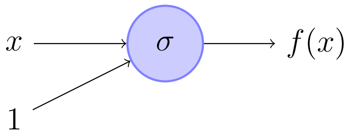
1. Derive the expression for the derivative of the inner product of two vectors $\mathbf{r}, \mathbf{s}$ w.r.t. θ, both of which are functions of parameter θ. Note that θ is a scalar.

$$\frac{d\langle \mathbf{r}.\mathbf{s}\rangle}{d\theta} = \left\langle \frac{d\mathbf{r}}{d\theta}.\mathbf{s}\right\rangle + \left\langle \frac{d\mathbf{s}}{d\theta}.\mathbf{r}\right\rangle$$

Solution:

2. **Partial derivatives**

(a) Consider the following computation,



$$f(x) = \frac{1}{1+e^{-(w\cdot x+b)}}$$

The value L is given by,

$$L = \frac{1}{2}(y - f(x))^2$$

Here, x and y are constants and w and b are parameters that can be modified. In other words, L is a function of w and b.

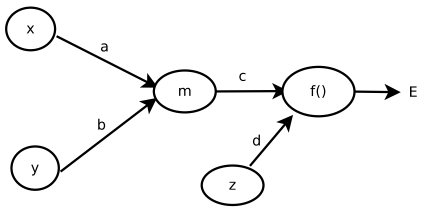
Derive the partial derivatives, $\frac{\partial L}{\partial w}$ and $\frac{\partial L}{\partial b}$.

Solution:

(b) Consider the evaluation of E as given below,

$$E = g(x, y, z) = f(c(ax + by) + dz)$$

Represented as a graph, we have



Here x, y, z are inputs (constants) and a, b, c, d are parameters (variables). m is an intermediate computation and f is some differentiable function. Specifically, let us consider f to be the $tanh$ function.

$$tanh(x) = \frac{e^x - e^{-x}}{e^x + e^{-x}}$$

Note that here E is a function of a, b, c, d.

Compute the partial derivatives of E with respect to the parameters a, b and d i.e. $\frac{\partial E}{\partial a}$, $\frac{\partial E}{\partial b}$ and $\frac{\partial E}{\partial d}$.

Solution:

3. The first order derivative of a function f is defined by the following limit,

$$\frac{df(x)}{dx} = \lim_{h \to 0} \frac{f(x+h) - f(x)}{h} \tag{1}$$

On observing the above definition we see that the derivative of a function is the ratio of change in the function value to the change in the function input, when we change the input by a small quantity (infinitesimally small).

Consider the function $f(x) = x^2$. The derivative of $f(x)$ is

$$\frac{df(x)}{dx} = 2x$$

The function evaluates to 4 at 2 i.e. $f(2) = 4$.

Say we wanted to estimate the value of $f(2.01)$ and $f(2.5)$ without using the definition of $f(x)$. We could think of using the definition of derivative to "extrapolate" the value of $f(2)$ to obtain $f(2.01)$ and $f(2.5)$.

A first degree approximation based on 1 would be the following.

$$f(x+h) \approx f(x) + h\frac{df(x)}{dx} \tag{2}$$

(a) Estimate $f(2.01)$ and $f(2.5)$ using the above formula.

Solution:

(b) Compare it to the actual value of $f(2.01) = 2.01^2 = 4.0401$, and $f(2.5) = 2.5^2 = 6.25$.

Solution:

(c) Explain the discrepancy from the actual value ? Why does it increase/decrease when we move further away from 2?

Solution:

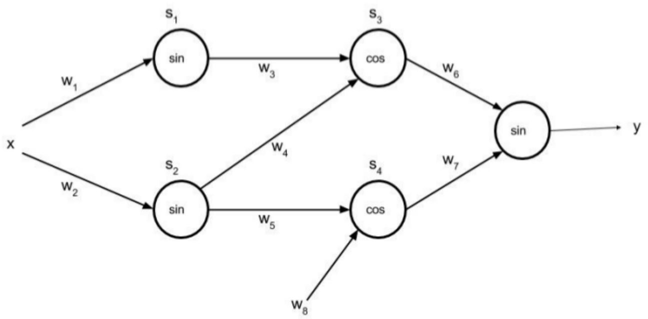
(d) Can we get a better estimate of $f(2.01)$ and $f(2.5)$ by correcting our estimate from part (a)?

