1. **Backpropagation for autoencoders.** Consider $\mathbf{x} \in \mathbb{R}^n$. Further, consider $\mathbf{W} \in \mathbb{R}^{m \times n}$ where $m < n$. Then $\mathbf{Wx}$ is of lower dimensionality that $\mathbf{x}$. One way to design $\mathbf{W}$ so that $\mathbf{Wx}$ still contains key features of $\mathbf{x}$ is to minimize the following expression.

$$\mathcal{L} = \frac{1}{2}\|\mathbf{W^TWx} - \mathbf{x}\|^2$$

(a) In words, describe why this minimization finds a $\mathbf{W}$ that ought to preserve information about $\mathbf{x}$.

Response: By minimizing the loss, $\mathbf{W}$ is trained such that the hidden representation $\mathbf{Wx}$ preserves the information about $\mathbf{x}$.

(b) Draw the computational graph for $\mathcal{L}$.

Response:

(c) In the computational graph, there should be two paths to $\mathbf{W}$. How do we account for these two paths when calculating $\nabla_{\mathbf{W}}\mathcal{L}$?

Response: We can use the law of total derivatives. Consider the following example: $a \to b \to d$ and $a \to c \to d$. Then, the total derivative of d with respect to a is given as:

$$\frac{\partial d}{\partial a} = \frac{\partial d}{\partial b} \cdot \frac{\partial b}{\partial a} + \frac{\partial d}{\partial c} \cdot \frac{\partial c}{\partial a}$$

(d) Calculate the gradient: $\nabla_{\mathbf{W}}\mathcal{L}$.

Response: Given the computational graph drawn in (b),

$$\begin{aligned}
\frac{\partial \mathcal{L}}{\partial(\mathbf{W^TWx} - \mathbf{x})} &= \mathbf{W^TWx} - \mathbf{x} \\
\frac{\partial \mathcal{L}}{\partial(\mathbf{W^T})} &= (\mathbf{W^TWx} - \mathbf{x})(\mathbf{Wx})^{\mathbf{T}} && (*) \\
\frac{\partial \mathcal{L}}{\partial(\mathbf{Wx})} &= \mathbf{W}(\mathbf{W^TWx} - \mathbf{x}) \\
\frac{\partial \mathcal{L}}{\partial \mathbf{W}} &= \mathbf{W}(\mathbf{W^TWx} - \mathbf{x})\mathbf{x^T} && (**)
\end{aligned}$$

so $\nabla_{\mathbf{W}}\mathcal{L} = (*) + (**) = \mathbf{W}(\mathbf{W^TWx} - \mathbf{x})\mathbf{x^T} + \left((\mathbf{W^TWx} - \mathbf{x})(\mathbf{Wx})^{\mathbf{T}}\right)^{\mathbf{T}} = \mathbf{W}(\mathbf{W^TWx} - \mathbf{x})\mathbf{x^T} + (\mathbf{Wx})(\mathbf{W^TWx} - \mathbf{x})^{\mathbf{T}}$.

2. **Backpropagation for Gaussian-process latent variable model. (Optional for students in C147: Please write 'I am a CS147 student' in the solution and you will get full credit for this problem).**

Response: *I am a CS147 student.*

3. **NNDL to the rescue!!** The Swish activation function for any scalar input K is defined as,

$$\text{swish}(k) = \frac{k}{1+e^{-k}} = k\sigma(k),$$

where $\sigma(k)$ is the sigmoid activation function you have seen in lecture.

(a) Draw the computational graph for the 2-layer FC net.

Response:

(b) Compute $\nabla_{W_2} L, \nabla_{b_2} L$.

Response: Using the computational graph drawn in (a),

$$\begin{aligned}
\frac{\partial L}{\partial b_2} &= \frac{\partial L}{\partial z_2} \\
\frac{\partial L}{\partial (W_2 h_1)} &= \frac{\partial L}{\partial z_2} \\
\frac{\partial L}{\partial W_2} &= \frac{\partial L}{\partial (W_2 h_1)} (h_1)^T
\end{aligned}$$

so $\nabla_{W_2} L = \frac{\partial L}{\partial (W_2 h_1)} h_1^T, \nabla_{b_2} L = \frac{\partial L}{\partial z_2}$

(c) Compute $\nabla_{W_1} L, \nabla_{b_1} L$.

Response: Using the computational graph drawn in (a) and noting that

$$\frac{\partial(\text{swish}(z))}{\partial z} = \sigma(z)(z - z\sigma(z) + 1)$$

we get

$$\begin{aligned}
\frac{\partial L}{\partial b_1} &= \frac{\partial(\text{swish}(z))}{\partial z} \odot W_2^T \frac{\partial L}{\partial z_2} = \sigma(z)(z - z\sigma(z) + 1) \odot W_2^T \frac{\partial L}{\partial z_2} \\
\frac{\partial L}{\partial (W_1 x)} &= \frac{\partial L}{\partial b_1} = \sigma(z)(z - z\sigma(z) + 1) \odot W_2^T \frac{\partial L}{\partial z_2} \\
\frac{\partial L}{\partial W_1} &= \frac{\partial L}{\partial (W_1 x)} x^T = \left(\sigma(z)(z - z\sigma(z) + 1) \odot W_2^T \frac{\partial L}{\partial z_2}\right) x^T
\end{aligned}$$

so $\nabla_{W_1} L = \left(\sigma(z)(z - z\sigma(z) + 1) \odot W_2^T \frac{\partial L}{\partial z_2}\right) x^T, \nabla_{b_1} L = \sigma(z)(z - z\sigma(z) + 1) \odot W_2^T \frac{\partial L}{\partial z_2}$