1 Backpropagation Practice

(a) Chain rule of multiple variables: Assume that you have a function given by \( f(x_1, x_2, \ldots, x_n) \), and that \( g_i(w) = x_i \) for a scalar variable \( w \). What is its computation graph? Sketch out a diagram of what the computation graph would look like. How would you compute \( \frac{d}{dw}f(g_1(w), g_2(w), \ldots, g_n(w)) \)?

(b) Let \( w_1, w_2, \ldots, w_n \in \mathbb{R}^d \), and we refer to these weights together as \( W \in \mathbb{R}^{n \times d} \). We also have \( x \in \mathbb{R}^d \) and \( y \in \mathbb{R} \). Consider the function

\[
f(W, x, y) = \left( y - \sum_{i=1}^n \phi(w_i^T x + b_i) \right)^2.
\]

Write out the function computation graph (also sometimes referred to as a pictorial representation of the network). This is a directed graph of decomposed function computations, with the output of the function at one end, and the input to the function, \( x \) at the other end, where \( b \) are the bias terms corresponding to each weight vector, i.e. \( b = [b_1, \cdots, b_n] \).

(c) Suppose \( \phi(x) \) (from the previous part) is the sigmoid function, \( \sigma(x) \). Compute the partial derivatives \( \frac{\partial f}{\partial w_i} \) and \( \frac{\partial f}{\partial b_i} \). Use the computational graph you drew in the previous part to guide you.

(d) Write down a single gradient descent update for \( w_i^{(t+1)} \) and \( b_i^{(t+1)} \), assuming step size \( \eta \). Your answer should be in terms of \( w_i^{(t)}, b_i^{(t)}, x, \) and \( y \).

(e) Define the cost function

\[
\ell(x) = \frac{1}{2} \| W^{(2)} \Phi \left( W^{(1)} x + b \right) - y \|_2^2,
\]

where \( W^{(1)} \in \mathbb{R}^{d \times d}, W^{(2)} \in \mathbb{R}^{d \times d} \), and \( \Phi : \mathbb{R}^d \to \mathbb{R}^d \) is some nonlinear transformation. Compute the partial derivatives \( \frac{\partial \ell}{\partial x}, \frac{\partial \ell}{\partial W^{(1)}}, \frac{\partial \ell}{\partial W^{(2)}}, \) and \( \frac{\partial \ell}{\partial b} \).

(f) Suppose \( \Phi \) is the identity map. Write down a single gradient descent update for \( W_i^{(1)} \) and \( W_i^{(2)} \), assuming step size \( \eta \). Your answer should be in terms of \( W_i^{(1)}, W_i^{(2)}, b_i \) and \( x, y \).