CS 189/289A Introduction to Machine Learning
Fall 2022 Jennifer Listgarten, Jitendra Malik **DIS4**

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 $\frac{\mathrm{d}}{\mathrm{d}w} f(g_1(w), g_2(w), \ldots, g_n(w))$?

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

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

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

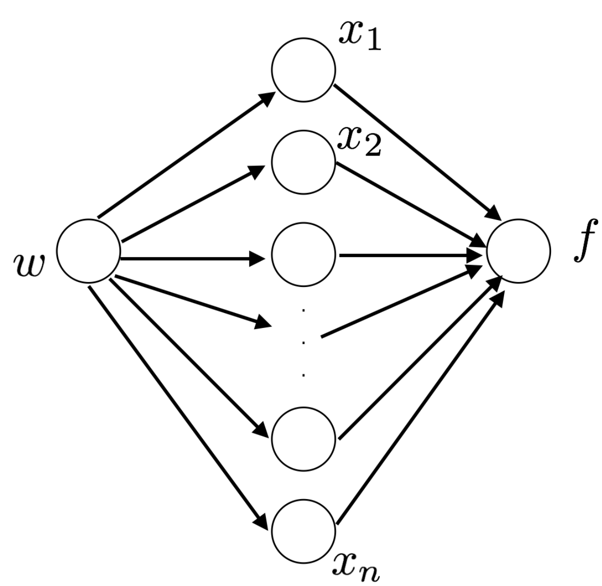

Figure 1: Example function computation graph

(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

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

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. $b = [b_1, \cdots, b_n]$.

Solution:

See Figure 2.

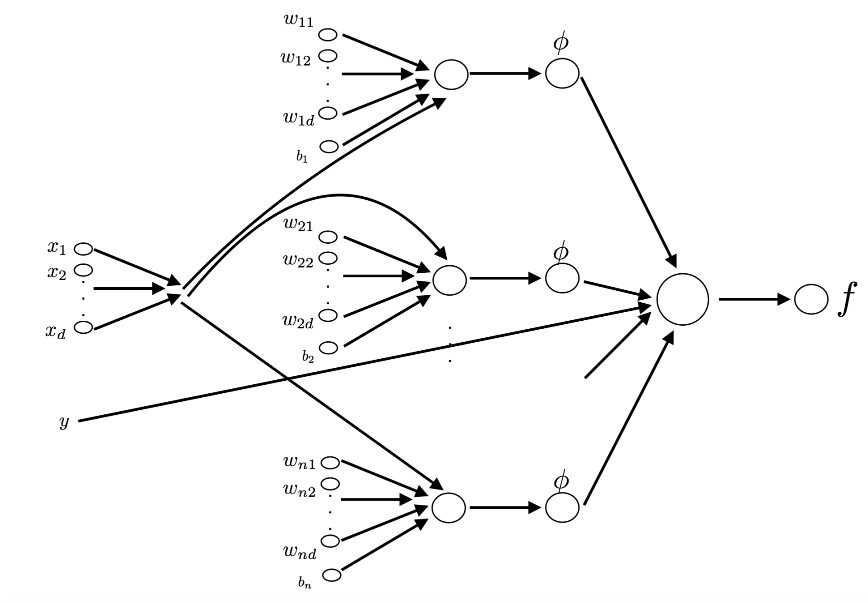

Figure 2: Example function computation graph

(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.

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

To remind ourselves, this is the 'forward' computation:

$$f = r^2$$
$$r = y - \sum_{i=1}^{n} \sigma(z_i)$$
$$z_i = w_i^\top x + b_i$$

Now the backward pass:

$$\frac{\partial f}{\partial r} = 2r$$
$$\frac{\partial r}{\partial z_i} = -\sigma(z_i)(1 - \sigma(z_i))$$

$$\frac{\partial z_i}{\partial w_i} = x^\top$$
$$\frac{\partial z_i}{\partial b_i} = 1$$

By applying chain rule

$$\frac{\partial f}{\partial w_i} = 2(\sum_{j=1}^{n} \sigma(w_j^\top x + b_j) - y)\sigma(w_i^\top x + b_i)(1 - \sigma(w_i^\top x + b_i))x^\top$$
$$\frac{\partial f}{\partial b_i} = 2(\sum_{j=1}^{n} \sigma(w_j^\top x + b_j) - y)\sigma(w_i^\top x + b_i)(1 - \sigma(w_i^\top x + b_i))$$

(d) 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.

Solution:

$$w_i^{(t+1)} \leftarrow w_i^{(t)} - 2\eta(\sum_{j=1}^{n} \sigma(w_j^{(t)\top} x + b_j^{(t)}) - y)\sigma(w_i^{(t)\top} x + b_i^{(t)})(1 - \sigma(w_i^{(t)\top} x + b_i^{(t)}))x$$
$$b_i^{(t+1)} \leftarrow b_i^{(t)} - 2\eta(\sum_{j=1}^{n} \sigma(w_j^{(t)\top} x + b_j^{(t)}) - y)\sigma(w_i^{(t)\top} x + b_i^{(t)})(1 - \sigma(w_i^{(t)\top} x + b_i^{(t)}))$$

(e) 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 $\frac{\partial \ell}{\partial b}$.

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

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

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

$$\frac{\partial \ell}{\partial x^{(4)}} = x^{(4)\top}$$
$$\frac{\partial \ell}{\partial x^{(3)}} = \frac{\partial \ell}{\partial x^{(4)}}\frac{\partial x^{(4)}}{\partial x^{(3)}} = \frac{\partial \ell}{\partial x^{(4)}}$$

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

The easy trick to solve the derivatives with respect to (each element of) a matrix is to "guess" the ordering of the expression so that the dimensions match up on both sides. More formally, we could express it as follows:

$$
\frac{\partial \ell}{\partial x^{(3)}} \frac{\partial x^{(3)}}{\partial W^{(2)}} = \frac{\partial \ell}{\partial x^{(3)}} x^{(2)} = \mathrm{Tr}\left(\frac{\partial \ell}{\partial x^{(3)}} (\cdot) x^{(2)} \right) = \mathrm{Tr}\left(x^{(2)} \frac{\partial \ell}{\partial x^{(3)}} (\cdot) \right) = x^{(2)} \frac{\partial \ell}{\partial x^{(3)}} \tag{2}
$$

(f) 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.

Solution:

$$
\begin{aligned}
W_{t+1}^{(1)} &\leftarrow W_t^{(1)} - \eta (W_t^{(2)})^\top \left(W_t^{(2)} \left(W_t^{(1)} x + b_t \right) - y \right) x^\top \\
W_{t+1}^{(2)} &\leftarrow W_t^{(2)} - \eta \left(W_t^{(2)} \left(W_t^{(1)} x + b_t \right) - y \right) (W_t^{(1)} x + b)^\top
\end{aligned}
$$

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

For numerical differentiation, we use the following first order formula:

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

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

