f)(D_f \times 1)
$$

Problem 7.4 Calculate the derivative $\partial l_i / \partial f[\mathbf{x}_i, \boldsymbol{\phi}]$ for the least squares lost function:

$$
l_i = (y_i - f[\mathbf{x}_i, \boldsymbol{\phi}])^2
$$

$$
\frac{\partial l_i}{\partial f[\mathbf{x}_i, \boldsymbol{\phi}]} = 2(f[\mathbf{x}_i, \boldsymbol{\phi}] - y_i)
$$

Problem 7.5 Calculate the derivative $\partial l_i / \partial f[\mathbf{x}_i, \boldsymbol{\phi}]$ for the binary classification loss function:

$$
l_i = -(1 - y_i)\log[1 - \text{sig}[f[\mathbf{x}_i, \boldsymbol{\phi}]]] - y_i \log[\text{sig}[f[\mathbf{x}_i, \boldsymbol{\phi}]]]
$$

where the function $\text{sig}[\bullet]$ is the logistic sigmoid and is defined as:

$$
\text{sig}[z] = \frac{1}{1 + \exp[-z]}
$$

Note that the derivative of the sigmoid with respect to its input is:

$$\frac{\partial \text{sig}[z]}{\partial z} = \frac{\exp[-z]}{(1+\exp[-z])^2} = \text{sig}[z](1-\text{sig}[z])$$

Therefore

$$\begin{aligned}\frac{\partial l_i}{\partial f[\mathbf{x}_i,\boldsymbol{\phi}]} &= -(1-y_i)\frac{-\text{sig}[f[\mathbf{x}_i,\boldsymbol{\phi}]](1-\text{sig}[f[\mathbf{x}_i,\boldsymbol{\phi}]])}{1-\text{sig}[f[\mathbf{x}_i,\boldsymbol{\phi}]]} - y_i\frac{\text{sig}[f[\mathbf{x}_i,\boldsymbol{\phi}]](1-\text{sig}[f[\mathbf{x}_i,\boldsymbol{\phi}]])}{\text{sig}[f[\mathbf{x}_i,\boldsymbol{\phi}]]}\\ &= (1-y_i)\text{sig}[f[\mathbf{x}_i,\boldsymbol{\phi}]] - y_i(1-\text{sig}[f[\mathbf{x}_i,\boldsymbol{\phi}]])\\ &= \text{sig}[f[\mathbf{x}_i,\boldsymbol{\phi}]] - y_i\end{aligned}$$

Problem 7.6 Show that for $\mathbf{z} = \boldsymbol{\beta} + \boldsymbol{\Omega}\mathbf{h}$:

$$\frac{\partial \mathbf{z}}{\partial \mathbf{h}} = \boldsymbol{\Omega}^T$$

where $\partial\mathbf{z}/\partial\mathbf{h}$ is a matrix containing the term $\partial z_i/\partial h_j$ in its i-th column and j-th row. To do this, first find an expression for the constituent elements $\partial z_i/\partial h_j$, and then consider the form that the matrix $\partial\mathbf{z}/\partial\mathbf{h}$ must take.

Since $z_i = \sum_j \Omega_{ij}h_j$, for every i, j we can see that $\partial z_i/\partial h_j = \Omega_{ij}$. Constructing $\partial\mathbf{z}/\partial\mathbf{h}$ with $\partial z_i/\partial h_j$ at its i-th column and j-th row would produce Ω^T.

Problem 7.7 Consider the case where we use the logistic sigmoid (see equation 7.37) as an activation function, so $h = \text{sig}[f]$. Compute the derivative $\partial h/\partial f$ for this activation function. What happens to the derivative when the input takes (i) a large positive value and (ii) a large negative value?

The derivative of the sigmoid activation function is

$$\frac{\partial h}{\partial f} = h(1-h)$$

Thus for both large positive and large negative values, $h \to 1$ or $h \to 0$ so the derivative is close to zero. This is referred as gradient vanishing for sigmoid activation and causes problems in training process. Because of this, often ReLU is a more preferred choice for nonlinear activation.

Problem 7.8 Consider using (i) the Heaviside function and (ii) the rectangular function as activation functions:

$$\mathrm{Heaviside}[z] = \begin{cases} 0 & z < 0 \\ 1 & z \geq 0 \end{cases}$$

and

$$\mathrm{rect}[z] = \begin{cases} 0 & z < 0 \\ 1 & 0 \leq z \leq 1 \\ 0 & z > 1 \end{cases}$$

Both functions have zero derivative in almost everywhere, so we cannot update the model parameters based on the gradient descent method.

