Section 08: Solutions

1. The Chain Rule

(a) Let $f\colon \mathbb{R}^n \to \mathbb{R}^m$, $g\colon \mathbb{R}^\ell \to \mathbb{R}^n$. Write the Jacobian of $f \circ g$ as a matrix in terms of the Jacobian matrix $\frac{\partial f}{\partial y}$ of f and the Jacobian matrix $\frac{\partial g}{\partial x}$ of g. Make sure the matrix dimensions line up. What conditions must hold in order for this formula to make sense?

Solution:

The Chain Rule theorem states that:

$$\frac{\partial (f \circ g)}{\partial x}(x) = \frac{\partial f}{\partial y}(g(x)) \cdot \frac{\partial g}{\partial x}(x)$$

In order for the dimensions to line up for matrix multiplication, we must have $\frac{\partial f}{\partial y} \in \mathbb{R}^{m\times n}$ and $\frac{\partial g}{\partial x} \in \mathbb{R}^{n\times \ell}$, since $f \circ g\colon \mathbb{R}^\ell \to \mathbb{R}^m$. Note that by this convention, the gradient of a vector-valued function is:

$$\frac{\partial f}{\partial y}(y) = \begin{bmatrix} \frac{\partial f_1}{\partial y_1}(y) & \cdots & \frac{\partial f_1}{\partial y_n}(y) \\ \vdots & & \vdots \\ \frac{\partial f_m}{\partial y_1}(y) & \cdots & \frac{\partial f_m}{\partial y_n}(y) \end{bmatrix}.$$

In order to apply the chain rule, f must be differentiable at $g(x)$ and g must be differentiable at x.

(b) Let $f : \mathbb{R}^n \to \mathbb{R}$ and $g : \mathbb{R} \to \mathbb{R}^n$. Write the derivative of $f \circ g$ as a summation between the partial derivatives $\frac{\partial f}{\partial y_i}$ of f and the partial derivatives $\frac{\partial g_i}{\partial x}$ of g.

Solution:

$$\frac{\partial f \circ g}{\partial x} = \sum_{i=1}^{n} \frac{\partial f}{\partial y_i}(g(x)) \cdot \frac{\partial g_i}{\partial x}(x).$$

(c) What if instead the input of g is a matrix $W \in \mathbb{R}^{p\times q}$? Can we still represent the derivative $\frac{\partial g}{\partial W}$ of g as a matrix?

Solution:

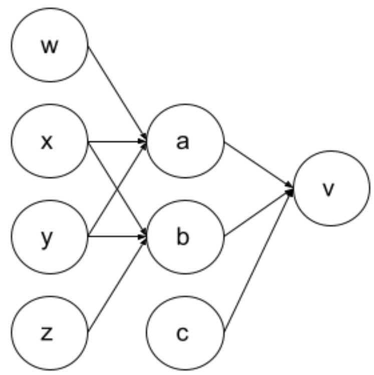
No, we cannot. The derivative of $g\colon \mathbb{R}^{p\times q} \to \mathbb{R}^n$ would be represented as a three-dimensional $n \times p \times q$ tensor. In practice, people often *flatten* the input matrix W to a vector $\text{vec}(W) \in \mathbb{R}^{pq}$. Then we can write the derivative of g as a Jacobian matrix, $\frac{\partial g}{\partial\, \text{vec}(W)} \in \mathbb{R}^{n\times pq}$. Then we must remember to un-flatten the derivative later when we update the matrix W.

2. Neural Network Chain Rule Warm-Up

Consider the following equations:

$$v(a,b,c) = c(a-b)^2$$
$$a(w,x,y) = (w+x+y)^2$$
$$b(x,y,z) = (x-y-z)^2$$

The way variables are related to each other can be represented as the network:



(a) Using the multi-variate chain rule(part 1.b), write the derivatives of the output v with respect to each of the input variables: c, w, x, y, z using only partial derivative symbols.

Solution:

$$\frac{\partial v}{\partial c} = \frac{\partial v}{\partial c}$$
$$\frac{\partial v}{\partial w} = \frac{\partial v}{\partial a} \cdot \frac{\partial a}{\partial w}$$
$$\frac{\partial v}{\partial x} = \frac{\partial v}{\partial a} \cdot \frac{\partial a}{\partial x} + \frac{\partial v}{\partial b} \cdot \frac{\partial b}{\partial x}$$
$$\frac{\partial v}{\partial y} = \frac{\partial v}{\partial a} \cdot \frac{\partial a}{\partial y} + \frac{\partial v}{\partial b} \cdot \frac{\partial b}{\partial y}$$
$$\frac{\partial v}{\partial z} = \frac{\partial v}{\partial b} \cdot \frac{\partial b}{\partial z}$$

(b) Compute the values of all the partial derivatives on the RHS of your results to the previous question. Then use them to compute the values on the LHS.

Solution:

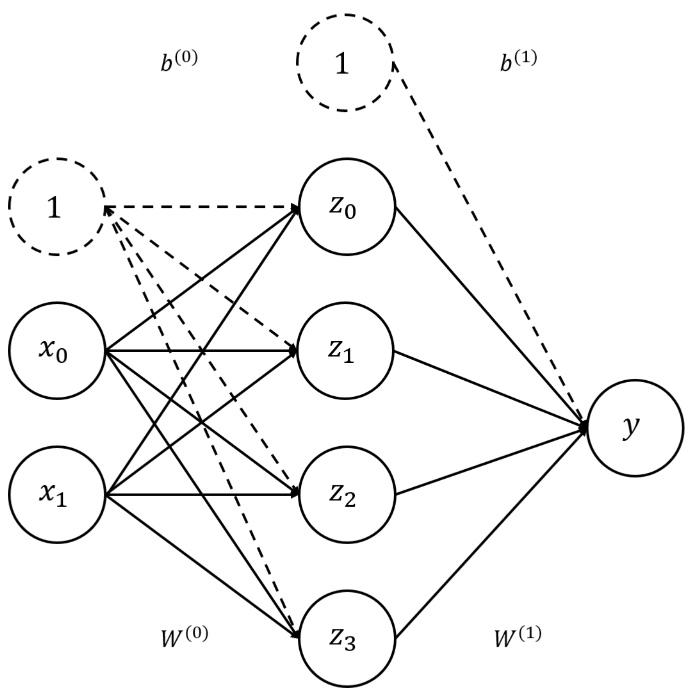
$$\frac{\partial v}{\partial a} = 2c(a-b) \quad \frac{\partial v}{\partial b} = -2c(a-b) \quad \frac{\partial v}{\partial c} = (a-b)^2$$
$$\frac{\partial a}{\partial w} = 2(w+x+y) \quad \frac{\partial a}{\partial x} = 2(w+x+y) \quad \frac{\partial a}{\partial y} = 2(w+x+y)$$
$$\frac{\partial b}{\partial x} = 2(x-y-z) \quad \frac{\partial b}{\partial y} = -2(x-y-z) \quad \frac{\partial b}{\partial z} = -2(x-y-z)$$

$$\frac{\partial v}{\partial c} = (a-b)^2$$
$$\frac{\partial v}{\partial w} = \frac{\partial v}{\partial a} \cdot \frac{\partial a}{\partial w} = 4c(a-b)(w+x+y)$$
$$\frac{\partial v}{\partial x} = \frac{\partial v}{\partial a} \cdot \frac{\partial a}{\partial x} + \frac{\partial v}{\partial b} \cdot \frac{\partial b}{\partial x} = 4c(a-b)(w+x+y) \cdot -4c(a-b)(x-y-z) = 4c(a-b)(w+2y+z)$$
$$\frac{\partial v}{\partial y} = \frac{\partial v}{\partial a} \cdot \frac{\partial a}{\partial y} + \frac{\partial v}{\partial b} \cdot \frac{\partial b}{\partial y} = 4c(a-b)(w+x+y) \cdot 4c(a-b)(x-y-z) = 4c(a-b)(w+2x-z)$$
$$\frac{\partial v}{\partial z} = \frac{\partial v}{\partial b} \cdot \frac{\partial b}{\partial z} = 4c(a-b)(x-y-z)$$

3. 1-Hidden-Layer Neural Network Gradients and Initialization

3.1. Forward and Backward pass

Consider a 1-hidden-layer neural network with a single output unit. Formally the network can be defined by the parameters $W^{(0)} \in \mathbb{R}^{h \times d}$, $b^{(0)} \in \mathbb{R}^h$; $W^{(1)} \in \mathbb{R}^{1 \times h}$ and $b^{(1)} \in \mathbb{R}$. The input is given by $x \in \mathbb{R}^d$. We will use sigmoid activation for the first hidden layer z and no activation for the output y. Below is a visualization of such a neural network with $d = 2$ and $h = 4$.



(a) Write out the forward pass for the network using $x, W^{(0)}, b^{(0)}, z, W^{(1)}, b^{(1)}, \sigma$ and y.

Hint: Write $z = \dots$ and $y = \dots$

Solution:

$$z = \sigma\left(W^{(0)}x + b^{(0)}\right)$$
$$y = W^{(1)}z + b^{(1)}$$

(b) Find the partial derivatives of the output with respect $W^{(1)}$ and $b^{(1)}$, namely $\frac{\partial y}{\partial W^{(1)}}$ and $\frac{\partial y}{\partial b^{(1)}}$.

Solution:

$$\frac{\partial y}{\partial W^{(1)}} = z$$
$$\frac{\partial y}{\partial b^{(1)}} = 1$$

(c) Now find the partial derivative of the output with respect to the output of the hidden layer z, that is $\frac{\partial y}{\partial z}$

Solution:

$$\frac{\partial y}{\partial z} = W^{(1)}$$

(d) Finally find the partial derivatives of the output with respect to $W^{(0)}$ and $b^{(0)}$, that is $\frac{\partial y}{\partial W^{(0)}}$ and $\frac{\partial y}{\partial b^{(0)}}$.

Hint: First find $\frac{\partial z_i}{\partial W_i^{(0)}}$ and $\frac{\partial z_i}{\partial b_i^{(0)}}$, where $W_i^{(0)}$ denotes the i-th row of $W^{(0)}$. Then note that $\frac{\partial y}{\partial W_i^{(0)}} = \sum_{j=1}^{h} \frac{\partial y}{\partial z_j}\frac{\partial z_j}{\partial W_i^{(0)}} = \frac{\partial y}{\partial z_i}\frac{\partial z_i}{\partial W_i^{(0)}}$ and $\frac{\partial y}{\partial b_i^{(0)}} = \sum_{j=1}^{h} \frac{\partial y}{\partial z_j}\frac{\partial z_j}{\partial b_i^{(0)}} = \frac{\partial y}{\partial z_i}\frac{\partial z_i}{\partial b_i^{(0)}}$ using the chain rule for multi-variate functions(1.b).

Solution:

$$\frac{\partial z_i}{\partial W_i^{(0)}} = z_i(1-z_i)x^\top \in \mathbb{R}^d$$
$$\frac{\partial y}{\partial W_i^{(0)}} = \frac{\partial y}{\partial z_i}\frac{\partial z_i}{\partial W_i^{(0)}} = W_i^{(1)} \cdot z_i(1-z_i)x^\top \in \mathbb{R}^d$$
$$\frac{\partial y}{\partial W^{(0)}} = \left[W^{(1)\top} \circ z \circ (1-z)\right] x^\top \in \mathbb{R}^{h\times d}\ ,$$

$$\frac{\partial z_i}{\partial b_i^{(0)}} = z_i(1-z_i) \in \mathbb{R}$$
$$\frac{\partial y}{\partial b_i^{(0)}} = \frac{\partial y}{\partial z_i}\frac{\partial z_i}{\partial b_i^{(0)}} = W_i^{(1)} \cdot z_i(1-z_i) \in \mathbb{R}$$
$$\frac{\partial y}{\partial b^{(0)}} = W^{(1)\top} \circ z \circ (1-z) \in \mathbb{R}^h.$$

We have provided the shapes of the matrix representations of derivatives. Try to reason about why it is of the given shape.

3.2. Weight initialization

Suppose we initialize all weights and biases in the network to 0 before performing gradient descent.

(a) For all $x \in \mathbb{R}^d$, find z and y after the forward pass.

Solution:

$$z_i = \sigma(W_i^{(0)} x + b^{(0)_i}) = \sigma(\mathbf{0}x + 0) = \sigma(0) = \frac{1}{2}$$

$$y = W^{(1)} z + b^{(1)} = \mathbf{0} \cdot \frac{\mathbf{1}}{\mathbf{2}} + 0 = 0$$

(b) Now find the values of the gradients $\frac{\partial y}{\partial W^{(1)}}$, $\frac{\partial y}{\partial b^{(1)}}$, $\frac{\partial y}{\partial W^{(0)}}$ and $\frac{\partial y}{\partial b^{(0)}}$. Note that some of the gradients will be in terms of x.

Solution:

$$\begin{aligned}
\frac{\partial y}{\partial W^{(1)}} &= z = \frac{\mathbf{1}}{\mathbf{2}} \\
\frac{\partial y}{\partial b^{(1)}} &= 1 \\
\frac{\partial y}{\partial W^{(0)}} &= \left[W^{(1)} \circ z \circ (1 - z)\right] x^\top \\
&= (\mathbf{0} \circ \frac{\mathbf{1}}{\mathbf{2}} \circ \frac{\mathbf{1}}{\mathbf{2}}) x^\top = \mathbf{0} \\
\frac{\partial y}{\partial b^{(0)}} &= W^{(1)} \circ z \circ (1 - z) \\
&= \mathbf{0} \circ \frac{\mathbf{1}}{\mathbf{2}} \circ \frac{\mathbf{1}}{\mathbf{2}} = \mathbf{0}\,.
\end{aligned}$$

(c) Observe the values of each z_i and observe each $\frac{\partial y}{\partial W_i^{(l)}}$ and $\frac{\partial y}{\partial b_i^{(l)}}$. What do you notice? And what does this imply for the expressiveness of the network? (Note that there is nothing special about the value 0 here, it just simplifies the calculations. The same can be shown for initialization with any constant c)

Solution:

The key insight is that if we initialize the weights to all have the same value, all z_i are the same. Similarly all $W_i^{(l)}$ and $b_i^{(l)}$ are the same too and so the output y could be expressed with just a single z_i instead of h. Thus the neural network boils down to just having a single hidden unit. The same holds for the gradients, so during a step of gradient descent, $W_i^{(l)}$ and $b_i^{(l)}$ are updated in the same way. Thus after a step of gradient descent, all $W_i^{(l)}$ and $b_i^{(l)}$ are still the same. By induction, the same holds after an arbitrary number of steps of gradient descent.