1 Derivatives of simple functions

Compute the derivatives of the following simple functions used as non-linearities in neural networks.

1. $\sigma(x) = \frac{1}{1+e^{-x}}$
2. $\text{ReLu}(x) = \max(x, 0)$
3. $\tanh(x) = \frac{\sinh(x)}{\cosh(x)} = \frac{e^x - e^{-x}}{e^x + e^{-x}}$

2 Backpropagation Practice (self-study)

1. 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. How would you compute $\frac{\mathrm{d}}{\mathrm{d}w} f(g_1(w), g_2(w), \ldots, g_n(w))$? What is its computation graph?(also sometimes referred to as a pictorial representation of the network). This is a directed graph of decomposed function computations, with the function at one end, and the variables W, b, x, y at the other end, where $b = [b_1, \cdots, b_n]$

2. Let $w_1, w_2, \ldots, w_n \in \mathbb{R}^d$, and we refer to these variables 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

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

   Write out the function computation graph.

3. Suppose $\phi(x) = \sigma(x)$ (from problem 1a). 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.

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

5. (optional) Define the cost function

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

   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 $\frac{\partial \ell}{\partial b}$.

6. (optional) Suppose Φ is the identity map. Write down a single gradient descent update for $W_{t+1}^{(1)}$ and $W_{t+1}^{(2)}$ assuming step size η. Your answer should be in terms of $W_t^{(1)}$, $W_t^{(2)}$, b_t and x, y.