Problem 7.9 Consider a loss function $l[\mathbf{f}]$, where $\mathbf{f} = \boldsymbol{\beta} + \boldsymbol{\Omega}\mathbf{h}$. We want to find how the loss l changes when we change $\boldsymbol{\Omega}$, which we'll express with a matrix that contains the derivative $\partial l/\partial \Omega_{ij}$ at the i-th row and j-th column. Find an expression for $\partial f_i/\partial \Omega_{ij}$ and, using the chain rule, show that:

$$\frac{\partial l}{\partial \boldsymbol{\Omega}} = \frac{\partial l}{\partial \mathbf{f}}\mathbf{h}^T$$

Since $f_i = \sum_j \Omega_{ij} h_j$, $\partial f_i/\partial \Omega_{ij} = h_j$.

Using the chain rule, we can see that

$$\frac{\partial l}{\partial \Omega_{ij}} = \frac{\partial l}{\partial f_i}\frac{\partial f_i}{\partial \Omega_{ij}} = \frac{\partial l}{\partial f_i} h_j$$

which returns to the expression $\partial l/\partial \boldsymbol{\Omega} = (\partial l/\partial \mathbf{f})\mathbf{h}^T$ in its matrix representation.

Problem 7.10 Derive the equations for the backward pass of the backpropagation algorithm for a network that uses leaky ReLU activations, which are defined as:

$$a[z] = \begin{cases} \alpha \cdot z & z < 0 \\ z & z \geq 0 \end{cases}$$

where α is a small positive constant (typically 0.1).

The backward update can be computed with

$$\frac{\partial l_i}{\partial \mathbf{f}_{k-1}} = \mathbb{I}[\mathbf{f}_{k-1} > 0] \odot (\boldsymbol{\Omega}_k^T \frac{\partial l_i}{\partial \mathbf{f}_k}) + \mathbb{I}[\mathbf{f}_{k-1} < 0] \odot \alpha(\boldsymbol{\Omega}_k^T \frac{\partial l_i}{\partial \mathbf{f}_k})$$

Problem 7.11 Consider training a network with fifty layers, where we only have enough memory to store the pre-activations at every tenth hidden layer during the forward pass. Explain how to compute the derivatives in this situation using gradient checkpointing.

We can reconstruct the values of the pre-activations required for backpropagation by computing the forward pass again from the layers that are checkpointed.

Problem 7.12 This problem explores computing derivatives on general acyclic computational graphs. Consider the function:

$$y = \exp[\exp[x] + \exp[x]^2] + \sin[\exp[x] + \exp[x]^2]$$

We can break this down into a series of intermediate computations so that

$$\begin{aligned} f_1 &= \exp[x] \\ f_2 &= f_1^2 \\ f_3 &= f_1 + f_2 \\ f_4 &= \exp[f_3] \\ f_5 &= \sin[f_3] \\ y &= f_4 + f_5 \end{aligned}$$

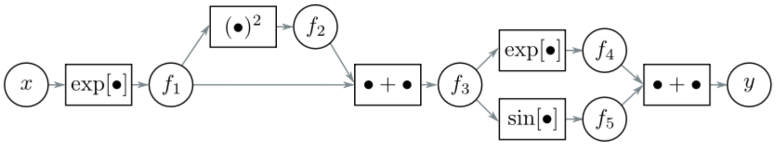


Figure 7.9 Computational graph for problem 7.12 and problem 7.13. Adapted from Domke (2010).

The associated computational graph is depicted in figure 7.9. Compute the derivative $\partial y/\partial x$ by reverse mode differentiation. In other words, compute in order:

$$\frac{\partial y}{\partial f_5}, \frac{\partial y}{\partial f_4}, \frac{\partial y}{\partial f_3}, \frac{\partial y}{\partial f_2}, \frac{\partial y}{\partial f_1}, \text{and} \frac{\partial y}{\partial x}$$

using the chain rule in each case to make use of the derivatives already computed.