Solution:

4. **Differentiation of function of multiple variables**

$$\begin{aligned} s_1 &= sin(w_1 x) \\ s_2 &= sin(w_2 x) \\ s_3 &= cos(w_3 s_1 + w_4 s_2) \\ s_4 &= cos(w_5 s_2 + w_8) \\ y &= sin(w_6 s_3 + w_7 s_4) \end{aligned} \tag{3}$$

An alternative representation of the function y is given in the figure below.



Compute the derivatives $\frac{dy}{dw_1}$ and $\frac{dy}{dw_2}$.

Solution:

5. **Differentiation of vectors**
Consider vectors $\boldsymbol{u}, \boldsymbol{x} \in \mathbb{R}^d$, and matrix $\boldsymbol{A} \in \mathbb{R}^{n \times n}$.
The derivative of a scalar f w.r.t. a vector $\boldsymbol{x}$ is a vector by itself, given by

$$\nabla f = \left(\frac{\partial f}{\partial x_1}, \frac{\partial f}{\partial x_2}, \ldots, \frac{\partial f}{\partial x_n} \right)$$

Derive the expressions for the following derivatives (gradients).

$$\nabla \boldsymbol{u}^{\boldsymbol{T}} \boldsymbol{x} \quad , \quad \nabla \boldsymbol{x}^{\boldsymbol{T}} \boldsymbol{x} \quad \text{and} \quad \nabla \boldsymbol{x}^{\boldsymbol{T}} \boldsymbol{A} \boldsymbol{x}$$

(Aside: Compare your results with derivatives for the scalar equivalents of the above expressions ax and x^2.

The derivative of a scalar f w.r.t. a matrix $\boldsymbol{X}$, is a matrix whose (i, j) component is $\frac{\partial f}{\partial X_{ij}}$, where X_{ij} is the (i, j) component of the matrix $\boldsymbol{X}$.)

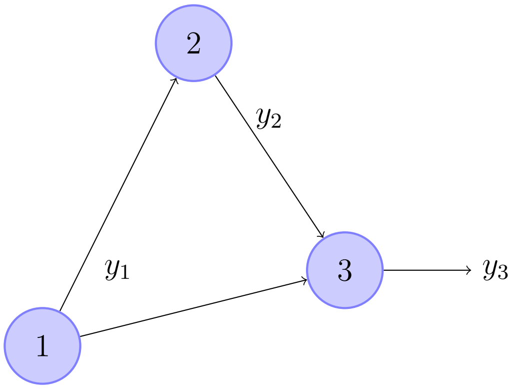
Solution:

6. **Ordered derivatives**

An ordered network is a network where the state variables can be computed one at a time in a specified order.

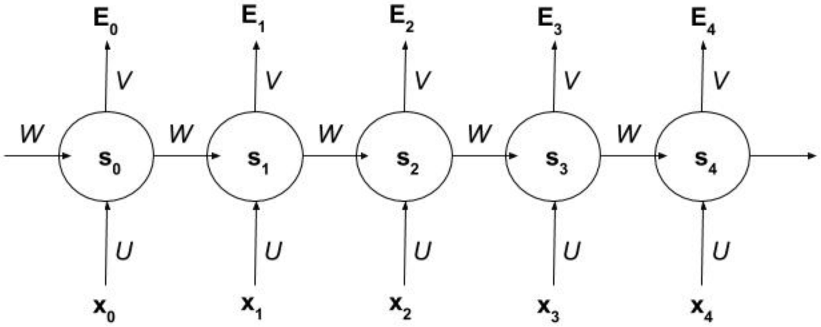
Answer the following questions regarding such a network.

