1. Taylor series gives a formula for approximating the value of the function $f(x)$ in a small neighborhood around it. Given that $3^3 = 27$ can you calculate the value of $(3.0001)^3$ using :

   (a) (**0.5 marks**) the first order approximation given by Taylor series

   Solution:

   (b) (**0.5 marks**) the second order approximation given by Taylor series

   Solution:

2. Consider the task of assigning one of the following 3 labels to an image : apple, banana, mango. Further, consider we are given n images for training such that each image belongs to one of these three categories. We have a model which assigns a probability distribution to each training example: $\mathbf{q} = [q_{apple}, q_{banana}, q_{mango}]$. We train the model by minimizing the cross entropy between the true distribution and the predicted distribution for each training example.

   (a) (**1 mark**) Show that this is the same as maximizing the log likelihood of the training data.

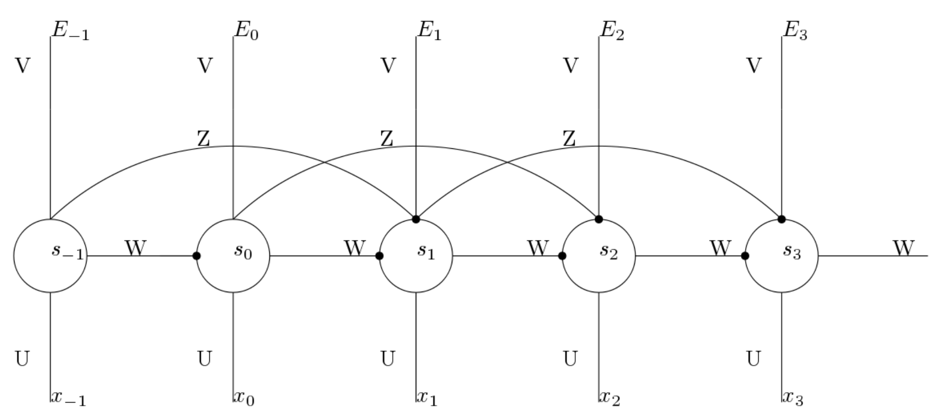
   Solution:

3. Consider $x \in \mathbb{R}^n$ and $A \in \mathbb{R}^{n \times n}$

   (a) (**1 mark**) Compute the gradient of $x^T Ax$ w.r.t. A

   Solution:

4. The figure belows shows a network, similar to the one we saw in the assignment. In addition to the dependencies that we saw before, the graph presented here has an additional dependency where s_i depends not just on the immediate ancestor, but on the one before that as well.



Here,

$$s_i = \sigma(Ws_{i-1} + Zs_{i-2} + Ux_i)$$

(a) (**1 mark**) Draw the dependency graph involving the variables s_1, s_2, s_3, W, Z

Solution:

(b) (**2 marks**) Based on the dependency graph from above, give a formula for computing $\frac{\partial s_3}{\partial W}$ and $\frac{\partial s_3}{\partial Z}$. Assume that s_1, s_2, s_3, W, Z are scalars

Solution:

5. We saw the proof of convergence for the Perceptron Learning algorithm in class. The proof relied on the fact that, **cos** of any angle is bounded by ± 1 and the fact that $cos\beta \propto \delta\sqrt{k}$, where β is an appropriately defined angle and k is the number of iterations of the learning algorithm. One question raised in the class was about the boundary case where $\delta = 0$. Recall that $\delta = \min w^* p_i \ \forall i$, where w^* is the optimal separating hyperplane and i runs over all points in the dataset.

   It would then seem like we cannot arrive at a contradiction when the number of iterations (k) of the Perceptron Learning Algorithm tends to ∞ (specifically, even when $k \to \infty, \delta\sqrt{k} = 0[\because \delta = 0]$ and hence $cos\beta$ is still bounded).

   (a) (**1 mark**) Does the Perceptron learning algorithm **not** converge if $\delta = 0$?

Solution:

   (b) (**1 mark**) If it does converge, how do you reconcile with the situation of $\delta = 0$?

Solution: