1 Paper Problems

1.

$$
\begin{aligned}
z &= \sigma(y_1^2 + y_2 y_3) \\
\frac{\partial z}{\partial x} &= \frac{\partial}{\partial x}\sigma(y_1^2 + y_2 y_3) \\
&= \sigma(y_1^2 + y_2 y_3)(1 - \sigma(y_1^2 + y_2 y_3)) * \frac{\partial}{\partial x}\left[y_1^2 + y_2 y_3\right] \\
\frac{\partial}{\partial x}\left[y_1^2 + y_2 y_3\right] &= 2y_1 * \frac{\partial}{\partial x}3x + \left[\frac{\partial}{\partial x}y_2 * y_3 + y_2 * \frac{\partial}{\partial x}y_3\right] \\
&= 6x + \left[\left(-e^{-x}\right) * \sin(x) + \cos(x) * e^{-x}\right] \\
&= 6x + e^{-x}\left[\cos(x) - \sin(x)\right] \\
\frac{\partial z}{\partial x} &= \sigma(y_1^2 + y_2 y_3)(1 - \sigma(y_1^2 + y_2 y_3)) * \left(6x + e^{-x}\left[\cos(x) - \sin(x)\right]\right) \\
\frac{\partial z}{\partial x}(0) &= \sigma(0)(1 - \sigma(0)) * (0 + (1)(1 - 0)) \\
&= \frac{1}{2} * (1 - \frac{1}{2}) * 1 \\
&= \frac{1}{4}
\end{aligned}
$$

2.

Layer One:

$$
\begin{aligned}
z_1^1 &= \sigma\left(w_{01}^1[x_0] + w_{11}^1[x_1] + w_{21}^1[x_2]\right) \\
&= \sigma\left((-1 * 1) + (-2 * 1) + (-3 * 1)\right) \\
&= \sigma\left(-6\right) = 0.002473 \\
z_2^1 &= \sigma\left(w_{02}^1[x_0] + w_{12}^1[x_1] + w_{22}^1[x_2]\right) \\
&= \sigma\left((1 * 1) + (2 * 1) + (3 * 1)\right) \\
&= \sigma\left(6\right) = 0.997527
\end{aligned}
$$

Layer Two:

$$
\begin{aligned}
z_1^2 &= \sigma\left(w_{01}^2[z_0^1] + w_{11}^2[z_1^1] + w_{21}^2[z_2^1]\right) \\
&= \sigma\left((-1 * 1) + (-2 * 0.002473) + (-3 * 0.997527)\right) \\
&= \sigma\left(-3.997527\right) = 0.018030 \\
z_2^2 &= \sigma\left(w_{02}^2[z_0^1] + w_{12}^2[z_1^1] + w_{22}^2[z_2^1]\right) \\
&= \sigma\left((1 * 1) + (2 * 0.002473) + (3 * 0.997527)\right) \\
&= \sigma\left(3.997527\right) = 0.981970
\end{aligned}
$$

Layer Three:

$$
\begin{aligned}
y &= w_{01}^3[z_0^2] + w_{11}^3[z_1^2] + w_{21}^3[z_2^2] \\
&= (-1 * 1) + (2 * 0.018030) + (-1.5 * 0.981970) \\
&= -2.436895
\end{aligned}
$$

3. Bolded values were from previously calculated and cached values.
Step One:

$$\begin{aligned}
\frac{\partial L}{\partial y} &= (y - y^*) = -2.436895 - 1 \\
&= -3.436895 \\
\frac{\partial L}{\partial w_{01}^3} &= \frac{\partial L}{\partial y}\frac{\partial y}{\partial w_{01}^3} = \textbf{-3.436895} * 1 \\
&= -3.436895 \\
\frac{\partial L}{\partial w_{11}^3} &= \frac{\partial L}{\partial y}\frac{\partial y}{\partial w_{11}^3} = \frac{\partial L}{\partial y} * z_1^2 = \textbf{-3.436895} * \textbf{0.018030} \\
&= -0.061967 \\
\frac{\partial L}{\partial w_{21}^3} &= \frac{\partial L}{\partial y}\frac{\partial y}{\partial w_{21}^3} = \frac{\partial L}{\partial y} * z_2^2 = \textbf{-3.436895} * \textbf{0.981970} \\
&= -3.374928
\end{aligned}$$

Just gonna stick each step on its own page since they just barely fit.

