matrix algebra

- roughly, we've seen linearity to be something where we have variables only raised to the first power. We've seen that linear functions are convenient for inverting → this is because we have nice one-to-one mappings, instead of many to one, like we see in quadratic and trig functions.
- As we've seen (but not emphasized), a lot of functions are "locally" linear. That is we can write them as $f(x+\Delta x)-f(a)=\Delta f \approx \ f'(a)\Delta x$ (with error $\epsilon \to 0$ as $\Delta x \to 0$). That is, if f is continuously differentiable at $x=a$, then locally (i.e. near $x=a$), f behaves linearly, that is $f(x) \approx f(a)+f'(a)(x-a)$. (that is, $f(x)$ is a approximately a linear function of x)
- In the 2-variable case, if u is a continuously differentiable function of x and y near (x_0, y_0) then $\Delta u \approx u_x(x_0,y_0)\Delta x + u_y(x_0,y_0)\Delta y$. (the error terms go to 0 since its continuously differentiable) → this is again the same local linearity as we saw in the 1-variable case.
- More generally, if $w=f(x_1,x_2,..x_n)$, then if w is continuously differentiable at $\vec{x}=\vec{a}$, then $\Delta w \approx \sum_{k=1}^{n} f_{x_k}(\vec{a})\Delta x_k$.
- Formally, a linear system of equations is a system where all the variables are raised to the first power. The following linear system has m equations with n unknowns:

  $a_{11}x_1+a_{12}x_2+...a_{1n}x_n=b_1$

  .

  .

  .

  $a_{m1}x_1+a_{m2}x_2+....+a_{mn}x_n=b_n$
- Solutions of such systems are controlled by the coefficients a_{ij} and the constants b_k.
- We can capture the coefficients in what we call a "matrix". By an m by n matrix, we mean a rectangular array of numbers arranged in m rows and n columns. (Note its rows $\times$ columns)
- Matrix multiplication is defined as follows.
  - The product of a $m \times n$ matrix and a $n \times p$ matrix is a $m \times p$ matrix. For two matrices to be compatible for multiplication, the number of columns of the first must be equal to the number of rows of the second and that's the only requirement.
  - Dot the i^{th} row of the first matrix with the j^{th} column of the second matrix to obtain the term in the i^{th} row and j^{th} column of the product matrix.
  - This sounds complex in words but is actually quite straightforward.
  - In other words, $p_{ij}=\sum_{k=1}^{n} a_{ik}b_{kj}$.
- The reason we invented this definition this way is simply because it turned out to be convenient and very useful for solving systems of linear equations. Since then, its actually been useful in lots of other cases → like training and inference in neural nets!
- A matrix where $m=n$ is called square matrix (where there are as many equations as unknowns).
- S_n is the set of all $n \times n$ matrices. Two matrices are only equal if all their entries are the same that is $a_{ij}=b_{ij}$. Addition is also defined term wise.
- The reason matrix product is not defined term by term is because this definition turned out be more useful for solving linear systems and even later it has been useful in other applications.
- For two matrices $AB \neq BA$, in general. (not associative)
  - Addition is associative though.
  - The 0 matrix is the matrix with all entries 0's.

    - Scalar multiplication is defined as usual.
    - Turns out $A(BC) = (AB)C$ and $A(B+C) = AB + AC$.
- The identity matrix is a matrix that's 0 everywhere except the left to right diagnol. You can easily show that multiplying by this preserves the matrix, hence the name. we call the identity matrix I_n
- Also, turns out $AI_n = I_nA = A$
- We can't always find a X such that $AX = I$. Proof: Think of $A = \left[\begin{smallmatrix} 1 & 2 \\ 2 & 4 \end{smallmatrix}\right]$. Let $X = \begin{smallmatrix} a & c \\ b & d \end{smallmatrix}$ be the solution. We can easily show this set of equations does not have a solution. Finding a X such that $AX = I$ is called finding the inverse of A. So we just showed that not all matrices are invertible.
- *so, we must beware of results that require finding the inverse (since not all matrices are invertible).*
- It is also possible that $AB = 0$ and $A \neq 0$ and $B \neq 0$.
- It's also possible that $AB = AC$ and $A \neq 0$ and $B \neq C$.
- A matrix that is invertible (for which A^{-1} exists) is called non-singular. for a non-singular matrix, we can show that since $AX = I$ is solvable, $AB = AC$ means that
  $A^{-1}AB = A^{-1}AC => IB = IC => B = C$.
- If a matrix is not non-singular, we call it singular.
- We will show here a simple theorem that will be very useful going forward. We will only show for the 2 x 2 case that if a matrix's determinant is not 0, only then it has a solution. Consider the matrix $AX = B$. For the 2x2 case, on expanding out $AX = B$ we get two sets of two equations each. The first set has equations with the 2 unknowns being the first column and the second set has equations with the 2 unknowns being the second column. Each set is basically describing two lines. The equations have solutions if the two lines are not parallel, that is if $\frac{a}{b} \neq \frac{c}{d}$. That gives us the condition that $ad - bc \neq 0$, which means the determinants must not be 0. We will show this for the general case other than just 2x2 later, but this is the proof for the 2x2 case.
- So we've shown above for the 2x2 case that if the determinant is 0 then its a singular matrix. The matrix is invertible if and only if the determinant is 0. We haven't shown this for bigger matrices but we will ignore that for a bit and come back to that later.
- Imagine 3 equations with 3 unknowns x_1, x_2, x_3 and the knowns y_1, y_2, y_3. Lets say we have:
  $y_1 = x_1 + x_2 + x_3$
  $y_2 = 2x_1 + 3x_2 + 4x_3$
  $y_3 = 3x_1 + 4x_2 + 6x_3$
- Now we know that we can add one equation to another and multiply. This preserves them. Now if we want to find the x's in terms of the known y's. To do this, we subtract a multiple of equation 1 from equation 2 and 3 so that we get rid of the x_1's in equation 2 and 3. Next we subtract a multiple of equation 2 from equation 3 to get rid of the x_2 in equation 3. Now we're left with x_3 in equation 3, which we can easily compute since its only expressed in terms of the y's. Then, using that we can plug it into equation 2 to find x_2. Then we can use that to find x_1. This process is called row reducing.
- After row reducing, the above system will look something like:

x_1	x_2	x_3	y_1	y_2	y_3
1	0	0	2	2	1
0	1	0	0	3	-2
0	0	1	-1	-1	1

  The matrix on the right side (the last 3 columns) is called the inverse since given $Y = AX$, and since the above system is equivalent to $Y = AX$, the left matrix has been reduced to the identity matrix, we can think of the above process as representing $A^{-1}Y = X$.
- A and A^{-1} represent equivalent systems of equations. If we think of $Y = AX$ as a function from $R^3 \rightarrow R^3$ ($(y_1, y_2, y_3) = f(x_1, x_2, x_3)$), we can think of $X = A^{-1}Y$ as f^{-1}. One takes x's to y's. the other is

like the inverse function.

- When we try to row reduce, it is possible that we get a matrix such that we have all 0's in the third row. In this case, we can't uniquely express the x's in terms of the y's. So we say that the matrix isn't invertible. That is, if the function f isn't invertible then the corresponding A won't have an inverse. In this case we get a constraint in terms of the y's (basically, a relationship between the y's so that the y's are not independent of each other. If an assignment of y's follows the constraint, we will have infinite solutions. If not, that assignment of y's has no solutions.
- Invertible matrices encode a 1 → 1 function. Non-invertible matrices encode a many to one, onto function (where many x's map to a y and some y's don't have a corresponding x) provided the derived constraint is met.
- Now to invert *any* system of equations

  $y_1 = f_1(x_1, ...x_n)$

  .

  .

  .

  $y_2 = f_n(x_1, ...x_n)$

  Now if these are *continuously differentiable*, we can approximate these as

  $\Delta y_k = \frac{\partial y_k}{\partial x_1}\Delta x_1 + ... + \frac{\partial y_k}{\partial x_n}\Delta x_n$etc near $\vec{x} = \vec{a}$.
- Notice how this linearizes the functions locally.
- Now the coefficients are the partial derivatives. We get a matrix of partial derivatives and if its determinant is not 0, it is invertible.
- Now, notice that the linear equations are in terms of the changes (Δ's). Also, since the above is an approximation it is only valid in the neighborhood of $\vec{x} = \vec{a}$.
- This matrix formed by the derivatives is called the Jacobian. (some people call the determinant of this the Jacobian).
- the rigorous proofs for invertibility and determinants and further discussion about the results that follow from the matrix definitions are discussed in the notes in the linear algebra course.
- For a function from $R^n \rightarrow R$ a local maxima (or minima) of f is such that there exists a neighborhood N of $\vec{a}$ such that $f(\vec{a}) \geq f(\vec{x})$ (or $f(\vec{a}) \leq f(\vec{x})$) for every $x \epsilon N$.
- To test for max/min:
  - solve:

    $f_{x_1}(x_1, ...x_n) = 0$

    .

    .

    .

    $f_{x_n}(x_1, ...x_n) = 0$
  - also find where f is not differentiable, not defined etc.
  - check the boundaries of the domain of x
  - then out of the above candidate points, investigate the sign of $f(x_1 + \Delta x_1 ...) - f(x_1, ...x_n)$. If it is always positive, then its a local maxima, if its always negative then its a local minima. Otherwise it's just a saddle point.
- There is a test to help investigate the sign of $f(x_1 + \Delta x_1 ...) - f(x_1, ...x_n)$, based on f_{xx}, f_{yy}, f_{xy}, that can be derived from Taylor's formula for multiple variables → not a hard derivation, just algebraic manipulation of the taylor formula.
- Taylor's formula for two variables: I won't give the full formula here, you can look that up. The idea for the derivation is described below. Other than this idea, its "just" algebraic manipulation.

We start with $F(t) = f(a + th, b + tk)$.
the chain rule gives $F'(t) = f_x \frac{dx}{dt} + f_y \frac{dy}{dt} = hf_x + kf_y$. Since f_x and f_y are continuously differentiable (they have continuous partial derivatives).
F' is a differentiable function of t so
$F'' = \frac{\partial F'}{\partial x}\frac{dx}{dt} + \frac{\partial F'}{\partial y}\frac{dy}{dt} = \frac{\partial}{\partial x}(hf_x + kf_y)h + \frac{\partial}{\partial y}(hf_x + kf_y)k = h^2 f_{xx} + 2hkf_{xy} + k^2 f_{yy}$

We just go on like this and we'll have to use the binomial theorem and we'll get there eventually 😀. I'm not as interested in the result as I am about telling you how we get there.

- Maxima/minima with constraints: we might want to maximize $f(x, y, z)$ subject to the constraint $g(x, y, z) = 0$, which implicitly defines $z = k(x, y)$. So now, Using this, we can express the original expression to minimize $g(x, y, z)$ as $g(x, y, k(x, y))$ which is some function h of x and y which we can minimize. We know that this would obey the constraint since z's dependence is coded into this setup now.