$$
\begin{aligned}
\frac{\partial y}{\partial f_5} &= 1 \\
\frac{\partial y}{\partial f_4} &= 1 \\
\frac{\partial y}{\partial f_3} &= \frac{\partial y}{\partial f_5}\frac{\partial f_5}{\partial f_3} + \frac{\partial y}{\partial f_4}\frac{\partial f_4}{\partial f_3} = \cos[f_3] + \exp[f_3] \\
\frac{\partial y}{\partial f_2} &= \frac{\partial y}{\partial f_3}\frac{\partial f_3}{\partial f_2} = \cos[f_3] + \exp[f_3] \\
\frac{\partial y}{\partial f_1} &= \frac{\partial y}{\partial f_3}\frac{\partial f_3}{\partial f_1} = (\cos[f_3] + \exp[f_3])(1 + 2f_1) \\
\frac{\partial y}{\partial x} &= \frac{\partial y}{\partial f_1}\frac{\partial f_1}{\partial x} = (\cos[f_3] + \exp[f_3])(1 + 2f_1)(\exp[x])
\end{aligned}
$$

Problem 7.13 For the same function as in problem 7.12, compute the derivative $\partial y/\partial x$ by forward-mode differentiation. In other words, compute in order:

$$
\frac{\partial f_1}{\partial x}, \frac{\partial f_2}{\partial x}, \frac{\partial f_3}{\partial x}, \frac{\partial f_4}{\partial x}, \frac{\partial f_5}{\partial x}, \text{and} \frac{\partial y}{\partial x}
$$

using the chain rule in each case to make use of the derivatives already computed. Why do we not use forward-mode differentiation when we calculate the parameter gradient for deep networks?

$$
\begin{aligned}
\frac{\partial f_1}{\partial x} &= \exp[x] \\
\frac{\partial f_2}{\partial x} &= \frac{\partial f_2}{\partial f_1}\frac{\partial f_1}{\partial x} = 2f_1 \exp[x] \\
\frac{\partial f_3}{\partial x} &= \frac{\partial f_3}{\partial f_1}\frac{\partial f_1}{\partial x} + \frac{\partial f_3}{\partial f_2}\frac{\partial f_2}{\partial x} = \exp[x](1 + 2f_1) \\
\frac{\partial f_4}{\partial x} &= \frac{\partial f_4}{\partial f_3}\frac{\partial f_3}{\partial x} = \exp[f_3]\exp[x](1 + 2f_1) \\
\frac{\partial f_5}{\partial x} &= \frac{\partial f_5}{\partial f_3}\frac{\partial f_3}{\partial x} = \cos[f_3]\exp[x](1 + 2f_1) \\
\frac{\partial y}{\partial x} &= \frac{\partial y}{\partial f_5}\frac{\partial f_5}{\partial x} + \frac{\partial y}{\partial f_4}\frac{\partial f_4}{\partial x} = (\exp[f_3] + \cos[f_3])\exp[x](1 + 2f_1)
\end{aligned}
$$

Forward mode differentiation is not what we need for training deep networks because gradient descent is about computing the gradient of the loss with respect to the parameters, not computing the gradient of the hidden units with respect to the input.

Problem 7.14 Consider a random variable a with variance $\mathrm{Var}[a] = \sigma^2$ and a symmetrical distribution around the mean $\mathbb{E}[a] = 0$. Prove that if we pass this variable through the ReLU function:

$$b = \mathrm{ReLU}[a] = \begin{cases} 0 & a < 0 \\ a & a \geq 0 \end{cases}$$

then the second moment of the transformed variable is $\mathbb{E}[b^2] = \sigma^2/2$.

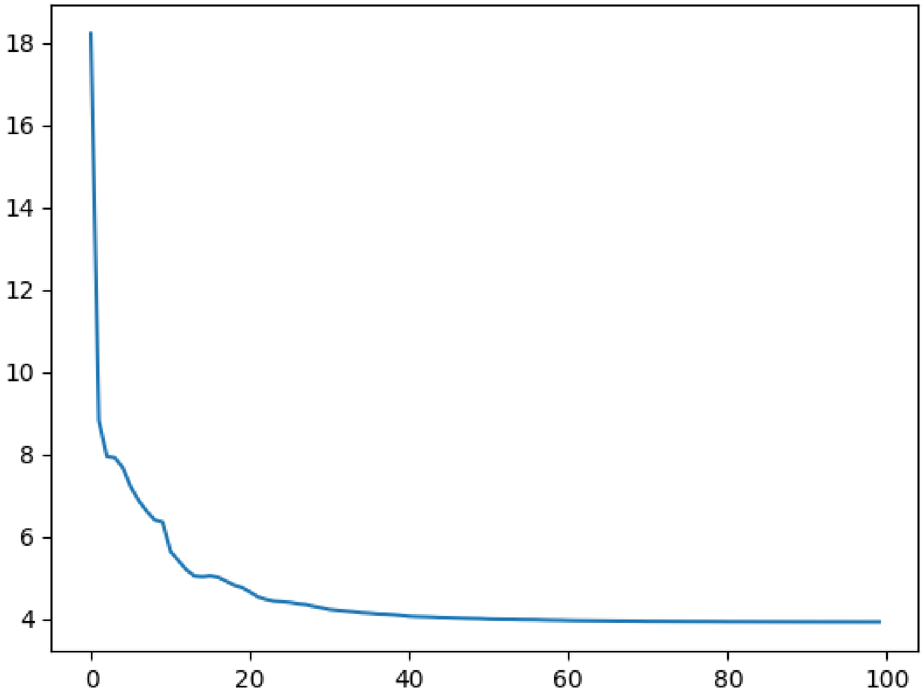
First we compute the second moment of the original variable, $\mathbb{E}[a^2] = \mathrm{Var}[a] + \mathbb{E}[a]^2 = \sigma^2$. Since a is derived from a symmetrical distribution around zero, we can say that

$$\mathbb{E}[b^2] = \frac{1}{2}\mathbb{E}[b^2|a \geq 0] + \frac{1}{2}\mathbb{E}[b^2|a < 0] = \frac{1}{2}\mathbb{E}[a^2] = \sigma^2/2$$

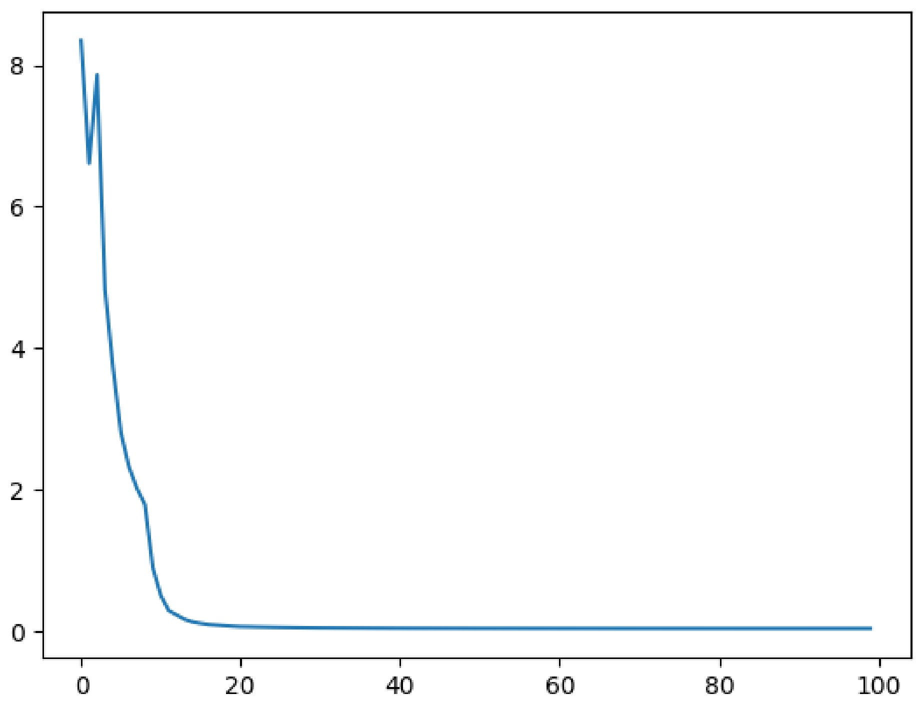
Problem 7.15 What would you expect to happen if we initialized all the weights and biases in the network to zero?

All the values of the hidden units and output will be zero, and the gradient will be also zero (though it depends how the gradient of ReLU at the zero is defined in the implementation). Thus, the model parameters cannot be updated with gradient descent algorithm.

Problem 7.16 Implement the code in figure 7.8 in PyTorch and plot the training loss as a function of the number of epochs.



Problem 7.17 Change the code in figure 7.8 to tackle a binary classification problem. You will need to (i) change the targets y so they are binary, (ii) change the network to predict numbers between zero and one (iii) change the loss function appropriately.


8
6
4
2
0
0
20
40
60
80
100