Mathematical Minimum - Final

Note : these are **math exercises** (nothing programming related) that will help you succeeding in this course and beyond (Computational Data Analysis..etc)

Fun Begins

Ex 1 : Sums and Products

Sums :

1. $\sum_{i=0}^{N} 1 =$
2. $2 \cdot \sum_{k=1}^{K} \sum_{t=1}^{T} 1 =$
3. $\sum_{k=1}^{K} \sum_{t=1}^{T} 0.5^k =$
4. $\sum_{k=1}^{\infty} \sum_{t=1}^{T} 0.5^k =$

Products :

The notation $\prod_{i=1}^{N} p_i$ denotes the product with N factors:

$$\prod_{i=1}^{N} p_i = p_1 p_2 \cdots p_N$$

1. $\prod_{i=1}^{M} \frac{1}{\theta} =$
2. $\prod_{k=1}^{K} \frac{k}{k+1} =$
3. $\ln\left(\prod_{k=1}^{K} e^k\right) =$

If you struggle for Ex 1, watch these :

Arithmetic & Geometric series : https://www.khanacademy.org/math/precalculus

Notes on notations :

- Geometric summation starting from i=0 or i=1 will only have different numerator for finite sum, with one have power to N and one to N+1. How about infinite sum of the geometric series.. What is the difference in the numerator ?
  For geometric summation the index is not important. It is more the first term and the number of terms. The infinite sum is just the limit when K goes to infinity of the finite version.

- You can think of K and T as some given constants whereas k is a dummy variable. https://en.wikipedia.org/wiki/1/2_%2B_1/4_%2B_1/8_%2B_1/16_%2B_%E2%8B%AF

Ex 2 : Function Properties

For each of the following functions f(x) below :

Find its limits $\lim_{x\to\pm\infty} f(x)$ as x approachs $\pm\infty$.

Choose the values of x where $f(x)$ is differentiable, i.e. $f'(x)$ exists

Choose the values of x where $f(x)$ is also strictly increasing, i.e. $f'(x) > 0$

For $f(x) = \max(0, x)$
(lf the limit diverges to infty, enter inf for ∞, and -inf for $-\infty$

$$\lim_{x\to-\infty} f(x) =$$
$$\lim_{x\to+\infty} f(x) =$$

Choose the intervals of x where

f(x) differentiable: **f'(x)>0 :**

(The left column is for "f(x) differentiable", and the right one is where "f'(x)>0.)

$x < 0$	$\square x < 0$
$x = 0$	$\square x = 0$
$x > 0$	$\square x > 0$

(Graph this function on a piece of paper!)

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

(Enter inf for ∞ and similarly -inf for $-\infty$ if the limit diverges to infty.)

$$\lim_{x\to-\infty} f(x) =$$
$$\lim_{x\to+\infty} f(x) =$$

Choose the intervals of x where
f(x) differentiable: f'(x)>0

$x < 0$	$\square x < 0$
$x = 0$	$\square x = 0$
$x > 0$	$\square x > 0$

Notes on functions : you can use https://www.wolframalpha.com/ or https://www.derivative-calculator.net to calculate and plot derivatives !

- The domain of $f(x) = \max(0, x)$ is the entire real line, i.e. $(-\infty, \infty)$. It asks what the function value is when $x \to -\infty$

- The condition of max (0,x) means y=0 for x<0. So what is the slope for x<=0?
  A continuous function would be an unsegmented straight line or unbroken curve. Maybe another way to say it is that the formula that describes its behavior continues across all x. In this case the formula for x<0 is f(x)=0 and the formula for x>0 is f(x)=x. y=0 does not hold for all x. A continuous function would imply that f(x)=0 over all x, but that's not the case.

- why f(x)=max(0,x) isn't differentiable at x>0
  If this function is differentible at x=0, then the left-hand limit and the right-hand limit of the derivative of f(x) at x=0 to must be the same. If you draw a graph of this function, you can see that the graph is not "smooth" at x=0. If a function is differentiable, the its graph should be smooth.

- If a function is not differentiable at certain point x0 , then f'(x0) does not exist. This means you cannot make any statement about f'(x) at x0 , which naturally eliminates itself from answering questions like what x makes f'(x)>0

- A function which is differentiable and strictly increasing is not necessarily has positive derivative at all points: take for example f(x)=x^3, which is strictly increasing and differentiable at all points, but has derivative 0 at the origin

Ex 3 : Points and Vectors

A list of n numbers can be thought of as a point or a vector in n -dimensional space. we will think of

n- dimensional vectors $\begin{bmatrix} a_1 \\ a_2 \\ \vdots \\ a_n \end{bmatrix}$ flexibly as points and as vectors.

Recall the dot product of a pair of vectors a and b :

$a \cdot b = a_1 b_1 + a_2 b_2 + \cdots + a_n b_n$ where $a = \begin{bmatrix} a_1 \\ a_2 \\ \vdots \\ a_n \end{bmatrix}$ and $b = \begin{bmatrix} b_1 \\ b_2 \\ \vdots \\ b_n \end{bmatrix}$

When thinking about a and b as vectors in n -dimensional space, we can also express the dot product as

$a \cdot b = \| a \| \| b \| \cos \alpha$

where α is the angle formed between the vectors a and b in n-dimensional Euclidean space.
Here, ‖a‖ refers to the length, also known as **norm**, of a:

$$\| a \| = \sqrt{a_1^2 + a_2^2 + \cdots + a_n^2}$$

What is the length of the vector $\begin{bmatrix} 0.4 \\ 0.3 \end{bmatrix}$?

What is the length of the vector $\begin{bmatrix} -0.15 \\ 0.2 \end{bmatrix}$?

What is the angle (in radians) between $\begin{bmatrix} 0.4 \\ 0.3 \end{bmatrix}$ and $\begin{bmatrix} -0.15 \\ 0.2 \end{bmatrix}$?
?

Choose the answer that lies between 0 and π.

Given 3-dimensional vectors$x^{(1)} = \begin{bmatrix} a_1 \\ a_2 \\ a_3 \end{bmatrix}$ and $x^{(2)} = \begin{bmatrix} a_1 \\ -a_2 \\ a_3 \end{bmatrix}$, when is $x^{(1)}$ orthogonal to $x^{(2)}$ the angle between them is $\frac{pi}{2}$?

when $2a_1 + 2a_3 = 0$
when $a_1^2 - a_2^2 + a_3^2 = 0$
when $a_1^2 + a_2^2 + a_3^2 = 0$

A unit vector is a vector with length 1. The length of a vector is also called its norm. Given any vector x, write down the unit vector pointing in the same direction as x?

Recall from linear algebra the definition of the projection of one vector onto another. As before, we have 3-dimensional vectors $x^{(1)} = \begin{bmatrix} a_1 \\ a_2 \\ a_3 \end{bmatrix}$ and $x^{(2)} = \begin{bmatrix} a_1 \\ -a_2 \\ a_3 \end{bmatrix}$

Which of these vectors is in the same direction as the projection of x(1) onto x(2)?

$x^{(1)}$
$x^{(2)}$
$x^{(1)} + x^{(2)}$

What is the signed magnitude c of the projection px(1)→x(2) of x(1) onto x(2)? More precisely, let u be the unit vector in the direction of the correct choice above, find a number c such that px(1)→x(2)=cu.

Notes :

- Check this for the last question http://sites.science.oregonstate.edu/math/home/programs/undergrad/CalculusQuestStudyGuides/vcalc/dotprod/dotprod.html

- On vector projections :

Let's be a and b two (not necessary unit) vectors. We want to compute the vector c being the projection of a on b and its $1-2$ norm (or length).
Let's start from the lenght. We know from a well-known trigonometric equation that
$\|c\| = \|a\| * \cos(\alpha)$, where α is the angle between the two vectors a and b
But we also know that the dot product $a \cdot b$ is equal to $\|a\| * \|b\| * \cos(\alpha)$
By substitution we find that $\|c\| = \frac{a \cdot b}{\|b\|}$. This quantity is also called the component of a in the direction of b.
To find the vector c we now simply multiply $\|c\|$ by the unit vector in the direction of b, $\frac{b}{\|b\|}$, obtaining $c = \frac{a \cdot b}{\|b\|^2} * b$
If b is already a unit vector, the above equations reduce to:
$\|c\| = a \cdot b$ and $c = (a \cdot b) * b$

If you reading the https://en.wikipedia.org/wiki/Vector_projectionVector Projection article on Wikipedia, and you may be confused by the different formulas that all seem to be describing the same thing. In the opening paragraph, it says that the scalar projection is defined as:

$$a_1 = \| \mathbf{a} \| \cos\theta = \mathbf{a} \cdot \hat{\mathbf{b}} = \mathbf{a} \cdot \frac{\mathbf{b}}{\| \mathbf{b} \|}$$

However, in the section "Definitions in terms of a and b: Vector Projection", it says that the definition of vector projection of a onto b is:

$$\mathbf{a}_1 = \frac{\mathbf{a} \cdot \mathbf{b}}{\| \mathbf{b} \|^2} \mathbf{b} = \frac{\mathbf{a} \cdot \mathbf{b}}{\mathbf{b} \cdot \mathbf{b}} \mathbf{b}$$

The first one is a scalar projection, meaning it is only a number. The second is a vector projection, which represents the projected vector. a_1 is the length of a1.

Ex 4 : Probability Density Functions

BEFORE DOING THE NEXT EXERCICES, READ THESE :

https://ermongroup.github.io/cs228-notes/preliminaries/probabilityreview/

https://tutorial.math.lamar.edu/classes/calcii/probability.aspx

And watch these (-20 min) :

https://www.youtube.com/watch?v=QKA4HNEw3aY

https://www.khanacademy.org/math/statistics-probability/random-variables-stats-library/random-variables-continuous/v/probability-density-functions

Let X be a **continuous** random variable with probability **density** function (pdf) $f_X(x)$

- Is the value of $f_X(x)$ always $\in [0,1]$?

  Yes/No

- For $a < b, \int_a^b f_X(x)dx \in$ $[0,1]$ and represents the probability that the value of X falls between a and b

  Yes/No

- Is the value of $f_X(x)$ always non-negative?

  Yes/No

- The value of integral $\int_{-\infty}^{\infty} f_X(x)dx$ of $f_X(x)$ from $-\infty$ to ∞ is a finite, undetermined value.

  Yes/No

Notes :

- Note that PDF values are **probability density**, not **probability.** On top of this, "... probability is given by the integral of this variable's PDF over that range—that is, it is given by the area under the density function but above the horizontal axis and between the lowest and greatest values of the range. The probability density function is nonnegative everywhere, and its integral over the entire space is equal to 1." Read more here : https://en.wikipedia.org/wiki/Probability_density_function

- The fourth question might be unclear. Providing $f_X(x)$ is valid PDF, you know what this integral is $\int_{-\infty}^{\infty} f_X(x)dx$ integrated to.

Example :

PMT9

Recall that the probability density of a *normal* or *Gaussian* distribution with mean μ and variance σ^2 is,

$$g(x) \equiv \frac{1}{\sigma\sqrt{2\pi}} \exp\left[-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right]$$

While σ^2 denotes the variance, the *standard deviation* is σ. You may assume $\sigma > 0$.

In statistics, one technique to fit a function to data is a procedure known as *maximum likelihood estimation (MLE)*. At the heart of this method, one needs to calculate a special function known as the *likelihood function*, or just the *likelihood*. Here is how it is defined :

Let $x_0, x_1, \ldots, x_{n-1}$ denote a set of nn input data points. The likelihood of these data, $L(x_0, \ldots, x_{n-1})$ is defined to be

$$L(x_0, \ldots, x_{n-1}) \equiv \prod_{k=0}^{n-1} p(x_i)$$

Ex 5: Univariate Gaussians

Before doing these take a look at :

- For quantile qs - https://www.youtube.com/watch?v=TzKeCv4S7nY
- For normal distri qs
  - https://web.stanford.edu/class/archive/cs/cs109/cs109.1178/lectureHandouts/110-normal-distribution.pdf
- For last qs, calculator. READ the qs carefully. Standard deviation is not the same as variance
  - http://onlnestatbook.com/2/calculators/normal_dist.html (from Joseph_Yiin)
- https://ocw.mit.edu/courses/mathematics/18-05-introduction-to-probability-and-statistics-spring-2014/readings/MIT18_05S14_Reading6a.pdf

A univariate Gaussian or normal distributions can be completely determined by its mean and variance.

Gaussian distributions can be applied to a large numbers of problems because of the central limit theorem (CLT). The CLT posits that when a large number of independent and identically distributed ((i.i.d.) random variables are added, the cumulative distribution function (cdf) of their sum is approximated by the cdf of a normal distribution.

Recall the probability density function of the univariate Gaussian with mean μ and variance $\sigma^2, \mathcal{N}(\mu, \sigma^2)$

$$f_X(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-(x-\mu)^2/(2\sigma^2)}$$

Probability review: PDF of Gaussian distribution

In practice, it is not often that you will need to work directly with the probability density function (pdf) of Gaussian variables. Nonetheless, we will make sure we know how to manipulate the (pdf) in the next two problems.

The pdf of a Gaussian random variable X is given by

$$f_X(x) = \frac{n}{3\sqrt{2\pi}} \exp\left(-\frac{n^2(x-2)^2}{18}\right)$$

then what is the mean μ and variance σ^2 of X?
μ =

σ^2 =

Let $X \sim \mathcal{N}(\mu, \sigma^2)$, i.e. the pdf of X is

$$f_X(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)$$

Let Y=2X. Write down the pdf of the random variable Y. (Your answer should be in terms of y, σ and μ.)

fY(y)=

Argmax

Let $f_X(x; \mu, \sigma^2)$ denote the probability density function of a normally distributed variable X with mean μ and variance σ^2 What value of x maximizes this function?

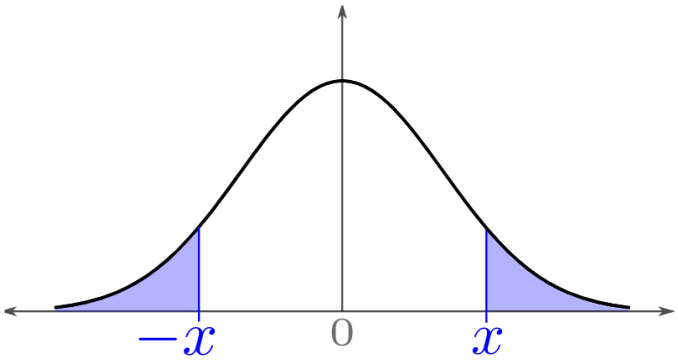
Maximum of pdf

As above, let $f_X(x; \mu, \sigma^2)$denote the probability density function of a normally distributed variable X with mean μ and variance σ^2

What is the maximum value of $f_X(x; \mu, \sigma^2)$

Quantile :

The quantile of order $1-\alpha$ of a variable X, denoted by q_α i X)
(specific to a particular X), is the number such that $\mathbf{P}(X \leq q_\alpha) = 1-\alpha$



The total area of the two shaded regions is 0.03.

P(|X|≤0.03)

- P(|X|≤0.015)
- 0.97
- 0.985
- q0.03
- q0.015

Let $X \sim \mathcal{N}(1,2)$, i.e., the random variable X is normally distributed with mean 1 and variance 2. What is the probability that $X \in [0.5,2]$?

Notes :

- There is no need to do integration in any of the problems. Recall what the Normal distribution is parametrized by and then do some pattern matching on the provided PDF.

- How to find the maximum of a normal PDF?
  One method would be to take find the derivative of the pdf and set it equal to zero. This gives you a critical point and then you can look at the sign of the derivative for other values and see if it is the maximum.

- On Quantiles. The curve is a probability density, it integrates to one (100%). If they say the blue area is 0.03 (3%), what is the probability that a random variable distributed like that falls into that blue region? x is the number that limits that slices that region. For example, if it was a standard normal, and the blue are integrated to 5%, x would result in 1.96, meaning that the probability of a r.v. X being less than -1.96 or higher than 1.96 is 5%. We call 1.96 the 2.5% quantile, given its a two tailed test. Use this to fiddle around
  http://onlinestatbook.com/2/calculators/normal_dist.html

- Notice the definition of qα is defined so that the **right** tail has probability α. This is different from how the quantile function is defined in python or R

- For the $Y=2X$ question, Note that the random variable is Y not X. So you should use Y instead of X. What kind of distribution does Y follow given by the expression Y = 2X? What is the description of the distribution? (i.e. parameters). Question 2 is actually quite easy to answer, given that X is a gaussian variable (mu, sigma). This, provided you went through a Probability course or read material regarding normal distributions somewhere else.
  You could also start with a CDF, convert it to a known CDF, then convert it to a PDF. You could, but there is no need to do something that sophisticated. Knowing that X is a gaussian variable of mean μ and variance σ2 has a direct implication on the nature of the distribution of Y=2X and the values of Y's mean and variance. Just remember that a linear combination of a gaussian is still a gaussian, and that if you multiply a random variable by a constant, its expected value is multiplied accordingly and its variance is multiplied by the square of that constant.

- For the last question you can use scipy :

```
from scipy.stats import norm
X = norm(3, scale=sqrt(9.))
X.cdf(10)
```

Ex 6 : Polynomials

Recall a degree n polynomial in x1,x2,…,xk are all linear combinations of monomials in x1,x2,…,xk , where monimials in x1,x2,…,xk are unordered words using x1, x^2,…, xk as the letters.

A degree 2, also known as quadratic, polynomial in the 1 variable x is of the form :

$$ax^2 + bx + c$$

for some numbers a,b,c . The polynomial is determined by the 3 coefficients a,b,c , and different choices of (a,b,c) result in different polynomials.
In linear algebraic terms, the space of degree 2 polynomials in 1 variable is of dimension 3 since it consists of all linear combinations of 3 linearly independent vectors x^2,x, and 1 .

A degree 2 polynomial in 2 variables x1,x2 is of the form :

$$ax_1^2 + bx_2^2 + cx_1x_2 + dx_1 + ex_2 + f$$

or some numbers a,b,c,d,e,f . Different choices of (a,b,c,d,e,f) result in different polynomials.
In linear algebraic terms, the space of degree 2 polynomials in 2 variables is of dimension 6 since it consists of all linear combinations of 6 linearly independent vectors $x_1^2, x_2^2, x_1x_2, x_1, x_2$, and 1

Consider degree 2 polynomials in 3 variables x1,x2,x3 . How many coefficients are needed to completely determine such a polynomial? Equivalently, what is the dimension of the space of polynomials in 3 variables such polynomials?

Notes :

- Here you want to collect all the ways you can combine the variables to create terms of degree ≤ n. Looking at example number 2, we have degree =n=2. We collect the following terms

$n = 2: ax_1^2 = ax_1x_1, bx_2^2 = bx_2x_2, cx_1x_2$
$n = 1: dx_1, ex_2$
$n = 0: f$

The equation for the given polynomial is the sum of all of these terms. I know that's not a resource so much as my own process for solving the problem, but hopefully it helps a bit. See Khan Academy videos on polynomials.

- Distributing the degree across the variables is analogous to distributing balls among urns, which is a classical combinatorial setting.
  https://artofproblemsolving.com/wiki/index.php/Ball-and-urn will give you a good intro. Be careful with the mapping of variables between the exercise and their explanation, in particular, consider how to account for unused parts of the degree in terms like x_1 or the constant term.
  Check also these :
  https://math.stackexchange.com/questions/2705636/formula-for-discriminant-of-a-polynomial-of-degree-2-in-3-variables
  More intuitive : https://en.wikipedia.org/wiki/Triangular_number

- What is dimension of the polynomials of degree N in K variables? - proof versus heuristic patterning ?

  By writing out cases up to N=3 and K=4, you can arrive at a simple combinatoric formula that matches all the cases, including the given examples and the graded exercise. What are the ways to show that it is true for arbitrary N and K?

  There are a few ways to do this. One is to first restrict yourself to looking at degree-j polynomials for certain j, and notice that there is a constraint on the exponents of the variables. You can count the number of satisfying combinations using the stars and bars technique. Then you can add these up for all degree possibilities j, yielding a combinatorial sum which you may or may not recognize (you can simplify the sum by writing out a few terms and guessing the sum+using induction, or maybe a combinatorial argument, but I'm not aware how to do that). The wikipedia page for "complete homogeneous symmetric polynomial" also has the answer hidden in a lot of math and notation. But the answer is nice in terms of binomial coefficients.

  There is also a very clever way to do it that bypasses the need to do a summation of binomial coefficients. The beauty of this way is that you don't have to look at degree-j polynomials one at a time and sum them up, you can count all the possibilities right from the beginning.

EX 7 Planes

A hyperplane in n dimensions is a $n-1$ dimensional subspace. For instance, a hyperplane in 2-dimensional space can be any line in that space and a hyperplane in 3-dimensional space can be any plane in that space. A hyperplane separates a space into two sides.

In general, a hyperplane in n-dimensional space can be written as $\theta_0 + \theta_1 x_1 + \theta_2 x_2 + \cdots + \theta_n x_n = 0$

For example, a hyperplane in two dimensions, which is a line, can be expressed as

$Ax_1 + Bx_2 + C = 0$

Using this representation of a plane, we can define a plane given an n-dimensional vector

$\boldsymbol{\theta} = \begin{bmatrix} \theta_1 \\ \theta_2 \\ \vdots \\ \theta_n \end{bmatrix}$ and offset θ_0. This vector and offset combination would define the

plane $\theta_0+\theta_1 x_1+\theta_2 x_2+\cdots+\theta_n x_n=0$. One feature of this representation is that the vector θ is normal to the plane.

Given a d-dimensional vector θ and a scalar offset θ_0 which describe a hyperplane $P:\ \theta\cdot x+\theta_0=0$. How many alternative descriptions θ' and θ'_0 are there for this plane P?

- ☐ 0
- ☐ 1
- ☐ ∞

Orthogonality Check

To check if a vector x is orthogonal to a plane P characterized by θ and θ_0, we check whether

$1. x = \alpha\theta \text{ for some } \alpha \in \mathbb{R}$
$2. x \cdot \theta = 0$
$3. x \cdot \theta + \theta_0 = 0$

Perpendicular Distance to Plane

Given a point x in n-dimensional space and a hyperplane described by θ and θ_0, find the **signed distance between the hyperplane and x**. This is equal to the perpendicular distance between the hyperplane and x, and is positive when x is on the same side of the plane as θ points and negative when x is on the opposite side.
(Enter **theta_0** for the offset θ_0.
Enter **norm(theta)** for the norm $\|\theta\|$ of a vector θ.
Use ***** to denote the dot product of two vectors, e.g. enter **v*w** for the dot product $v \cdot w$ of the vectors v and w.)

Orthogonal Projection onto Plane

Find an expression for the **orthogonal projection** of a point v onto a plane P that is characterized by θ and θ_0. Write your answer in terms of v, θ and θ_0.
(Enter **theta_0** for the offset θ_0.
Enter **norm(theta)** for the norm $\|\theta\|$ of a vector θ.
Use ***** to denote the dot product of two vectors, e.g. enter **v*w** for the dot product $v \cdot w$ of the vectors v and w.)

Perpendicular Distance to Plane

Let P_1 be the hyperplane consisting of the set of points $x \begin{bmatrix} x_1 \\ x_2 \end{bmatrix}$ for which $3x_1+x_2-1=0$. (Note that this hyperplane is in fact a line, since it is 1-dimensional.)
What is the signed perpendicular distance of point $a=[-1,-1]$ from P_1?

What is the signed perpendicular distance of the origin from P_1?

What is the orthogonal projection of point $a=[-1,-1]$ onto p_1?

First coordinate:

Second coordinate:

Consider a hyperplane in a d-dimensional space. If we project a point onto the plane, can we recover the original point from this projection?

Notes :

- If you are lost read these first :

  https://www.math.kit.edu/ianm2/lehre/am22016s/media/distance-harvard.pdf

  http://cms.ysu.edu/sites/default/files/documents/mathematics-assistance-center/Calculus%203_Cheat_Sheet_Reduced.pdf

  https://math.stackexchange.com/questions/1755856/calculate-arbitrary-points-from-a-plane-equation

  http://faculty.bard.edu/belk/math213/PlanesAndHyperplanes.pdf

- Links for projection of a plane https://www.maplesoft.com/support/help/Maple/view.aspx?path=MathApps/ProjectionOfVectorOntoPlane there is a good explanaition about the "Projection of a Vector onto a Plane"

- $\hat{\theta}$ is the unit vector along θ and is equal to $\frac{\theta}{\|\hat{\theta}\|}$

- The equivalent solution without using a specific symbol for the unit vector is then: $\overrightarrow{v_p} = \vec{v} - \frac{\vec{v}\cdot\vec{\theta}+\theta_0}{\|\vec{\theta}\|} * \frac{\vec{\theta}}{\|\theta\|} = \vec{v} - \vec{\theta} * \frac{\vec{v}\cdot\vec{\theta}+\theta_0}{\|\vec{\theta}\|^2}$

- The hyper-plane will be a sub-space if C and θ_0 are equal to zero. Otherwise it will NOT contain the null-vector thereby, it won't be a sub-space. https://math.stackexchange.com/questions/1380886/generalization-of-lines-and-planes/1381493#1381493

- How can we find all points which belongs to the plane ? It is enough to take any one point x0 which belongs to the plane and a vector Theta perpendicular to the plane. Indeed if we draw a vector from x0 to any point on the plane x, that is x-x0, it will be perpendicular to the vector Theta. Al points x satisfying those conditions will actually form the plane. Formally (x-x0)*Theta*=||*x-x0*||||Theta||*cos(alpha) = 0, that is x*Theta-x0*Theta=0, lets x0*Theta be Theta0, so we have x*Theta+Theta0=0. Now the question is how many points x0 and how many different vectors Theta can we take for the plane still be orthogonal to Theta and contain the point x0?

- A plane is defined by any initial point on the plane and any vector Perpendicular to that plane. How many such points and vectors can we find for exactly the same plane?

- Orthogonal projection of a point v on plane. Should you use transpose ? Definitely no need to transpose! The only vector operations is scalar multiplication (dot) and plus/minus. All dimensions of all vectors are identical. Am I right to understand that Transpose comes from matrix (tensor) multiplication, where we want to consider scalar product of two vectors as a particular case of arbitrary dimension tensor multiplication? That is what we would write in Python. But in this context just plain kindergarden dot product notation is used defined for two vectors without transpose.

- What does the orthogonal projection of a point onto a plane mean? Lets represent points (p the projection, and x the original point) in the space with vectors from the origin: pv,xv ,then: $p_v: (p_v, x_v, \theta \in V, \theta_0 \in F \mid \theta \cdot p_v + \theta_0 = 0, \exists s \in F: x_v - p_v = s\theta)$ Orthogonal projection of a point A onto a plane π is a point B∈π closest to point A (that is, the vector AB→ would be orthogonal to the plane π). Orthogonal projection of a point v to a plane is a point v0 on the plane such as the vector between points v and v0 is orthogonal to the plane. Just drop the point to the plane in the direction perpendicular to the plane Check https://math.stackexchange.com/questions/2874812/orthogonal-projection-of-a-point-to-plane and https://www.khanacademy.org/math/linear-algebra/alternate-bases/orthogonal-projections/v/linear-alg-visualizing-a-projection-onto-a-plane

- Orthogonal Projection onto Plane : how this theta representation on a plane and proofs to get the projection of a point on the plane ? Check https://ocw.mit.edu/courses/mathematics/18-06sc-linear-algebra-fall-2011/least-squares-determinants-and-eigenvalues/projections-onto-subspaces/MIT18_06SCF11_Ses2.2sum.pdf and https://textbooks.math.gatech.edu/ila/projections.html and https://mathinsight.org/distance_point_plane#distance_formula_2

- what is Offset of a vector : Offset is in regard to a (hyper)plane not to a vector. It is the θ_0 in the representation $\theta \cdot x + \theta_0 = 0$.

- Perpendicular Distance to Plane : They ask for the signed perpendicular distance. Don't forget the sign in your answers. Check https://www.youtube.com/watch?v=HW3LYLLc60I&t=385s

- Number of Representations problem :
  It is known from analytical geometry fact, that d-dimensional vector θ identifies indefinitely many parallel hyperplanes in d-dimensional space, while pair of d-dimensional vector θ and a certain value θ_0 uniquely identify a single hyperplane in d-dimensional space. As it has been stated before, the equation of the plane is $\theta_0+\theta_1 x_1+\theta_2 x_2+\cdots+\theta_n x_n=0$.
  For any value $a \neq 0$ it is possible to re-define a plane $a \cdot \theta_0+a \cdot \theta_1 x_1+a \cdot \theta_2 x_2+\cdots+a \cdot \theta_n x_n=0$, this new equation defines the same plane, but with scaled coefficients $\eta_i=a \cdot \theta_i$. Some people treat equations $\theta_0+\theta_1 x_1+\theta_2 x_2+\cdots+\theta_n x_n=0$ and $\eta_0+\eta_1 x_1+\eta_2 x_2+\cdots+\eta_n x_n=0$ as different representations of the same plane, some treat them as the same representation, arguing that second representation is the linear combination of the first. Somehow, it is more treated as convention within group.

  So,

  $$\theta_0+\theta_1 x_1+\theta_2 x_2+\cdots+\theta_n x_n=0$$

  and

  $$a \cdot \theta_0+a \cdot \theta_1 x_1+a \cdot \theta_2 x_2+\cdots+a \cdot \theta_n x_n=0$$

  are the same representations of the same hyperplane, or they are different?
  The hyperplane is indeed unique once θ and θ_0 are fixed. Note that the problem is asking for the number of *alternative descriptions* of θ, meaning for any $\theta' \neq \theta$, they are considered different.

- For the question : “Consider a hyperplane in a d -dimensional space. If we project a point onto the plane, can we recover the original point from this projection?". If we assume orthogonal projection then for d -dimensional space we have d inhomogeneous linear algebraic equations from which all d components of the original point can be computed, right? the issue is that every point P in the hyperplane is the orthogonal projection of infinitely many points, namely those on the line perpendicular to the plane through P.

If we only know P (and pretend we have forgotten how we got there!!), then we have no way of knowing which point on the line we came from.

- How to multiply a scalar and a vector in standard notation?
  https://www.geogebra.org/classic/2d is very useful for visualising
  https://mathworld.wolfram.com/Point-LineDistance2-Dimensional.html is good to understand the 2-D version before considering the 3-D or N-D planes

- **One nice thing about linear algebra is that for many situations, one can get a decent intuition of what's going on by looking at a simpler case. In this case, imagine the point is just some point, p in $\mathbb{R}^3$ given by $\begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix}$ and the plane is just the xy-plane. A perpendicular projection onto the plane will yield a new point given by $\begin{bmatrix} x_1 \\ x_2 \\ 0 \end{bmatrix}$ You've taken something described by three coordinates and essentially reduced it to something described (in a new 2-D space) by only 2 coordinates. Note, if you didn't use a perpendicular projection you would just end up with a new point given by something like $\begin{bmatrix} x'_1 \\ x'_2 \\ 0 \end{bmatrix}$**

- "A projection of a vector onto a plane is on the plane, meaning it satisfies the equation that defines the plane" shouldn't this be true only for planes passing through the origin? Because I think in any arbitrary plane a perpendicular vector to plane's normal vector can only satisfy the plane equation for theta_0 = 0. Any vector on a plane is perpendicular to this plane's normal vector, but a vector completely on a plane (and perpendicular to plane's normal vector) will not satisfy the plane's equation necessarily. Yes, there is some ambiguity here. Let's assume there are two vectors v and u that satisfies the equation. If you view the vector representation as point coordinates, then the points are on the plane. On the other hand, the vector v−u is also on/parallel to the plane and v−u⊥normal. Again check this
  https://math.stackexchange.com/questions/2874812/orthogonal-projection-of-a-point-to-plane

-

Ex 8 : Matrices and Vectors

(Chances that there are matrices in MT1 are very low but the subject is VITAL beyond MT1, so this just an appetizer) :

Let $\mathbf{A} = \begin{bmatrix} 1 & 2 & 3 \\ 4 & 5 & 6 \\ 1 & 2 & 1 \end{bmatrix}$

Let $g = [2 \quad 1 \quad 3]$
Can we compute $a\mathbf{A}$?

Let g and A be as above. Can we compute Ag?

Let $\mathbf{B} = \begin{bmatrix} 2 & 1 & 0 \\ 1 & 4 & 4 \\ 5 & 6 & 4 \end{bmatrix}$

Determine the rank of B. Recall that the rank of a matrix is the number of linearly independent rows or columns.

Let $\mathbf{M}^{-1}$ denote the inverse of a matrix M. Let A be as defined above. Compute $\mathbf{A}^{-1}$. What matrix does the product $\mathbf{A}\mathbf{A}^{-1}$produce?

Notes :

- The matrix A defined above is invertible, you could try to compute it via python or R.
- **How do quickly verify linear independence ?**
  For instance in Python you can use numpy with **`np.np.linalg.matrix_rank(A)`**), which is the simplest "mental model" to recognise linear independence in matrix rows/columns. However See there are others methods. First see if you can find a row or column that can be obtained by a linear combination of other columns.
  For example if you had the matrix $\begin{pmatrix} 1 & 2 & 3 \\ 1 & 1 & 2 \\ 3 & 4 & 7 \end{pmatrix}$

  you could (with some checking) notice that the last row is the sum of the first row and twice the second row (or that the third column is the sum of the first two columns) and thus the row or column vectors aren't independent. If the pattern isn't obvious with just short visual inspection, you could do it formally by setting up a system of equations: $\alpha[1,2,3] + \beta[1,1,2] + \gamma[3,4,7] = 0$. If the only possible solution is $\alpha=\beta=\gamma=0$ then the vectors are independent, otherwise they're dependent (and in this case we have a solution $\alpha=1$, $\beta=2$, $\gamma=1$ so the vectors are dependent).

Another way to test for independence, which only works if you have a square matrix, is to compute the determinant of the matrix. If the determinant is 0 the vectors are dependent, otherwise they're independent.

Another method, in case it's helpful, is to reduce the matrix B to it's row echelon form (pretty much an upper triangular matrix with non zero entries on the diagonal). Once you've completed this task you count the pivots, which are the leftmost nonzero entry in each row where this entry has a 0 below it. The number of pivots equal the rank of the matrix.

This method is slower but can be used to get a definitive answer when you are having trouble confirming the existence or non-existence of linearly dependent rows/columns.

There are 3 pivots in the matrix below and it has a rank of 3.

$$\begin{pmatrix} 1 & 2 & 3 & 4 \\ 0 & 3 & 4 & 5 \\ 0 & 0 & 5 & 3 \\ 0 & 0 & 0 & 0 \end{pmatrix}$$

Check UT Austin's LAFF linear algebra course, this is one of the best LA courses available !

Notes :

If you are lost about matrix multiplication

https://www.mathsisfun.com/algebra/matrix-multiplying.html

EX 9 Matrix Multiplication

Let $\mathbf{A} = \begin{pmatrix} 1 & -1 & 2 \\ 0 & 3 & -4 \end{pmatrix}$ and let $\mathbf{B} = \begin{pmatrix} -1 & 0 & 0 \\ 2 & 0 & 1 \\ 0 & 1 & 3 \end{pmatrix}$. The dimensions of the product $\mathbf{AB}$ are:

☐ Rows * ☐ columns

More generally, let $\mathbf{A}$ be an $m \times n$ matrix and $\mathbf{B}$ be an $n \times k$ matrix. What is the size of $\mathbf{AB}$?

☐ Rows * ☐ columns

In addition, it C is a $k \times \jmath$ matrix, what is the size of $\mathbf{ABC}$?

☐ Rows * ☐ columns

Vector Product :

Suppose $\mathbf{u} = \begin{pmatrix} 1 \\ 3 \end{pmatrix}$ and $\mathbf{v} = \begin{pmatrix} -1 \\ 1 \end{pmatrix}$. The product $\mathbf{u}^T\mathbf{v}$ evaluates the **inner product** (also called the **dot product**) of u and v, which evaluates to

$\mathbf{u}^T\mathbf{v}$ = ?

The inner product of u and v is sometimes written as $\langle \mathrm{u}, \mathrm{v} \rangle$.

Vector Outer product :

Suppose $\mathbf{u} = \begin{pmatrix} 1 \\ 3 \end{pmatrix}$ and $\mathbf{v} = \begin{pmatrix} -1 \\ 1 \end{pmatrix}$. The product $\mathbf{uv}^T$ valuates the **outer product** of u and v, which is a 2×2 matrix in this case.

What is $(\mathbf{uv}^T)_{1,1}$?

What is $(\mathbf{uv}^T)_{1,2}$?

What is $(\mathbf{uv}^T)_{2,1}$?

What is $(\mathbf{uv}^T)_{2,2}$?

Notes :

- The outer product of vectors u and v is a matrix. The indices stand for the elements of this matrix. $\mathbf{uv}^T = \begin{pmatrix} (\mathbf{uv}^T)_{1,1} & (\mathbf{uv}^T)_{1,2} \\ (\mathbf{uv}^T)_{2,1} & (\mathbf{uv}^T)_{2,2} \end{pmatrix}$

EX 10 Linear Independence, Subspaces and Dimension

Vectors $v_1,\ldots,v_n$ are said to be **linearly dependent** if there exist scalars $c_1,\ldots,c_n$ such that (1) not all c_i's are zero and (2) $c_1v_1+\cdots+c_nv_n=0$. Otherwise, they are said to be **linearly independent** : the only scalars $c_1,\ldots,c_n$ that satisfy $c_1v_1+\cdots+c_nv_n=0$ are $c_1=\cdots=c_n=0$.
The collection of non-zero vectors $v_1,\ldots,v_n\in R^m$ determines a **subspace** of R^m, which is the set of all linear combinations $c_1v_1+\cdots+c_nv_n$ over different choices of $c_1,\ldots,c_n\in R$. The **dimension** of this subspace is the size of the **largest possible, linearly independent** sub-collection of the (non-zero) vectors $v_1,\ldots,v_n$.

Row and Column Rank

Suppose $\mathbf{A}=\begin{pmatrix}1 & 3\\ 2 & 6\end{pmatrix}$ The rows of the matrix, $(1,3)$ and $(2,6)$, span a subspace of dimension

[ ] Is the row rank of A

The columns of the matrix $\begin{pmatrix}1\\2\end{pmatrix}$ and $\begin{pmatrix}3\\6\end{pmatrix}$ span a subspace of dimension

[ ] Is the column rank of A

Rank of a matrix

In general, row rank is always equal to the column rank, so we simply refer to this common value as the **rank** of a matrix.

What is the largest possible rank of a 2×2 matrix?

What is the largest possible rank of a 5×2 matrix?

In general, what is the largest possible rank of an $m\times n$ matrix?

m ?

n ?

min (m,n) ?

max(m,n)?

Find the rank for these matrices :

$$\begin{pmatrix} \mathbf{1} & \mathbf{1} \\ \mathbf{1} & \mathbf{1} \end{pmatrix}?$$

$$\begin{pmatrix} 1 & -1 \\ 1 & 0 \end{pmatrix}?$$

$$\begin{pmatrix} 0 & 0 \\ 0 & 0 \end{pmatrix}?$$

$$\begin{pmatrix} 1 & 1 & 1 \\ 1 & 0 & 1 \\ 0 & 1 & 0 \end{pmatrix}?$$

$$\begin{pmatrix} 1 & 1 & 0 \\ 0 & -3 & 2 \\ 0 & 0 & 1 \end{pmatrix}?$$

Invertibility of a matrix

An $n \times n$ matrix $\mathbf{A}$ is invertible if and only if $\mathbf{A}$ has full rank, i.e. $\text{rank}(\mathbf{A}) = n$.

Which of the following matrices are invertible? Choose all that apply.

$$\mathbf{A} = \begin{pmatrix} 1 & 3 \\ 2 & 6 \end{pmatrix}$$
$$\mathbf{B} = \begin{pmatrix} 1 & 2 \\ 2 & 1 \end{pmatrix}$$
$$\mathbf{C} = \begin{pmatrix} 1 & 1 & 0 \\ 0 & 1 & 1 \\ 0 & 0 & 1 \end{pmatrix}$$
$$\mathbf{D} = \begin{pmatrix} 2 & -1 & -1 \\ -1 & 2 & -1 \\ -1 & -1 & 2 \end{pmatrix}$$

Notes :

if you are lost about ranks check this
https://en.wikipedia.org/wiki/Linear_independence https://people.math.osu.edu/costin.9/264H/Rank

EX 11 : Determinant

Given a matrix, $\mathbf{A}$, we denote its transpose as $\mathbf{A}^T$The transpose of a matrix is equivalent to writing its rows as columns, or its columns as rows. Then $\mathbf{A}^T_{i,j} = \mathbf{A}_{j,i}$

Recall that the **determinant** $\det(\mathrm{A})$ of a square matrix A indicates whether it is invertible. For 2×2 matrices, it has the formula

$$\det\begin{pmatrix} a & b \\ c & d \end{pmatrix} = ad - bc$$

For larger matrices, the formula is a bit more complicated.

Compute the Determinant

Let $\mathbf{A} = \begin{bmatrix} 1 & 2 & 3 \\ 4 & 5 & 6 \\ 1 & 2 & 1 \end{bmatrix}$

Compute $\det(\mathbf{A}^T)$.

Compute $\det(\mathbf{A})$

Note :

- If you are lost : https://www.youtube.com/watch?v=OyFBYQUFgKI
- Formula to find determinants of any n x n matrix :

The simplest "a bit more complicated" formula to compute the determinants of a generic $\mathrm{n}\times\mathrm{n}$ matrix is the Laplace Expansion, where the determinants are computed recursively summing up the elements of any row or column, each multiplied by its *cofactor* (the determinant of the sub-matrix obtained removing the element row and column multiplied by the sign given by $(-1)^{\mathrm{i+j}}$, where i and j are the element row and column index).
Each cofactor is then computed using the same method, until arriving to the basic formula for the 2×2 matrix reported.

While simple this method is not efficient at all and is indeed much more computationally expensive than other methods.

To find the determinant in Python:

Python: `np.linalg.det(A)`

EX 12 : Eigenvalues, Eigenvectors and Determinants

Eigenvalues and Eigenvectors of a matrix

Let $\mathbf{A} = \begin{pmatrix} 3 & 0 \\ 1 & 2 \\ \frac{1}{2} & \end{pmatrix}$, $\mathbf{v} = \begin{pmatrix} 2 \\ 1 \end{pmatrix}$ and $\mathbf{w} = \begin{pmatrix} 0 \\ 1 \end{pmatrix}$

$\mathbf{Av} = \lambda_1 \mathbf{v}$, where $\lambda_1 =$

$\mathbf{Aw} = \lambda_2 \mathbf{w}$, where $\lambda_2 =$

herefore, v is an eigenvector of A with eigenvalue $\lambda 1$, and w is an eigenvector of A with eigenvalue $\lambda 2$.

Geometric Interpretation of Eigenvalues and Eigenvectors

Let $\mathbf{A} = \begin{pmatrix} 3 & 0 \\ 1 & 2 \\ \frac{1}{2} & \end{pmatrix}, \mathbf{v} = \begin{pmatrix} 2 \\ 1 \end{pmatrix}$ and $\mathbf{w} = \begin{pmatrix} 0 \\ 1 \end{pmatrix}$ Recall from the previous exercise that v and w are eigenvectors of A.

Suppose $\mathbf{x} = \mathbf{v} + 2\mathbf{w} = \begin{pmatrix} 2 \\ 3 \end{pmatrix}$. Then $\mathbf{Ax} = s\mathbf{v} + t\mathbf{w}$, where:

s =

And

t =

In particular, s describes the amount that A stretches x in the direction of v, and $t2$ (note the “2" in front of w in x) describes the amount that A stretches x in the direction of w.

Determinant and Eigenvalues

What is the determinant of the matrix A $\mathbf{A} = \begin{pmatrix} 3 & 0 \\ \frac{1}{2} & 2 \end{pmatrix}$?

On the other hand, what is the product of the eigenvalues λ_1, λ_2 of A? (We already computed this in the previous exercises.)

Trace and Eigenvalues

Recall that the **trace** of a matrix is the sum of the diagonal entries.

What is the trace of the matrix $\mathbf{A} = \begin{pmatrix} 3 & 0 \\ \frac{1}{2} & 2 \end{pmatrix}$?

On the other hand, what is the sum of the eigenvalues λ_1, λ_2 of A? (We already computed this in the previous exercises.)

Nullspace

If a (nonzero) vector is in the nullspace of a square matrix A, is it an eigenvector of A?

Which of the following are equivalent to the statement that 0 is an eigenvalue for a given square matrix A? (Choose all that apply.)

There exists a nonzero solution to $Av=0$.

- [ ] $\det(A)=0$
- [ ] $\det(A)\neq 0$
- [ ] $NS(A)=0$
- [ ] $NS(A)\neq 0$

Notes :

- Check blue brown video for eigen stuff
  https://www.youtube.com/watch?v=PFDu9oVAE-g

- If a (nonzero) vector is in the nullspace of a square matrix A, is it an eigenvector of A ?

A vector in null space can qualify as eigenvector only when corresponding eigenvalue =0.. to satisfy **A.V=lambda.V=0.V=0** (also the definition of a vector in null space). Is it wrong ?

f $v\neq 0$ and $Av=0v$, by definition $\lambda=0$ is an eigenvalue of matrix A. v is also in the null space of A.

- NS(A)=0 meaning
  zero vector with the dimension of the column of A.

- What does NS(A) mean? null space of A. (means all vectors B such that AB=0 and B is not 0.

EX 13 Gradient and Optimization

Multivariable Calculus Review: Simple Gradient

Let :

$$f: \quad \mathbb{R}^d \to \mathbb{R}$$
$$\theta = \begin{pmatrix} \theta_1 \\ \theta_2 \\ \vdots \\ \theta_d \end{pmatrix} \mapsto f(\theta)$$

denote a **differentiable** function. The **gradient** of f is the vector-valued function

$$\nabla_\theta f: \mathbb{R}^d \to \mathbb{R}^d$$
$$\theta = \begin{pmatrix} \theta_1 \\ \theta_2 \\ \vdots \\ \theta_d \end{pmatrix} \mapsto \left. \begin{pmatrix} \frac{\partial f}{\partial \theta_1} \\ \frac{\partial f}{\partial \theta_2} \\ \vdots \\ \frac{\partial f}{\partial \theta_d} \end{pmatrix} \right|_\theta$$

Consider

$f(\theta) = \theta_1^2 + \theta_2^2$

Compute the gradient ∇f

Geometric Picture of the Function

As above, consider $f(\theta) = \theta_1^2 + \theta_2^2$. Let us visualize $f(\theta)$ as a surface on the (θ_1, θ_2) -plane. We will use the usual horizonal plane as the (θ_1, θ_2) -plane, and the vertical axis as the $f(\theta)$ -axis.

What is the shapes of such a curve? (parabola, curve..etc)

Consider how the level curves $\theta_1^2 + \theta_2^2 = +K$ change as K increases from 0 to ∞ .
Does the graph (surface) of $f(\theta)$ have a global maximum, or global minimum, or neither?

At each point $\theta = (\theta_1, \theta_2)$ in the (θ_1, θ_2) -plane, $f(\theta)$ decreases in the direction of....

$$\nabla_\theta f(\theta)$$

Or

$$-\nabla_\theta f(\theta)$$

Gradient ascent/descent methods are typical tools for maximizing/minimizing functions.
Consider the function $L(x, \theta)$ where $\theta = [\theta_1, \theta_2, \ldots, \theta_n]^T$ and $x = [x_1, x_2, \ldots, x_n]^T$
Our goal is to select θ such to maximize/minimize the value of L while keeping x fixed.

Compute the Gradient

The gradient $\nabla_\theta L(x, \theta)$ is a vector with n components:

$$\nabla_\theta L(x, \theta) = \begin{pmatrix} \frac{\partial}{\partial \theta_1} L(x, \theta) \\ \vdots \\ \frac{\partial}{\partial \theta_n} L(x, \theta) \end{pmatrix}$$

(Note that we are treating x as a constant and also differentiating w.r.t. to θ.)

Evaluate the gradient $\nabla_\theta L(x, \theta)$

Which of the following is its j^{th} component?

$$\frac{\exp(-\theta \cdot x)}{1+\exp(-\theta \cdot x)}$$

$$\frac{-x_j \exp(-\theta \cdot x)}{1+\exp(-\theta \cdot x)}$$

$$\frac{-x_j}{1+\exp(-\theta \cdot x)}$$

Gradient Ascent or Descent

The direction of the derivative of a function gives us the direction of the largest change in the function as the independent variables vary.

In gradient ascent/descent methods, we make an educated guess about the next values of θ, with consecutive updates that will hopefully eventually converge to the global minimum of $L(x,\theta)$ (if it exists).

If

$$\theta' = \theta + \epsilon \cdot \nabla_\theta L(x, \theta)$$

where ϵ is a small positive real number, Which of the following is true?

$L(x, \theta') > L(x, \theta)$

$L(x, \theta') < L(x, \theta)$

Notes :

- if you are lost about gradients check this https://betterexplained.com/articles/vector-calculus-understanding-the-gradient/

- https://www.khanacademy.org/math/multivariable-calculus/multivariable-derivatives/partial-derivative-and-gradient-articles/a/the-gradient
- https://www.symbolab.com/solver/gradient-calculator/gradient%20x%5E%7B2%7D%2By%5E%7B2%7D
- https://www.monroecc.edu/faculty/paulseeburger/calcnsf/CalcPlot3D/
- https://youtu.be/sDv4f4s2SB8

- https://ocw.mit.edu/courses/mathematics/18-02sc-multivariable-calculus-fall-2010/2.-partial-derivatives/part-a-functions-of-two-variables-tangent-approximation-and-optimization/session-25-level-curves-and-contour-plots/level-curves/

- https://ocw.mit.edu/courses/mathematics/18-02sc-multivariable-calculus-fall-2010/2.-partial-derivatives/part-b-chain-rule-gradient-and-directional-derivatives/session-38-directional-derivatives

- for « compute the gradient » problem One trick with problems like this one is to expand the vector operations. Something like $e^{\wedge}[-(T\_1^{\star}x\_1 + T\_2^{\star}x\_2)]$ and then you choose the correct variables to calculate the gradient.

- why some are xj an others are just x ? Because theta*x is a scalar product. xj is a vector component: x and theta are vectors. theta*x = theta_1*x_1 + theta_2*x_2+...+theta_n*x_n = Σ theta_j * x_j

- Remember that both x and θ are vectors and the exponent is an inner product.you can write explicitly the inner product; thus:

$$L(\vec{x}, \vec{\theta}) = \log\left[1 + \exp\left(-\vec{x} \cdot \vec{\theta}\right)\right] = \log\left[1 + \exp\left(-(x_1\theta_1 + x_2\theta_2 + \cdots + x_n\theta_n)\right)\right]$$

Now differentiate w.r.t. θ_j .

In general, for differentiating multivariate functions, there are two possibilities: you either know the specific rules of differentiation for multivariate calculus (which you can find on the internet or in any multivariate calculus text) or expand the vector and matrix operations and then differentiate.

- For Geometric Picture of the Function : For function $z=f(x,y)$, level curve of value c is the curve in x, y space, such that $f(x,y)=c$. For different values of c you get different level curves. See this to get some intuition: https://en.wikipedia.org/wiki/Level_set#/media/File:Himmelblau_contour.svg

- For Gradient Ascent or Descent: sign of x. x is a vector. $x = [x_1, x_2, \dots, x_n]^T$ If of dimension 1 (in which case a real number), its sign would not matter.

- Gradient Ascent or Descent : If we dont know what the function is (convex or concave), how can we answer this question? I mean, convex and concave functions will give different answers right? ∇f always moves in the direction of steepest ascent while $-\nabla f$ moves in the direction of steepest descent.