1 Definitions

(a) **Convexity**: A function $f : \mathbb{R}^M \to \mathbb{R}$ is convex if and only if $\forall \alpha \in [0, 1]$, i.e., $0 \leq \alpha \leq 1$,

$$f(\alpha \mathbf{x}_1 + (1 - \alpha)\mathbf{x}_2) \leq \alpha f(\mathbf{x}_1) + (1 - \alpha) f(\mathbf{x}_2)$$

You can think of a convex function as a curve or surface that opens upward like a smiley face. It could also be completely flat or flat in some places, but it definitely can't curve/bend downward.

(b) **Multivariate Chain Rule:** Let $f : \mathbb{R}^N \to \mathbb{R}$. Let $g_i : \mathbb{R} \to \mathbb{R}$ for all $i \in \{1, 2, \ldots, N\}$. Let $x \in \mathbb{R}$, $z_i = g_i(x)$, and $y = f(g_1(x), g_2(x), \ldots, g_N(x))$. Then, the multivariate chain rule states that:

$$\frac{d}{dx} f(g_1(x), g_2(x), \ldots, g_N(x)) = \sum_{i=1}^{N} \frac{dy}{dz_i} \frac{dz_i}{dx}$$

If instead, we combined all of the g_i functions into one function $\mathbf{g} : \mathbb{R} \to \mathbb{R}^N$, $\mathbf{z} = \mathbf{g}(x)$, we effectively have the same thing but now using vectors (and partial derivatives). Given $y = f(\mathbf{g}(x))$

$$\frac{\partial}{\partial x} f(\mathbf{g}(x)) = \sum_{i=1}^{N} \frac{\partial y}{\partial z_i} \frac{\partial z_i}{\partial x} = \frac{\partial y}{\partial \mathbf{z}} \frac{\partial \mathbf{z}}{\partial x}$$

Note that the above chain rule is written according to numerator layout. If specified to use denominator layout, the order of the derivatives is opposite that of numerator layout.

	Numerator layout	Denominator layout	Notes
$\frac{d}{dt}\ f\left(g(t), h(t)\right)$	$\frac{df}{dg}\frac{dg}{dt} + \frac{df}{dh}\frac{dh}{dt}$	Same	$f : (\mathbb{R} \times \mathbb{R}) \to \mathbb{R}$, $t \in \mathbb{R}$ $g : \mathbb{R} \to \mathbb{R}$, $h : \mathbb{R} \to \mathbb{R}$
$\frac{d}{dt}\ f\left(g_1(t), \ldots, g_N(t)\right)$	$\sum_{i=1}^{N} \frac{df}{dg_i}\frac{dg_i}{dt}$	Same	$f : (\mathbb{R} \times \cdots \times \mathbb{R}) \to \mathbb{R}$, $t \in \mathbb{R}$ $f_i : \mathbb{R} \to \mathbb{R}\ \ \forall i \in \{1, \ldots, N\}$
$\frac{d}{dt}\ f(\mathbf{g}(t))$	$\frac{\partial f}{\partial \mathbf{g}}\frac{\partial \mathbf{g}}{\partial t}$	$\frac{\partial \mathbf{g}}{\partial t}\frac{\partial f}{\partial \mathbf{g}}$	$f : \mathbb{R}^N \to \mathbb{R}$, $\mathbf{g} : \mathbb{R} \to \mathbb{R}^N$ $t \in \mathbb{R}$
$\frac{\partial}{\partial \mathbf{v}}\ f(\mathbf{g}(\mathbf{v}))$	$\frac{\partial f}{\partial \mathbf{g}}\frac{\partial \mathbf{g}}{\partial \mathbf{v}}$	$\frac{\partial \mathbf{g}}{\partial \mathbf{v}}\frac{\partial f}{\partial \mathbf{g}}$	$f : \mathbb{R}^N \to \mathbb{R}$, $\mathbf{g} : \mathbb{R}^M \to \mathbb{R}^N$ $\mathbf{v} \in \mathbb{R}^M$
$\frac{\partial}{\partial \mathbf{v}}\ f(\mathbf{g}(\mathbf{v}), \mathbf{h}(\mathbf{v}))$	$\frac{\partial f}{\partial \mathbf{g}}\frac{\partial \mathbf{g}}{\partial \mathbf{v}} + \frac{\partial f}{\partial \mathbf{h}}\frac{\partial \mathbf{h}}{\partial \mathbf{v}}$	$\frac{\partial \mathbf{g}}{\partial \mathbf{v}}\frac{\partial f}{\partial \mathbf{g}} + \frac{\partial \mathbf{h}}{\partial \mathbf{v}}\frac{\partial f}{\partial \mathbf{h}}$	$h : \mathbb{R}^K \times \mathbb{R}^N \to \mathbb{R}$, $\mathbf{v} \in \mathbb{R}^M$ $f : \mathbb{R}^M \to \mathbb{R}^K$, $\mathbf{g} : \mathbb{R}^M \to \mathbb{R}^N$

(c) **Neural Network:** A machine learning model that aims to approximate some function through the composition of both linear and nonlinear functions. There are two parts of neural networks: forward pass and backpropagation.

(a) **Forward Pass**: The process of calculating the predicted output of your network (and corresponding loss), given data, weights, and network structure. Given the input data $\mathbf{x}$, we can 1) transform the data using the weights associated with the first layer, W, then 2) apply the corresponding activation function to the output, and then 3) pass the result to the next layer and repeat. The forward pass does not involve taking derivatives and proceeds from the input layer to the output layer.

(b) **Backpropagation**: Given a neural network and a corresponding loss function, backpropagation gives us the gradient of the loss function with respect to the weights of the neural network. The method is called backward propagation because to efficiently apply the chain rule repeatedly, we calculate the dervatives of the final layer first, then proceed backward to the first layer.

(d) **Activation Function**: A nonlinear function, that is added to a neural network in order to help the network learn more complex patterns in the data. A few common ones are listed below.

Name	Function Definition	Derivative
logistic (sigmoid)	$g(z) = \frac{1}{1+e^{-z}}$	$g'(z) = g(z)(1 - g(z))$
tanh	$g(z) = \frac{e^z - e^{-z}}{e^z + e^{-z}}$	$1 - g^2(z)$
ReLU	$g(z) = \max(0, z)$	$g'(z) = \begin{cases} 1 & \text{if } z > 0 \\ 0 & \text{if } z < 0 \end{cases}$

2 Chain Rule Is All You Need

Note: Just generic math notation here. So, for example, $h(x)$ definitely doesn't refer to a hypothesis function.

Let $m : \mathbb{R}^2 \to \mathbb{R}$, $g : \mathbb{R} \to \mathbb{R}$, and $h : \mathbb{R} \to \mathbb{R}$ be functions defined as follows:

- $m(x_1, x_2) = x_1 x_2$ Note: we used m for multiply :)
- $g(x) = \sin(x)$
- $h(x) = x^2$

Suppose $x \in \mathbb{R}$ is given. We define the composite function $y = m(u, v) = m(g(x), h(x))$, where $u = g(x)$ and $v = h(x)$.

1. Apply multivariate chain rule to write $\frac{dy}{dx}$ in terms of $\frac{dy}{du}$, $\frac{dy}{dv}$, $\frac{du}{dx}$, and $\frac{dv}{dx}$.

$$\frac{dy}{dx} = \frac{d}{dx} m(u, v) = \frac{dy}{du}\frac{du}{dx} + \frac{dy}{dv}\frac{dv}{dx}$$

2. Find $\frac{dy}{dx}$ using the equation from the previous part.

$$\begin{aligned}\frac{dy}{dx} &= \frac{dy}{du}\frac{du}{dx} + \frac{dy}{dv}\frac{dv}{dx} \\ &= v(\cos x) + u(2x) \\ &= x^2 \cos x + 2x \sin(x)\end{aligned}$$

3. Rewrite the equation from part 1 to show that $\frac{d(uv)}{dx} = u\frac{dv}{dx} + v\frac{du}{dx}$.

$$\begin{aligned}\frac{d(uv)}{dx} &= \frac{dy}{dx} \\ &= \frac{dy}{du}\frac{du}{dx} + \frac{dy}{dv}\frac{dv}{dx} \\ &= v\frac{du}{dx} + u\frac{dv}{dx}\end{aligned}$$

4. Note that the statement proven in part 3 is the product rule. Now, try to prove the quotient rule in a similar way. Let $q : \mathbb{R}^2 \to \mathbb{R}$ and $y \in \mathbb{R}$ such that $q(x_1, x_2) = \frac{x_1}{x_2}$ and $y = q(u, v) = q(g(x), h(x))$. Prove the following equation:

$$\frac{dy}{dx} = \frac{v\frac{du}{dx} - u\frac{dv}{dx}}{v^2}$$

$$\begin{aligned}\frac{dy}{dx} &= \frac{dy}{du}\frac{du}{dx} + \frac{dy}{dv}\frac{dv}{dx} \\ &= \left(\frac{1}{v}\right)\left(\frac{du}{dx}\right) - \frac{u}{v^2}\frac{dv}{dx} \\ &= \frac{v\frac{du}{dx} - u\frac{dv}{dx}}{v^2}\end{aligned}$$

3 Neural Networks Are Fun

Assume we have the following data point $\mathbf{x}$:

$$\mathbf{x} = \begin{bmatrix} 2 \\ 5 \end{bmatrix}$$

with corresponding binary label:

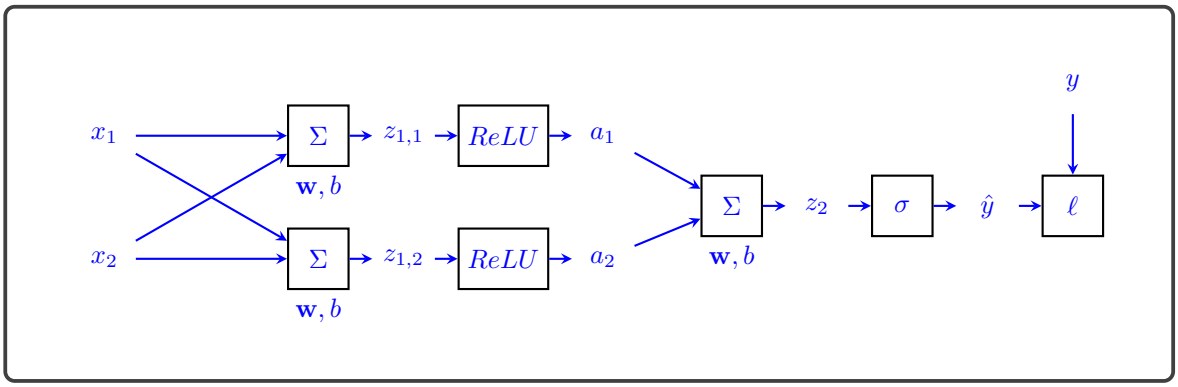
$$y = 1$$

which is part of a larger dataset X with binary labels $\mathbf{y}$.

Pat has tasked you with creating a neural network to solve this classification problem. For the loss function, he wants you to use the squared error: $\ell(y, \hat{y}) = (y - \hat{y})^2$. (Pat would like to note that using squared error for classification problems is very strange!) He also requests you to use stochastic gradient descent to train the network by minimizing the loss. Finally, he specifies that the neural network should have only one hidden layer with two neurons, a ReLU activation function at the hidden layer, a sigmoid activation function at the output layer, and all bias terms should be included.

3.1 Forward Pass

1. Draw what the neural network will look like.



2. We often group weights into a single matrix for each layer of neurons. How many weight matrices are there in the neural network?

2

Assume our weights and biases are as follows with $W_{l,i,j}$ and b_l where l represents the layer and i, j represent the row and column index, respectively:

$$W_1 = \begin{bmatrix} W_{1,1,1} & W_{1,1,2} \\ W_{1,2,1} & W_{1,2,2} \end{bmatrix} = \begin{bmatrix} 3 & -1 \\ -3 & 1 \end{bmatrix}$$
$$b_1 = \begin{bmatrix} b_{1,1} \\ b_{1,2} \end{bmatrix} = \begin{bmatrix} 2 \\ -1 \end{bmatrix}$$
$$W_2 = \begin{bmatrix} W_{2,1,1} & W_{2,1,2} \end{bmatrix} = \begin{bmatrix} -2 & 2 \end{bmatrix}$$
$$b_2 = 1$$

Here are some intermediate values that we'll use going forward (haha, "forward" and "backward"):

For the hidden layer

$$\mathbf{z_1} = W_1\mathbf{x} + \mathbf{b_1}$$
$$\mathbf{a} = \text{ReLU}(\mathbf{z_1})$$

For the output layer

$$z_2 = W_2\mathbf{a} + b_2$$
$$\hat{y} = \sigma(z_2)$$

3. What are the values in $\mathbf{z_1}$, the output of the hidden neurons before applying the activation? Recall we need to add an extra one to the start of $\mathbf{x}$ vector to account for the bias term in this layer!

$$\mathbf{z_1} = W_1\mathbf{x} + \mathbf{b_1} = \begin{bmatrix} 3 & -1 \\ -3 & 1 \end{bmatrix} \begin{bmatrix} 2 \\ 5 \end{bmatrix} + \begin{bmatrix} 2 \\ -1 \end{bmatrix} = \begin{bmatrix} 6-5 \\ -6+5 \end{bmatrix} + \begin{bmatrix} 2 \\ -1 \end{bmatrix} = \begin{bmatrix} 3 \\ -2 \end{bmatrix}$$

4. What is $\mathbf{a}$, the output values of our hidden neurons?

We need to apply the ReLU function to this vector to get the output. Recall that the ReLU function is applied to each element individually and is defined as:

$$g(x) = \max(0, x)$$

Our output will be $\mathbf{a} = \begin{bmatrix} 3 \\ 0 \end{bmatrix}$

5. What is z_2, the values from the output layer (before activation)?

$z_2 = W_2\mathbf{a} + b_2 = \begin{bmatrix}-2 & 2\end{bmatrix}\begin{bmatrix}3\\0\end{bmatrix} + 1 = -6 + 0 + 1 = -5$

6. What is our final output $\hat{y}$?

We need to apply the sigmoid function to the value z_2.

$\hat{y} = \frac{1}{1+e^5} = 0.00669$

7. What does our model classify $\mathbf{x}$ as?

0

8. What is the loss $\ell(y, \hat{y})$ on this example? (Note: recall that Pat admitted that using squared error with classification is strange, but we can still do it.)

$\ell(y, \hat{y}) = (y - \hat{y})^2 = (1 - 0.00669)^2 = 0.98666$

Objective Function

When we are considering all of the training data, as we do in gradient descent, our objective function is:

$$J(W_1, W_2; \mathbf{y}, X) = \frac{1}{N}\sum_{i=1}^{N} \ell\left(y^{(i)}, \hat{y}^{(i)}\right) \tag{1}$$

$$= \frac{1}{N}\sum_{i=1}^{N} \left(y^{(i)} - \sigma\left(W_2 \text{ReLU}\left(W_1\mathbf{x}^{(i)} + \mathbf{b_1}\right) + b_2\right)\right)^2 \tag{2}$$

However, when we are using stochastic gradient descent, our objective function is with respect to just one training point, $\mathbf{x}, y$, at a time and thus, the output of the loss, ℓ, is our objective function for that one point:

$$J(W_1, W_2; y, \mathbf{x}) = \ell\left(y, \hat{y}\right) \tag{3}$$

$$= \left(y - \sigma\left(W_2 \text{ReLU}\left(W_1\mathbf{x} + \mathbf{b_1}\right) + b_2\right)\right)^2 \tag{4}$$

3.2 Parameter Counting

1. How many parameters are learned in the first hidden linear layer?

   The parameters are the weights and biases learned. Since the first weight matrix is $W_1 \in \mathbb{R}^{2\times 2}$ and the bias is $\mathbf{b_1} \in \mathbb{R}^{2\times 1}$, there are 6 parameters in the first linear layer.

2. How many parameters are learned in the first activation layer?

   0. Activation layers do not learn parameters.

3. How many parameters are learned in the entire neural network?

   9. There are 6 parameters from the first linear layer and 3 parameters from the second linear layer.

3.3 Backpropagation

Since our goal is to find the best weights that optimize our neural network, now we'll do backpropagation to update the weights we were given. You will need to calculate the derivative of J with respect to the weights W_1 and W_2, then you use gradient descent to update them. Throughout this assignment, we will use numerator layout for our derivatives!

In the interest of time, let's just calculate the derivatives: $\frac{\partial J}{\partial W_{2,1,1}}$ and $\frac{\partial J}{\partial W_{1,1,2}}$.

Let's start with $\frac{\partial J}{\partial W_{2,1,1}}$, where $W_{2,1,1}$ is the $1,1$ entry in W_2 weight matrix.

1. Using the multivariate chain rule, write the derivative chain expression for $\frac{\partial J}{\partial W_{2,1,1}}$.

   Hint: think about writing out the chain rule as d-out/d-in for each layer: $\frac{\partial out}{\partial in}\frac{\partial out}{\partial in}\cdots\frac{\partial out}{\partial weight}$

   $\frac{\partial J}{\partial W_{2,1,1}} = \frac{\partial \ell}{\partial \hat{y}}\frac{\partial \hat{y}}{\partial z_2}\frac{\partial z_2}{\partial W_{2,1,1}}$

2. What is $\frac{\partial \ell}{\partial \hat{y}}$? Write in terms of y and $\hat{y}$

   $-2(y-\hat{y})$

3. What is $\frac{\partial \hat{y}}{\partial z_2}$? Write in terms of z_2 (but then simplify it to write in terms of just $\hat{y}$).

> The derivative of the sigmoid function is $\sigma(z_2)(1-\sigma(z_2))$.
>
> We can simplify this further as $\hat{y}(1-\hat{y})$.

4. What is $\frac{\partial z_2}{\partial W_{2,1,1}}$?

> $z_2 = W_2\mathbf{a} + b_2 = W_{2,1,1}a_1 + W_{2,2,1}a_2 + b_2$
>
> $\frac{\partial z_2}{\partial W_{2,1,1}} = a_1 = 3$

5. Finally, what is $\frac{\partial J}{\partial W_{2,1,1}}$? Hint: remember the chain rule that we wrote for this above.

> $$\begin{aligned}\frac{\partial J}{\partial W_{2,1,1}} &= \frac{\partial \ell}{\partial \hat{y}}\frac{\partial \hat{y}}{\partial z_2}\frac{\partial z_2}{\partial W_{2,1,1}} \\ &= -2(y-\hat{y})\hat{y}(1-\hat{y})\cdot 3 \\ &= -2(1-0.00669)(0.00669)(1-0.00669)(3) = -0.03960\end{aligned}$$

Now we will calculate the derivative with respect to $W_{1,1,2}$: $\frac{\partial J}{\partial W_{1,1,2}}$, where $W_{1,1,2}$ is the $1,2$ entry in W_1 weight matrix.

6. What is the derivative chain expression for $\frac{\partial J}{\partial W_{1,1,2}}$? Reminder: d-out/d-in, d-out/d-in, ...

> $\frac{\partial J}{\partial W_{1,1,2}} = \frac{\partial \ell}{\partial \hat{y}}\frac{\partial \hat{y}}{\partial z_2}\frac{\partial z_2}{\partial \mathbf{a}}\frac{\partial \mathbf{a}}{\partial \mathbf{z_1}}\frac{\partial \mathbf{z_1}}{\partial W_{1,1,2}}$
>
> Note: Here's a huge shortcut that we'll take. Given that there are some zeros in various partial derivatives when we only need the derivative with respect to $W_{1,1,2}$ (not requiring the full vector $\mathbf{a}$ but only a_1), we can do:
>
> $\frac{\partial J}{\partial W_{1,1,2}} = \frac{\partial \ell}{\partial \hat{y}}\frac{\partial \hat{y}}{\partial z_2}\frac{\partial z_2}{\partial a_1}\frac{\partial a_1}{\partial z_{1,1}}\frac{\partial z_{1,1}}{\partial W_{1,1,2}}$

We already have the first two derivatives calculated from the previous question.

7. What is $\frac{\partial z_2}{\partial a_1}$?

> $z_2 = W_2\mathbf{a} + b_2 = W_{2,1,1}a_1 + W_{2,1,2}a_2 + b_2$
>
> $\frac{\partial z_2}{\partial a_1} = W_{2,1,1} = -2$

8. What is $\frac{\partial a_1}{\partial z_{1,1}}$?

   This is the derivative of the ReLU function, which is equal to 1 if the input is greater than 0, else it is equal to 0 if the input is less than 0. In this case, since $z_{1,1} = 3$, the derivative is 1.

9. What is $\frac{\partial z_{1,1}}{\partial W_{1,1,2}}$?

   $z_{1,1} = W_{1,1,1}x_1 + W_{1,1,2}x_2 + b_{1,1}$
   $\frac{\partial z_{1,1}}{\partial W_{1,1,2}} = x_2 = 5$

10. Finally, what is $\frac{\partial J}{\partial W_{1,1,2}}$?

    $\frac{\partial J}{\partial W_{1,1,2}} = -2(y - \hat{y})\hat{y}(1 - \hat{y})(-2)(1)(5)$
    $= -2(1 - 0.00669)(0.00669)(1 - 0.00669)(-2)(1)(5) = 0.13202$

11. What is our updated $W_{2,1,1}$ and $W_{1,1,2}$ if we use learning rate $\alpha = 2$?

    Recall gradient descent rule: $\theta_j^{(t+1)} = \theta_j^{(t)} - \alpha \frac{\partial J}{\partial \theta_j}$
    $W_{2,1,1} = -2 - 2 \cdot (-0.03960) = -1.9208$
    $W_{1,1,2} = -1 - 2 \cdot (0.13202) = -1.2640$

4 Conceptually Understanding Neural Networks

1. Compare neural networks with linear regression and logistic regression. How are the models similar, and how are they different?

   One similarity between all three models is that they can be trained using gradient descent. Only linear regression has a closed-form solution.

   Neural networks are the only model within the three that can model nonlinear data without using feature engineering. This is due to nonlinear activation functions. On the other hand, linear regression and logistic regression are simpler models. If they perform well enough, they may be preferred because they are more interpretable.