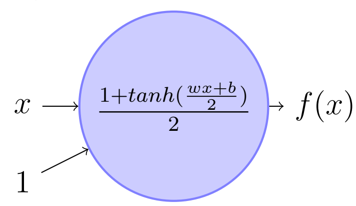
Instructions:

- This assignment is meant to help you grok certain concepts we will use in the course. Please don't copy solutions from any sources.
- Avoid verbosity.
- Questions marked with * are relatively difficult. Don't be discouraged if you cannot solve them right away!
- The assignment needs to be written in latex using the attached tex file. The solution for each question should be written in the solution block in space already provided in the tex file. **Handwritten assignments will not be accepted.**

---

1. **Partial Derivatives**

   (a) Consider the following computation ,



where $f(x) = \frac{1+tanh(\frac{wx+b}{2})}{2}$ and by definiton : $tanh(z) = \frac{e^z-e^{-z}}{e^z+e^{-z}}$

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: We are given $f(x) = \frac{1+tanh(\frac{wx+b}{2})}{2}$, and $tanh(z) = \frac{e^z-e^{-z}}{e^z+e^{-z}}$.
The loss functions L is given by, $L = \frac{1}{2}(y - f(x))^2$. Computing $\frac{\partial L}{\partial w}$ and $\frac{\partial L}{\partial b}$.

Using chain rule: $\frac{\partial L}{\partial w} = (y - f(x))(\frac{\partial(y-f(x))}{\partial w})$
Since y is a constant, the above equation becomes, $\frac{\partial L}{\partial w} = (y - f(x))(\frac{\partial(-f(x))}{\partial w})$

$$\frac{\partial f(x)}{\partial w} = \frac{\partial[\frac{tanh(wx+b)}{2}]}{\partial w}$$

$$\frac{\partial f(x)}{\partial w} = \frac{\partial[\frac{tanh(\alpha)}{2}]}{\partial w}, where \alpha = \frac{wx+b}{2}$$

Using chain rule, $\frac{\partial f(x)}{\partial w} = (\frac{\partial[\frac{tanh(\alpha)}{2}]}{\partial w})(\frac{\partial \alpha}{\partial w})$

$\frac{\partial f(x)}{\partial w} = (\frac{1-tanh^2(\alpha)}{2})(\frac{x}{2}) = (\frac{x}{4})(1 - tanh^2(\frac{wx+b}{2}))$

Substituting the value of $\frac{\partial f(x)}{\partial w}$ into $\frac{\partial L}{\partial w} = (y - f(x))(\frac{\partial(-f(x))}{\partial w})$, we get the following:

$\frac{\partial L}{\partial w} = (y - f(x))(\frac{x}{4})(tanh^2(\frac{wx+b}{2}) - 1)$

$\frac{\partial L}{\partial w} = (y - \frac{1+tanh(\frac{wx+b}{2})}{2})(\frac{x}{4})(tanh^2(\frac{wx+b}{2}) - 1)$

Similarly,

$\frac{\partial L}{\partial b} = (y - \frac{1+tanh(\frac{wx+b}{2})}{2})(\frac{1}{4})(tanh^2(\frac{wx+b}{2}) - 1)$

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

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

Here x, y, z are inputs (constants) and a, b, c, d are parameters (variables). σ is the logistic sigmoid function defined as:

$$\sigma(x) = \frac{1}{1 + e^{-x}}$$

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: The equation of E is given as :

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

where σ is the logistic sigmoid function: $\sigma(x) = \frac{1}{1+e^{-x}}$

Therefore,

$$E = g(x, y, z) = \sigma(c(ax + by) + dz) = \frac{1}{1 + e^{-(c(ax+by)+dz)}}$$

Then the partial derivative of E with respect to the parameter 'a':

$$\frac{\partial E}{\partial a} = \frac{\partial}{\partial m}\frac{1}{(1 + e^- m)}\frac{\partial m}{\partial a}$$

where, $m = (c(ax + by) + dz)$

$$\frac{\partial E}{\partial a} = \sigma(m)(1 - \sigma(m))cx$$
$$= \frac{1}{(1 + e^{-}m)}[1 - \frac{1}{(1 + e^{-}m)}]cx$$
$$= [\frac{1}{(1 + e^{-}m)} - \frac{1}{(1 + e^{-}m)^2}]cx$$
$$= \frac{e^{-}m}{(1 + e^{-}m)^2}cx$$

$$\frac{\partial E}{\partial a} = \frac{e^{-}(c(ax + by) + dz)}{(1 + e^{-}(c(ax + by) + dz))^2}cx$$

Similarily, the partial derivative of E with respect to the parameters 'b' and 'c':

$$\frac{\partial E}{\partial b} = \frac{e^{-}(c(ax + by) + dz)}{(1 + e^{-}(c(ax + by) + dz))^2}cy$$

$$\frac{\partial E}{\partial c} = \frac{e^{-}(c(ax + by) + dz)}{(1 + e^{-}(c(ax + by) + dz))^2}z$$

2. **Erroneous Estimates**

The first order derivative of a real valued function f is defined by the following limit (if it exists),

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

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 $f(x) = x^2 - 2x + 1$.

(a) Using the limit definition of derivative, show that the derivative of $f(x)$ is $\frac{df(x)}{dx} = 2x - 2$.

Solution: Given the limit is : $\frac{df(x)}{dx} = \lim_{h\to 0} \frac{f(x+h)-f(x)}{h}$, with $f(x) = x^2 - 2x + 1$,

To show that $\frac{df(x)}{dx} = 2x - 2$:

We got,

$$\frac{df(x)}{dx} = \lim_{h \to 0} \frac{[(x+h)^2 - 2(x+h) + 1] - [x^2 - 2x + 1]]}{h}$$

$$\frac{df(x)}{dx} = \lim_{h \to 0} \frac{(x^2 + 2xh + h^2 - 2x - 2h + 1) - x^2 + 2x - 1}{h}$$

$$\frac{df(x)}{dx} = \lim_{h \to 0} \frac{2xh + h^2 - 2h}{h}$$

$$\frac{df(x)}{dx} = \lim_{h \to 0} \frac{2x + h - 2}{1}$$

$$\frac{df(x)}{dx} = \lim_{h \to 0} 2x + \lim_{h \to 0} h - \lim_{h \to 0} 2$$

Thus,

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

(b) The function evaluates to 0 at 1 i.e. $f(1) = 0$.

Say we wanted to estimate the value of $f(1.01)$ and $f(1.5)$ without using the definition of $f(x)$. We could think of using the definition of derivative to "extrapolate" the value of $f(1)$ to obtain $f(1.01)$ and $f(1.5)$.

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

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

Estimate $f(1.01)$ and $f(1.5)$ using the above formula.

Solution:

The given equation is:

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

,
Therefore, we got the following:

$$f(1.01) = f(1+0.01) \approx f(1) + 0.01[2(1) - 2]$$

$$f(1.01) \approx 0 + 0.01(2(1) - 2])$$

$$f(1.01) \approx 0$$

Similarly,

$$f(1.5) = f(1+0.5) \approx f(1) + 0.5[2(1) - 2]$$

$$f(1.5) \approx 0 + 0.5(2(1) - 2])$$

$$f(1.5) \approx 0$$

Therefore,

$$f(1.01) = 0 and f(1.5) = 0$$

(c) Compare it to the actual value of $f(1.01) = 0.0001$, and $f(1.5) = 0.25$.

Solution:
Here, Actual value of $f(1.01) = 0.0001$ and $f(1.5) = 0.25$, but the estimated value is 0 for both of them . Still the estimated value for x = 1.01 is somewhat closer to its actual value, but for x = 1.5, there is a large difference in its estimated and actual value.

(d) Explain the discrepancy from the actual value. Why does it increase/decrease when we move further away from 1?

Solution:
At x=1, the value of the function is equal to its first derivative at that point, both of them is 0. This limts the estimated function to estimate the values of the function at x = 1.01 and x = 1.5.
and as we noticed here that the first derivative of the function linearly increases. which is nearly equal to the the value of the function at x = 1

(e) Can we get a better estimate of $f(1.01)$ and $f(1.5)$ by "correcting" our estimate from part (a)? Can you suggest a way of doing this?

Solution: By making small changes in the estimate function in part(b), we can get a better estimate the values of $f(1.01)$ and $f(1.5)$.

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

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

Gradient of a scalar function w.r.t a matrix is a matrix.

$$\nabla_{\mathbf{A}} f = \begin{bmatrix} \frac{\partial f}{\partial A_{11}} & \frac{\partial f}{\partial A_{12}} & \cdots & \frac{\partial f}{\partial A_{1n}} \\ \frac{\partial f}{\partial A_{21}} & \frac{\partial f}{\partial A_{22}} & \cdots & \frac{\partial f}{\partial A_{2n}} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial f}{\partial A_{n1}} & \frac{\partial f}{\partial A_{n2}} & \cdots & \frac{\partial f}{\partial A_{nn}} \end{bmatrix}$$

Gradient of the gradient of a function w.r.t. a vector is a matrix. It is referred to as Hessian.

$$\mathbf{H}_{\mathbf{x}} f = \nabla^2_{\mathbf{x}} f = \begin{bmatrix} \frac{\partial^2 f}{\partial^2 x_1} & \frac{\partial^2 f}{\partial x_1 \partial x_2} & \cdots & \frac{\partial^2 f}{\partial x_1 \partial x_n} \\ \frac{\partial^2 f}{\partial x_2 \partial x_1} & \frac{\partial^2 f}{\partial^2 x_2} & \cdots & \frac{\partial^2 f}{\partial x_2 \partial x_n} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial^2 f}{\partial x_n \partial x_1} & \frac{\partial^2 f}{\partial x_n \partial x_2} & \cdots & \frac{\partial^2 f}{\partial^2 x_n} \end{bmatrix}$$

(a) Derive the expressions for the following gradients.

1. $\nabla_{\mathbf{x}} \boldsymbol{u}^{\boldsymbol{T}} \boldsymbol{x}$
2. $\nabla_{\mathbf{x}} \boldsymbol{x}^{\boldsymbol{T}} \boldsymbol{x}$
3. $\nabla_{\mathbf{x}} \boldsymbol{x}^{\boldsymbol{T}} \boldsymbol{A} \boldsymbol{x}$
4. $\nabla_{\mathbf{A}} \boldsymbol{x}^{\boldsymbol{T}} \boldsymbol{A} \boldsymbol{x}$
5. $\nabla^2_{\mathbf{x}} \boldsymbol{x}^{\boldsymbol{T}} \boldsymbol{A} \boldsymbol{x}$

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

The gradient 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:

1.

$$\begin{aligned}
\nabla_{\mathbf{x}} \boldsymbol{u}^{\boldsymbol{T}} \boldsymbol{x} &= \nabla_{\mathbf{x}} (\sum_i u_i x_i) \\
&= (\frac{\partial \sum_i u_i x_i}{\partial x_1}, \frac{\partial \sum_i u_i x_i}{\partial x_2}, ..., \frac{\partial \sum_i u_i x_i}{\partial x_n}) \\
&= (u_1, u_2, ..., u_n) \\
&= \boldsymbol{u}
\end{aligned}$$

2.

$$\begin{aligned}
\nabla_{\mathbf{x}} \boldsymbol{x}^{\boldsymbol{T}} \boldsymbol{x} &= \nabla_{\mathbf{x}} (\sum_i x_i^2) \\
&= (\frac{\partial \sum_i x_i^2}{\partial x_1}, \frac{\partial \sum_i x_i^2}{\partial x_2}, ..., \frac{\partial \sum_i x_i^2}{\partial x_n}) \\
&= (2x_1, 2x_2, ..., 2x_n) \\
&= 2\boldsymbol{x}
\end{aligned}$$

$$\boldsymbol{x}^{\boldsymbol{T}} \boldsymbol{A} \boldsymbol{x} = \sum_i x_i \sum_j A_{ij} x_j$$

From this we can write its general equation for the term of x_i term

$$(A_{1i}x_1 + A_{i1}x_1) + (A_{2i}x_2 + A_{i2}x_2) + ... + A_{ii}x_i^2 + ... + (A_{in}x_n + A_{ni}x_n)$$

Similarly, for the A_{ij} term

$$x_i x_j$$

3.

$$\begin{aligned}
\nabla_{\mathbf{x}} \boldsymbol{x}^{\boldsymbol{T}} \boldsymbol{A} \boldsymbol{x} &= \nabla_{\mathbf{x}} \sum_i x_i \sum_j A_{ij} x_j \\
&= (..., i^{th} term\ from\ above\ equation, ...) \\
&= (\boldsymbol{A} + \boldsymbol{A}^{\boldsymbol{T}}) \boldsymbol{x}
\end{aligned}$$

4.

$$\begin{aligned}
\nabla_{\mathbf{A}} \boldsymbol{x^T A x} &= \nabla_{\mathbf{A}} \sum_i x_i \sum_j A_{ij} x_j \\
&= \begin{bmatrix} x_1^2 & x_1x_2 & ... & x_1x_n \\ x_2x_1 & x_2^2 & ... & x_2x_n \\ . & . & ... & . \\ . & . & ... & . \\ . & . & ... & . \\ x_nx_1 & x_nx_2 & ... & x_n^2 \end{bmatrix} \\
&= \boldsymbol{xx^T}
\end{aligned}$$

5.

$$\begin{aligned}
\nabla_{\mathbf{x}}^2 \boldsymbol{x^T A x} &= \nabla_{\mathbf{x}}(\nabla_{\mathbf{x}} \boldsymbol{x^T A x}) \\
&= \nabla_{\mathbf{x}}(\boldsymbol{A^T} + \boldsymbol{A})\boldsymbol{x} \\
&= \boldsymbol{A^T} + \boldsymbol{A}
\end{aligned}$$

All these were the derived expression for the given gradients after interpretation.

(b) Use the equations obtained in the previous part to get the Linear regression solution that you studied in ML or PR. Suppose X as input example-feature matrix, Y as given outputs and $\mathbf{w}$ as weight vector.

Solution: From the previous part of the question we got the following equations:

$$(Y - Xw)^T(Y - Xw) = Y^TY - Y^TXw - w^TX^TY + w^TX^TXw$$

$$\frac{\partial}{\partial w}(Y^TY - Y^TXw - w^TX^TY + w^TX^TXw) = 0$$

$$-X^TY - X^TY + 2X^TXw = 0$$

$$2X^TXw = 2X^TY$$

$$X^TXw = X^TY$$

$$w = (X^TX)^{-1}X^TY$$

(c) By now you must have the intuition. Gradient w.r.t. a 1 dimensional array was 1 dimensional. Gradient w.r.t. a 2-dimensional array was 2 dimensional. Higher order arrays are referred to as tensors. Let T be a 3 dimensional tensor. Write the expression of $\nabla_{\mathbf{T}} f$. You can use gradients w.r.t. a vector or a matrix in the expression.

Solution:

$$\begin{bmatrix} \nabla_{\mathbf{T}} f_1 & \nabla_{\mathbf{T}} f_2 & \dots & \nabla_{\mathbf{T}} f_n \end{bmatrix}$$

in z direction.

where,

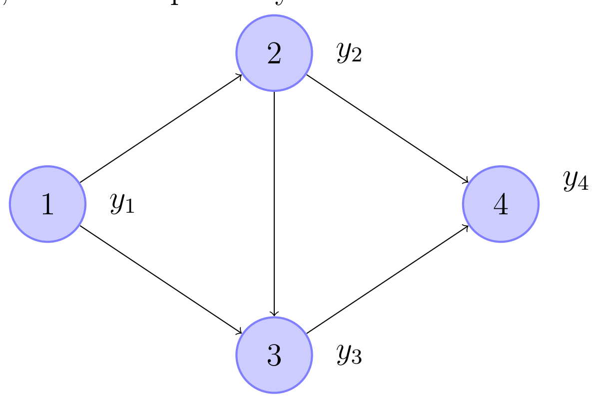
$$\nabla_{\mathbf{T}} f_1 = \begin{bmatrix} \frac{\partial f}{\partial T_{111}} & \frac{\partial f}{\partial T_{121}} & \cdots & \frac{\partial f}{\partial T_{1n1}} \\ \frac{\partial f}{\partial T_{211}} & \frac{\partial f}{\partial T_{221}} & \cdots & \frac{\partial f}{\partial T_{2n1}} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial f}{\partial T_{n11}} & \frac{\partial f}{\partial T_{n21}} & \cdots & \frac{\partial f}{\partial T_{nn1}} \end{bmatrix}$$

4. **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.

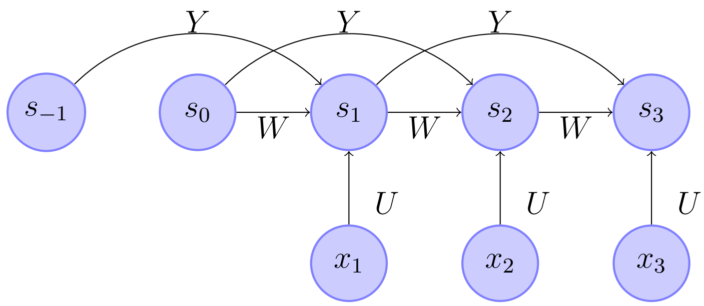
   (a) Given the ordered network below, give a formula for calculating the ordered derivative $\frac{\partial y_4}{\partial y_1}$ in terms of partial derivatives w.r.t. y_1 and y_2 where y_1, y_2 and y_3 are the outputs of nodes 1, 2 and 3 respectively.



Solution:

$$\frac{\partial y_4}{\partial y_1} = \frac{\partial y_4}{\partial y_2}\frac{\partial y_2}{\partial y_1} + \frac{\partial y_4}{\partial y_3}\frac{\partial y_3}{\partial y_1} + \frac{\partial y_4}{\partial y_3}\frac{\partial y_3}{\partial y_2}\frac{\partial y_2}{\partial y_1}$$

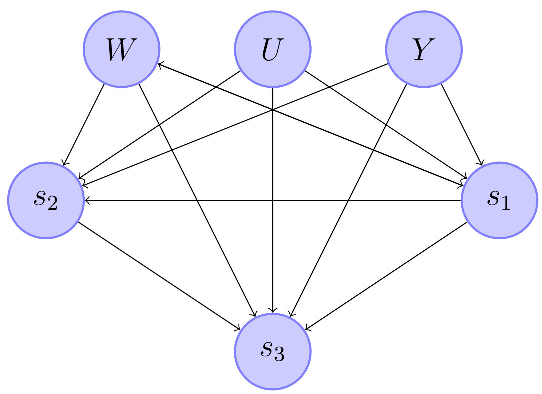
   (b) The figure above can be viewed as a dependency graph as it tells us which variables in the system depend on which other variables. For example, we see that y_3 depends on y_1 and y_2 which in turn also depends on y_1. Now consider the network given below,



$$\textit{Here,}\quad s_i = \sigma(Ws_{i-1} + Ys_{i-2} + Ux_i + b) \quad (\forall i \geq 1).$$

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

Solution:



$$\begin{aligned} \textit{Here,}\quad s_3 &= \sigma(Ws_2 + Ys_1 + Ux_3 + b) \\ s_2 &= \sigma(Ws_1 + Ys_0 + Ux_2 + b) \\ s_1 &= \sigma(Ws_0 + Ys_{-1} + Ux_1 + b) \end{aligned}$$

(c) Give a formula for computing $\frac{\partial s_3}{\partial W}$, $\frac{\partial s_3}{\partial Y}$ and $\frac{\partial s_3}{\partial U}$ for the network shown in part (b)

Solution: Formulas for computing the $\frac{\partial s_3}{\partial W}$, $\frac{\partial s_3}{\partial Y}$ and $\frac{\partial s_3}{\partial U}$ for the networks are :

$$\frac{\partial s_3}{\partial W} = \frac{\partial s_3}{\partial W} + \frac{\partial s_3}{\partial s_2}\frac{\partial s_2}{\partial W} + \frac{\partial s_3}{\partial s_1}\frac{\partial s_1}{\partial W} + \frac{\partial s_3}{\partial s_2}\frac{\partial s_2}{\partial s_1}\frac{\partial s_1}{\partial W}$$

$$\frac{\partial s_3}{\partial Y} = \frac{\partial s_3}{\partial Y} + \frac{\partial s_3}{\partial s_2}\frac{\partial s_2}{\partial Y} + \frac{\partial s_3}{\partial s_1}\frac{\partial s_1}{\partial Y} + \frac{\partial s_3}{\partial s_2}\frac{\partial s_2}{\partial s_1}\frac{\partial s_1}{\partial Y}$$

$$\frac{\partial s_3}{\partial U} = \frac{\partial s_3}{\partial U} + \frac{\partial s_3}{\partial s_2}\frac{\partial s_2}{\partial U} + \frac{\partial s_3}{\partial s_1}\frac{\partial s_1}{\partial U} + \frac{\partial s_3}{\partial s_2}\frac{\partial s_2}{\partial s_1}\frac{\partial s_1}{\partial U}$$

5. **Baby Steps**

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) Does your method always lead to the global minima (smallest value) for non convex functions (which may have multiple local minima)? If yes, can you explain (prove or argue) why? If not, can you give a concrete example of a case where it fails?

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

Solution:

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

Solution:

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

Solution:

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

Solution:

6. **Constrained Optimization** Let $f(x, y)$ and $g(x, y)$ be smooth (continuous, differentiable etc.) real valued functions of two variables. We want to minimize f, which is a convex function of x and y.

(a) Argue that at the minima of f, the partial derivative $\frac{\partial f}{\partial x}$ and $\frac{\partial f}{\partial y}$ will be zero. Thus setting the partial derivatives to zero is a possible method for finding the minima.

Solution: Let (x^m, y^m) be a minimum of f(x,y)
Let us consider

$$f'(x^a, y^a) > 0$$

$$\frac{\partial f(x, y)}{\partial x} > 0$$

and

$$\frac{\partial f(x, y)}{\partial y} > 0$$

$$f(x^a - h, y^a) = f(x^a, y^a) - h\frac{\partial f}{\partial x}|_(x^a, y^a)$$

Thus,

$$f(x^a - h, y^a) < f(x^a, y^a)$$

, which indicates that the original assumption is wrong.

Similarly,

$$f(x^a y^a - h) < f(x^a, y^a)$$

this also indicates that the original assumption is wrong.
Therefore,

$$\frac{\partial f}{\partial x}, \frac{\partial f}{\partial y} < 0$$

Therefore,

$$\frac{\partial f}{\partial x} = \frac{\partial f}{\partial y} = 0$$

, at minima.

(b) Suppose we are only interested in minimizing f in the region where $g(x, y) = c$, where c is some constant. Suppose this region is a curve in the x-y plane. Call this the feasible curve. Will our previous technique still work in this case? Why or why not?

Solution:

(c) What is the component of ∇g along the feasible curve, when computed at points lying on the curve?

Solution:

(d) * At the point on the feasible curve, which achieves minimum value of f, what will be the component of ∇f along the curve?

Solution:

(e) Using the previous answers, show that at the point on the feasible curve, achieving minimum value of f, $\nabla f = \lambda \nabla g$ for some real number λ. Thus, this equation, combined with the constraint $\nabla g = 0$ should enable us to find the minima.

Solution:

(f) * Using the insights from discussion so far, solve the the following optimization problem:

$$\max_{x,y,z} x^a y^b z^c$$

where

$$x + y + z = 1$$

and given $a, b, c > 0$.

Solution: A function $f(x, y, z)$ is given as:

$$f(x, y, z) = x^a y^b z^c$$

where the constraints are:

$$x + y + z = 1$$

and

$$a, b, c > 0$$

To solve this optimization problem we need to maximize the given function $f(x, y, z)$ at x, y, z:

$$\frac{\partial f}{\partial x} = ax^{a-1} y^b z^c = 0$$

$$\frac{\partial f}{\partial y} = bx^a y^{b-1} z^c = 0$$

$$\frac{\partial f}{\partial z} = cx^a y^b z^{c-1} = 0$$

Now equating :

$$\frac{\partial f}{\partial x} = \frac{\partial f}{\partial y}$$

and

$$\frac{\partial f}{\partial y} = \frac{\partial f}{\partial z}$$

we got the values of x and z:

$$x = \frac{Ya}{b}$$

$$z = \frac{Yc}{b}$$

Now using the above values of x and z we calculated y from the given constraint $x + y + z = 1$:

$$y = \frac{b}{a + b + c}$$

Similarly,

$$x = \frac{a}{a + b + c}$$

and

$$z = \frac{c}{a+b+c}$$

7. **Billions of Balloons**
Consider a large playground filled with 1 billion 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 billion 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 $1000000 * \hat{k}_1$, $1000000 * \hat{k}_2$ and $1000000 * \hat{k}_3$.

   (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:

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

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

   Would you prefer the distribution $\mathbf{q} = [q_1, q_2, ..., q_n]$ over $\mathbf{p} = [p_1, p_2, ..., p_n]$ for the above task? Give reasons and provide an example to support your choice.

   Solution:

8. ** Let X be a real-valued random variable with p as its probability density function (PDF). We define the cumulative density function (CDF) of X as

$$F(x) = \Pr(X \leq x) = \int_{y=-\infty}^{y=x} p(y)dy$$

What is the value of $\mathbb{E}_X[F(X)]$ (the expected value of the CDF of X)? **The answer is a real number** (Hint: The expectation can be formulated as a double integral. Try to plot the area over which you need to integrate in the x-y plane. Now look at the area over which you are not integrating. Do you notice any symmetries?)

Solution: Given CDF of X is: $F(x) = \Pr(X \leq x) = \int_{y=-\infty}^{y=x} p(y)dy$

The expected value of $F(x)$,

$$\mathbb{E}_X[F(X)] = \int_{y=-\infty}^{y=+\infty} F(x)p(y)dx$$

As we know that the differentiation of cumulative density function is the probability density function:

$$p(y) = \frac{dF(x)}{dx}$$

Therefore, $p(y)dx = dF(x)$

Now the expected value is :

$$\mathbb{E}_X[F(X)] = \int_{y=0}^{y=1} F(x)dF(x) = \left[\frac{[F(x)]^2}{2}\right]_0^1 = \frac{1}{2}$$

9. * **Intuitive Urns**

An urn initially contains 3 red balls and 3 blue balls. One of the balls is removed without being observed. To find out the color of the removed ball, Alice and Bob independently perform the same experiment: they randomly draw a ball, record the color, and put it back. This is repeated several times and the number of red and blue balls observed by each of them is recorded.

Alice draws 6 times and observes 6 red balls and 0 blue balls.

Bob draws 600 times and observes 303 red balls and 297 blue balls.

Obviously, both of them will predict that the removed ball was blue.

(a) Intuitively, who do you think has stronger evidence for claiming that the removed ball was blue, and why? (**Don't cheat by computing the answer. This subquestion has no marks, but is compulsory!**)

Solution:

(b) What is the exact probability that the removed ball was blue, given Alice's observations? (Hint: Think Bayesian Probability)

Solution:
A : Event that the ball removed was blue
X : Alice's observation
From Bayes' theorem,

$$\Pr(A|X) = \frac{\Pr(X|A)\Pr(A)}{\Pr(X)}$$

Since there are equal number of balls in the beginning, $\Pr(A) = \frac{1}{2}$
Since there are only two possible outcomes, we can model Alice's observation as a Binomial random variable with $n = 6$ and p as probability of picking a blue ball.

$$\Pr(X) = \frac{1}{2}\binom{6}{0}(\frac{2}{5})^6 + \frac{1}{2}\binom{6}{0}(\frac{3}{5})^6$$
$$\Pr(X|A) = \binom{6}{0}(\frac{3}{5})^6$$
$$\Pr(A|X) = \frac{3^6}{3^6 + 2^6}$$

(c) What is the exact probability that the removed ball was blue, given Bob's observations? (Hint: Think Bayesian Probability)

Solution:
A : Event that the ball removed was blue
Y : Bob's observation
From Bayes' theorem,

$$\Pr(A|Y) = \frac{\Pr(Y|A)\Pr(A)}{\Pr(Y)}$$

Since there are equal number of balls in the beginning, $\Pr(A) = \frac{1}{2}$
Since there are only two possible outcomes, we can model Bob's observation as a Binomial random variable with $n = 600$ and p as probability of picking a blue ball.

$$\Pr(Y) = \frac{1}{2}\binom{600}{297}(\frac{2}{5})^{297}(\frac{3}{5})^{303} + \frac{1}{2}\binom{600}{297}(\frac{3}{5})^{297}(\frac{2}{5})^{303}$$
$$\Pr(Y|A) = \binom{600}{297}(\frac{3}{5})^{303}(\frac{2}{5})^{297}$$
$$\Pr(A|Y) = \frac{3^6}{3^6 + 2^6}$$

(d) Computationally, who do you think has stronger evidence for claiming that the removed ball was blue?

Solution:

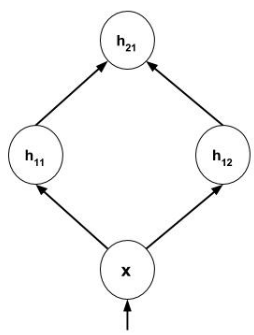
Did your intuition match up with the computations? If yes, awesome! If not, remember that probability can often be seem deceptively straightforward. Try to avoid intuition when dealing with probability by grounding it in formalism.

10. **Plotting Functions for Great Good**

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

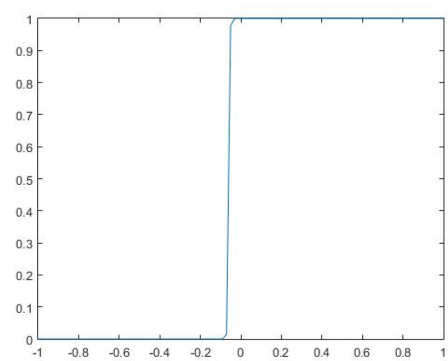
$$\begin{aligned} h_{11}(x) &= \frac{1}{1+e^{-(400x+24)}} \\ h_{12}(x) &= \frac{1}{1+e^{-(400x-24)}} \\ h_{21} &= h_{11}(x) - h_{12}(x) \end{aligned}$$

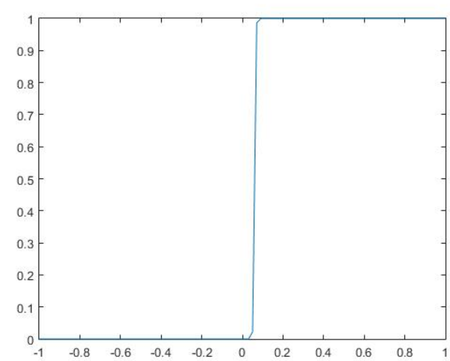
The above set of functions are summarized in the graph below.

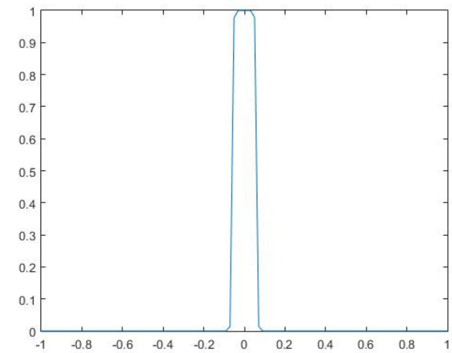


Plot the following functions: $h_{11}(x)$, $h_{12}(x)$ and $h_{21}(x)$ for $x \in (-1, 1)$

Solution:



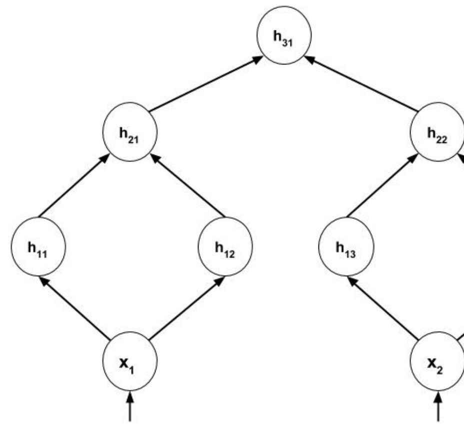




(b) Now consider the variables x_1, x_2 and the functions $h_{11}(x_1, x_2), h_{12}(x_1, x_2), h_{13}(x_1, x_2), h_{14}(x_1, x_2)$, $h_{21}(x_1, x_2), h_{22}(x_1, x_2), h_{31}(x_1, x_2)$ and $f(x_1, x_2)$ such that

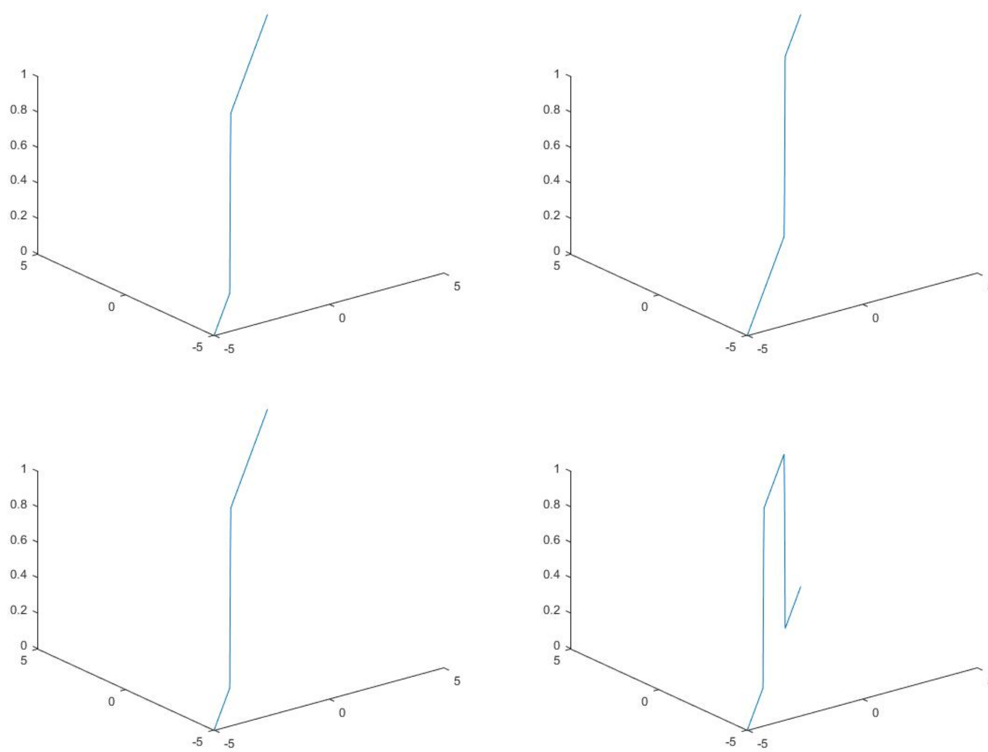
$$\begin{aligned}
h_{11}(x_1, x_2) &= \frac{1}{1 + e^{-(x_1 + 100x_2 + 200)}} \\
h_{12}(x_1, x_2) &= \frac{1}{1 + e^{-(x_1 + 100x_2 - 200)}} \\
h_{13}(x_1, x_2) &= \frac{1}{1 + e^{-(100x_1 + x_2 + 200)}} \\
h_{14}(x_1, x_2) &= \frac{1}{1 + e^{-(100x_1 + x_2 - 200)}} \\
h_{21}(x_1, x_2) &= h_{11}(x_1,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

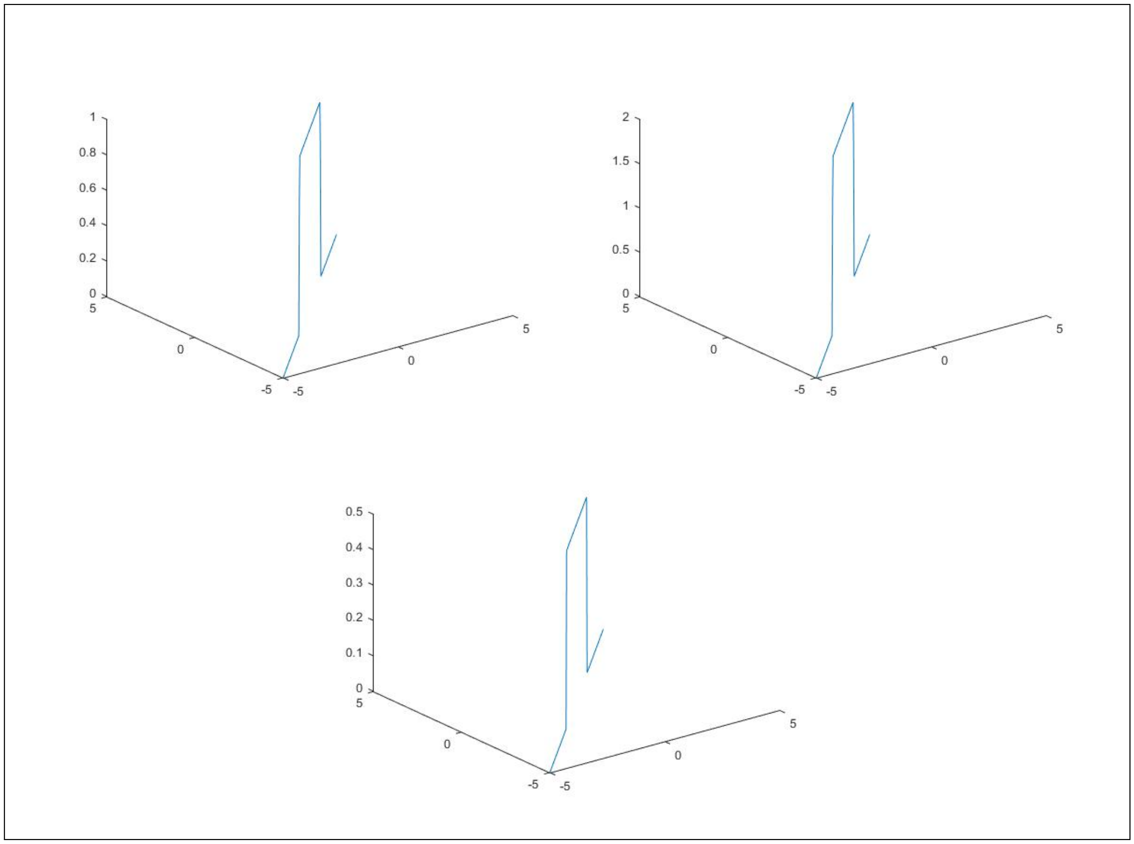
The above set of functions are summarized in the graph below.



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:




1
0.8
0.6
0.4
0.2
0
5
0
-5
-5
0
5
2
1.5
1
0.5
0
5
0
-5
-5
0
5
0.5
0.4
0.3
0.2
0.1
0
5
0
-5
-5
0
5