(a) Given the ordered network below, give a formula for calculating the ordered derivative $\frac{\partial y_3}{\partial y_1}$ in terms of partial derivatives w.r.t. y_1 and y_2 where y_1, y_2 and y_3 are the outputs of nodes 1, 2 and 3 respectively.



Solution:

(b) The figure above can be viewed as a dependency graph as it tells us which variables in the system depend on which other variables. For example, we see that y_3 depends on y_1 and y_2 which in turn also depends on y_1. Now consider the network given below,



$$\textit{Here,}\quad s_i = \sigma(Ws_{i-1} + Ux_i + b) \quad (\forall i > 0)$$

Can you draw a dependency graph involving the variables s_4, s_3, s_2, s_1, W?

Solution:

(c) Give a formula for computing $\frac{\partial s_4}{\partial W}$ for the network shown in part (b)

Solution:

7. **Minimizing functions**

From basic calculus, we know that we can find the minima (local and global) of a function by finding the first and second order derivatives. We set the first derivative to zero and verify if the second derivative at the same point is positive. The reasoning behind the following procedure is based on the interpretation of the derivative of a function as the slope of the function at any given point.

The above procedure, even though correct can be intractable in practice while trying to minimize functions. And this is not just a problem for the multivariable case, but even for single variable functions. Consider minimizing the function $f(x) = x^5 + 5sin(x) + 10tan(x)$. Although the function f is a contrived example, the point is that the standard derivative approach, might not always be a feasible way to find minima of functions.

In this course, we will be routinely dealing with minimizing functions of multiple variables (in fact millions of variables). Of course we will not be solving them by hand, but we need a more efficient way of minimizing functions. For the sake of this problem, consider we are trying to minimize a convex function of one variable $f(x)$, [1] which is guaranteed to have a single minima. We will now build an iterative approach to finding the minima of functions.

The high level idea is the following:
Start at a (random) point x_0. Verify if we are at the minima. If not, change the value so that we are moving closer to the minima. Keep repeating until we hit the minima.

(a) Use the intuition built from Q.3 to find a way to change the current value of x while still ensuring that we are improving (i.e. minimizing) the function.

Solution:

(b) How would you use the same idea, if you had to minimize a function of multiple variables ?

Solution:

(c) (Extra) Can you think of the number of steps needed to reach the minima ?

Solution:

[1] `https://en.wikipedia.org/wiki/Convex_function`

(d) (Extra) Do you think this procedure always works ? (*i.e.*, are we always guaranteed to reach the minima)

Solution:

(e) (Extra) Can you think of ways to improve the number of steps needed to reach the minima ?

Solution:

8. Consider two discrete distributions p and q where p is the true distribution (and hence cannot be changed) whereas q is the estimated distribution. Let θ be the parameters of q. One way of estimating the parameters of q is to find a θ which minimizes the KL-divergence between p and q. Show that this is equivalent to finding a θ which minimizes the cross entropy between p and q.

Solution:

9. Consider a large playground filled with 1 million balloons. Of these there are k_1 blue, k_2 green and k_3 red balloons. The values of k_1, k_2 and k_3 are not known to you but you are interested in estimating them. Of course, you cannot go over all the 1 million balloons and count the number of blue, green and red balloons. So you decide to randomly sample 1000 balloons and note down the number of blue, green and red balloons. Let these counts be $\hat{k}_1$, $\hat{k}_2$ and $\hat{k}_3$ respectively. You then estimate the total number of blue, green and red balloons as $1000 * \hat{k}_1$, $1000 * \hat{k}_2$ and $1000 * \hat{k}_3$.

(a) Your friend knows the values of k_1, k_2 and k_3 and wants to see how bad your estimates are compared to the true values. Can you suggest some ways of calculating this difference? [Hint: Think about probability!]

Solution:

(b) Consider two ways of converting $\hat{k}_1$, $\hat{k}_2$ and $\hat{k}_3$ to a probability distribution:

$$p_i = \frac{\hat{k}_i}{\sum_i \hat{k}_i}$$

$$q_i = \frac{e^{\hat{k}_i}}{\sum_i e^{\hat{k}_i}}$$

Why (if at all) would you prefer the distribution $\mathbf{q} = [q_1, q_2, ..., q_n]$ over $\mathbf{p} = [p_1, p_2, ..., p_n]$?

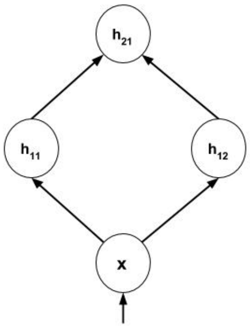
Solution:

10. Plotting functions

(a) Consider the variable x and functions $h_{11}(x)$, $h_{12}(x)$ and $h_{21}(x)$ such that

$$\begin{aligned} h_{11}(x) &= \frac{1}{1+e^{-(400x+24)}} \\ h_{12}(x) &= \frac{1}{1+e^{-(400x-24)}} \\ h_{21} &= h_{11}(x) - h_{12}x) \end{aligned}$$

The above set of functions are summarized in the graph below.



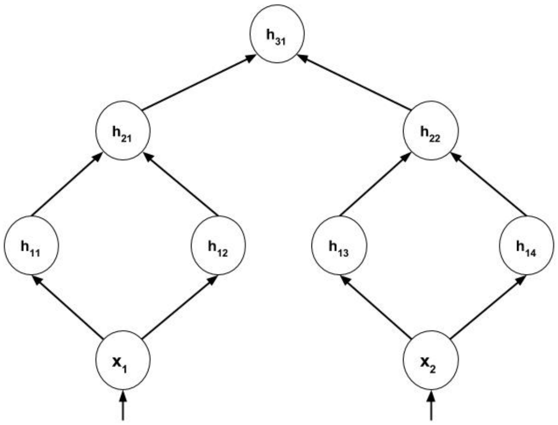
Plot the following functions: $h_{11}(x)$, $h_{12}(x)$ and $h_{21}(x)$ for $x \in (-1, 1)$

Solution:

(b) Now consider the variables x_1, x_2 and the functions $h_{11}(x_1, x_2), h_{12}(x_1, x_2), h_{13}(x_1, x_2), h_{14}(x_1, x_2)$, $h_{21}(x_1, x_2), h_{22}(x_1, x_2), h_{31}(x_1, x_2)$ and $f(x_1, x_2)$ such that

$$
\begin{aligned}
h_{11}(x_1, x_2) &= \frac{1}{1 + e^{-(x_1 + 100x_2 + 200)}} \\
h_{12}(x_1, x_2) &= \frac{1}{1 + e^{-(x_1 + 100x_2 - 200)}} \\
h_{13}(x_1, x_2) &= \frac{1}{1 + e^{-(100x_1 + x_2 + 200)}} \\
h_{14}(x_1, x_2) &= \frac{1}{1 + e^{-(100x_1 + x_2 - 200)}} \\
h_{21}(x_1, x_2) &= h_{11}(x_1, x_2) - h_{12}(x_1, x_2) \\
h_{22}(x_1, x_2) &= h_{13}(x_1, x_2) - h_{14}(x_1, x_2) \\
h_{31}(x_1, x_2) &= h_{21}(x_1, x_2) + h_{22}(x_1, x_2) \\
f(x_1, x_2) &= \frac{1}{1 + e^{-(50h_{31}(x) - 100)}}
\end{aligned}
$$

The above set of functions are summarized in the graph below.



Plot the following functions: $h_{11}(x_1, x_2), h_{12}(x_1, x_2), h_{13}(x_1, x_2), h_{14}(x_1, x_2), h_{21}(x_1, x_2)$, $h_{22}(x_1, x_2), h_{31}(x_1, x_2)$ and $f(x_1, x_2)$ for $x_1 \in (-5, 5)$ and $x_2 \in (-5, 5)$

Solution: