CS 189 Fall 2018 Introduction to Machine Learning DIS8

1 Derivatives of simple functions

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

(a) $\sigma(x) = \frac{1}{1+e^{-x}}$

Solution: Taking the derivative via chain rule, we have

$$\sigma'(x) = -\frac{1}{(1+e^{-x})^2}(-e^{-x}) = \frac{1}{1+e^{-x}}\left(1 - \frac{1}{1+e^{-x}}\right) = \sigma(x)(1-\sigma(x)).$$

(b) $\mathsf{ReLu}(x) = \max(x, 0)$

Solution: The derivative here is equal to 1 if $x > 0$, and 0 if $x < 0$. At 0, the function is not differentiable, so we must pick a "subgradient", which is some tangent to the function. It is typical to pick either 0 or 1.

(c) $\tanh(x) = \frac{\sinh(x)}{\cosh(x)} = \frac{e^x - e^{-x}}{e^x + e^{-x}}$

Solution: Notice that $\tanh(x) = \frac{1-e^{-2x}}{1+e^{-2x}} = \sigma(2x) - (1 - \sigma(2x)) = 2\sigma(2x) - 1.$
Hence, by chain rule, it is clear that the derivative is just $4\sigma'(2x) = 4\sigma(2x)(1-\sigma(2x))$.

2 Backpropagation

In this discussion, we will explore the chain rule of differentiation, and provide some algorithmic motivation for the backpropagation algorithm.

(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. 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?

Solution: This is the chain rule for multiple variables. In general, we have

$$\frac{\mathrm{d}f}{\mathrm{d}w} = \sum_{i=1}^{n} \frac{\partial f}{\partial x_i}\frac{\partial x_i}{\partial w} = \frac{\partial f}{\partial \mathbf{x}}\frac{\partial \mathbf{x}}{\partial w}.$$

The function graph of this computation is given in Figure 1.

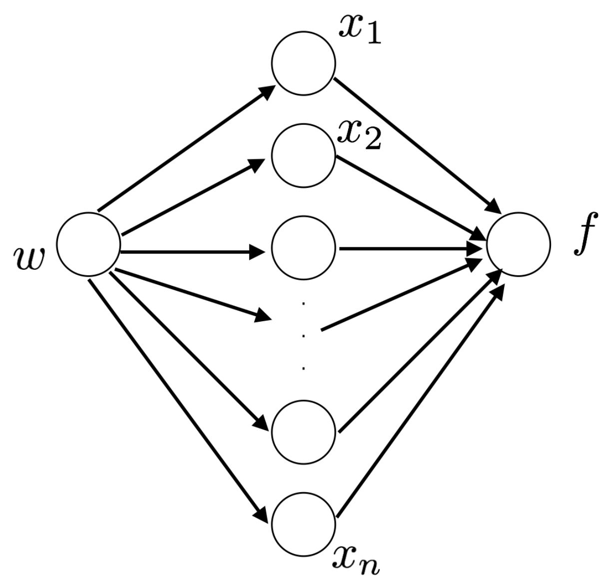

Figure 1: Example function computation graph

(b) Let $\mathbf{w}_1, \mathbf{w}_2, \ldots, \mathbf{w}_n \in \mathbb{R}^d$, and we refer to these variables together as $\mathbf{W} \in \mathbb{R}^{n \times d}$. We also have $\mathbf{x} \in \mathbb{R}^d$ and $y \in \mathbb{R}$. Consider the function

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

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 function at one end, and the variables $\mathbf{W}, \mathbf{b}, \mathbf{x}, y$ at the other end, where $\mathbf{b} = [b_1, \cdots, b_n]$.

Solution:

See Figure 2.

(c) Suppose $\boldsymbol{\phi}(x) = \sigma(x)$ (from problem 1a). Compute the partial derivatives $\frac{\partial f}{\partial \mathbf{w}_i}$ and $\frac{\partial f}{\partial b_i}$. Use the computational graph you drew in the previous part to guide you.

Solution: Denote $r = y - \sum_{i=1}^{n} \sigma(\mathbf{w}_i^\top \mathbf{x} + b_i)$ and $z_i = \mathbf{w}_i^\top \mathbf{x} + b_i$.

To remind ourselves, this is the ‘forward’ computation:

$$\begin{aligned} f &= r^2 \\ r &= y - \sum_{i=1}^{n} \sigma(z_i) \\ z_i &= \mathbf{w}_i^\top \mathbf{x} + b_i \end{aligned}$$

Now the backward pass:

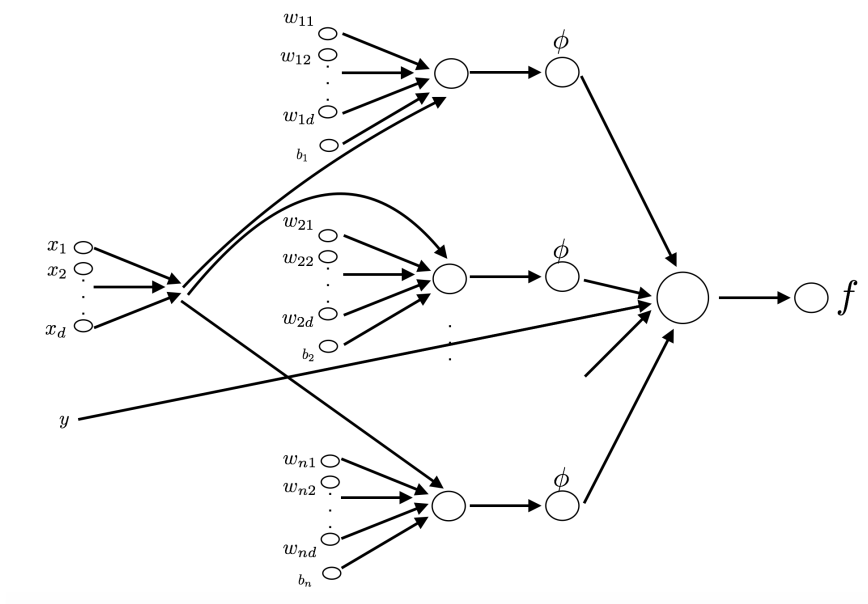

Figure 2: Example function computation graph

$$\frac{\partial f}{\partial r} = 2r$$
$$\frac{\partial r}{\partial z_i} = -\sigma(z_i)(1-\sigma(z_i))$$
$$\frac{\partial z_i}{\partial \mathbf{w}_i} = \mathbf{x}^\top$$
$$\frac{\partial z_i}{\partial b_i} = 1$$

By applying chain rule

$$\frac{\partial f}{\partial \mathbf{w}_i} = 2(\sum_{i=1}^{n}\sigma(\mathbf{w}_i^\top\mathbf{x}+b_i)-y)\sigma(\mathbf{w}_i^\top\mathbf{x}+b_i)(1-\sigma(\mathbf{w}_i^\top\mathbf{x}+b_i))\mathbf{x}^\top$$
$$\frac{\partial f}{\partial b_i} = 2(\sum_{i=1}^{n}\sigma(\mathbf{w}_i^\top\mathbf{x}+b_i)-y)\sigma(\mathbf{w}_i^\top\mathbf{x}+b_i)(1-\sigma(\mathbf{w}_i^\top\mathbf{x}+b_i))$$

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

Solution:

$$\mathbf{w}_i^{(t+1)} \leftarrow \mathbf{w}_i^{(t)} - 2\eta(\sum_{i=1}^{n}\sigma(\mathbf{w}_i^{(t)}i^\top\mathbf{x}+b_i^{(t)})-y)\sigma({\mathbf{w}_i^{(t)}}^\top\mathbf{x}+b_i^{(t)})(1-\sigma({\mathbf{w}_i^{(t)}}^\top\mathbf{x}+b_i^{(t)}))\mathbf{x}$$

$$b_i^{(t+1)} \leftarrow b_i^{(t)} - 2\eta(\sum_{i=1}^{n} \sigma(\mathbf{w}_i^{(t)}i^\top \mathbf{x} + b_i^{(t)}) - y)\sigma({\mathbf{w}_i^{(t)}}^\top \mathbf{x} + b_i^{(t)})(1 - \sigma({\mathbf{w}_i^{(t)}}^\top \mathbf{x} + b_i^{(t)}))$$

(e) (optional) Define the cost function

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

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

Solution: First, we write out the intermediate variable for our convenience.

$$\begin{aligned}
\mathbf{x}^{(1)} &= \mathbf{W}^{(1)}\mathbf{x} + \mathbf{b} \\
\mathbf{x}^{(2)} &= \mathbf{\Phi}(\mathbf{x}_1) \\
\mathbf{x}^{(3)} &= \mathbf{W}^{(2)}\mathbf{x}^{(2)} \\
\mathbf{x}^{(4)} &= \mathbf{x}^{(3)} - \mathbf{y} \\
\ell &= \frac{1}{2}\|\mathbf{x}^{(4)}\|_2^2.
\end{aligned}$$

Remember that the superscripts represents the index rather than the power operators. We have

$$\begin{aligned}
\frac{\partial \ell}{\partial \mathbf{x}^{(4)}} &= \mathbf{x}^{(4)\top} \\
\frac{\partial \ell}{\partial \mathbf{x}^{(3)}} &= \frac{\partial \ell}{\partial \mathbf{x}^{(4)}}\frac{\partial \mathbf{x}^{(4)}}{\partial \mathbf{x}^{(3)}} = \frac{\partial \ell}{\partial \mathbf{x}^{(4)}} \\
\frac{\partial \ell}{\partial \mathbf{x}^{(2)}} &= \frac{\partial \ell}{\partial \mathbf{x}^{(3)}}\frac{\partial \mathbf{x}^{(3)}}{\partial \mathbf{x}^{(2)}} = \frac{\partial \ell}{\partial \mathbf{x}^{(3)}}\mathbf{W}^{(2)} \\
\frac{\partial \ell}{\partial \mathbf{W}^{(2)}} &= \frac{\partial \ell}{\partial \mathbf{x}^{(3)}}\frac{\partial \mathbf{x}^{(3)}}{\partial \mathbf{W}^{(2)}} = \mathbf{x}^{(2)}\frac{\partial \ell}{\partial \mathbf{x}^{(3)}} \\
\frac{\partial \ell}{\partial \mathbf{x}^{(1)}} &= \frac{\partial \ell}{\partial \mathbf{x}^{(2)}}\frac{\partial \mathbf{\Phi}}{\partial \mathbf{x}^{(1)}} \\
\frac{\partial \ell}{\partial \mathbf{x}} &= \frac{\partial \ell}{\partial \mathbf{x}^{(1)}}\frac{\partial \mathbf{x}^{(1)}}{\partial \mathbf{x}} = \frac{\partial \ell}{\partial \mathbf{x}^{(1)}}\mathbf{W}^{(1)} \\
\frac{\partial \ell}{\partial \mathbf{b}} &= \frac{\partial \ell}{\partial \mathbf{x}^{(1)}}\frac{\partial \mathbf{x}^{(1)}}{\partial \mathbf{b}} = \frac{\partial \ell}{\partial \mathbf{x}^{(1)}} \\
\frac{\partial \ell}{\partial \mathbf{W}^{(1)}} &= \frac{\partial \ell}{\partial \mathbf{x}^{(1)}}\frac{\partial \mathbf{x}^{(1)}}{\partial \mathbf{W}^{(1)}} = \mathbf{x}\frac{\partial \ell}{\partial \mathbf{x}^{(1)}}.
\end{aligned}$$

A more formal way to solve these requires doing the expansion. For example, $\frac{\partial \ell}{\partial \mathbf{W}^{(1)}} = \frac{\partial \ell}{\partial \mathbf{x}^{(1)}}\frac{\partial \mathbf{x}^{(1)}}{\partial \mathbf{W}^{(1)}}$. However, the right hand side is a tensor, in which the matrix codebook does not provide us a useful information. We need to do that manually. Notice that $x_k^{(1)} = \sum_l W_{kl}^{(1)} x_l + b_k$, we have

$$\frac{\partial \ell}{\partial W_{ij}^{(1)}} = \sum_k \frac{\partial \ell}{\partial x_k^{(1)}}\frac{\partial x_k^{(1)}}{\partial W_{ij}^{(1)}}$$

$$= \sum_k \sum_l \frac{\partial \ell}{\partial x_k^{(1)}} \left(\epsilon_{ik}\epsilon_{jl} x_l\right)$$
$$= \frac{\partial \ell}{\partial x_i^{(1)}} x_j$$

so that

$$\frac{\partial \ell}{\partial \mathbf{W}^{(1)}} = \frac{\partial \ell}{\partial \mathbf{x}^{(1)}} \frac{\partial \mathbf{x}^{(1)}}{\partial \mathbf{W}^{(1)}} = \mathbf{x} \frac{\partial \ell}{\partial \mathbf{x}^{(1)}}. \tag{2}$$

(f) (optional) Suppose $\mathbf{\Phi}$ is the identity map. Write down a single gradient descent update for $\mathbf{W}_{t+1}^{(1)}$ and $\mathbf{W}_{t+1}^{(2)}$ assuming step size η. Your answer should be in terms of $\mathbf{W}_t^{(1)}$, $\mathbf{W}_t^{(2)}$, $\mathbf{b}_t$ and $\mathbf{x}, \mathbf{y}$.

Solution:

$$\mathbf{W}_{t+1}^{(1)} \leftarrow \mathbf{W}_t^{(1)} - \eta (\mathbf{W}_t^{(2)})^\top \left(\mathbf{W}_t^{(2)} \left(\mathbf{W}_t^{(1)} \mathbf{x} + \mathbf{b}_t \right) - \mathbf{y} \right) \mathbf{x}^\top$$

$$\mathbf{W}_{t+1}^{(2)} \leftarrow \mathbf{W}_t^{(2)} - \eta \left(\mathbf{W}_t^{(2)} \left(\mathbf{W}_t^{(1)} \mathbf{x} + \mathbf{b}_t \right) - \mathbf{y} \right) (\mathbf{W}_t^{(1)} \mathbf{x} + \mathbf{b})^\top$$

Side note: The computation complexity of computing the $\frac{\partial \ell}{\partial \mathbf{W}}$ for Equation (1) using the analytic derivatives and numerical derivatives is quite different!

For numerical differentiation, what we do is to use the following first order formula

$$\frac{\partial \ell}{\partial W_{ij}} = \frac{\ell\left(W_{ij} + \epsilon, \cdot\right) - \ell\left(W_{ij}, \cdot\right)}{\epsilon}.$$

We need $O(d^4)$ operations in order to compute $\frac{\partial \ell}{\partial \mathbf{W}}$. On the other hand, it only takes $O(d^2)$ operations to compute it analytically.