I never want to have to do this by hand ever again. I have no doubt that I made a single mistake somewhere and it propagated up through the rest of it. There is clearly a reason we make the computers do this.

Also, I really hope I'm not wrong in thinking that $\frac{\partial \sigma}{\partial s_z}$ is just $z\,(1 - z)$. I'm pretty sure that's what I'm seeing when looking at the derivative of the activated node...

Step Two:

Node One:

$$\frac{\partial L}{\partial z_1^2} = \frac{\partial L}{\partial y} * \frac{\partial y}{\partial z_1^2} = \frac{\partial L}{\partial y} * w_{11}^3$$
$$= \textbf{-3.436895} * 2$$
$$= -6.87379$$
$$\frac{\partial \sigma}{\partial s_{z_1^2}} = z_1^2 \left(1 - z_1^2\right)$$
$$= \textbf{0.018030} \left(1 - \textbf{0.018030}\right)$$
$$= 0.017705$$
$$\frac{\partial L}{\partial w_{01}^2} = \frac{\partial L}{\partial z_1^2} * \frac{\partial z_1^2}{\partial w_{01}^2} = \frac{\partial L}{\partial z_1^2} * \frac{\partial \sigma}{\partial s_{z_1^2}} z_0^1$$
$$= \textbf{-6.87379} * [\textbf{0.017705} * \textbf{1}]$$
$$= -0.121700$$
$$\frac{\partial L}{\partial w_{11}^2} = \frac{\partial L}{\partial z_1^2} * \frac{\partial z_1^2}{\partial w_{11}^2} = \frac{\partial L}{\partial z_1^2} * \frac{\partial \sigma}{\partial s_{z_1^2}} z_1^1$$
$$= \textbf{-6.87379} * [\textbf{0.017705} * \textbf{0.002473}]$$
$$= -0.000301$$
$$\frac{\partial L}{\partial w_{21}^2} = \frac{\partial L}{\partial z_1^2} * \frac{\partial z_1^2}{\partial w_{21}^2} = \frac{\partial L}{\partial z_1^2} * \frac{\partial \sigma}{\partial s_{z_1^2}} z_2^1$$
$$= \textbf{-6.87379} * [\textbf{0.017705} * \textbf{0.997527}]$$
$$= -0.121399$$

Node Two:

$$\frac{\partial L}{\partial z_2^2} = \frac{\partial L}{\partial y} * \frac{\partial y}{\partial z_2^2} = \frac{\partial L}{\partial y} * w_{21}^3$$
$$= \textbf{-3.436895} * -1.5$$
$$= 5.155343$$
$$\frac{\partial \sigma}{\partial s_{z_2^2}} = z_2^2 \left(1 - z_2^2\right)$$
$$= \textbf{0.981970} \left(1 - \textbf{0.981970}\right)$$
$$= 0.017705$$
$$\frac{\partial L}{\partial w_{02}^2} = \frac{\partial L}{\partial z_2^2} * \frac{\partial z_2^2 2}{\partial w_{02}^2} = \frac{\partial L}{\partial z_2^2} * \frac{\partial \sigma}{\partial s_{z_2^2}} z_0^1$$
$$= \textbf{5.155343} * [\textbf{0.017705} * \textbf{1}]$$
$$= 0.91275$$
$$\frac{\partial L}{\partial w_{12}^2} = \frac{\partial L}{\partial z_2^2} * \frac{\partial z_2^2}{\partial w_{12}^2} = \frac{\partial L}{\partial z_2^2} * \frac{\partial \sigma}{\partial s_{z_2^2}} z_1^1$$
$$= \textbf{5.155343} * [\textbf{0.017705} * \textbf{0.002473}]$$
$$= 0.000226$$
$$\frac{\partial L}{\partial w_{22}^2} = \frac{\partial L}{\partial z_2^2} * \frac{\partial z_2^2}{\partial w_{22}^2} = \frac{\partial L}{\partial z_2^2} * \frac{\partial \sigma}{\partial s_{z_2^2}} z_2^1$$
$$= \textbf{5.155343} * [\textbf{0.017705} * \textbf{0.997527}]$$
$$= 0.09150$$

Step Three:

Node One:

$$\begin{aligned}
\frac{\partial L}{\partial z_1^1} &= \frac{\partial L}{\partial z_1^2} * \frac{\partial z_1^2}{\partial z_1^1} + \frac{\partial L}{\partial z_2^2} * \frac{\partial z_2^2}{\partial z_1^1} \\
&= \frac{\partial L}{\partial z_1^2} * w_{11}^2 + \frac{\partial L}{\partial z_2^2} * w_{12}^2 \\
&= \mathbf{6.87379} * -2 + \mathbf{5.155343} * 2 \\
&= 24.058266 \\
\frac{\partial \sigma}{\partial s_{z_1^1}} &= z_1^1 \left(1 - z_1^1\right) \\
&= \mathbf{0.002473} \left(1 - \mathbf{0.002473}\right) \\
&= 0.002467 \\
\frac{\partial L}{\partial w_{01}^1} &= \frac{\partial L}{\partial z_1^1} * \frac{\partial z_1^1}{\partial w_{01}^1} = \frac{\partial L}{\partial z_1^1} * \frac{\partial \sigma}{\partial s_{z_1^1}} x_0 \\
&= \mathbf{24.058266} * [\mathbf{0.002467} * \mathbf{1}] \\
&= 0.059349 \\
\frac{\partial L}{\partial w_{11}^1} &= \frac{\partial L}{\partial z_1^1} * \frac{\partial z_1^1}{\partial w_{11}^1} = \frac{\partial L}{\partial z_1^1} * \frac{\partial \sigma}{\partial s_{z_1^1}} x_1 \\
&= \mathbf{24.058266} * [\mathbf{0.002467} * \mathbf{1}] \\
&= 0.059349 \\
\frac{\partial L}{\partial w_{21}^1} &= \frac{\partial L}{\partial z_1^1} * \frac{\partial z_1^1}{\partial w_{21}^1} = \frac{\partial L}{\partial z_1^1} * \frac{\partial \sigma}{\partial s_{z_1^1}} x_2 \\
&= \mathbf{24.058266} * [\mathbf{0.002467} * \mathbf{1}] \\
&= 0.059349
\end{aligned}$$

Node Two:

$$\begin{aligned}
\frac{\partial L}{\partial z_2^1} &= \frac{\partial L}{\partial z_1^2} * \frac{\partial z_1^2}{\partial z_2^1} + \frac{\partial L}{\partial z_2^2} * \frac{\partial z_2^2}{\partial z_2^1} \\
&= \frac{\partial L}{\partial z_1^2} * w_{21}^2 + \frac{\partial L}{\partial z_2^2} * w_{22}^2 \\
&= \mathbf{6.87379} * -3 + \mathbf{5.155343} * 3 \\
&= 36.087399 \\
\frac{\partial \sigma}{\partial s_{z_2^1}} &= z_2^1 \left(1 - z_2^1\right) \\
&= \mathbf{0.997527} \left(1 - \mathbf{0.997527}\right) \\
&= 0.002467 \\
\frac{\partial L}{\partial w_{02}^1} &= \frac{\partial L}{\partial z_2^1} * \frac{\partial z_2^2 2}{\partial w_{02}^1} = \frac{\partial L}{\partial z_2^1} * \frac{\partial \sigma}{\partial s_{z_2^1}} x_0 \\
&= \mathbf{36.087399} * [\mathbf{0.002467} * \mathbf{1}] \\
&= 0.089023 \\
\frac{\partial L}{\partial w_{12}^1} &= \frac{\partial L}{\partial z_2^1} * \frac{\partial z_2^1}{\partial w_{12}^1} = \frac{\partial L}{\partial z_2^1} * \frac{\partial \sigma}{\partial s_{z_2^1}} x_1 \\
&= \mathbf{36.087399} * [\mathbf{0.002467} * \mathbf{1}] \\
&= 0.089023 \\
\frac{\partial L}{\partial w_{22}^1} &= \frac{\partial L}{\partial z_2^1} * \frac{\partial z_2^1}{\partial w_{22}^1} = \frac{\partial L}{\partial z_2^1} * \frac{\partial \sigma}{\partial s_{z_2^1}} x_2 \\
&= \mathbf{36.087399} * [\mathbf{0.002467} * \mathbf{1}] \\
&= 0.089023
\end{aligned}$$

Could be wrong since the weights of all of the first layer ended up the same for the individual nodes, I'm pretty sure that's caused by having the same values for all of $\mathbf{x}$ though. However, this was a nightmare and I do not intend on doing anything more with this problem. I wash my hands of it, I should have just done one of the extra credit coding problems and skipped this. Would have saved me hours of my life.

4.

a) **Objective Function**

$$J(\mathbf{w}) = 3 * \log\left(1 + \exp(-y_i \mathbf{w}^T \mathbf{x}_i)\right) + \mathbf{w}^T \mathbf{w}$$

I'm assuming that since we're using stochastic gradient descent rather than taking a sum I can just use the number of samples, ie 3.

Gradient

Please note that given that its 3 samples I am not shuffling them.

$$\begin{aligned}
\frac{\partial J}{\partial w_j} &= \frac{\partial}{\partial w_j} 3 * \log\left(1 + \exp(-y_i \mathbf{w}^T \mathbf{x}_i)\right) + \mathbf{w}^T \mathbf{w} \\
&= 2w_j + \frac{\partial}{\partial w_j} 3 * \log\left(1 + \exp(-y_i \mathbf{w}^T \mathbf{x}_i)\right) \\
&= 2w_j + \frac{3}{(1 + \exp(-y_i \mathbf{w}^T \mathbf{x}_i))} * \frac{\partial}{\partial w_j}\left[1 + exp(-y_i \mathbf{w}^T \mathbf{x}_i]\right) \\
&= 2w_j - \frac{3 y_i x_i}{(1 + \exp(-y_i \mathbf{w}^T \mathbf{x}_i))} \\
\nabla J &= [\mathbf{w}_0; 0] - \frac{3 y_i}{(1 + \exp(-y_i \mathbf{w}^T \mathbf{x}_i))} \mathbf{x}_i
\end{aligned}$$

b)

I learned from the last assignment that this is not worth 3 points.

2 Practice

2.

b.

I ran it several times and my results varied slightly (massively one of the times, not entirely sure what happened there). The following table is a list of results that seemed to be fairly standard from my runs. One of the runs the 100 width spiked up to an error of 30%, not sure what happened there and I was unable to reproduce it.

Width	Training Error	Test Error
5	0.0	0.0
10	0.0	0.0
25	0.0	0.0
50	0.0	0.002
100	0.009	0.012

c.

Initializing with zero weights preformed slightly (but consistently) worse, the following table shows typical results for it. That being said I think a test error of 1% is pretty impressive.

Width	Training Error	Test Error
5	0.009	0.008
10	0.009	0.01
25	0.009	0.01
50	0.009	0.008
100	0.011	0.012

d.

My SVM was consistently seeing errors around 3%. So the neural network preforms significantly better than it, if I recall it also took longer to run but I'm not certain. In terms of implementation SVMs were a lot more friendly.

e.

I attempted it, got frustrated when I couldn't get it to work and just deleted it. I really probably should have kept it to get at least partial credit but I didn't think about it before deleting it...

3.

Didn't do

4.

I prefer sklearn, which I think uses tensorflow in the background, not sure about that though. Starting out I tried to emulate how they have it set up but kinda lost my way at some point. I will say that implementing it all ourselves taught me quite a bit on how they work behind the scenes.